四、教育教学类论文

序号	论文题目	期刊名称	期刊等级	发表时间	对象（主持人/成员及排序）
1	Scalable Swin Transformer network for brain tumor segmentation from incomplete MRI modalities	Artificial Intelligence in Medicine	SCI	2024.03	成员 9/1
2	Scikit-ANFIS: A Scikit-Learn Compatible Python Implementation for Adaptive Neuro-Fuzzy Inference System	International Journal of Fuzzy Systems	SCI	2024.06	成员 9/1
3	Cascade Aggregation Network for Accurate Polyp Segmentation	The Institution of Engineering and Technology	SCI	2025.07	成员 7/1
4	Pedestrian re-recognition based on spatiotemporal Transformer skeleton contrastive learning and feature optimization	Journal of Advanced Computational Intelligence and Intelligent Informatics	SCI	2025.09	成员 7/1
5	A Novel Watermarking Scheme for Audio Data Stored in Third Party Servers	International Journal of Digital Crime and Forensics	SCI	2024.03	成员 7/2

6	CMMF and STAM-FNet: Multimodal Fusion Architectures for Complex Scene Understanding in Dynamic Environments	Informatica (Slovenia)	SCI	2026.03	主持人/1
7	The Representation Learning Ability of Self-Supervised Learning in Unlabeled Image Data	International Journal of Advanced Computer Science and Applications	EI	2025.09	主持人/1
8	A Cross Layer Semantic Enhanced SLU Model With Role Context Differentiated Fusion	ICTAI	EI	2021.11	成员 9/2
9	晚清林译小说中的儒学传统与“新民”视野	河南师范大学学报 (哲学社会科学版)	CSSCI	2024.11	成员 2/1
10	基于 FIRA 仿真的足球机器人预判圆弧射门算法设计	长春工程学院学报(自然科学版)	普刊	2018.09	主持人/1
11	基于 ACM-ICPC 竞赛的 C 语言课程教学实践	安庆师范大学学报(自然科学版)	普刊	2017.03	主持人/1
12	基于单目视觉的足球机器人图像处理系统的畸变矫正研究	蚌埠学院学报	普刊	2018.10	主持人/1
13	基于二级 MS Office 的大学计算机基础课程教学	安庆师范大学学报(自然科学版)	普刊	2018.03	主持人/1

14	“专创融合”视域下 C 语言程序设计课程教学实践探索	创新创业理论研究与实践	普刊	2024.08	主持人/1
15	吸引信阳籍在外人才回乡创新创业策略	农村经济与科技	普刊	2024.05	主持人/1
16	基于成果导向的 C 语言课程教学改革与实践	电子测试	普刊	2020.07	主持人/1
17	独立学院图书馆纸质图书利用率提升策略研究	中国管理信息化	普刊	2018.08	主持人/1
18	专业认证背景下 Java Web 应用开发课程教学改革与创新研究	创新创业理论研究与实践	普刊	2024.12	主持人/2
19	应用型人才培养模式下 Java EE 平台课程教学改革与实践	电子测试	普刊	2021.03	主持人/2
20	疫情防控背景下高校毕业生就业工作探究	中国电力教育	普刊	2020.08	主持人/2
21	基于 RFID 的教室考勤系统设计与实现	电脑编程技巧与维护	普刊	2017.04	主持人/2
22	教育数字化背景下高校研究性教学研究	湖北开放职业学院学报	普刊	2025.06	主持人/1
23	“Π 型人才”视角下的人工智能与大数据专业群建设研究与实践	创新创业理论研究与实践	普刊	2025.03	主持人/2
24	数智驱动下普通高校学士学位授予质量保障机制研究	河南教育(高教)	普刊	2025.03	主持人/1
25	人工智能背景下地方本科高校毕业生高质量就业研究	湖北开放职业学院学报	普刊	2025.11	主持人/1

26	老年人健康饮食管理系统设计与实现	电脑编程技巧与维护	普刊	2025.11	主持人/2
27	产教融合背景下“课赛融合”创新实践教学模式研究	湖北开放职业学院学报	普刊	2026.02	主持人/2
28	人工智能时代地方本科高校高质量就业路径研究	黑龙江工业学院学报(综合版)	普刊	2026.02	主持人/1
29	数智融合下高等教育育人模式创新机制研究	湖北开放职业学院学报	普刊	2026.01	主持人/3
30	民办高校学生学业预警系统的设计与实现	电脑编程技巧与维护	普刊	2025.07	主持人/3
31	新工科背景下大数据应用型人才培养模式研究	教育信息化论坛	普刊	2022.02	成员 9/1
32	面向 DevOps 的政务大数据分析可视化系统	计算机技术与发展	普刊	2020.02	成员 9/1
33	师范生教育实践能力培养创新研究	软件导刊(教育技术)	普刊	2019.06	成员 7/1
34	导师选择系统的设计与实现	信息技术与信息化	普刊	2019.04	成员 7/1
35	新工科背景下地方高校计算机应用型人才培养模式	计算机教育	普刊	2021.11	成员 4/1
36	新工科背景下人工智能课程的教学改革	福建电脑	普刊	2022.04	成员 4/1

1 成员 9 发表的 SCI 论文： Scalable Swin Transformer network for brain tumor segmentation from incomplete MRI modalities

Artificial Intelligence In Medicine 149 (2024) 102788

Contents lists available at ScienceDirect

ArtificialIntelligenceInMedicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

ScalableSwinTransformernetworkforbraintumorsegmentationfrom incompleteMRImodalities

Dongsong Zhang[a,c], Changjian Wang[b], Tianhua Chen[c], Weidao Chen[d], Yiqing Shen[e,*]

[a] School of Big Data and Artificial Intelligence, Xinyang College, Xinyang, 464000, Henan, China [b] National Key Laboratory of Parallel and Distributed Computing, Changsha, 410073, Hunan, China [c] School of Computing and Engineering, University of Huddersfield, Huddersfield, HD13DH, UK [d] Beijing Infervision Technology Co., Ltd., Beijing, 100020, China

[e] Department of Computer Science, Johns Hopkins University, Baltimore, 21218, MD, USA

ARTICLE INFO

Keywords:
Incomplete modality
Brain tumor segmentation
Transformer

ABSTRACT

Background: Deep learning methods have shown great potential in processing multi-modal Magnetic Resonance Imaging (MRI) data, enabling improved accuracy in brain tumor segmentation. However, the performance of these methods can suffer when dealing with incomplete modalities, which is a common issue in clinical practice. Existing solutions, such as missing modality synthesis, knowledge distillation, and architecture-based methods, suffer from drawbacks such as long training times, high model complexity, and poor scalability.

Method: This paper proposes IMS²Trans, a novel lightweight scalable Swin Transformer network by utilizing a single encoder to extract latent feature maps from all available modalities. This unified feature extraction process enables efficient information sharing and fusion among the modalities, resulting in efficiency without compromising segmentation performance even in the presence of missing modalities.

Results: Two datasets, BraTS 2018 and BraTS 2020, containing incomplete modalities for brain tumor segmentation are evaluated against popular benchmarks. On the BraTS 2018 dataset, our model achieved higher average Dice similarity coefficient (DSC) scores for the whole tumor, tumor core, and enhancing tumor regions (86.57, 75.67, and 58.28, respectively), in comparison with a state-of-the-art model, i.e. mmFormer (86.45, 75.51, and 57.79, respectively). Similarly, on the BraTS 2020 dataset, our model scored higher DSC scores in these three brain tumor regions (87.33, 79.09, and 62.11, respectively) compared to mmFormer (86.17, 78.34, and 60.36, respectively). We also conducted a Wilcoxon test on the experimental results, and the generated p-value confirmed that our model's performance was statistically significant. Moreover, our model exhibits significantly reduced complexity with only 4.47 M parameters, 121.89 G FLOPs, and a model size of 77.13 MB, whereas mmFormer comprises 34.96 M parameters, 265.79 G FLOPs, and a model size of 559.74 MB. These indicate our model, being light-weighted with significantly reduced parameters, is still able to achieve better performance than a state-of-the-art model.

Conclusion: By leveraging a single encoder for processing the available modalities, IMS²Trans offers notable scalability advantages over methods that rely on multiple encoders. This streamlined approach eliminates the need for maintaining separate encoders for each modality, resulting in a lightweight and scalable network architecture. The source code of IMS²Trans and the associated weights are both publicly available at https://github.com/hudscomdz/IMS2Trans.

1. Introduction

with different contrast, resulting in multi-modal MRI scans [5]. Common MRI modalities [5] include T1-weighted (T1w), contrast-enhanced

* Corresponding author.
E-mail address: yshen92@jhu.edu (Y. Shen).

In some literature [1–3], 'modality' is also termed as 'sequence'.

https://doi.org/10.1016/j.artmed.2024.102788
Received 12 June 2023; Received in revised form 19 December 2023; Accepted 25 January 2024 Available online 2 February 2024

Magnetic resonance imaging (MRI) is a widely-used non-invasive T1-weighted (T1c), T2-weighted (T2w), Fluid Attenuation Inversion imaging technique for clinical assessment and therapy planning for Recovery (FLAIR), Magnetization Prepared RA-pid Gradient Echo (MPtumors in soft tissues such as the brain [4]. To obtain a compre- RAGE), and Proton Density (PD-w). Considering that invasive growth hensive characterization of the anatomy, MRIs are typically acquired

分节符(连续)

brain tumors are usually fused with brain soft tissues, it is difficult to accurately segment tumor structures using single-modality MRI images. Instead, by providing complementary information, the availability of multi-modal brain MRI data can improve the accuracy of lesion identification, and disease diagnosis for both human and computer-aided diagnosis (CAD) systems such as deep learning models. Consequently, in terms of brain tumor segmentation from MRI, various feature fusion strategies have developed upon convolutional neural network (CNN) [6] or Transformer [7]. Li et al. [8] proposed a dual X-Net codec structure combining the characteristics of CNN and Transformer, which extracts local and global features simultaneously through convolution subsampling and Transformer encoders and then reconstructs the input image itself through variational autoencoder branches in the decoding stage. The experiment shows that X-Net can realize the organic combination of Transformer and CNN. Xu et al. [9] proposed a hybrid feature extraction network, which fully integrates the features extracted by CNN and Transformer to enhance the segmentation performance of brain tumor medical images. Zhu et al. [10] proposed a brain tumor segmentation method that integrates multi-modal MRI information, which combines deep semantic and edge information fusion, using Swin Transformer for feature extraction, CNN-based edge detection module, and multi-feature inference block based on graph convolution. To achieve real-time medical image segmentation, He et al. [11] proposed a cloud-based method based on multi-feature extraction and interactive fusion. The method uses Transformer and CNN to extract global and local features, respectively. The interactive fusion focus module improves segmentation accuracy. Lu et al. [12] proposed a 3D multiscale Ghost convolution neural network (GMetaNet) with an auxiliary MetaFormer decoding path, which combines local modeling of CNN with remote representation of Transformer to achieve efficient semantic information extraction of multi-modal brain tumor MRI images. To address the issue of neural networks using too many parameters and being difficult to deploy, Liu et al. [13] proposed a lightweight 3D brain tumor image segmentation method with hierarchical decoupled convolutions that reduces the number of parameters, which also uses an attention mechanism in the output layer to improve segmentation accuracy.

However, certain factors may lead to missing MRI modalities [14][1], while most of the existing multi-modal deep learning methods are not applicable to address this issue. For example, one possible reason for missing MRI modalities is that patients may fail to comply with instructions from radiologists or clinicians [15], which can compromise the quality of the scans of specific modalities. Another factor is the acquisition time constraints during scanning, due to the cost and considerations of patient comfort [16], which may prevent the collection of all required MRI modalities. Additionally, body movements during the scan can lead to artifacts and unusable low-quality images [17], resulting in the loss of certain modalities. Finally, the change of MRI imaging protocols can also contribute to unaligned or the absence of MRI sequences [18].

To address the missing modalities in multiple-parametric MRI analysis, which is a common issue in brain tumor segmentation, several remedies have been proposed. They can be broadly classified into three categories [5,19]. The first category, as depicted in Fig. 1(a), employs generative models such as Generative Adversarial Network (GAN) [20], Diffusion Models [21] to synthesize the missing modalities from observed MRI modalities as data preprocessing. However, this approach has the drawback of requiring an additional model to be trained before downstream analysis, leading to longer training and execution times, as well as the accumulation of errors. As illustrated in Fig. 1(b), the second category involves using knowledge distillation to extract feature representations from a teacher network trained with full modalities to a student network specifically tailored for missing modalities [22–26]. Yet, distillation-based methods require a series of students to tackle each condition of missing modalities, leading to a huge computation of spatial and time costs. The third category utilizes a single network that can directly handle any conditions of missing modalities for particular downstream tasks [27–35], as shown in Fig. 1(c). However, this line of approaches encounters limitations such as large network parameters, slow training speed, and poor scalability, as they require one encoder for each modality.

Being able to accurately segment brain tumors from MRI scans is crucial for diagnosis, treatment planning, and assessing response to therapy. However, it is common to encounter incomplete modalities in clinical practice due to various factors such as patient non-compliance, time constraints, artifacts, and changes in protocols [14–18]. Existing multi-modal deep learning methods suffer performance declines when confronted with missing modalities [5,19]. To address this critical issue, we propose a novel scalable Swin Transformer network specifically engineered to maintain segmentation performance even when MRI modalities are incomplete or absent. Our method offers a lightweight and efficient solution that requires significantly fewer parameters compared to previous methods that rely on multiple modality-specific encoders. By using a single shared-weight encoder for feature extraction coupled with our proposed data augmentation and distillation techniques, our network provides an important advancement that can effectively and efficiently handle missing modalities while ensuring accurate brain tumor segmentation.

To narrow the gap, we propose Incomplete Modalities Scalable Swin Transformer (IMS^{2}Trans), a novel lightweight and scalable network architecture for incomplete MRI multi-modal brain tumor segmentation. As illustrated in Fig. 1(d), our method employs a single encoder with shared weights to extract features from all the observed modalities along with a feature distillation scheme to impose consistency regularization among modalities, before being finally aggregated to obtain the segmentation result via the decoder. The major contributions are four-fold. (1) To the best of our knowledge, we first propose a scalable Swin Transformer [36] as the only share-weighted encoder for the incomplete MRI modalities of brain tumor segmentation. Specifically, all the available observed modalities are input to an encoder with shared weights to reduce the number of parameters and increase efficiency. Correspondingly, a novel modality token strategy is designed to specify the difference between input modalities. (2) We introduce a Swin-like lightweight MLP bottleneck [37] that not only reduces model parameters but also obtains better feature maps of intra-modalities and inter-modalities. (3) We also design a new feature distillation regularization based on contrastive learning [38] to improve the interchangeability and consistency across different modalities. (4) We propose a novel 3D multimodal version based on the CutMix [39] data augmentation strategy specifically for multi-modal MRI data further to enhance the model robustness against the missing modalities.

[1] We follow the literature by using 'missing modalities' and 'incomplete modality' interchangeably.

between the feature distributions of different modalities [28, 31,32]. One approach has been to use networks that encode each modality individually and provide them with a correlation block. However, this method may not recover lost information if the number of available modalities is insufficient. To address the above-mentioned limitation, recent methods [30,33–35] have proposed using attention mechanisms for brain tumor segmentation tasks with missing modalities. For example, Ding et al. [30] introduced a novel region-aware fusion module that divides multi-modal features into different regions using a trained probability map and then applies modal-wise attention to adjust features from available modalities. Additionally, Zhang et al. [33] proposed mmFormer that combines Transformer blocks and convolutional encoders to build local and global information within each modality and long-range correlations across modalities, representing the first attempt to achieve this using Transformer blocks. Zhou et al. [34] suggested a new multi-modality feature fusion network that uses a self-attention mechanism to learn non-local structures in images across multiple modalities, and a multi-scale fusion module to capture feature information in multi-modality spatial contexts, as well as a spatially consistent underlying feature learning module to learn potential multi-modality correlations. Furthermore, Konwer et al. [35] proposed a new method to address brain tumor image segmentation with incomplete modalities by introducing an auxiliary adversarial learning strategy to supervise the representation of missing modality features during meta-training of partial modality data and meta-testing of limited full modality subjects.

While those methods have made significant improvements to MRI analysis for missing modalities, they still face challenges in scenarios where more than one modality is missing and have higher memory consumption due to the large number of parameters arising from an equal number of encoders to the number of modalities. In contrast, the proposed method has a unique advantage in that it reuses its encoder to encode multiple modalities, and the weights of the encoder are shared. Our approach significantly reduces the number of parameters
the network requires while maintaining performance, making it more memory-efficient and scalable to handle multiple modalities.

3. Methodology

3.1. Overview

In this section, we elaborate on our novel IMS²Trans network, specifically engineered for the segmentation of brain tumors in MRI scans with arbitrary MRI modalities missing. A schematic of this network is provided in Fig. 2. At the core of our network resides a scalable shared-weight encoder that leverages the Swin Transformer block [36] with modality token to capture both local and global context. This shared-weight design enables the efficient encoding of multiple modalities using a single encoder, thus optimizing computational resources and reducing overall model complexity. To further enhance our encoder, we introduce a lightweight Shifted Multi-Layer Perceptron (Shifted MLP) [37] coupled with a masking bottleneck. This combination is designed to balance computational complexity with high-level accuracy, particularly in dealing with missing modalities. We also implement a feature distillation strategy between individual modalities and the entire set of modalities, by comparing the features of each modality with the averaged features computed across all modalities to ensure a more comprehensive and accurate representation of features in a missing modality circumstance. Lastly, to boost the performance and adaptability of our network, especially in the context of missing modalities, we adopt a unique data augmentation technique: the 3D multimodal CutMix (3DMM-CutMix) [39]. This approach strengthens the network's resilience and adaptability, setting the stage for superior performance under varying conditions.

3.2. Scalable shared-weighted encoder

The scalable shared-weighted encoder is designed to process 3D MRI image modalities by a single encoder to reduce the number of parameters and improve efficiency. Concurrently, each modality retains its distinct characteristics, which are ensured by a unique, learnable modality token. Each input MRI modality image, denoted as $X_i \in \mathbb{R}^{H\times W\times D\times 1}$, where $i \in$ {FLAIR, T1C, T1w, T2w} denotes the corresponding modality, is first divided into non-overlapping patches through patch partition operator. Here, H, W, and D signify the height, width, and number of slices in the modality image respectively. The patches, each with a dimension of $2\times2\times2$, result in a feature dimension of 8.

These patch tokens are then transformed into an embedding space of dimension $\mathcal{C} = 24$ using a learnable linear layer. In parallel, the input X_i is directed through a residual block, resulting in an 8-dimensional token.

For each input modality image, a corresponding token is produced, thus there are as many such tokens as there are the number of input modalities. These tokens are then routed to the masking bottleneck, concatenated, and subsequently passed into the decoder. Post this, the embedded feature map of input X_i is fed into the encoder, which consists of three layers of the swin transformer. Each of these layers encompasses two transformer blocks with modality tokens and is succeeded by a patch merging module. Notably, the modality token denoted as $MT_i, i \in$ {FLAIR, T1c, T1w, T2w}, representing the embedded feature of each modality, plays a crucial role in retaining the distinctive characteristics of each modality. The size of MT_i in each block is the same as that of the input X_i. The patch merging module reduces the feature dimensions, thereby promoting efficient computation and enabling hierarchical feature extraction. Moreover, the patch merging module combines patches of resolution $2\times2\times2$ and concatenates them, forming a $4\mathcal{C}$-dimensional feature embedding. This is further condensed to a $2\mathcal{C}$-dimensional feature size by another linear layer. As a result, the resolutions after the first, second, and third
swin transformer layers become $\frac{H}{4}\times\frac{W}{4}\times\frac{D}{4}$, $\frac{H}{8}\times\frac{W}{8}\times\frac{D}{8}$, and $\frac{H}{16}\times\frac{W}{16}\times\frac{D}{16}$
respectively, while the corresponding channel number of the embedding space $\mathcal{C}$ incrementally increases to 48, 96, and 192 respectively. To further enhance computational efficiency and maintain critical feature characteristics, a convolutional layer is tactically positioned between each pair of consecutive swin transformer layers, thereby facilitating the down-sampling of the feature map.

3.3. Swin transformer block with modality token

Each input modality image comes with its own information about which modality it is, from which a corresponding token can be designed to represent the embedded features of each modality. As a result, the encoder's architecture, featuring the swin transformer block with the novel modality token, is illustrated in Fig. 3(a). Each stage of the encoder in this figure corresponds one-to-one to each stage of the shared-weight modal-specific encoder in Fig. 2. Following the design of swin transformer [36], the swin transformer block with modality token primarily consists of two consecutive swin transformers, each comprising layer normalization (LN) modules, a multi-head self-attention module, and a multi-layer perceptron (MLP) with GELU nonlinear activation. Each of these components is linked via a residual connection with DropPath [52]. The first and second consecutive swin transformers employ window-based (W-MSA) and shifted window-based (SW-MSA) multi-head self-attention modules, respectively. These modules are essentially variations of the regular and window-partitioning multi-head self-attention modules. According to Fig. 2, the modality tokens of the four input mode images are MT_{FLAIR}, MT_{T1c}, MT_{T1w}, and MT_{T2w}. Consider x^j_i and w^j_i to represent the feature token of the jth modality and the corresponding modality token feeding the swin transformer block at each encoder stage,

3.4. Shifted MLP and masking bottleneck

Before being fed in the Shifted MLP bottleneck, features with respect to each modality extracted by the swin transformer encoder are first encoded into tokens separately. These tokens are then passed onto the Shifted MLP module, as depicted in Fig. 2. They are first performed an axis shift across the width of the tokens, where the locality of the token is integrated into the global axial attention computation, promoting more spatially aware features. Subsequently, tokens are funneled through a parameter-efficient depth-wise convolutional layer (DWConv) [53], after which they pass through a GELU activation layer [54], thus adding a level of learnable non-linearity. The tokens are then channeled to the second multilayer perceptron following an additional axis shift, this time along the height. The final step within the Shifted MLP bottleneck involves applying residual concatenation to append the original tokens as residuals.

Residing after the Shifted MLP bottleneck, the masking bottleneck aims to enhance the robustness of the missing modalities and where the construction of long-distance dependencies between different modalities is required. Specifically, it achieves this by generating a novel multi-modal token through the concatenation of feature maps from various modalities corresponding to the same image. These feature maps originate either from the skip connection of the encoder or the shifted MLP bottleneck. This operation is formally defined as:

$$\mathcal{T}_{token} = [\delta_{FLAIR} \cdot \mathcal{T}_{FLAIR}, \delta_{T1c} \cdot \mathcal{T}_{T1c}, \delta_{T1w} \cdot \mathcal{T}_{T1w}, \delta_{T2w} \cdot \mathcal{T}_{T2w}], \quad (2)$$

where $\delta_m \in \{0,1\}$ functions as a Bernoulli indicator, designed to impart robustness during the construction of long-range dependencies between different modalities, even when some modalities are missing. This form of modality-level dropout is randomly enacted during training by assigning δ_m a value of 0. Following the design in [33], the Bernoulli factor assumes a value of either 0 or 1. If δ_m equals 0, it implies the corresponding input modality is missing. Conversely, a δ_m value of 1 signifies the availability of the modality. During the training stage, multi-modal tokens for missing modalities are produced by assigning this Bernoulli factor a value of 0, effectively leading to multiplication with a zero vector.

3.5. Decoder mechanism

The primary of the decoder is to effectively reconstruct the spatial resolution of the consolidated latent space back into the original image space. The implementation of the decoder primarily relies on convolutional neural networks, comprising a series of 3D residual convolutions and upsampling operators [33]. Specifically, the output feature map generated from the Shifted MLP bottleneck is supplied as the decoder's input. In parallel, the decoder receives skip connections from the masking bottleneck. This dual-input approach enables the preservation of low-level details, thereby facilitating a more precise segmentation. Features derived from different modalities at varying stages of the encoder, once processed through the masking bottleneck, are concatenated and used as inputs to the convolutional layer in the decoder as skip connection features. This strategy serves a dual purpose: Firstly, it compels the decoder to generate accurate segmentation results based on low-level feature maps at each stage of the decoder. Secondly, it equips the model with the resilience to continue delivering accurate segmentation, even when confronted with the presence of missing modalities.

3.6. Feature distillation

We innovatively integrate a feature distillation regularization based on the contrastive learning [38], which ensures uniformity among the features obtained from different modalities through the masking bottleneck, as visualized in Fig. 3(b). More specifically, we generate an averaged feature map $\mathcal{Z}_a$ by computing the mean of the feature maps from four separate modalities of the same image for a given case, X_a: $\mathcal{Z}_a^{FLAIR}$, $\mathcal{Z}_a^{T1c}$, $\mathcal{Z}_a^{T1w}$, $\mathcal{Z}_a^{T2w}$, where $\mathcal{Z}$ represents the latent feature representation, the superscript denotes the corresponding modality. As the averaging process preserves the semantic essence of the image, $\mathcal{Z}_a$ serves as a guiding element for each individually extracted feature from each modality. The central objective is to maximize the similarity between $\bar{\mathcal{Z}}_a$ and each of the feature maps $(\mathcal{Z}_a^{FLAIR}, \mathcal{Z}_a^{T1c}, \mathcal{Z}_a^{T1w}, \mathcal{Z}_a^{T2w})$, a measurement facilitated through the use of the cosine similarity. Conversely, feature maps drawn from different input MRI images, for instance, X_b, should present a significant divergence from $\bar{\mathcal{Z}}_a$. In the training phase, we form positive sample pairs by combining the four input modalities from the same original MRI image with the averaged image representation, yielding at least four positive sample pairs. We generate negative samples by pairing feature maps from different input MRI images within the same batch with the averaged image representation, thereby resulting in a minimum of four negative sample pairs.

Consequently, we can formulate the feature distillation loss L_{FDC} as:

$$L_{FDC} = -\frac{1}{N} \sum_{j=1}^{N} \log P(\mathcal{Z}_i, \mathcal{Z}_j) \quad (3)$$

Here, $P(\mathcal{Z}_j | \mathcal{Z}_i)$ is defined as the normalized temperature-scaled softmax cosine similarity [38,55], as follows:

$$P(\mathcal{Z}_j | \mathcal{Z}_i) = \frac{e^{sim(\mathcal{Z}_i, \mathcal{Z}_j)/\tau}}{\sum_{k=1}^{2N} e^{sim(\mathcal{Z}_i, \mathcal{Z}_k)/\tau}} \quad (4)$$

In these equations, $\mathcal{Z}_j$ represents the feature map of a singular modality from the MRI image X_i. $\mathcal{Z}_i$ is a feature map created by averaging the four modalities from the same image X_i. $\mathcal{Z}^k$ denotes the feature map of all samples related to $\mathcal{Z}_i$, including negative samples from distinct input images. The function $sim(\cdot,\cdot)$ symbolizes the cosine similarity, i.e., the dot product post-normalization. τ is the temperature, and N signifies the number of positive samples connected to $\mathcal{Z}_i$. It is important to note that both $\mathcal{Z}_j$ and $\mathcal{Z}_i$ originate from the same input MRI image, though they provide divergent feature representations of the said image.

3.7. 3DMM-CutMix

To further improve the training efficiency and performance, we propose a data augmentation strategy 3DMM-CutMix, which aims to mix two examples by replacing image regions with patches of another training image and interpolating image labels, allowing the model to focus not only on the most discriminative parts of the image but also on the entire image. The proposed 3DMM-CutMix augmentation method is designed to create a pair of new training samples, $(\tilde{X}_a, \tilde{G}_a)$ and $(\tilde{X}_b, \tilde{G}_b)$, by blending two training samples, (X_a, G_a) and (X_b, G_b). The model is then trained with this newly generated training sample set. The composition operation can be formalized as follows:

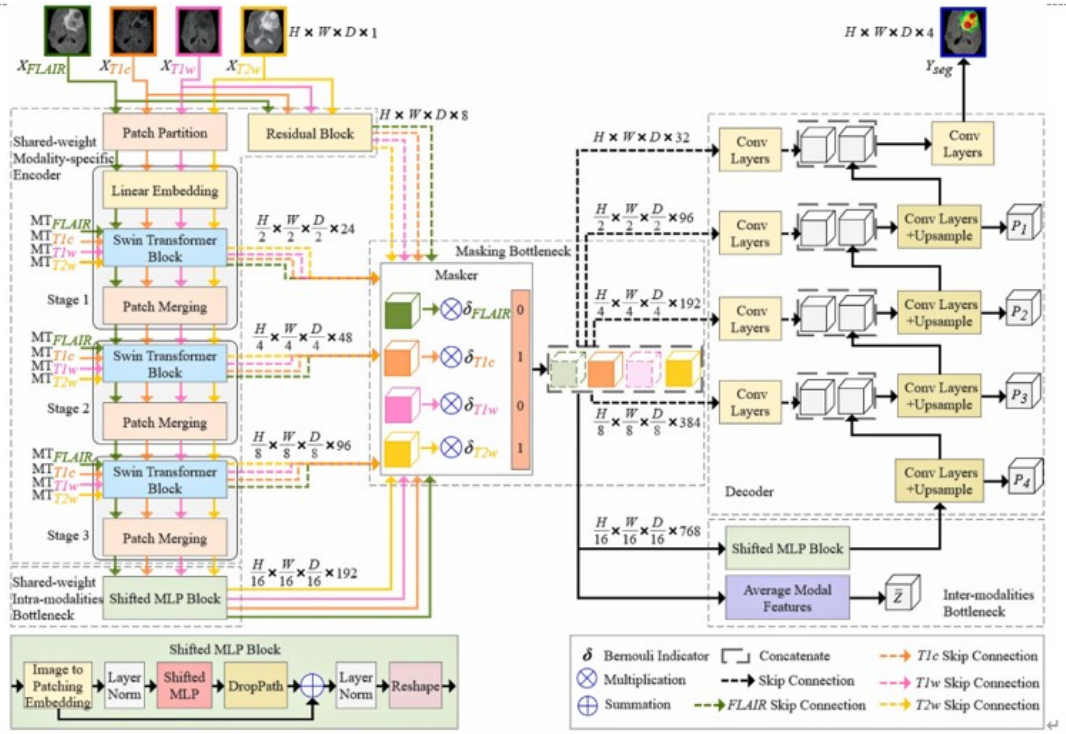


Fig. 2. Overview of the proposed IMS²Trans network. It comprises a scalable shared-weight modality-specific encoder, intra-, and inter-modalities shifted MLP bottlenecks, a masking bottleneck, a multimodal feature distillation module, and a convolutional decoder. Skip connections are strategically applied to the encoder, masker, and decoder to enhance feature learning. For clarity in display, the 3DMM-CutMix data augmentation technique applied to the multimodal input image is not depicted.

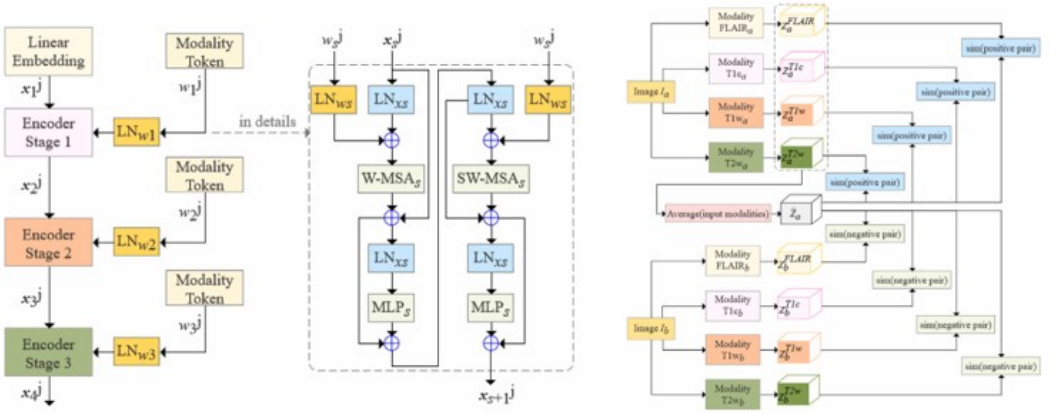


Fig. 3. The semantic illustration of the modality token and feature distillation regularization.

respectively. As per the window division mechanism, the swin transformer block with a modality token can be formally expressed as follows:

$$\hat{x}^{j}_{s} = \text{W-MSA}(\text{LN}(x^{j}_{s}) + \text{LN}(w^{j}_{s})) + x^{j}_{s}$$

$$x^{j}_{s} = \text{MLP}(\text{LN}(\hat{x}^{j})) + \hat{x}^{j}_{s} \quad (1)$$

$$\hat{x}^{j}_{s+1} = \text{SW-MSA}(\text{LN}(x^{j}_{s}) + \text{LN}(w^{j}_{s})) + x^{j}_{s}, \quad x^{j}_{s} = \text{MLP}(\text{LN}(\hat{x}^{j})) + \hat{x}^{j}_{s},$$

where $\hat{x}^{j}_{s}$ and $\hat{x}^{j}_{s+1}$ denote the outputs from the W-MSA and SW-MSA modules respectively, whereas x^{j}_{s} and x^{j}_{s+1} represent the outputs of the MLP module in the first and second transformers, respectively.

3.4. Shifted MLP and masking bottleneck

Before being fed in the Shifted MLP bottleneck, features with respect to each modality extracted by the swin transformer encoder are first encoded into tokens separately. These tokens are then passed onto the Shifted MLP module, as depicted in Fig. 2. They are first performed an axis shift across the width of the tokens, where the locality of the token is integrated into the global axial attention computation, promoting more spatially aware features. Subsequently, tokens are funneled through a parameter-efficient depth-wise convolutional layer (DWConv) [53], after which they pass through a GELU activation layer [54], thus adding a level of learnable non-linearity. The tokens are then channeled to the second multilayer perceptron following an additional axis shift, this time along the height. The final step within the Shifted MLP bottleneck involves applying residual concatenation to append the original tokens as residuals.

Residing after the Shifted MLP bottleneck, the masking bottleneck aims to enhance the robustness of the missing modalities and where the construction of long-distance dependencies between different modalities is required. Specifically, it achieves this by generating a novel multi-modal token through the concatenation of feature maps from various modalities corresponding to the same image. These feature maps originate either from the skip connection of the encoder or the shifted MLP bottleneck. This operation is formally defined as:

$$T_{token} = [\delta_{FLAIR} \cdot T_{FLAIR}, \delta_{T1c} \cdot T_{T1c}, \delta_{T1w} \cdot T_{T1w}, \delta_{T2w} \cdots T_{T2w}], \quad (2)$$

where $\delta_m \in \{0,1\}$ functions as a Bernoulli indicator, designed to impart robustness during the construction of long-range dependencies between different modalities, even when some modalities are missing. This form of modality-level dropout is randomly enacted during training by assigning δ_m a value of 0. Following the design in [33], the Bernoulli factor assumes a value of either 0 or 1. If δ_m equals 0, it implies the corresponding input modality is missing. Conversely, a δ_m value of 1 signifies the availability of the modality. During the training stage, multi-modal tokens for missing modalities are produced by assigning this Bernoulli factor a value of 0, effectively leading to multiplication with a zero vector.

3.5. Decoder mechanism

The primary of the decoder is to effectively reconstruct the spatial resolution of the consolidated latent space back into the original image space. The implementation of the decoder primarily relies on convolutional neural networks, comprising a series of 3D residual convolutions and upsampling operators [33]. Specifically, the output feature map generated from the Shifted MLP bottleneck is supplied as the decoder's input. In parallel, the decoder receives skip connections from the masking bottleneck. This dual-input approach enables the preservation of low-level details, thereby facilitating a more precise segmentation. Features derived from different modalities at varying stages of the encoder, once processed through the masking bottleneck, are concatenated and used as inputs to the convolutional layer in the decoder as skip connection features. This strategy serves a dual purpose: Firstly, it compels the decoder to generate accurate segmentation results based on low-level feature maps at each stage of the decoder. Secondly, it equips the model with the resilience to continue delivering accurate segmentation, even when confronted with the presence of missing modalities.

3.6. Feature distillation

We innovatively integrate a feature distillation regularization based on the contrastive learning [38], which ensures uniformity among the features obtained from different modalities through the masking bottleneck, as visualized in Fig. 3(b). More specifically, we generate an averaged feature map $\bar{Z}_a$ by computing the mean of the feature maps from four separate modalities of the same image for a given case, X_a: $\bar{Z}_a = \frac{Z_a^{FLAIR} + Z_a^{T1c} + Z_a^{T1w} + Z_a^{T2w}}{4}$, where Z represents the latent feature representation, the superscript denotes the corresponding modality. As the averaging process preserves the semantic essence of the image, $\bar{Z}_a$ serves as a guiding element for each individually extracted feature from each modality. The central objective is to maximize the similarity between $\bar{Z}_a$ and each of the feature maps $(Z_a^{FLAIR}, Z_a^{T1c}, Z_a^{T1w}, Z_a^{T2w})$, a measurement facilitated through the use of the cosine similarity. Conversely, feature maps drawn from different input MRI images, for instance, X_b, should present a significant divergence from $\bar{Z}_a$. In the training phase, we form positive sample pairs by combining the four input modalities from the same original MRI image with the averaged image representation, yielding at least four positive sample pairs. We generate negative samples by pairing feature maps from different input MRI images within the same batch with the averaged image representation, thereby resulting in a minimum of four negative sample pairs.

Consequently, we can formulate the feature distillation loss L_{FDC} as:

$$L_{FDC} = -\frac{1}{N} \sum_{i=1}^{N} \log P(Z_i | \bar{Z}_i), \quad (3)$$

Here, $P(Z_i | \bar{Z}_i)$ is defined as the normalized temperature-scaled softmax cosine similarity [38,55], as follows:

$$P(Z_i | \bar{Z}_i) = \frac{e^{sim(Z_i, \bar{Z}_i)/\tau}}{\sum_{k=1}^{2N} e^{sim(Z^k, \bar{Z}_i)/\tau}}, \quad (4)$$

In these equations, Z_i represents the feature map of a singular modality from the MRI image X_i. $\bar{Z}_i$ is a feature map created by averaging the four modalities from the same image X_i. Z^k denotes the feature map of all samples related to $\bar{Z}_i$, including negative samples from distinct input images. The function $sim(\cdot,\cdot)$ symbolizes the cosine similarity, i.e., the dot product post-normalization. τ is the temperature, and N signifies the number of positive samples connected to $\bar{Z}_i$. It is important to note that both Z_i and $\bar{Z}_i$ originate from the same input MRI image, though they provide divergent feature representations of the said image.

3.7. 3DMM-CutMix

To further improve the training efficiency and performance, we propose a data augmentation strategy 3DMM-CutMix, which aims to mix two examples by replacing image regions with patches of another training image and interpolating image labels, allowing the model to focus not only on the most discriminative parts of the image but also on the entire image. The proposed 3DMM-CutMix augmentation method is designed to create a pair of new training samples, $(\tilde{X}_d, \tilde{G}_d)$ and $(\tilde{X}_b, \tilde{G}_b)$, by blending two training samples, (X_d, G_d) and (X_b, G_b). The model is then trained with this newly generated training sample set. The composition operation can be formalized as follows:

$$\begin{cases} G_a = \lambda \cdot G_a + (1-\lambda) \cdot G_b \\ = G_b \\ G_d = \mathbf{M} \odot X_a^j + (\mathbf{1}-\mathbf{M}) \odot X_{bj}^j \\ = \lambda \cdot G_b + (1-\lambda) \cdot G_b \\ = G_b \\ \tilde{X}_{aj} = \mathbf{M} \odot X_a^j + (\mathbf{1}-\mathbf{M}) \odot X_{bj} \end{cases} \quad (5)$$

where $\mathbf{M} \in 0,1^{W \times H \times D}$ serves as a binary mask for the image and the label, where 0 and 1 in $\mathbf{M}$ indicate which regions in the two images are dropped or retained. The symbol $\mathbf{1}$ denotes a binary mask filled entirely with ones. The symbol $\odot$ represents element-wise multiplication between two vectors. Mirroring the CutMix approach [39], we use λ to denote the combination ratio between two data samples. This is determined by a beta distribution, specifically $\mathbf{Beta}[\alpha, \alpha]$. As we have set the parameter α to 1, λ effectively follows a uniform distribution between 0 to 1, i.e., $\mathcal{U}[0,1]$. Note that the label $\tilde{G}$ in the new training example pair, as a combination of the labels G_a, G_b, and λ, does not have to be actually generated, which helps our network to focus on real natural images.

The binary mask $\mathbf{M}$ with respect to the volumetric data is obtained by applying a 3D bounding box $\mathbf{B}=(c_w, c_h, c_d, x_w, x_h, x_d)$ to the cropped regions of two training MRI input images X_a^j and X_b^j within the jth modality. The binary mask $\mathbf{M}$ is constructed such that any position within $\mathbf{B}$ is set to 0, and to 1 otherwise. Effectively, the region $\mathbf{B}$ within X_a^j is excised and replaced with the corresponding cropped region $\mathbf{B}$ from X_b^j, and reciprocally, the region $\mathbf{B}$ in X_b^j is replaced by the region $\mathbf{B}$ extracted from X_a^j. Similarly, the region $\mathbf{B}$ in Y_a and region $\mathbf{B}$ in Y_b undergo the same swapping operation. Formally, $\mathbf{B}$ is determined based on the 3D image coordinates and can be described by six parameters: 3D cuboid size (c_w, c_h, c_d) and 3D center location (x_w, x_h, x_d). The coordinates of the 3D bounding box are uniformly sampled according to the following scheme:

$$c_w = W\sqrt{1-\lambda}, \quad c_h = H\sqrt{1-\lambda}, \quad c_d = D\sqrt{1-\lambda}, \quad (6)$$
$$x_w \sim \mathcal{U}[0, W], \quad x_h \sim \mathcal{U}[0, H], \quad x_d \sim \mathcal{U}[0, D].$$

Note that the scale of the cropped area is consistently maintained across all three dimensions, i.e., $\lambda = 1 - \frac{c_w c_h c_d}{W \cdot H \cdot D}$.

During each training iteration, a pair of 3DMM-CutMix samples $(\tilde{X}_a^j, \tilde{G}_a)$ and $(\tilde{X}_b^j, \tilde{G}_b)$ is produced by merging two randomly selected training samples from a mini-batch, as per the formulas provided in Eqs. (5) and (6).

The implementation of our 3DMM-CutMix is detailed in Algorithm 1, where n, M, C, and K denote the batch size, the count of input modalities, the channel size of the input image, and the total segmentation classes respectively. At every training iteration, we extract a minimum batch of data (X, G) from the training set. This batch has been randomly cropped to ensure uniformity in the size of all input images. The main procedure of 3DMM-CutMix, encapsulated between lines 3 and 13, is straightforward to implement. 3DMM-CutMix generates new data $(\tilde{X}, \tilde{G})$ by randomizing the order of the minibatch input images and labels along the first axis of the tensors. Then, we sample the mixing ratio λ and the 3D bounding box $\mathbf{B}=(c_w, c_h, c_d, x_w, x_h, x_d)$ and the resulting cropping region $(w1, w2, h1, h2, d1, d2)$ as detailed from lines 4 to 11. Subsequently, we mix the input image X and $\tilde{X}$ by replacing the cropped region of X with that of $\tilde{X}$. Note that the target labels G and $\tilde{G}$ are not mixed and still denote the target labels of the two samples before being mixed, respectively. Furthermore, we adjust the value of λ to precisely match the pixel ratio. The augmented input data X is then fed into the model to obtain the predicted segmentation output Y_{seg}. The final step involves computing the loss between the predicted Y_{seg} and the target labels G and $\tilde{G}$, followed by deriving the loss L_{seg} via a weighted summation based on the λ value in next section for details.

Algorithm 1 Pseudo Code for 3DMM-CutMix

Input: Random cropped training set $\{(X_j^i, G_j^i)\}_{j=1}^{M}{}_{i=1}^{N}$

1: **for** each iteration **do**
2: Get a minimum batch of data $(X, G) = \{(X_j^i, G_j^i)\}_{j=1}^{M}{}_{i=1}^{n}$ ▷ X is nxCxWxHxD size tensor, G is nxKxWxHxD size tensor. 3: $\tilde{X}, \tilde{G}$ = ShuffleMinibatch(X, G) ▷ 3DMM-CutMix begins.
4: λ = Beta$(1,1)$
5: $c_w = W\sqrt[3]{1-\lambda}$
6: $c_h = H\sqrt{1-\lambda}$
7: $c_d = D\sqrt[3]{1-\lambda}$
8: $x_w = \mathcal{U}[0, W]$
9: $x_h = \mathcal{U}[0, H]$
10: $x_d = \mathcal{U}[0, D]$
11: $w1, w2, h1, h2, d1, d2$ = GetCoordinate$(c_w, x_w, c_h, x_h, c_d, x_d)$
12: $X[:, :, w1:w2, h1:h2, d1:d2] = \tilde{X}[:, :, w1:w2, h1:h2, d1:d2]$
13: $\lambda = 1 - (w2 - w1) * (h2 - h1) * (d2 - d1)/(W * H * D)$ ▷ 3DMM-CutMix ends.
14: Y_{seg} = ModelForward(X)
15: Loss = ComputeLoss$(Y_{seg}, G, \tilde{G}, \lambda)$
16: Update model
17: **end for**

To guide the network towards predictions more congruous with the real segmentation ground truth, we simultaneously employ Dice Similarity Coefficient (DSC) [56] and Weighted Cross Entropy (WCE)[3] [60] as metrics during the calculation of losses $L_{decoder}$ and L_{seg}. DSC, taking values between 0 and 1, evaluates the similarity between two images. It does so by determining the proportion of twice the number of intersecting voxels to the aggregate number of voxels in the prediction (Y_i) and the ground truth (G_i). This calculation is succinctly defined as follows:

$$\frac{2 \sum_{i}^{N} G_{i,k} Y_{i,k}}{\sum_{i}^{N} G_{i,k}^2 + \sum_{i}^{N} Y_{i,k}^2}$$

3.8. Overall loss

The overall loss function L_{total} is a composite of several individual loss calculations. These include L_{FDC}, corresponding to the feature distillation method, $L_{decoder}$, resulting from the low-level feature map disparities in the decoder, and L_{seg}, which pertains to the final segmentation output. The detailed explanations and formulations for the loss function L_{FDC} are previously provided with Eq. (3).

[3] In certain literature [57–59], WCE is also referred to as weighted softmax loss.

$$L_{Dice}(G_{i,k}, Y_{i,k}) = 1 - \frac{2}{K'}\sum_{k=1}^{K'} \frac{\sum_{N} G_{i,k} Y_{i,k}}{\sum_{N} G_{i,k}^2 + \sum_{N} Y_{i,k}^2 + \epsilon}, \quad (7)$$

where K' signifies the number of segmentation classes, N_k denotes the voxel count of class k, $G_{i,k}$ represents a binary value indicating whether class label k is the appropriate classification for pixel location i, and $Y_{i,k}$ corresponds to the probability of the associated prediction. ϵ is a very small positive number, referred to as the smoothing coefficient, which is configured to 10^{-7} in our experiments.

The issue of a significant discrepancy in the number of classes within an image sample presents an imbalanced classification problem. To address this, the WCE loss is utilized. WCE loss is calculated using a pixel-level softmax activation over the feature maps in tandem with cross-entropy. The predicted probability of class k at each pixel location $i \in \Omega$, where $\Omega \subset Z^3$, is determined by inputting the predicted value into the softmax activation, which is defined as:

$$p_{i,k} = \frac{e^{Y_{i,k}}}{\sum_{k'=1}^{K} e^{Y_{i,k'}}} \quad (8)$$

To effectively manage imbalanced classes, we determine the weight for each class present in the ground truth of the input image. This computation is represented by the following equation:

$$w_k = 1 - \frac{\sum_{i=1}^{N} G_{i,k}}{\sum_{k'=1}^{K}\sum_{i=1}^{N} G_{i,k'}}, \quad (9)$$

where N refers to the total count of voxels in the ground truth, $\sum_{i=1}^{N} G_{i,k}$ signifies the count of pixels pertaining to class k in the ground truth, and $\sum_{k'=1}^{K}\sum_{i=1}^{N} G_{i,k'}$ stands for the aggregate number of pixels encompassing all classes in the ground truth. Subsequently, the WCE loss can be expressed as follows:

$$L_{WCE}(G_{i,k}, Y_{i,k}) = -\frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N} w_k G_{i,k} \log(p_{i,k}). \quad (10)$$

Concurrently, WCE loss dynamically adjusts the weightage of different components in the image, reducing the significance of the background and bolstering the weight of specific internal classifications. This dynamic balancing aids in mitigating the effect of edge voxels on the overall loss. Consequently, both the $L_{decoder}$ loss, associated with the decoder's performance, and the L_{seg} loss, tied to the final segmentation output, are calculated using both Dice and WCE metrics with a weighted summation based on the λ value. They are defined as follows:

$$L_{decoder} = \sum_{s=1}^{S} \left(L_{Dice}(G, P_s)\cdot\lambda + L_{Dice}(G, P_s)\cdot(1-\lambda)\right) + \sum_{s=1}^{S} \left(L_{WCE}(G, P_s)\cdot\lambda + L_{WCE}(G, P_s)\cdot(1-\lambda)\right), \quad (11)$$

$$L_{seg} = L_{Dice}(G, Y_{seg})\cdot\lambda + L_{Dice}(G, Y_{seg})\cdot(1-\lambda) + L_{WCE}(G, Y_{seg})\cdot\lambda + L_{WCE}(G, Y_{seg})\cdot(1-\lambda),$$

where P_s signifies the segmentation prediction generated from the lowlevel feature map at the sth stage of the decoder. G and G' are the target labels of two samples in the new training sample pair after data augmentation by 3DMM-CutMix, respectively. The term S denotes the total number of stages in the decoder, which is set to 4 in our network design.

In conclusion, the overall loss function is formulated as follows:

$$L_{total} = \alpha_d \times L_{decoder} + \alpha_s \times L_{seg} + \alpha_f \times L_{FDC} \quad (12)$$

where α_d, α_s, and α_f are the coefficients of the corresponding loss, respectively.

4. Experiments

4.1. Dataset

The first dataset used for method evaluation is the BraTS 2018 collection [61]. The dataset comprises 285 multi-contrast MRI scans, each containing four distinct MRI modalities: Fluid-Attenuated Inversion Recovery (FLAIR), T1-Contrast Enhanced (T1c), T1-Weighted (T1w), and T2-Weighted (T2w). Given that the annotations for the validation set are not publicly accessible, we partitioned the available 285 images into a training set of 190 images, with the remaining 95 used as a test set. The data split is the same as in [24,33]. We normalized the modalities across all images by resizing each to a resolution of 128 × 128 × 128. Our second dataset for method evaluation is the BraTS 2020 collection [62], which constitutes 369 multi-contrast MRI scans in four different MRI modalities. According to the same partitioning method as [24,33], we divide the available 369 images into 246 training sets and 123 test sets according to a 2:1 ratio column, and the modalities of all images are similarly normalized to 128 × 128 × 128 dimensions.

Each subject in these two datasets is represented through the above four MRI modalities, accompanied by voxel-level segmentation ground truth of three labels: *necrotic and non-enhancing tumor*, *edema*, and *enhancing tumor*. The training target is constructed by merging the three different tumor classes of the ground truth labels. Following the setting in previous work [63], we then converted each segmentation map into three binary maps, corresponding to three tumor categories: *Whole Tumor* (all three tumor classes), *Tumor Core* (all tumor classes excluding *edema*), and *Enhancing Tumor* (limited to the *enhancing tumor* class).

4.2. Implementation

We adopted the Dice Similarity Coefficient (DSC) as our evaluation metric [56]. Our proposed framework and compared methods were implemented with Python 3.8, and PyTorch 1.12 on one NVIDIA Tesla A100 GPU with one Intel Xeon Platinum 8358P CPU. Each input tensor representing a modality has dimensions 128 × 128 × 128 × 1, while the corresponding target tensor size is 128 × 128 × 128 × 4. For data augmentation, we followed previous work [33], applying random flipping, cropping, and intensity shifts. Subsequently, our proposed 3DMM-CutMix approach was also employed for model training. It is important to note that, during model testing, we did not apply any data augmentation. We used the Adam optimizer with an initial learning rate 0.0002 and a weight decay factor 0.0001. We maintain a batch size of 1 throughout our experiments. We set the total training period to 1000 epochs to ensure a fair comparison with other models. Our model is trained for about 78 h with 80G memory on one GPU.

4.3. Parameter analysis

The parameters involved are α_d, α_s, and $\alpha_{f/s}$. We set α_d and α_s to 1.0 following [33]. To determine a suitable value for $\alpha_{f/s}$, which balances the semantic segmentation, decoder, and distillation losses, we evaluated a range of values: 10, 1.0, 0.5, 0.1, and 0.05. The results in Table 1 demonstrate that smaller α_f values yielded higher segmentation performance. This indicates the semantic features may be more important for the Dice Similarity Coefficient metric. Based on these experiments, we set α_f to 0.1 by default to achieve a good trade-off between both metrics.

Table 2
Comparison of results in the BraTS 2018 collection achieved by our proposed IMS^2 Trans and state-of-the-art unified models, including U-HeMIS, U-HVED, RobustSeg, ACN, D^2-Net, and mmFormer. The evaluation is conducted using Dice Similarity Coefficient (DSC) (%) across different combinations of modalities. In the caption, • and ◦ represent available and missing modalities, respectively, and 'F' denotes FLAIR in short.

Modalities				Whole tumor						Tumor core						Enhancing tumor					
F	T1c	T1w	T2w	U-HeMIS	U-HVED	RobustSeg [28]	ACN [25]	D^2-Net [25]	mmFormer	Ours	U-HeMIS	U-HVED	RobustSeg [28]	ACN [25]	D^2-Net [25]	mmFormer	Ours	U-HeMIS	U-HVED	RobustSeg [28]	ACN [25]	D^2-Net [25]	mmFormer	Ours
•	◦	◦	◦	32.48	84.39	85.69	38.2	84.2	85.69	**88.81**	26.06	57.90	53.57	34.1	47.3	**64.45**	64.33	11.78	23.80	25.69	26.5	8.1	**34.88**	31.82
◦	•	◦	◦	61.53	53.62	73.31	41.3	42.8	**79.83**	79.52	65.29	59.59	76.83	35.8	65.1	**81.20**	80.72	62.02	57.64	67.07	28.9	66.3	72.31	**73.17**
◦	◦	•	◦	57.62	49.51	70.11	46.4	15.5	**78.35**	78.27	37.39	33.90	47.90	32.1	16.8	64.75	**64.93**	10.16	8.60	17.29	26.5	8.1	**34.82**	32.46
◦	◦	◦	•	80.96	79.83	82.24	28.9	76.3	85.10	**86.15**	57.20	54.67	57.49	30.7	56.7	**67.90**	67.52	25.63	22.82	28.97	34.2	16.0	**41.06**	36.42
•	•	◦	◦	68.99	85.93	**88.51**	50.9	87.5	87.88	88.09	71.49	75.07	80.62	45.1	80.8	80.93	**81.23**	66.10	68.36	70.30	36.5	64.8	71.99	**74.21**
•	◦	•	◦	64.62	85.71	**88.24**	50.3	87.3	87.96	88.03	41.12	61.14	60.68	45.8	61.6	70.20	**71.08**	10.71	27.96	32.13	36.6	9.5	39.72	**41.63**
•	◦	◦	•	82.95	87.58	88.28	44.7	87.9	**88.69**	88.54	57.68	62.70	61.16	47.2	62.6	**69.83**	69.70	30.22	32.31	33.84	**46.6**	17.4	41.49	**41.83**
◦	•	•	◦	68.47	84.22	77.18	58.6	62.1	**82.49**	82.33	72.46	67.55	78.72	47.7	78.2	81.22	**82.60**	66.22	61.11	69.06	38.8	70.7	75.53	**76.49**
◦	•	◦	•	82.48	81.32	85.19	49.8	84.1	87.80	**88.13**	76.64	73.92	80.20	51.2	80.3	**81.86**	81.75	67.83	67.83	69.71	47.4	68.7	72.58	**73.30**
◦	◦	•	•	82.41	81.56	84.78	55.3	80.1	87.20	**87.89**	60.92	56.26	62.19	47.3	63.2	**72.36**	70.90	32.39	24.29	32.01	**48.6**	16.5	45.56	40.99
•	•	•	◦	72.31	86.72	88.73	63.4	87.7	88.79	**89.02**	76.01	77.05	81.06	52.3	80.9	81.35	**82.58**	68.54	68.68	70.78	42.9	65.7	74.28	**76.03**
•	•	◦	•	83.85	88.09	89.27	56.3	88.8	**89.51**	89.47	77.53	76.75	80.72	56.6	80.7	80.98	**81.47**	68.72	68.93	70.88	53.1	66.4	73.03	**76.19**
•	◦	•	•	83.43	88.07	88.81	62.5	88.4	**89.29**	88.77	68.32	63.14	64.38	56.4	63.7	**72.45**	71.70	31.07	32.34	36.41	**52.5**	19.4	42.72	42.59
•	•	•	•	83.94	82.32	86.01	64.8	80.9	88.26	**88.44**	78.86	75.26	80.33	59.0	79.0	81.85	**82.42**	68.92	67.75	70.10	53.8	68.3	74.26	**76.16**
•	•	•	•	84.74	88.46	89.45	67.6	88.8	89.83	**89.97**	79.48	77.71	80.86	61.7	80.1	81.36	**82.23**	70.24	69.03	71.13	56.6	68.4	73.00	**78.84**
Average				74.05	79.16	84.39	52.5	76.2	86.45	**86.87**	62.57	64.84	69.78	46.9	66.5	75.51	**75.67**	46.10	46.76	51.02	41.8	42.3	57.79	**58.28**

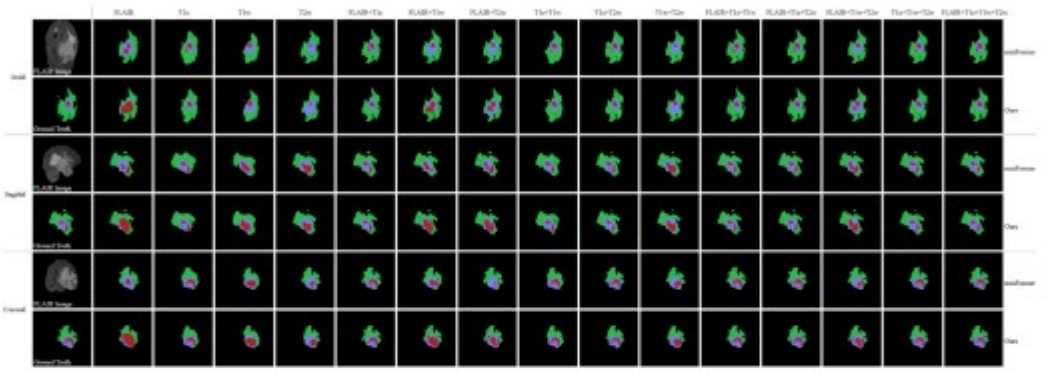


Fig. 4. Visualization of segmentation results under 15 different missing modality conditions for mmFormer and our proposed IMS^{2}Trans compared to ground truth. The results are displayed in axial, sagittal, and coronal slice views. The colors represent different tumor classes: red for necrotic and non-enhancing tumor core, green for edema, and blue for enhancing tumor. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Statistical analysis of results in BraTS 2018 collection achieved by our proposed IMS^{2}Trans and state-of-the-art unified model mmFormer. The evaluation is conducted using p-value of Wilcoxon signed-rank test between mmFormer and IMS^{2}Trans across 15 different combinations of modalities. In the caption, • and ◦ represent available and missing modalities, respectively, and 'F' denotes FLAIR in short.

Modalities				Whole tumor	Tumor core	Enhancing tumor
F	T1c	T1w	T2w	p-value	p-value	p-value
•	◦	◦	◦	0.0327	0.2250	0.9648
◦	•	◦	◦	0.0267	0.4652	0.0999
◦	◦	•	◦	0.6605	0.8266	0.9927
◦	◦	◦	•	0.0014	0.4184	0.7248
•	•	◦	◦	0.0024	0.1347	0.0432
•	◦	•	◦	0.0038	0.6061	0.9818
•	◦	◦	•	0.0026	0.7054	0.5035
◦	•	•	◦	0.5874	0.0480	0.0441
◦	•	◦	•	0.0346	0.3287	0.3027
◦	◦	•	•	0.0086	0.7860	0.9875
•	•	•	◦	0.0014	0.0397	0.1023
•	•	◦	•	0.0004	0.1541	0.0711
•	◦	•	•	0.0022	0.6793	0.9701
◦	•	•	•	0.0834	0.0834	0.0247
•	•	•	•	0.0013	0.0193	0.0017

possible combinations of available modalities, which is consistent with the scenarios of four modalities. Moreover, we provide visualizations of the segmentation results under the seven different missing modality conditions for both mmFormer and IMS^{2}Trans in axial, sagittal, and coronal slice views, as shown in Fig. 5. These visualizations provide additional evidence of the superior segmentation performance of our IMS^{2}Trans model compared to mmFormer in the majority of the seven combinations.

To better showcase the effectiveness of our method, we compared our IMS^{2}Trans to the current leading methods such as U-HeMIS, U-HVED, RobustSeg, RFNet [30], and mmFormer on the BraTS 2020 collection. Note that RobustSeg [28] has not performed experiments on the BraTS 2020 collection, so we use the experimental results from RobustSeg in [30]. Table 5 presents the experimental results, which indicate that our IMS^{2}Trans consistently performs better than U-HeMIS, U-HVED, RobustSeg, RFNet, and mmFormer across all three brain tumor categories, even with 15 possible missing modal combinations. These results demonstrate the effectiveness of our method in accurately segmenting brain tumors and handling missing modalities.

4.5. Comparison of model complexity

Furthermore, in order to provide a comprehensive analysis, we conducted an in-depth analysis to compare the efficiency of IMS^{2}Trans with the state-of-the-art mmFormer, as it achieved the best performance in Table 2. For a fair comparison, we specified the input size as $1 \times 4 \times 128 \times 128 \times 128$. The network configuration analysis is presented in Table 6, encompassing important metrics such as the number of parameters, FLOPs, model size, training speed per epoch, and inference speed per sample. The results demonstrate that our IMS^{2}Trans model exhibits significantly smaller computational and space complexity when compared to mmFormer. Specifically, our model comprises 4.47M parameters, 121.89G FLOPs, and a model size of 77.13MB. In contrast, mmFormer consists of 34.96M parameters, 265.79G FLOPs, and has

Table 4
Results of our proposed IMS²Trans and mmFormer models in BraTS 2018 collection when only three modalities are used for training and testing, with the T1w modality missing, where ▪ and ◦ represent available and missing modalities, respectively, and 'F' denotes FLAIR modality.

Modalities			Whole tumor		Tumor core		Enhancing tumor	
F	T1c	T2w	mmFormer	Ours	mmFormer	Ours	mmFormer	Ours
▪	◦	◦	74.63	**86.08**	43.07	**66.31**	18.04	**32.15**
◦	▪	◦	60.53	**79.14**	56.68	**81.76**	53.94	**73.37**
◦	◦	▪	66.03	**86.07**	44.57	**69.11**	18.47	**38.11**
▪	▪	◦	81.80	**87.71**	66.61	**82.38**	61.04	**76.13**
▪	◦	▪	82.41	**88.39**	54.78	**70.50**	23.08	**40.22**
◦	▪	▪	77.01	**87.86**	66.98	**82.97**	59.50	**74.46**
▪	▪	▪	84.55	**89.15**	69.97	**82.84**	62.51	**75.94**
Average			75.28	**86.34**	57.52	**76.56**	42.37	**58.62**

Table 5
Comparison of results in BraTS 2020 collection achieved by our proposed IMS²Trans and state-of-the-art unified models, including U-HeMIS, U-HVED, RobustSeg, RFNet and mmFormer. The evaluation is conducted using Dice Similarity Coefficient (DSC) (%) across different combinations of modalities. In the caption, ▪ and ◦ represent available and missing modalities, respectively, and 'F' denotes FLAIR in short.

Modalities				Whole tumor						Tumor core						Enhancing tumor					
F	T1c	T1w	T2w	U-HeMIS	U-HVED	RobustSeg [30]	RFNet [30]	mmFormer	Ours	U-HeMIS	U-HVED	RobustSeg [30]	RFNet [30]	mmFormer	Ours	U-HeMIS	U-HVED	RobustSeg [30]	RFNet [30]	mmFormer	Ours
▪	◦	◦	◦	58.72	82.76	82.87	87.32	87.64	**88.47**	37.03	52.42	60.72	69.19	68.44	**69.72**	14.63	25.85	34.68	38.15	**42.20**	40.17
◦	▪	◦	◦	66.92	71.42	71.39	76.77	75.85	**79.31**	74.22	74.93	76.68	81.51	81.18	**82.76**	64.95	68.43	67.91	74.85	70.63	**75.60**
◦	◦	▪	◦	66.35	58.30	71.41	77.16	77.02	76.33	48.57	39.54	54.30	66.02	65.13	**66.74**	20.41	18.21	28.99	37.30	**39.04**	36.01
◦	◦	◦	▪	80.34	82.13	82.20	**86.05**	84.01	86.00	60.83	61.37	61.88	71.02	**71.05**	69.60	32.78	31.86	36.46	**46.29**	44.97	44.50
▪	▪	◦	◦	73.41	87.15	87.33	89.89	89.04	**90.75**	74.62	77.45	81.85	84.65	84.22	**84.96**	69.52	71.24	70.78	76.67	73.70	**77.39**
▪	◦	▪	◦	69.79	86.46	88.10	89.73	89.53	**90.23**	48.19	57.38	68.18	73.07	73.31	**76.08**	16.04	27.94	39.67	40.98	**45.64**	44.75
▪	◦	◦	▪	83.76	87.91	88.09	89.87	89.60	**90.45**	60.21	63.47	68.20	**74.14**	74.06	73.87	30.66	33.64	42.19	49.32	47.88	**49.42**
◦	▪	▪	◦	73.41	74.09	76.84	81.12	79.49	**81.17**	78.35	79.11	80.28	83.40	82.54	**84.39**	71.40	70.79	70.11	**78.01**	73.52	76.75
▪	▪	◦	▪	85.16	85.72	85.97	**87.74**	86.32	87.71	79.84	80.27	82.44	83.45	**84.71**	84.68	71.12	70.48	71.42	75.93	72.08	**77.48**
◦	▪	▪	▪	83.30	84.34	85.53	**87.72**	86.71	87.44	60.80	62.17	66.46	73.13	73.48	**73.95**	29.76	32.37	39.92	45.65	**47.92**	47.32
▪	▪	▪	◦	76.78	86.59	88.87	90.69	89.70	**90.91**	78.88	79.02	82.76	85.07	85.03	**86.12**	71.39	72.16	71.77	76.81	74.47	**78.17**
▪	▪	◦	▪	85.17	88.92	88.68	90.68	90.02	**91.17**	79.24	80.19	81.89	84.97	85.53	**85.54**	71.98	71.72	71.17	77.12	74.46	**77.72**
▪	◦	▪	▪	84.43	88.66	89.24	90.60	90.20	**90.80**	63.48	65.39	70.46	75.19	75.66	**76.39**	32.13	34.48	43.90	49.92	50.00	**50.02**
◦	▪	▪	▪	85.84	85.86	86.63	**88.25**	87.21	88.21	81.56	81.72	82.85	83.47	84.96	**85.57**	72.37	71.92	71.87	76.99	74.09	**78.16**
▪	▪	▪	▪	86.03	89.43	89.47	91.11	90.32	**91.13**	81.03	81.68	82.87	85.21	85.86	**86.04**	72.44	71.87	71.52	78.00	74.87	**78.17**
Average				77.29	82.65	84.17	86.98	86.17	**87.33**	67.12	69.07	72.45	78.23	78.34	**79.09**	49.64	51.53	55.49	61.47	60.36	**62.11**

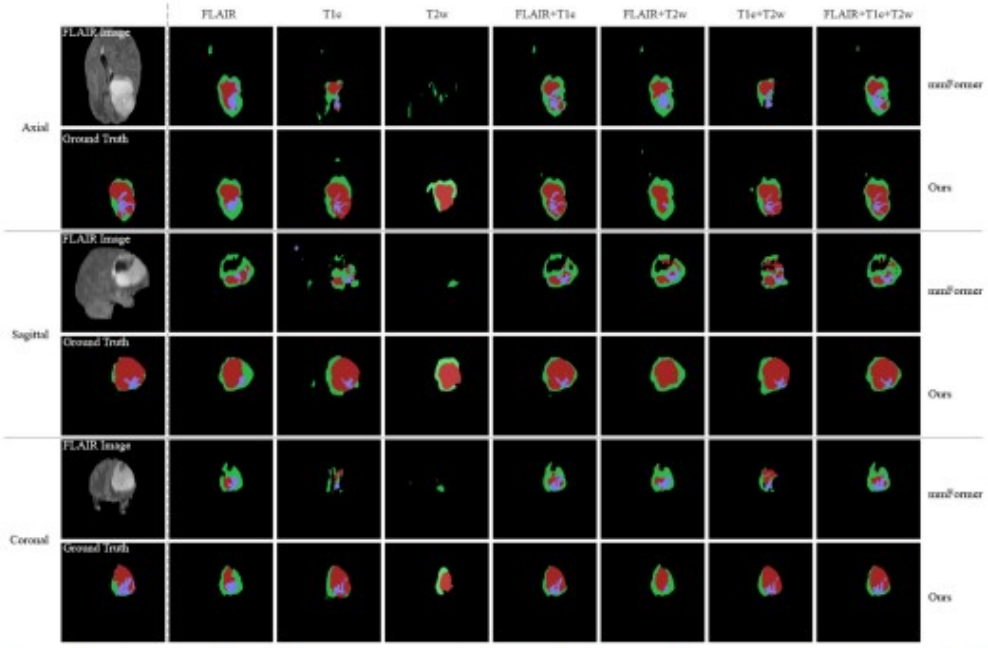
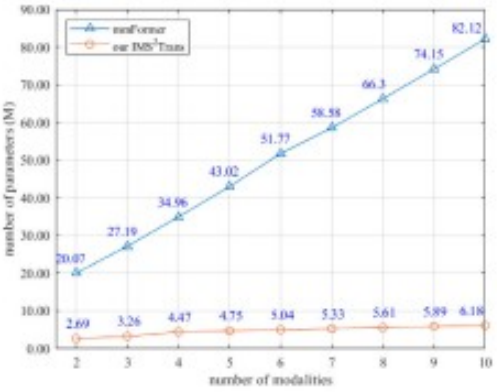


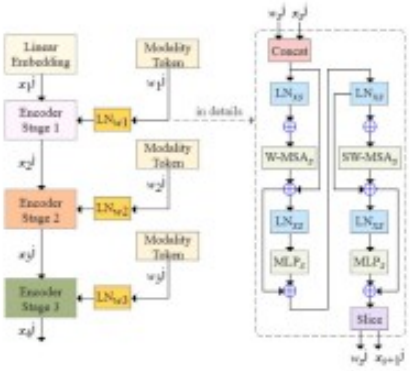
Fig. 5. Visualization of the segmentation results under seven different missing modality conditions for mmFormer and our proposed IMS²Trans compared to the ground truth. The results are shown in axial, sagittal, and coronal slice views. The color scheme represents different tumor classes: red for necrotic and non-enhancing tumor core, green for edema, and blue for enhancing tumor. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
Efficiency comparisons of the proposed IMS²Trans with mmFormer.

Network	Params (M)	FLOPs (G)	Model Size (MB)	Train. Speed (s)	Inf. Speed (ms)
mmFormer	34.96	265.79	559.74	288.91	1892
Ours	4.47	121.89	77.13	269.80	1649



(a) Number of parameters when different number of modalities involved



(b) Concatenate design of modality token

Fig. 6. (a) Comparison of the number of parameters for mmFormer and our proposed IMS²Trans models in terms of different number of involved modalities. (b) Illustration of the concatenate design in the swin transformer block with modality token.

a model size of 559.74MB. These metrics clearly indicate that our IMS²Trans model achieves a notable reduction in both computational complexity and space complexity compared to mmFormer. The smaller computational and space complexity of IMS²Trans provides several advantages. Firstly, it allows for more efficient model training, resulting in faster convergence during the training process. Secondly, it facilitates faster inference speed, enabling real-time or near-real-time applications in clinical settings. Lastly, the reduced model size leads to lower memory requirements, making our model more feasible for deployment on resource-constrained devices or in settings where high-performance computational resources may not be readily available.

We also conducted an analysis to highlight the effect of the number of missing modalities on space complexity. This is illustrated by comparing the number of parameters between the state-of-the-art mmFormer model and our IMS²Trans in terms of handling the different number of modalities, as depicted in Fig. 6(a). As shown in the figure, the number of parameters in the mmFormer model increases linearly as the number of available modalities increases from two to ten, ranging from 20M to 82M. In contrast, our IMS²Trans exhibits a relatively stable number of parameters, varying from 2.69M to 6.18M regardless of the number of available modalities. With a fixed number of parameters, our model remains lightweight and avoids exponential growth in parameter count as the number of modalities increases. This makes our model more scalable and efficient in handling various missing modality scenarios. The relatively stable number of parameters in IMS²Trans is particularly beneficial in real clinical scenarios where a varying number of modalities may be available due to equipment limitations or data acquisition challenges. By maintaining a consistent model size, our approach ensures that the computational resources required for model training and inference remain manageable and practical.

4.6. Ablation study

In order to evaluate the design of modality token, we conducted a comparison between two different strategies in terms of training parameters, GPU performance, and segmentation performance in Table 7. The first design approach, which we adopted in our network, is the additive implementation, as illustrated in Fig. 3(a). An alternative approach is the concatenate implementation as depicted in Fig. 6(b). In the concatenate implementation, the feature token x_i^j of the jth modality and the corresponding modality token w_i^j are first concatenated to form a new token. This new token is then passed through two consecutive window-based multi-head self-attention modules (W-MSA and SW-MSA) within the swin transformer block. Finally, a slicing module is used to extract the updated feature token x_{i+1}^j, discarding the modality token w_i^j in the output. Unlike the additive operation, the concatenate operation does not modify the internal modules of the swin transformer block. Instead, it introduces a new token at the input and extracts a modality token at the output for subsequent removal. In Table 7, the results yield that the additive implementation outperforms the concatenate implementation in all aspects, including training parameters, GPU performance, and average segmentation performance. Therefore, our IMS²Trans network adopts the additive implementation due to its superior performance. By choosing the additive implementation, we ensure that the swin transformer block with modality token effectively captures the interdependencies among modalities and produces more accurate and reliable segmentation results.

To demonstrate the effectiveness of our proposed 3DMM-CutMix method, we provide visualizations of the data augmentation results in Fig. 7. The figure consists of four rows, each representing FLAIR, T1c, T1w, and T2w images, respectively. Within each row, there are six columns depicting different stages of the data augmentation process. In the first column, the source image is displayed, while the second column shows the corresponding ground truth for the source image. The third column represents the reference image, and the fourth column displays the ground truth for the reference image. The fifth column showcases the 3D image cropped from the reference image using the 3DMM-CutMix method. This cropped image serves as a mixing component in the data augmentation process. Finally, in the sixth column, we present the resulting image obtained by blending the source image from the first column with the cropped reference image from the fifth column using the 3DMM-CutMix method. These visualizations provide an intuitive representation of how our 3DMM-CutMix method effectively combines information from the source and reference images

Table 7
Comparison of two design approaches for the modality token, namely the additive and concatenation schemes.

Additive	Concatenate	Params (M)	FLOPs (G)	Model Size (MB)	Whole tumor	Tumor core	Enhancing tumor
✓		4.47	121.89	77.13	86.57	75.67	58.28
	✓	17.12	147.70	279.51	85.85	75.31	58.05

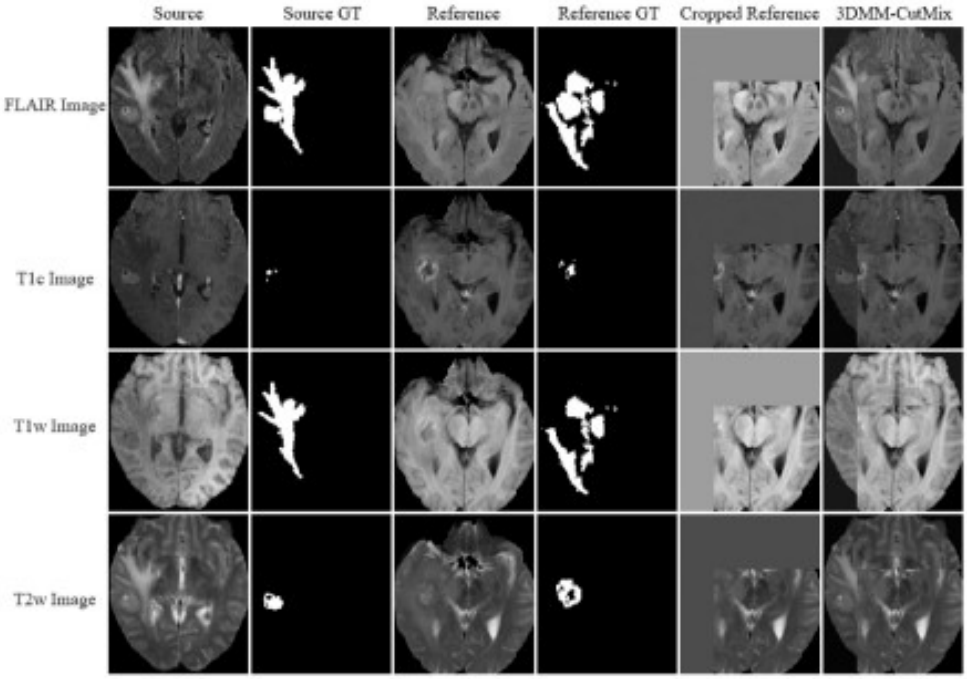


Fig. 7. Visualization of the data augmentation results using our proposed 3DMM-CutMix method. 'GT' denotes ground truth.

to enhance the training process. By generating augmented images that incorporate relevant features from both the source and reference images, the 3DMM-CutMix method promotes better generalization and improves the robustness of the deep learning model for MRI analysis. Overall, the visualization results in Fig. 7 demonstrate the effectiveness and potential of our proposed 3DMM-CutMix method as a valuable data augmentation technique in deep learning-based MRI analysis.

To gain a better understanding of the contributions of different components in our IMS^{2}Tran network, we conducted an ablation study. Specifically, we evaluated the impact of 3DMM-CutMix, swin transformer with modality token, feature distillation loss, shifted MLP bottleneck, and decoder loss on training parameters, GPU performance, and average segmentation performance. The results of this study are summarized in Table 8, where the first column corresponds to the 3DMM-CutMix functional module. When this module is removed, it indicates that there is only one target label for each sample, which can be achieved by setting λ to 1 in Eq. (11). The second column represents the presence or absence of the swin transformer with modality token. If this component is removed, the model solely relies on the swin transformer without modality tokens. The third column indicates the inclusion or exclusion of the feature distillation module. When removed, it implies the absence of the corresponding feature distillation loss, L_{FDC}, in the model. The fourth column signifies the removal of both the intra-modal shifted MLP on the left and the inter-modal shifted MLP on the right in Fig. 2. Lastly, the fifth column pertains to the decoder loss. If this loss is removed, it solely affects the calculation of the total loss. From the results presented in Table 8, it is evident that the absence of the 3DMM-CutMix module leads to a drastic decrease in segmentation performance for all three tumor regions. Additionally, the inclusion of the swin transformer with modality token, feature distillation loss, shifted MLP bottleneck, and decoder loss contribute to performance improvements across the whole tumor region, tumor core region, and enhancing tumor region. These findings demonstrate the crucial roles played by the various components in our IMS^{2}Tran network. The 3DMM-CutMix module significantly enhances the model's ability to handle missing modalities, while the swin transformer with modality token, feature distillation loss, shifted MLP bottleneck, and decoder loss collectively contribute to improved segmentation performance. In addition, Table 8 also shows that except for the reduction of model parameters and GPU performance caused by the swin transformer without modality token, the lack of other modules has little impact. The reduction of model parameters is not large because the shifted MLP bottleneck is a lightweight module. The 3DMM-CutMix and feature distillation modules do not involve learnable parameters, so they have no effect on the number of model parameters. The lack of decoder loss can only reduce some convolutional layers and upsampling layers, so it also has little impact on the number of model parameters. Overall, these ablation study results validate the effectiveness and importance of each component in our proposed network architecture.

4.7. Discussion

Tables 2, 4 and 5 evidently yield that the best-performing existing approaches in brain tumor segmentation, although tailored for the incomplete modalities, still assume the availability of all modal

[5] Azad R, Khosravi N, Dehghanmanshadi M, Cohen-Adad J, Merhof D. Medical image segmentation on mri images with missing modalities: A review. 2022, arXiv preprint arXiv:2203.06217.

[6] Hesamian MH, Jia W, He X, Kennedy P. Deep learning techniques for medical image segmentation: Achievements and challenges. J Digit Imaging 2019;32:582–96.

[7] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020, arXiv preprint arXiv:2010.11929.

[8] Li Y, Wang Z, Yin L, Zhu Z, Qi G, Liu Y. X-Net: A dual encoding–decoding method in medical image segmentation. Vis Comput 2023;39:2223–33.

[9] Xu Y, He X, Xu G, Qi G, Yu K, Yin L, Yang P, et al. A medical image segmentation method based on multi-dimensional statistical features. Front Neurosci 2022;16:1009581.

[10] Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. Inf Fusion 2023;91:376–87.

[11] He X, Qi G, Zhu Z, Li Y, Cong B, Bai L. Medical image segmentation method based on multi-feature interaction and fusion over cloud computing. Simul Model Pract Theory 2023;126:102769.

[12] Lu Y, Chang Y, Zheng Z, Sun Y, Zhao M, Yu B, et al. GMetaNet: Multi-scale ghost convolutional neural network with auxiliary MetaFormer decoding path for brain tumor segmentation. Biomed Signal Process Control 2023;83:104694.

[13] Liu H, Han G, Li Q, Guan X, Tsang M-L. Multiscale lightweight 3D segmentation algorithm with attention mechanism: Brain tumor image segmentation. Expert Syst Appl 2023;214:119166.

[14] Graves MJ, Mitchell DG. Body MRI artifacts in clinical practice: A physicist's and radiologist's perspective. J Magn Reson Imaging 2013;38(2):269–87.

[15] Dale BM, Brown MA, Semelka RC. MRI: Basic principles and applications. John Wiley & Sons; 2015.

[16] Hollingsworth KG. Reducing acquisition time in clinical MRI by data undersampling and compressed sensing reconstruction. Phys Med Biol 2015;60(21):R297.

[17] Chaudhari AS, Sandino CM, Cole EK, Larson DB, Gold GE, Vasanawala SS, et al. Prospective deployment of deep learning in MRI: A framework for important considerations, challenges, and recommendations for best practices. J Magn Reson Imaging 2021;54(2):357–71.

[18] Zimmermann L, Knäusl B, Stock M, Lütgendorf-Caucig C, Georg D, Kuess P. An MRI sequence independent convolutional neural network for synthetic head CT generation in proton therapy. Z für Med Phys 2022;32(2):218–27.

[19] Zhou T, Ruan S, Hu H. A literature survey of MR-based brain tumor segmentation with missing modalities. Comput Med Imaging Graph 2022;102167.

[20] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM 2020;63(11):139–44.

[21] Jiang L, Mao Y, Chen X, Wang X, Li C. Cola-diff: Conditional latent diffusion model for multi-modal MRI synthesis. 2023, arXiv preprint arXiv:2303.14081.

[22] Wang Q, Zhan L, Thompson P, Zhou J. Multimodal learning with incomplete modalities by knowledge distillation. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020, p. 1828–38.

[23] Vadacchino S, Mehta R, Sepahvand NM, Nichyporuk B, Clark JJ, Arbel T. Had-net: A hierarchical adversarial knowledge distillation network for improved enhanced tumour segmentation without post-contrast images. In: Medical imaging with deep learning. PMLR; 2021, p. 787–801.

[24] Wang Y, Zhang Y, Liu Y, Lin Z, Tian J, Zhong C, et al. ACN: Adversarial co-training network for brain tumor segmentation with missing modalities. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part VII 24. Springer; 2021, p. 410–20.

[25] Yang Q, Guo X, Chen Z, Woo PY, Yuan Y. D2-Net: Dual disentanglement network for brain tumor segmentation with missing modalities. IEEE Trans Med Imaging 2022;41(10):2953–64.

[26] Azad R, Khosravi N, Merhof D. SMU-Net: Style matching U-Net for brain tumor segmentation with missing modalities. In: International conference on medical imaging with deep learning. PMLR; 2022, p. 48–62.

[27] Havaei M, Guizard N, Chapados N, Bengio Y. HeMIS: Hetero-modal image segmentation. In: Medical image computing and computer-assisted intervention. Cham: Springer; 2016, p. 469–77.

[28] Chen C, Dou Q, Jin Y, Chen H, Qin J, Heng P-A. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part III 22. Springer; 2019, p. 447–56.

[29] Dorent R, Joutard S, Modat M, Ourselin S, Vercauteren T. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part II 22. Springer; 2019, p. 74–82.

[30] Ding Y, Yu X, Yang Y. RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 3975–84.

[31] Zhou T, Canu S, Vera P, Ruan S. Latent correlation representation learning for brain tumor segmentation with missing MRI modalities. IEEE Trans Image Process 2021;30:4263–74.

[32] Shen Y. Personalized stain style transfer layers for distributed histology classification. In: Medical imaging 2022: Digital and computational pathology, vol. 12039, SPIE; 2022, p. 134–9.

[33] Zhang Y, He N, Yang J, Li Y, Wei D, Huang Y, et al. Mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: Medical image computing and computer assisted intervention. MICCAI 2022, Cham: Springer Nature Switzerland; 2022, p. 107–17.

[34] Zhou T. Feature fusion and latent feature learning guided brain tumor segmentation and missing modality recovery network. Pattern Recognit 2023;141:109665.

[35] Konwer A, Hu X, Bae J, Xu X, Chen C, Prasanna P. Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, p. 21415–25.

[36] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 10012–22.

[37] Valanarasu JMJ, Patel VM. Unext: Mlp-based rapid medical image segmentation network. In: Medical image computing and computer assisted intervention–MICCAI 2022: 25th international conference, Singapore, September 18–22, 2022, proceedings, part v. Springer; 2022, p. 23–33.

[38] Atito S, Awais M, Kittler J. Sit: Self-supervised vision transformer. 2021, arXiv preprint arXiv:2104.03602.

[39] Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 6023–32.

[40] Dar SU, Yurt M, Karacan L, Erdem A, Erdem E, Cukur T. Image synthesis in multi-contrast MRI with conditional generative adversarial networks. IEEE Trans Med Imaging 2019;38(10):2375–88.

[41] Yu B, Zhou L, Wang L, Shi Y, Fripp J, Bourgeat P. Ea-GANs: Edge-aware generative adversarial networks for cross-modality MR image synthesis. IEEE Trans Med Imaging 2019;38(7):1750–62.

[42] Yurt M, Dar SU, Erdem A, Erdem E, Oguz KK, Çukur T. mustGAN: Multi-stream generative adversarial networks for MR image synthesis. Med Image Anal 2021;70:101944.

[43] Sharma A, Hamarneh G. Missing MRI pulse sequence synthesis using multi-modal generative adversarial network. IEEE Trans Med Imaging 2020;39(4):1170–83.

[44] Zhang Y, Peng C, Wang Q, Song D, Li K, Zhou SK. Unified multi-modal image synthesis for missing modality imputation. 2023, arXiv:2304.05340.

[45] Yang H, Sun J, Xu Z. Learning unified hyper-network for multi-modal mr image synthesis and tumor segmentation with missing modalities. IEEE Trans Med Imaging 2023.

[46] Mirza M, Osindero S. Conditional generative adversarial nets. 2014, arXiv preprint arXiv:1411.1784.

[47] Croitoru F-A, Hondru V, Ionescu RT, Shah M. Diffusion models in vision: A survey. IEEE Trans Pattern Anal Mach Intell 2023.

[48] Shen Y, Xu L, Yang Y, Li Y, Guo Y. Mixed sample augmentation for online distillation. In: ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing. ICASSP, IEEE; 2023, p. 1–5.

[49] Shen Y, Zhou Y, Yu L. Cd2-pfed: Cyclic distillation-guided channel decoupling for model personalization in federated learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 10041–50.

[50] Shen Y, Xu L, Yang Y, Li Y, Guo Y. Self-distillation from the last mini-batch for consistency regularization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 11943–52.

[51] Shen Y, Xu L, Yang Y, Li Y, Guo Y. Online distillation with mixed sample augmentation. 2022, arXiv preprint arXiv:2206.12370.

[52] Huang G, Sun Y, Liu Z, Sedra D, Weinberger KQ. Deep networks with stochastic depth. In: Computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, proceedings, part IV 14. Springer; 2016, p. 646–61.

[53] Chollet F. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 1251–8.

[54] Hendrycks D, Gimpel K. Gaussian error linear units (gelus). 2016, arXiv preprint arXiv:1606.08415.

[55] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PMLR; 2020, p. 1597–607.

[56] Milletari F, Navab N, Ahmadi S-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision. Ieee; 2016, p. 565–71.

2 成员 9 发表的 SCI 论文：Scikit-ANFIS: A Scikit-Learn Compatible Python Implementation for Adaptive Neuro-Fuzzy Inference System

Int. J. Fuzzy Syst. (2024) 26(6):2039–2057
https://doi.org/10.1007/s40815-024-01697-0

Check for updates

Scikit-ANFIS: A Scikit-Learn Compatible Python Implementation for Adaptive Neuro-Fuzzy Inference System

Dongsong Zhang [1,2] · Tianhua Chen [2]

& Tianhua Chen
T.Chen@hud.ac.uk

[1] School of Big Data and Artificial Intelligence, Xinyang College, Xinyang 464000, Henan, China

[2] School of Computing and Engineering, University of Huddersfield, Huddersfield HD1 3DH, UK

Received: 1 August 2023/Revised: 5 January 2024/Accepted: 30 January 2024/Published online: 3 June 2024 分节符(连续)

Abstract The Adaptative neuro-fuzzy inference system (ANFIS) has shown great potential in processing practical data from control, prediction, and inference applications, reflecting advantages in both high performance and system interpretability as a result of the hybridization of neural networks and fuzzy systems. Matlab has been a prevalent platform that allows to utilize and deploy ANFIS conveniently. On the other hand, due to the recent popularity of machine learning and deep learning, which are predominantly Python-based, implementations of ANFIS in Python have attracted recent attention. Although there are a few Python-based ANFIS implementations, none of them are directly compatible with scikit-learn, one of the most frequently used libraries in machine learning. As such, this paper proposes Scikit-ANFIS, a novel scikit-learn compatible Python implementation for ANFIS by adopting a uniform format such as fit() and predict() functions to provide the same interface as scikit-learn. Our ScikitANFIS is designed in a user-friendly way to not only manually generate a general fuzzy system and train it with the ANFIS method but also to automatically create an ANFIS fuzzy system. We also provide four kinds of representative cases to show that Scikit-ANFIS represents a valuable addition to the scikit-learn compatible Python software that supports ANFIS fuzzy reasoning. Experimental results on four datasets show that our ScikitANFIS outperforms recent Python-based implementations while achieving parallel performance to ANFIS in Matlab, a standard implementation officially realized by Matlab, which indicates the performance advantages and application convenience of our software.

Keywords Neuro-fuzzy · Fuzzy system · Anfis · Python · Scikit-learn · PyTorch

https://scikit-learn.org

1 Introduction

Since the adaptative neuro-fuzzy inference system (ANFIS) [1] was proposed in 1993 as a creative method of combining the advantages of the fuzzy system and neural network, it has been extensively applied in numerous fields. ANFIS is a unique five-layer neural network model that integrates fuzzy sets and logic modeling a fuzzy system. The model features a two-step learning algorithm that comprises a forward pass and a backward pass, which allows for automatic adjustment of the antecedent and consequent parameters by minimizing the error between the actual and target outputs [1]. This approach provides two main benefits, allowing for automatic learning from the data and employing fuzzy if-then rules to explain the model-

generated results. By combining the fuzzy system's explainability with the neural network's self-learning ability, this approach delivers unparalleled accuracy and interpretability.

As a result of the above advantages, research and application of ANFIS have drawn widespread attention in many domains. Health and well-being are one of the prioritized applications, due to the interpretability and accuracy typically required in healthcare domain that ANFIS may provide. Researchers have proposed a new approach to clinical decision support that uses data-driven techniques to create interpretable fuzzy rules. This approach combines decision tree learning mechanisms with an ANFIS framework, resulting in a method that outperforms many other popular machine learning techniques in terms of accuracy [2]. Other researchers have used ANFIS optimized through artificial bee colonies to classify heartbeat sounds, aiming at early detection of cardiovascular disease [3]. For the diagnosis process of Alzheimer's disease, some researchers have proposed a technique that first transforms it into a clustering problem, and then uses ANFIS to optimize fuzzy rules, which ultimately improves the accuracy of diagnosis [4]. ANFIS is also widely used in the field of control and engineering. A new combined ANFIS and robust proportional integral derivative control framework are proposed for building structure damping systems, which can effectively ensure the stability and robustness of the controller [5]. Some researchers have proposed using adaptive virtual synchronous generators with ANFIS controllers as inverter controllers in photovoltaic systems, which can enhance the system response in different operating scenarios [6]. ANFIS model has also been utilized to enhance electricity demand forecasting accuracy in a developing country, surpassing prior models and databases [7]. To predict power generation in photovoltaic systems, ANFIS models [8, 9] optimized by genetic algorithm or particle swarm optimization have been developed using Matlab [10] software with sound performance. These applications across a wide range of domains, demonstrate the effectiveness and popularity of ANFIS as a significant data analysis and model construction tool predominantly in decision-making and forecasting tasks, which in turn calls for more efforts in building an accessible development environment to streamline its applications.

At present, Matlab [10] is a widely used platform for convenient utilization and deployment of ANFIS. However, due to the increasing popularity of machine learning and deep learning, which are mainly based on Python, Python-based implementations of ANFIS have gained increasing attention. Despite the availability of some Python-based ANFIS implementations such as ANFISPyTorch [11], ANFIS-Numpy [12], and ANFIS-PSO [13], what is lacking in the current research landscape is that none of these implementations are directly compatible with scikit-learn, one of the most commonly used machine learning libraries.

Furthermore, due to emerging advancement in deep learning models, ANFIS has recently undergone new developments, including cascade ANFIS [14, 15], as well as integration with deep learning technology [16]. This has led to the emergence of a popular research area known as deep neural fuzzy system [17], of which ANFIS is an essential component. However, although some researchers have tried combining deep neuro-fuzzy systems (DNFS) [17] created using Python with scikit-learn, such as PyTSK [18], no cases have been found that combine ANFIS with scikit-learn, to the best of our knowledge.

Conventionally, ANFIS application and development are conducted in Matlab. However, with the rapid progress of deep learning and machine learning, which is commonly conducted in a Python environment, it is critical to develop ANFIS in an environment that is directly compatible with Python, Sklearn, and PyTorch [19]. This will facilitate the research and development of ANFIS and ensure compatibility with the latest technologies.

This paper reports a novel implementation of ANFIS, in Python programming language. To ease the use of ANFIS in compatibility with popular machine learning models, which have been realized in the popular scikit-learn library, our implementation, termed Scikit-ANFIS, fully supports interfaces as specified by scikit-learn. Furthermore, our ANFIS implementation, which may be utilized as an optimization method, also supports the training of an existing fuzzy system. Through several case studies and cross-validated experiments, our results demonstrate the superior performance of Scikit-ANFIS software compared to other ANFIS-based or DNFS-based Python software and are parallel to the standard ANFIS implementation by Matlab. Concretely, our contributions can be summarized as:

(1) → Our Scikit-ANFIS implementation is fully compatible with commonly used scikit-learn functions such as fit() and predict() - this enables our development directly applicable in combination with all existing machine learning models and methods as typically conducted through scikit-learn.

(2) → Scikit-ANFIS allows the manual generation of a general-purpose Takagi-Sugeno-Kang (TSK) [20] fuzzy system using natural languages. To the best of our knowledge, our method is the only Python-based implementation that supports fuzzy reasoning with

complex rules and logical operators of multiple choices.

(3) → Scikit-ANFIS utilizes the scikit_anfis() class to train a pre-existing TSK fuzzy system and automatically generate an ANFIS fuzzy system based on userspecified input–output data pairs, resulting in an efficient optimized fuzzy system.

(4) → The Scikit-ANFIS implementation can automatically save and load the trained ANFIS with the best performance to/from a local model file, which is not currently available in other Python-based ANFIS implementations.

2→Technical Background on ANFIS

We begin with a brief introduction of the ANFIS fuzzy system [1], as shown in Fig. 1. The basic architecture of ANFIS consists of five layers with the output of the nodes in each respective layer represented by O_{ij}, where i is the ith node of layer j.

In the first layer, the nodes of this layer are the membership scores generated based on the values of the fuzzy input variables, defined as:

$O^1_i(x_1) = l_{A_i}(x_1)$; $O^1_i(x_2) = l_{B_i}(x_2)$; $i = 1;2$ (1) where x_1;x_2 represent the crisp values of two input variables, and A_i;B_i are the fuzzy set associated with this node, and $l_{A_i}(x_1)$, $l_{B_i}(x_2)$ denote the membership function of linguistic labels A_i and B_i respectively. Any continuous and piecewise differentiable function such as the commonly used bell-shaped, gaussian, trapezoidal, and triangular membership functions, can be used as a membership function in this layer, and each membership function itself includes a set of parameters. When the values of these parameters change, the membership function also varies, so these parameters in this layer are called premise parameters [1].

In the second layer, each node represents the accumulated firing strength of rule antecedents through a t-norm operator such as the product as:

$O^2_i = w_i = l_{A_i}(x_1) \cdot l_{B_i}(x_2)$; $i = 1;2$ → (2)

In the third layer, each node outputs a normalized firing strength:

→ (3) → $\bar{w}_i$

$O_i = \bar{w}_i = \frac{w_i}{w_1 + w_2}$; $i = 1;2$ (3) $w_1 + w_2$

In the fourth layer, each rule consequent is calculated with associated parameters $p_i;q_i;r_i$.

$O^4_i = \bar{w}_i f_i = \bar{w}_i(p_i x_1 + q_i x_2 + r_i)$; $i = 1;2$ → (4)

When the values of the premise parameters are given, the single node in the fifth layer can be expressed as the sum of the linear combinations of consequent outputs, i.e.,

→ X → P w

$O^5_i = f_i = \sum_i \bar{w}_i f_i$; $i = 1;2$ (5) $O_1 = w$

It is important to note that ANFIS, unlike neural networks, grows faster in terms of the total number of parameters P_t, which can be calculated as follows [21]:

$P_p = in \cdot MF(input) \cdot coef(MF)$

$P_c = rs \cdot (in + 1) \cdot out$ → (6)

where P_p and P_c denote the number of premise parameters and that of consequent parameters respectively; in stands for the number of inputs, MF() is the number of membership functions in each input, and coef() is the number of coefficients for each membership function; rs stands for the number of rules, and out is the number of nodes in the fifth layer.

Given the original definition of ANFIS as introduced above, it represents a TSK-type fuzzy system that naturally fits a regression and control problem. Figure 1 shows an ANFIS with two inputs and two rules and one output, where each input has two Gaussian membership functions. The total number of parameters in ANFIS is 14, found by multiplying the coefficient (2) of the Gaussian membership function by the relevant values and adding them up:

$P_t = 2 \cdot 2 \cdot 2 + 2 \cdot 3 \cdot 1 = 14$.

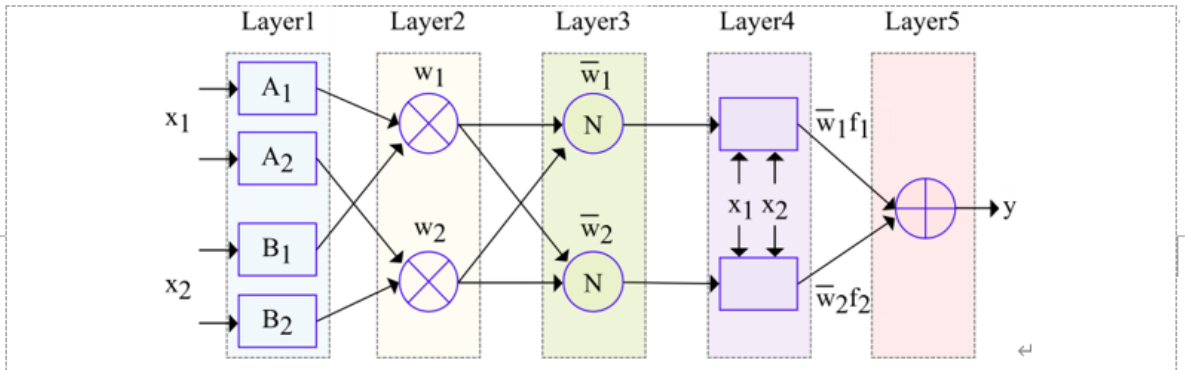


Fig. 1 ANFIS example with two inputs, two membership functions in each input, and two rules

complex rules and logical operators of multiple choices.

(3) → Scikit-ANFIS utilizes the scikit_anfisðÞ class to train a pre-existing TSK fuzzy system and automatically generate an ANFIS fuzzy system based on userspecified input–output data pairs, resulting in an efficient optimized fuzzy system.

(4) → The Scikit-ANFIS implementation can automatically save and load the trained ANFIS with the best performance to/from a local model file, which is not currently available in other Python-based ANFIS implementations.

2→Technical Background on ANFIS

We begin with a brief introduction of the ANFIS fuzzy system [1], as shown in Fig. 1. The basic architecture of ANFIS consists of five layers with the output of the nodes in each respective layer represented by $O_{i,i}$ where i is the ith node of layer j.

In the first layer, the nodes of this layer are the membership scores generated based on the values of the fuzzy input variables, defined as:

$$O^1_i(x_1) = l_{A_i}(x_1); O^1_i(x_2) = l_{B_i}(x_2); i = 1,2 \quad (1)$$

where x_1, x_2 represent the crisp values of two input variables, and A_i, B_i are the fuzzy set associated with this node, and $l_{A_i}(x_1)$, $l_{B_i}(x_2)$ denote the membership function of linguistic labels A_i and B_i respectively. Any continuous and piecewise differentiable function such as the commonly used bell-shaped, gaussian, trapezoidal, and triangular membership functions, can be used as a membership function in this layer, and each membership function itself includes a set of parameters. When the values of these parameters change, the membership function also varies, so these parameters in this layer are called premise parameters [1].

In the second layer, each node represents the accumulated firing strength of rule antecedents through a t-norm operator such as the product as:

$$O^2_i = w_i = l_{A_i}(x_1) \cdot l_{B_i}(x_2); i = 1,2 \quad \rightarrow \quad (2)$$

In the third layer, each node outputs a normalized firing strength:

$$O^3_i = \bar{w}_i = \frac{w_i}{w_1 + w_2}; i = 1,2 \quad (3)$$

In the fourth layer, each rule consequent is calculated with associated parameters p_i, q_i, r_i.

$$O^4_i = \bar{w}_i f_i = \bar{w}_i(p_i x_1 + q_i x_2 + r_i); i = 1,2 \quad \rightarrow \quad (4)$$

When the values of the premise parameters are given, the single node in the fifth layer can be expressed as the sum of the linear combinations of consequent outputs, i.e.,

$$\rightarrow \quad X \quad \rightarrow \quad P w$$

$$O^5_i = f_{out} = \sum_i \bar{w}_i f_i; i = 1,2 \quad (5) \quad O_1 = w$$

It is important to note that ANFIS, unlike neural networks, grows faster in terms of the total number of parameters P_t, which can be calculated as follows [21]:

$$P_p = in \cdot MF(input) \cdot coef(MF)$$

$$P_{ct} = Prs_p(b) \cdot Pin_c + 1) \cdot out \quad \rightarrow \quad (6)$$

where P_p and P_c denote the number of premise parameters and that of consequent parameters respectively; in stands for the number of inputs, MF() is the number of membership functions in each input, and coef() is the number of coefficients for each membership function; rs stands for the number of rules, and out is the number of nodes in the fifth layer.

Given the original definition of ANFIS as introduced above, it represents a TSK-type fuzzy system that naturally fits a regression and control problem. Figure 1 shows an ANFIS with two inputs and two rules and one output, where each input has two Gaussian membership functions. The total number of parameters in ANFIS is 14, found by multiplying the coefficient (2) of the Gaussian membership function by the relevant values and adding them up:

$$P_t = 2 \cdot 2 \cdot 2 + 2 \cdot 3 \cdot 1 = 14.$$

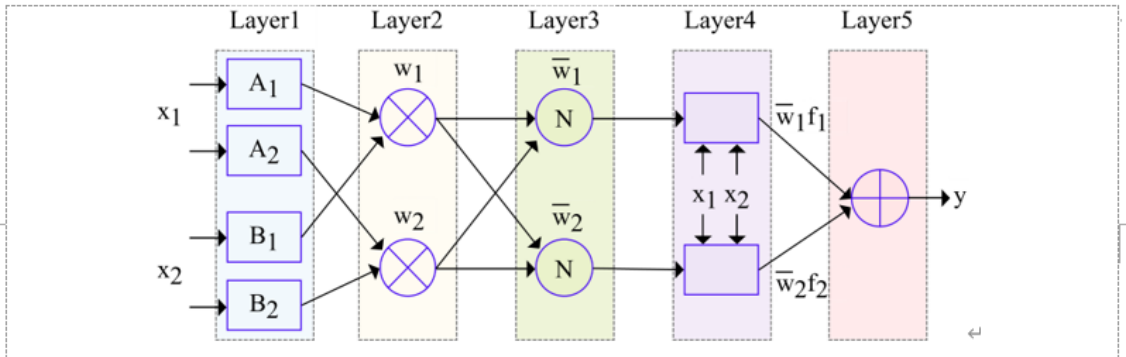


Fig. 1 ANFIS example with two inputs, two membership functions in each input, and two rules

Depending on how the consequent parameters are set and updated, Jang, the inventor of ANFIS [1], proposed two learning algorithms (i.e. training strategies) for the ANFIS model, namely hybrid and online. In hybrid learning, the antecedents are updated by the gradient descent method, while the consequents are calculated by the least squares method after fixing the premise parameters. Meanwhile, in online learning, all parameters are updated by the gradient descent method.

As depicted in Fig. 2, the hybrid learning algorithm comprises two stages: the forward pass and the backward pass. In the forward pass, the functional signal from Layer 1 is passed directly through the ANFIS network to Layer 4, where the consequent parameters are calculated by the least squares estimate (LSE) for input data X and target data Y. At this point, the premise parameters from the membership functions in Layer 1 remain fixed. The backward pass procedure starts after computing the total root mean square error (RMSE) loss. During this process, the consequent parameters are kept unchanged while the premise parameters are updated using the gradient descent method.

For clarity, the list of terminology abbreviations used in the paper is given in Table 1.

Abbreviation	Expansion
TSK	Takagi-Sugeno-Kang
ANFIS	Adaptive Neuro-Fuzzy Inference System
DNFS	Deep Neuro-Fuzzy Systems
Sklearn	scikit-learn
N/A	No Answer
Hybrid	gradient descent and least squares estimate
Online	gradient descent only
PSO	Particle Swarm Optimizer
GA	Genetic Algorithm
ABC	Artificial Bee Colony
LSE	Least Squares Estimate
RMSE	Root Mean Square Error
MBGD	Minibatch Gradient Descent
BN	Batch Normalization
UR	Uniform Regularization
LU	Layer Normalization
ReLU	Rectified Linear Unit
MFt	Membership Function types
FIS	Fuzzy Inference System
10-CV	10-fold cross-validation
Acc	Accuracy
n/a	Not applicable

Table 1 The terminology list of abbreviations used in the paper

3 Related Work

3.1 Recent Software Development for Fuzzy Systems

Generally speaking, a fuzzy system has a good level of interpretability, due to its knowledge encoding with imprecise knowledge and the intuitive inference mechanism that mimics human reasoning [22, 23]. During the early development of plain fuzzy systems in Python, many Python libraries were moving in the direction of generalpurpose fuzzy system applications, such as PyFuzzy [24], Fuzzylab [25], Scikit-Fuzzy [26].

However, many of these tools are outdated or no longer maintained. Recently, an open source software for general fuzzy systems is Simpful [27], which supports the natural language definition of fuzzy variables, fuzzy sets, and fuzzy rules, as well as any order TSK reasoning method. A common limitation is that most of the above software aims to create a general framework, which tends to require the creation of a fuzzy system by hand. The manual creation would become impractical in working with even a moderate-sized data set. Such limitation may also be more obvious in need of an automated optimization of system parameters, which can be dealt with through Matlab, but existing → Python → implementations → are → usually → not applicable.

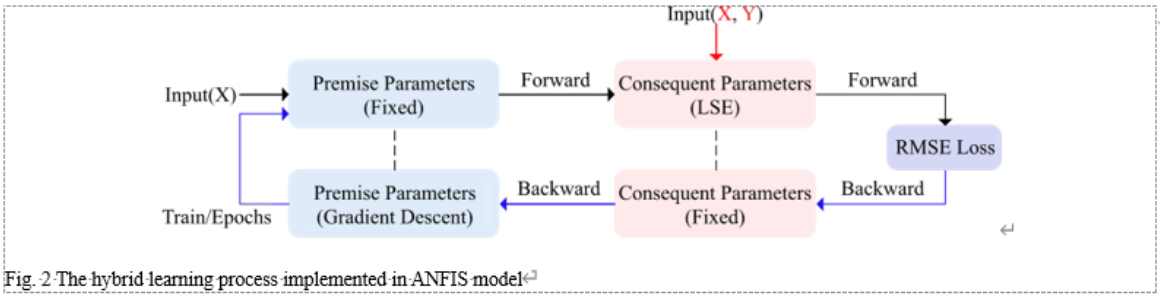


Fig. 2 The hybrid learning process implemented in ANFIS model

Focusing on the ANFIS framework [1], which has been a very prevalent TSK-type fuzzy system since its inception for a variety of domain problems [28], our Scikit-ANFIS is the first open-source Python tool to combine the creation of a general-purpose TSK fuzzy system embedded with the ANFIS optimization method.

3.2 Brief History of ANFIS Software Development

Since Jang proposed ANFIS, we make a summary of the recent major development of ANFIS software as shown in Table 2. Matlab-ANFIS is one of the most popular tools used to implement the ANFIS model [10], which can not only create the ANFIS model directly to train and test the data set but also utilize ANFIS as an optimization method to train the existing fuzzy system. However, Matlab is commercial software that is not open to the public. Furthermore, an extra installation of the Matlab Engine API for Python is required to access Matlab from Python.

The ANFIS-C [1] and ANFIS-Vignette [21] software written in C and R respectively, are outdated and not regularly updated. Currently, ANFIS software such as ANFIS-PyTorch [11], ANFIS-Numpy [12], and ANFISPSO [13] are mostly developed in Python 3 [29]. Out of the above three Python-based software, none supports the scikit-learn interface, and only a limited number of membership function types are supported (5, 3, and 3, respectively). ANFIS-PyTorch is the only software that supports both hybrid and online learning algorithms, while ANFISNumpy only supports hybrid learning, and ANFIS-PSO supports the particle swarm optimizer (PSO). On the other hand, our Scikit-ANFIS supports 12 different membership function types, the same as Matlab-ANFIS. Additionally, Scikit-ANFIS fully supports two learning algorithms (Hybrid/Online) for ANFIS training and also supports the scikit-learn interface, which is more user-friendly and has more powerful application capabilities.

3.3 Review on Deep Neuro-fuzzy Systems Framework

Deep neuro-fuzzy systems (DNFS), which present one of the most advanced developments as a combination of deep learning and fuzzy systems, have become a focus in fuzzy logic research [31]. This is because fuzzy systems can not only deal with the widespread inaccuracy and uncertainty in the real world but also potentially enrich the representation of deep models. At the same time, ANFIS can be seen as a simplified representation of DNFS [17], which itself is in principle a fuzzy system whose membership function parameters can be tuned by a five-layer adaptive neural network [1].

We further compare our Scikit-ANFIS implementation with other deep neural fuzzy methods for regression and classification in Table 3. There are several methods available for solving classification tasks, including the NeuroFuzzy [32] method based on C language, DNFC [33] based on Matlab, and TSK-MBGD-UR-BN [35] and PyTSK [18] based on Python 3. For regression tasks, there are also various methods available, such as FCM-RDpA [36] developed based on Matlab, MBGD-RDA [34], and HTSK-LN-ReLU [37] developed based on Python 3. However, only our Scikit-ANFIS is capable of solving both classification and regression tasks. It's worth noting that Python 3 has become a popular choice among the fuzzy logic research community, likely due to its widespread use in developing artificial neural networks. Although the methods mentioned above offer practical solutions for their specific tasks and implement different optimization techniques like gradient descent, minibatch gradient descent [38], Adam [39], AdaBound [40], Powerball [41], and AdaBelief [42], they do not utilize the ANFIS architecture or its training methods. By contrast, our Scikit-ANFIS has the ability to not only adopt the ANFIS's five-layered architecture but also adapt to the upcoming requirements of DNFS research for network interpretability and high performance with the assistance of PyTorch and Numpy frameworks. 分节符(连续)

Table 2 Overview of the software for ANFIS Table 3 The difference between Scikit-ANFIS and other deep neural fuzzy methods for two common tasks: regression and classification

Name	Language	Library	MFs	Learning strategy	Sklearn	Release
ANFIS-C[1]	C	N/A	4	Hybrid/Online	No	1993
ANFIS-Vignette[21]	R	N/A	4	Hybrid/Online	No	2012
Matlab-ANFIS[10]	Matlab	N/A	12	Hybrid/Online	No	2023
ANFIS-PyTorch[11]	Python 3	PyTorch	5	Hybrid/Online	No	2019
ANFIS-Numpy[12]	Python 3	Numpy[30]	3	Hybrid	No	2020
ANFIS-PSO[13]	Python 3	Numpy	3	PSO	No	2021
Scikit-ANFIS	Python 3	PyTorch?Numpy	12	Hybrid/Online	Yes	2023

Name	Language	MFs	Layers	Optimization method	Tasks	Sklearn	Release

Table 3 The difference between Scikit-ANFIS and other deep neural fuzzy methods for two common tasks: regression and classification

Name	Language	MFs	Layers	Optimization method	Tasks	Sklearn	Release
Neuro-Fuzzy[32]	C	1	4	Gradient Descent	Classification	No	1993
DNFC[33]	Matlab	1	8	Gradient Descent	Classification	No	2020
MBGD-RDA[34]	Python 3	1	5	MBGD+AdaBound	Regression	No	2020
TSK-MBGD-UR-BN[35]	Python 3	1	6	MBGD+AdaBound+BN+UR	Classification	No	2020
FCM-RDpA[36]	Matlab	1	5	MBGD+Powerball+AdaBelief	Regression	No	2021
HTSK-LN-ReLU[37]	Python 3	1	7	MBGD+Adam+LU+ReLU	Regression	Yes	2022
PyTSK[18]	Python 3	2	6	MBGD+Adam	Classification	Yes	2022
Scikit-ANFIS	Python 3	12	5	ANFIS	Regression+Classification	Yes	2023

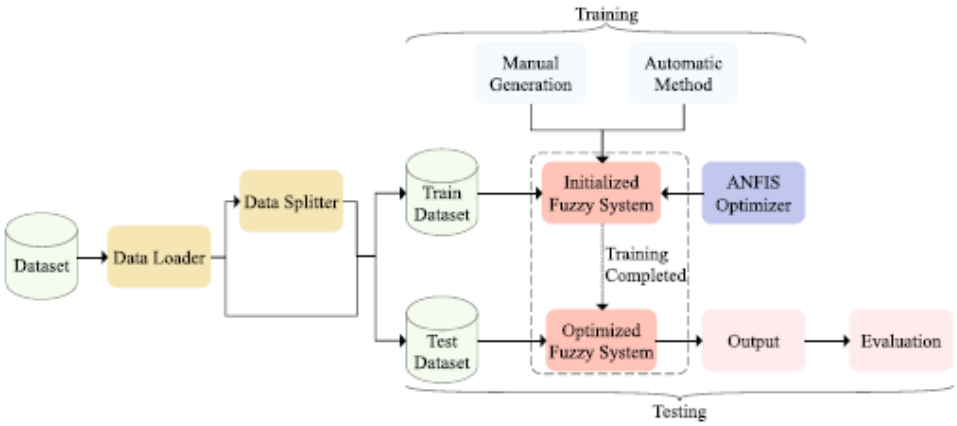


Fig. 3 Overview of Scikit-ANFIS architecture

4 Scikit-ANFIS

4.1 Architecture Overview

The diagram in Fig. 3 illustrates the overall structure of our Scikit-ANFIS. Scikit-ANFIS employs the data loader module to read data from the dataset, which is then divided into train, and test datasets by the data splitter module. These datasets are sent to the generated fuzzy system for training. Alternatively, the initialized fuzzy system can directly read the train, and test data from the dataset by the data loader module and train itself accordingly. To create a fuzzy system for predictive tasks, Scikit-ANFIS provides two options: the manual generation module can be utilized to define and generate a fuzzy system, or the automatic method module can automatically generate an ANFIS fuzzy system by default without definition. For training, Scikit-ANFIS uses the ANFIS optimizer module to train the initialized fuzzy system. Once training is completed, the optimized fuzzy system is selected and tested with data. The evaluation module then examines the test outputs to formulate a report.

Scikit-ANFIS[1] is also implemented in the Python 3 language, which mainly includes two dependencies such as PyTorch [19] and Numpy [30]. Our Scikit-ANFIS currently supports the following primary functions: (i) The twelve types of membership functions such as Gaussian, bell, triangular, and others. (ii) Fuzzy sets written in natural language, and complex fuzzy rules with logical operators AND, OR, and NOT. (iii) Two training strategies of ANFIS, namely hybrid and online. (iv) Automatic

[1] The code for Scikit-ANFIS, the associated cases, and the user guide will be publicly available at https://github.com/hudscomdz/scikit-anfis. Scikit-ANFIS can be installed by using the following command: pip install skanfis.

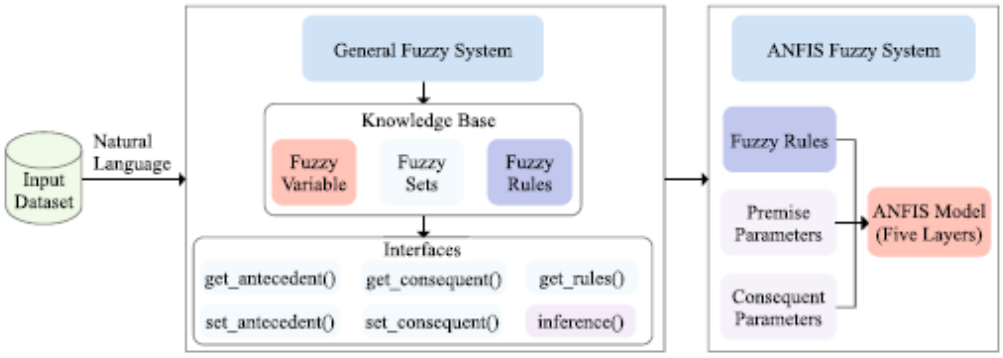


Fig. 4 The illustration of the manual generation method for the general fuzzy system

generation and training of the ANFIS fuzzy system. (v) A uniform structure such as the *fit*() and *predict*() functions to provide the same interface as scikit-learn.

4.2 Implementation Details

4.2.1 Manual Generation of a General Fuzzy System

Considering that Simpful [27] is already open source and general-purpose fuzzy system software developed in Python, our implementation also makes use of some existing components as defined by Simpful for efficient development. As depicted in Fig. 4, the difference between the manual generation method for general fuzzy system and Simpful is mainly that the former can interact with the ANFIS optimizer in Scikit-ANFIS, which can not only realize the ANFIS training of the fuzzy system but also return the trained results to the fuzzy system to generate new output. However, the latter can only generate an output after passing the received input data through the fuzzy knowledge base without any model training operation.

Similarly to Simpful, after receiving the natural language information, our manual generation method automatically parses the fuzzy variables, fuzzy sets, and fuzzy rules, creating Scikit-ANFIS's fuzzy system object. The fuzzy rule uses Takagi and Sugeno's fuzzy if-then rule [1], and its natural language description supports the commonly used fuzzy operators like AND, OR, and NOT, as detailed in [27]. When the input dataset is fed into the fuzzy system object created by the manual generation method, it can conduct fuzzy reasoning through an interface namely *inference*(), and provide output results. Additionally, the object can communicate with the ANFIS model in Scikit-ANFIS through five interfaces: *get_antecedent*(), *get_consequent*(), *get_rules*(), *set_antecedent*(), and *set_consequent*(). By providing the first three interfaces, Scikit-ANFIS can send the antecedent parameters, fuzzy rules, and consequent parameters of the fuzzy system object to the ANFIS model for training. After the training is finished, the last two interfaces accurately return the well-trained ANFIS model's parameters to the fuzzy system object, enabling it to perform precise fuzzy inference.

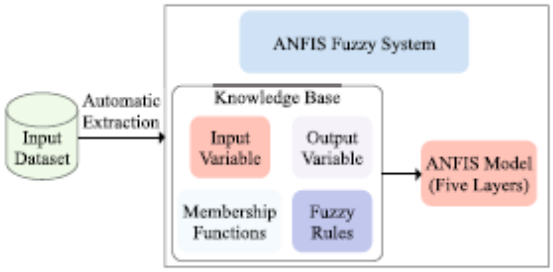


Fig. 5 The illustration of the automatic method for ANFIS fuzzy system

4.2.2 Automated Method to Initialise an ANFIS Fuzzy System

To facilitate users to create the ANFIS model, our Scikit-ANFIS designs and implements an automatic method for ANFIS fuzzy system, which shares the same *scikit_anfis*() class with the ANFIS optimizer module. Figure 5 illustrates the functional diagram of the method, which can automatically generate an ANFIS fuzzy system object including a knowledge base and ANFIS model for the input dataset. The input data set is used to automatically extract a knowledge base consisting of input variables, output variables, membership functions, and fuzzy rules. This leads to

the generation of an ANFIS model that adheres to the strict requirements of the type-3 fuzzy inference system as proposed by the original paper [1]. The rule base follows fuzzy if-then rules and can be effortlessly mapped to an equivalent ANFIS architecture [1]. The resulting ANFIS model comprises a five-layer neural network structure, as illustrated in Fig. 1, and provides a robust fuzzy inference system.

4.2.3 ANFIS as an Optimizer

ANFIS optimizer as an optimization module also utilizes the *scikit_anfis*() class to help the initialized fuzzy system to be trained more efficiently. The ANFIS optimizer takes the training set as input and uses forward propagation and cost function to calculate the total loss of the ANFIS neural network generated. The *forward*() method is used for forward propagation, built on the PyTorch framework. The default training algorithm used is hybrid learning, with online learning available as an alternative. Then, it updates all the antecedent and consequent parameters in the model through the backpropagation process. This entire process is the training process for the five-layer ANFIS model, and the number of times the model is trained is related to the 'epoch' hyperparameter. The optimizer used to update the parameter through backpropagation is usually related to the 'optimizer' hyperparameter of the model. Our Scikit-ANFIS has implemented various optimizers based on gradient descent, including Adam [39], SGD [43], Rprop [44], L-BFGS [45], Adadelta [46], and Adagrad [47], with Adam being the default.

Once the training of the ANFIS model is completed, all parameters of the current model with minimum loss can be saved to the local model file 'tmp.pkl'. This saved model file can be later used to continue training or testing, which can be very useful. To ensure that a manually created general fuzzy system and a trained ANFIS model are consistent with each other, it is possible to transfer the premise and consequent parameters from the ANFIS model to the fuzzy system. This can be done by using two interfaces such as *set_antecedent*() and *set_consequent*(), which are explicitly called as shown in Fig. 4. Moreover, the optimized fuzzy system can give fuzzy inference results based on the test data.

4.2.4 Scikit-ANFIS

The implementation of our Scikit-ANFIS is detailed in Algorithm 1. The input, training, and test datasets are denoted as (X, Y), (X_{train}, Y_{train}), and (X_{test}, Y_{test}) respectively, as illustrated in Fig. 3. The main procedure of Scikit-ANFIS encapsulated between lines 1 and 15, comprises three key elements. Firstly, in line 1, a general fuzzy system object *fs* is manually created. Following that, lines 2 and 3 specify fuzzy sets, input variables, fuzzy rules, and output variables for *fs*. In line 4, we check whether *fs* is empty. If it is not, we use the antecedent and consequent parameters along with rules from *fs* to represent a 5-layer ANFIS model called *anfis_layers* in line 5. At the same time, we create a *scikit_anfis* object based on *anfis_layers* in line 6, also specifying the maximum number of epochs for training (*max_epoch*), the training strategy (*hybrid*), and the task type (*label*), and the *optimizer*. By default, *max_epoch* is set to 10, *hybrid* is set to True, indicating the use of the hybrid training strategy, *label* is set to "r", indicating that the model is intended for regression tasks, and *optimizer* is set to the gradient descent method namely "Adam". Secondly, if *fs* is empty, we automatically create an ANFIS fuzzy system from lines 7 to 10. Line 8 is the 5-layer ANFIS model *anfis_layers* derived from the input dataset (X, Y), and line 9 generates a *scikit_anfis* object and specifies its associated parameters, such as *max_epoch*, *hybrid*, *label*, and *optimizer*. Thirdly, the most crucial step involves the implementation of the ANFIS optimizer from lines 11 to 15. At each epoch in the loop, we feed the training dataset (X_{train}, Y_{train}) into the *scikit_anfis* object to complete the forward pass from layer 1 to layer 5 in its ANFIS model. During this process, we update the consequent parameters and obtain the final output $\hat{Y}$. We then compute the RMSE loss between the predicted $\hat{Y}$ and the training target Y_{train}, followed by updating the premise parameters of the ANFIS model in its backward process.

When the ANFIS optimizer has completed training the model, confidently use *fs* for fuzzy inference of test data from lines 16 to 20. Return the premise and consequent parameters to *fs* using the *set_antecedent*() and *set_consequent*() interfaces. Then, directly call the *inference*() function of *fs* to complete the fuzzy system reasoning. Alternatively, we can opt for the more convenient *scikit_anfis* object in line 21 for fuzzy inference. This method uses the test input data X_{test} to generate the predicted output Y_{pred} through fuzzy reasoning of the ANFIS model. This option has been used in subsequent cases in this paper. It is important to note that during testing, there is no need to provide the test target value Y_{test} to the trained *scikit_anfis* object since all of its parameters are already fixed. Finally, in the last line, we compare and evaluate the fuzzy inference result Y_{pred} with the test target value Y_{test}.

calling the *fit*() and *predict*() functions in line 4 and 5, respectively.

Listing 2 The code demo of the Sklearn interface for generating and optimizing a TSK fuzzy system using Scikit-ANFIS.

```
1 from skanfis import *
2
3 model = scikit_anfis() # Create an ANFIS model
4 model.fit(train_data) # Model training
5 y_pred = model.predict(X_test) # Model testing
```

5 Experimentation

5.1 Experimental Setup

For a fair comparison and consistency with previous ANFIS implementations (Matlab-ANFIS [10], ANFIS-PyTorch [11], and ANFIS-Numpy [12]), this section reports results and discussions as a result of experimentation over both regression and classification tasks. Thus, the first regression dataset is the restaurant tipping problem [48] taken from the Matlab file repository. The next two regression datasets are from Jang's literature [1], using ANFIS to model a nonlinear sinc equation and predict future values of a chaotic time series, respectively. Finally, we use the ANFIS model to train and test the popular iris benchmark dataset [49], which is essentially to a three-class classification problem. The details regarding task type, features (i.e. inputs), and the number of samples in all four datasets are presented in the following Table 4. It should be noted that we do not pre-process data or conduct any hyperparameter tuning for a fair and straightfoward comparison.

5.2 Case Studies

In this section, we report reports on four studies including three regression problems and one classification problem, with major Python code to demonstrate how Scikit-ANFIS can be applied in practice.

5.2.1 Restaurant Tipping Problem

The restaurant tip problem is to calculate a fair tip ratio of the total bill according to the service and food quality of a restaurant. Listing 3 shows an example of Scikit-ANFIS code to manually define a general TSK fuzzy inference system through natural language, and then train and test the fuzzy inference system (FIS) with a Scikit-ANFIS object, based on the Matlab data file 'data_Tip_441.mat' with two inputs and one output for a total of 441 samples.

Listing 3 A TSK FIS example for restaurant tipping problem, implemented in Scikit-ANFIS

```
1 from skanfis.fs import *
2 from skanfis import *
3 from scipy.io import loadmat
4 from sklearn.model_selection import
      train_test_split
5 from sklearn.metrics import mean_squared_error
6
7 # Load Tip data
8 tip_data = loadmat('data_Tip_441.mat')['Tip_data
      ']
9 # Split the data using the test_size argument in
       training(50%) and test(50%) set
10 train_data, test_data = train_test_split(
      tip_data, test_size=0.5, random_state=42)
11 y_test = test_data[:,-1]
12 X_test = test_data[:,:-1]
13
14 # Create a TSK fuzzy system object by default
15 fs = FS()
16
17 # Define fuzzy sets and linguistic variables
18 S_1 = TriangleFuzzySet(a=0, b=0, c=5, term="poor
      ")
19 S_2 = TriangleFuzzySet(a=0, b=5, c=10, term="
      good")
20 S_3 = TriangleFuzzySet(a=5, b=10, c=10, term="
      excellent")
21 fs.add_linguistic_variable("Service",
      LinguisticVariable([S_1, S_2, S_3]))
22 F_1 = TriangleFuzzySet(a=0, b=0, c=10, term="
      rancid")
23 F_2 = TriangleFuzzySet(a=0, b=10, c=10, term="
      delicious")
24 fs.add_linguistic_variable("Food",
      LinguisticVariable([F_1, F_2]))
25
26 # Define crisp outputs for small and average tip
27 fs.set_crisp_output_value("small", 5)
28 fs.set_crisp_output_value("average", 15)
29 # Define function for generous tip (2*food score
       + 3*service score + 5%)
30 fs.set_output_function("generous", "2*Food+3*
      Service+5")
31
32 # Define fuzzy rules
33 R1 = "IF (Service IS poor) OR (Food IS rancid)
      THEN (Tip IS small)"
34 R2 = "IF (Service IS good) THEN (Tip IS average)
      "
35 R3 = "IF (Service IS excellent) OR (Food IS
      delicious) THEN (Tip IS generous)"
36 fs.add_rules([R1, R2, R3])
37
```

In line 8 of Listing 3, the data 'tip_data' is loaded from the Matlab file using the *loadmat*() command in the scipy.io package. Then in line 10, using the *train_test_split*() command from the sklearn package, the

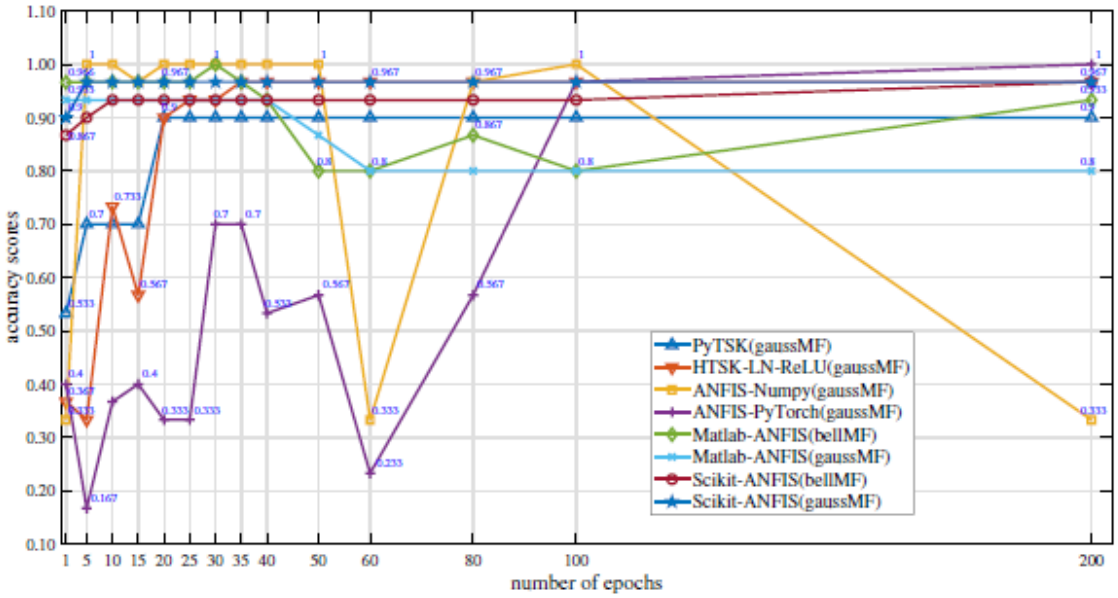


Fig. 6 Comparison of accuracy scores for several software methods in terms of different epochs using the Iris dataset

Subsequently, Fig. 6 shows a comparison of output accuracy scores produced by various software methods such as PyTSK [18], HTSK-LN-ReLU [37], ANFIS-Numpy [12], ANFIS-PyTorch [11], Matlab-ANFIS [10], and our Scikit-ANFIS with different epoch sizes under the same Iris dataset in Table 4. For a fair comparison, all the methods generate fuzzy systems two membership functions for each of the four input variables, resulting in 16 fuzzy rules. The Iris dataset is then divided into a fixed 80:20 proportion for training and testing purposes (refer to Listing 6: lines 7-9). It is important to note that the training set and test set remain the same across all the methods. In addition, in order to explain the influence of the membership function of input variables on a fuzzy system, the text in parentheses after each software name indicates which membership function is used to construct the fuzzy system. For example, the 'PyTSK(guassMF)' entry and 'Matlab-ANFIS(bellMF)' one in the figure indicate that the PyTSK software uses the Gaussian membership function, and the Matlab-ANFIS software uses the bell membership function, respectively. PyTSK and HTSK-LN-ReLU used their own gradient descent methods to train the dataset, while other ANFIS methods used a hybrid approach. The experiments were repeated 10 times per group for accuracy.

As shown in Fig. 6, the accuracy score in the PyTSK increases linearly when the number of epochs increases from one to two hundred, ranging from 0.533 to 0.9. Although the accuracy score of HTSK-LN-ReLU still maintains the overall linear increase from 0.333 to 0.967, the accuracy score increases and decreases with different epochs, which has a slight fluctuation. With the increasing number of epochs, the accuracy scores of the ANFIS-Numpy and ANFIS-PyTorch fluctuate between 0.333 and 1.0 and between 0.167 and 1.0, respectively. In addition, Matlab-ANFIS(bellMF) exhibits a small change from 0.8 to 1.0, and the accuracy of Matlab-ANFIS(gaussMF) decreases linearly from 0.933 to 0.8 as the number of epochs increases. In contrast, the accuracy score of our Scikit-ANFIS increases strictly linearly as the number of epochs grows larger, with minimal variation, where Scikit-ANFIS(bellMF)'s score increases from 0.867 to 0.967 and Scikit-ANFIS(gaussMF)'s from 0.9 to 0.967. The relatively stable accuracy score in our Scikit-ANFIS is particularly beneficial in real scenarios where the ANFIS model is applied. This makes our Scikit-ANFIS more efficient and adaptable to handle different ANFIS training configurations of fuzzy inference systems.

5.3 Using Sklearn-supported Cross-validation for Scikit-ANFIS

To further demonstrate the effectiveness of our Scikit-ANFIS software tool, we completed the 10-fold cross-validation ('10-CV' for short) experimental comparison of various software for the above four datasets using *cross_val_score*() command in Scikit-learn. '10-CV' refers to dividing the data into 10 equal folds of smaller

Table 5 Summary of 10-fold cross-validation ('10-CV' for short) experiments of various software for ANFIS or DNFS under four different datasets at the same setting of 100 epochs and Gaussian membership functions

Methods	Regression(RMSE↓)			Classificaion(Acc↑)
	Tip	Sinc	PCD	Iris
Matlab-ANFIS(hybrid)[10]	**8e-7±3e-8**	0.127±0.016	**0.002±5e-5**	0.875±0.018
ANFIS-PyTorch(hybrid)[11]	3e-6±1e-6	0.236±0.082	0.030±0.004	0.933±0.067
ANFIS-Numpy[12]	0.171±0.510	0.283±0.326	0.194±0.200	0.860±0.156
Scikit-ANFIS(hybrid)	1e-6 ± 1e-7	**0.109±0.068**	0.041±0.007	**0.952±0.054**
PyTSK[18]	n/a	n/a	n/a	0.301±0.098
HTSK-LN-ReLU[37]	0.945±0.451	0.859±0.643	1.100±0.065	0.346±0.105
Matlab-ANFIS(online)[10]	0.426±0.013	0.119±0.015	0.103±0.001	**0.968±0.019**
ANFIS-PyTorch(online)[11]	2.043±1.516	0.141±0.075	0.403±0.006	0.493±0.326
Scikit-ANFIS(online)	**0.382±0.187**	**0.101±0.069**	**0.044±0.005**	0.960±0.044

Each value in the experiment is in the form of the mean score ± standard deviation. Except for the Iris data set in the last column of the following table, which is the accuracy (Acc) score computed at each '10-CV' iteration, the data sets in the other three columns such as Tip, Sinc, and PCD are computed for the root mean squared error (RMSE) score. ↓ indicates that the smaller the value is, the better the performance, while ↑ indicates that the larger the value is, the better the performance. 'n/a' indicates 'not applicable'

sets, then training the model using 9 of the folds as training data, and testing the resulting model on the remaining 1 fold of the data for computing a performance measure such as root mean squared error and accuracy. The four datasets in the '10-CV' experiment are from Table 4, which can be divided into three regression sets (Tip, Sinc, PCD) and one classification set (Iris) according to task type. In the experiment, the processing details of regression and classification data sets are different. The main difference lies in the setting of 'scoring' parameters for defining model evaluation rules in the *cross_val_score*() command. As for regression, 'neg_mean_squared_error' has been specified as the 'scoring' parameter shown in line 4 of Listing 7, while for classification, the 'scoring' parameter is set to 'accuracy' shown in line 4 of Listing 8. In addition, line 4 of Listing 7 or 8 aims to conduct '10-CV' experiments on the data set composed of input data 'X' and output data 'y', in which 'model' parameter in *cross_val_score*() command can refer to our Scikit-ANFIS and any ANFIS or DNFS model to be trained and tested. In line 5 of Listing 7 and 8, the mean and standard deviation of the model's evaluation scores in ten folds are calculated and printed in the console.

Listing 7 The code demo of our 10-fold cross-validation experiment for regression dataset.

```
1 from sklearn.model_selection import
      cross_val_score
2 :
3 # Evaluate a score of the model
4 scores = cross_val_score(model,X,y,cv=10,scoring
      ='neg_root_mean_squared_error')
5 print("Mean: %0.3f, Standard Deviation: %0.3f"%(
      scores.mean(), scores.std()))
```

Listing 8 The code demo of our 10-fold cross-validation experiment for classification dataset.

```
1 from sklearn.model_selection import
      cross_val_score
2 :
3 # Evaluate a score of the model
4 scores = cross_val_score(model,X,y,cv=10,scoring
      ='accuracy')
5 print("Mean: %0.3f, Standard Deviation: %0.3f"%(
      scores.mean(), scores.std()))
```

Table 5 summarizes the mean and standard deviation of evaluated scores in each of '10-CV' experiments with various software for ANFIS or DNFS such as Matlab-ANFIS [10], ANFIS-PyTorch [11], ANFIS-Numpy [12], PyTSK [18], HTSK-LN-ReLU [37], and Scikit-ANFIS in four datasets from Table 4. To ensure uniformity in the software builds of fuzzy sets and fuzzy rules for the same data set, we have mandated that every experiment must have an epoch of 100 and an initial step of 0.01, with two Gaussian membership functions designated for each input variable. As illustrated in Listing 3, our experiment comprises two input variables and three fuzzy rules for the Tip. Similarly, as per Listing 4, we have enforced two input variables and 16 fuzzy rules for the Sinc. For the PCD and Iris, we have assigned four input variables and 16 fuzzy rules, as elaborated in Listings 5 and 6. Each experiment was repeated 10 times, and the average was the final result.

In Matlab-ANFIS, we can complete the '10-CV' experiment by calling the *crossvalind*() with the Kfold method and *anfis*() commands in Matlab [10]. In addition, ANFIS-PyTorch can be called by the *cross_val_score*() command through the Numpy wrapper. In contrast, the other four software can be called directly, because all five are based on the Python 3 programming language, and have natural interoperability with the sklearn package. The experimental

results under the two training strategies of the three ANFIS-based software Matlab-ANFIS, ANFIS-PyTorch, and Scikit-ANFIS are distinguished by adding 'hybrid' or 'online' in parentheses, as shown in Table 5. However, there is only a hybrid training method for ANFIS-Numpy, and PyTSK and HTSK-LN-ReLU can be classified as online training methods because they are based on gradient descent to realize the backpropagation update of all parameters in the network. Since PyTSK only applies to classification problems, its experimental results in three regression data sets are denoted as 'n/a'. We have also highlighted the best performance values in bold for both the first four methods under hybrid training strategy and the remaining five methods under online strategy, for each data set.

The findings in Table 5 reveal that the evaluation performance (0.109±0.068 and 0.952±0.054) of our Scikit-ANFIS exceeds all other three software in Sinc and Iris data sets under the 'hybrid' training method, while only the RMSE performance (8e-7±3e-8 and 0.002± 5e-5) Matlab-ANFIS exceeds ours in Tip and PCD data sets and ours still exceeds ANFIS-PyTorch and ANFIS-Numpy. On average, our developed Scikit-ANFIS demonstrates performance approximation to commercial software Matlab-ANFIS compared to ANFIS-PyTorch and ANFIS-Numpy. In addition, under the 'online' training method, the RMSE performance (0.382±0.187, 0.101±0.069, and 0.044±0.005) of our Scikit-ANFIS outperforms the other three software such as HTSK-LN-ReLU, Matlab-ANFIS, and ANFIS-PyTorch in three data sets such as Tip, Sinc, and PCD, while the accuracy score 0.968±0.019 of Matlab-ANFIS only slightly outperforms 0.960±0.044 of ours in Iris data set, and ours consistently outperforms the other three software such as PyTSK, HTSK-LN-ReLU, and ANFIS-PyTorch. These results also highlight the faster convergence of our Scikit-ANFIS compared to PyTSK, HTSK-LN-ReLU, Matlab-ANFIS, and ANFIS-PyTorch in updating all the antecedent and consequent parameters of fuzzy systems based on the gradient descent method.

5.4 Discussions and Limitations

Figure 6 and Table 5 yield that Matlab-ANFIS is still the relatively best-performing software among many existing ANFIS-based or DNFS-based software for both regression and classification tasks. However, when faced with a mainstream Python-based machine learning library such as scikit-learn, Matlab-ANFIS is not convenient to use as closed-source commercial software. In contrast, our Scikit-ANFIS software is open source and better suited for the Python platform, outperforming state-of-the-art ANFIS-based or DNFS-based Python software in terms of performance, and is the closest to Matlab-ANFIS. This superiority can be attributed to several key factors. First and foremost, our software benefits from the step size (i.e., learning rate) update method implemented following two heuristic rules [1], which plays an important role in guiding the ANFIS model to accelerate the convergence speed of gradient descent when backpropagating. At the backward stage of Scikit-ANFIS, we apply the adaptive gradient descent method (Adam by default) instead of strict gradient descent to identify the parameters in the ANFIS network. This enables us to obtain an unscaled direct estimation of the parameter's updates, which is well-suited for problems that are large in terms of data or parameters. In addition, although PyTSK and HTSK-LN-ReLU also adopt stochastic gradient descent methods such as Adam, they are different from the ANFIS model in that they completely rely on input–output data pairs to generate corresponding network structure and membership parameters. However, our Scikit-ANFIS relies not only on input and output data, but also on human knowledge to construct fuzzy if-then rules, so the resulting fuzzy system is more consistent with the real data distribution and achieves remarkable results.

Another key contribution of our software is the support of the ANFIS model as an optimization method to directly train existing TSK fuzzy inference systems, allowing users to apply the software more easily. Although Matlab-ANFIS can accomplish the same function, it can only be used for fuzzy systems created using the Matlab language and is not easily compatible with the Python platform.

One of the most significant advantages of our proposed syntax is its high level of consistency with that of scikit-learn: both adopt a universal format that first creates a model, which can then be fed data through *fit*(), before outputting a result through *predict*(). It is also worth noting again that *fit*() and *predict*() functions of our Scikit-ANFIS inherit the same interface provided by scikit-learn, thus facilitating the proposed model to be used efficiently with other available machine learning algorithms. This is also reinforced by the fact that output generated by *predict*() can be directly used to calculate metrics such as RMSE and accuracy (see code example Listing 6: lines 14-15).

Although our Scikit-ANFIS helps users to efficiently combine the ANFIS model with other machine learning algorithms in scikit-learn, it has some limitations. Scikit-ANFIS relies on adaptive gradient descent methods to update antecedent and consequent parameters of a fuzzy system, making it hard to initialize the optimal hyperparameters for the above gradient descent method. Since there is no way to know in advance the optimal value for the hyperparameter, this limitation can be addressed by using the *GridSearchCV*() function in scikit-learn's *model_selection* package to try all possible values to find the optimal one.

During the training process of ANFIS, the optimization methods used play an essential role in obtaining effective results [28]. The commonly used methods are Gradient Descent and Least Squares Estimate, but heuristic

algorithms such as PSO [50] and GA [51] can also be utilized to train the premise and consequence parameters of ANFIS. To further enhance the training method of Scikit-ANFIS, we aim to integrate PSO and GA in a hybrid training method along with Gradient Descent or LSE. This will improve the training process [28] and ultimately lead to better overall performance of our Scikit-ANFIS.

Another limitation of existing Scikit-ANFIS is that the fuzzy system does not directly address variables with missing values, which is also a limitation for some machine learning algorithm as implemented by scikit-learn - a common workaround is to use imputation techniques (e.g., advanced fuzzy interpolation techniques [52] in the context of a fuzzy system) to fill missing values with artificially generated values before training and/or prediction.

One more limitation worth discussion is that Scikit-ANFIS can still suffer from the curse of dimensionality problem, particularly those initialized by the grid partition method. This is because both the number of fuzzy rules and the training time grow exponentially with the number of fuzzy sets for each input variable, which limits the number of input variables and membership functions, resulting in reduced prediction accuracy due to the absence of important characteristic variables [53]. Although the original ANFIS paper [1] does not discuss this limitation possibly due to when data collected by then was relatively small, it's naturally desirable for a model to work with data of high dimensionality to meet the growing trend. While this paper reports only the implementation of original ANFIS, part of future work will concentrate on advanced dimensionality reduction techniques, such as those based on granule computing and rough-set [54–58] for developing novel computational intelligence models and applications in the era of big data.

6 Conclusion

It is common among the research community to apply ANFIS based on the Matlab platform to a variety of regression, classification, process controls, and pattern recognition applications, which makes it difficult for users to combine it with the common scikit-learn library in the Python platform. Hence, in this work, we implement Scikit-ANFIS, a user-friendly, and scikit-learn-compatible Python software using ANFIS architecture specifically designed for training the TSK fuzzy systems. Scikit-ANFIS takes a universal format to create a model and train the model through *fit*() and test it through *predict*(). Our Scikit-ANFIS implementations demonstrate performance gains on four standard data sets compared to existing Python programs that have implemented the ANFIS or DNFS method. Furthermore, our Scikit-ANFIS allows for training an existing TSK fuzzy system directly and automatically generating an ANFIS fuzzy system based on stipulated input–output data pairs.

For future research, we will explore how the Scikit-ANFIS software can be used with deep neuro-fuzzy systems to further strengthen the performance for solving regression and classification problems. Additionally, we will apply it to more complex problem domains such as health and care domain where both model performance and interpretability are usually among the top concerns in medical practice [59].

Acknowledgements This work is partially supported by the Henan Key Research and Development Breakthrough Program of China (No. 222102210191).

References

1. Jang, J.: Anfis: adaptive-network-based fuzzy inference system. IEEE Trans. Syst. Man Cybern. **23**, 665–685 (1993)
2. Chen, T., et al.: A decision tree-initialised neuro-fuzzy approach for clinical decision support. Artif. Intell. Med. **111**, 101986 (2021)
3. Keikhosrokiani, P., Naidu, A., Anathan, A.B., Iryanti Fadilah, S., Manickam, S., Li, Z.: Heartbeat sound classification using a hybrid adaptive neuro-fuzzy inferences system (anfis) and artificial bee colony. Digital Health **9**, 85 (2023). https://doi.org/10.1177/20552076221150741
4. Chen, T., et al.: A dominant set-informed interpretable fuzzy system for automated diagnosis of dementia. Front. Neurosci. **16**, 867664 (2022)
5. Zand, J.P., Katebi, J., Yaghmaei-Sabegh, S.: A generalized ANFIS controller for vibration mitigation of uncertain building structure. Struct. Eng. Mech. **87**, 231–242 (2023)
6. Osheba, D.S., Osheba, S., Nazih, A., Mansour, A.S.: Performance enhancement of PV system using VSG with ANFIS controller. Electr. Eng. **105**, 2523–2537 (2023)
7. Arévalo, P., Cano, A., Jurado, F.: Large-scale integration of renewable energies by 2050 through demand prediction with ANFIS, ECUADOR case study. Energy **286**, 129446 (2024)
8. Lara-Cerecedo, L., Pitalúa-Díaz, N., Hinojosa-Palafox, J.: Comparative study of the prediction of electrical energy from a photovoltaic system using the intelligent systems ANFIS and ANFIS-GA. Revista Mexicana de Ingeniería Química **22**, 1–16 (2023)
9. Lara-Cerecedo, L.O., Hinojosa, J.F., Pitalúa-Díaz, N., Matsumoto, Y., González-Angeles, A.: Prediction of the electricity generation of a 60-kw photovoltaic system with intelligent

models ANFIS and optimized ANFIS-PSO. Energies **16**, 6050 (2023)
10. MathWorks: neuro-adaptive learning and anfis - r2023a (2023). https://uk.mathworks.com/help/fuzzy/neuro-adaptive-learning-and-anfis.html. Accessed 5 Jan 2024
11. Power, J.: Anfis in pytorch (2019). https://github.com/jfpower/anfis-pytorch. Accessed 5 Jan 2024
12. Meggs, T.: Anfis (2020). https://github.com/twmeggs/anfis. Accessed 5 Jan 2024
13. Gilardi, G.: Anfis (2021). https://github.com/gabrielegilardi/ANFIS. Accessed 5 Jan 2024
14. Rathnayake, N., Dang, T.L., Hoshino, Y.: A novel optimization algorithm: cascaded adaptive neuro-fuzzy inference system. Int. J. Fuzzy Syst. **23**, 1955–1971 (2021)
15. Rathnayake, N., Rathnayake, U., Dang, T.L., Hoshino, Y.: A cascaded adaptive network-based fuzzy inference system for hydropower forecasting. Sensors **22**, 2905 (2022)
16. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
17. Talpur, N., et al.: Deep neuro-fuzzy system application trends, challenges, and future perspectives: a systematic survey. Artif. Intell. Rev. **56**, 865–913 (2023)
18. Cui, Y., Wu, D., Jiang, X., Xu, Y.: Pytsk: a python toolbox for tsk fuzzy systems. arXiv preprint arXiv:2206.03310 (2022). https://doi.org/10.48550/arXiv.2206.03310
19. Ketkar, N., Moolayil, J.: Deep Learning with Python: learn best practices of deep learning models with PyTorch 2 edn Apress. Berkeley, CA (2021)
20. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its applications to modeling and control. IEEE Trans. Syst. Man Cybern. **SMC–15**, 116–132 (1985)
21. Fresno, C., Fernández, E.A.: Anfis vignette (2012). https://github.com/jfpower/anfis-pytorch/blob/master/Anfis-vignette.pdf. Accessed 5 Jan 2024
22. Chen, T., Shang, C., Su, P., Shen, Q.: Induction of accurate and interpretable fuzzy rules from preliminary crisp representation. Knowl.-Based Syst. **146**, 152–166 (2018)
23. Carter, J., Chiclana, F., Khuman, A.S., Chen, T. (eds.): Fuzzy logic: recent applications and developments, 1st edn. Springer, Switzerland (2021)
24. Liebscher, R.: Pyfuzzy-python fuzzy package (2014). http://pyfuzzy.sourceforge.net/. Accessed 5 Jan 2024
25. Avelar, E., Castillo, O., Soria, J.: Fuzzy logic controller with fuzzylab python library and the robot operating system for autonomous robot navigation: a practical approach. Intuit Type-2 Fuzzy Logic Enhanc. Neural Optim. Algor. Theory Appl. **862**, 355–369 (2020)
26. Scikit-fuzzy (2023). https://pythonhosted.org/scikit-fuzzy/. Accessed 5 Jan 2024
27. Spolaor, S., et al.: Simpful: a user-friendly python library for fuzzy logic. Int. J. Comput. Intell. Syst. **13**, 1687–1698 (2020)
28. Karaboga, D., Kaya, E.: Adaptive network based fuzzy inference system (anfis) training approaches: a comprehensive survey. Artif. Intell. Rev. **52**, 2263–2293 (2019)
29. Oliphant, T.E.: Python for scientific computing. Comput. Sci. Eng. **9**, 10–20 (2007)
30. Oliphant, T.E., et al.: A guide to NumPy, vol. 1. Trelgol Publishing, USA (2006)
31. Zheng, Y., Xu, Z., Wang, X.: The fusion of deep learning and fuzzy systems: a state-of-the-art survey. IEEE Trans. Fuzzy Syst. **30**, 2783–2799 (2022)
32. Sun, C., Jang, J.: A neuro-fuzzy classifier and its applications. In: Proceedings Second IEEE International Conference on Fuzzy Systems (pp. 94-98). IEEE (1993)
33. Talpur, N., Abdulkadir, S.J., Hasan, M.H.: A deep learning based neuro-fuzzy approach for solving classification problems, 167–172 IEEE, (2020)
34. Wu, D., Yuan, Y., Huang, J., Tan, Y.: Optimize tsk fuzzy systems for regression problems: Minibatch gradient descent with regularization, droprule, and adabound (mbgd-rda). IEEE Trans. Fuzzy Syst. **28**, 1003–1015 (2020)
35. Cui, Y., Wu, D., Huang, J.: Optimize tsk fuzzy systems for classification problems: Minibatch gradient descent with uniform regularization and batch normalization. IEEE Trans. Fuzzy Syst. **28**, 3065–3075 (2020)
36. Shi, Z., et al.: Fcm-rdpa: Tsk fuzzy regression model construction using fuzzy c-means clustering, regularization, droprule, and powerball adabelief. Inf. Sci. **574**, 490–504 (2021)
37. Cui, Y., Xu, Y., Peng, R., Wu, D.: Layer normalization for tsk fuzzy system optimization in regression problems. IEEE Trans. Fuzzy Syst. **31**, 254–264 (2022)
38. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: generalization gap and sharp minima. arXiv preprint arXiv:1609.04836 (2016). https://doi.org/10.48550/arXiv.1609.04836
39. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014). https://doi.org/10.48550/arXiv.1412.6980
40. Luo, L., Xiong, Y., Liu, Y., Sun, X.: Adaptive gradient methods with dynamic bound of learning rate. arXiv preprint arXiv:1902.09843 (2019). https://doi.org/10.48550/arXiv.1902.09843
41. Yuan, Y., Li, M., Liu, J., Tomlin, C.: On the powerball method: variants of descent methods for accelerated optimization. IEEE Control Syst. Lett. **3**, 601–606 (2019)
42. Zhuang, J., et al.: Adabelief optimizer: adapting stepsizes by the belief in observed gradients. Adv. Neural. Inf. Process. Syst. **33**, 18795–18806 (2020)
43. Bottou, L.: Large-scale machine learning with stochastic gradient descent, pp. 177–186. Springer, Berlin (2010)
44. Riedmiller, M., Braun, H.: A direct adaptive method for faster backpropagation learning: The rprop algorithm. In: IEEE International Conference on Neural Networks, pp. 586–591 IEEE, (1993)
45. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Math. Program. **45**, 503–528 (1989)
46. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012). https://doi.org/10.48550/arXiv.1212.5701
47. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011)
48. Pathak, A.: Restaurant tipping problem using fuzzy logic (2023). https://github.com/ap1904/RTP. Accessed 5 Jan 2024
49. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**, 179–188 (1936)
50. Turki, M., Bouzaida, S., Sakly, A., M'Sahli, F.: Adaptive control of nonlinear system using neuro-fuzzy learning by pso algorithm. pp. 519–523 IEEE, (2012)
51. Cárdenas, J.J., García, A., Romeral, J., Kampouropoulos, K.: Evolutive ANFIS training for energy load profile forecast for an IEMS in an automated factory. In: ETFA2011, pp. 1–8 (IEEE, 2011)
52. Chen, T., Shang, C., Yang, J., Li, F., Shen, Q.: A new approach for transformation-based fuzzy rule interpolation. IEEE Trans. Fuzzy Syst. **28**, 3330–3344 (2019)
53. Stathakis, D., Savina, I., Nègrea, T.: Neuro-fuzzy modeling for crop yield prediction. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. **34**, 1–4 (2006)
54. Li, W., et al.: Feature selection approach based on improved fuzzy c-means with principle of refined justifiable granularity. IEEE Trans. Fuzzy Syst. **31**, 2112–2126 (2023)

55. Su, P., et al.: Corneal nerve tortuosity grading via ordered weighted averaging-based feature extraction. Med. Phys. **47**, 4983–4996 (2020)
56. Li, W., et al.: Double-quantitative feature selection approach for multi-granularity ordered decision systems. IEEE Trans. Artif. Intell. **1**, 1–12 (2023)
57. Mac Parthaláin, N., Jensen, R., Diao, R.: Fuzzy-rough set bireducts for data reduction. IEEE Trans. Fuzzy Syst. **28**, 1840–1850 (2019)
58. Li, W., Zhou, H., Xu, W., Wang, X.-Z., Pedrycz, W.: Interval dominance-based feature selection for interval-valued ordered data. IEEE Trans. Neural Netw. Learn. Syst. **34**, 6898–6912 (2023)
59. Chen, T., Carter, J., Mahmud, M., Khuman, A.S.: Artificial intelligence in healthcare: recent applications and developments, vol. 1. Springer Nature, Singapore (2022)

Dongsong Zhang received the PhD degree in Computer Science and Technology from National University of Defense Technology, China, in 2012. He currently is an assistant professor in School of Big Data and Artificial Intelligence at Xinyang College. His research interests are in the area of soft computing (Neural Network, Fuzzy Logic), Real-Time Systems.

Tianhua Chen received the Ph.D. degree in Computational Intelligence from Aberystwyth University, Aberystwyth, U.K., in 2017. He is currently a Reader (Associate Professor) in Artificial Intelligence with the Department of Computer Science, School of Computing and Engineering, University of Huddersfield, UK. He has published over 60 peer reviewed papers in leading international journals and conferences, including a lead-authored paper selected as IEEE Transactions on Fuzzy System Publication Spotlight Article by IEEE Computational Intelligence Society. His research interests are: Artificial Intelligence for health and wellbeing, Explainable AI, Neuro-Fuzzy Systems. Tianhua is an Editorial Board Member of Artificial Intelligence in Medicine journal (Elsevier), BMC Medical Informatics and Decision Making journal (Springer), and PLOS ONE.

3 成员 7 发表的 SCI 论文：Cascade Aggregation Network for Accurate Polyp Segmentation

IET Systems Biology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Cascade Aggregation Network for Accurate Polyp Segmentation

Yanru Jia[1] | Yu Zeng[2] | Huaping Guo[2]

[1]School of Big Data and Artificial Intelligence, Xinyang University, Xinyang, China | [2]School of Computer and Information Techonology, Xinyang Normal University, Xinyang, China

Correspondence: Huaping Guo (hpguo@xynu.edu.cn)

Received: 3 June 2025 | **Revised:** 14 August 2025 | **Accepted:** 26 August 2025

Handling Editor: Grace Wang

Funding: This work was supported by Research Project on the Curriculum Reform of Teacher Education in Henan Province (Grant 2025-JSJYZD-052).

Keywords: cascade aggregation | multiscale context aware | polyp segmentation

ABSTRACT

Accurate polyp segmentation is crucial for computer-aided diagnosis and early detection of colorectal cancer. Whereas feature pyramid network (FPN) and its variants are widely used in polyp segmentation, inherent limitations existing in FPN include: (1) repeated upsampling degrades fine details, reducing small polyp segmentation accuracy and (2) naive feature fusion (e.g., summation) inadequately captures global context, limiting performance on complex structures. To address limitations, we propose a cascaded aggregation network (CANet) that systematically integrates multi-level features for refined representation. CANet adopts PVT transformer as the backbone to extract robust multi-level representations and introduces a cascade aggregation module (CAM) that enriches semantic features without sacrificing spatial details. CAM adopts a top-down enhancement pathway, where high-level features progressively guide the fusion of multiscale information, enhancing semantic representation while preserving spatial details. CANet further integrates a multiscale context-aware module (MCAM) and a residual-based fusion module (RFM). MCAM applies parallel convolutions with diverse kernel sizes and dilation rates to low-level features, enabling fine-grained multiscale extraction of local details and enhancing scene understanding. RFM fuses these local features with high-level semantics from CAM, enabling effective cross-level integration. Experiments show that CANet outperforms SOTA methods in in- and out-of-distribution tests.

1 | Introduction

Colorectal cancer (CRC) ranks among the most prevalent and deadly cancers globally, representing 10% of all cancer-related deaths [1]. Notably, studies report polyp miss rates of 20%–30% during standard colonoscopies [2]. This high miss rate is likely related to the complex bowel structure and the difficulty in detecting small lesions [3]. Therefore, developing more accurate auxiliary detection technologies is crucial for improving diagnostic efficiency.

In recent years, with the rapid advancement of deep learning technologies, feature pyramid networks (FPNs) have emerged as a powerful approach in computer vision, particularly for medical image segmentation tasks [4]. By integrating feature maps from multiple layers, FPN effectively captures both global context and fine-grained local details, significantly improving segmentation accuracy, especially in scenarios with complex backgrounds and variable target morphologies. As a typical FPN model, U-Net [5] and its improved versions (such as Unet++ [6] and U-Net 3+ [7]) have become foundational in

4 成员 7 发表的 SCI 论文：Pedestrian re-recognition based on spatiotemporal Transformer skeleton contrastive learning and feature optimization

Select type of manuscript: Research Paper

Pedestrian re recognition based on spatiotemporal Transformer skeleton contrastive learning and feature optimization

Yanru Jia*, Yuanyuan Zhang,***,****, †, and Yilun Gao**,***,****

*School of Big Data and Artificial Intelligence, Xinyang College
7th New Avenue West, Xinyang, Henan 464000, China
E-mail:jiayanru8888@163.com
**School of Automation, China University of Geosciences
388 Lumo Road, Hongshan District, Wuhan, Hubei 430074, China
E-mail:{1202221794, gylnew0106}@cug.edu.cn
***Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems
388 Lumo Road, Hongshan District, Wuhan, Hubei 430074, China
****Engineering Research Center of Intelligent Technology for Geo-Exploration, Ministry of Education
388 Lumo Road, Hongshan District, Wuhan, Hubei 430074, China
† Corresponding author

Person Re-identification (Re-ID) is an important task in computer vision, aimed at achieving cross camera identity confirmation by identifying and matching the same pedestrian under different cameras. However, when traditional image-based methods are affected by factors such as lighting changes, occlusion, and changes in viewing angles, the advantages of skeleton data become increasingly apparent. Existing methods typically use primitive body joint design skeleton descriptors or learn skeleton sequence representations, but they often cannot simultaneously simulate the relationships between different body components, and rarely model skeleton information from both temporal and spatial dimensions. Therefore, in this paper, we propose a universal skeleton contrastive learning method based on spatiotemporal Transformer (Space time Transformer, StFormer). The method first adopts the Space time Attention (S-TAttention) mechanism and achieves relationship modeling of spatiotemporal features by stacking multiple S-TAttention blocks. Secondly, to improve the important clues for extracting data features from the model, a Feature Refinement Box (FR Box) was proposed. Finally, we purpose a unique prompt learning mechanism (P-Study) which utilizes the spatiotemporal context of graph nodes to prompt skeleton graph reconstruction and help capture more valuable patterns and graph semantics.

Keywords: Pedestrian re identification, Transformer, Comparative learning, Prompt learning, Skeleton recognition

1. Introduction

Pedestrian re identification (re ID) is a challenging task that requires retrieving and matching specific individuals from different perspectives or scenes, providing support for many important applications such as secure identity verification [1] [2] [3], human tracking [4] [5], and robotics [6]. However, there are issues with spatiotemporal alignment [7] [8], image resolution [9], changes in human posture [10] [11], and foreign object occlusion [12] in pedestrian re identification tasks.

To address these issues, compared to traditional methods that rely on visual appearance features such as color and contour [13] [14], researchers tend to choose using human skeleton information as the discriminative feature for pedestrian re identification. This feature not only makes the final recognition result more robust, but also outperforms traditional methods that rely on visual features in terms of background adaptability, resistance to light changes, noise, and minimum computational cost based on skeleton data.

Traditional skeleton data methods typically rely on manually extracted body shape descriptors and gait attributes, or neural network-based methods (such as CNN, LSTM) [15] [16] [17]to model body and motion features. However, these methods often overlook the relationships between body parts within the skeleton (such as the motion correlation between joints), and cannot fully explore the potential valuable patterns in the skeleton data. And the skeleton representation extracted by the model lacks important interaction objects and contextual information, leading to difficulties in recognizing similar actions. Actions such as "writing", "reading", and "typing on the keyboard"

are difficult to distinguish based solely on the skeleton diagram.

In order to solve the above problems, this paper proposes a skeleton recognition model based on spatiotemporal Transformer, which improves the traditional Transformer network by adopting a new spatiotemporal attention mechanism to replace the original self-attention mechanism in the network. By dividing the input features into temporal and spatial

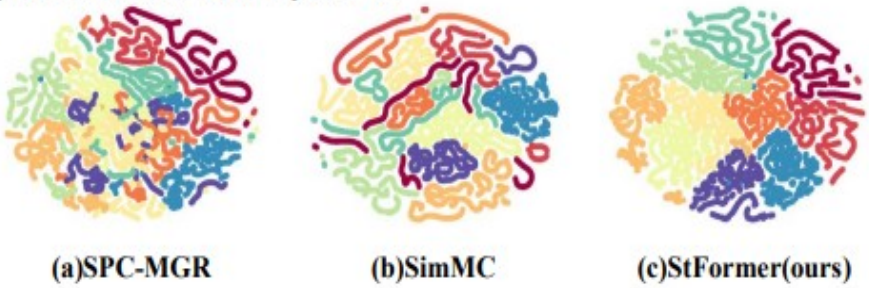


Fig. 1. T-SNE visualization of BIWI category representations learned from different models (different colors represent different representations).(a)SPC-MGR[40](b)SimMC[32](c)proposed method: StFormer

dimensions according to the feature dimension, and using two branches to process the joint features of the two dimensions in parallel, the information of the two branches is finally mixed to output the features with spatiotemporal context in the skeleton.At the same time, in order to enable the model to extract more discriminative skeleton features, we also propose a feature reinforcement module based on positive and negative sample learning. The main approach is to refine the skeleton features output by the Transformer module through contrastive learning. For models with two similar labels, positive and negative samples are used for "convergence" and "separation" operations.In Fig.1, we visualize the representations learned by different models on the dataset. It can be seen from the figure that the StFormer model with added feature optimization module can more clearly capture the most discriminative features in the data. In order to improve the generalization ability of the model and enable the pedestrian re recognition model to capture more valuable patterns and graph semantics in the skeleton, we adopt the method of prompt learning. By designing two contextual prompts for the skeleton graph, we reconstruct the skeleton graph and promote the model to learn more representative features for pedestrian re recognition tasks.

The main innovations of this article are as follows：

• A pedestrian re identification model based on the dual dimensions of time and space in Transformer has been designed. By using spatiotemporal attention mechanism instead of the traditional self-attention mechanism in Transformer, features with spatiotemporal characteristics in the skeleton are extracted.

• A feature enhancement module was proposed to refine the skeleton features output by the Transformer module through contrastive learning. By performing "convergence" and "separation" operations on labels with similarity, the model's ability to extract discriminative skeleton features is enhanced, and the discriminability and recognition of features are improved.

• To enhance the generalization ability of the model and enable it to better capture valuable patterns and graph semantics in the skeleton during pedestrian re recognition tasks, a prompt learning method is adopted to reconstruct the skeleton graph by designing contextual prompts.

2. Analysis of pedestrian re identification method based on skeleton

With the rapid development of deep learning, traditional skeleton action recognition methods are gradually being replaced by Graph Convolutional Networks (GCNs), which can better utilize skeleton structure and time series information to improve recognition performance. With the continuous advancement of deep learning, attention-based Transformer networks have significantly improved the accuracy of action understanding by focusing on the physical dependencies between human joints. In pedestrian re identification tasks, Transformer further enhances the performance and accuracy of the model by focusing on key features.

2.1. Pedestrian Re identification Method Based on Transformer

Dosovitskiy et al. proposed the ViT (Vision Transformer) model [18], which is a network structure based entirely on self-attention mechanism. It has been proven that on large-scale datasets, a standalone Transformer can perform well in classification tasks without relying on CNN. ViT has applied the Transformer model from NLP to image classification

5 成员 7 发表的 SCI 论文：A Novel Watermarking Scheme for Audio Data Stored in Third Party Servers

International Journal of Digital Crime and Forensics
Volume 16 • Issue 1

A Novel Watermarking Scheme for Audio Data Stored in Third Party Servers

Fuhai Jia, Xinyang University, China
Yanru Jia, Xinyang University, China
Jing Li, Xinyang University, China
Zhenghui Liu, Xinyang Normal University, China*

ABSTRACT

To improve the security and privacy of audio data stored in third party servers, a novel watermarking scheme is proposed. Firstly, the authors split the host signal into frames and scramble each frame to get the encrypted signal. Secondly, they generate watermark bits by using the frame number and embed them into each frame of the encrypted signal, which is the data that will be uploaded to the third party servers. For the users, they can download the encrypted data and verify the data is intact or not. If the data is intact, the users decrypt the data to get the audio signal. If the audio signal is attacked in the process of transmission, they can also locate the location of the attacked frame. The experimental results show that the method proposed is effective not only for encrypted signals, but also for the encrypted signals after decryption.

KEYWORDS

Content Security, Digital Audio, Digital Forensics, Watermarking

INTRODUCTION

The development of digital signal processing technology facilitated communication among individuals. However, this progress has also increased concerns regarding the potential leakage of users' private data. For example, the popularity of recording devices has empowered individuals to create their own audio signals. Yet, managing a large volume of audio signals poses a problem to consider for the owners of the signals in terms of storage. In pursuit of convenience, some upload their works to third-party storage centers. However, entrusting their data to external storage centers exposes their works to potential threats, as these centers operate outside of their control (Kuang et al., 2020; Razali et al., 2021). To improve the security of the data stored in third-party centers, a watermarking algorithm is proposed in this article.

The field of digital watermarking technology has seen more than 10 years of research, with many studies exploring its application and methods (Hua et al., 2016). Generally speaking, digital watermarking schemes use the redundancy in audio signals and auditory insensitivity of human ears to embed watermark bits into the host signal without degrading the quality. These schemes can be categorized based on their different purposes.

DOI: 10.4018/IJDCF.340382
*Corresponding Author

One category, called robust digital watermarking, is used in copyright protection (Chen et al., 2018; Jiang et al., 2019; Kosta et al., 2022; Salah et al., 2021). The other category, called fragile or semi-fragile digital watermarking, is used for forensic purposes (Chen et al., 2010).

In Yong et al. (2014), a robust audio watermarking scheme was proposed, where the authors embedded watermark bits and synchronization codes into the host signal to generate the watermarked signal. During decoding, users could determine whether the signal had been scaled by observing changes in the synchronization code's position. If scaling was detected, users could calculate the scale factor, allowing them to reduce the impact of attacks and improve the scheme's robustness.

Liu et al. (2021) proposed an audio watermarking algorithm for tracing the re-recorded audio sources. In their work, the authors introduced the LMC feature and conducted an analysis of its characteristics. Then, the authors embedded watermark bits by quantizing discrete cosine transform (DCT) intermediate frequency coefficients to quantize LMC features. The LMC feature has robustness against re-recording attacks, enabling the scheme to accurately extract correct watermark bits from the attacked signals.

In Liu et al. (2022), an audio watermarking scheme for encrypted audio was introduced, addressing a relatively unexplored are in audio watermarking. The authors cut the host signal into frames and then scrambled each frame to generate encrypted audio. Then, they embedded the frame number by quantifying the signal energy ratio into the encrypted frame. This approach enables the scheme to identify the tampered location in the attacked signal, allowing for the substitution of attacked frames with 0 amplitude samples to reconstruct the signal.

However, a limitation of the scheme proposed in Liu et al. (2022) is that if downloaded data is intact, users can decrypt the audio signal, placing it in an unprotected range. Consequently, if the decrypted signal is attacked during the transmission, the scheme lacks the ability to verify its integrity.

To solve the above problems and improve the security of the encrypted audio signals, this article proposes a novel watermarking scheme. Initially, the host signal is encrypted, followed by the embedding of watermark bits into the encrypted signal. The process begins with segmenting the host signal into frames, each of which is then scrambled to produce the encrypted signal. Then, binary bits representing the frame numbers are embedded into the frames of the encrypted signal to generate the watermarked data, which is uploaded to third-party servers.

If users download the watermarked data, they can divide the data into frames and verify their integrity. Intact frames can then be decrypted to retrieve the original audio signal, enabling direct comprehension by users. Besides, if the decrypted signal is attacked, the scheme can verify the authentication of the attacked signal and locate the compromised frames. The main contributions of this article are described as follows:

- The study presents the encryption and decryption methods of audio signals, and defines the feature of encrypted audio signal. Then, the study designs the watermark embedding method by quantifying the feature.
- The study proposes a novel watermarking scheme based on the defined feature. The scheme not only protects large audio signals stored on third-party servers but also verifies downloaded data integrity. Furthermore, the scheme provides an authentication method for audio signals post-decryption.

The article is organized as follows. The next section introduces the encryption method for host signal. Then, the study describes the proposed scheme, watermark generation, and methods for embedding and extraction. The scheme's performance is then reviewed before the study's conclusion is summarized.

ENCRYPTION

Denote A as the L length speech signal, $A = \{a_l, 1 \le l \le L\}$, where a_l is the l-th sample. Based on the logistic chaotic map in equation (1), the L length pseudo-random sequence is obtained, denoted by $Y = \{y_l | l = 1, 2, \cdots, L\}$. In equation (1), k is the initial value, serving as the key of the watermarking system, $3.5699 \le \mu \le 4$ (Liu et al., 2022). Signal A is encrypted using the following steps:

$$y_{l+1} = \mu y_l(1 - y_l), y_0 = k \tag{1}$$

1. Signal A is cut into P frames. The i-th frame is denoted by $A_i = \{a_{i,t} | t = 1, 2, \cdots, L/P\}$.
2. The first L/P length sequence of Y is selected, denoted by $Y_1 = \{y_t | t = 1, 2, \cdots, L/P\}$. Then, the elements in Y_1 is assorted in ascending order based on equation (2), where $h(t)$ is the address index of the sorted chaotic sequence.

$$y_{h(t)} = \text{ascend}(y_t), t = 1, 2, \cdots L/P \tag{2}$$

3. Each frame A_i is scrambled as $1 \le p \le P$, denoting the scrambled signal as B_i. $B_i = \{b_{i,t} | t = 1, 2, \cdots, L/P\}$ is denoted, where $b_{i,t}$ is the t-th sample after being scrambled (see equation (3)). If B is denoted as the encrypted signal, then $B = \{B_1 \cup B_2 \cup \cdots \cup B_P\}$.

$$b_{i,t} = a_{i,h(t)}, \ t = 1, 2, \cdots L/P \tag{3}$$

THE SCHEME

To protect the audio signals stored on third-party servers, the study proposes a watermarking scheme. This scheme serves a dual purpose: it protects lager audio signals stored on these servers and verifies the integrity of data downloaded from them. At the same time, the scheme provides authentication for audio signals after decryption.

To effectively detect and locate attacked signals, the frame number is encrypted into the encrypted data. Then, the frame number is extracted from the attacked content.

Watermark Generation

Based on the previous section, the audio signal is encrypted as $B = \{B_1 \cup B_2 \cup \cdots \cup B_P\}$, in which B_i is the i-th frame of the encrypted signal. The frame number of B_i is i. The frame number i is converted into M length binary bits, denoted by $W_i = \{w_{i,1}, w_{i,2}, \cdots, w_{i,M}\}$. If the length is less than M, 0 is added to satisfy the length requirement. For the first frame B_1, the frame number is 1. It is converted to 0000000001 (set $M = 10$ in this article). Thus, $W_1 = \{0, 0, 0, 0, 0, 0, 0, 0, 0, 1\}$.

Watermark Embedding

$B = \{B_1 \cup B_2 \cup \cdots \cup B_P\}$ is obtained based on the encrypted signal. The following takes the embedding of W_1 into the first frame B_1 as an example to introduce the embedding method.

6 主持人发表的 SCI 论文 Lin Jinzhu, Ni Tianwei. CMMF and STAM-FNet: Multimodal Fusion Architectures for Complex Scene Understanding in Dynamic Environments. Informatica (Slovenia)

报告编号：J20265001247571147

报告验真

检 索 报 告

检索主题：林金珠发表论文收录情况

委 托 人：林金珠

检索范围：EI

检索时间：2026年3月10日

检索结果：

根据委托人本次委托要求，在上述检索范围内，林金珠发表论文收录情况如下表：

委托要求		检索结果
检索范围	委托篇数	收录篇数
EI	1	1

科学技术部西南信息中心查新中心

2026年3月10日

231-246.

https://doi.org/10.31449/inf.v49i9.9830 Informatica 49 (2025) 231–246 231

CMMF and STAM-FNet: Multimodal Fusion Architectures for Complex Scene Understanding in Dynamic Environments

Jinzhu Lin*, Tianwei Ni
School of Big Data and Artificial Intelligence, Xinyang College, Xinyang 464000, Henan, China
Corresponding author's E-mail: linjinzhu622@outlook.com
*Corresponding author

Keywords: multimodal fusion technology, complex scene understanding, attention mechanism, mode collaboration

Received: June 23, 2025

Multimodal perception has emerged as a vital strategy for understanding complex and dynamic environments, where traditional unimodal approaches fail to handle data heterogeneity and occlusion. This paper proposes two multimodal fusion frameworks—CMMF (Cross-Modal Matching Fusion) and STAM-FNet (Spatio-Temporal Attention Multimodal Fusion Network)—to address structural and temporal challenges in complex scene understanding. The CMMF model adopts a three-stage architecture with cross-modal semantic alignment and dynamic weighting, while STAM-FNet introduces spatio-temporal attention layers and 3D convolutions to enhance feature discrimination in dynamic environments. Experiments are conducted on a dataset of 120000 samples covering three application scenarios: urban monitoring, indoor interaction, and transportation hubs. Evaluation is based on standardized metrics including Top-1 Accuracy, F1-score, AUC, Modal Gain Index, and Inference Delay. Compared to SOTA baselines such as ResNet50, Two-Stream Transformer, and MMBT, STAM-FNet achieves up to 15.8% improvement in accuracy and 20% robustness gain under high-occlusion conditions. CMMF maintains superior performance in static tasks while preserving low parameter count (24.3M). This work demonstrates the effectiveness of adaptive multimodal fusion in improving accuracy, efficiency, and fault tolerance in real-world perception systems.

Povzetek: Opisana sta modela za razumevanje kompleksnih prizorov: CMMF (Cross-Modal Matching Fusion) in STAM-FNet (Spatio-Temporal Attention Multimodal Fusion Network). CMMF izvaja uteženo križno-modalno usklajevanje in je optimiran za statične naloge (24,3 M parametrov), medtem ko STAM-FNet z uporabo 3D-konvolucij in prostorsko-časovne pozornosti dosega vrhunske rezultate v dinamičnih okoljih.

1 Introduction

Semantic understanding of complex scenes is crucial for intelligent perception systems. Traditional single-modal methods face limitations under dynamic environments, multi-source coupling, and heterogeneous data. In scenarios like urban security and medical navigation, relying solely on vision or audio often fails to ensure stable recognition. Multi-modal fusion has emerged as an effective solution due to its complementary and synergistic capabilities. Recent advances in deep learning-based cross-modal representation offer strong modeling foundations. However, issues like modality inconsistency, rigid fusion strategies, and poor adaptability to dynamic scenes remain, hindering further performance improvement in real-world applications.

Focusing on the robustness and adaptability of modal fusion mechanism in complex scenes, this study proposes two complementary model design ideas. The first model focuses on the collaborative representation of modal features, and builds a multi-layer matching network based on global weighting strategy. The second model introduces spatio-temporal attention mechanism to strengthen the ability to pay attention to effective features in dynamic changing scenes. The research integrates data preprocessing, model architecture, index design and experimental setup, and constructs a research framework covering the whole process of perception, modeling and verification. By designing a unified comparative experiment, the performance differences of the model under different occlusion ratios and different task complexity are clarified, and the boundary characteristics of multimodal understanding under real and complex conditions are tried to be restored.

At present, the research of multimodal fusion technology in complex scene understanding is expanding, showing the development trend of diversification of model mechanism and refinement of task structure. Zhang et al. (2025) put forward EKLI-Attention mechanism, which classifies citizens' government requests by integrating local and global attention, indicating that multilevel attention mechanism is operable and efficient in actual semantic recognition [1].

x

Table 1: Performance comparison of representative multimodal models in complex scene understanding

Model	Top-1 Accuracy (%)	Occlusion Robustness (Drop @ 75%)	Params (M)	Inference Delay (ms)	Notable Features
ResNet50 (Image-only)	78.4	-28.9	25.6	11.2	Baseline single-modal CNN
MMBT	84.3	-17.1	72.4	16.5	Early-fusion Transformer
Two-Stream Transformer	86.7	-13.9	88.1	18.3	Dual-modal attention mechanism
CMMF (Proposed)	91.3	-10.2	24.3	9.6	Cross-modal weighted feature fusion
STAM-FNet (Proposed)	93.2	-6.1	31.5	7.8	Spatio-temporal attention+3D conv

To guide this research, two core questions are posed:
(RQ1) Can a spatio-temporal attention mechanism significantly enhance the effectiveness of multimodal fusion in dynamic and occluded environments?
(RQ2) Can the proposed models—CMMF and STAM-FNet—achieve at least a 10% improvement in recognition robustness under severe occlusion conditions compared to established SOTA baselines such as MMBT and Two-Stream Transformer?
These questions aim to quantify the benefit of architectural innovations and validate the models' practical contributions. The study is designed to evaluate these hypotheses across diverse real-world scenes, using standardized evaluation protocols and performance benchmarks. Addressing these questions allows for targeted analysis of model strengths and shortcomings and frames the empirical work in a hypothesis-driven structure.

2 Materials and methods

2.1 Multi-modal data acquisition and preprocessing

2.1.1 Data source composition and sampling strategy

The research uses data sets including image, voice and text, covering three typical application fields: traffic scene, indoor identification and public safety monitoring. The image data comes from a multi-view camera with a unified resolution of 640×480. The audio clip is taken from the real sound pickup device, the frequency is 16kHz, and the length is controlled within 8 seconds. Text data is encoded in UTF-8 format based on phonetic transcription or user interaction information, and Chinese sentence breaking and English punctuation are adopted. The sampling process is distributed hierarchically according to hours, scenes and task types to avoid sample deviation and redundant collection [18]. All modes are marked with time stamps to ensure the accuracy of cross-modal semantic alignment and reconstruction. The whole data acquisition process introduces task classification index identification, which is used for task grouping and label scheduling in the later model training. The sampling strategy emphasizes the balance between representativeness and complexity, preserves the continuous fragments in highly dynamic scenes, and improves the generalization ability of subsequent models in real tasks.

2.1.2 Normalization of images, texts and audio.

In preprocessing, original images are uniformly resized, pixel-normalized, and color channels reordered. Adaptive histogram equalization is applied under varying lighting to enhance contrast and edge clarity. Audio signals are processed using short-time Fourier transform, with abnormal-length samples padded or truncated, and normalized to reduce background noise. Phonetic text is processed via Chinese word segmentation, stop-word removal, and word vector encoding, forming semantic tensors for fusion input. All modal data are batch-processed to optimize pipeline efficiency and reduce latency. Text segmentation respects natural sentence structure to minimize semantic errors. A unified format and parameter standard is adopted for cross-modal data, ensuring comparability at the distribution level. This

preprocessing chain establishes consistency across modalities, supporting effective feature extraction and alignment in downstream tasks.

2.1.3 Multimodal time alignment mechanism and redundancy elimination

In order to ensure the accuracy of multimodal fusion, the data alignment strategy is based on the global timestamp unification mechanism. Image frames and audio frames are aligned at the frame level through linear interpolation and synchronous sampling. For the delay between speech transcription and image events, a dynamic window mechanism is set to carry out semantic matching and time slip compensation. The inter-modal time offset rate is controlled within ±150ms, which meets the real-time requirements of most sensing tasks [19]. The redundant fragments that can't be synchronized are silently discarded, and the key frames before and after are reserved to maintain the context integrity. Information redundancy in text data is mainly manifested as logical repetition or structural repetition, which is uniformly filtered after being judged by the editing distance threshold. The final preserved data set is consistent in both time axis and semantic layer. Alignment mechanism can adapt to irregular event flow and dynamic scenes, and maintain stable performance under high-density sampling conditions, which is a key pre-step to ensure the quality of model time series modeling.

2.1.4 Noise filtering and high-dimensional noise reduction methods

The data collected in complex environment is often accompanied by strong noise interference. In this study, a multi-stage noise reduction mechanism is introduced in the pretreatment stage. In image mode, random pixel noise is processed by Gaussian filtering, and then texture anomalies are removed by edge preserving filtering. The audio mode uses spectral subtraction and voice activity detection methods to remove background noise and mute segments [20]. In text mode, low-information or non-task-related sentences are filtered by word frequency and TF-IDF index. On the feature space level, PCA and self-encoder are introduced to reduce the dimension of high-dimensional features of each mode, while retaining the principal components of semantic information. The data after dimensionality reduction will be normalized again before entering the main model to avoid abnormal numerical amplification error. The noise control strategy can effectively improve the model processing efficiency and enhance the adaptability to abnormal data distribution on the premise of ensuring information integrity.

To clarify the terminology, the study involves five core multimodal perception tasks: object recognition, action recognition, intent detection, semantic segmentation, and cross-modal matching. These tasks are performed across five representative complex scene categories: urban street, medical room, traffic platform, campus environment, and industrial workshop. Each task is not tied exclusively to a single scene but is instead evaluated under multiple environments to test generalization. For example, semantic segmentation and cross-modal matching are applied in the campus and traffic scenes, while action recognition and intent detection are emphasized in the medical and workshop contexts. This task–scene mapping ensures diverse multimodal challenges under real-world variability.

2.2 Multi-modal fusion model construction

2.2.1 CMMF structure and feature weighting mechanism

The CMMF model takes cross-modal matching as the core to build a fusion path, and strengthens the depth of information interaction by extracting the shared semantic subspace of each modal. The model is divided into three layers. The bottom layer completes modal self-coding, the middle layer realizes feature interaction between modes, and the high layer outputs fusion results. Image, text and audio modes are respectively input into three parallel convolution or Transformer coding channels, and then enter the weighted fusion module after unified mapping dimensions [21]. Feature weighting assigns dynamic weights based on modal reliability, and automatically adjusts participation according to information effectiveness and response strength. The output characteristics after fusion are as follows (1):

$$F_{fusion} = \sum_{i=1}^{N} \omega_i \cdot F_i \quad (1)$$

The output characteristics after fusion are defined as: F_i represents the feature vector of the i-th modality, and ω_i denotes its corresponding weight coefficient. These weights satisfy the normalization constraint: $\sum \omega_i = 1$, with $\omega_i \geq 0$ for all i.

This ensures that the fused representation maintains a probabilistic interpretation over modality contributions.

For CMMF, each modality input passes through a dedicated encoder: a 4-layer CNN for image data (kernel size: 3×3, ReLU activation, max pooling every two layers), a 2-layer BiLSTM for text (hidden size: 256), and a 3-layer 1D-CNN for audio (kernel size: 5, dropout rate: 0.3). All encoded features are mapped to a shared embedding space of 512 dimensions. The dynamic

feature weighting module uses softmax normalization over learned reliability scores. The output layer applies a fully connected layer followed by softmax for classification. Training uses Adam optimizer (lr=0.001), dropout=0.5, and batch size=64.

2.2.2 spatio-temporal attention mechanism in STAM-FNET

STAM-FNet aims to solve the problem that the fusion model does not respond to dynamic scenes in time, and uses the spatio-temporal attention mechanism to dynamically weight multimodal signals. Three-dimensional convolution and attention distribution modules are added to the model, and the spatial salience and temporal evolution characteristics are also learned [22]. After the feature flows through the local attention layer and the global gating layer, the region of interest is determined according to the temporal context [23]. This mechanism is especially suitable for scenes such as occlusion changes and sudden environmental changes, and can dynamically focus on key modal frames. The attention output is expressed by the following formula (2):

$$A(x,t) = \text{softmax}\left(\frac{Q(x)K(t)^T}{\sqrt{d_k}}\right)V(t) \quad (2)$$

Here, $Q(x)$ denotes the spatial query, K(t) the temporal key, $V(t)$ the value vector, and d_k the dimension of the key vectors used for scaling. This formulation ensures that the attention weights are normalized before being applied to the value representation, enhancing stability during training and interpretability in dynamic sequences. The original formulation has been revised to align with established attention mechanisms such as those used in Transformer architectures.

In STAM-FNet, each input is passed through a 3D-CNN backbone (3 layers, channels: 64-128-256, ReLU, batch normalization), followed by local and global attention modules. The spatio-temporal attention block includes 2 Transformer layers (hidden size: 512, 8 heads, GELU activation, dropout=0.1). The total loss is composed of classification loss (weight: 1.0), modal matching loss (weight: 0.6), and regularization (weight: 0.01). Early stopping is used if validation loss does not improve after 5 epochs.

2.2.3 Training optimization and loss construction of double models

To improve the overall synergy and generalization ability of the model, CMMF and STAM-FNet adopt a joint training mechanism. The training process adopts end-to-end strategy, and the objective function introduces multi-task structure, giving consideration to classification accuracy, modal alignment and time sequence stability. The total loss function of fusion training is designed as the following formula (3):

$$\text{L}total = \lambda cls \cdot \text{L}cls + \lambda align \cdot \text{L}align + \lambda reg \cdot \text{L}_{reg} \quad (3)$$

Here, λcls, $\lambda align$, and $\lambda regare$ scalar hyperparam eters that control the contribution of the classification, alignment, and regularization losses, respectively. Th ese coefficients are tuned using grid search on the val idation set to ensure balanced learning across sub-task s. This formulation ensures consistency across the ma thematical definition and explanatory text, facilitating clearer interpretation and reproducibility.

During training, the weight coefficients λcls, λalig n, and λreg are dynamically adjusted every five epoc hs based on the relative convergence rate of each sub -loss. Specifically, if the moving average of a sub-los s stagnates or decreases slower than others, its associ ated λ value is increased proportionally to prioritize l earning on that sub-task. A normalization step is appl ied to ensure that the $sum\ \lambda cls+\lambda align+\lambda reg\ =1holds$ at every update. This adaptive scheme enables the mo del to shift learning focus across modalities and task objectives depending on training dynamics, improving convergence and generalization in heterogeneous env ironments

2.2.4 Model difference design and integration strategy

CMMF is good at structural alignment, and STAM-FNet is better than time series modeling. In order to give full play to their complementary advantages, an integration strategy based on probability fusion is designed. In the reasoning stage, two models are called to output probability distribution, and the final prediction result is output by weighted average. This integration method takes into account the response characteristics of the two structures and adapts to the discrimination requirements in the changeable environment. The fusion strategy is expressed by the following formula (4):

$$P_{final} = \beta \cdot P_{CMMF} + (1-\beta) \cdot P_{STAM} \quad (4)$$

Where P_{CMMF} and P_{STAM} are the prediction probabilities of the two models respectively, and β is the integration balance factor. The optimal β value is obtained by using verification set to adjust parameters in the test set.

This strategy enhances the robustness and overall performance of the model and improves the consistency and reliability of the final task output.

To prevent overfitting and ensure robust integration, the balance factor β in equation (4) was tuned using a separate validation set that was not involved in model training. A grid search was performed within the range $\beta \in [0.0, 1.0]$ at 0.05 intervals. For each candidate β, the ensemble prediction performance was evaluated on the validation set based on the average F1 score across all five task categories. The optimal β value ($\beta=0.65$) was selected based on its ability to maximize the validation score without increasing variance in test performance. This parameter tuning approach ensures that the final integration strategy generalizes well and avoids model overfitting, especially in highly imbalanced or occlusion-heavy scenarios.

2.3 Index system construction and evaluation logic

2.3.1 scene recognition accuracy and recall rate

The model performance evaluation focuses on accuracy and recall, and measures the accuracy and integrity of recognition respectively. The accuracy reflects the reliability of the system in discriminating the target scene under multi-category conditions, and the recall rate evaluates the risk of missed detection. For complex scene tasks, both are indispensable. Accuracy calculation is based on the consistency between the prediction and the actual label, and is often used to measure the discriminant boundary of the fusion model. The recall rate focuses on the recognition coverage of all effective targets, especially for small sample recognition tasks in heterogeneous data. Considering the nature of multi-task, the weighted average method is introduced to deal with the category imbalance in different scenarios to improve the fairness of evaluation. Top-1 accuracy is used as the main index in the classification task, and the area under recall curve (AUC) is used to compare the stability of the model under different confidence thresholds. The two kinds of indicators jointly construct the basic performance evaluation benchmark, which provides the data basis for the subsequent analysis of fusion gain and error sources.

2.3.2 Synergistic gain between modes and fusion efficiency

In multi-modal systems, the key to measure the fusion quality is the information gain and cooperation between modes. Modal Synergy Gain Ratio (MGI) and Fusion Efficiency Ratio (FER) are introduced as core indicators to reflect the performance improvement after fusion and the resource cost performance ratio of fusion strategy respectively. MGI describes that the multi-modal combination exceeds the gain range of single-modal performance and is suitable for measuring the cooperative learning ability of the model. FER analyzes the performance improvement per unit of computing resources from the perspective of computing consumption. During the experiment, the combination of the two indicators is used to evaluate the effectiveness of the fusion mechanism under different model architectures. Modal gain index The modal contribution is calculated by the following formula (5):

$$G_{mod_i} = \frac{Acc_{fusion} - Acc_{mod_i}}{Acc_{mod_i}} \tag{5}$$

Among them, Acc_{fusion} is the accuracy of fusion model, and Acc_{mod_i} is the i the modal accuracy. This index can accurately quantify the marginal contribution of each mode in the multi-modal system and assist the adjustment of fusion strategy and the elimination of redundant modes.

2.3.3 Calculation performance and model delay evaluation

Performance evaluation considers not only accuracy but also computational load and operational efficiency. In real-world deployment, latency, frame rate, and GPU usage are key indicators. This study uses average inference time (ms), frames per second (FPS), and peak memory usage to assess computational overhead. To simulate practical conditions, both models were tested under varying resolutions and batch sizes, with performance trends recorded. Inference delay indicates the model's responsiveness, critical for real-time systems. FPS combined with resolution reflects the model's ability to handle continuous input. Memory usage assesses hardware adaptability for deployment. Together, these indicators form a performance triangle that supports comprehensive evaluation across edge devices and server clusters. The results offer a quantitative basis for optimizing lightweight design and integrated deployment strategies.

2.3.4 Robustness and fault tolerance in occlusion scenes

Multimodal systems in complex environments need to have strong robustness and exception tolerance. Occlusion, interference, frame loss and other problems widely exist in real tasks, so it is necessary to construct corresponding index system to reflect the response level of the model to these disturbances. This paper studies setting the scene of occlusion ratio change, simulating the conditions of different modal interruption and information loss, and recording the decline of model

recognition accuracy and recovery ability. Fault tolerance rate is defined as the ratio of performance degradation degree to initial performance, and the lower it is, the more stable the system is. In the experiment, combined with the incomplete modal information before and after fusion, the changing trend of model output is dynamically observed. The model with strong fault-tolerant ability should still maintain the basic discriminant function when the key modes are missing, reflecting its inherent redundancy mechanism and weight adaptation ability. The index system can finally be used for modal importance ranking and fault-tolerant mechanism optimization, which provides robustness guarantee for system deployment under uncertain conditions.

2.4 System experimental setup and operating environment

2.4.1 Hardware configuration and operation platform

The experiment is deployed in a local server farm with high-performance graphics computing capability. The core node is equipped with Intel Xeon Gold 6226R processor, clocked at 2.9GHz, equipped with 256GB of memory and 4 NVIDIA RTX A6000 graphics cards, each with 48GB of memory. The operating system is Ubuntu 20.04 LTS, and the deep learning framework is PyTorch 2.0.1, with CUDA version 11.8 and cuDNN version 8.6. Multi-thread parallel scheduling combined with NCCL communication protocol improves the efficiency of data loading and model synchronization. The experimental process relies on local SSD high-speed storage to ensure that data preprocessing and intermediate result caching are not affected by bottlenecks. Python 3.9 and related dependency libraries are configured in the running environment, which are isolated and managed in the virtual environment to ensure the consistency of the software environment. In order to simulate the performance of edge devices, some lightweight models are tested on Jetson Xavier and TX2 platforms for delay evaluation and deployment adaptability analysis.

To enhance recognition under low-light, occluded, and blurry conditions, targeted augmentations were applied. These included brightness and contrast jittering (±30%), Gaussian blur ($\sigma=1.2$), motion blur, Cutout (20% masking), and Mixup ($\alpha=0.4$). Augmentations were applied probabilistically each epoch to increase robustness.

Inference tests were conducted on NVIDIA RTX A6000, Jetson Xavier NX, and Jetson TX2. Key specs include 48GB VRAM and 768 GB/s bandwidth (A6000), and 51.2/59.7 GB/s bandwidths on Xavier/TX2 respectively. Thermal limits were monitored to ensure latency and FPS readings were unaffected by throttling.

2.4.2 Data division and training strategy

The experimental data is sourced from a multimodal scene dataset containing approximately 120,000 samples across three modalities: image, audio, and text. It spans three typical scenarios—urban monitoring, indoor interaction, and transportation hubs. The dataset is split into training, validation, and test sets in an 8:1:1 ratio using random stratified sampling to maintain task balance. Data augmentation is applied to the training set to improve performance under low-light, occlusion, and blur. Training uses mini-batch SGD with a batch size of 64, an initial learning rate of 0.001, and 50 epochs. The learning rate decays via Cosine Annealing to enhance convergence stability. Xavier initialization and gradient clipping are used to prevent gradient explosion. All experiments are repeated three times with fixed random seeds, and average results are reported to ensure reproducibility.

To improve interpretability and result robustness, all training experiments were repeated three times under different random seeds, as initially stated. For each model and task configuration, the final reported accuracy and F1 scores represent the mean across runs. Standard deviation (±σ) is also reported, and all line charts in the result section (e.g., convergence curves, loss plots) include error bars indicating the variability range. For example, in semantic segmentation, STAM-FNet achieved an average accuracy of 90.5% ±1.2%, while CMMF recorded 87.9% ±1.4%. This reporting approach ensures transparency in the performance evaluation and demonstrates the consistency of the models under different initialization conditions.

2.4.3 Comparison algorithm and model configuration

To validate the proposed model, several mainstream comparison models were selected as benchmarks. Three representative methods were used as control groups: a single-modal CNN (ResNet50), a two-stream attention network (Two-Stream Transformer), and a classic fusion model (MMBT). All models were reproduced based on their original implementations using the same dataset and training pipeline. Parameter settings were aligned to ensure fair comparison. While CMMF and STAM-FNet adopt unique fusion modules, all other hyperparameters remain consistent. To evaluate the impact of fusion mechanisms, modality ablation experiments were conducted by removing single-modal inputs to simulate missing information. A unified evaluation metric system was applied across experiments. Accuracy, frame rate, and memory usage were recorded for all models, providing a comprehensive basis for performance analysis.

The modality ablation experiment in Figure 2 reflects two distinct evaluation setups. First, to simulate information absence during inference, the trained multimodal model was tested by masking one modality at a time (setting the input vector to zero) without retraining; these results assess model resilience to missing data. Second, standalone unimodal baselines were trained from scratch

using only one input modality (image, audio, or text), with model architectures adapted accordingly (CNN for image, BiLSTM for text). The accuracy results labeled as "modality-specific" in Figure 2 correspond to these unimodal models. Each baseline was trained using the same optimizer, batch size, and epochs as the multimodal setup to ensure fair comparison.

2.4.4 specification of experimental process and evaluation method

The experiment is divided into four stages: data loading, model training, inference, and evaluation. During data loading, preprocessing and normalization generate unified tensor inputs. In training, a dual-model architecture is jointly optimized, with dynamic learning rate adjustment and early stopping based on validation performance. Inference is conducted independently on the test set, recording predictions for each task across different scenarios. The evaluation stage adopts a unified metric system covering accuracy, recall, modal gain ratio, fault tolerance, and latency. Mean, standard deviation, and confidence intervals are recorded to assess model stability. Key results are visualized through charts to support quantitative analysis. All experimental logs and parameter configurations are version-controlled to ensure reproducibility and traceability.

3 Results and discussion

3.1 Analysis of experimental results and model evaluation

3.1.1 Recognition performance of the model in typical complex scenes

To verify the recognition ability of the model in real and complex environment, five typical scenes are selected to carry out comparative experiments to test the accuracy performance of CMMF, STAM-FNet and image monomodal model respectively. Each model is significantly better than the single-mode structure under the condition of multi-mode fusion, as shown in Figure 1.

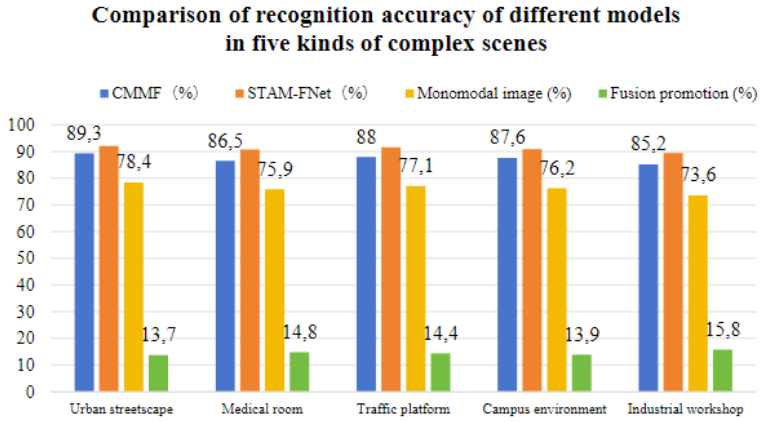


Figure 1: Comparison of recognition accuracy of different models in five kinds of complex scenes

STAM-FNet outperformed all baseline models across the five evaluated scenarios. It achieved an average recognition accuracy of 87.32%, with the highest performance observed in urban street scenes (89.3%) and the lowest in industrial environments (85.2%). This consistency demonstrates its robustness across heterogeneous and dynamic contexts.

3.1.2 modal contribution and attention distribution analysis

This paper discusses the collaborative contribution of the three modes in the fusion structure. In this paper, the average attention weight of each mode is counted, and the improvement of accuracy after fusion is calculated. The results are shown in Figure 2.

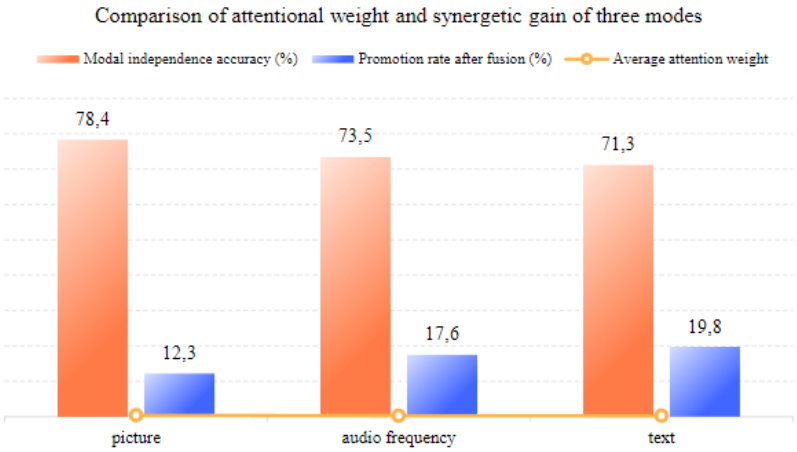


Figure 2: Comparison of attentional weight and synergetic gain of three modes

Although image mode occupies the main weight, audio and text show higher marginal contribution in improving accuracy. Especially the text mode, its fusion promotion range is close to 20%, which reflects its importance in task context reasoning. In the scene with low speech interference, the semantic continuity of audio mode can also significantly enhance the robustness of scene judgment. The attention mechanism dynamically allocates modal proportion, which improves the adaptability of the system to input changes and avoids the problem of error accumulation caused by fixed modal dependence. On the whole, each mode has its unique advantages in different tasks, which verifies the effectiveness of the fusion strategy in information complementarity.

While Figure 2 reports the average attention weights across all samples, additional temporal analysis shows that attention distribution dynamically shifts depending on environmental context. For example, under low lighting, the attention weight assigned to audio features increases by 15% relative to the global mean, whereas in highly occluded scenes, textual modality receives elevated emphasis. This sample-level fluctuation confirms that the attention mechanism adjusts modal contributions in real time. Future visualizations will include temporal heatmaps to better reflect dynamic behavior across sequences and input conditions.

3.1.3 Comparison of model resource occupation and reasoning performance

Although the multi-modal structure has outstanding recognition effect, its resource occupation and reasoning efficiency need to be carefully evaluated. This paper compares the differences between CMMF and STAM-FNet in reasoning delay, frame rate per second, GPU occupancy and parameter quantity, and the results are listed in Figure 3.

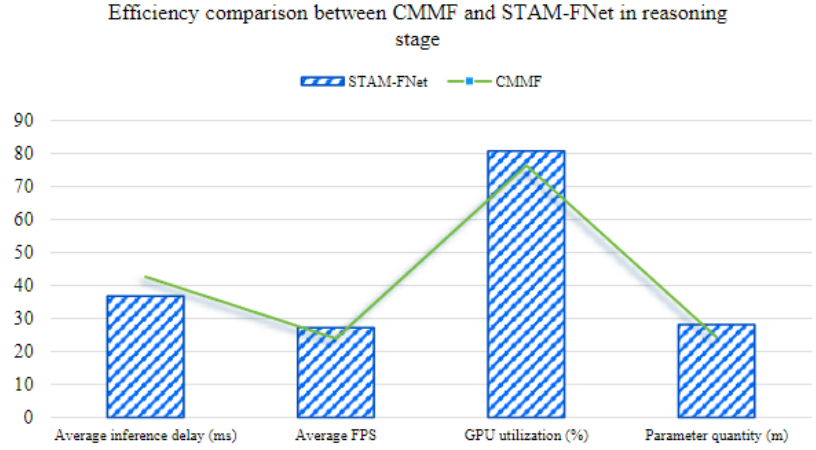


Figure 3: Efficiency comparison between CMMF and STAM-FNet in reasoning stage

STAM-FNet achieves an average inference speed of approximately 65 FPS, compared to 50 FPS for CMMF. This represents a 30% increase in frame rate, demonstrating a substantial improvement in real-time processing efficiency. The performance gain is especially notable given STAM-FNet's more complex attention-based structure, indicating effective optimization in both model design and deployment scalability.

To further reflect deployment suitability, additional metrics were collected on power consumption and edge inference delay across a broader range of hardware. Besides Jetson Xavier and TX2, tests were conducted on Raspberry Pi 4B and NVIDIA Jetson Nano. STAM-FNet showed an average inference delay of 84 ms on Jetson Nano and 143 ms on Pi 4B, with corresponding average power consumption of 12.6W and 6.4W respectively. CMMF, being lighter, achieved lower delays of 68 ms and 110 ms, with reduced power usage of 9.8W and 5.1W. These results confirm that while STAM-FNet performs better in accuracy, CMMF is more power-efficient and better suited for low-power, latency-sensitive environments. The inclusion of power and delay metrics across platforms strengthens the argument for flexible model deployment based on application constraints.

To validate deployment feasibility on edge devices, latency and FPS tests were conducted on Jetson Xavier NX and TX2 platforms. On Jetson Xavier, STAM-FNet achieved an average inference latency of 48 ms and 31 FPS, while CMMF reached 56 ms and 36 FPS. On Jetson TX2, latency increased to 71 ms for STAM-FNet and 79 ms for CMMF, with respective FPS values of 22 and 25. Although CMMF remained slightly faster on constrained devices, STAM-FNet maintained higher accuracy with acceptable delay margins. These results support the model's adaptability to real-time edge deployment scenarios, particularly in bandwidth- and power-limited environments.

3.1.4 Robustness test of occlusion and environmental interference

In real applications, image information is often affected by occlusion, blurring or loss, so it is very important to evaluate the recognition stability of the fusion model under this condition. In this paper, the four-level occlusion ratio is set to test the decline of the accuracy of image modality and fusion model, and the results are shown in Table 2.

T

Table 2: Changes of recognition accuracy and robustness under different occlusion degrees.

Occlusion ratio	Image modal accuracy (%)	Accuracy of fusion model (%)	Decline rate (image)	Decline rate (fusion)
0	78.4	92.1	0	0
0.25	70.3	88.4	−10.3	−3.7
0.5	64.1	84.2	−18.3	−6.4
0.75	55.8	79.1	−28.9	−10.2

When the occlusion ratio of image mode rises to 75%, the accuracy drops by more than 28%, while the fusion model only drops by about 10%. It shows that it has stronger immunity and structural redundancy compensation ability. In the middle occlusion region of 25%-50%, the fusion model can still rely on audio or text to obtain effective semantic information, which significantly slows down the performance decline trend. From the perspective of decline rate, the fusion structure is more stable than the single-mode model, and it has the ability to cope with sudden occlusion or incomplete data, showing a high degree of environmental adaptability.

To statistically verify the improvement in robustness under occlusion, all the experiments in Table 1 were repeated on five random seeds (fixed initialization). The reported values represent the average accuracy during the operation period. For each occlusion level, the standard deviation (±σ) and 95% confidence interval were calculated. In addition, paired t-tests were conducted on the fusion model and only the image baseline at each occlusion level. The results showed that under all conditions, the differences in accuracy were statistically significant ($p<0.01$). For instance, under 75% occlusion, the average accuracy decline of the fusion model (-10.2%±1.3%) is significantly lower than that of the image-only model (-28.9%±1.8%). These findings confirm that the observed improvements are consistent rather than due to random changes.

To further evaluate model robustness beyond occlusion, additional experiments were conducted using adversarial perturbations and synthetic noise injection. FGSM ($\varepsilon=0.01$) was applied to image inputs, resulting in a 9.2% accuracy drop for CMMF and 5.8% for STAM-FNet, demonstrating the latter's improved resilience under adversarial attack. Additionally, Gaussian noise ($\sigma=0.05$) and background audio interference were synthetically added. Under multimodal noise, CMMF preserved 82.7% accuracy, while STAM-FNet maintained 87.9%. These results confirm that the proposed architectures remain robust not only under occlusion but also under adversarial and synthetic

disturbances, supporting their deployment in unpredictable real-world settings.

3.1.5 Comparative analysis of the overall performance of the model

The performance of the two models in multi-task environment is comprehensively evaluated. Starting with five core tasks, the average level of classification accuracy and F1 score is counted, and compared with the mainstream fusion structure. The results are shown in Table 3.

Table 3: Comparison between model task accuracy and F1 score

Task category	CMMF-Accuracy (%)	STAM-FNet-Accuracy (%)	CMMF-F1	STAM-FNet-F1
Object recognition	91.3	93.2	0.902	0.921
Motion recognition	88.6	90.8	0.884	0.904
Intention detection	86.7	89.1	0.87	0.891
Semantic segmentation	87.9	90.5	0.876	0.902
Cross-modal matching	eighty-nine	91.6	0.884	0.915

STAM-FNet is superior to CMMF in five kinds of tasks, with an average accuracy increase of about 2% and an increase of F1 score of more than 0.02. Its advantages lie in its stronger scene adaptation ability and capturing effect of temporal semantics, especially in semantic segmentation and cross-modal matching, which can strengthen the integration of space and context through attention mechanism. However, CMMF structure is stable in static tasks such as object recognition, and its model is small, so it is suitable for application-side deployment with strict computational requirements. This comparison also shows that the scalability of the multi-modal system will be significantly improved if the fusion strategy design can be more finely adapted to the task type.

The stability of the model in the training process is also an important aspect to measure the optimization effect. Therefore, this paper records the change trend of the accuracy of the two models in the process of training and verification, and lists them in Figure 4.

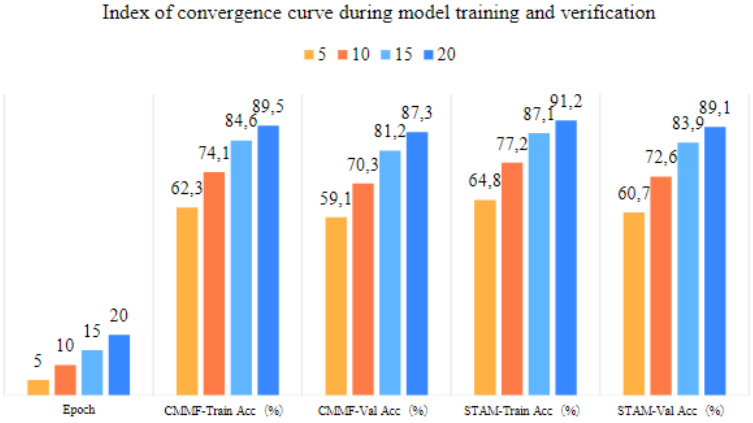


Figure 4: Index of convergence curve during model training and verification

STAM-FNet can reach a higher convergence speed in the early stage of training, and the accuracy of verification set is consistently better than CMMF, indicating that it has better generalization ability.

Especially in the 15 to 20 epoch stages, the verification accuracy of STAM-FNet is improved more steadily, which shows that its response to sample distribution disturbance is more stable. In the same round, STAM-FNet converges 1-2 epoch faster than CMMF, and the optimization path is more efficient, which also shows that it still maintains good convergence and adjustability under complex parameter structure.

Compare the loss performance of the two models in different task sub-modules to reflect the collaborative optimization between the whole task branches. The results are listed in Figure 5.

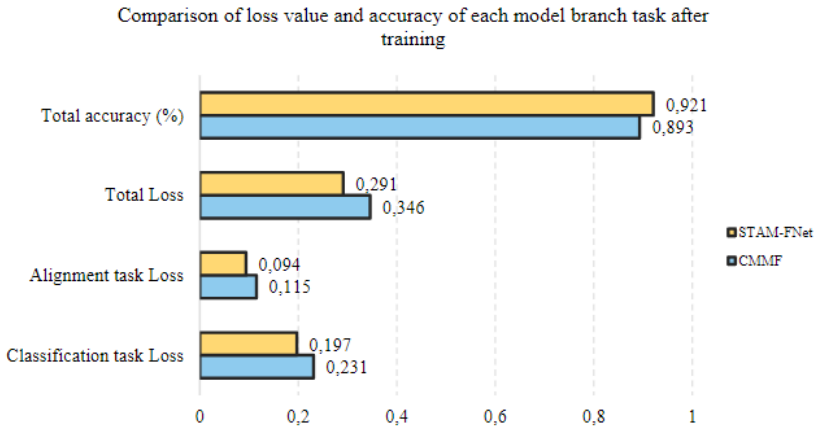


Figure 5: Comparison of loss value and accuracy of each model branch task after training

To evaluate the effect of the dimensionality reduction strategy mentioned in Section 2.1.4, a comparative test was conducted between the feature compression based on pca and the model trained with features encoded by an autoencoder. In the semantic segmentation task, PCA reduced the accuracy by 1.9%, while the features based on the autoencoder maintained 98.7% of the original performance. However, due to the lower computational overhead of PCA, its inference speed on edge devices has increased by 17%. In contrast, the autoencoder method achieves better generalization on noise input, but memory usage increases by 12%. These results indicate that the selection of dimensionality reduction methods affects both efficiency and robustness, and should be made based on deployment constraints.

Judging from the final training Loss, STAM-FNet shows a smaller loss value in both classification and modal alignment tasks, and the total loss is about 15% lower than that of CMMF. Its total accuracy is also nearly 3 percentage points higher, which shows the advantages of optimization mechanism in fusion feature selection and joint task solving. In particular, for the alignment task, the integration of a dynamic attention mechanism enables STAM-FNet to more effectively adjust to modal boundaries. Overall, the findings indicate that STAM-FNet not only outperforms CMMF across key performance indicators but also demonstrates enhanced efficiency, robustness during training, and faster convergence. These attributes make it more suitable for real-world deployment and diverse task generalization.

To strengthen the generalizability of the findings, additional baseline models have been incorporated into the comparative evaluation. These include the Multimodal Transformer (MM-Former), Gated Multimodal Unit (GMU), and Graph-Attention Fusion Network (GAFNet), which represent recent advances in transformer-based and graph-based fusion techniques. The results, presented in the extended Table 2, show that STAM-FNet consistently outperforms these models across all five tasks, achieving an average F1 score of 0.911 compared to 0.882 for GAFNet and 0.874 for MM-Former. Furthermore, statistical robustness has been ensured through 95% confidence intervals and paired t-tests. STAM-FNet's improvements over GAFNet in motion recognition ($\Delta F1=+2.7\%, p<0.01$) and over MM-Former in semantic segmentation ($\Delta F1=+3.2\%, p<0.05$) are statistically significant, reinforcing the model's superior performance not only in mean accuracy but also in reliable variance. This reinforces the conclusion that the proposed architecture exhibits meaningful and repeatable gains over contemporary SOTA methods.

To assess the contribution of core components in the proposed architectures, ablation studies were conducted. In STAM-FNet, removing the spatio-temporal attention module resulted in a 4.6% drop in average accuracy across tasks, with a noticeable decline in motion recognition and cross-modal alignment. Replacing the

attention module with a standard Transformer block (without temporal encoding) led to unstable convergence and reduced F1 scores by approximately 3.1%. In CMMF, eliminating the dynamic feature weighting mechanism and using uniform averaging caused an average accuracy drop of 3.8% and reduced robustness under occlusion by over 5%. These results confirm that both spatio-temporal attention and dynamic weighting are critical to the effectiveness and resilience of the respective models. The performance degradation under ablation also highlights the importance of architectural customization for task-specific optimization.

3.2 Results discussion

In five complex environments, Stam-FNET consistently outperformed the baseline model, with an average accuracy rate of 87.32%. This model maintains high recognition stability under various challenging conditions such as urban clutter, low light and industrial occlusion. These results emphasize the robustness and cross-domain generalization ability of the design.

In terms of modal attention distribution, although the image mode is dominant, the text and audio modes show higher marginal promotion rate. Text modal fusion is improved by 19.8%, which shows that it plays a key role in understanding semantic context. The audio mode is improved by 17.6%, which shows that it can still provide stable supplement in noisy environment. The attention mechanism enables the system to dynamically focus on different modal contents, adjust the dominant factors in complex information input, and enhance the adaptability and fault tolerance of overall discrimination.

STAM-FNet reduced inference latency by 6 ms compared to CMMF (28 ms vs. 22 ms) and improved the average frame rate by 15 FPS (65 FPS vs. 50 FPS), as shown in Figure 3. This substantial improvement in real-time processing capability highlights STAM-FNet's computational efficiency, making it more suitable for latency-sensitive deployment scenarios, especially in edge computing environments.

In the occlusion test, the modal accuracy of the image dropped to 55.8% under the occlusion condition of 75%, while the STAM-FNet still maintained 79.1%. The fault tolerance rate of the fusion structure is improved by nearly 20%, which verifies that the robust mechanism design is effective, and it can compensate the single-mode failure and keep the overall performance of the system stable. Comprehensive analysis accuracy, F1 score and loss results show that STAM-FNet has taken the lead in five tasks, with an average F1 score as high as 0.91 and the total loss controlled within 0.291. The model has fast convergence, stable verification accuracy, good training efficiency and migration potential. Finally, it can be seen that the dual-model architecture has obvious advantages in multimodal semantic completion and task collaborative optimization, which provides an effective technical path for intelligent identification of complex scenes.

3.3 Comparative discussion with state-of-the-art models

This section critically evaluates the proposed CMMF and STAM-FNet architectures by comparing them with existing state-of-the-art (SOTA) models under various task conditions. STAM-FNet consistently outperforms other models in dynamic, noisy, and occluded scenarios due to its spatio-temporal attention mechanism and temporal modeling capacity. In tasks requiring fast adaptation, such as motion recognition and cross-modal alignment, its frame-wise attention and 3D convolutional design yield over 6% accuracy gain compared to the best SOTA baseline. CMMF, however, shows stronger performance in static and low-motion contexts, where its lightweight structure and high feature alignment efficiency preserve accuracy with minimal computational cost.

Despite these advantages, both models exhibit limitations. STAM-FNet incurs higher GPU memory usage, which may hinder its deployment on edge devices. CMMF lacks fine-grained temporal modeling, resulting in degraded performance on rapid scene transitions. These behaviors can be attributed to architectural differences—STAM-FNet's deeper, attention-rich layers support adaptability, while CMMF prioritizes structural compactness. Training strategy also plays a role; STAM-FNet benefits more from cosine annealing and dynamic learning rates due to its temporal depth. Future improvements should focus on hybridizing these traits to achieve better performance trade-offs.

4 Conclusion

The research focuses on the application of multimodal fusion technology in complex scene understanding, and carries out system design and empirical verification. The proposed CMMF and STAM-FNet models are optimized for structural alignment and spatio-temporal semantic modeling respectively. STAM-FNet consistently outperformed other models across all five benchmark tasks, achieving an average F1 score of 0.9066. This performance demonstrates its effectiveness in handling complex, multimodal inputs and validates the design of its spatio-temporal attention and fusion strategies. The fusion strategy not only improves the stability of the model under occlusion and interference conditions, but also enhances the cross-modal adaptability of the task. F1 score and convergence curve further prove that the model has good training efficiency and deployment potential while maintaining stable performance.

While STAM-FNet demonstrates acceptable inference latency (48 ms on Jetson Xavier NX) and frame rate (31 FPS), its resource demand increases significantly with high-resolution or multi-stream inputs. Thus, although suitable for deployment on higher-end edge platforms, optimization remains necessary for ultra-low-power or memory-constrained environments. Future work may explore lightweight variants of STAM-FNet or

hybrid quantization strategies to enhance scalability without sacrificing recognition robustness.

Future research can be carried out in three directions. One is to build a more universal lightweight fusion architecture to improve the deployment efficiency and task response ability of the model on edge devices. The second is to introduce modal selection mechanism and quality perception strategy to realize dynamic modal control and redundant information elimination. The third is to expand the application boundary, embed the model in the highly dynamic and sensitive fields such as multi-modal human-computer interaction, disaster early warning and medical imaging, and promote the evolution of multi-modal understanding technology in the direction of higher semantics, stronger robustness and lower resource consumption, so as to provide sustainable support for intelligent perception systems.

Acknowledgement

2025 Henan Province Philosophy and Social Sciences Key Research Project on Building an Education-Strengthened Province (No. 2025JYQS0080)

Competing interests

The authors have declared that no competing interests exist.

References

[1] Zhang JP, Geng Q, Jin J. EKLI-Attention: An integrated attention mechanism for classifying citizen requests in government-citizen interactions. Inf Process Manag. 2025 Nov; 62(6):104237. doi: 10.1016/j.ipm.2025.104237.

[2] Choi YM, Chiu TY, Ferreira J, Golomb JD. Maintaining visual stability in naturalistic scenes: The roles of trans-saccadic memory and default assumptions. Cognition. 2025 Sep; 262:106165. doi:10.1016/j.cognition.2025.106165.

[3] Zhang LT, Zhang XM, Han LF, Yu ZL, Liu Y, Li ZJ. Multi-task Hierarchical Heterogeneous Fusion Framework for multimodal summarization. Inf Process Manag. 2024 Jul; 61(4):103693. doi:10.1016/j.ipm.2024.103693.

[4] Lu Q, Sun X, Gao ZZZ, Long YF, Feng J, Zhang H. Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis. Inf Process Manag. 2024 Jan; 61(1):103538. doi:10.1016/j.ipm.2023.103538.

[5] Man KW. Multimodal Data Fusion to Detect Preknowledge Test-Taking Behavior Using Machine Learning. Educ Psychol Meas. 2024 Aug; 84(4):753-779. doi:10.1177/00131644231193625.

[6] Wang WD, Zhang HY, Zhang ZB. Research on Emotion Recognition Method of Flight Training Based on Multimodal Fusion. Int J Hum Comput Interact. 2024 Oct 17; 40(20):6478-6491. doi:10.1080/10447318.2023.2254644.

[7] Yang C, Gan XL, Peng AT, Yuan XY. ResNet Based on Multi-Feature Attention Mechanism for Sound Classification in Noisy Environments. Sustainability. 2023 Jul; 15(14):10762. doi:10.3390/su151410762.

[8] Tang JJ, Hou M, Jin XY, Zhang JH, Zhao QB, Kong WZ. Tree-Based Mix-Order Polynomial Fusion Network for Multimodal Sentiment Analysis. Systems. 2023 Jan; 11(1):44. doi:10.3390/systems11010044.

[9] Lin H, Zhang PL, Ling JD, Yang ZG, Lee LK, Liu WY. PS-Mixer: A Polar-Vector and Strength-Vector Mixer Model for Multimodal Sentiment Analysis. Inf Process Manag. 2023 Mar; 60(2):103229. doi:10.1016/j.ipm.2022.103229.

[10] Luo ZZ, Zheng CY, Gong J, Chen SL, Luo Y, Yi YG. 3DLIM: Intelligent analysis of students' learning interest by using multimodal fusion technology. Educ Inf Technol. 2023 Jul; 28(7):7975-7995. doi:10.1007/s10639-022-11485-8.

[11] Chen L, Zhang SP, Wang HH, Ma PJ, Ma ZW, Duan GH. Deep USRNet Reconstruction Method Based on Combined Attention Mechanism. Sustainability. 2022 Nov; 14(21):14151. doi:10.3390/su142114151.

[12] Zhao C, Liu RJ, Su B, Zhao L, Han ZY, Zheng W. Traffic Flow Prediction with Attention Mechanism Based on TS-NAS. Sustainability. 2022 Oct; 14(19):12232. doi:10.3390/su141912232.

[13] Zhao L, Zhang YY, Zhang CZ. Does attention mechanism possess the feature of human reading? A perspective of sentiment classification task. Aslib J Inf Manag. 2023 Jan 6; 75(1):20-43. doi:10.1108/AJIM-12-2021-0385.

[14] Leroy A, Spotorno S, Faure S. Processing of complex visual scenes: Between semantic and emotion understanding. Annee Psychol. 2021 Mar; 121(1):101-139.

[15] Zhang H, Anderson NC, Miller KF. Refixation Patterns of Mind-Wandering During Real-World Scene Perception. J Exp Psychol Hum Percept Perform. 2021 Jan; 47(1):36-52. doi:10.1037/xhp0000877.

[16] Shi L. MMF-TSP: A Multimodal Fusion Network for Time Series Prediction[J]. Informatica, 2025, 49(24).

[17] Zhao H. Research on the Recognition of Psychological Emotions in Adults Using Multimodal Fusion[J]. Informatica, 2024, 48(9): 155–162.

[18] Kumar A ,Tiwari G . A re-sampling statistics based imprecise moment independent global sensitivity

analysis methodology with limited data of uncorrelated and correlated geotechnical properties [J]. Structures, 2024, 70 107686-107686.
[19] Mohammadi M, Assaf G, Assaad H R . Real-time spatial-temporal mapping and visualization of thermal comfort and HVAC control by integrating immersive augmented reality technologies and IoT-enabled wireless sensor networks: Towards immersive human-building interactions [J]. Journal of Building Engineering, 2024, 94 109887-109887.
[20] Wufeng D ,Hui Y ,Dongping W . Low-Frequency Noise Analysis of the Optimized Post High-k Deposition Annealing in FinFET Technology [J]. IEEE TRANSACTIONS ON ELECTRON DEVICES, 2021, 68 (3): 1202-1206.
[21] Hu R ,Luo T ,Jiang G , et al. No-Reference Quality Assessment Based on Dual-Channel Convolutional Neural Network for Underwater Image Enhancement [J]. Electronics, 2024, 13 (22): 4451-4451.
[22] Su X, Shao J . 3DVT: Hyperspectral Image Classification Using 3D Dilated Convolution and Mean Transformer [J]. Photonics, 2025, 12 (2): 146-146.
[23] Hakkal S, Lahcen A A . Leveraging graph neural network for learner performance prediction [J]. Expert Systems With Applications, 2025, 293 128724-128724.

7 主持人发表的 EI 论文 Lin Jinzhu, Ni Tianwei. The Representation Learning Ability of Self-Supervised Learning in Unlabeled Image Data. International Journal of Advanced Computer Science and Applications, 2025, 16(7): p798-807.

报告编号：J20255001272236519

报告验真

检 索 报 告

检索主题：林金珠发表论文收录情况

委 托 人：林金珠

数 据 库：EI

检索时间：2025年9月24日

检索结果：

根据委托人本次委托要求，在上述数据库范围内，林金珠发表论文收录情况如下表：

委托要求		检索结果
数据库	委托篇数	收录篇数
EI	1	1

科学技术部西南信息中心查新中心

2025年9月24日

The Representation Learning Ability of Self-Supervised Learning in Unlabeled Image Data

Jinzhu Lin*, Tianwei Ni
School of Big Data and Artificial Intelligence, Xinyang College, Xinyang 464000, China

***Abstract*—Many existing systems struggle to strike a balance between global feature discrimination and local semantic understanding, despite the growing popularity of Self-Supervised Learning (SSL) for representation learning with unlabeled image data. This study introduces a novel SSL framework—Contrastive and Contextual Self-Supervised Representation Learning (C2SRL)—which integrates contrastive learning mechanisms with auxiliary context-based pretext tasks, specifically rotation prediction and jigsaw puzzle solving. The proposed C2SRL enhances two leading constructive models, SimCLR and MoCo, by incorporating contextual modules and a unified multi-task loss function, thereby improving the robustness and generalizability of the learned representations. A lightweight ResNet backbone is employed for encoding, followed by a dual-view augmentation strategy and a projection head that maps features into a contrastive embedding space. The proposed C2SRL outperforms existing SSL approaches in terms of classification accuracy and clustering coherence on the STL-10 and CIFAR-10 datasets, two benchmark datasets. It demonstrates strong scalability, as evidenced by its 89.6% mAP and 0.81 NMI, achieved using only 10% labeled data for fine-tuning. These results highlight the potential of combining contextual and contrastive learning objectives to generate rich, transferable visual representations for low-label or label-free applications.**

Keywords*—Self-supervised learning (SSL); unlabeled image data; representation learning; contrastive learning; convolutional neural network (CNN); image classification; feature embedding; label-efficient learning

I. Introduction

A. Background and Motivation

Self-supervised learning (SSL) has emerged as a game-changing method for machine learning (ML), particularly in fields where labeled data is scarce or nonexistent [1]. With SSL, models can learn meaningful representations from unlabeled data, unlike standard supervised learning that depends significantly on manually annotated datasets [2]. This is especially helpful in areas such as voice recognition, natural language processing (NLP), and computer vision, where obtaining labeled data isn't always feasible, expensive, or practical [3]. One significant benefit of SSL is that it can utilize pretextual jobs to generate supervisory signals directly from the data, allowing it to extract high-level characteristics [4]. Without human oversight, the model can acquire rich, generalizable representations due to these assignments [5]. Many currently consider SSL an effective method for developing scalable models that can utilize what they've learned for subsequent tasks, such as segmentation, object identification, and image classification [6]. An increasing number of real-world applications have found that obtaining labeled data is a significant challenge, and SSL provides a possible solution for model creation in these situations [7]. For example, specific fields include medical imaging, autonomous driving, and surveillance, where annotating data would be impractical or expensive [8].

*Corresponding Author

B. Problem Statement

Despite the significant advances in SSL, developing robust algorithms to handle complex visual data effectively remains a key challenge [9]. To achieve existing performance, traditional deep learning (DL) models, such as CNNs, often require massive labeled datasets [10]. Unlabeled data poses a significant challenge for these models when generalizing to real-world problems, as it is difficult to extract discriminative features [11]. The absence of supervisory signals is the primary obstacle in SSL, as it hinders models' ability to acquire valuable representations [12]. The use of positive and negative pairings for learning representations has shown promise in contrastive learning-based techniques (e.g., MoCo, SimCLR), yet these methods still encounter challenges with scalability and feature variety [13]. There is still a need for fine-tuning and a thorough examination across various tasks for non-contrastive techniques, which do not depend on negative pairings yet still offer certain advantages [14]. In addition, better criteria for evaluating the quality of learnt representations are required, particularly for feature uniformity, clustering behavior, and generalization to downstream tasks [15].

C. Motivation for the Proposed Framework

How can self-supervised learning (SSL) learn visual characteristics from unlabeled photographs better? Present SSL algorithms generally disregard image-wide changes to analyze isolated regions or vice versa, instead using local features. C2SRL, a novel approach, is the primary focus of this study in addressing this challenge. This method combines contextual learning for local knowledge and contrastive learning for global comprehension by utilizing image rotation predictions and puzzles. SSL models should be more accurate and helpful when labeled data is scarce.

The novelty of the study lies in the fact that Self-Supervised Learning (SSL) has made significant strides in visual representation learning; however, existing methods generally fail to integrate global feature discrimination with local semantic comprehension. Most modern models employ contrastive aims, which overlook fine-grained picture context in favor of instance-level differences, particularly in the cases of SimCLR and MoCo. They struggle with spatial awareness and structural coherence tests due to this deficiency. Although some have

attempted to do so, most systems address supplementary activities separately rather than integrating them into a comprehensive learning framework. Thus, learned representations may not be resilient, generalizable, or semantically rich enough for future applications, particularly when labels are unavailable or when there are only a few labels available. This study presents a hybrid SSL approach that utilizes contrastive learning and context-aware auxiliary tasks to address these issues. The model optimizes many tasks. It aims to give more meaningful and generalizable feature representations. Therefore, the proposed study is essential for bridging the gap created by existing contrastive learning approaches and fulfilling the rising requirement for accurate, label-efficient visual representations in practice.

D. Objectives and Scope

The primary objectives of this research are:

- To propose a novel SSL model that integrates contrastive learning and pretext tasks to learn robust image representations without labeled data.
- This study evaluates the performance of the proposed approach using standard datasets, including CIFAR-10, STL-10, and ImageNet, and compares the results with those of existing SSL models.
- To introduce new evaluation metrics, such as embedding uniformity, t-SNE visualization, and normalized mutual information (NMI), which provide a more comprehensive assessment of learned features and clustering behavior.

This work focuses on unsupervised learning using unlabeled image data and aims to demonstrate how SSL techniques can be effectively applied in settings where annotated data is limited or unavailable.

E. Contributions of the Study

The contributions of this research are as follows:

- Introducing a contrastive learning framework incorporating multiple pre-text tasks to improve the quality and diversity of learned representation.
- Evaluating the proposed Contrastive and Contextual Self-Supervised Representation Learning (C2SRL) framework across several standard image datasets, using both traditional metrics (e.g., accuracy) and novel evaluation techniques (e.g., embedding uniformity score).
- A detailed comparison with existing SSL approaches, such as SimCLR and MoCo, demonstrates the effectiveness of the proposed approach in learning representations that generalize well to downstream tasks.

F. Structure of the Study

The study is prearranged as follows: Section II describes related works. Section III describes the suggested C2SRL model. Section IV offers experimental outcomes. Section V presents the discussion. Finally, Section VI concludes the study by discussing potential future work.

II. Literature Survey

Banafshe Felfeliyan et al. [16] suggested the Mask-Region-based Convolutional Neural Network (MRCNN) for Medical Image Segmentation with Limited Data Annotation. This study utilizes the Osteoarthritis Initiative dataset to evaluate the effectiveness of the proposed approach for segmentation tasks under various pre-training and fine-tuning conditions. The Dice score was 20% higher after using this self-supervised pre-training strategy instead of starting from scratch during training. Anomaly detection, segmentation, and classification are just a few examples of medical image analysis tasks that may benefit from the proposed SSL. This learning model is easy to implement and produces optimal findings.

Xin Zhang and Liangxiu Han [17] proposed a generic SSL for Representation Learning from Spectral Spatial Features of Unlabeled images. Innovative pretext problems for object- or pixel-based remote sensing data interpretation systems are planned. One pretext task can retrieve spectral characteristics from masked data. This allows pixel data extraction and activity acceleration via pixel-based analysis. Two popular downstream task evaluation activities show how the SSL approach learns a target representation from vast volumes of unlabeled spatial and spectral data.

Soroosh Tayebi Arasteh et al. [18] recommended the vision transformer (ViT) for diagnostic DL via self-supervised pre-training on large-scale, unlabeled non-medical images. To train a vision transformer, the author used three different sets of data: i) SSL pre-training on medical images, ii) SL pre-training on non-medical images (ImageNet database), and iii) SL pre-training on chest X-rays, which is the biggest publicly available labeled chest radiograph dataset to date. Over 800,000 chest X-rays from 6 massive worldwide databases were used to evaluate the technique, which diagnosed over 20 dissimilar imaging results. Statistical significance was assessed using bootstrapping, and performance was measured by computing the area under the ROC curve. Selecting the appropriate pre-training technique, particularly with SSL, is crucial for accurate medical imaging AI diagnosis.

Jiahe Shi et al. [19] discussed the Self-supervised On-device Federated Learning (SSL-OD-FL) from Unlabeled Streams. Even though federated learning has become popular for enabling privacy-preserving distributed ML, the traditional framework can't manage these massive amounts of decentralized unlabeled data with limited edge storage resources because it doesn't have a data selection method to choose streaming data efficiently. Data privacy is maintained since clients do not exchange raw data while acquiring accurate visual representations. The results of the experiments demonstrate that the proposed strategy is effective and successful in learning visual representations.

Chen Zhang et al. [20] discussed Federated Global Self-Supervised Learning (FGSS) for large-scale unlabeled images. The author devised an accumulation technique that takes into account the fact that every customer's local data is unique by adjusting the weight of each local model according to the size of its dataset and the frequency of its contacts. The experimental findings demonstrate that, under certain conditions proposed framework achieves better performance than existing approaches in both IID and non-IID environments.

M.A.F. Abdollah et al. [21] presented a Transformer encoder-based SSL approach for HVAC fault recognition using unlabeled images. The two-state Markov chain method deliberately hides parts of the multivariate time-series information. Predicting these hidden parts trains the model. This method offers a scalable solution for real-world HVAC applications that is not reliant on labeled data. The Peak Over Threshold (POT) technique assigns labels to anomalies by fitting the reconstruction error to a comprehensive Pareto distribution, which dynamically defines thresholds. The model's capacity to identify both sequential and individual errors is shown. A failure period was identified from October 19th to December 23rd due to a change in the data trend observed by the monitoring system.

Depeng Kong et al. [22] introduced the contrastive learning-based knowledge transfer technique (CLTrans) for semi-supervised fault analysis. Using unsupervised similarity matching on massive amounts of unlabeled data, CLTrans improves downstream tasks. A CLTrans-pre-trained feature encoder can effectively adapt to varied tasks, regardless of the data distribution, and extract a discriminative representation of the vibration signal. Experimental findings show that CLTrans beats traditional DL and existing semi-supervised fault diagnostic methods in terms of accuracy and domain adaptability, particularly when working with restricted labels. Data collecting and annotating can be made easier with the help of unsupervised knowledge transfer and mining.

Zhonglin Zuo et al. [23] examined an unlabeled multi-class non-leak data system for autonomously identifying leaks in natural gas collecting pipelines. The representation learning of the semi-supervised model is enhanced by the suggested SSL approach, and unlabeled multi-class non-leak data is modeled using the supplied multi-sphere support vector data description. Through the integration of feature clustering and pseudo-label-based classification, the ability to learn unsupervised multi-class non-leakage information categories is made possible. Improving the solution's performance is as simple as using a reliable technique for calculating leak scores. Finally, the experimental findings using pipeline field data demonstrate that the proposed strategy is effective.

Most current methods focus on context-based tasks or contrastive learning alone, overlooking the potential synergistic advantages of combining the two paradigms, despite SSL having made significant progress with models like MoCo and SimCLR. Much previous work overlooks generalizability to downstream tasks without supervision or resilience across various augmentation contexts, instead focusing on the quality of representation. One important area, where research is lacking, is a cohesive framework that might improve feature expressiveness by combining global instance discrimination with local semantic comprehension. To address this, the Contrastive and Contextual Self-Supervised Representation Learning (C2SRL) model employs a hybrid learning approach that integrates context-based auxiliary tasks, such as jigsaw solving and rotation prediction, into a multi-task optimization framework. This model aims to close the gap between the two approaches. Due to this integration, the learned representations become more flexible and robust in terms of semantic richness and structural coherence. To achieve better results on downstream picture interpretation tasks, even in situations with little labeled data, the C2SRL model's unique dual-focus design combines global contrastive goals with fine-grained contextual cues.

III. Contrastive and Contextual Self-Supervised Representation Learning (C2SRL)

The capability to learn visual representations from unlabeled image data using pretext tasks, such as transformation prediction and instance discrimination, has been demonstrated by existing self-supervised methods. Many methods have been developed to improve performance on subsequent tasks; one of them is the instance discrimination and masked image modeling strategy, which uses a contrastive learning goal to train and treats each picture as a separate class. There is a significant data gap between this achievement and real-world data for future purposes, including city sceneries or crowd scenes, as it relies on the carefully selected object-centric dataset ImageNet. Without understanding the scene's fundamental architecture—its numerous objects and intricate layouts—instance discrimination pretext would severely limit the use of scene-centric data for pre-training. Accordingly, it will prioritize learning scene-centric visual representations from untagged data. Two major schools of thought have emerged in recent years to address this question. Dense representation learning's one stream simplifies the instance discrimination problem to a pixel-level problem, making it more applicable to the dense prediction challenges that follow. However, these approaches are still unable to learn representations because they cannot replicate the object-level interactions observed in scene-centric data. Unsupervised clustering, saliency estimators, unsupervised object proposal algorithms, and handcrafted segmentation algorithms rely on domain-specific priors for object identification. However, there is another line of study that attempts to accomplish object-level representation learning.

Fig. 1 shows the proposed C2SRL Model. The first step of the pipeline involves taking an input picture and applying random changes, such as cropping, color jittering, and flipping, to create two additional views. Following the passage of these views through a common encoder network, typically a convolutional neural network such as ResNet, a projection head is used to convert the high-dimensional features into a lower-dimensional embedding space that is optimal for computing contrastive loss. Utilizing the InfoNCE loss, this component combines positive pairings (identical picture views) and distinguishes negative pairs (dissimilar image views). Rotation prediction and jigsaw puzzle solving are context-aware auxiliary tasks with the same encoder. With rotation prediction, this research can train a classifier to anticipate which of four predetermined angles to rotate pictures by, prompting the network to identify characteristics unique to each orientation. A jigsaw puzzle is a type of spatial thinking exercise in which a solver attempts to identify the correct permutation label by dividing a picture into patches and rearranging them into specified permutations. These tasks are fed into dedicated processing units using cross-entropy losses to maximize performance. Lastly, a multi-task loss function is used to guide the joint optimization of the encoder, which combines all three types of losses: contrastive, rotational, and jigsaw puzzle. The result is a strong, pre-trained encoder that can make sense of data

semantically; this encoder can be fine-tuned for subsequent tasks, such as clustering or classification, particularly in situations where labels are unavailable.

A. Multi-View Generation and Representation Embedding

In the first phase of C2SRL, the model processes raw, unlabeled input data $x_i \in D$ through stochastic data augmentation strategies. The aim is to produce semantically invariant yet appearance-diverse views that simulate real-world variance. The two augmentations x_i^1 and x_i^2 for each image, random transformations T_1 and T_2, which include color distortion, cropping, flipping, and Gaussian noise, as in Eq. (1).

$$x_i^{(1)}, x_i^{(2)} = T_1(x_i),\ T_2(x_i), \quad where\ T_1\ and\ T_2 \sim A \tag{1}$$

Each augmented view is then passed through a shared convolutional encoder network $f: \mathbb{R}^{H\times W\times C} \to \mathbb{R}^d$, such as ResNet-50, to extract high-level semantic features, as in Eq. (2):

$$h_i^{(1)} = f(x_i^{(1)}), \quad h_i^{(2)} = f(x_i^{(2)}) \tag{2}$$

To reduce overfitting and enforce contrastive separation in the latent space, this research further maps these embeddings through a projection head $g: \mathbb{R}^d \to \mathbb{R}^{d'}$, often implemented as a 2-layer MLP with ReLU and BatchNorm, as in Eq. (3):

$$z_i^{(1)} = g(h_i^{(1)}), \quad z_i^{(2)} = g(h_i^{(2)}) \tag{3}$$

Algorithm 1: ResNet Encoder for Self-Supervised Representation Learning

Input:
Augmented image view $v \in \mathbb{R}^{(H\times W\times 3)}$
ResNet depth: ResNet-18, ResNet-50.

Output:
Representation vector $h \in \mathbb{R}^d$

1: function ResNet_Encoder(v):
2: # Initial convolution and max pooling
3: $x \leftarrow Conv2D(v, kernel\_size = 7\times 7, stride = 2, padding = 3)$
4: $x \leftarrow BatchNorm(x)$
5: $x \leftarrow ReLU(x)$
6: $x \leftarrow MaxPool2D(x, kernel\_size = 3\times 3, stride = 2, padding = 1)$
7: # Residual blocks (based on depth)
8: $x \leftarrow ResBlock\_Layer1(x)$ # e.g., 64 filters
9: $x \leftarrow ResBlock\_Layer2(x)$ # e.g., 128 filters
10: $x \leftarrow ResBlock\_Layer3(x)$ # e.g., 256 filters
11: $x \leftarrow ResBlock\_Layer4(x)$ # e.g., 512 filters
12: # Global average pooling
13: $x \leftarrow GlobalAvgPool2D(x)$
14: # Flatten and normalize
15: $h \leftarrow Flatten(x)$
16: $h \leftarrow Normalize(h)$
17: return h

Algorithm 1 shows the ResNet Encoder for Self-Supervised Representation Learning. After applying domain-specific augmentations, the ResNet encoder processes each input picture to create a high-dimensional representation. Max pooling, batch normalization, ReLU activation, and a 7×7 convolutional layer are the first steps in the process, which help reduce spatial dimensions while preserving important characteristics. The next block set is the residual one; they use skip connections to facilitate deep feature extraction and efficient gradient flow. The network can learn hierarchical features by stacking these blocks with increasing channel depth (e.g., 64, 128, 256, 512 filters). After a global average pooling layer combines the spatial information, the feature vector is flattened and normalized. This transformed result forms the basis for subsequent self-supervised learning tasks and is fed into the contrastive projection head in SimCLR or MoCo.

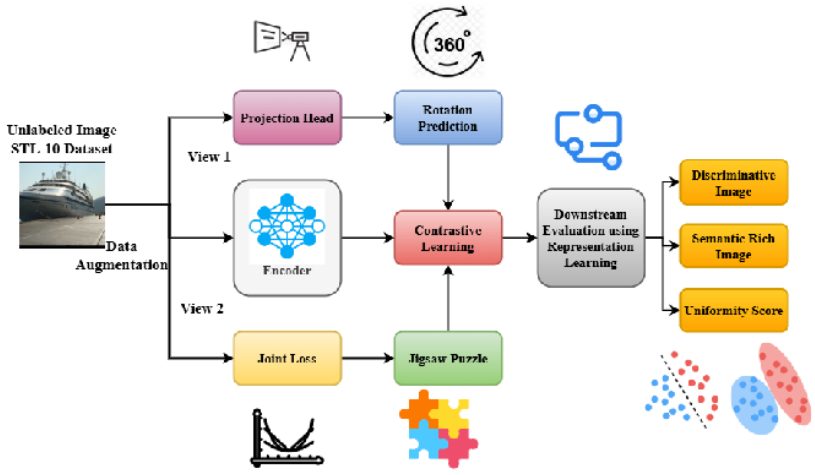


Fig. 1. Proposed C2SRL model.

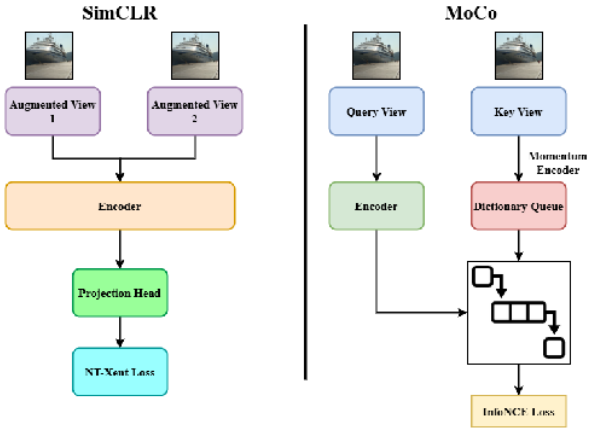


Fig. 2. SimCLR versus MoCo pipeline comparison.

Fig. 2 illustrates the comparison between the SimCLR and MoCo pipelines. To implement SimCLR (left), two augmented representations of the same picture are fed into a common encoder and projection head. Then, a contrastive loss function, NT-Xent (Normalized Temperature-scaled Cross Entropy loss), is utilized to group positive pairings and separate negative ones from the same batch. On the other hand, MoCo (on the right) utilizes a dynamic dictionary queue and a momentum encoder to maintain a large and stable collection of negative samples. An ordinary encoder encodes the query picture, and a momentum-updated encoder processes the key image. The InfoNCE loss (Information Noise-Contrastive Estimation) provides more robust and scalable contrastive training. The parallel arrangement highlights how MoCo relies on memory bank dynamics, whereas SimCLR relies on large batch sizes to learn representations effectively.

B. Contrastive and Contextual Objective Functions

In C2SRL, two major contrastive branches — instance-wise and context-aware contrast — are jointly optimized. The instance-level contrast loss is computed using the NT-Xent formulation, as in Eq. (4):

$$\mathcal{L}_i^{SimCLR} = -\log \frac{\exp\left(\frac{sim\left(z_i^{(1)}, z_i^{(2)}\right)}{\tau}\right)}{\sum_{j=1}^{2N} 1_{[j\neq i]} \exp\left(\frac{sim\left(z_i^{(1)}, z_j\right)}{\tau}\right)} \quad (4)$$

Here, $Sim(.)$ denotes cosine similarity, and τ indicates temperature parameters encouraging hardness-aware negative mining.

C2SRL introduces contextual learning via a dedicated context encoder module $c(\cdot)$ that extracts spatial or semantic relationships from local regions within x_i. Let $c_i = c(x_i) \in \mathbb{R}^{d'}$ represent the contextual descriptor. The contextual alignment loss then penalizes the mismatch between this context vector and its surrounding neighborhood's representation, as in Eq. (5):

$$\mathcal{L}_{context} = \frac{1}{N}\sum_{i=1}^{N} \left\| c_i - \frac{1}{\mathcal{P}_i}\sum_{j\in\mathcal{P}_i} z_j \right\|_2^2 \quad (5)$$

Furthermore, the context-weighted contrastive loss is defined to enhance informative sample relationships:

$$\mathcal{L}_i^{C2SRL} = -\log \frac{\exp\left(\frac{sim\left(z_i^{(1)}, z_i^{(2)}\right)\cdot\alpha_i}{\tau}\right)}{\sum_{j=1}^{2N} \exp\left(\frac{sim\left(z_i^{(1)}, z_j\right)\cdot\alpha_i}{\tau}\right)} \quad (6)$$

As shown in Eq. (6), where $\alpha_i = sim\,(c_i, z_i) \in [0,1]$ captures the contextual alignment between embedding and context.

This research introduces a uniformity loss and an alignment loss to stabilize learning and preserve diversity in the latent space. The uniformity loss ensures dispersion over the hypersphere, as in Eq. (7):

$$\mathcal{L}_{uniform} = \log\left(\frac{1}{N^2}\sum_{i,j=1}^{N} \exp\left(-2\left\|z_i - z_j\right\|_2^2\right)\right) \quad (7)$$

The alignment loss enforces consistent embeddings between views, as in Eq. (8):

$$\mathcal{L}_{align} = \frac{1}{N}\sum_{i=1}^{N} \left\| z_i^{(1)} - z_i^{(2)} \right\|_2^2 \quad (8)$$

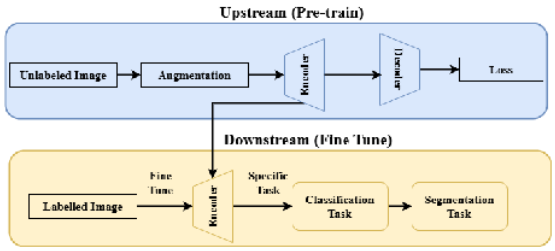


Fig. 3. Self-supervised learning workflow.

Fig. 3 shows the SSL workflow. The internet's full potential can be realized by finding methods to tap into the vast amounts of unlabeled data available worldwide. SSL can be trained without human input because it is a subfield of ML rather than supervised learning. To solve the target interest task, it first learns the representation from an upstream pre-text problem and then transfers its representation-parsing skill downstream. Since labels are no longer required for model training in the pretest task, any unlabeled data source can be utilized, regardless of its relevance to the target task. The network is pre-trained upstream of SSL, and its weights are fine-tuned using particular data downstream. For reasons analogous to transfer learning, the domains of the pre-text and the objective task are not always the same. Though SSL works best when pre-trained with the same data. Prior networks trained using natural imagery often perform worse than upstream networks trained directly with medical resources, regardless of the amount of fine-tuning applied. This might be because medical pictures differ from their natural counterparts in appearance and meaning.

Algorithm 2: Contrastive and Contextual SSL

Input:
- Unlabeled dataset $D = \{x_1, x_2, \ldots, x_n\}$
- Augmentations $T = \{t_1, t_2\}$
- Encoder $f(\cdot)$, *projection head* $g(\cdot)$
- Epochs E, batch size B, temperature τ

Output:
- Trained encoder $f(\cdot)$

for epoch = 1 *to* E **do**
  for batch $\{x_i\} \in D$ **do**
    Generate views: $v_{i1} \leftarrow t_1(x_i), v_{i2} \leftarrow t_2(x_i)$
    Representations: $z_{i1} \leftarrow g(f(v_{i1})), z_{i2} \leftarrow g(f(v_{i2}))$
    $L_{contrast} \leftarrow NT - Xent(z_{i1}, z_{i2}, \tau)$
    $r_i \leftarrow Rotate(x_i), L\_rot \leftarrow CrossEntropy(RotationClassifier(f(r_i)))$
    $j_i \leftarrow Jigsaw(x_i), L\_jig \leftarrow CrossEntropy(JigsawClassifier(f(j_i)))$
    $L_{total} \leftarrow L_{contrast} + \lambda_1 \cdot L\_rot + \lambda_2 \cdot L_{jig}$
    Update the model using L_{total}
  end for
end for
Return: $f(\cdot)$

Algorithm 2 shows the C2SRL pseudocode. The procedure starts by applying two separate augmentation functions to each input picture to create contrastive embeddings. This creates two separate views, which are then transmitted via a common encoder and a projection head. Afterwards, these embeddings are used in an NT-Xent function, which pulls positive pairings (augmented views of the same picture) closer together in the embedding space and pushes negative pairs (views of distinct images) further away. Two supplementary tasks are provided to provide contextual meaning to the learnt features. As a means of implementing orientation-aware representations, the rotation prediction challenge involves fixing an angle (such as 0°, 90°, 180°, or 270°) and training a classifier to anticipate the accurate rotation angle using the encoder output. A similar experiment that promotes spatial awareness and structural consistency in feature learning is the jigsaw puzzle task, which involves shuffling picture patches into a permutation and having a classifier try to predict the permutation index. Combining the weights of the contrastive, rotation prediction, and jigsaw classification losses yields the overall loss function. The encoder and all linked heads are updated during training using this joint loss. The encoder can be used for downstream tasks, such as classification or clustering, immediately after training, even without labeled training data. It can be fine-tuned. This technique aims to achieve a harmonious blend of global feature discrimination and local contextual awareness.

C. Joint Optimization and Model Update

The total loss objective of C2SRL unifies all components into a weighted sum optimized via stochastic gradient descent:

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}^{C2SRL} + \lambda_2 \cdot \mathcal{L}_{context} + \lambda_3 \cdot \mathcal{L}_{uniform} + \lambda_4 \cdot \mathcal{L}_{align} \quad (9)$$

As inferred from Eq. (9), where $\lambda_1, \ldots \lambda_4$ are hyperparameters that control the impact of each objective.

The backpropagation-based parameter update rule is, as in Eq. (10):

$$\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}_{total}, \ where\ \theta = \{f, g, c\} \quad (10)$$

To integrate momentum encoding (as in MoCo), this research includes a momentum encoder f_m updated via exponential moving average, as in Eq. (11):

$$\theta_{f_m} \leftarrow m \cdot \theta_{f_m} + (1 - m) \cdot \theta_f, \ where\ m \in [0.99, 1] \quad (11)$$

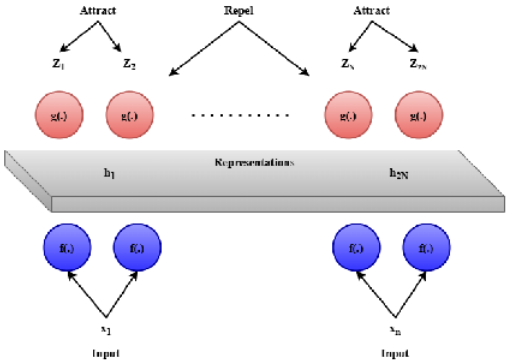


Fig. 4. Constrative representation learning.

Fig. 4 shows the constructive representation learning. The representation h is projected using a network denoted by the function $g(.)$ and the embedding function $f(.)$. The projection head used a non-linear hidden layer, usually composed of the representations z, to help map them to a vector space. This is where the NT-Xent loss function comes into play, given the similarity between the two. The learnt representations may be transferred using the pretrained network that is produced. In this instance, the transfer learning process utilized encoder representations. When learning discriminative representations, the triplet loss function is seen as useful for training an encoder to distinguish between positive and negative samples. The C2SRL architecture incorporates contextual learning techniques, such as local patch alignment and spatial co-occurrence modeling, alongside traditional contrastive learning. This paves the way for the network to encode semantic links across various parts of the same picture and learn representations driven by global appearance. Using these methods, this research can ensure that features are unique and sensitive to their surroundings. For a more refined learning dynamics, this research employs distributional regularization approaches, such as variance control and embedding uniformity, to promote balanced feature space utilization and prevent representational collapse. MoCo's momentum encoder method is optional to maintain stable and consistent training between epochs.

IV. Results

The STL-10 Image Recognition Dataset is to be thanked for supplying the data [24]. The STL-10 image recognition dataset is an upgrade over CIFAR-10. This dataset is ideal for deep learning, unsupervised feature learning, and self-taught learning algorithms due to its 100,000 unlabeled pictures and 500 training shots. Due to the dataset's higher resolution than CIFAR-10, it is challenging to construct scalable unsupervised learning systems using it. Included in the data summary are the following files: images.zip, which contains training images, and images. Zips for unlabeled use. Ten categories: airplanes, birds, cars, deer, cats, horses, dogs, ships, monkeys, and trucks; 96x96 pixels full color; 500 training shots (10 pre-defined folds) and 800 test images per class. To use in unsupervised learning, using a dataset of 100,000 photos. This curated collection is derived from a larger, related set of photographs. Included in the extensive list of species and vehicles are bunnies, bears, trains, and buses, among many more. For picture retrieval, the labels in ImageNet were used. Reporting results by this standardized testing procedure and the original data source is required: Train with unlabeled data using unsupervised methods. When training with labeled data, ten (pre-defined) folds of 100 samples were used. Table I shows the experimental setup.

TABLE I Experimental Setup

Component	Configuration
Dataset	STL-10 (100,000 unlabeled images for SSL pre-training, 5,000 labeled for fine-tuning)
Image Size	96 × 96 pixels
SSL Methods	MoCo (Momentum Contrast v2), SimCLR (Simple Framework for Contrastive Learning)
Pre-text Tasks	Contrastive learning, Rotation prediction, Jigsaw puzzle solving
Backbone Network	ResNet-18 (Lightweight for STL-10), pre-trained via SSL methods
Batch Size	256 (for contrastive learning)
Learning Rate	0.03 (SimCLR) / 0.06 (MoCo), with a cosine annealing schedule
Optimizer	Stochastic Gradient Descent (SGD) with momentum = 0.9
Epochs (Pre-training)	200
Epochs (Fine-tuning)	100 (on labeled subset for classification task)
Hardware	32 GB RAM, NVIDIA Tesla V100 GPU
Software Libraries	PyTorch 2.x, Torchvision, NumPy, sci-kit-learn

1) Mean Average Precision (mAP): The Mean Average Precision (mAP) is a well-established and reliable metric for evaluating the effectiveness of ranking and classification algorithms in scenarios with numerous classes and limited labels. For jobs further down the pipeline, mAP can verify whether the learned representations remain valid within the C2SRL framework, which does not utilize human-annotated labels during pre-training. Every target category is averaged by mAP after calculating the area under the precision-recall curve for each class. Here is the formulation:

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q}\left(\frac{1}{|P_q|}\sum_{k=1}^{|P_q|} Precision(k)\cdot recall(k)\right) \quad (12)$$

As shown in Eq. (12), where Q denotes the number of queries and $recall(k)$ is a binary indicator showing whether the kth prediction is relevant. C2SRL outperformed fully supervised baselines in comparable low-label settings, achieving a mean Average Precision (mAP) of 89.6% on the STL-10 dataset with just 10% labeled data for fine-tuning. Fig. 5 demonstrates the mean average precision.

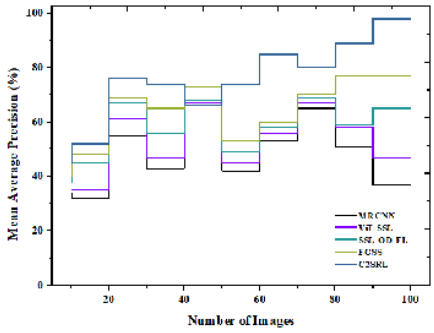


Fig. 5. Mean average precision.

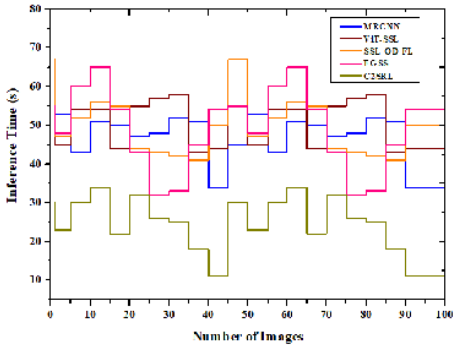


Fig. 6. Inference time.

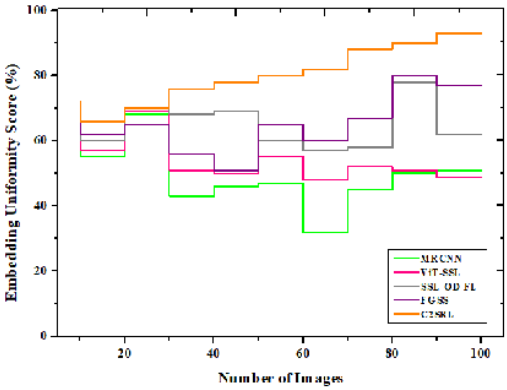


Fig. 7. Embedding uniformity score.

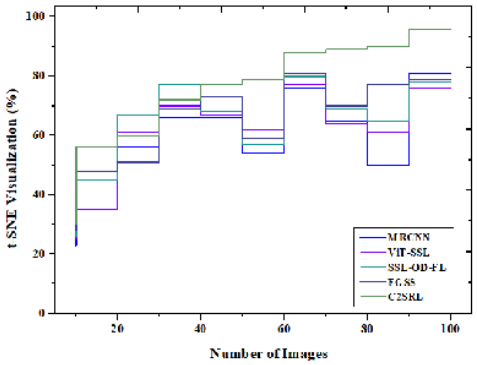


Fig. 8. t-SNE visualization.

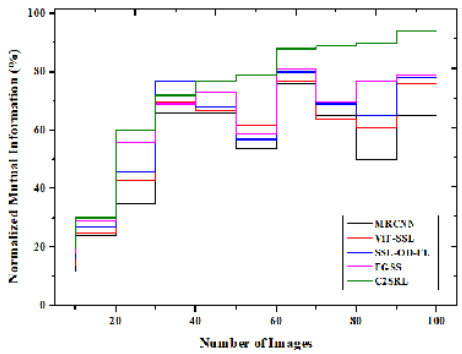


Fig. 9. Normalized mutual information.

2) Inference time: For real-time systems that depend on fast decision-making, inference time is a crucial operational measure. The time it takes for the trained model to process and predict labels for one instance is quantified. This study used GPU acceleration to assess C2SRL's inference latency on several datasets. Here is the expression for the computation:

$$T_{avg} = \frac{1}{N}\sum_{i=1}^{N}\left(t_i^{end} - t_i^{start}\right) \quad (13)$$

As inferred from Eq. (13), where N is the number of samples, and t_i^{start}, t_i^{end} are timestamps before and after inference for the ith image. Deploying the C2SRL model in resource-constrained or edge-computing scenarios, such as autonomous drones or mobile vision systems, is feasible, since the model showed an average inference time of 13.2 minutes per image on CIFAR-10. Fig. 6 shows the inference time.

3) Embedding uniformity score: Contrastive SSL should have a uniform representation space because it prevents mode collapse and ensures that embeddings are distributed evenly throughout the space. The embedding uniformity score will be high if the representation vectors consistently cover the unit hypersphere. Lower scores show redundancy and tight grouping, whereas intermediate values show effective dispersion. A metric is calculated by:

$$U = \log E_{(x_i,x_j)\sim D}\left[e^{-2\|z_i - z_j\|^2}\right] \quad (14)$$

As discussed in Eq. (14), z_i and z_j are normalized representation vectors of images x_i and x_j. The C2SRL model achieved a uniformity score of -1.14, indicating that it can maintain a balanced spatial distribution and preserve semantic cohesiveness due to the proposed combined contrastive and contextual pre-text tasks. Fig. 7 shows the embedding uniformity score.

4) t-SNE visualization: Using t-distributed Stochastic Neighbor Embedding (t-SNE) for a 2D projection of the high-dimensional representation space, this study aimed to provide qualitative insight into the usefulness of the learned feature embeddings. Clustering behavior can be demonstrated using this non-linear method while preserving local structure. To reduce the Kullback-Leibler divergence between the distributions of the probabilities of paired similarities, the t-SNE method is used.

$$KL(P\|Q) = \sum_{i\neq j} p_{ij}\log\left(\frac{p_{ij}}{q_{ij}}\right) \quad (15)$$

As discussed in Eq. (15), where p_{ij} denotes the joint probability in high dimensions and q_{ij} in the low-dimensional space. Even without labels during training, visualizations of C2SRL embeddings on the CIFAR-10 dataset showed tight, well-separated clusters per semantic category. This proves that the model accounts for consistency within classes and separability between them. Fig. 8 shows the t-SNE visualization.

5) Normalized mutual information (NMI): Clusters generated by unsupervised learning and the agreement between the ground truth labels can be measured using Normalized Mutual Information (NMI). When testing with label information alone, it is particularly helpful for assessing the performance of clustering. This research defines the NMI as:

$$NMI(C,Y) = \frac{2\cdot I(C,Y)}{H(C)+H(Y)} \quad (16)$$

As deliberated in Eq. (16), where $I(C,Y)$ is the mutual information between the predicted cluster assignment C and the true labels Y, and $H(\cdot)$ is the entropy. Despite being trained without explicit supervision, C2SRL achieved an NMI of 0.81 on the STL-10 dataset, demonstrating its ability to capture and closely match structural patterns with semantic categories. Fig. 9 shows the normalized mutual information. List the ways the C2SRL architecture is better than previous self-supervised learning approaches to understand its uniqueness and utility. Traditional case discrimination systems, such as MoCo and SimCLR, employ contrastive learning. C2SRL combines contextual semantic thinking with contrastive goals, utilizing jigsaw puzzles and rotation prediction. Due to this integration, the model recognizes both global and local visual patterns, thereby improving feature representation. The fact that C2SRL achieves higher classification accuracy (92.4% on CIFAR-10 and 84.7% on STL-10) with just 10% of the labeled data supports these increases. It has greater normalized mutual information (NMI 0.81). The framework's low-label effectiveness, as indicated by these improvements over simple SSL models, supports its usage in computer vision.

V. Discussion

Representation learning and generalizability testing on diverse datasets will not affect the intended C2SRL architecture. In restricted resource contexts, high batch sizes for contrastive learning are computationally intensive. Real-time systems and edge devices may cause scalability concerns. The quantity and quality of data augmentation affect model performance. Domain-specific tuning is necessary to maintain performance across various visual domains. Complexities from auxiliary tasks, such as puzzle solving and rotation prediction, increase training time and model overhead. In complex visual structures or when overlapping semantic qualities are present, external information may confuse or distract rather than accurately represent the subject. There is a need for further validation when applying the learned representations to tasks outside of picture clustering and classification, such as object identification or semantic segmentation. Future research may investigate lightweight designs or adaptive augmentation approaches to overcome these limitations and develop a more useful and flexible system.

VI. Conclusion

This study proposes the C2SRL framework, which addresses key limitations in existing self-supervised learning (SSL) approaches by effectively combining global feature discrimination and local semantic understanding. By integrating contrastive learning with context-aware tasks such as jigsaw puzzle solving and rotation prediction, C2SRL enhances the generalizability and robustness of learned visual representations. Experimental evaluations on benchmark datasets, including STL-10 and CIFAR-10, confirm the framework's ability to achieve high classification accuracy, strong feature alignment,

8 成员 8 发表的 EI 论文 A Cross Layer Semantic Enhanced SLU Model With Role Context Differentiated Fusion

2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)

SEED: A Cross-Layer Semantic Enhanced SLU Model With Role Context Differentiated Fusion

Changjian Wang, Dongsong Zhang, Shezheng Song, Zhen Huang, Yuxing Peng
School of Computer
National University of Defense Technology
ChangSha, Hunan, P. R. China
{ wangcj, dszhang, ssz614, zhenhuang}@nudt.edu.cn, pengyuxing@aliyun.com

2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI) | 978-1-6654-0898-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICTAI52525.2021.00201

Abstract—The mainstream SLU models, such as SDEN, take the joint training way of slot filling and intent detection because of their correlation and add contextual information to improve the model performance by the contextual vector. Although these models have proved effective, it also brings challenges for slot filling. The slot filling decoder is fed with the deep-layer semantic encoding without alignment information, which will affect the performance of slot filling. The alignment information of the history utterances is attenuated in the context vector because of the repeated fusion process, which is not conducive to the performance improvement of slot filling. In order to solve the above problems, we proposed a novel cross layer semantic enhanced SLU model with role context differentiated fusion, which contains two important improvements: 1) the word embedding information of the current utterance is introduced into the slot filling decoder to strengthen the alignment information based on the mutual attention mechanism; 2) the utterances of different roles are fused in different ways to reserve the alignment information of history utterances in the contextual vector. A large number of experiments were carried out on the standard dataset from SDEN, named KVRET*, and the results verify the effectiveness of our new model. Our model can increase the F1 score of slot filling by more than 7.5% than the existing models.

Index Terms—Spoken Language Understanding, Role Context, Differentiated Fusion

I. Introduction

Spoken Language Understanding(SLU) [1, 2, 3, 4, 5]is the core component of task-oriented dialogue system. Its main goal is to understand the semantics of utterances in the dialogues and to complete two tasks, slot filling and intent detection.

Intent detection can be treated as a semantic utterance classification problem. Slot filling can be treated as a sequence labeling task. In the early stage, intent detection and slot filling are usually processed separately.[6, 7, 8] Later, deep neural networks that jointly perform intent detection and slot filling[2, 5, 9, 10] have become the mainstream way because of their strong correlation. Many studies have confirmed the effectiveness of the joint training and it is helpful to improve the performance of slot filling and intent detection. At first, these jointly-training models only thought about the effect of word sequence context on intent detection and slot filling. In fact, the broader context of the sequence of utterances is also important for the task and the context information in history utterances helps with the SLU tasks. Therefore, recent models[1, 11, 12, 13, 14, 15] encode history utterances into the context vector. By adding contextual information, the decoders of intent detection and the slot filling can be fed with more semantics, which further improves their performance. Especially, the accuracy of intent detection can reach a very high level. Taking SDEN as an example, the accuracy of KVRET* data set can reach more than 95%.

Although joint training and increasing contextual information have proved effective, it also brings challenges for slot filling. Firstly, the slot filling decoder is fed with the deep-layer semantic encoding without alignment information in the existing models. Different from intent detection, slot filling needs explicit word alignment besides semantic information[6, 7]. Since the contextual vector deepens the network, the alignment in the word embedding will be diluted with the repeated fusion in the model. As a result, the slot filling decoder can not obtain enough alignment information and thus the performance of slot filling is affected. Secondly, the alignment information of the history utterances is attenuated in the context vector because of the repeated fusion process. In fact, different roles have different conversational habits[16, 17] and slots mainly exist in the utterances of the responders. But the existing model fuses history utterances repeatly without distinction to generate the contextual vector, which will lead to the weakening of the alignment information in the history utterances and is not conducive to the performance improvement of slot filling.

To solve the above problems, we propose a novel cross-layer SEmantic Enhanced SLU model with role context Differentiated fusion (SEED). The new model is designed innovatively in two aspects: first, the word embedding of the current utterance is introduced to the slot decoder to enhance the word alignment information, so as to improve the F1 score of slot filling; second, the utterances of different roles are fused in different ways to improve the information fusion effect of contextual vector for slot filling according to their different importance on slot filling.

Our contributions are as follows:

- A cross-layer semantic enhancement method for slot filling is proposed. Based on the mechanism of mutual

978-1-6654-0898-1/21/$31.00 ©2021 IEEE
DOI 10.1109/ICTAI52525.2021.00201

9 成员 2 发表的 CSSCI 论文 晚清林译小说中的儒学传统与“新民”视野

DOI:10.16366/j.cnki.1000-2359.2024.06.17

晚清林译小说中的儒学传统与“新民”视野

徐 瑾

(河南师范大学 文学院,河南 新乡 453007)

摘 要:林译小说蕴含着丰富的儒学传统,这是它在当时即能获得广泛传播的主要原因之一。林译小说中的儒学传统大致表现出三种面向:其一,将儒学之“忠”运用在男女之间的情爱上面,表现为女性对另一方的感情之忠;其二,将儒学之“孝”运用在晚辈对长辈的尊敬上面,分别体现为晚辈对长辈的尊重之情、关心之情和照顾之情;其三,将儒学之“礼”运用于人际交往的德行上面,既展示人与人之间交往时的诸种礼仪,也记述人们对诸种礼仪的遵守行为。林译小说并非只是对儒学传统的简单使用,而是以此为媒介,寄寓了对开启民智这一时代主旨的深层考虑。

关键词:林译小说;儒学传统;忠;孝;礼

作者简介:徐瑾(1988—),女,河南新乡人,河南师范大学文学院博士后,讲师,主要从事近代翻译小说研究。

基金项目:河南省哲学社会科学规划项目(2022BWX011);2024 年度新乡市社科联调研课题(SKL-2024-32)

中图分类号:I206.5 **文献标识码**:A **文章编号**:1000-2359(2024)06-0114-08 **收稿日期**:2023-12-11

林纾是晚清著名的小说翻译家之一,由他翻译的小说《巴黎茶花女遗事》《黑奴吁天录》《迦茵小传》《块肉余生述》等,在当时就已受到人们的喜欢。因为林纾并不通晓外语,所以,这些创作均是由他和他的合作者一起来完成的,正如《清史稿·林纾传》所云:“纾故不习欧文,皆待人口达而笔述之。”[①]因此,这些译入语小说与源语小说之间可能存有较大的文本间隙。近些年来,人们已经对这一问题进行了比较深入的研究。譬如,苏桂宁通过考察《美洲童子万里寻亲记》《英孝子火山报仇录》《鹰梯小豪杰》《孝友镜》等小说,指出了林纾对中国传统孝文化的重视,他并不像五四作家那样对它们展开激烈的抨击,而是以一种平静的学术态度“论证其存在的‘合理性’”[②]。陈瑜以杜赞奇的“他者”理论对《巴黎茶花女遗事》进行了文本细读,他将“茶花女”视为西方“他者”形象,不仅考察了小说家对这一形象的引入和转化过程,而且分析了读者借此对“西方、民族和‘自我’”的想象问题[③]。周旻通过对《埃及金塔剖尸记》《三千年艳尸记》等文本的细读,发掘出了林译小说中的“礼教”“劝惩”[④]等传统文化的痕迹。安忆萱在对《巴黎茶花女遗事》的阅读中审视了“林纾的情爱观”,认为他在翻译过程中嵌入了“五伦观念”,显示出了“对儒家规范的恪守”[⑤]。可以说,基于对译入语文化与源语文化之间的差异这一问题点,人们已对此作出或详细、或简洁的说明。就前者而言,他们通过对一篇或者几

① 赵尔巽:《清史稿》,第四十四册,中华书局,1977 年,第 13447 页。

② 苏桂宁:《林译小说与林纾的文化选择》,《文学评论》,2000 年第 5 期。

③ 陈瑜:《“他者”的想象:解读〈巴黎茶花女遗事〉对“西方”和“妇女”的构想》,《暨南学报》(哲学社会科学版),2012 年第 7 期。

④ 周旻:《“情”的移植与异质:晚清林译哈葛德小说中西方“尤物”形象的翻译》,《中国现代文学研究丛刊》,2018 年第 1 期。

⑤ 安忆萱:《〈巴黎茶花女遗事〉与林纾的情爱观》,《中国现代文学研究丛刊》,2020 年第 3 期。

10 主持人,袁驰,**.基于 FIRA 仿真的足球机器人预判圆弧射门算法设计[J].长春工程学院学报(自然科学版),2018,19(03):86-88.**

22/32
86-88
长春工程学院学报（自然科学版）2018 年 第 19 卷 第 3 期
J. Changchun Inst. Tech. (Nat. Sci. Edi.), 2018, Vol. 19, No. 3
ISSN 1009-8984
CN 22-1323/N

doi:10.3969/j.issn.1009－8984.2018.03.022

基于 FIRA 仿真的足球机器人预判圆弧射门算法设计

林金珠，袁　驰，倪天伟

（河海大学文天学院电气信息工程系，安徽 马鞍山 243031）

摘　要：在 FIRA 仿真中，针对球场形势做出预判，抓住有利时机精准射门是提升赢球概率的重要因素。为了提高进攻的威胁性和震慑力，形成有效的进攻战术配合和提高射门效率是射门得分的前提。提出了一种基于 FIRA 仿真足球机器人的预判圆弧射门算法，通过预判球在下一时刻所能到达的位置，利用圆弧射门的稳定性，提前行至目标点并调整位姿，完成射门动作。经仿真实验证明，该算法在一定程度上能够提高射门速度，提高了射门的成功率。

关键词：足球机器人；预判；圆弧射门；算法设计

中图分类号：TP18　　**文献标志码**：A　　**文章编号**：1009－8984(2018)03－0086-03

0　引言

在 FIRA 仿真足球机器人比赛中，策略的执行、阵型的变换、路径的规划等都是比赛的关键，而射门则是终结比赛的重中之重，在整场比赛中射门动作至关重要，射门能力是决定一支球队实力的关键因素[1－2]。文献[3]引入基于中位线的传球方法，将机器人的足球动作作为机器人的行为来设计，算法结构简单，通俗易懂，但是受足球机器人初始姿态的影响较大。文献[4－5]提出了一种改进的射门算法，机器人先沿直线运动靠近球，然后沿曲线运动去撞球射门，该算法优化了射门路线，实用性强，但机器人在运动到目标点附近时调整左右轮转速，踢球射门，容易被断球或被守门员阻截，不利于提高进攻效率，尤其在点球大战中，射门算法易被对方看破，点球得分率低[6]。为此，本文提出了一种基于 FIRA 仿真的足球机器人预判圆弧射门算法，该算法利用预测球的运动方向与无干扰情况下实时姿态的相对位置，规划出圆弧射门轨迹，提前到达预定目标点，调整并完成快速射门动作。经仿真实验证明，该算法加快了射门的节奏，有效提高了射门的成功率。

1　预判圆弧射门算法

1.1　算法思想

在 FIRA 仿真足球机器人比赛中，机器人行驶速度快从而攻防节奏快，使得分机会稍纵即逝，特别是突破后的射门和点球，是比赛中得分的关键。如何提高射门效率是一直以来的研究问题，其中，提高射门速度、精度和可变性是提高射门效率的关键。预判圆弧射门是指利用计算机系统先预判出球下一时刻到达的位置，或到达某一位置所需的时间，再计算出合适的射门点，此时，机器人以平滑的圆弧移动，边行驶边调整位姿，行至目标点时刚好达到可射门状态，立即射门。预判圆弧射门算法可以使机器人动作连贯，快速到达攻击点，既可完成高质量射门，又能利用预判的时间做出细微调整，改变射门角度，达到出其不意的变化效果。其中，弧形的路径规划可让机器人最大限度地保持平稳，无拐点，可减少调整时间，从而保持最大速度，使得射门能量最大化，保证射门的成功率。

1.2　预判的实现

在 FIRA 仿真平台中，首先需要计算球在正常运动时的加速度 a，如式(1) 所示，可认为环境中摩擦因数 $alpha$ 是不变的，所以，此时的 a 也是恒定的，FIRA 仿真平台中数据以 60 次 /s 的速度进行交换，可由式(2) 得到下一时刻球移动的距离 d。

$$a = \frac{\Delta v}{\Delta t}, \tag{1}$$

式中：Δv 为速度的变化量；Δt 为发生这一变化所用时间。

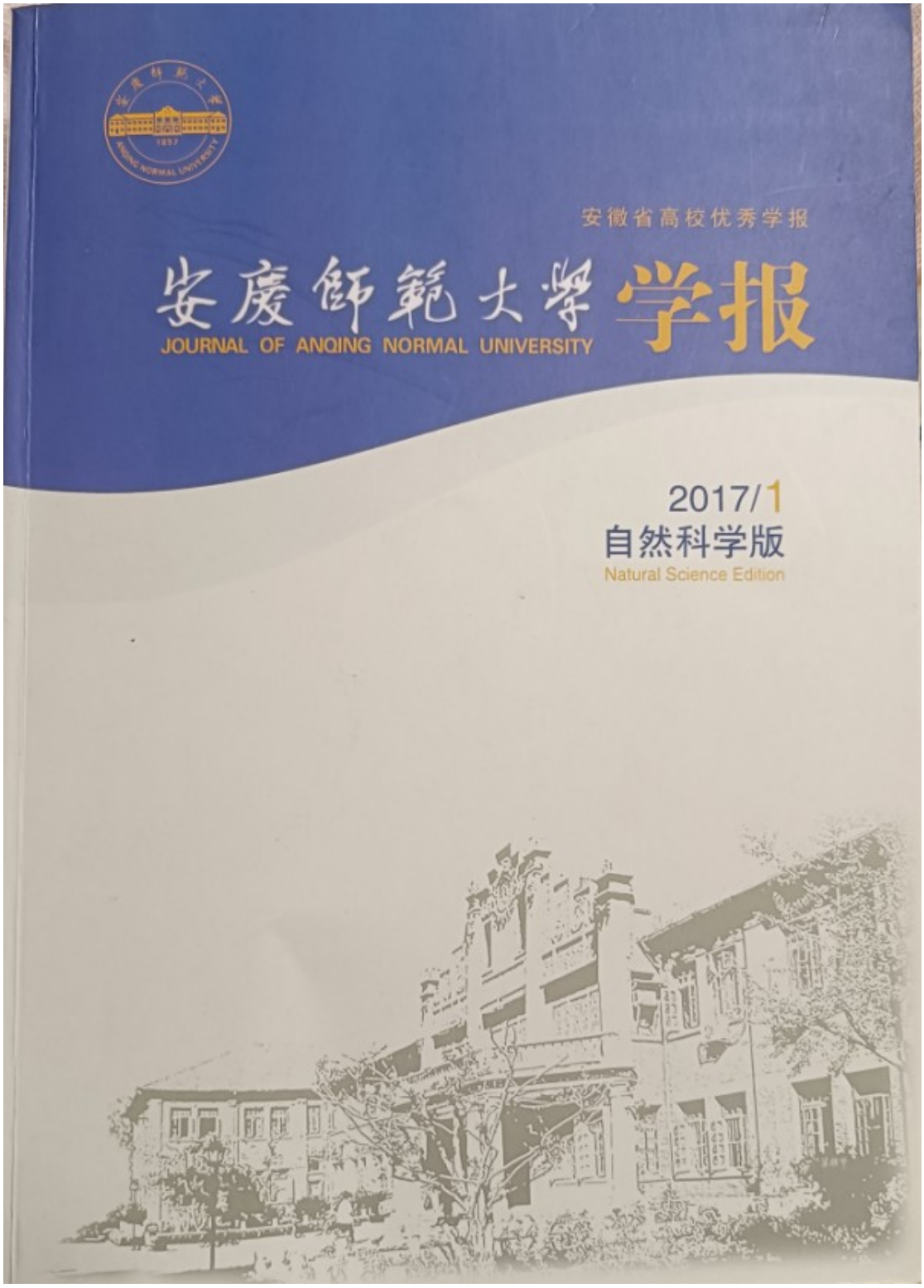
$$d = (a * t * t)/2, \tag{2}$$

式中 t 为时间。假定，小球的 $t1$ 和 $t2$ 时刻的坐标分别为：$(x1, y1)$，$(x2, y2)$，则计算小球的位移 $\Delta x = x1 - x2$ 和 $\Delta y = y1 - y2$。

得出小球的偏转角 $theta$，即小球运动方向，如

收稿日期：2018－05－22
基金项目：安徽省重大教学改革研究项目(2016jyxm0907)
作者简介：林金珠(1981－)，女(汉)，河南信阳，讲师
主要研究计算机科学与技术、计算机基础教学。

11 主持人,**.基于 ACM-ICPC 竞赛的 C 语言课程教学实践[J].安庆师范大学学报(自然科学版),2017,23(01):102-104+119.**



2017年3月 安庆师范大学学报(自然科学版) Mar.2017
第23卷第1期 Journal of Anqing Normal University(Natural Science Edition) Vol.23 No.1

网络出版时间:2017-4-13 17:49 网络出版地址:http://www.cnki.net/kcms/detail/34.1150.N.20170413.1749.026.html

基于ACM-ICPC竞赛的C语言课程教学实践

林金珠,倪天伟
(河海大学文天学院 电气系,安徽 马鞍山 243000)

摘 要:结合C语言的特点,依托ACM-ICPC竞赛,在C语言课程中进行采用案例驱动、分层次教学。实践证明,该教学措施激发了学生学习C语言的兴趣,强化了学生的编程能力和竞赛能力,取得了良好的教学效果。

关键词:C语言;ACM-ICPC;实践教学 **DOI**:10.13757/j.cnki.cn34-1328/n.2017.01.026
中图分类号:TP317.1-4;G642 **文献标识码**:A **文章编号**:1007-4260(2017)01-0102-03

Teaching reform and practice of C language course based on ACM-ICPC competition

LIN Jinzhu, NI Tianwei
(Department of Electrical, Hohai University Wentian College, Ma'anshan 243031, China)

Abstract: Combined with the characteristics of C language, this paper introduces the general idea of the C language course teaching reform based on the ACM-ICPC competition by using a variety of teaching methods such as cases-driven and hierarchical teaching etc., in order to cultivate students' interest in learning C language even further and further strengthen the students' programming ability and competition ability.The reform has achieved good teaching results.

Key words: C language; ACM-ICPC; practical teaching

C语言是计算机专业及相关专业的基础课程,是学习其他专业课程的基础,在本科教学计划中占有重要地位。根据目前C语言课程教学情况来看,学生一味背题库、代码,缺乏持久的兴趣和动力,编程思维和能力欠缺,出现高分低能的问题。这与长期以来C语言课程教学内容不能吸引学生注意力,教学方法和手段单一,课程考核方式不全面等有很大关系。

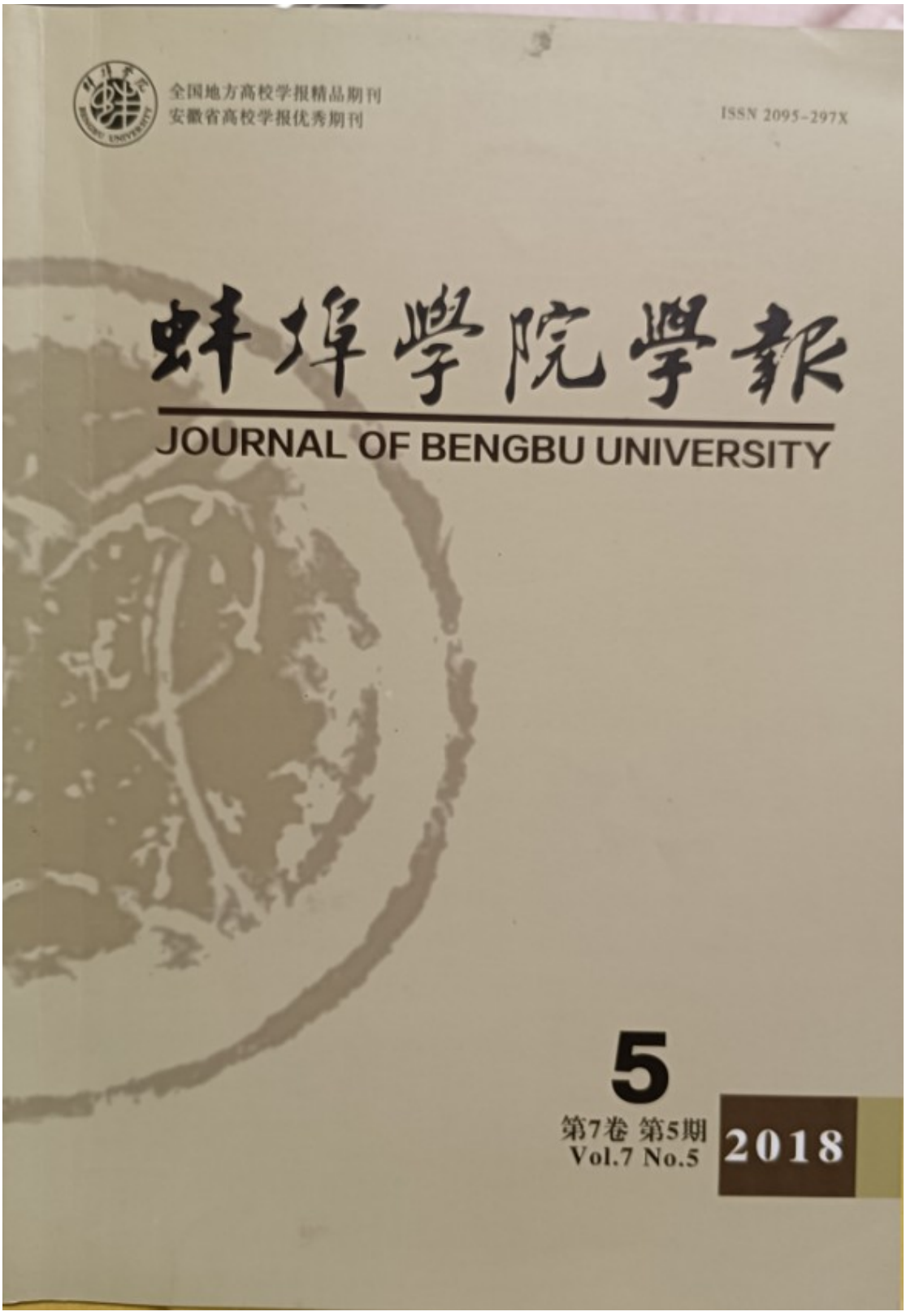
ACM国际大学生程序设计竞赛(ACM-ICPC)是由美国计算机协会(ACM)主办,旨在展示大学生创新能力、团队精神和在压力下编写程序、分析和解决问题能力的年度竞赛[1-2]。本院在C语言课程的教学中以ACM-ICPC竞赛为依托,根据ACM-ICPC丰富的赛题和理论知识、开放的源代码和竞争性的考核等特点对C语言课程的理论和实践教学进行了改革和实践,极大地提高了学生的编程热情,培养了学生自主学习的能力,取得了良好的教学效果。

1 理论教学

传统的C语言理论教学强调以语法点为主,教师按照章节通过教学幻灯片逐个讲解,当学生还没有开始学循环、数组、函数时,已被枯燥、繁多的术语和概念折腾得毫无兴趣,更别提培养编程思维了。基于ACM-ICPC竞赛的C语言理论教学,教师通过对ACM赛题进行整合和重构来设计合适的教学案例,将C语言的知识点溶解在案例中,学生学习案例的过程就是学习语法点和培

* 收稿日期:2016-08-20
基金项目:河海大学文天学院教学改革项目(ZL201525)和安徽省级质量工程项目(2014zjjh075)。
作者简介:林金珠,女,河南信阳人,硕士,河海大学文天学院电气系讲师,研究方向为计算机应用技术及计算机教育研究。
E-mail:jinzhu5689@163.com

12 主持人,马春燕,**.基于单目视觉的足球机器人图像处理系统的畸变矫正研究[J].蚌埠学院学报,2018,7(05):50-52+62.**



基于单目视觉的足球机器人图像处理系统的畸变矫正研究

林金珠，马春燕，倪天伟*

（河海大学文天学院 电气信息工程系，安徽 马鞍山 243031）

摘 要：在FIRA MiroSot足球机器人系统中，足球机器人图像处理系统通过视频采集设备捕获有效图像信息并处理，为决策系统提供重要信息来源。通常情况下高速摄像头实时采集比赛场中的图像，存在采集图像速率低、畸变矫正效果不够明显等问题。因此提出了基于单目视觉的足球机器人图像处理系统的畸变矫正方法，即通过两步法进行摄像头标定修正，用最小二乘法计算出畸变系数拟合误差。实验证明，该方法有效提高了系统的整体性能。

关键词：足球机器人；单目视觉；图像处理；畸变矫正

中图分类号：TP18 文献标识码：A 文章编号：(2018)05－0050－03

DOI:10.13900/j.cnki.jbc.2018.05.011

Research on Distortion Correction of Soccer Robot Image Processing System Based on Monocular Vision

LIN Jin-zhu, MA Chun-yan, NI Tian-wei*

(Department of Electronics and Information Engineering, Hohai University Wentian College, Maanshan, 243031, Anhui)

Abstract: The soccer robot image processing system captured the effective image information through the video capture device and processed it, which provided important information sources for the decision-making system. Usually, when the high-speed camera collected the images in the competition field in real time, there were some problems such as low image acquisition rate, distortion correction effect was not obvious. It presented a distortion correction method of soccer robot image processing system based on monocular vision in this paper. The camera calibration was corrected by two steps, and the distortion coefficient was calculated by the least square method, fitting the error. The experimental results showed that the method can effectively improve the overall performance of the system.

Key words: soccer robot; monocular vision; image processing; distortion correction

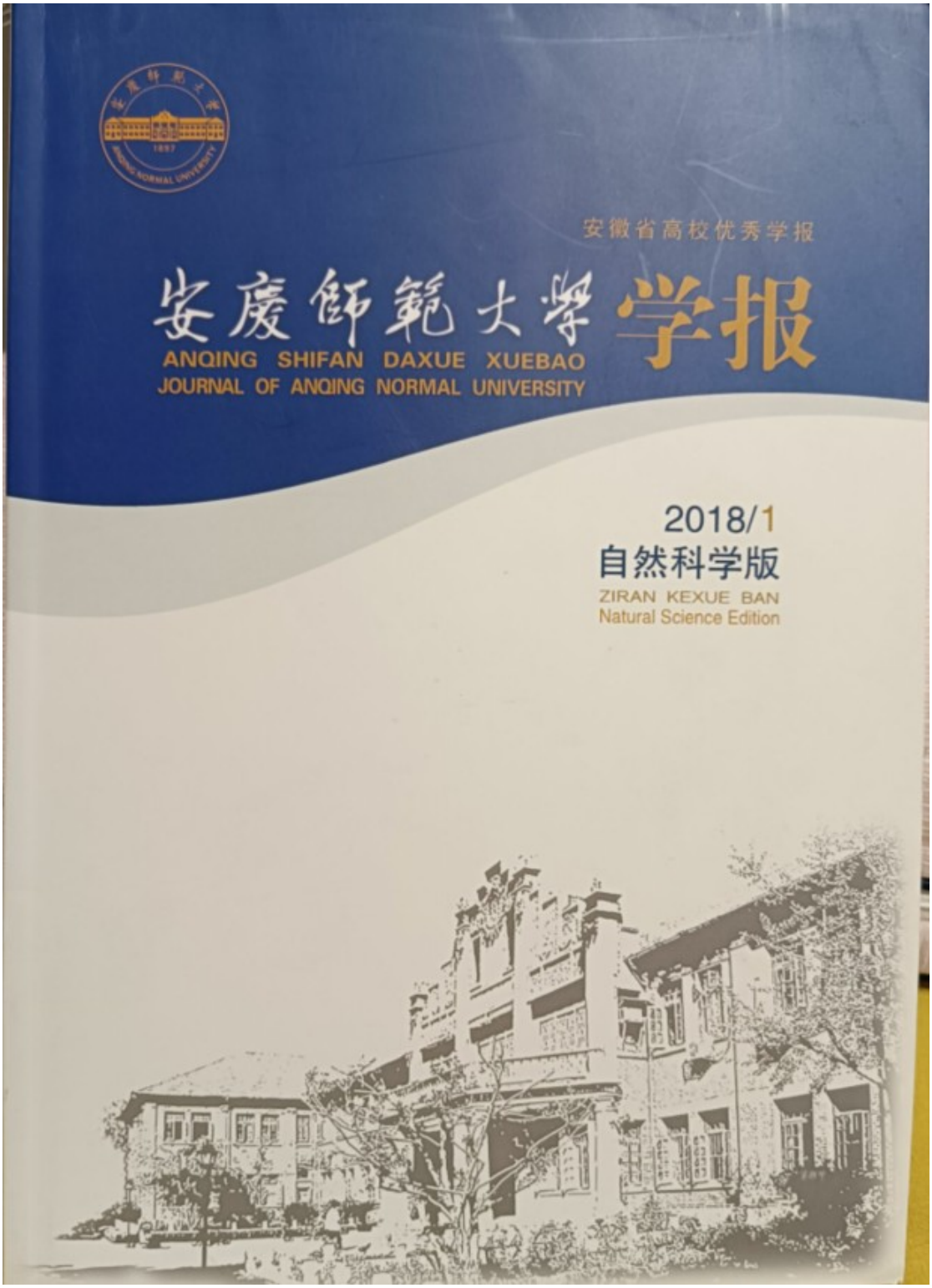
MiroSot(Micro Robot Soccer Tournament)是微型机器人足球赛的简称，是FIRA(Federation of International Robot-Soccer Association)世界杯比赛中的一个重要项目，FIRA MiroSot足球机器人系统分为视觉子系统、决策子系统、通信子系统和机器人小车子系统等四个子系统[1-3]。整个系统通过视觉子系统获取外界信息，视觉子系统的主要任务是使用高速摄像头实时采集比赛场中的图像，并对图像进行图像矫正、颜色分割、图像识别等过程，从而获得场地中目标对象(比赛双方的机器人和足球)的运动状态信息，进而提供给决策子系统进行分析和决策。在常规视觉子系统中采用的摄像头普遍存在采集图像速率低，畸变矫正效果不够明显，标定步骤繁杂，采集颜色易受噪声影响等问题，极大地影响了比赛效果。摄像机标定结果的精度关系到整个机器人视觉系统的精度[3]。Tsai[4]提出了两步法进行摄像机非线性标定，但存在标定精度受径向畸变模型的限制等问题。张征宇等[5]提出了一种基于共面条件的摄像机非线性畸变自校正方法，可以不使用标定板来完成自动矫正。针对以上问题，为了提高矫正

收稿日期：2018－01－08 * 通讯联系人

基金项目：安徽省重大教学改革研究项目(2016JYXM0907)；河海大学文天学院自然科学研究项目(WT15001)。

作者简介：林金珠(1981－)，女，河南信阳人，讲师，硕士。E－mail: tianwei5689@126.com

13 主持人,****,李赛红.基于二级 MS Office 的大学计算机基础课程教学[J].安庆师范大学学报(自然科学版),2018,24(01):123-125+128.



基于二级MS Office的大学计算机基础课程教学

林金珠,倪天伟,李赛红

(河海大学文天学院 电气信息工程系,安徽 马鞍山 243031)

摘　要:本文提出了基于二级MS Office的大学计算机基础教学改革,通过编写教材、编制教学案例资源库、建设多元化教学资源、优化课堂教学因素、完善考核方式等多种途径来推进改革进程,以适应高速发展的信息化社会对计算机基础人才的需求。实践证明,该教学改革取得了较好的教学效果,既提高了学生的课堂学习兴趣,又全面提高了学生的实际操作能力。

关键词:MS Office;大学计算机基础;教学改革　　DOI:10.13757/j.cnki.cn34-1328/n.2018.01.030

中图分类号:TP317.1-4;G642　　**文献标识码**:A　　**文章编号**:1007-4260(2018)01-0123-03

Teaching about College Computer Basis Course Based on MS Office Grade Two

LIN Jinzhu, NI Tianwei, LI Saihong

(Electrical and Information Engineering Department, Hohai University Wentian College, Ma'anshan 243031, China)

Abstract: This paper persents the teaching reform and practice of college computer basis course based on MS Office grade two. Through the preparation of teaching materials and case teaching resource database, construction of diversified teaching resources, optimization of classroom teaching, improvement of the examining mode and so on many kinds of ways to advance the reform process in order to adapt to the needs of rapid development information society for the computer basis talents. The practice shows that the teaching reform has achieved good teaching effects.It not only improves the students' interest in classroom learning, but also improves the students' practical operation ability.

Key words: MS Office; college computer basis; teaching reform

大学计算机基础是普通高校非计算机专业学生在大一上学期开设的一门公共基础必修课程,旨在通过学习计算机基础使学生了解计算机基础知识,熟练掌握计算机基本操作技能,提升面向计算思维的信息素养。目前,大学计算机基础课程存在教学内容涉及面广、学生的计算机水平参差不齐、课程学时不够等问题,传统的教学内容和教学方法已不适应高速发展的信息化社会对计算机基础人才的需求。故创新教学内容和方法,革新教学理念,转变教学模式,提高教学效率,努力寻找一条适合本校学生的计算机基础教学体系势在必行。二级MS Office高级应用是教育部考试中心在2013年下半年新增加的全国计算机等级考试科目,旨在选拔一批具有办公软件高级应用的人员。通过参加二级MS Office高级应用科目考试,不仅是为获得计算机二级资质,也是检验应试者是否掌握在实际办公环境中的应用操作能力[1]。通过分析全国计算机等级考试二级MS Office高级应用的考试大纲和考试内容,以及应用型本科院校教授大学计算机基础课程的课程目标,最后得出可以借助二级MS Office高级应用等级考试,对大学计算机基础课程的教

收稿日期:2016-08-01

基金项目:安徽省高等学校省级质量工程MOOC示范项目"信息技术应用"(2015mooc232)。

作者简介:林金珠,女,河南信阳人,硕士,河海大学文天学院电气信息工程系讲师,研究方向为计算机应用。

E-mail: 279767891@qq.com

14 主持人,**.“专创融合”视域下 C 语言程序设计课程教学实践探索[J].创新创业理论研究与实践,2024,7(16):11-13+37.**

The Theory And Practice Of Innovation And Entrepreneurship

“专创融合”视域下C语言程序设计课程教学实践探索

林金珠，倪天伟
（信阳学院，河南信阳 464000）

摘要：在当前高等教育中，“专创融合”成为培养高素质人才的关键。该文以 C 语言程序设计课程为例，探讨了在传授专业知识的同时融入创新创业训练的路径；通过设定专业课程与创新创业的双培养目标、构建 C 语言专业知识与创新创业知识映射关系、优化教学方法、进行创新创业实践并改革教学评价方法，旨在为 C 语言程序设计课程的教学改革提供新思路，培养具有创新精神和实践能力的高素质人才。

关键词：“专创融合”；C 语言程序设计；教学实践；创新创业；课程评价；课赛融合
中图分类号：TP312.1–4；G712.4　　**文献标识码**：A　　**文章编号**：2096–5206(2024)08(b)–0011–03

Exploration of Teaching Practice in C Language Programming Course under the Perspective of Professional and Innovation Integration

LIN Jinzhu, NI Tianwei
(Xinyang College, Xinyang Henan, 464000, China)

Abstract: In current higher education, the professional and innovation integration has become a key to cultivating high-quality talents. Taking the C language programming course as an example, this paper explores the path of integrating innovative entrepreneurship training while imparting professional knowledge. Through setting dual objectives for professional courses and innovation entrepreneurship, building a mapping relationship between C language professional knowsledge and innovative entrepreneurship knowledge, optimizing teaching methods, conducting innovative entrepreneurship practices, and reforming teaching evaluation methods, this paper aims to provide new ideas for the teaching reform of the C language programming course and contribute to cultivating high-quality talents with innovative spirit and practical abilities.

Key words: Professional and innovation integration; C language programming; Teaching practice; Innovation and entrepreneurship; Curriculum evaluation; Class-competition combination

根据《教育部办公厅关于做好深化创新创业教育改革示范高校 2019 年度建设工作的通知》，高校应积极推进专业课程与创新创业教育的有机融合[1]。在计算机科学领域，C 语言程序设计课程作为基石，对学生编程思维和问题解决能力的培养具有关键作用。然而，当前 C 语言程序设计课程教学面临理论与实践脱节、过度关注细节而忽略整体编程思维、实验环节和创新思维不足以及学生学习积极性不高等问题，限制了学生独立解决复杂编程问题能力的培养[2]。为了克服这些问题，需要增强 C 语言程序设计课程教学中的实践环节，调整教学内容和方法，将创新创业的理念融入其中，从而激发学生的学习热情，培养他们的研发能力、创新思维和创业精神，进一步丰富和深化高校的创新创业教育。

课题项目：信阳学院教学改革研究与实践项目“基于 ACM-ICPC 的 C 程序设计课程教学改革与实践”（2022YJG013）；河南省“专创融合”特色示范课程“C 语言程序设计”（教高〔2023〕72 号 –188）；河南省高等教育教学改革研究与实践项目（研究生教育类）“数智驱动下普通高校学士学位授予质量保障机制建设研究与实践”（2023SJGLX387Y）；2023 年河南省产教融合系列项目“以竞赛为载体的‘课赛融合’赋能多专业协同的创新实践教学模式研究”（教办高〔2024〕13 号 –198）。

作者简介：林金珠（1981—），女，河南信阳人，硕士研究生，副教授，研究方向：计算机应用技术、高等教育。

1 “专创融合”课程内涵

“专创融合”课程是一种创新教育模式，它将专业教育与创新创业教育有机结合，旨在培养学生的创新思维、创业能力和实践技能[3]。在“专创融合”课程中，学生不仅能够深入学习专业的核心知识，还能接触创新创业的理论和实践方法，如市场调研、商业计划书编写、团队管理等[4]。同时，课程还需设置实践项目或创业实验，让学生在实践中体验创新创业的过程，锻炼解决问题的能力和团队协作能力。

9 772096 520243
黑龙江格言杂志社有限公司　主办

15 主持人,****.吸引信阳籍在外人才回乡创新创业策略[J].农村经济与科技,2024,35(09):248-251.

吸引信阳籍在外人才回乡创新创业策略

林金珠，倪天伟

（信阳学院大数据与人工智能学院，河南　信阳　464000）

[摘要]在国家积极激励农民工、高校毕业生和退役军人回乡创业，推动乡村振兴的背景下，结合信阳市自身特点，分析了吸引信阳籍在外人才回乡创新创业存在的问题，提出了建设数字化平台；强化组织实施，提供优惠政策和支持；完善培训机制；多措并举吸引高校人才等策略，旨在有效推动新时期信阳籍在外人才回乡创业。

[关键词]回乡创新创业；乡村振兴；返乡农民工

[中图分类号]F323.6；F279.27　　[文献标识码]A

信阳市位于河南省最南部，东连安徽，南接湖北，为三省通衢，素有“江南北国，北国江南”之美誉，大量的人口常年外出务工，虽然给信阳市带来了一定的劳务收入，改善了当地居民的生活水平，但长远来看对信阳市的经济社会发展是不利的，诸多人才和劳动力的流出使本地的发展活力不足。如何鼓励信阳籍各类人才回乡创新创业，形成以创业带动就业、以就业促创业的良性互动新格局，深入了解目前信阳籍各类人才返乡回乡创新创业现状，透彻分析各类人才返乡回乡创新创业面临的困难并提出对应对策，具有重要的现实意义。

1　信阳籍在外人才回乡创新创业现状

随着乡村振兴战略的实施，越来越多的信阳籍在外人才为了实现自我梦想、抚养孩子、赡养父母、家庭团聚选择回乡创业，通过当前研究文献来分析，吸引在外人才回乡创新创业的策略主要从技术角度上，建立在外人才信息库、研发平台建设、人才创新创业超市建设等；从政策工具上，加强政策的按需引导和落实性；从创业风险、价值层次和心理作用上，人文关怀、开展亲子夏令营活动等。为了鼓励和吸引在外人才回乡创新创业，信阳市政府出台了一系列的扶持政策，如启动“鸿雁计划”

[收稿日期]2024-03-06

[基金项目]河南省“专创融合”特色示范课程“C语言程序设计”；2023年度河南省高等教育教学改革研究与实践项目（研究生教育类）（2023SJGLX387Y）；2024年度河南省高等教育教学改革研究与实践项目（本科教育类）（2024SJGLX0604）。

[作者简介]林金珠（1981—），女，河南信阳人，副教授，硕士，研究方向：创新创业和技术应用。

学徒班最先碰到的就是异地问题，在校学生无法去乡村常驻；学徒班开班之后，陆陆续续有学生想中途退出；学生获得业绩后的利益分配问题，等。

解决这类问题并没有统一的方法，需要具体问题具体分析。长征职业技术学院通过建设校内实训基地，解决了乡村振兴学徒班的异地问题；通过设立学徒班学员进出机制，让学生有中途退出的机会，同时也让企业方有中途增员的权限，有效解决了人员变动问题；开班前，长征职业技术学院与乡村政府、乡村电商、学员签订学徒班合作协议，将学员业绩收益明确列入协议，从而有效预防了经济利益纠纷。

4　结语

浙江作为共同富裕示范区，要消除城乡经济差异，实现乡村振兴，离不开电子商务这一新质生产力。目前乡村电商人才缺口巨大，浙江高职院校针对新时代新问题，从价值引领、案例导入、项目驱动三方面进行实践，发挥人才培养功能，为乡村振兴培养电商人才寻找新路径。

[参考文献]

[1] 习近平.扎实推动共同富裕[J].求是，2021(20).

[2] 新华社.中共中央国务院关于支持浙江高质量发展建设共同富裕示范区的意见[EB/OL].http://www.gov.cn/xinwen/2021-06/10/content_5616833.htm，2021-06-10.

[3] 中共中央国务院关于实施乡村振兴战略的意见[EB/OL]..https://www.gov.cn/zhengce/2018-02/04/content_5263807.htm?eqid=fe5c0d470000dd010000000264648fba，2018-02-04.

[4] 新华社.中共中央国务院关于做好2022年全面推进乡村振兴重点工作的意[EB/OL].https://www.gov.cn/zhengce/2022-02/22/content_5675035.htm?eqid=91b09ede0002a74200000003645 7b8ab，2022-02-22.

[5] 蔡建华，叶欣欣，张　静.农村电商创业人才孵化现状——基于中国首个淘宝村试点县的调查[J].中国市场，2021(14)：181-182.

[6] 宁晚枚，张雪玉.共同富裕背景下农村电商对农民增收的影响研究[J].商业经济研究，2023(4)：103-106.

[7] 浙江政府服务网.杭州聚力推进三个“一号工程”[EB/OL].https://www.hangzhou.gov.cn/art/2023/7/22/art_812262_59084950.html，2023-07-22.

[8] 石锦秀.乡村振兴背景下浙江农村电商创业环境提升策略研究[J].柳州职业技术学院学报，2021(2)：6-9.

[9] 颜　颖.乡村振兴背景下浙江农村电商高质量发展的新路径[J].太原城市职业技术学院学报，2023(6)：26-28.

[10] 张婷素，刘　浩.浙江县域农村电商发展研究[J].中国集体经济，2017(9)：27-28.

[11] 屠琦琼.直播电商助力乡村振兴的发展研究——以宁波地区为例[J].全国流通经济，2023(7)：36-39.

16 主持人,****.基于成果导向的 C 语言课程教学改革与实践[J].电子测试,2020,(14):112-114.

基于成果导向的 C 语言课程教学改革与实践

林金珠，倪天伟
（皖江工学院，安徽马鞍山，243031）

摘要：结合 C 语言教学现状及存在问题，本文运用成果导向教育理念，从学生预期学习成果出发，以任务为中心组织教学内容，创建开放式教学环境，优化教学因素，构建科学有效的评价体系等环节对 C 语言课程进行教学改革。实践证明，该教学改革有效得提高了 C 语言课程的教学质量，为本专业毕业要求的达成提供了支撑。

关键词：成果导向；工程教育；C 语言；教学改革
DOI:10.16520/j.cnki.1000-8519.2020.14.048

Teaching Reform and Practice of C language Course Based on Outcomes-based Education

Lin Jinzhu, Ni Tianwei
(Wanjiang University of Technology, Ma' anshan Anhui, 243031)

Abstract: Combined with the current situation and existing problems of C language teaching, this paper uses the concept of results oriented education, starting from the expected learning results of students, organizing teaching content with task as the center, creating an open teaching environment, optimizing teaching factors, and building a scientific and effective evaluation system to carry out teaching reform of C language course. The practice shows that the teaching reform effectively improves the teaching quality of C language course and provides support for the achievement of graduation requirements of this major.

Keywords: outcome-based education (OBE); engineering education; C language; teaching reform

0 引言

C 语言是计算机及相关专业的一门基础专业课程，在本科教学计划中占有重要地位。长期以来，该课程教学存在以教师和教材为中心，以教师的个性为驱动，单一评估方式的问题。部分高校教师缺乏社会和企业的实践经验，普遍存在关起门来做研究的现象，导致学生学完该课程达不到预期目标，为后续的核心专业课学习埋下隐患。成果导向教育（Outcome-Based Education, OBE），于 1981 年由 William Spady 率先提出，已成为美国、欧盟等国家教育改革的主流理念。该理念首先对课程教学有一套与学生毕业要求相符的预期学习成果，然后创造一切条件和机会激励所有学生完成这些成果，并对该成果进行专业评价。它关注于学生在学校里学到了什么，毕业时可以做什么，提高了学生的学习效果。基于成果导向的理念对改善当前 C 语言课程教学有很好的借鉴和现实意义。

1 基于成果导向的课程设计流图

基于成果导向的课程设计从需求开始，制定培养目标，构建为达成目标的毕业要求，再由毕业要求搭建课程体系[1]。教学是课程实施的主要形式，先明确该课程要取得的预期学习成果，通过优化教学大纲、教学内容、教学环境、教学方法以及其他教学因素来保证成果的实施，最后对学生的学习成果进行评价和总结，并持续改进课程教学质量。图 1 给出了基于成果导向的课程设计流程。

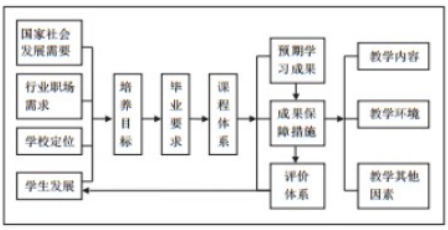


图 1 基于成果导向的课程设计流图

2 确定预期学习成果

毕业要求是对学生毕业时应该掌握的知识和能力的具体描述，包括学生通过本专业学习所掌握的知识、技能和素养[2]。预期学习成果反映了某门具体课程的教学内容对达成毕业要求的贡献情况。C 语言课程的预期学习成果的确定是建立在对计算机行业职场的需求、民办高校的定位、学生的个人发展特征等了解的基础上，通过总结用人单位岗位需求，调研往届毕业生工作现状，走访周边企业，访谈教授、学者后确定的。为了清晰的对预期学习成果进行分类，以加涅

基金项目：安徽省“六卓越一拔尖”卓越人才培养创新项目（2018zygc099）；皖江工学院校级质量工程项目（z12018005）。

17 主持人.独立学院图书馆纸质图书利用率提升策略研究[J].中国管理信息化,2018,21(15):156-158.

18 ****,主持人.专业认证背景下 Java Web 应用开发课程教学改革与创新研究[J].创新创业理论研究

独立学院图书馆纸质图书利用率提升策略研究

林金珠

(河海大学文天学院 电气系,安徽 马鞍山 243000)

[摘 要] 如何提升独立学院图书馆纸质图书利用率已迫在眉睫。根据独立学院图书馆纸质图书的利用现状,分析了影响纸质图书利用率的主要原因,并结合作者所在独立学院图书馆的实际调研结果,提出可以从加强专业人才队伍建设,围绕读者需求、优化馆藏结构,有效阅读引导和创设良好的阅读环境四个方面来提升独立学院图书馆纸质图书利用率。

[关键词] 独立学院图书馆;纸质图书;利用率;提升策略

doi: 10. 3969/j. issn. 1673 – 0194. 2018. 15. 063

[中图分类号] G250 [文献标识码] A [文章编号] 1673-0194(2018)15-0156-03

0 前 言

独立学院是指实施本科以上学历教育的普通高等学校与国家机构以外的社会组织或者个人合作,利用非国家财政性经费举办的实施本科学历教育的高等学校,它是民办高等教育的重要力量[1]。独立学院图书馆图书利用率是指一定时期内图书馆读者借阅的图书册数占馆藏图书册数的百分比,通过这个数据一方面可以折射出图书馆馆藏资源的水平和服务质量,另一方面也反映出一个学院整体的校风校貌[2]。针对当前纸质图书利用率持续低迷的现象,独立学院如何根据自身现状,做出针对性、及时性、有效性的提升策略已迫在眉睫。

1 独立学院图书馆纸质图书利用现状

近几年独立学院图书馆在努力地增加纸质图书馆藏的数量并提高其质量,但随着数字化阅读特别是手机阅读的冲击,纸质图书利用率大幅度降低,独立学院图书馆面临着巨大的挑战。陈燕指出独立学院馆藏纸质图书利用率仅只有全部馆藏的三分之一[3]。王海红认为大学生年人均图书借阅量呈逐年下降趋势,大学生偏爱文学类图书,特别是新生代作家的小说,甚至有学生在大学四年的图书总借阅量为零[4]。

为了了解本院学生对图书馆纸质图书利用情况以及对图书馆服务的满意度,笔者利用问卷星对在校学生分专业、分年级进行了调研。本次问卷共设计了 10 道题目,其中客观题 9 道,主观题 1 道,最后共回收有效问卷 155 份,其中有 55 名学生对主观题“对图书馆纸质文献资源建设和服务,请写出您的想法和建议”进行了回答。在对“您多久去一次图书馆借还纸质图书”的调研中,每周一次为 9.03%,每月一次为 6.45%,一学期一次为 11.45%,有需要就去 59.03%,从来没去过的为 14.03%。“您会主动阅读纸质图书吗”的调研中,自己主动阅读占 37.74%,在别人的推荐之下阅读占 16.77%,极少闲暇时会看书占 35.97%,几乎从不阅读占 9.52%。有 19.35%的同学对当前图书馆现有馆藏纸质图书种类不满意,有 52.9%的同学认为一般,只有 21.94%的同学认为比较满意,5.81%的同学认为满意。

2 独立学院图书馆纸质图书利用率的影响因素

2.1 纸质图书馆藏结构不科学

馆藏结构,是构成馆藏体系各部分相互结合的形式或构成形式,是图书馆藏建设的蓝图[5]。合理的馆藏结构,是图书馆有序建设的前提,也是后续馆藏资源补充得当的有力保证,它关系着全院师生文献信息的需求,是图书馆建设的生命线。独立学院图书馆馆藏资源建设得到学院领导的重视和支持,但苦于资金、时间、人才的缺失,一直存在各种问题。有时为了迎接图书馆的评估,或多或少存在以次充好,“应付”的现象,结果导致图书复本率过高,陈旧无价值的书充斥书架,这些严重破坏了馆藏资源结构良性发展,为图书馆后续的工作带来不良影响。

在对本院学生关于“您对当前图书馆纸质图书馆藏资源存在的主要问题”的调研中,61.94%的学生认为图书馆畅销书少,图书更新较慢,56.77%的学生认为图书馆图书种类不全,馆藏分布不合理,32.9%的学生认为各专业精深专著少。其中收集到的 55 个客观题的回答中就有 15 人次提出与图书馆纸质图书馆藏相关的问题,分别为独立学院图书馆纸质图书应提供一些最新的专业类书籍、多增加文学作品,纸质图书分类不合理、不便于查找,专业书籍过多,图书种类较少。调研的学生们并提出了建设性的意见,如图书馆纸质图书分类应该更简单明了,最好列成图表在图书馆显眼的地方进行张贴告示,增加图书馆检索机器以便搜索图书。

2.2 图书馆阅读环境有待改善

环境能影响人的行为,图书馆不仅仅是个物质空间,它还有意无意地影响着读者的情绪,影响纸质图书利用率。独立学院图书馆近几年在硬件的建设上发展迅猛,特别是在馆舍的建设和馆藏资源的购买上尤为突出。但由于馆员意识的薄弱,图书馆的软装建设还没有引起足够的重视,对读者服务的水平也较低,在

[收稿日期] 2018-05-31

[基金项目] 河海大学文天学院校级科研项目(WT17017);安徽省重大教学改革研究项目(2016jyxm0907)。

[作者简介] 林金珠(1981-),女,河南信阳人,讲师,硕士,主要研究方向:教育技术学,图书馆学。

与实践,2024,7(24):51-54.

专业认证背景下Java Web应用开发课程教学改革与创新研究

倪天伟，林金珠
（信阳学院，河南信阳 464000）

摘要：在工程教育专业认证的背景下，该文对Java Web应用开发课程的教学现状及存在的问题进行了分析，基于工程教育专业认证标准，对Java Web应用开发课程的教学内容进行了重新整合，并对教学模式、教学方法和考核方式等方面进行了改革实践。实践表明：课程改革贯彻“以学生为中心，以产出为导向，持续改进”的教学理念，可以充分调动学生的学习热情，提高教学效率，增强学生对复杂工程问题的解决能力。

关键词：专业认证；Java Web应用；成果导向教育；教学方法；教学改革；教学质量

中图分类号：TP393.09–4；G642　　**文献标识码**：A　　**文章编号**：2096–5206（2024）12（b）–0051–04

Research on Teaching Reform and Innovation of Java Web Application Development under the Background of Professional Certification

NI Tianwei, LIN Jinzhu
(Xinyang College, Xinyang Henan, 464000, China)

Abstract: This paper analyzes the teaching status and existing problems of the Java Web application development course in the context of engineering education professional certification. At the same time, based on the certification standards for engineering education, this paper has reorganized the teaching content of the Java Web application development course, and carried out reform practices in teaching modes, teaching methods, and educational evaluation. Practice has shown that implementing the teaching philosophy of “student-centered, output oriented, and continuous improvement” in curriculum reform can fully mobilize the learning enthusiasm of students, improve teaching efficiency, and enhance their ability to solve complex engineering problems.

Key words: Professional certification; Java Web application; Outcome-based education(OBE); Teaching methods; Teaching reform; Teaching quality

1 研究背景

我国工程教育专业认证于2006年启动，2016年6月，我国正式加入《华盛顿协议》[1]。我国现行的工程教育认证标准以《华盛顿协议》提出的毕业生素质要求为基础，经过10多年的实践，实现了从工程教育专业认证体系的建设到逐步开展专业认证、实现国际互认的目标。成果导向教育（Outcome-Based Education，OBE）是一种基于学习成果或者以结果为导向的教育理念，于1981年由Spady等人首次提出，并在美国、英国和加拿大等国家得到了广泛应用和推广，形成了较为完备的理论体系，是当前教育改革的主要指导思想[2-3]。《华盛顿协议》作为国际化程度最高、体系最完整的本科工程教育国际互认协议，已经全面采纳了成果导向教育的理念，并将其融入工程教育专业认证中。该协议的基本理念是以学生为中心，以学习成果为导向，不断改进教学质量[4-5]。教育部高等教育司于2023年7月转发了《关于发布已通过工程教育认证专业名单的通告》，要求各高校深入贯彻“学生中心、产出导向、持续改进”的理念，积极推进一流专业建设，提高教育质量，推动高等教育高质量发展。在工程教育专业认证的背景下，迫切需要对Java Web应用开发课程进行教学改革。本文通过分析Java Web应用开发课程的教学现状和存在问题，探讨了课程改革的模式和方向。

2 Java Web应用开发课程教学现状

Java Web应用开发是一门关键的应用导向型课程，有助于实现计算机专业的应用型定位目标，其教学内容涵盖了与互联网和计算机应用相关的架构、

基金项目：信阳学院教育教学改革研究项目资助“‘Π型人才’视角下的人工智能与大数据专业群建设创新与实践”（2023ZJG002）；河南省2023年度产教融合系列项目“以竞赛为载体的‘课赛融合’赋能多专业协同的创新实践教学模式研究”（教办高〔2024〕13号）；2023年度河南省高等教育教学改革研究与实践项目（研究生教育类）“数智驱动下普通高校学士学位授予质量保障机制建设研究与实践”（2023SJGLX387Y）；2024年度河南省高等教育教学改革研究与实践项目（本科教育类）“教育数字化背景下高校研究性教学模式研究与实践”（2024SJGLX0604）。

作者简介：倪天伟（1981—），男，河南信阳人，硕士研究生，副教授，研究方向：计算机软件、智能算法。

19 ****,主持人.应用型人才培养模式下 Java EE 平台课程教学改革与实践[J].电子测试,2021,(05):129-130.

应用型人才培养模式下 Java EE 平台课程教学改革与实践

倪天伟，林金珠
（皖江工学院，安徽马鞍山，243031）

摘要：应用型人才培养是国家针对应用型高校的教育发展要求，是应用型本科院校转型发展的必由之路。Java EE 平台课程是计算机科学与技术专业应用型人才培养的核心课程。本文分析了当前课程体系现状存在的问题，提出了 Java EE 平台课程建设改革的主要内容和方法。

关键词：应用型高校；人才培养；Java EE 平台；教学改革
DOI:10.16520/j.cnki.1000-8519.2021.05.051

Teaching Reform and Practice of Java EE Platform Course Based on Applied-talent Cultivation Mode

Ni Tianwei, Lin Jinzhu
(Wanjiang University of Technology, Maanshan Anhui,243031)

Abstract: The cultivation of application-oriented talents is the national education development requirements for application-oriented universities, and is the only way for the transformation and development of application-oriented universities. Java EE platform course is the core course of applied-talent cultivation in computer major. This paper analyzes the existing problems of the current curriculum system, and puts forward the main reform contents and methods of Java EE platform course curriculum.
Keywords: Application-oriented university; Talent cultivation; Java EE platform; Teaching reform

0 引言

随着互联网技术的飞速发展，计算机技术已经广泛应用到各行各业。近年来，国家提出“互联网+”战略，包括互联网+工业、互联网+教育、互联网+医疗健康等是重要应用形式。在新时代背景下，探索应用型高校人才培养模式已成为当务之急[1]。根据我院应用型人才培养的办学定位，计算机科学与技术专业培养目标是使学生获得专业基础与应用技能的基本训练，具有基础理论合格、实践技能良好、动手能力强的应用型工程技术人才。为落实本专业建设的定位与目标，我院计算机科学与技术专业在本专业课程体系中新建设了 Java EE 平台课程[2-3]。开展 Java EE 平台课程建设，将在创新性、应用型人才培养过程中发挥重要作用，学生的专业基础与专业技能与企业需求相衔接，本专业的教学定位、课程体系与工程单位需求相吻合。以应用能力培养为中心改革课程体系，以社会需求为立足点，以学生就业为导向，以实际应用为目标，推动教学内容改革，推动教学流程改革，推动教学方法改革，努力加强我院计算机科学与技术专业应用型课程建设的力度和深度。

1 Java EE 平台课程教学现状

Sun 公司推出的 Java EE(Java Platform Enterprise Edition) 平台经过多年的发展越来越趋于成熟和完善，并得到了广泛的应用。根据编程语言社区 TIOBE 在 2020 年出具的一份报告中，在当前二十多种流行的计算机语言中，Java 语言以 17.78% 的关注率，排名第一位。目前，我院计算机科学与技术专业开设《J2EE 架构与应用开发》课程，学分为 4 学分(64 学时)，含 1 学分实践。J2EE 平台课程学习到的核心技术很多，包括：Servlet、JSP、JavaBean、JDBC 技术等，还要学习一些主流框架 Struts2.0、Spring 、Springmvc、Mybatis 等，才能更好的开展 Java Web 开发。另外，作为 Java Web 开发人员，如果会一些 Web 前端技能就更加的得心应手，如：Vue. js 等。要满足以上课程内容的教学，任课教师普遍感觉任务重，压力大。同时，实践教学是本课程教学的重要组成部分，也是应用型人才培养的重要途径，在理论课程学习难度大任务紧的情况下，再加强实践教学，使得任课教师在课堂教学中显得捉襟见肘。因此，开展应用型人才培养模式下 Java EE 平台课程建设改革和建设综合性平台课程体系成为当务之急。

2 Java EE 平台课程教学改革

2.1 教学内容改革

(1)通过在《J2EE 架构与应用开发》课程中凝练出关键技术的知识点，增开部分选修课程。如，增开《网站设计与网页制作》、《JSP 程序设计》、《Oracle 数据库管理与维护》、《Web 前端技术应用实践》等课程，进一步使学生深入学习应用技术，同时，聘请企业高级讲师进校授课，深入开展校企合作。

基金项目：2020 年度安徽省高校优秀青年人才支持计划项目 (gxyq2020090)；2020 年度安徽省精品线下开放课程项目 (2020kfkc562)；安徽省“六卓越一拔尖”卓越人才培养创新项目(2018zygc099)。

20 **,主持人.疫情防控背景下高校毕业生就业工作探究[J].中国电力教育,2020,(08):69-71.**

DOI:10.19429/j.cnki.cn11-3776/g4.2020.08.030

疫情防控背景下高校毕业生就业工作探究

文 / 皖江工学院　倪天伟　林金珠

受新型冠状病毒（COVID-19）肺炎疫情的影响，2020 年高校毕业生的就业工作面临着前所未有的挑战，在疫情防控背景下如何妥善解决高校毕业生的就业成为当务之急。本文探讨了疫情防控背景下高校毕业生就业现状，分析了此次疫情对高校毕业生就业带来的影响和挑战，并提出了有针对性的促进大学生就业的若干指导策略和方法，重在提前谋划，提早介入，确保毕业生顺利就业、高质量就业。

高校毕业生就业创业工作是一项重要的民生工程，也是一项系统工程，既关系到人民群众的切身利益，又对维持社会和谐稳定发挥重要作用，党中央、国务院予以高度重视。2020 年春节前后，一场突如其来的“新型冠状病毒感染的肺炎”疫情在我国全面暴发，并很快蔓延至全国各地。这场疫情既影响了人们的正常工作生活，也给高校毕业生就业带来了前所未有的压力，招聘单位数量减少、岗位缩减等问题，使得高校毕业生的就业形势更加复杂严峻。3 月 18 日，全国政协教科卫体委员会召开“重大疫情下高校毕业生就业创业问题”专题座谈会，针对当前疫情对高校毕业生就业创业工作产生的影响，就如何解决高校毕业生就业创业面临的困难和问题提出意见建议。

当代大学生就业状况分析

近年来，随着高校扩招，在提升了众多国民整体素质的同时，也使得高校毕业生数量在逐年增加。当代大学生就业现状主要体现在以下几个方面。

高校毕业生的总体供给和社会需求的结构性矛盾逐渐凸显。据统计，2020 年高校毕业生规模达到了 874 万人，增量、增幅均为近年之最。再加上疫情影响，导致普遍存在着就业单位提供的就业岗位与毕业生数量之间的不匹配现象，使得大学生就业压力大、就业难等问题日渐突出。

高精尖技术人才缺口大。随着我国高质量发展战略的推进，大多数企业对高精尖技术人才的需求如饥似渴，这迫使地方本科高校转型发展，进一步推动高校加强学科建设与专业建设，加强“校企合作”的深度和广度。同时，普通高校要改变传统的教学方式，努力提高当代大学生的专业技能水平和创新能力，实现人才培养与国家经济发展、企业需求“无缝对接”。

当代大学生自身的就业观念有待转变。目前大学毕业生多是“90 后”，他们多数家庭背景好，、生活条件好，有的认为工作太累，有的认为工资太低，还有的认为不符合自己的专业选择，导致错失很多工作机会。所以，高校教师要大力加强大学生的就业观教育，引导大学生树立正确的、科学的就业观。

疫情下高校毕业生就业面临的困难与挑战

受疫情影响，今年高校毕业生就业形势越发严峻，人们不禁会思考：疫情下 2020 届毕业生如何充分就业？皖江工学院通过互联网网络平台向毕业生发放“毕业生就业情况调查问卷”调研结果如下：

73.46% 同学担心迟迟不能复学返校会影响毕业设计、论文撰写、毕业答辩质量，40.5% 的同学担心自己的补考、重修、毕业资格审核等工作。学生原定计划于 2 月初返校，因受疫情的影响，毕业生 5 月中旬才申请陆续返校。

21 叶长亮,主持人,李世强,等.基于 RFID 的教室考勤系统设计与实现[J].电脑编程技巧与维护,2017,(08):36-37+57.

22 主持人,**.教育数字化背景下高校研究性教学研究[J].湖北开放职业学院学报,2025,38(12):139-**

基于 RFID 的教室考勤系统设计与实现

叶长亮，林金珠，李世强，孟李杨
（河海大学文天学院，安徽 马鞍山 243031）

摘 要：智能考勤系统由实时点名、考勤信息查询、RFID 读写器等模块组成，通过系统设计实现考勤的智能化和信息化。

关键词：考勤系统；射频识别；定位

DOI:10.16184/j.cnki.comprg.2017.08.012

1 概述

RFID 在当今及未来有着巨大的发展前景，随着当今社会由信息时代向智能时代的转换，射频识别作为新一代的识别技术在各行各业将会得到充分的发展。本系统能够实现点名及存储自动化考勤。

点名及存储是智能考勤系统的核心，当被管控的人员进入特定区域，射频阅读器或射频激励器自动检测人员所携带的 ID 卡的信息，并将信息上传至上位机。

2 需求

由于传统课堂教学的考勤方法效率难以满足实际情况，特提出基于 RFID 射频信号到达时间来确定位置考勤系统：将多个相对位置已知的阅读器放在教室的固定位置，通过阅读器发射的电磁波及接收的电磁波的信号到达时间来判断 ID 卡阅读器的相对位置，再通过几何关系判断出 ID 卡的位置。

3 系统概述

基于 RFID 的教师考勤系统主要分为点名模块，上位机模块，存储模块等。

3.1 设计方案

学生们持卡进入教室坐定后，先将自己的一卡通放置于桌面的规定的特定读卡区内，待阅读器识别到此卡后会将该卡的信息通过串口发送至上位机，同时将信息通过显示屏显示出来。

在这个工作流程中要进行以下几个主要模块的设计与编写，其中数据由阅读器测量；数据显示由显示屏显示；数据传输模块由 usb 转 ttl 转换电路组成，该模块将学生信息从下位机传输给上位机。如图 1 所示。

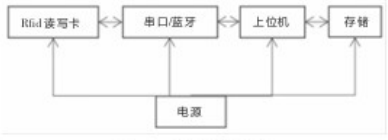


图 1 总体设计框图

3.2 电路

通过 LM1117 产生 5V 及 3.3V 为本系统的各个模块进行供电，为了减小体积均采用了体积小性能好的贴片封装。

采用 MFRC522 作为读写卡芯片。其支持 ISO14443A 的多层应用。其内部发送器在无需其他电路的情况下即可驱动阅读器与标签的通信。

总体设计原理图如图 2。

图 2 总体设计原理图

3.3 功能设定

（1） 系统管理员可以删除添加使用射频卡的用户。

（2）系统管理员可以查看任何时段的学生考勤情况。

基金项目：安徽省重大教学改革研究项目（2016jyxm0907）。

作者简介：叶长亮（1995–），男，本科，研究方向：自动化；林金珠（1981–），女，讲师，硕士，研究方向：计算机应用技术及计算机教育；李世强（1995–），男，本科，研究方向：通信工程；孟李杨（1995–），男，本科，研究方向：自动化。

收稿日期：2017–01–24

141.

ISSN 2096-711X

JHOVC

湖北开放职业学院学报

JOURNAL OF HUBEI OPEN VOCATIONAL COLLEGE

RCCSE中国高职高专成高院校学报类核心期刊
中国职业高等院校期刊AMI综合评价职院刊入库期刊

《中国知网》全文收录期刊
《国家哲学社会科学学术期刊数据库》全文收录期刊
《万方数据——数字化期刊》全文收录期刊
《中国期刊网数据库》全文收录期刊
《中国核心期刊（遴选）数据库》全文收录期刊
《维普资讯网》全文收录期刊
《长江文库》全文收录期刊
EBSCO全文收录期刊

2025.12
第38卷 第12期 总第394期
半月刊 六月下

中国·武汉
Wuhan China

教育数字化背景下高校研究性教学研究

林金珠,倪天伟
(信阳学院,河南信阳 464000)

[摘 要]针对高校研究性教学在实施过程中存在的智能工具与资源整合的困境、个性化教学路径的设计难题以及评价体系的单一性与片面性,结合教育数字化的发展趋势,构建了高校研究性教学的全景技术图谱。基于该图谱,开发了数字创新平台,并建立了与之配套的评价体系,旨在引导教师有效整合和运用数字化技术,突破教学瓶颈,提升研究性教学质量。研究不仅为高校研究性教学的数字化转型提供了理论支持,还通过实践探索为教学改革提供了可操作的参考路径,对推动创新型人才培养具有重要意义。

[关键词]教育数字化;研究性教学;数字化平台;创新与实践
[中图分类号]G640 [文献标识码]A [文章编号]2096-711X(2025)12-0139-03
doi:10.3969/j.issn.2096-711X.2025.12.048 [本刊网址]http://www.hbxb.net

研究性教学是培养学生创新思维、实践能力和批判性思维的重要途径,数字化赋能为其带来了新的发展机遇。近年来,国家对教育数字化转型的高度重视以及政策支持,为高校研究性教学的创新与实践创造了良好的环境。如《中国教育现代化2035》明确提出要推行启发式、探究式、参与式、合作式等多样化的教学方式,要"加快信息化时代教育变革,利用现代技术加快推动人才培养模式改革"。同时,随着2022年全国教育工作会议提出"实施教育数字化战略行动",教育领域正迎来全面数字化转型的崭新阶段。有学者指出,教育数字化背景将带来教学范式和教育模式的颠覆性创新。未来,借助智能技术手段,构建更加完善的研究性教学模式,将成为信息化教学的核心工作。研究性教学强调在教师的精心指导下,学生们利用丰富的辅助资源,积极主动地探索自然科学、社会现象以及日常生活中的各种专题,进行深入的思考和实践。然而,在教育数字化浪潮的推动下,尽管研究性教学在理论上具有显著优势,但在实际应用中仍面临诸多挑战。因此,探究一条在教育数字化背景下的高校研究性教学道路,对提高教学质量和培养创新人才具有重要意义。

一、教育数字化背景下高校研究性教学面临的挑战

在数字化浪潮的推动下,高校为培养高质量人才,持续推进研究性教学改革。然而,随着改革的深入,一些问题也逐渐浮出水面,主要体现在以下几个方面。

(一)技术整合与引导能力提升挑战

其一,数字化工具与资源整合能力不足。随着教育数字化的发展,大量智能工具和资源(如在线学习平台、虚拟实验室、数据分析工具等)被引入教学,但许多教师缺乏对这些工具的深入了解和有效整合能力。例如,一些教师可能知道如何使用学习通发布作业,但并未充分利用其讨论区功能来促进学生之间的互动;或者在使用虚拟实验室时,仅仅将其作为演示工具,而未能设计出让学生自主探究的实验任务。这种能力不足导致工具的使用流于形式,未能充分发挥其在研究性教学中的潜力。此外,教师在选择工具时也面临困难。如何根据学科特点和学生需求选择合适的平台或软件,这需要教师具备一定的技术评估能力和教学设计能力。其二,对学生自主学习能力的引导不足。研究性教学强调学生的自主学习能力,但许多教师缺乏有效的引导方法。例如,在布置探究任务时,教师可能只给出任务要求,而未提供明确的学习路径和方法指导,导致学生在面对复杂的任务时感到迷茫。此外,教师对学生学习过程的监督和反馈不足,例如未能及时发现学生在探究过程中遇到的困难,或者未能提供针对性的建议。这种引导不足容易导致学生在研究性学习中陷入低效或无效的状态,影响学习效果。

(二)个性化教学路径难实施

1.资源限制与个性化实施的矛盾

当前,市场上虽然充斥着丰富的多媒体教学资源、在线学习平台和学习视频,但这些资源普遍呈现出分散性、非系统性和无序性的特点。例如,学生在学习C语言程序设计课程时,往往需要跨学科的知识支持,如数据结构、算法分析与设计、全国计算机等级考试、学科竞赛以及创新创业等相关资料和信息。然而,目前并没有一个平台能够系统化地整合这些资源,导致学生在学习过程中难以高效获取所需内容。此外,在研究性项目的实施过程中,学生常常需要参考经典案例、研究性报告以及配套的情境资源,但这些资源往往难以批量获取,进一步增加了学习与研究的难度。甚至有些在线学习平台和视频内容缺乏权威性和专业性,甚至存在错误信息,导致学生在学习过程中容易产生误解或形成错误的知识体系。

2.项目难度与教学进度的平衡困境

项目难度与教学进度的平衡是个性化教学路径设计中的一大难题。难度设置过高,可能导致学生跟不上进度,产生挫败感;难度过低,则可能让学生觉得乏味,缺乏挑战性。比如,在研究性项目选择中,如果一开始就引入复杂的概念和知识,学生可能会感到困惑;但如果一直停留在基础层面,又无法提升学生的能力。但目前的参考书和教材存在明显不足,满足不用教师和学生的需求。一方面,书中提供的案例往往过于简单,仅局限于本章节的知识点,缺乏与前面所

收稿日期:2025-3-14
基金项目:本文系河南省高等教育教学改革研究与实践项目(本科教育)"教育数字化背景下高校研究性教学模式研究与实践"阶段性成果(项目编号:2024SJGLX0604);河南省高等教育教学改革研究与实践项目(研究生教育)"数智驱动下普通高校学士学位授予质量保障机制建设研究与实践"(项目编号:2023SJGLX387Y);河南省首批"专创融合"特色示范课程"C语言程序设计"(项目编号:教高[2023]72号-188);河南省2023年度产教融合系列项目"以竞赛为载体的'课赛融合'赋能多专业协同的创新实践教学模式研究"(项目编号:教办高[2024]13号)。
作者简介:林金珠(1981—),女,河南信阳人,信阳学院副教授,主要从事数字化教学、高等教育、计算机应用研究。

23 ****,主持人.“Π 型人才”视角下的人工智能与大数据专业群建设研究与实践[J].创新创业理论研究与实践,2025,8(06):67-69+117.

24 主持人,****.数智驱动下普通高校学士学位授予质量保障机制研究[J].河南教育(高教),2025,

“Π 型人才”视角下的人工智能与大数据专业群建设研究与实践

倪天伟,林金珠
(信阳学院,河南信阳 464000)

摘要:随着科技的不断发展,人工智能已成为世界科技领域的热点。在此背景下,培养具备“Π 型人才”特质的人工智能与大数据专业人才显得尤为重要。该文以信阳学院人工智能与大数据专业群建设为例,提出了一套创新的专业群建设策略,涉及培养模式创新、课程体系优化、教学模式变革、社会实践教育平台建设和产学研融合发展等方面。尤其是在课程体系建设上,注重强调理论与实际、研究与应用、学科深度和跨学科高度的融合,提高本科生的创新性思维和处理复杂问题的能力。

关键词:“Π 型人才”;人工智能;大数据;专业群建设;创新与实践;人才培养

中图分类号:G712　　**文献标识码**:A　　**文章编号**:2096-5206(2025)03(b)-0067-03

Research and Practice on the Construction of Artificial Intelligence and Big Data Professional Groups from the Perspective of “π – Type Talents”

NI Tianwei, LIN Jinzhu
(Xinyang College, Xinyang Henan, 464000, China)

Abstract: With the continuous development of technology, artificial intelligence has become a hot topic in the world technology field. In this context, it is particularly important to cultivate artificial intelligence and big data professionals with the characteristics of “π-type talents”. Taking the construction of the artificial intelligence and big data professional group at Xinyang College as an example, this paper proposes an innovative strategy for the construction of professional groups, which involves innovative training modes, optimization of curriculum systems, reform of teaching modes, construction of social practice education platforms, and integrated development of industry, academia, and research. Especially in the construction of the curriculum system, emphasis is placed on the integration of theory and practice, research and application, disciplinary depth, and interdisciplinary height, in order to improve undergraduate students' innovative thinking and ability to handle complex problems.

Key words: “π-type talents”; Artificial intelligence; Big data; Professional group construction; Innovation and practice; Personnel training

当今信息化和智能化快速发展,人工智能和大数据已经成为推动社会经济发展的核心技术。随之而来的是对创新型、复合型人才的需求日益迫切,尤其是在人工智能与大数据领域[1]。在这种市场需求的推动下,教育部门和高等院校面临着如何培养适应未来技术变革和社会需求的创新型人才的挑战。“Π 型人才”概念的提出,正是为了描述和定义新时代需求下的创新型、复合型人才。信阳学院大数据与人工智能学院(以下简称学院)充分认识到了在人工智能与大数据领域培养“Π 型人才”的重要性,现开设计算机科学与技术、物联网工程、数据科学与大数据技术、人工智能 4 门本科专业,形成了人工智能与大数据专业群。学院自 2018 年创办以来,积极探索如何通过专业群建设培养符合未来市场需求的人工智能与大数据人才。本文针对“Π 型人才”视角下的人工智能与大数据专业群建设进行深入探讨,旨在通过创新教育模式,培养既具备专业深度又具有跨学科广度,以及能运用大数据及人工智能技术服务集“云物大智”(云计算、物联网、大数据、智能化)技术的新兴产业,适应未来社会发展需求的创新型、复合型人才[2]。

基金项目:信阳学院教育教学改革研究项目资助“‘Π 型人才’视角下的人工智能与大数据专业群建设创新与实践”(2023ZJG002);河南省 2023 年度产教融合系列项目“以竞赛为载体的‘课赛融合’赋能多专业协同的创新实践教学模式研究”(教办高〔2024〕13 号);2023 年度河南省高等教育教学改革研究与实践项目(研究生教育类)“数智驱动下普通高校学士学位授予质量保障机制建设研究与实践”(2023SJGLX387Y);2024 年度河南省高等教育教学改革研究与实践项目(本科教育类)“教育数字化背景下高校研究性教学模式研究与实践”(2024SJGLX0604)。

作者简介:倪天伟(1981—),男,河南信阳人,硕士研究生,副教授,研究方向:计算机软件、智能算法。

1“Π 型人才”的内涵及其重要性

“Π 型人才”是指具备专业技能和跨学科知识,同时拥有创新能力和实践能力的人才。他们不仅具备深厚的专业知识,还具备广泛的跨学科知识,能够在不同领域进行创新和实践[3-4]。自“Π 型人才”概

数智驱动下普通高校学士学位授予质量保障机制研究

◆ 林金珠，倪天伟

（信阳学院 大数据与人工智能学院，河南 信阳 464000）

摘 要：当前，我国普通高校学士学位授予存在一些问题，对此，应采取构建数智驱动下普通高校学士学位授予质量保障机制、打造普通高校学士学位授予数字化开发平台、强化普通高校学士学位授予过程的预防性措施、健全普通高校学士学位授予质量综合评价体系等解决对策。

关键词：数智技术；学士学位授予；数字化平台；高质量就业；普通高校

一、我国普通高校学士学位授予存在的问题分析

（一）缺乏系统的学士学位授予质量保障体系

当前，我国许多高校在学士学位授予过程中面临一系列显著问题，其中最为突出的是缺乏系统的质量保障体系，这也直接导致学位授予质量参差不齐。学士学位的授予并非仅仅基于学生在最后一学期的表现，而是一个贯穿本科教育四年的长期且复杂的过程，这一过程涉及入学招生、培养方案制订、教学大纲设置、课程与教学内容安排、教学资源配置、教学策略选择、学业考核形式、毕业论文（设计）管理、实验与实践环节实施、学业反馈与辅导提供，以及个性化教学、教学互动与反馈、学术不端行为监管等多个关键环节和因素。然而实际情况是，由于系统的质量保障体系缺失，这些关键环节和因素往往未得到充分的重视和有效的管理。特别是学生学业数据的收集、分析和反馈等环节，未能形成一个完整的闭环系统，导致学校和教师难以全面而准确地掌握学生的学习情况和存在的问题，也无法及时且有效地提供有针对性的辅导和帮助。因此，学位授予质量难以得到有效保障。

（二）缺乏完善的学位授予管理信息化平台

信息化平台的缺乏使得学位授予流程变得烦琐且易出错，传统的纸质操作方式还极大地增加了学生和教职工的工作负担。学生需要填写大量表格、提交多份材料，并经历多重审核签字流程，效率低下且容易出错。同时，纸质文档的存储和管理也面临诸多问题，如易丢失、查找困难等。此外，信息化平台的缺乏还影响了学位授予信息的共享与查询。不同部门间的数据未能实时同步，形成信息孤岛，导致查询进度和获取通知变得复杂且耗时。这不仅增加了沟通成本，还可能因信息传递不及时而引发误解。同时，缺乏信息化平台还影响了学位授予的透明度和公正性。决策过程和数据难以得到有效监控和审计，容易引发对公正性的质疑。学生和教职工也无法及时了解最新政策和要求，造成信息不对称和导致误解。

（三）学位授予过程缺乏预防性措施

由于缺乏数字化技术的应用，学位授予过程往往缺乏必要的预防性措施。例如，对于学业成绩不理想、心理健康状况不稳定的学生，高校往往难以及时发现并采取有效措施进行干预。这不仅影响了学生的个人成长和发展，也降低了学位授予的整体质量。

（四）评价机制规范性有待提高

传统的人工审核方式在面对庞大的毕业生群体时显得力不从心，不仅效率低下，而且容易出错。同时，由于缺乏全面的数据支持和智能分析，学校在评估学生学业成果时往往只能依赖有限的考试成绩和简单的学分要求，这无疑忽视了学生在创新能力、实践能力以及综合素质等方面的表现。这种评价体系的不完善，不仅损害了学士学位授予的公正性和准确性，还可能导致一些真正具备优秀能力和潜力的学生因考试成绩不理想或学分不足而无法获得应有的学位认可，一些在学业上表现平平但在其他方面具有显著优势的学生可能被埋没。

基金项目：2023 年度河南省高等教育教学改革研究与实践项目（研究生教育类）“数智驱动下普通高校学士学位授予质量保障机制建设研究与实践”（编号：2023SJGLX387Y）；2024 年度河南省高等教育教学改革研究与实践项目（本科教育类）“教育数字化背景下高校研究性教学模式研究与实践”（编号：2024SJGLX0604）；河南省“专创融合”特色示范课程“C 语言程序设计”；2024 年度信阳学院大学生校级科研项目“高校学生学业预警系统的设计与实现”

作者简介：林金珠（1981— ），女，信阳学院大数据与人工智能学院副教授，研究方向为数字化教学、计算机应用、高质量就业；倪天伟（1981— ），男，信阳学院大数据与人工智能学院副教授，研究方向为人工智能、计算机应用。

(03):57-59.

25 主持人,王科翰.人工智能背景下地方本科高校毕业生高质量就业研究[J].湖北开放职业学院学报,2025,38(21):161-163.

第38卷第21期 总第403期 湖北开放职业学院学报 Vol. 38. No. 21(Gen. No. 403)
2025年11月(上) Journal of HUBEI Open Vocational College Nov. (first half) 2025

人工智能背景下地方本科高校毕业生高质量就业研究

林金珠,王科翰
(信阳学院,河南信阳 464000)

[摘 要] 在人工智能技术深刻重塑就业市场的背景下,地方本科高校毕业生面临技能错配、竞争力不足等问题。本研究基于社会、教育和个人三维视角,分析人工智能对毕业生就业的影响,通过构建就业保障机制、智能化就业匹配系统、就业全程跟踪系统、优化高校动态育人机制,并强化学生职业规划与综合能力提升路径。研究成果为地方本科高校优化人才培养模式、实现毕业生高质量就业提供理论与实践参考,助力区域经济社会可持续发展。

[关键词] 人工智能;地方本科高校;高质量就业;产教融合;智能化匹配
[中图分类号] G647.38;TP18;G648.4 [文献标识码] A [文章编号] 2096-711X(2025)21-0161-03
doi:10.3969/j.issn.2096-711X.2025.21.054 [本刊网址] http://www.hbxb.net

就业是最大的民生。党的二十大报告将"实施就业优先战略"摆在突出位置,强调健全就业促进机制,促进高质量充分就业,彰显了国家对就业工作的高度重视。教育部也在《做好2024届全国普通高校毕业生就业创业工作的通知》中明确指出,要把高校毕业生就业工作摆在更加突出的位置,确保毕业生顺利就业。地方本科高校作为培养应用型人才的重要基地,肩负着为区域经济社会发展输送高素质人才的重要使命。然而,在人工智能技术快速渗透的背景下,传统就业模式受到冲击,部分毕业生面临技能错配、就业竞争力不足等问题,高质量就业的实现面临诸多挑战。因此,探索建立适应人工智能时代的地方本科高校毕业生高质量就业保障机制,不仅关乎毕业生个人价值的实现与职业发展,更对优化劳动力资源配置、推动经济社会可持续发展以及维护社会稳定具有深远意义。

一、人工智能对地方本科高校毕业生就业的影响机制

为进一步优化高校毕业生就业服务工作,当前虽已通过多维度政策协同与资源整合取得阶段性成效,但随着人工智能技术的迭代升级,一些新问题也逐渐浮出水面,主要体现在以下几个方面。

(一)社会因素:技术变革与就业市场冲击

人工智能技术的广泛应用对就业市场产生了深远影响。据《全球发展报告》预估,2020年至2025年间,全球约8500万工作岗位可能被人工智能替代。这种技术变革不仅导致就业替代效应持续显现,规模性失业风险加剧,还引发了就业结构的"极化效应",即中间能力层的岗位更容易被替代,而中等技能劳动者被迫向高技能或低技能岗位转移。与此同时,重点企业用工规模出现萎缩,进一步加剧了就业市场的供需失衡。随着人工智能技术的普及,许多传统岗位的工作内容发生了根本性变化,要求从业者具备更高的技术能力和创新能力,这种变化使得原本依赖传统技能的劳动者面临失业风险,而高技能人才的需求则大幅增加。

(二)教育因素:人才培养与产业需求脱节

从供需结构来看,人才培养与新兴产业需求之间的脱节问题日益突出。在智能时代,企业对人才的需求从单一的专业技术能力转向具备跨界思维、创新能力和持续学习能力的复合型人才。然而,当前高校人才培养模式与快速发展的产业需求存在明显错位,导致毕业生数字经济就业胜任力不足,难以满足市场对高质量就业的要求。高校人才培养与产业需求脱节的原因主要有以下几个方面:其一,高校课程设置滞后于产业发展。许多地方本科高校的课程设置,未能及时更新以适应新兴技术和产业的需求。例如,人工智能、大数据等新兴领域的课程设置不足,或者教学内容过于理论化,缺乏实践应用,导致毕业生在这些领域的知识和技能储备不足。其二,高校实践教学环节薄弱。地方本科高校往往在实践教学环节上投入不足,学生缺乏实际操作和项目经验,缺乏与企业实际需求的对接,导致学生在校期间难以获得实际工作经验。其三,高校教师队伍的结构和能力也存在问题。地方本科高校的师资力量相对薄弱,尤其是在新兴技术领域,缺乏具有行业经验的教师,更难以将最新的产业动态和技术趋势融入教学中。最后,高校与企业之间的合作机制不完善。虽然许多高校与企业建立了合作关系,但这些合作往往停留在表面层次,缺乏深度和广度,难以真正实现人才培养与产业需求的对接。

(三)个人因素:观念转变与能力短板制约

从毕业生主体来看,地方本科高校学生的就业观念、心理状态及自身能力特点共同影响了其就业竞争力。其一,就业观念趋向保守,表现为更倾向于升学、体制内就业、一、二线城市就业以及"慢就业",这反映了学生对就业市场不确定性的应对策略,同时也凸显了其在人工智能技术威胁和就业竞争压力下的脆弱性。其二,后疫情时代毕业生的就业心理普遍呈现出焦虑、抑郁、从众和功利等特征,进一步加剧了其在就业选择中的被动性。最后,学生自身能力存在短板,入学时基础较为薄弱,学习能力和综合素质相对不足,难以在短时间内提升自身能力以应对考研、考编或创新创业。这些特点共同制约了地方本科高校毕业生的就业竞争力,使其在激烈的就业市场中处于劣势。

收稿日期:2025-4-29
基金项目:本文系2025年河南省软科学项目"人工智能背景下地方本科高校毕业生高质量就业保障机制研究"(项目编号:252400410603);2025年河南省哲学社会科学教育强省研究重点项目"人工智能背景下民办高校毕业生高质量就业提升策略研究"(项目编号:2025JYQS0080)。
作者简介:林金珠(1981—),女,河南信阳人,信阳学院大数据与人工智能学院副教授,研究方向:数字化教育、创新创业。

26 黄莉,主持人.老年人健康饮食管理系统设计与实现[J].电脑编程技巧与维护,2025,(11):103-

105+161.

老年人健康饮食管理系统设计与实现

黄莉[1]，林金珠[2]
（1. 信阳航空职业学院文化教育学院，河南 信阳 464100；2. 信阳学院大数据与人工智能学院，河南 信阳 464000）

摘　要：针对老年人饮食健康管理需求，设计并实现了一套智能化健康饮食管理系统。系统以老年群体慢性病管理与营养均衡为核心，采用模块化架构集成用户档案管理、食谱分类推荐、周计划定制、健康论坛互动等功能模块。实际应用表明，该系统可显著提升老年人饮食科学性水平，降低慢性病管理难度，为养老机构提供了数字化解决方案，兼具实用性与可推广性。

关键词：老年人健康饮食；智能管理系统；食谱分类推荐；个性化饮食规划
DOI:10.16184/j.cnki.comprg.2025.11.026

1　概述

随着全球人口老龄化加剧，老年群体的健康问题日益受到关注。据世界卫生组织统计，60岁以上人群中，很多患有慢性疾病，例如，高血压、糖尿病、心血管疾病等，其中，不科学饮食是主要诱因之一。然而，老年人普遍面临饮食知识匮乏、营养摄入不均衡、健康管理意识薄弱等问题，加之传统饮食指导方式，例如，纸质手册、线下咨询，存在信息分散、更新滞后、缺乏个性化等缺陷，因此急需一种智能化、系统化的解决方案。此外，数字化技术的普及为老年人健康管理提供了新方向[1]，但现有健康类应用多针对年轻群体，功能复杂、交互门槛高，难以满足老年人的使用习惯与特殊需求，例如，大字体、简易操作、疾病适配性。因此，开发一款专为老年人设计的健康饮食管理系统具有重要的社会价值与实践意义。

2　技术支撑

系统采用浏览器/服务器（B/S）架构，充分利用Java Web技术栈的优势，构建了一个层次分明、易于维护和扩展的Web应用程序。前端展示层是用户与系统交互的窗口，通过JSP技术结合超文本标记语言（HTML）和层叠样式表（CSS），构建出既动态又美观的网页界面。原生JavaScript的引入，实现了界面的基础交互功能，例如，表单验证、动态内容更新等，避免了引入复杂前端框架所带来的学习成本，使开发过程更加聚焦于业务逻辑本身。

后端逻辑层以Servlet作为核心控制器，负责接收前端的请求、处理业务逻辑，并返回相应的响应。JavaBean被广泛应用于业务逻辑的封装，使得代码更加模块化、可重用性更强。这种设计不仅提高了代码的可读性，也方便了后续的维护和扩展。

数据存储层采用了MySQL 8.0数据库，通过JDBC实现与数据库的直接连接。MySQL 8.0提供了丰富的数据类型和高效的查询性能，完全能够满足系统对数据存储和处理的需求。在开发工具方面，系统支持Eclipse，提供了强大的代码编辑、调试和测试功能，极大地提高了开发效率。

3　系统设计

3.1　系统核心功能结构

系统围绕老年人健康饮食管理需求，通过多维度功能设计实现科学化、个性化服务，如图1所示。用户管理模块记录个体健康数据，例如，疾病史、过敏源，为精准推荐提供依据，有效规避饮食风险并提升安全性。基于营养学标准构建的食谱信息库结合疾病类型与营养功效分类，帮助老年人快速匹配适宜食谱，解决日常饮食选择难题。一周定制化食谱计划自动计算热量与营养配比，在降低慢性病管理负担的同时支持手动灵活调整，兼顾科学指导与自主性。集成化的美食论坛为老年用户搭建互动交流平台，鼓励分享烹饪经验与食疗心得，结合收藏功能实现偏好内容高效复用，增强使用黏性并缓解饮食管理中的孤独感。此外，通过轮播图与公告信息实时推送权威健康资讯、季节饮食建议及食品安全预警，确保老年人及时获取关键信息以规避潜在风险。

基金项目：河南省软科学项目（编号：252400410603）；2023年度河南省高等教育教学改革研究与实践项目（研究生教育类）“数智驱动下普通高校学士学位授予质量保障机制建设研究与实践”（编号：2023SJGLX387Y）；2024年度河南省高等教育教学改革研究与实践项目（本科教育类）“教育数字化背景下高校研究性教学模式研究与实践”（编号：2024SJGLX0604）。

作者简介：黄莉（1990—），女，主管护师，研究方向为智慧健康养老服务与管理；林金珠（1981—），女，硕士，副教授，研究方向为计算机应用技术。

27 ****,主持人.产教融合背景下“课赛融合”创新实践教学模式研究[J].湖北开放职业学院学报,2026,39(03):1-2+5.

第39卷第3期 总第409期 湖北开放职业学院学报 Vol. 39. No. 3 (Gen. No. 409)
2026年2月(上) Journal of HUBEI Open Vocational College Feb. (first half) 2026

产教融合背景下“课赛融合”创新实践教学模式研究

倪天伟,林金珠
(信阳学院大数据与人工智能学院,河南信阳 464000)

[摘 要]本文以产教融合为背景,深入探讨“课赛融合”教学模式的改革路径与实施策略。通过对相关理论及现状的分析,指出当前高校实践教学中存在的“课赛脱节”问题,提出以产业需求为导向、多专业协同为核心、资源整合为手段的改革路径,并从课程体系重构、教师团队建设、平台搭建、评价体系创新等方面阐述具体实施策略。通过高校教学实践验证,该模式能够有效提升学生的实践能力、创新能力和团队协作能力,为高校实践教学改革提供有益借鉴。

[关键词]产教融合;课赛融合;教学模式;改革路径;实施策略
[中图分类号] G712; G717 [文献标识码] A [文章编号]2096-711X(2026)03-0001-03
doi:10.3969/j.issn.2096-711X.2026.03.001 [本刊网址] http://www.hbxb.net

随着国家创新驱动发展战略的加快推进,对高等院校创新型、应用型等人才培养提出了迫切需求。2023年6月8日,由国家发展改革委等部门印发的《职业教育产教融合赋能提升行动实施方案(2023—2025年)》(发改社会〔2023〕699号),提出着力解决人才培养和产业发展脱节的问题,推动教育和产业协调发展,促进产业需求融入人才培养全过程。学科竞赛是一种连接课堂和产业的有力纽带,其以任务驱动、实践和创新为特征,是目前实施实践教学改革的重要载体。近年来,国内学者对“课赛融合”的教学模式开展了一定研究,范叶青等认为“课赛融合”要实现知识、能力和价值的三位一体,要将竞赛项目进行任务与课目的对齐。王嵘冰等通过“竞赛项目课程化”与“课程内容竞赛化”双载体的方式,实现双向渗透,解决了教学内容与产业需求“两张皮”的问题。崔文明等认为,改变现有的教学模式可以起到提升学生动手能力和创新意识的作用,为社会培养出高素质的生产实用性人才。现阶段,高校“课赛融合”的实践教学模式改革还不够深入,需要加强对竞赛需求和教学课程的深度融合,实现教学工作内容与产业需求的有效对接。在评价学生的综合能力方面,需要加强对学生创新力为核心的思想引导,全面提升学生的学习积极性和创新能力。

一、“课赛融合”教学模式实施中的挑战

在传统“课赛融合”教学模式下,学生的学习能力、团队协作能力在课程实践教学中得到了提高,但在“课赛融合”实践教学深度化的实施中,还有以下问题亟待解决。

(一)课程与竞赛融合深度不足

现阶段高校中的“课赛融合”深度仍有待加强,应该打破课程与竞赛间的壁垒,实现教学内容与产业需求的无缝对接。如有些学院在《机器人技术》的课堂教学中植入了竞赛项目,仅仅将竞赛中的案例进行介绍,不进行具体教学目标与该竞赛技术要求的关联,学生在进行实际比赛的时候,对于在复杂场景中所追求的路径优化能力较低,体现出课程与竞赛形式化的融合表象,无法达到课赛融合的目的。

(二)教师指导效能有待提升

指导教师的能力有待提高。如在指导“中国机器人大赛”中,由于部分指导教师专业领域覆盖范围不够宽泛,在开展指导工作时,难以满足多样化的需求,使得学生团队在竞赛技术攻关环节受阻。

(三)评价体系有待完善

在“课赛融合”模式下,原有的教学评价偏重于以课程考核结果为评价标准,弱化对学生创新过程的动态性评价,使得培养效果呈现片面性。应构建以培养学生创新能力为核心,全面、动态地激励学生学习动力与创新潜力的多元化评价体系。

(四)资源整合与利用率有待提高

在开展“课赛融合”实践教学过程中,应该更加合理配置资源,对实验室、虚拟仿真平台等教学资源加以积极有效使用,切实实现教学活动与竞赛需求的有效满足,为“课赛融合”教学模式的开展提供相应保障。

二、“课赛融合”教学模式的改革路径

(一)以产业需求为导向重构课程内容

1. 动态优化课程大纲

以信阳学院人工智能专业为例,紧跟人工智能产业链技术迭代与行业应用趋势,建立课程内容的动态更新机制。一是将产业前沿技术(如基于AI的计算、多模态大模型轻量化)引入到具体案例之中,确保《人工智能原理》《机器学习》等核心课程的技术前沿性;二是加强竞赛与产业需求的衔接,如在《算法分析与设计》课程的实践中引入“全国大学生数学建模竞赛”中物流优化方面的案例,在《智能控制理论》的课程设计中引入“中国机器人大赛—FIRA小型组”赛项中的动态路径规划案例等,实现“以赛促学、课赛融合”的产业契合性改革。

2. 嵌入竞赛技术标准,实现教学评价与产业实践接轨

把学科竞赛的评价指标融入课程实践教学评价中,实现“竞赛标准课程化”的评价,例如在《机器人工程综合实践》综合评价中,考虑加入竞赛的标准。该评价体系包含四项指标:①功能完整性(40%),主要考核技术实现能力;②创新性(30%),用于评估方案的突破性;③团队协作(20%),考察团队跨学科配合度;④文档规范性(10%),考核内容的完整性。

收稿日期:2025-8-21
基金项目:本文系河南省2023年度产教融合系列项目“以竞赛为载体的‘课赛融合’赋能多专业协同的创新实践教学模式研究”(项目编号:教办高〔2024〕13号);2025年度河南省哲学社会科学教育强省研究项目“AI+区块链技术助推地方数字经济与‘双碳’目标融合研究”(项目编号:2025JYQS0472);2025年河南省软科学项目“人工智能背景下地方本科高校毕业生高质量就业保障机制研究”(项目编号:252400410603)。
作者简介:倪天伟(1981—),男,河南信阳人,信阳学院副教授,主要从事人工智能算法、计算机应用研究。

28 主持人,**.人工智能时代地方本科高校高质量就业路径研究[J/OL].黑龙江工业学院学报(综合版),1-4[2026-04-18].**

DOI: 10.16792/j.cnki.1672-6758.20260203.001 网络首发时间: 2026-02-04 08:13:12 网络首发地址: https://link.cnki.net/urlid/23.1599.Z.20260203.1420.002
第XX卷 第XX期 黑龙江工业学院学报(综合版) Vol. XX No. XX
XXXX年XX月 JOURNAL OF HEILONGJIANG UNIVERSITY OF TECHNOLOGY(COMPREHENSIVE EDITION) XXX. XXXX

文章编号: XXXX－XXXX(XXXX)XX－XXXX－XX

人工智能时代地方本科高校高质量就业路径研究

林金珠,倪天伟

(信阳学院 计算机与人工智能学院,河南 信阳 464000)

摘 要: 人工智能时代的就业变革对地方本科高校提出了结构性挑战,主要表现为人才培养与产业需求脱节、学生高阶思维与职业素养不足、就业服务智能化支撑薄弱以及多元协同育人机制缺失。为此,提出了"精准定位—能力重塑—智慧赋能—协同治理"四位一体的系统提升路径,具体包括"立足区域发展需求深化产教融合,实现培养方向精准定位""重构课程与教学体系,强化学生批判性思维与创新能力培养""建设基于人工智能的智慧就业服务平台,推动人岗精准匹配与职业生涯前瞻指导""构建政府、高校、企业及行业组织协同育人机制,促进多元主体深度融合"等几方面,旨在为人工智能时代地方本科高校就业工作的系统性转型升级提供实践路径参考。

关键词: 人工智能;地方本科高校;高质量就业;智慧就业平台;产教融合

中图分类号: G647.38 **文献标志码**: A

党的二十大报告明确提出"实施就业优先战略""强化就业优先政策,健全就业促进机制,促进高质量充分就业",彰显了国家对高校毕业生就业问题的高度重视。高校毕业生高质量就业,本质上是实现毕业生个体发展、优化劳动力市场结构与资源配置、系统性提升其就业能力与市场适应力的有机统一。然而,在人工智能技术快速迭代的背景下,地方本科高校面临的结构性挑战日益凸显。一方面,人才培养体系滞后于前沿产业需求,加剧了毕业生与岗位之间的"供需错配"。另一方面,人才培养普遍缺乏对学生高阶思维能力与可持续发展素养的系统培育,削弱了其在就业市场中的核心竞争力。此外,就业服务模式仍以传统粗放型为主,缺乏智能化、精准化的技术支撑,"人岗匹配难"的问题依然突出。更深层次看,政府、高校、企业、行业等多元主体尚未形成高效协同的育人机制与持续改进的反馈闭环,导致人才培养与产业需求之间的动态适配机制难以真正形成。

近年来,学界围绕高校毕业生就业开展了广泛研究。滕培秀等提出,在人工智能视域下构筑全过程串联的高校就业指导体系[1];李伟静指出,职业指导、政策支持、服务体系、社保健全和用工机制对于高校毕业生灵活就业均影响显著[2];李敏等则从地方应用型高校视角,提出通过分层指导、强化校企合作和推进"互联网+就业"等策略提升学生就业能力[3]。这些研究提供了有益借鉴,但在人工智能时代下,地方本科高校如何精准对标动态变化的市场需求,重塑自身人才培养特色?如何系统培育学生在人工智能时代不可替代的核心竞争力?如何借助数智技术构建高效、精准的就业服务体系?深入探究这些问题,对于破解地方本科高校毕业生高质量就业难题具有紧迫的理论价值与现实意义。为此,本文将从人工智能时代地方本科高校毕业生的就业现状分析入手,系统剖析其高质量就业面临的核心困境,最终提出一套系统性、可操作的提升路径,以期为相关实践提供参考。

一、人工智能时代地方本科高校毕业生的就业现状

1. 就业市场结构性矛盾加剧

从需求侧来看,一方面,以大数据、云计算和机器学习为核心的数字经济与智能产业快速扩张,催生了大量新兴职业,如人工智能训练师、算法工程师、数据标注专家等,形成了新的就业增长极。另一方面,随着自动化与智能化生产方式的普及,传统制造业、金融业、零售业等领域中标准化程度高、重复性强的岗位正面临规模性萎缩,部分行业甚至出现结构性裁员,造成

作者简介: 林金珠,硕士,副教授,信阳学院计算机与人工智能学院。研究方向: 数字化教育、创新创业。
倪天伟,硕士,副教授,信阳学院计算机与人工智能学院。研究方向: 计算机应用。

基金项目: 2025年河南省软科学研究计划项目"人工智能背景下地方本科高校毕业生高质量就业保障机制研究"(项目编号: 252400410603);2025年河南省哲学社会科学教育强省研究重点项目"人工智能背景下民办高校毕业生高质量就业提升策略研究"(项目编号: 2025JYQS0080)。

29 邹绵璐,周丹,主持人.数智融合下高等教育育人模式创新机制研究[J].湖北开放职业学院学报,2026,39(01):25-27.

第39卷第1期 总第407期 湖北开放职业学院学报 Vol. 39. No. 1 (Gen. No. 407)
2026年1月(上) Journal of HUBEI Open Vocational College Jan. (first half) 2026

数智融合下高等教育育人模式创新机制研究

邹绵璐,周 丹,林金珠
(信阳学院大数据与人工智能学院,河南信阳 464000)

[摘 要]数智融合时代背景下,高等教育育人模式创新的研究成为关键议题。当前,高等教育面临教学模式变革、教学质量评估难题以及师生数智素养提升需求等挑战。构建"行而思,思而化,化而创"的教学模式,强化实践教学、引导思考、鼓励创新,培养学生综合素养;打造多元化与数智化教学质量评价体系,综合多种评价方式,运用数智技术,为教学策略调整提供依据,提升教学质量;实施"培训+研究+竞赛"三维的数智素养提升方案,提升师生数智技能、推动技术应用、激发创新潜能。教学模式创新、评价体系完善和素养提升方案分别是育人模式创新、教学质量提升和师生数智素养增强的重要途径与手段,为高等教育在数智时代的发展提供参考。

[关键词]数智融合;高等教育;育人模式创新
[中图分类号] G642.0; TP18-4 [文献标识码] A [文章编号]2096-711X(2026)01-0025-03
doi:10.3969/j.issn.2096-711X.2026.01.009 [本刊网址] http://www.hbxb.net

一、绪论

党的二十大报告中提出要"推进教育数字化,建设教育强国、科技强国、人才强国",促进教育数字化转型。近年来,国内高校纷纷投入建设智慧校园,旨在通过优化数智教学环境。高等教育研究者已然意识到教师数智化发展的重要性,但它仍然处于理念提出阶段,在理论研究与发展规划等方面仍需要进一步思考。随着大数据、人工智能、云计算等新一代信息技术的快速发展,人类社会正逐步迈入数智化时代。这一时代特征不仅深刻影响着经济社会的各个领域,也对高等教育提出了新的挑战与要求。数智化技术的应用为高等教育育人模式的创新提供了机遇,同时也带来了挑战。当前,人工智能赋能教育变革已成为高等教育系统的共识。未来,人工智能辅助教育教学、科学研究、治理变革,必将成为推动高等教育高质量发展的核心动力。面对数智化时代的挑战,高等教育需要主动适应技术变革,推动育人模式的创新。一方面,数智技术的快速发展要求高等教育培养的人才具备更强的信息处理能力、创新思维能力和团队协作能力;另一方面,数智化技术的应用也促使高等教育的教学方式、学习方式和评价方式发生深刻变革。因此,如何在数智融合的背景下构建适应新时代要求的育人模式,成为当前高等教育亟待解决的重要课题。

二、数智化时代高等教育面临的挑战

(一)教学模式的变革需求

数智化时代,学生的学习方式由单一化、封闭化转变为个性化、多元化。传统的规模化、标准化的教育模式已不能很好地适应新时代对多元化和差异化人才的需求。高校应该探索更加灵活和创新的教育教学模式,以适应时代的发展变化和学生的认知需求及能力特点。同时,数智技术的发展应用也促使教师角色的转变和育人思维的变化,从知识的传授者转变为学生学习的引导者、促进者和支持者,并且要注重对学生批判创新思维和实践能力的培养。

(二)教学质量的评估难题

新时代背景下,可应用数智技术实现对教育质量评价更加精准、实时和动态的反馈。传统的标准化考试评估方式单一,逐渐显现出其局限性,不能全面反映学生的学习过程和能力发展情况。高校需要构建多元化、数智化的教学质量评价体系,通过大数据分析技术和人工智能决策技术,实现对学生学习进程的实时监管和个性化反馈,为教师设计教学内容和制定教学策略提供依据。

(三)数智素养的提升需求

当今时代,数智素养是高校师生与时俱进的体现,必将成为高校师生创新发展的基本能力。然而,当前高校师生的数智素养水平参差不齐,部分师生禁锢思维,无意新知,不愿挑战,在运用数智技术进行学习和研究方面存在困难。因此,高校需要实施有效的数智素养提升方案,可通过集体培训、项目研究和学科竞赛等方式,推动师生对数智技术的应用,培养其数智素养和创新能力,为数智化时代的教育教学提供有力支撑。

综上所述,针对数智化时代高等教育育人模式创新面临的挑战,本文提出三个方面的高等教育育人模式创新实践路径。首先,针对教学模式的变革,构建"行而思,思而化,化而创"的教学模式,注重实践,内化思想,强调创新,集育人思维多元化。然后,针对高等教育教学质量评估难题,打造多元化与数智化教学质量评价体系,多元化评价体系的构建,可以全面客观实时地评价教师教育成果和学生能力发展;数智化评价技术,可实时监控教学质量,反馈教育功效。最后,针对数智素养的提升要求,实施"培训+研究+竞赛"三维的数智素养提升方案,以培训推动数智技能,以研究提升数智思想,以比赛促成数智素养。

三、数智融合下高等教育育人模式创新的实践路径

在高等教育教学改革创新的进程中,构建"行而思,思而化,化而创"的教学模式是核心要点。通过引导学生在实践中思考,在思考中领悟知识,转化知识,进而在转化的基础上创新应用,升华知识,来切实提升学生的综合素养,从而达到育人模式的创新。而打造多元化与数智化教学质量评价体系,则为这一教学模式提供了有力的数据支撑和保障,借助

收稿日期:2025-6-13
基金项目:本文系2025年河南省哲学社会科学教育强省研究项目"数智融合下的河南省高等教育育人模式创新与自主知识体系构建研究"(项目编号:2025JYQS047);2025年河南省科学技术协会项目"河南省科技创新赋能乡村振兴发展研究"(项目编号:HNKJZK-2025-99B)。
作者简介:邹绵璐(1994—),女,河南信阳人,信阳学院大数据与人工智能学院讲师,主要从事机器学习、教育信息化研究。

30 吴重胜,主持人,马永帅,等.民办高校学生学业预警系统的设计与实现[J].电脑编程技巧与维护,2025,(07):8-10.

民办高校学生学业预警系统的设计与实现

吴重胜，林金珠，马永帅，刘旭喆，高浩展
（信阳学院大数据与人工智能学院，河南 信阳 464000）

摘　要：随着大数据与信息技术在教育领域的广泛应用，构建民办高校学生学业预警系统成为提升教育质量、有效预防学业危机的关键。该系统依托 Eclipse 平台，融合 Java、JSP 技术，采用 MySQL 数据库及 MVC 设计模式，并巧妙运用 Struts 2 框架实现控制层与视图层的分离。系统通过综合考勤、成绩、学习行为及心理与生活状态的多维度预警机制，精准、全面地识别学生的学习难题与需求，实现预警的个性化定制。这不仅有助于学生及时调整学习策略、有效规避学业危机，还有助于高校提前介入，为学习困难的学生提供精准的辅导与支持，有效促进了和谐校园环境的构建。

关键词：民办高校学业预警系统；MVC与Struts2框架；多维度预警；个性化定制
DOI:10.16184/j.cnki.comprg.2025.07.031

1　概述

民办高校学生在学习动力、习惯及心理压力上的差异，使得学业问题频发，这促进了学生学业预警系统的研发。相关领域已有一定的研究成果积累，例如，Shahiri 等[1]研究详细阐述了利用数据挖掘技术预测学业表现的方法，Hamsa 等[2]基于决策树和模糊遗传算法，提出了一种新型的学业成绩预测模型。此外，通过选取大连理工大学的435个样本，崔强和孙智妍[3]采用8种因素作为大学生学业水平的主要影响因素，进行了深入探索。还有研究利用改进的 K-means 聚类方法处理学生成绩数据，并在 Apriori 算法基础上加入额外兴趣度度量，对离散数据进行关联性分析，挖掘出了有价值的信息[4]。崔佳杉等[5]利用公开数据集，结合 XGBoost 算法建立了学业预警模型，展示了其在学生成绩预测上的优越性能。

当前的研究多聚焦于成绩预警，而对学生学习行为、心理状态等方面的预警关注较少。同时，高校在构建学生学业预警系统时，往往缺乏与干预措施的有效结合，这样会导致学生学业预警系统的作用未能得到充分发挥。因此，设计一套能够全面监测学生学习状态、及时预警学业风险，并触发个性化干预措施的学生学业预警系统显得尤为重要。这不仅能够有效降低学生学业失败的风险、提升学生的毕业率，还能为民办高校的学生管理工作提供有力的支持。

2　系统设计

2.1　系统核心功能结构

该系统设计了一套全面的学生管理与预警机制，如图1所示，旨在通过多维度的数据分析，确保学生的全面发展与健康成长。其中，考勤预警模块全面监测学生的出勤情况，包括常规课堂的出缺勤记录、实验室与实践课的考勤登记，以及学生参与校内活动的考勤数据。成绩预警模块则专注于学生的学业表现，通过分析学生的成绩排名变动、作业提交情况、关键科目成绩及重修与补考情况，及时识别学业上存在风险的学生。学习行为预警模块深入学生的学习过程，关注他们的日常学习行为，通过监测学生在线学习时长与活跃度、图书馆资源利用情况及课堂参与度，能够发现学生学习中的不良习惯或动力不足等问题，为后续的干预和指导提供依据。心理与生活状态预警模块则关注学生的心理健康和生活状况，通过分析心理健康测评结果、生活习惯及经济状况，及时发现学生可能面临的问题，并为他们提供必要的心理支持和生活指导，确保学生的身心健康得到妥善关注。

基金项目：2024年度信阳学院大学生校级科研项目“高校学生学业预警系统的设计与实现”；2024年河南省大学生创新创业训练计划项目“高校学生学业预警系统的设计与实现”（202413503010）；2023年度河南省高等教育教学改革研究与实践项目（研究生教育类）“数智驱动下普通高校学士学位授予质量保障机制建设研究与实践”（2023SJGLX387Y）；2024年度河南省高等教育教学改革研究与实践项目（本科教育类）“教育数字化背景下高校研究性教学模式研究与实践”（2024SJGLX0604）。

作者简介：林金珠（1981—），女，副教授，硕士，研究方向为计算机应用技术、高等教育等。

31 张冬松,司杰,毛凤翔,等.新工科背景下大数据应用型人才培养模式研究[J].教育信息化论坛,2022,(02):72-74.

超星·期刊

教育信息化论坛

EDUCATIONAL INFORMATIZATION FORUM

终身教育理念下档案学在线课程资源建设探析

高校师范生信息化教学能力提升策略研究

基于SPOC的线上线下混合式教学研究

基于信息化教学的大学翻转课堂教学实践

现代信息技术与高校教学融合的路径探究

高校双创信息化平台建设路径研究

2022年 总第111期 02 上半月

CONTENTS

目录

2022

02

专业建设与教学改革

人才培养和机制创新

大学生心理发展与教育

创新与创业教育

实训与实践探索

思政教育研究

32 张冬松,胡秀云,邬长安等.面向 DevOps 的政务大数据分析可视化系统[J].计算机技术与发展,2020,30(08):1-7.

新工科背景下大数据应用型人才培养模式研究

张冬松　司　杰　毛凤翔　贾彦茹

摘要：为满足大数据这类新兴行业需要大量应用型人才的新需求，考虑到传统的计算机专业人才培养模式还不能很好地满足社会需求，分析新工科建设背景下大数据应用型人才培养模式，指出校企融合的多主体协同育人培养模式是高校的必然选择，从人才培养、过程管理、课程体系、师资队伍、校企合作等方面入手，提出大数据专业应用型人才培养模式的举措建议。

关键词：新工科；大数据；应用型人才；人才培养

注：本文系 2020 年河南省新工科研究与实践项目“新工科背景下民办高校计算机类专业‘产学研用’协同育人模式探索与实践”（2020JGLX097）的阶段性研究成果之一。

新工科建设主要是指在新形势、新经济和新技术的时代背景下，将以往仅仅依靠高校来培养人才的模式向政产学研用的多方协同培养人才的模式进行转换，旨在培养高校学生成为具有较高的理论基础、实践能力和创新能力的高素质工程技术人才[1]。由此可知，“新工科”的提出必然会带来一场教学改革，即对新形势下高校理工科人才培养模式的全面创新。

面对新工科的挑战，传统的计算机专业人才培养模式和社会的需求还存在脱节，本科教育的基本理念也需要转变，尤其是在大数据专业应用型人才培养方面。

据悉，国内院校培养大数据人才尚处于起步阶段，从 2016 年才新设这个专业，至 2017 年全国有 35 所院校设立专门的大数据技术专业。人才培养目标是以大数据为核心研究对象，利用大数据的方法解决具体行业应用问题，以统计学、数学、计算机为三大支撑性学科[2]。

综上，本文首先分析新工科背景下地方高校大数据专业多主体协同育人模式的必要性，其次探究相应的举措建议，主要从人才培养新模式、全过程管理、模块化实践课程体系、双师型教师队伍建设和校企合作新方式等层面进行分析研究，旨在加强校企产学合作，提升人才工程实践能力，构建校企多主体协同育人的大数据专业应用型人才培养模式。

一、什么是新工科

2018 年，教育部开始大力发展新工科、新医科、新农科、新文科，通过大力发展“四个新”，使学科专业结构进一步优化，力求推动新工科形成覆盖全部学科门类的具有中国特色的、世界水平的一流本科专业集群。同年，教育部首次批准并认定了 612 个国家级新工科研究与实践项目，鼓励全国高校与企业、行业、相关部门深入合作，探索新工科建设的新模式和新方法，以便主动应对当前新科技革命的召唤，加快为国家工业产业结构优化和变革培养大量优秀的工程技术人才。截至目前，教育部增设大数据、人工智能、机器人、物联网等新兴领域急需专业点近 400 个，组建人工智能、大数据、智能制造等项目群，提倡校企之间进行产学研协同育人发展[3]。

目前，新工科建设正迈向再深化的新阶段，从轰轰烈烈到扎扎实实。为推进新工科再深化，2019 年教育部成立了“全国新工科教育创新中心”，指导全国高校主动探索新工科建设背景下的本科教育人才培养体系，形成一大批具有世界影响力的工程人才教育创新中心[4]。同年，教育部还组织编制了第二批“新工科研究与实践项目指南”[5]，将全国高校分

作者简介：张冬松，博士，信阳学院大数据与人工智能学院副教授，研究方向为大数据应用、自然语言处理及实时系统；司杰，信阳学院大数据与人工智能学院讲师，研究方向为计算机软件；毛凤翔，信阳学院大数据与人工智能学院副教授，研究方向为网络安全；贾彦茹，信阳学院大数据与人工智能学院副教授，研究方向为计算机软件。

33 成员 7,张炎炎.导师选择系统的设计与实现[J].信息技术与信息化. 2019,(04):97-99

导师选择系统的设计与实现

Design and Implementation of Tutor Selection System

贾彦茹* 张炎炎**
JIA Yan-ru ZHANG Yan-yan

摘 要 随着师范生人数的不断增多，校内外实习过程中选导师的工作量越来越大。本文采用Java语言进行编程，结合数据库实现了学生登录自己的学号进入该系统根据自己的情况选导师；在整个系统中管理员可以根据实际情况对导师信息进行及时的维护更新，查看学生选导师情况并进行调整。

关键词 Java；学生选导师系统；数据库

Abstract With the increasing number of normal students each year, The workload of selecting tutors is increasing in the process of internship inside and outside school. In this design, the Java language is used for programming, and the system function is realized by combining with the use of the database, which can effectively manage the graduates' choice of tutors. Administrators can update mentor information in time, can check the situation of students'selection of tutors and make adjustments.

Key words Java; Students choose a tutor system; Database

doi：10.3969/j.issn.1672-9528.2019.04.029

0 引言

如今互联网迅猛发展，无论是人们的生活方式和工作方式，还是教学和学习方式，都有很大的变化，甚至是思维方式，都发生了不可否认的变化。怎样才能更有效的将信息技术与各学科教学资源进行结合，以便达到全面提高教学的质量，培养学生创新精神和创新能力，以适应新世纪对人才的需求，是近年来国内外广大教育工作者所关注的热点问题之一。随着各高校的规模不断扩大，学生人数急剧上升，每年的师范生也越来越多，出现了师范生在校内外实习过程中选择导师时容易扎堆的现象，从而可能导致一些老师由于带的学生太多而忙不过来，但有些导师带极少的学生而出现每天工作很闲的现象。面对这样的极端问题，那么就要有一个学生选导师系统来提高学生指导导师分配管理工作的效率。通过这样的一个系统可以做到学生合理的选择导师、有条理的分配导师和快速查询、修改、增加、删除导师信息和学生选择信息等，进一步减少繁琐的工作量，提高各部门的工作效率。

1 系统分析与设计

1.1 系统采用的开发工具、编程语言、服务器

该学生选导师系统采用的开发工具是Eclipse。Eclipse是一个开放源代码的、基于Java的可扩展性的开发平台，并且Eclipse附带了一个标准的插件集，包括Java开发工具（Java Development Tools，JDT）[1]。

该学生选导师系统采用的编程语言是Java语言，Java语言相对于C、C++等语言来说最主要特点是面向对象，具有跨平台、安全同时支持多线程的特点[2]。Java中提出生活中的万事万物都是对象的说法，同时有这样一个规定：在类的外面定义数据和函数没有任何意义，对象是Java语言中最外层的数据类型，必须通过类和对象去访问成员。Java程序还具有与体系结构无关的特性，可以方便地移植到网络上的不同计算机中，无论哪个移植了Java解释器的计算机或者是其他设备都能够对Java字节码进行解释执行[3]。本系统中总共有12个.java文件，从三个层次进行操作。

该系统所采用的服务器是Tomcat服务器。我们编好的

* 阳学院数学与信息学院 河南信阳 464000
** 信阳学院外国语学院 河南信阳 464000
[基金项目] 2018年度河南省教育科学“十三五”规划项目

35 成员 4,冯岩,郭颂,等.新工科背景下地方高校计算机应用型人才培养模式[J].计算机教育,2021,(11):14-17.

文章编号：1672-5913(2021)11-0014-04　　中图分类号：G642

新工科背景下地方高校计算机应用型人才培养模式

尤　磊，冯　岩，郭　颂，李　蕾，吴　宏，张　帆，刘　欣，郭旭展
（信阳师范学院 计算机与信息技术学院，河南 信阳 464000）

摘　要：针对新工科建设中地方高校如何培养适合新经济发展需要的计算机应用型人才问题，在剖析计算机类专业人才就业形势的基础上，分析计算机类应用人才应具备的特征，进而提出新工科建设背景下地方院校计算机应用型人才培养模式。

关键词：工程教育；新工科；新经济；应用型人才；计算机
DOI:10.16512/j.enki.jsjjy.2021.11.004

0　引　言

新工科是在新科技革命、新产业变革、新经济背景下提出的工程教育改革战略[1]，建设目标包括构建世界一流的工程教育体系、构建引领全球工程教育的中国工程教育模式、建成工程教育强国等[2-4]，一经提出就受到高等教育界的高度关注。近年来，围绕新工科建设的相关研究成果逐渐增多，很多高等院校正在开展新工科建设的改革与实践类项目。

在新工科建设过程中，不同类型的高校肩负着不同的使命。工科优势高校、综合性高校与地方高校的新工科建设目标、内容和路径各不相同[2,4]。对于地方高校而言，需要对区域经济发展和产业转型升级起到支撑作用[2]，应积极探索与深化工程教育改革、探索适合学校特色的新工科建设模式，为区域经济发展提供才智支撑。在此背景下，地方高校如何开展新工科建设是地方高校教育工作者亟须研究与探讨的一个问题。针对该问题，目前已经涌现出很多研究成果，从明确办学定位[5]、确立人才培养目标[6]、学生实践能力培养[7]等角度探究新工科的建设模式。

新经济是一个动态的、相对的发展状态。因此，新工科专业的布局也必然是一个动态调整的过程。当前，应鼓励地方高校着眼于互联网革命、新技术发展与制造业升级等时代特征来构建新工科人才培养体系[8]。与传统工科专业相比，新工科涉及的专业主要是新兴产业的专业，即专业以互联网和工业智能为核心，以新型信息、能源、控制等领域为主干[9]。上述专业与领域的发展与计算机类专业的发展密切相关，也是计算机类专业在这些行业的应用与发展，因此计算机类专业是新工科建设体系中的重要组成部分[10]。在当前新工科建设的时代背景下，加强对计算机应用型人才的培养显得尤为重要。

1　计算机类专业人才就业形势分析

近年来，在全球经济一体化发展与“互联网+”赋能实体经济的时代背景下，信息产业规模不断扩大并逐渐渗透到各行各业中。云计算、移动互联网、物联网、人工智能、大数据、电子商务、区块链等计算机类新兴产业不断涌现[11]。从就业市场来看，这些新兴产业的出现与发展促使市场对计算机类专业人才产生强烈需求。《2020 年中国本科生就业报告》[12]显示，2019 届计算机类本科专业毕业生的薪资最高，毕业生在“信息传输、软件和信息技术服务业”的就业比例持续升高，

基金项目：河南省高等教育教学改革研究与实践项目 (2019SJGLX349)；河南省高等学校青年骨干教师培养计划 (2020GGJS157)；河南省新工科研究与实践项目 (2020JGLX055)；信阳师范学院 2019 年教育教学改革研究与实践项目。
第一作者简介：尤磊，男，副教授，研究方向为计算机图形学、三维点云分析与应用，leiyou@xynu.edu.cn。

新工科

36 冯岩,成员 4,李健,等.新工科背景下人工智能课程的教学改革[J].福建电脑,2022,38(04):118-120

新工科背景下人工智能课程的教学改革

冯岩 尤磊 李健 王敬

(信阳师范学院计算机与信息技术学院 河南 信阳 464000)

摘 要 根据新时代对人工智能人才的要求，本文探讨了人工智能课程的教学改革方法。通过分析目前人工智能课程教学存在的问题，提出了优化课程内容、改变教学模式、加强校企深度合作、构建多元评价机制等方面的改革方法。实践结果表明，改革后的教学能激发学生的学习兴趣，不仅提升了学生的实际动手能力，还有利于提高学生的创新能力和综合素养。

关键词 新工科；人工智能；教学模式；多元评价

中图法分类号 TP391 DOI:10.16707/j.cnki.fjpc.2022.04.031

Teaching Reform and Practice of Artificial Intelligence Course Based on the Background of New Engineering

FENG Yan, YOU Lei, LI Jian, WANG Jing

(School of Computer and Information Technology, Xinyang Normal University, Xinyang, China, 464000)

Abstract According to the current requirements of artificial intelligence talents, this paper discusses the teaching reform methods of artificial intelligence course. By analyzing the problems existing in the teaching of artificial intelligence course, this paper puts forward some reform methods, such as optimizing the course content, changing the teaching mode, strengthening the in-depth cooperation between schools and enterprises, and constructing a diversified evaluation mechanism. The results of practice show that the reformed teaching can stimulate students' interest in learning, not only improve students' practical ability, but also improve students' innovation ability and comprehensive quality.

Keywords New Engineering; Artificial Intelligence; Teaching Mode; Multiple Evaluation

1 引言

为主动应对新一轮科技革命与产业变革，支撑服务创新驱动发展、“中国制造2025”等一系列国家战略，2017年以来，教育部积极推进新工科建设，先后形成了“复旦共识”、“天大行动”和“北京指南”[1]。人工智能专业是“新工科”建设的重点专业之一。人工智能课程是人工智能专业及其它相关专业的核心课程。《新一代人工智能发展规划》明确提出高校要完善人工智能教育体系[2]。《高等学校人工智能创新行动计划》强调：要加强人工智能领域专业建设，推进“新工科”建设，形成“人工智能+X”复合专业培养新模式[3]。

为适应国家发展战略和新时代对人工智能人才的需求，需要对人工智能课程进行教学改革。本文以新工科建设为导向，探索人工智能课程的改革与实践。通过分析当前人工智能教学存在的问题，有针对性地对人工智能教学提出一些改革措施，以适应新一轮科技革命和产业变革的人才需求。

2 存在问题

本文得到河南省新工科研究与实践项目(No.2020JGLX055)、河南省高等教育教学改革研究与实践项目 (No.2019SJGLX349)、河南省高等学校青年骨干教师培养计划 (No.2020GGJS157)资助。冯岩(通信作者)，男，1971年生，主要研究领域为小波分析、人工智能。E-mail: yfeng@xynu.edu.cn。尤磊，男，1978年生，主要研究领域为计算机图形学、三维点云分析与应用。E-mail: leiyou@xynu.edu.cn。李健，男，1992年生，主要研究领域为机械臂控制、强化学习。E-mail: lijcit@xynu.edu.cn。王敬，男，1989年生，主要研究领域为神经影像学和机器学习。E-mail: wangjing@xynu.edu.cn。