中图法分类号:TP183　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-14

论文引用格式:Zhang Jin, Liang Fengmei, Liu Ajian, Wan Jun, Cai Zhanchuan, Lei Zhen. Multimodal prompt learning for domain generalization face anti-spoofing[J/OL]. Journal of Image and Graphics, XXXX:1-14. DOI: 10.11834/jig.260118.(张瑾, 梁凤梅, 刘阿建, 万军, 蔡占川, 雷震. 面向域泛化人脸活体检测的多模态提示学习[J/OL]. 中国图象图形学报,XXXX:1-14. DOI: 10.11834/jig.260118.)[DOI:10.11834/jig.260118]

面向域泛化人脸活体检测的多模态提示学习

张瑾[1],梁凤梅[1],刘阿建[2,3],万军[2,3],蔡占川[3],雷震[2,3]

1. 太原理工大学电子信息工程学院,山西省晋中市　030600; 2. 中国科学院自动化研究所,北京市　101190; 3. 澳门科技大学计算机科学与工程学院,澳门特别行政区　999078

摘　要:**目的**　人脸活体检测是保障人脸识别系统安全的核心技术之一。目前基于视觉-语言多模态的人脸活体检测算法借助语言模态的表达能力来提高模型泛化性,但由于任务特定的类别语义与域语义模糊问题,模型难以同时稳定刻画类别判定线索与跨域环境变化,性能提升受限。因此,提出面向域泛化人脸活体检测的多模态提示学习方法,通过预定义的语义信息引导提示学习,获得模型可理解的类别语义与域语义。**方法**　首先构建类别语义元素集和域语义元素集,将类别语义和域语义分解为可学习的语义元素。其中,类别语义元素用于表征真实或伪造类别概念,域语义元素用于表征光照变化、图像质量、传感器差异等域属性。其次,基于注意力机制选择并融合匹配的语义元素,为视觉分支与文本分支生成模型可理解的语义提示,并将提示以层级结构嵌入到各自模态序列,补充语义表示,增强模型对类别概念和域差异的理解。最后,使用分层多模态适配器使模型在理解人脸活体检测任务的同时对齐文本描述与视觉特征。**结果**　在人脸活体检测任务公开的四个基准数据集CASIA-FASD、Idiap-Replay Attack、MSU-MFSD和OULU-NPU上进行跨域场景实验,本方法的AUC分别达到99.99%、97.66%、99.77%、99.67%,HTER分别为0.11%、7.07%、1.67%、1.70%,平均HTER达到2.64%。**结论**　与最新的多种方法进行比较,本文所提模型性能优于其他模型,能够有效缓解域泛化人脸活体检测中的语义模糊问题,提升域泛化能力。

关键词:人脸活体检测;域泛化;视觉-语言模型;提示学习;语义元素建模;跨模态对齐

Multimodal prompt learning for domain generalization face anti-spoofing

Zhang Jin[1], Liang Fengmei[1], Liu Ajian[2,3], Wan Jun[2,3], Cai Zhanchuan[3], Lei Zhen[2,3]

1. College of Electronic Information Engineering, Taiyuan University of Technology, Jinzhong 030600, China; 2. Institute of Automation, Chinese Academy of Sciences, Beijing 101190, China; 3. School of Computer Science and Engineering, Macao University of Science and Technology, Macao Special Administrative Region 999078, China

Abstract: **Objective** In the current digital era, face recognition technology has been widely applied in many critical fields such as identity verification, intelligent access control, mobile payment, and public security, becoming a core supporting technology for ensuring information security and convenient services. However, face recognition systems are confronted with severe security threats. Malicious attackers can deceive systems by forging faces (e.g., printed photos, 3D masks, video replay), leading to security incidents such as identity impersonation and property loss. Accordingly, face anti-spoofing has emerged as the core defense against such attacks and for safeguarding the security of face recognition systems.

收稿日期:2026-03-05;**修回日期**:2026-05-09

基金项目:中国国家自然科学基金会项目(项目编号:62406320);澳门科学技术发展基金项目(项目编号:0140/2024/AGJ);广东省科学技术基金项目(项目编号:2025A0505010025)

Supported by: Chinese National Natural Science Foundation Projects (Project No.: 62406320); Science and Technology Development Fund of Macau Project (Project No.: 0140/2024/AGJ); Science and Technology Foundation of Guangdong Province Project (Project No.: 2025A0505010025)

© 中国图象图形学报版权所有

With the rapid development of deep learning, face anti-spoofing has been continuously upgraded, evolving from early methods based on unimodal visual features (e. g. , texture, depth, motion information) to vision-language multimodal fusion-based detection methods. Compared with unimodal approaches, vision-language multimodal face anti-spoofing algorithms leverage the powerful semantic representation capability of the language modality to associate visual features with textual descriptions. This effectively alleviates the insufficient generalization ability of a single modality in complex scenarios and improves the model's adaptability to different types of forgery attacks and various shooting environments to a certain extent. Although existing multimodal face anti-spoofing algorithms have achieved certain progress, obvious performance bottlenecks still exist in practical applications. The core problems focus on two aspects: task-specific category semantic ambiguity and domain semantic ambiguity. Category semantic ambiguity is mainly reflected in that the model struggles to accurately understand the essential semantic differences between the two core categories — "live" and "fake", and the semantic features of different forgery types cannot be effectively distinguished, resulting in decreased recognition accuracy when the model encounters unseen forgery types. Domain semantic ambiguity arises from the diversity of shooting environments: significant differences exist in shooting angles, lighting conditions, background complexity, device resolution, etc. , across different datasets. The model has difficulty capturing the common semantics between different domains and is prone to domain shift, which greatly degrades detection performance in cross-domain scenarios. The above semantic ambiguities severely limit the generalization ability of multimodal face anti-spoofing models and fail to meet the demands of complex scenarios and cross-domain detection in real-world applications. Therefore, aiming to address the shortcomings of existing algorithms, this paper proposes a domain-generalized face anti-spoofing model based on multimodal prompt learning. By employing predefined semantic information to guide the prompt learning process, the model is helped to clearly and accurately acquire understandable category semantics and domain semantics, effectively alleviating semantic ambiguity and improving detection performance in cross-domain scenarios, thus providing more reliable security guarantees for face recognition systems. **Method** We designs a complete multimodal prompt learning framework that enables the model to fully understand category semantics and domain semantics, thereby improving cross-domain generalization ability. First, two sets of learnable semantic elements are constructed, corresponding to category semantics and domain semantics respectively. The category semantic element set contains various semantic elements that describe the essential characteristics of categories, used to represent the concepts of live or fake faces. The domain semantic element set is employed to characterize domain attributes such as lighting variations, image quality, and sensor differences. All semantic elements are represented as learnable vectors and continuously optimized during model training to better adapt to task requirements. Second, for the input visual features and textual descriptions, semantic elements are selected and fused based on the attention mechanism to generate model-understandable semantic prompts. These prompts are embedded into the corresponding modal sequences in a hierarchical structure to supplement semantic representations and enhance the model's understanding of category concepts and domain discrepancies. Finally, a hierarchical multimodal adapter module is introduced to achieve precise alignment between textual descriptions and visual features, while helping the model deeply understand the core requirements of the face anti-spoofing task. The multimodal adapter adopts a lightweight structure, which does not require large-scale modifications to the original model and can quickly adapt to the fusion needs of vision-language multimodal features. **Result** To verify the effectiveness and superiority of the proposed model, this paper conducts extensive cross-domain experiments on four internationally recognized benchmark datasets for face anti-spoofing: CASIA-FASD, Idiap-Replay Attack, MSU-MFSD, and OULU-NPU. Experimental results under cross-domain scenarios show that the proposed model achieves AUC values of 99. 99%, 97. 66%, 99. 77%, and 99. 67% on the four benchmark datasets CASIA-FASD, Idiap-Replay Attack, MSU-MFSD, and OULU-NPU, respectively, with corresponding HTER values of 0. 11%, 7. 07%, 1. 67%, and 1. 70%. The average HTER reaches 2. 64%, demonstrating stable and outstanding cross-domain detection performance. **Conclusion** Experimental results demonstrate that our model achieves superior performance in cross-domain scenarios. This fully validates that the proposed multimodal prompt learning framework can effectively alleviate the semantic ambiguity problem, improve the model's ability to understand category semantics and domain semantics, and enhance the model's cross-domain generalization performance.

Key words: face anti-spoofing; domain generalization; vision-language model; prompt learning; semantic element model-

© 中国图象图形学报版权所有

ing; cross-modal alignment

论文引用格式：Zhangjin, Liang Fengmei, Liu Ajian, Wan Jun, Cai Zhanchuan, Lei Zhen. XXXX. Multimodal prompt learning for domain generalization face anti-spoofing. Journal of Image and Graphics, XX(XX):XXXX-XXXX(张瑾,梁凤梅,刘阿建,万军,蔡占川,雷震. XXXX. 面向域泛化人脸活体检测的多模态提示学习. 中国图象图形学报, XX(XX):XXXX-XXXX[DOI:10.11834/jig.260118]

0 引 言

人脸识别技术因人脸获取便捷、非接触的生物特性被广泛应用于智能解锁、电子支付、银行卡或证券开户等各种场景。然而,该技术面临着照片打印、视频重放或3D面具等呈现攻击的威胁,易造成身份冒充、财产损失等安全问题。为应对这一问题,人脸活体检测(Face Anti-Spoofing, FAS)技术(Xie 等, 2022)应运而生,通过区分采集到的人脸来自真实活体或者呈现攻击从而判断真实人脸与伪造人脸,以保障人脸识别系统的安全性。

早期的人脸活体检测方法主要依赖手工特征(Pereira 等,2012;Li 等,2016;Patel 等,2016)或深度学习特征(Yu 等,2020;Yu 等,2021),在训练和测试分布相似的域内场景中表现良好,但面对未知域时无法获得理想的性能。为提升模型在未知域上的性能,基于域泛化的人脸活体检测方法(Domain Generalization Face Anti-Spoofing, DG-FAS)被提出,通过从多个源域中学习与领域无关的判别特征来增强模型在未知测试域上的泛化能力。其中基于对抗学习(Shao 等,2019;Jia 等,2020)的方法通过生成器与判别器的对抗训练,使生成器不断欺骗判别器,从而学习到域不变的特征表示;元学习方法(Chen 等, 2021;Zhou 等,2022)则倾向于在训练时从源域模拟到未见域,提升模型跨域泛化能力;其他方法,如实例白化(Wang 等,2022;Zhou 等,2023)以自监督的方式实现特征对齐。这些方法在提升模型泛化性方面取得了一定进展,但它们仅使用单一视觉模态的图像数据训练模型,且依赖相应的图像标签引导其学习,这限制了模型捕获真实域分布的能力,进而限制模型对复杂未知场景的泛化能力。

而视觉-语言模型(Vision-Language Models, VLMs)(Radford 等, 2021; Jia 等, 2021; Yuan 等, 2021)为解决这一问题提供了新的思路,其利用文本描述的丰富语义信息来补充视觉特征,从而增强人脸活体检测算法对不同攻击类型和域变化的理解。其中,对比语言-图像预训练(Contrastive language-image pre-training, CLIP)模型(Radford 等,2021)凭借其强大的跨模态对齐能力,被视为提升人脸活体检测域泛化能力的有力工具。FLIP(Srivatsan 等, 2023)首次将CLIP引入FAS,通过分类标记编码全局图像信息。VL-FAS(Fang 等,2024)利用面部区域的细粒度自然语言描述去除无关噪声以增强泛化。但由于原始的CLIP并非专门为FAS训练,直接将其应用于人脸活体检测任务时,往往存在两个问题:1)类别语义模糊。CLIP的文本模态对FAS的任务特定类别存在语义模糊,即文本编码器无法理解“真实人脸对应live,伪造人脸对应fake”的语义内涵,从而限制了模型性能的进一步提升。2)域语义模糊。FAS通常将每个数据集视为一个独立的域,而基于CLIP的检测器往往难以使用确切语言来描述图像分辨率、相机传感器、光照变化等域信息,使得模型在理解和区分不同域信息时容易产生混淆和不确定性,进而导致模型泛化性能下降。

为缓解上述问题,先前研究尝试引入提示学习(Liu 等,2024;Guo 等,2025)来提升模型的域泛化能力。CFPL-FAS(Liu 等,2024)通过无类别的提示学习,利用内容和风格特征动态调整分类器权重。CCPE(Guo 等,2025)融合固有内容提示和可学习内容提示以解决单一提示形式无法描绘多样化攻击类型的问题。现有提示学习方法通常通过优化少量全局提示标记或模板参数来适配人脸活体检测,但这类方法往往共享同一组提示参数来服务全部样本和所有源域,提示优化主要依靠任务自身监督,缺少对样本级域偏移和类别差异进行针对性设计。对于人脸活体检测而言,伪造类别同时包含打印攻击、回放攻击和3D面具等差异明显的攻击形式,而不同数据域在光照变化、传感器响应、清晰度与压缩失真上又具有显著差异。若仍使用统一的可学习提示,类别语义与域语义容易在同一提示空间中耦合,模型最终更容易学习到源域的平均语义,而难以针对未知

© 中国图象图形学报版权所有

域生成具有区分性的解释。

因此，本文提出面向域泛化人脸活体检测的多模态提示学习方法(Multimodal Prompt Learning for Domain Generalization Face-anti Spoofing，MMPL)，通过预先定义的语义元素来生成模型可理解的类别语义提示和域语义提示，从而提升模型域泛化能力。具体来说，本文将语义知识扩展为可学习的语义元素集合，分别构建类别语义元素集和域语义元素集，通过多个语义元素来刻画不同层面的类别信息与域信息。接着通过注意力机制为每个输入动态挑选与融合相关语义元素，以生成适配当前样本的类别提示与域提示。这些提示以层级结构嵌入到对应的视觉与语言模态，增强模型对类别概念与域信息的理解。最后，使用分层多模态适配器对齐视觉特征与文本特征的语义空间，提升跨域场景下人脸活体检测模型的泛化性能。

1 本文方法

本文采用目前主流的视觉-语言预训练模型CLIP作为基础模型，整体框架如图1所示。设计可学习的语义元素集合，利用多个语义元素来刻画不同层面的类别信息和域信息。当语义元素足够丰富时，即便面对未知域或罕见攻击，模型也能够在集合中找到近似的语义表达，并通过加权融合生成新的提示以强化类别语义和域语义，解决语义模糊问题。

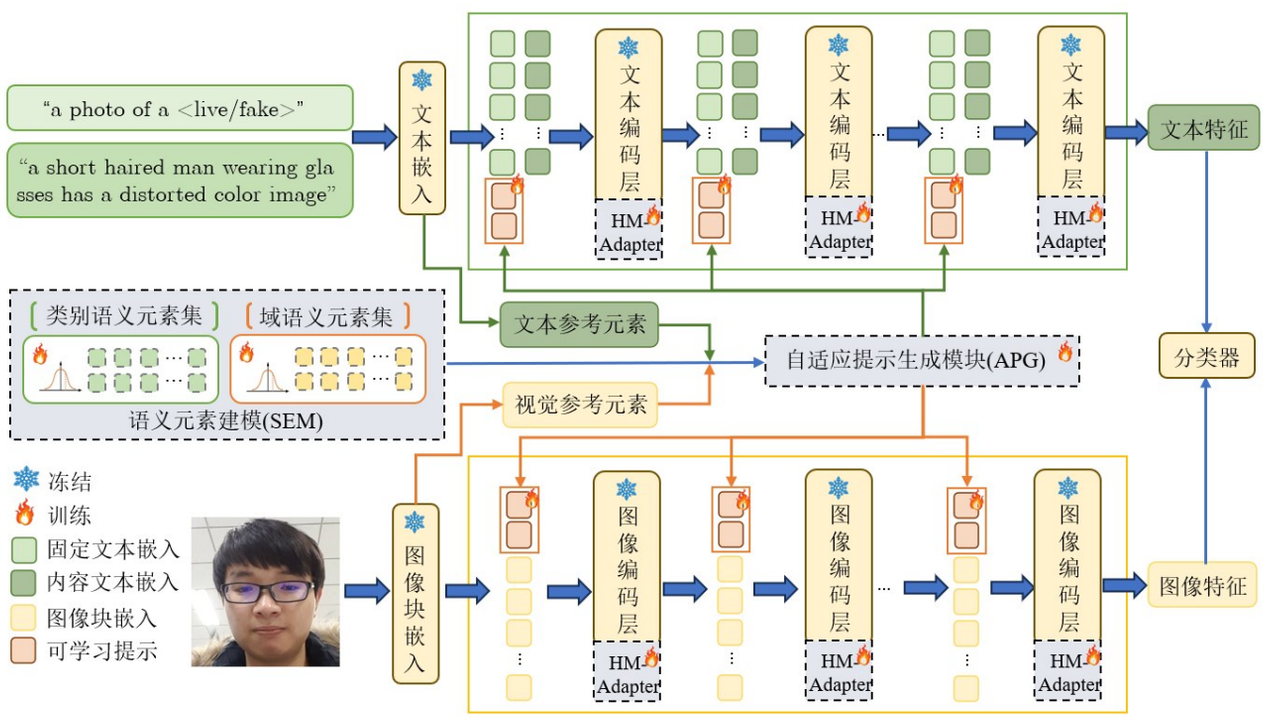


图1 模型整体架构

Fig. 1 Overall Architecture of the Model

1.1 语义元素建模

为克服人脸活体检测任务现有方法中存在的类别语义与域语义模糊问题，首先进行语义元素建模(Semantic Element Modeling，SEM)，将复杂的语义知识分解为可学习的语义元素，并构建可学习语义元素集。

可学习语义元素集包括类别语义元素集和域语义元素集。其中，语义元素通过均值为0、标准差为0.02的高斯分布进行初始化采样，并在模型训练中持续优化。定义 $\boldsymbol{U}_{class}=\{(p_{live}^{n},p_{fake}^{n})\}_{n=1}^{N}\in\mathbb{R}^{N\times d_t}$ 表示类别语义元素集，其中 N 表示语义元素的数量，d_t 为文本向量维度。p_{live}^{n} 和 p_{fake}^{n} 分别表示刻画“活体”和“伪造”类别的语义元素，用于描述真实人脸与欺骗人脸的内在属性。域语义元素集记为 $\boldsymbol{U}_{domain}=\{(\mu_b^n,\sigma_b^n)\}_{n=1}^{N}\in\mathbb{R}^{N\times d_v}$，其中 d_v 为视觉特征维度。μ_b^n 和 σ_b^n 分别表示刻画域信息的均值和标准差统计量，

© 中国图象图形学报版权所有

用于捕获与图像域信号相关的风格变化，如光照条件、相机设置和图像质量等。

通过将模糊的“真实/伪造”和复杂的“域信息”分解成更基础、更明确的语义元素后，本方法可以使用这些语义元素来理解类别概念和域概念，进一步为样本生成最适配的视觉提示与文本提示。

1.2 自适应提示生成

为生成更适配每个样本的特定提示，本方法引入自适应提示生成（Adaptive Prompts Generation，APG）模块。该模块以实例元素为参考，基于注意力机制选择并融合相关语义元素，从而为每个输入生成最合适的提示。

1.2.1 实例元素

本方法分别在文本端和视觉端引入实例元素，用于引导可学习语义元素进行动态选择与融合，使生成的语义提示更适配当前样本的个性化特征。

对于文本端，除了人脸活体检测任务中常用的固定文本提示L_t，如“a photo of a <live/fake>”以外，本方法还引入了内容描述提示L_c，例如“a short haired man wearing glasses has a distorted color image”。对文本提示L_t和L_c进行输入预处理，并在文本序列的首尾添加起始标记[SOS]和结束标记[EOS]。将经过文本嵌入层映射后的序列输入编码器，生成基础文本表示$\boldsymbol{P}_t$和$\boldsymbol{P}_c$，如式(1)-(2)所示。并对内容描述提示L_c取[EOS]标记位置经过层归一化后的输出$\boldsymbol{C}_\mu$作为文本端的实例元素。

$$\boldsymbol{P}_t = \mathrm{Transformer}(L_t)\#(1)$$

$$\boldsymbol{P}_c = \mathrm{Transformer}(L_c)\#(2)$$

对于输入人脸图像$\boldsymbol{I}$，首先通过Transformer编码器得到图像块嵌入$\boldsymbol{E}$，然后利用AdaIN（Huang 等，2017）操作计算图像块嵌入$\boldsymbol{E}$的均值$\boldsymbol{E}_\mu$和标准差$\mathbf{E}_\sigma$，并将它们作为视觉端的实例元素，计算公式如式(3)-(5)所示：

$$\boldsymbol{E} = \mathrm{Transformer}(\boldsymbol{I})\#(3)$$

$$\boldsymbol{E}_\mu = \frac{1}{\sqrt{H}\times\sqrt{W}}\sum_{h=1}^{\sqrt{H}}\sum_{w=1}^{\sqrt{W}}\boldsymbol{E}\#(4)$$

$$\boldsymbol{E}_\sigma = \sqrt{\frac{1}{\sqrt{H}\times\sqrt{W}}\sum_{h=1}^{\sqrt{H}}\sum_{w=1}^{\sqrt{W}}(\boldsymbol{E}-\boldsymbol{E}_\mu)^2+\epsilon}\#(5)$$

式中H和W分别表示预处理后输入人脸图像的高度和宽度。ϵ=10^{-6}，为极小常数，用来保证计算稳定性。均值和标准差能够直接反映图像亮度、对比度、色偏等风格统计信息。通过这两个统计量，模型能捕捉单张人脸图像的域相关信息。相比域嵌入、风格编码、对抗特征等方法，通过AdaIN直接从输入样本中获得均值和标准差，不用依赖额外域标签，也不会引入额外的对抗训练增加不稳定性。

1.2.2 提示生成

以内容描述提示的文本特征$\boldsymbol{C}_\mu$和图像块特征的均值$\boldsymbol{E}_\mu$、标准差$\boldsymbol{E}_\sigma$分别作为用于文本端和视觉端参考的实例元素，通过基于注意力机制的自适应提示生成模块动态选择语义元素，并将其加权结合成与当前样本最匹配的类别提示和域提示，如图2所示。生成的提示以层级结构嵌入到文本和视觉模态的序列中，引导模型捕捉不同的类别和域语义特征，从而解决语义模糊问题。

对于文本分支，内容描述提示的文本特征$\boldsymbol{C}_\mu$和类别语义元素集$\boldsymbol{U}_{class}$通过APG模块挑选与当前样本最匹配的类别语义元素，从而生成类别语义提示$\boldsymbol{P}_t'$。具体过程如式(6)-(9)所示：

$$[\boldsymbol{Q}_t,\boldsymbol{K}_t,\boldsymbol{V}_t] = [\boldsymbol{C}_\mu\boldsymbol{W}_t^q,\boldsymbol{U}_{class}\boldsymbol{W}_t^k,\boldsymbol{U}_{class}\boldsymbol{W}_t^v]\#(6)$$

$$A_t = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_t(\boldsymbol{K}_t)^{\mathrm{T}}}{\sqrt{d_t}}\right)\#(7)$$

$$I_t = \mathrm{argmax}(A_t,\mathrm{dim}=-1)\#(8)$$

$$\boldsymbol{P}_t' = \boldsymbol{U}_{class}[I_t]\#(9)$$

式中，$\boldsymbol{Q}_t$, $\boldsymbol{K}_t$, $\boldsymbol{V}_t$分别表示查询向量、键向量和值向量，$\boldsymbol{W}_t^q$, $\boldsymbol{W}_t^k$和$\boldsymbol{W}_t^v$为投影矩阵；Softmax(·)表示softmax函数，它应用于所有行以确保每个元素在空间位置上的注意力总和等于1。$\sqrt{d_t}$为缩放因子，目的是防止点积结果随着维度增大而过大，从而避免Softmax输出过于集中，保证训练的稳定性。A_t表示当前样本与各类别语义元素之间的匹配权重。$\mathrm{argmax}(A_t,\mathrm{dim}=-1)$表示在最后一个维度上选取注意力权重最大的元素索引，对应当前样本最需要的类别语义解释。将索引I_t对应的语义元素作为最终的类别语义提示$\boldsymbol{P}_t'$，因此$\boldsymbol{P}_t'$不再是对所有样本共享的固定提示，而是由实例元素为参考产生的样本特定类别提示。

$\boldsymbol{P}_t'$与固定文本提示$\boldsymbol{P}_t$在文本分支进行拼接，生成优化的类别提示P_t''。如式(10)所示：

$$\boldsymbol{P}_t'' = \boldsymbol{P}_t + \boldsymbol{P}_t'\#(10)$$

对于视觉分支，以实例元素$\boldsymbol{E}_\mu$和$\boldsymbol{E}_\sigma$为指导，结

© 中国图象图形学报版权所有

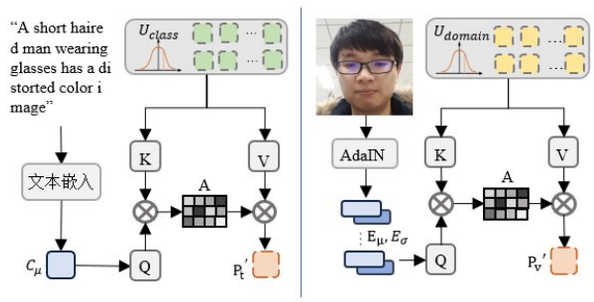


图2　自适应提示生成模块的结构

Fig. 2　Structure of the Adaptive Prompt Generation Module

合域语义元素集$\boldsymbol{U}_{domain}$，通过APG模块分别生成域语义提示$\boldsymbol{P}_\mu$和$\boldsymbol{P}_\sigma$。以生成提示$\boldsymbol{P}_\mu$为例，将图像特征均值$\boldsymbol{E}_\mu$和域语义元素$\boldsymbol{U}_{domain}$通过可学习参数$\boldsymbol{W}_i^q$，$\boldsymbol{W}_i^k$和$\boldsymbol{W}_i^v$分别投影，以获得查询向量$\boldsymbol{Q}_i$，键向量$\boldsymbol{K}_i$和值向量$\boldsymbol{V}_i$。然后，采用多头注意力机制计算注意力矩阵A_i和提示$\boldsymbol{P}_\mu$。以上过程如式(11)-(13)所示：

$$\left[\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i\right]=\left[\boldsymbol{E}_\mu\boldsymbol{W}_i^q,\boldsymbol{U}_{domain}\boldsymbol{W}_i^k,\boldsymbol{U}_{domain}\boldsymbol{W}_i^v\right]\#(11)$$

$$A_i=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_i(\boldsymbol{K}_i)^{\mathrm{T}}}{\sqrt{d_v}}\right)\#(12)$$

$$\boldsymbol{P}_\mu=A_i\cdot\boldsymbol{V}_i\#(13)$$

式中，$\boldsymbol{W}_i^q$，$\boldsymbol{W}_i^k$和$\boldsymbol{W}_i^v$是线性投影矩阵。注意力矩阵A_i实际上包含了每个域语义元素与实例元素的关联系数。将注意力矩阵A_i作用于向量$\boldsymbol{V}_i$，根据可学习语义元素和实例元素的相关性自适应分配权重，生成最适配当前样本域信息的提示$\boldsymbol{P}_\mu$。$\boldsymbol{P}_\sigma$通过类似的方法对$\boldsymbol{E}_\sigma$进行处理获得。接着，将$\boldsymbol{P}_\mu$和$\boldsymbol{P}_\sigma$进行拼接得到域语义提示$\boldsymbol{P}_v'$，如式(14)所示：

$$\boldsymbol{P}_v'=\boldsymbol{P}_\mu+\boldsymbol{P}_\sigma\#(14)$$

$\boldsymbol{P}_v'$与基础提示嵌入$\boldsymbol{E}$在视觉分支进行拼接，生成优化的域提示$\boldsymbol{P}_v''$。如式(15)所示：

$$\boldsymbol{P}_v''=\boldsymbol{E}+\boldsymbol{P}_v'\#(15)$$

1.3　多模态交互适配

在视觉-语言预训练模型向下游任务微调的过程中，使用适配器是一种常见且有效的方法。然而，早期适配器多为单模态设计，仅针对视觉或文本模态单独优化，忽略了模态间的关联信息。最近，多模态适配器方法(Yang 等，2024；Seputis 等，2024；Ghiasvands 等，2025)通过在视觉-语言模态分支插入适配器，在通用图像识别领域展现出良好的效果。Yang等人(2024)提出多模态适配器包含模态特定的上下投影层和全局共享投影层，以增强视觉与语言分支表征的对齐。MMA(Seputis 等，2024)基于多头注意力网络，通过嵌入下采样层、掩码多头注意力层和上采样层，将CLIP的文本与图像嵌入联合适配。pFedMMA(Ghiasvands 等，2025)通过模态特定投影层与全局共享投影层进行非对称优化也为跨域适配提供了参考。

为进一步提升人脸活体检测模型的域泛化能力，本文在文本编码器和图像编码器的Transformer层中通过残差连接插入分层多模态适配器(Hierarchical Multimodal Adapter，HM-Adapter)，保留模态特有知识的同时促进跨模态特征对齐，从而提升人脸活体检测模型在跨域场景下的鲁棒性与识别能力。本文提出的分层多模态适配器具体细节如图3所示。

对于图像编码器和文本编码器的每个Transformer层，首先将图像特征和文本特征分别通过线性层从原始维度降至瓶颈维度，再通过一个共享的中间层进行特征对齐，最后升维回原始维度。共享的中间层由线性层和ReLU组成，输入输出维度均为中间瓶颈维度。适配器模块实现如式(16)所示：

$$A(\mathrm{x})=\boldsymbol{W}^{up}\cdot\phi\left(\boldsymbol{W}^{share}\cdot\phi\left(\boldsymbol{W}^{down}(\boldsymbol{x})\right)\right)\#(16)$$

式中，$\boldsymbol{W}^{down}\in\mathbb{R}^{d\times r}$，$\boldsymbol{W}^{up}\in\mathbb{R}^{r\times d}$，$r$为瓶颈维度，本文设为64。$d$是特征维度。$\boldsymbol{W}^{share}$是共享投影层，在文本和图像两个分支之间共享。$\phi$是非线性激活函数ReLU。

1.4　损失函数

在模型训练过程中，文本和图像编码器的主干参数被冻结，而语义元素集、自适应提示向量与多模态适配器是可更新的。

1.4.1　约束损失

对于用来表示类别语义与域语义的可学习语义元素集，每个语义元素被要求可以独立表示一个语义信号，从而增强模型对不同语义的理解能力。这通过不相似性损失实现。以域语义元素集$\boldsymbol{U}_{domain}$的均值统计量$\boldsymbol{\mu}_b^n$为例，即通过最小化$\boldsymbol{\mu}_b^i(i\in[0,N])$与$\boldsymbol{\mu}_b^j(j\in[0,N],\ i\neq j)$的余弦相似度$<\cdot,\cdot>$来增大两者距离，计算如式(17)所示：

$$\mathcal{L}_{\mathrm{dis}}=\left\langle\frac{\boldsymbol{\mu}_b^i}{\left(\boldsymbol{\mu}_b^i\right)_2},\frac{\boldsymbol{\mu}_b^j}{\left(\boldsymbol{\mu}_b^j\right)_2}\right\rangle\#(17)$$

式中$/\!/\cdot/\!/_2$表示L_2范数。对域语义元素集的标准

© 中国图象图形学报版权所有

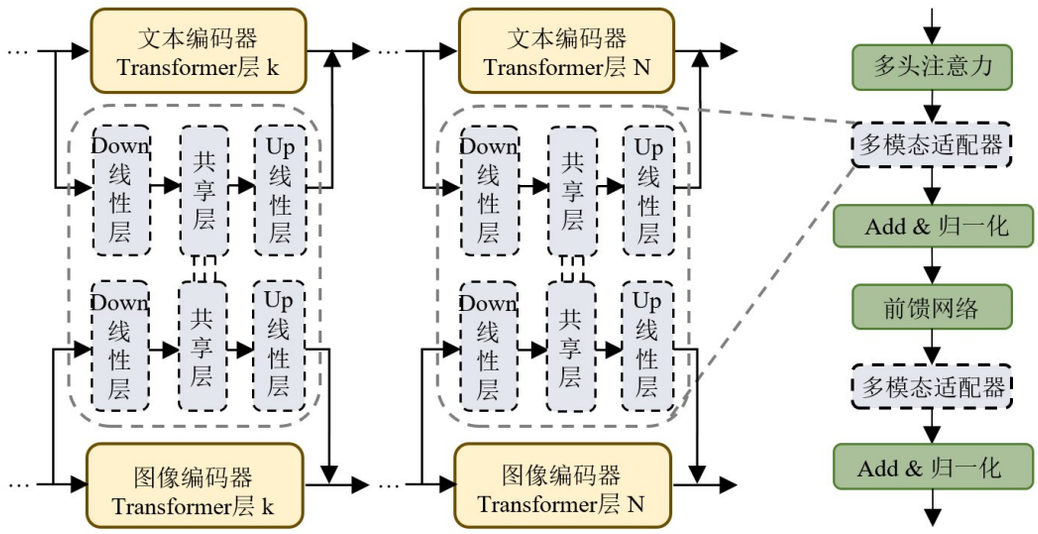


图3　分层多模态适配器的结构与相对位置

Fig. 3　Structure and relative position of the hierarchical multimodal adapter

差统计量与类别语义元素也进行同样的计算，并将它们相加得到总的不相似性损失。同时，本文使用最大均值差异（Maximum Mean Discrepancy，MMD）损失来对齐原始嵌入向量与自适应语义提示的分布，使这两个分布的差异尽可能小。如式(18)所示：

$$\mathcal{L}_{\mathrm{MMD}}=\frac{1}{L}\sum_{i=1}^{L}\left[\kappa\left(\boldsymbol{\mu}_i,\boldsymbol{p}_{\mu}^{i}\right)+\kappa\left(\boldsymbol{\sigma}_i,\boldsymbol{p}_{\sigma}^{i}\right)\right]\#(18)$$

式中，$\kappa(\cdot)$为多核MMD的核函数，用于计算两个样本在再生核希尔伯特空间中的相似性。$\boldsymbol{\mu}_i$、$\boldsymbol{\sigma}_i$分别表示原始特征的均值和标准差，$\boldsymbol{p}_{\mu}^{i}$、$\boldsymbol{p}_{\sigma}^{i}$表示生成的自适应语义提示。$L$表示Transformer层数。文本嵌入向量与类别语义提示的分布对齐同样使用MMD损失，然后相加得到总的MMD损失。

1.4.2　分类损失

分类损失函数使用标准的交叉熵损失，如式(19)所示：

$$\mathcal{L}_{\mathrm{ce}}=-\sum_{i=1}^{n}y_i\log\hat{y}_i\#(19)$$

式中n为类别数量，y_i为类别真值，$\hat{y}_i$为类别的预测概率值。

1.4.3　总损失

本文将分类损失和约束损失相结合，形成最终损失，如式(20)所示：

$$\mathcal{L}_{\mathrm{total}}=\mathcal{L}_{\mathrm{ce}}+\lambda 1\mathcal{L}_{\mathrm{dis}}+\lambda 2\mathcal{L}_{\mathrm{MMD}}\#(20)$$

式中$\lambda 1=1,\lambda 2=0.001$是损失函数的平衡因子，取值依据为多次超参数实验优化结果。当$\lambda 1$过小时，不同语义元素之间的区分性不足；当$\lambda 2$过大时，过强的分布对齐会抑制分类判别能力。

2　实验与分析

2.1　实验数据集与评估方法

本文使用常用于人脸活体检测及其域泛化场景的4个数据集进行评估，包括OULU-NPU（O）、CASIA-FASD（C）、Idiap-Replay Attack（I）和MSU-MFSD（M）。OULU-NPU数据集包含55位参与者录制的4950段高分辨率视频样本，涵盖多种光照和环境条件，能够有效测试模型在真实场景下的泛化能力；CASIA-FASD数据集由50位被试者的平板和摄像机拍摄的600段视频组成，包含打印照片攻击和视频重放攻击，用于验证模型对不同伪造手段的识别能力；Idiap-Replay Attack数据集收集了50名参与者的1300段视频，覆盖室内光照变化和低分辨率视频场景，同时包含打印和视频重放攻击，能够考察模型在低质量输入下的稳健性；MSU-MFSD数据集则包含55名参与者在不同拍摄条件下录制的440段视频，提供了35名参与者的公开视频样本，采集自不同摄像设备并包含多角度拍摄的视频，为跨设备和跨场景的泛化能力评估提供了丰富的样本。在实验中，本文将每个数据集视为一个域，采用留一法来评估所提方法，即A&B→C表示在A和B的组合数据集上作为源域进行训练，并在C上作为未见的目标域进行测试。

© 中国图象图形学报版权所有

采用两个标准评价指标：半总错误率(Half-Total Error Rate，HTER)和曲线下面积(Area Under the Curve，AUC)。

HTER是错误接受率(False Acceptance Rate，FAR)和错误拒绝率(False Rejection Rate，FRR)的平均值，衡量系统错误分类欺骗攻击和真实人脸的比例。HTER的值越小，表明模型的性能越好。HTER计算如式(21)-(23)所示：

$$FAR = \frac{FP}{FP + TN} \#(21)$$

$$FRR = \frac{FN}{TP + FN} \#(22)$$

$$HTER = \frac{FAR + FRR}{2} \#(23)$$

式中，FAR是指将非目标样本错误地识别为目标样本的概率，FRR是指将目标样本错误地识别为非目标样本的概率。False Positive(FP)表示模型将伪造人脸样本预测为活体人脸样本的数量；True Negative(TN)表示模型将伪造人脸样本预测为伪造人脸样本的数量；True Positive(TP)表示模型将活体人脸样本预测为活体人脸样本的数量；False Negative(FN)表示模型将活体人脸样本预测为伪造人脸样本的数量。

ROC曲线是根据不同的分类阈值，以真阳性率(True Positive Rate，TPR)为纵坐标、假阳性率(False Positive Rate，FPR)为横坐标绘制出的曲线，该曲线绘制了不同阈值下的真阳性率与假阳性率。AUC是ROC曲线下的面积，用来衡量模型的判别能力，取值范围为[0，1]，值越大表示模型的分类能力越强。

2.2 实验设置

本文实验操作使用PyTorch深度学习框架，采用编程语言为Python3.9，并在单个NVIDIA GTX 2080 Ti GPU上进行训练和测试。输入图像使用MTCNN算法(Zhang等，2016)检测人脸，裁剪并调整大小为224×224。本文使用的图像编码器是CLIP预训练的ViT-B/16，图像表示的维度为d_v=768。文本编码器采用Transformer，将77作为文本标记的最大数量，文本向量维度为d_t=512。联合嵌入空间的维度d_{vt}=512。训练过程中的优化器使用SGD，初始学习率为0.0001，设置的总训练轮数为100轮。对于模型的超参数，语义元素集的元素数量N设置为6，可学习提示的长度设为2，提示深度为12。

2.3 域泛化人脸活体检测对比试验

2.3.1 多源域跨数据集对比试验

在本小节实验中，随机选择1个数据库作为测试集，其余3个数据库作为训练集。具体来说，OCI→M表示在OULU-NPU、CASIA-FASD、Idiap-Replay Attack上训练，在MSU-MFSD上测试。OMI→C、OCM→I和ICM→O的定义类似。表1报告了本文提出的MMPL模型在四个跨域场景中执行域泛化评估的结果，并与当前一些先进算法进行了比较。

比较的方法分为两类。第一类为视觉单模态域泛化方法，主要通过对抗学习、元学习、风格投影或特征对齐来提升模型跨域泛化能力，但缺少显式的文本语义辅助。第二类为视觉-语言多模态方法，通过自然语言描述、条件提示或可学习提示增强视觉表示。其中InstructFLIP(Lin等，2025)通过内容与风格指令解耦缓解跨域训练冗余；SG-CLIP(Liang等，2025)利用语义约束增强跨域表示；AIM-FAS(Cheng等，2026)引入自适应策略以提升不平衡数据条件下的跨域鲁棒性。本文方法与这些工作相比，更强调类别语义与域语义的显式解耦，构建类别语义元素集和域语义元素集，并利用实例参考元素为每个样本动态组装特定提示，提升模型的域泛化能力。

从表1可以看出，本文所提MMPL模型在四个域泛化典型场景OCI→M、OMI→C、OCM→I、ICM→O及平均的性能指标上均取得了不错的性能，AUC分别达到99.77%、99.99%、97.66%、99.67%，平均AUC为99.27%，HTER分别为1.67%、0.11%、7.07%、1.70%，平均HTER为2.64%，四个协议的平均HTER为最佳结果。可以看出，表现最佳的视觉单模态方法S-Adapter的平均HTER为6.03%，本文所提MMPL模型与其相比，平均HTER降低了3.39个百分点，说明仅依赖视觉特征增强，仍难以充分建模跨域活体检测中复杂的类别差异与域差异，导致模型在跨域场景中泛化能力不足。与同样引入视觉-语言先验的多模态方法CFPL-FAS和S-CPTL相比较，CFPL-FAS和S-CPTL在实验中平均HTER分别为3.60%和3.33%，而本文所提MMPL模型将平均HTER进一步降低了0.96%和0.69%。分析认为与现有视觉-语言多模态方法相比，MMPL的优势并不单纯来自引入可学习提示，而在于提示构造方式的不同。现有普通可学习提示方法通常直接优化一

© 中国图象图形学报版权所有

表1　本文方法在多源域跨数据集上的比较结果

Table 1　Comparison results of the proposed methods across datasets in multiple source domains

方法	OCI→M		OMI→C		OCM→I		ICM→O		Avg.
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)
MADDG(Shao 等,2019)	17.69	88.06	24.50	84.51	22.19	84.99	27.98	80.02	23.09
SSDG(Jia 等,2020)	7.38	97.17	10.44	95.94	11.71	96.59	15.61	91.54	11.28
D^2AM(Chen 等,2021)	12.70	95.66	20.98	85.58	15.43	91.22	15.27	90.87	16.09
SSAN(Wang 等,2022)	6.67	98.75	10.00	96.67	8.88	96.79	13.72	93.63	9.81
PatchNet(Wang 等,2022)	7.10	98.46	11.33	94.58	13.40	95.67	11.82	95.07	10.91
AMEL(Zhou 等,2022)	10.23	96.62	11.88	94.39	18.60	88.79	11.31	93.96	13.01
IADG(Zhou 等,2023)	5.41	98.19	8.70	96.44	10.62	94.50	8.86	97.14	8.39
S-Adapter(Cai 等,2024)	3.43	99.50	6.32	97.82	7.16	97.61	7.21	98.00	6.03
TTDG-V(Zhou 等,2024)	4.16	98.48	7.59	98.18	9.62	98.18	10.00	96.15	7.84
VL-FAS(Fang 等,2024)	3.13	99.31	4.00	98.64	5.00	**98.90**	7.92	97.05	5.01
S-CPTL(Guo 等,2024)	**1.43**	99.17	0.89	99.00	6.86	98.63	4.12	99.02	3.33
CFPL-FAS(Liu 等,2024)	3.09	99.45	2.56	99.10	5.43	98.41	3.33	99.05	3.60
CCPE(Guo 等,2025)	3.10	99.21	1.33	99.36	6.08	94.36	5.57	98.49	4.02
SG-CLIP(Liang等,2025)	4.05	99.62	2.93	98.50	**4.21**	98.59	3.26	99.31	3.61
IustructFLIP(Lin等,2025)	5.52	98.12	1.47	99.79	9.12	96.17	14.33	94.79	7.61
AIM-FAS(Cheng等,2026)	3.42	99.53	3.02	99.33	6.27	98.61	7.99	98.80	5.16
MMPL(Ours)	1.67	**99.77**	**0.11**	**99.99**	7.07	97.66	**1.70**	**99.67**	**2.64**

注:加粗字体为每列最优值,下划线标出的数值为次优值。

组统一的提示向量,或基于文本模板、内容提示、风格提示对提示标记进行整体更新,其本质仍是对提示标记本身进行参数化学习,因此对类别语义与域语义的表达往往耦合在同一空间内,难以显式刻画类别概念与域属性之间的差异。相比之下,MMPL分别为类别语义和域语义建立独立的语义元素集,并利用实例参考元素通过注意力机制自适应选择和融合相关元素,从而生成针对当前样本的类别提示与域提示。在实验OCM→I中,本文模型的表现结果相对略差。造成这一现象的主要原因在于O、C、M三个源域均在可控光照或稳定帧率下采集,而I域的采集光照变化剧烈,进一步加剧了跨域分布偏移。因此,本文模型在I域测试时受到数据域特性的一定制约,这也提示未来可以进一步探索在更复杂采集条件下的域泛化机制,以提升模型在实际复杂环境中的稳定性与适应性。

2.3.2　受限源域跨数据集对比试验

为深入探究所提方法的泛化性能,本小节通过限制源域数量展开进一步评估。选取分布特性差异显著的MSU-MFSD和Idiap-Replay Attack数据集作为源域。同时,分别将CASIA-FASD和OULU-NPU数据集作为测试域,由此构建出MI→C和MI→O两个测试任务。表2展示了本文方法和多种方法在两个测试任务上的实验结果。

从表2中可以看出,在受限源域跨数据集设置下,本文所提MMPL模型在两个测试任务上均取得了最优结果。其中,在MI→C和MI→O任务上,HTER分别达到5.33%和6.11%,AUC分别达到98.02%和98.18%,平均HTER为5.72%,达到最佳。这表明在仅使用MSU-MFSD和Idiap-Replay Attack两个源域进行训练的条件下,MMPL模型仍能够学习到具有较强泛化能力的判别特征,并在未见目标域上保持稳定性能,验证了本文使用可学习的语义元素来组合类别提示和域提示的有效性。与早期基于对抗学习和元学习的方法MADDG和D^2AM相比,本文方法在HTER和AUC两个指标上均表现

© 中国图象图形学报版权所有

表2　本文方法在受限源域跨数据集上的比较结果

Table 2　Comparison results of the proposed method across datasets in the restricted source domain

方法	MI→C		MI→O		Avg.
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)
MADDG(Shao 等,2019)	41.02	64.33	39.35	65.10	40.19
SSDG(Jia 等,2020)	31.89	71.29	36.01	66.88	33.95
D^2AM(Chen 等,2021)	32.65	72.04	27.70	75.36	30.18
CLIP(Radford 等,2021)	15.42	93.51	15.31	90.30	15.37
CLIP-V(Radford 等,2021)	13.56	93.74	12.97	91.09	13.27
SSAN(Wang 等,2022)	30.00	76.20	29.44	76.62	29.72
IADG(Zhou 等,2023)	24.07	85.13	18.47	90.49	21.27
AMEL(Zhou 等,2022)	23.23	85.17	19.68	87.01	21.46
S-Adapter(Cai 等,2024)	17.93	89.56	19.76	86.87	18.85
TTDG-V(Zhou 等,2024)	17.77	86.69	17.70	90.09	17.77
MMPL(Ours)	**5.33**	**98.02**	**6.11**	**98.18**	**5.72**

注:加粗字体为每列最优值。

出显著优势,说明引入视觉-语言协同建模有助于提升跨域活体检测性能。与TTDG-V只进行视觉方面的风格投影相比,本文所提模型在类别与域两方面改善语义模糊问题,其平均HTER较TTDG降低了12.05%,模型具有良好的泛化性。进一步地,与多模态基础模型CLIP进行比较,本文方法仍取得更优结果,表明所构建的可学习语义元素集及提示生成机制能够更有效地缓解类别语义与域语义的模糊问题,从而增强模型在受限源域条件下的跨域泛化能力。

2.4　消融试验

2.4.1　各个组件的有效性

本小节实验在基础模型IVLP(Rasheed 等,2022)上逐步引入所提出的各个模块,以衡量所提方法中各个组件的有效性。其中IVLP是一种以CLIP为基础的多模态提示学习方法,仅组合深层文本提示与视觉提示,两分支间不进行协同作用。在本文中,为IVLP配备基于类别的人工提示以方便进行实验验证。实验评估的组成部分是类别语义元素集$\boldsymbol{U}_{class}$、域语义元素集$\boldsymbol{U}_{domain}$、自适应提示生成模块APG和分层多模态适配器HM-Adapter。用"√"表示在模型中使用该模块,"×"表示不使用该模块。在ICM→O协议上观察实验结果,如表3所示。由于语义元素本身不具备自适应生成提示的能力,只有与APG模块联合才能验证其对提示优化的真实贡献,因此本小节将二者联合进行消融实验。

表3　模型各个组件消融实验(ICM→O)

Table 3　Ablation experiment of each component of model (ICM→O)

基础模型	U_{class}	U_{domain}	APG	HM-Adapter	HTER	AUC
√	×	×	×	×	3.05%	99.39%
√	√	×	√	×	2.73%	99.47%
√	×	√	√	×	2.65%	99.53%
√	√	√	√	×	2.57%	99.61%
√	√	√	√	√	**1.70%**	**99.67%**

注:加粗字体为每列最优值。

2.4.2　不同损失函数的有效性

为验证各个损失函数的作用,在ICM→O协议的跨域场景下逐步引入各个损失函数。实验结果如表4所示。仅使用分类损失$\mathcal{L}_{cls}$时,模型HTER为3.05%。在此基础上引入不相似性损失$\mathcal{L}_{dis}$,HTER降至2.49%,这说明不相似性损失能够有效确保每个语义元素的独特性,避免信息重叠,提升了特征空间的表达能力。进一步引入$\mathcal{L}_{\mathrm{MMD}}$可以看出HTER得到更显著的改善,达到1.70%。这不仅表明分布对齐能够有效缩小不同语义特征的分布差距,增强模

© 中国图象图形学报版权所有

型对跨域数据的泛化能力，也证实了语义元素多样化与分布对齐的协同作用，能够同时优化类别判别性与域适应性。

2.4.3　语义元素集中的不同元素数量的有效性

本文构建的可学习的语义元素集包含N个语义元素，如图4所示，在ICM→O协议设置下对参数N进行探索实验，值从2到12，按步长为1均匀增加。实验结果表明，元素数量过少时，模型难以全面捕捉语义特征的多样性，导致模型泛化能力差；元素数量过多时，会增加不必要的复杂性，降低模型性能。经过综合考虑，本文最终将语义元素的个数N在所有实验中设置为6。

表4　不同损失函数的约束性分析（ICM→O）

Table 4　Constraint analysis of different loss functions（ICM→O）

$\mathcal{L}_{cls}$	$\mathcal{L}_{dis}$	$\mathcal{L}_{\mathrm{MMD}}$	HTER	AUC
√	×	×	3.05%	99.34%
√	√	×	2.49%	99.54%
√	√	√	**1.70%**	**99.67%**

注：加粗字体为每列最优值。

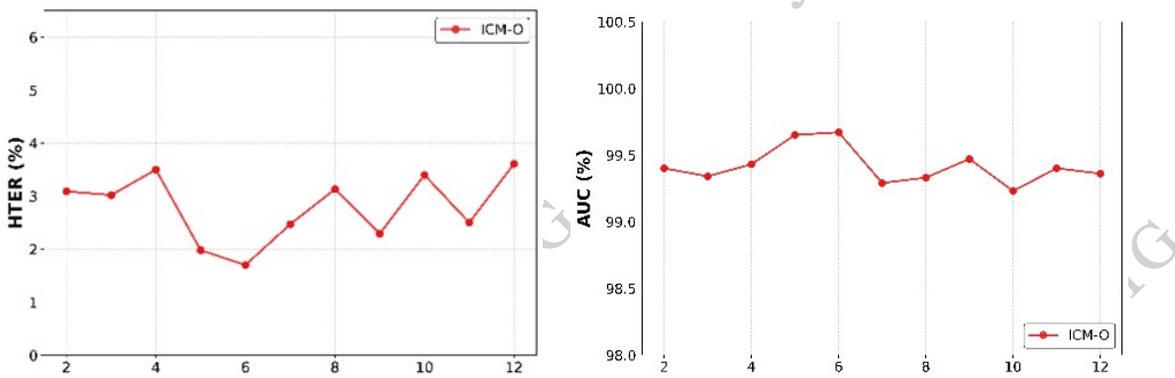


图4　语义元素数量N对HTER和AUC指标的影响研究

Fig. 4　Influence of the number of semantic elements N on HTER and AUC indicators

2.5　可视化分析

为了直观评估本文方法对人脸关键区域的聚焦能力，如图5所示，采用注意力模型可解释性工具（Chefer H等，2021）对输入图像进行视觉分析。图中第一行表示输入图像，第二行是使用基础模型的注意力图，第三行是本文所提方法展示的注意力图。可以看出，基础模型在检测人脸伪造痕迹方面存在一定缺陷，容易受到背景或者无关纹理的干扰。相比之下，本文提出的MMPL模型能够引导注意力聚焦在更具判别性的区域。比如在OMI→C协议下，使用本文方法，对于伪造图像，屏幕的边界区域是模型关注的区域；在OCI→M协议设置下，模型重点关注的区域是伪造人脸出现异常反光的区域。这证明了所提方法通过引入可学习语义元素并生成特定提示，能够引导模型关注更具判别力的语义区域，增强了模型的域泛化能力。

3　结　论

本文提出了一种面向域泛化人脸活体检测的多模态提示学习方法。该方法通过构建类别语义元素集和域语义元素集，使用可学习元素对人脸活体检测任务的真实或伪造类别语义与跨域环境语义进行表征。接着使用基于注意力机制的自适应提示生成模块动态选择并组合这些语义元素，为每个样本实例生成最适配的类别提示与域提示，在提升模型泛化性能的同时增强语义理解能力。最后引入分层多模态适配器对齐视觉特征与文本特征，进一步提高模型的判别性和泛化性。为验证所提方法的有效性，本文在常用于人脸活体检测研究的公开数据集上进行跨域测试与消融实验。实验结果表明，该方法能够有效缓解人脸活体检测中的语义模糊问题，并在多个评估协议中取得了领先表现。未来工作中，计划进一步探索域泛化人脸活体检测模型在复杂环境下的自适应能力，结合域自适应或一致性约

© 中国图象图形学报版权所有

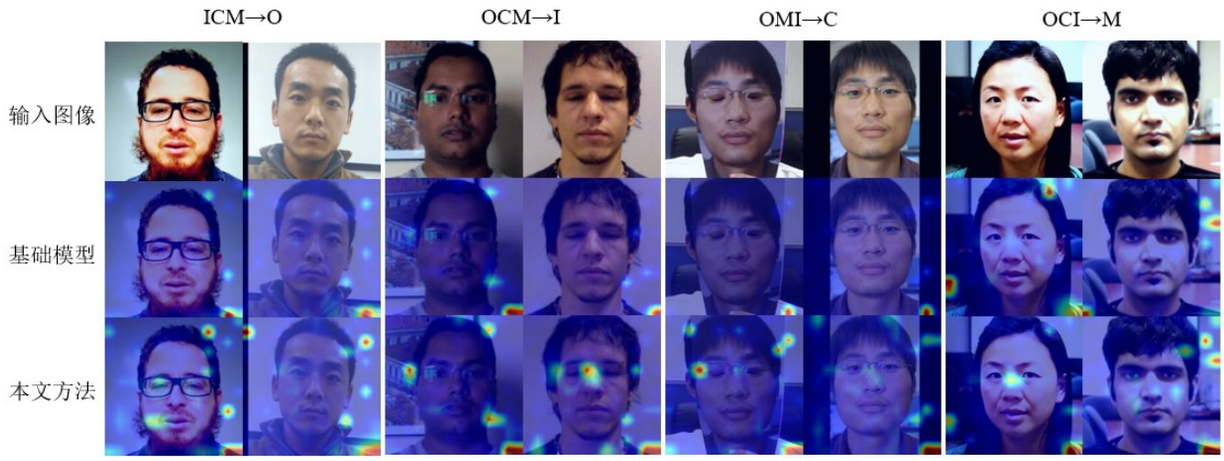


图5　模型的注意力可视化展示图

Fig. 5　Attention visualization of the model

束机制增强模型在极端光照、低分辨率和传感器偏移场景中的稳定性。同时结合轻量化设计，利用结构化剪枝、低秩分解与量化等方式压缩模块设计，以降低计算开销并提升其在实际移动端或嵌入式设备中的部署能力。

参考文献(References)

Cai R Z, Yu Z T, Kong C Q, Li H L, Chen C S, Hu Y J, et al. 2024. S-Adapter: Generalizing Vision Transformer for Face Anti-Spoofing With Statistical Tokens. IEEE Transactions on Information Forensics and Security, 19: 8385-8397 [DOI: 10.1109/TIFS.2024.3420699]

Chefer H, Gur S and Wolf L. 2021. Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers//Proceedings of the IEEE/CVF International Conference on Computer Vision. [DOI: 10.1109/ICCV48922.2021.00045]

Chen Z H, Yao T P, Sheng K K, Ding S H, Tai Y, Li J L, et al. 2021. Generalizable Representation Learning for Mixture Domain Face Anti-Spoofing. Proceedings of the AAAI Conference on Artificial Intelligence, 35(2): 1132-1139 [DOI:10.1609/aaai.v35i2.16199]

Cheng F, Qiao Y Z, Meng F J, Wang X L, Peng M S, Li K X, Wu Z Z and Chen M W. 2026. Vision-language adaptation with imbalance mitigation for generalizable face anti-spoofing//Pattern Recognition. : 113101 [DOI:10.1016/j.patcog.2026.113101]

Fang H, Liu A J, Jiang N, Lu Q, Zhao G Q and Wan J. 2024. VL-FAS: Domain Generalization via Vision-Language Model For Face Anti-Spoofing//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). : 4770-4774 [DOI:10.1109/ICASSP48485.2024.10448156].

Ghiasvand S, Alizadeh M and Pedarsani R. 2025. pFedMMA: Personalized Federated Fine-Tuning with Multi-Modal Adapter for Vision-Language Models[EB/OL].[2025-07-07]. https://arxiv.org/pdf/2507.05394.pdf

Guo J B, Liu A J, Diao Y F, Zhang J, Ma H, Zhao B. et. al 2025. Domain Generalization for Face Anti-spoofing via Content-aware Composite Prompt Engineering. IEEE Transactions on Multimedia [DOI:10.1109/TMM.2025.3618575]

Guo J B, Liu H, Luo Y Z, Hu X L, Zou H, Zhang Y. et. al 2024. Style-conditional Prompt Token Learning for Generalizable Face Anti-spoofing//Proceedings of the 32nd ACM International Conference on Multimedia (MM'24). New York, NY, USA: Association for Computing Machinery: 994-1003[DOI:10.1145/3664647.3680857]

Huang X and Belongie S. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization//Proceedings of the IEEE International Conference on Computer Vision (ICCV). [DOI: 10.1109/ICCV.2017.167]

Jia C, Yang Y, Xia Y, Chen Y T, Parekh Z, Pham H. et. al 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision//Proceedings of the 38th International Conference on Machine Learning. PMLR, 139: 4904-4916 [DOI: 10.48550/arXiv.2102.05918]

Jia Y P, Zhang J, Shan S G and Chen X L. 2020. Single-Side Domain Generalization for Face Anti-Spoofing//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). : [DOI:10.1109/CVPR42600.2020.00851]

Li L, Feng X Y, Boulkenafet Z, Xia Z Q, Li M M and Hadid A. 2016. An original face anti-spoofing approach using partial convolutional neural network//Proceedings of the 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA). : 1-6 [DOI:10.1109/IPTA.2016.7821013]

Liang F M, Zhang J, Jia Y L, Ma H and Liang Y Y. 2025. Domain Generalization in Face Anti-Spoofing Based on Vision-Language Semantic Awareness//Chinese Conference on Biometric Recognition (CCBR). : 504-514 [DOI:10.1007/978-981-95-6123-0_47]

© 中国图象图形学报版权所有

Lin K H, Tseng Y W, Huang K Y, Wu J C and Cheng W H. 2025. InstructFLIP: Exploring Unified Vision-Language Model for Face Anti-spoofing//Proceedings of the 33rd ACM International Conference on Multimedia (MM). : 2987-2996 [DOI: 10.1145/3746027.3754939]

Liu A J, Xue S, Gan J W, Wan J, Liang Y Y, Deng J K.et.al 2024. CFPL-FAS: Class Free Prompt Learning for Generalizable Face Anti-spoofing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). : [DOI: 10.1109/CVPR52733.2024.00029]

Patel K, Han H and Jain A K. 2016. Secure Face Unlock: Spoof Detection on Smartphones. IEEE Transactions on Information Forensics and Security, 11 (10) : 2268-2283 [DOI: 10.1109/TIFS. 2016.2578288]

Pereira T D F, Anjos A, De Martino J M and Marcel S. 2013. LBP-TOP Based Countermeasure against Face Spoofing Attacks//Park J I, Kim J, eds. Computer Vision - ACCV 2012 Workshops. Lecture Notes in Computer Science, vol 7728. Berlin, Heidelberg: Springer: 121-132 [DOI:10.1007/978-3-642-37410-4_11]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision//Proceedings of the 38th International Conference on Machine Learning. PMLR, 139: 8748-8763 [DOI:10.48550/arXiv.2103.00020]

Rasheed H A, Khattak M U, Maaz M, Khan S H and Khan F S. 2022. Fine-tuned CLIP Models are Efficient Video Learners//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). : 6545-6554 [DOI: 10.1109/CVPR52729.2023.00633]

Seputis D, Mihailov S, Chatterjee S and Xiao Z H. 2024. Multi-Modal Adapter for Vision-Language Models[EB/OL]. [2024-09-03]. https://arxiv.org/pdf/2409.02958.pdf

Shao R, Lan X Y, Li J W and Yuen P C. 2019. Multi-Adversarial Discriminative Deep Domain Generalization for Face Presentation Attack Detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). : 10023-10031 [DOI: 10.1109/CVPR.2019.01026]

Srivatsan K, Naseer M and Nandakumar K. 2023. FLIP: Cross-domain Face Anti-spoofing with Language Guidance//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV).:19628-19639 [DOI: 10.1109/ICCV51070.2023.01803]

Wang C Y, Lu Y D, Yang S T and Lai S H. 2022. Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).: 20281-20290 [DOI: 10.1109/CVPR52688.2022.01964]

Wang Z, Wang Z Z, Yu Z T, Deng W H, Li J H, Gao T T, et.al 2022. Domain Generalization via Shuffled Style Assembly for Face Anti-Spoofing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).: 4123-4133 [DOI: 10.1109/CVPR52688.2022.00409]

Xie X H, Bian J T and Lai J H. 2022. Review on face liveness detection. Journal of Image and Graphics, 27(1): 63-87 (谢晓华, 卞锦堂, 赖剑煌. 2022. 人脸活体检测综述. 中国图象图形学报, 27(1): 63-87) [DOI: 10.11834/ jig.210470]

Yang L X, Zhang R Y, Wang Y C and Xie X H. 2024. MMA: Multi-Modal Adapter for Vision-Language Models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).: 23826-23837 [DOI: 10.1109/CVPR52733.2024.02249]

Yuan L, Chen D D, Chen Y L, Codella N, Dai X Y, Gao J F, et al. 2021. Florence: A New Foundation Model for Computer Vision[EB/OL].[2021-11-22]. https://arxiv.org/pdf/2111.11432.pdf

Yu Z T, Zhao C X, Wang Z Z, Qin Y X, Su Z, Li X B, et al. 2020. Searching Central Difference Convolutional Networks for Face Anti-Spoofing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).: 5295-5305 [DOI: 10.1109/CVPR42600.2020.00534]

Yu Z T, Li X B, Shi J G, Xia Z Q and Zhao G Y. 2021. Revisiting Pixel-Wise Supervision for Face Anti-Spoofing. IEEE Transactions on Biometrics, Behavior, and Identity Science, 3 (3) : 285-295 [DOI: 10.1109/TBIOM.2021.3065526]

Zhou Q Y, Zhang K Y, Yao T P, Yi R, Ding S H and Ma L Z. 2022. Adaptive mixture of experts learning for generalizable face anti-spoofing//Proceedings of the 30th ACM International Conference on Multimedia.:6009-6018 [DOI: 10.1145/3503161.3544769]

Zhou Q Y, Zhang K Y, Yao T P, Lu X Q, Ding S H and Ma L Z. 2024. Test-Time Domain Generalization for Face Anti-Spoofing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).:175-187 [DOI: 10.1109/CVPR52733.2024.00025]

Zhou Q Y, Zhang K Y, Yao T P, Lu X Q, Yi R, Ding S H, et al. 2023. Instance-Aware Domain Generalization for Face Anti-Spoofing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) : 20453-20463 [DOI: 10.1109/CVPR52729.2023.01959]

作者简介

张瑾，女，硕士，主要研究方向为计算机视觉，人脸活体检测。E-mail：zjin2046@163.com

梁凤梅，通信作者，女，副教授，主要研究方向为计算机视觉，深度学习等。E-mail：fm_liang@163.com

刘阿建，男，副研究员，主要研究方向为计算机视觉，人脸活体检测等。E-mail：ajian.liu@ia.ac.cn

万军，男，研究员，主要研究方向为模式识别、计算机视觉等。E-mail：jun.wan@ia.ac.cn

© 中国图象图形学报版权所有

蔡占川,男,教授,主要研究方向为计算机图形学与图像处理,智能信息处理等。E-mail:zccai@must.edu.mo

雷震,男,研究员,主要研究方向为计算机视觉,模式识别,人脸识别等。E-mail:zhen.lei@ia.ac.cn

© 中国图象图形学报版权所有