中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)02-0435-16

论文引用格式:Wang Y J, He J, Zhang J X, Sun R H and Liu X L. 2025. Semi-supervised facial expression recognition robust to head pose empowered by dual consistency constraints. Journal of Image and Graphics, 30(02):0435-0450(王宇键, 何军, 张建勋, 孙仁浩, 刘学亮. 2025. 头姿鲁棒的双一致性约束半监督表情识别. 中国图象图形学报, 30(02):0435-0450)[DOI:10.11834/jig.240205]

头姿鲁棒的双一致性约束半监督表情识别

王宇键[1],何军[2],张建勋[1*],孙仁浩[2],刘学亮[3]

1. 重庆理工大学计算机科学与工程学院,重庆　400054; 2. 数据空间研究院,合肥　230088;
3. 合肥工业大学计算机与信息学院,合肥　230601

摘　要:**目的**　现有表情识别方法聚焦提升模型的整体识别准确率,对方法的头部姿态鲁棒性研究不充分。在实际应用中,人的头部姿态往往变化多样,影响表情识别效果,因此研究头部姿态对表情识别的影响,并提升模型在该方面的鲁棒性显得尤为重要。为此,在深入分析头部姿态对表情识别影响的基础上,提出一种能够基于无标签非正脸表情数据提升模型头部姿态鲁棒性的半监督表情识别方法。**方法**　首先按头部姿态对典型表情识别数据集AffectNet重新划分,构建了AffectNet-Yaw数据集,支持在不同角度上进行模型精度测试,提升了模型对比公平性。其次,提出一种基于双一致性约束的半监督表情识别方法(dual-consistency semi-supervised learning for facial expression recognition,DCSSL),利用空间一致性模块对翻转前后人脸图像的类别激活一致性进行空间约束,使模型训练时更关注面部表情关键区域特征;利用语义一致性模块通过非对称数据增强和自学式学习方法不断地筛选高质量非正脸数据用于模型优化。在无需对非正脸表情数据人工标注的情况下,方法直接从有标签正脸数据和无标签非正脸数据中学习。最后,联合优化了交叉熵损失、空间一致性约束损失和语义一致性约束损失函数,以确保有监督学习和半监督学习之间的平衡。**结果**　实验结果表明,头部姿态对自然场景表情识别有显著影响;提出AffectNet-Yaw具有更均衡的头部姿态分布,有效促进了对这种影响的全面评估;DCSSL方法结合空间一致性和语义一致性约束充分利用无标签非正脸表情数据,显著提高了模型在头部姿态变化下的鲁棒性,较MA-NET(multi-scale and local attention network)和EfficientFace全监督方法,平均表情识别精度分别提升了5.40%和17.01%。**结论**　本文提出的双一致性半监督方法能充分利用正脸和非正脸数据,显著提升了模型在头部姿态变化下的表情识别精度;新数据集有效支撑了对头部姿态对表情识别影响的全面评估。

关键词:表情识别(FER);头部姿态;双一致性约束;半监督学习;AffectNet;图像识别

Semi-supervised facial expression recognition robust to head pose empowered by dual consistency constraints

Wang Yujian[1], He Jun[2], Zhang Jianxun[1*], Sun Renhao[2], Liu Xueliang[3]

1. School of Computer Science and Engineering, Chongqing University of Technology, Chongqing　400054, China;
2. Institute of Dataspace, Hefei　230088, China; 3. School of Computer Science and Information Engineering, Hefei University of Technology, Hefei　230601, China

Abstract: **Objective** The field of facial expression recognition (FER) has long been a vibrant area of research, with a

收稿日期:2024-04-11;**修回日期**:2024-06-14;**预印本日期**:2024-06-21

*通信作者:张建勋　zjx@cqut.edu.cn;

基金项目:国家自然科学基金项目(61971078);重庆市教委科技重大项目(KJZD-M202301901)

Supported by: National Natural Science Foundation of China(61971078); Chongqing Municipal Education Commission Science and Technology Major Project(KJZD-M202301901)

focus on improving the accuracy of identifying expressions across a wide range of faces. However, despite these advancements, a crucial aspect that has not been adequately explored is the robustness of FER models to changes in head pose. In real-world applications, where faces are captured at various angles and poses, existing methods often struggle to accurately recognize expressions in faces with considerable pose variations. This limitation has created an urgent need to understand the extent to which head pose affects FER models and to develop robust models that can handle diverse poses effectively. First, the impact of head pose on FER was analyzed. Rigorous experimentation has provided strong evidence that existing FER approaches are indeed vulnerable when faced with faces exhibiting large head poses. This vulnerability not only limits the practical applicability of these methods but also highlights the critical need for research focused on enhancing the pose robustness of FER models. This challenge is addressed by introducing a semi-supervised framework that leverages unlabeled nonfrontal facial expression samples to increase the pose robustness of FER models. This framework aims to overcome the limitations of existing methods by exploring unlabeled data to supplement labeled frontal face data, allowing the model to learn representations that are invariant to head pose variations. Incorporating unlabeled data expands the model's exposure to a wider range of poses, ultimately enhancing robustness and accuracy in FER. This study highlights the importance of pose robustness in FER and proposes a semi-supervised framework to address this critical limitation. Rigorous experimentation and analysis provide insights into the impact of head pose on FER, and a robust model to accurately recognize facial expressions across diverse poses is developed. This approach paves the way for more practical and reliable FER systems in real-world applications. **Method** The AffectNet dataset was reorganized via a deterministic resampling procedure to examine the impact of head pose on FER. In this procedure, the same numbers of faces from different expression categories and head pose intervals were uniformly and randomly sampled to build a new challenging FER dataset called AffectNet-Yaw, whose test samples were balanced both in the category axis and the head pose axis. The AffectNet-Yaw dataset enables a deep investigation into how head pose affects the performance of an FER model. A semi-supervised framework for FER called dual consistency constraints (DCSSLs) was used to improve the robustness of the model to head poses. This framework leverages a spatial consistency module to force the model to produce consistent category activation maps for each face and its flipped mirror during training, which allows the model to prioritize the key facial regions for FER. A semantic consistency module is also employed to force the model to extract consistent features of two augmentations of the same face that exhibit similar semantics. In particular, two different data augmentations were applied to a face. One of the arguments is weak, and the other is strong. The weakly augmented face was flipped, and model predictions of this sample and its flipped mirror were obtained. Only those unlabeled nonfrontal faces for which the model makes the same prediction with high confidence are retained. Their predicted categories together with their strongly augmented variant comprise "date-label" pairs that are utilized for model training as pseudolabeled positives. This scheme increases data variation to benefit model optimization. Within the framework, a joint optimization target that can integrate cross-entropy loss, the spatial consistency constraint, and the semantic consistency constraint was devised to balance supervised learning and semi-supervised learning. Owing to the joint training, the proposed framework requires no manual labeling of nonfrontal faces; instead, it directly learns from labeled frontal faces and unlabeled nonfrontal faces, highly boosting its robustness and generalizability. **Result** The evaluation results of various fully supervised FER methods on both the AffectNet and AffectNet-Yaw datasets underscore the profound impact of head pose variability in real-world scenarios, emphasizing the critical need to increase the robustness of the FER model to such challenges. Empirical findings confirm that the AffectNet-Yaw dataset serves as a rigorous and effective platform for comprehensive investigations into how head pose influences FER model performance. Comparative analyses between baseline models and state-of-the-art methods on the AffectNet and AffectNet-Yaw datasets reveal compelling insights. In particular, the novel DCSSL framework considerably enhances the model's ability to adapt to head pose variations, showing substantial improvement over existing benchmarks. When MA-NET and EfficientFace are used as benchmarks, the DCSSL framework achieves considerable average performance improvements of 5.40% and 17.01%, respectively. In addition, the effectiveness of this approach was determined by comparing the models that have performed well in the field of expression recognition in the last three years on two separate datasets. In terms of weighting parameter settings, different weighting choices have a considerable effect on model performance. A series of parameter selection experiments were conducted via the control variable approach, and the model achieves optimal expression recogni-

tion performance on the AffectNet-Yaw test data when the weight of the spatial consistency constraint loss is set to 1 and that of the semantic consistency constraint loss is set to 5. These results highlight the efficacy of the proposed DCSSL framework in mitigating the detrimental effects of head pose variations on FER accuracy. When spatial and semantic consistency modules are integrated, the proposed approach can not only improve model robustness but also demonstrate its ability to adapt and generalize effectively across diverse head poses encountered in real-world applications. This study not only contributes to a challenging new dataset named AffectNet-Yaw for advancing FER research under realistic conditions but also establishes a novel methodology named DCSSL that sets a new standard for addressing head pose challenges in FER. These advancements are pivotal for enhancing the reliability and applicability of FER systems in practical settings where head pose variability is prevalent. **Conclusion** The proposed DCSSL framework in this work can efficiently exploit both frontal and nonfrontal faces, successfully increasing the accuracy of FER under diverse head poses. The new AffectNet-Yaw dataset has a more balanced data distribution, both along the category axis and the head pose axis, enabling a comprehensive study of the impact of head poses on FER. Both of these elements hold substantial value for building robust FER models.

Key words: facial expression recognition (FER); head pose; dual-consistency constraint; semi-supervised learning; AffectNet; image recognition

0 引　言

面部表情识别(facial expression recognition, FER)作为计算机视觉领域重要研究方向之一,在人机交互(Derbali等,2023)、自闭症检测(Jeong和Ko, 2018)和安全驾驶等领域(Moin等,2023)展现了巨大的应用潜力。然而,无约束自然场景下的表情识别由于受到人脸遮挡、头部姿态和光照等不可控因素影响,面临复杂挑战。具体到头部姿态,现有研究(Le等,2023;崔鑫宇 等,2024;赵明华 等,2024)通常假设使用正脸(frontal)或接近正脸(near-frontal)表情数据,其在识别如图1所示的非正脸(non-frontal)表情时并不总是成功。

图1　自然场景中不同头部姿态的人脸表情

Fig. 1　Facial expressions of different head poses in natural scenes

为提升表情识别头部姿态鲁棒性,Rudovic等人(2013)结合耦合缩放式高斯回归头部姿态归一化过程,提出具有头部姿态不变性的概率模型,以处理连续头部姿态变化下的表情识别。但该方法对特征点提取的准确性过度依赖,无法有效处理大幅度角度变化导致的特征点缺失问题。Rao等人(2015)采用加速稳定特征(speeded up robust features, SURF)算子提取面部局部区域特征,并在集成学习框架中选择最有利于表情识别的特征,能够识别一些大角度的表情。但该方法对于光照和特征变化敏感,导致特征提取不稳定。Vieriu等人(2015)和Yang等人(2023)通过将受头部姿态影响的3D人脸变换到与姿态无关的2D柱面,再划分成局部区域,应用随机森林等分类器进行表情识别。Wang等人(2020a)将人脸图像裁剪成子图,引入自注意力和交叉注意力机制,融合子图深度卷积特征以构建具有姿态不变性的人脸特征,在遮挡和头部姿态变化情况下都取得了较好的表情识别效果。但它们主要依赖实验室环境数据展开研究,可用数据量少、对表情数据的可控性有较高依赖,因此方法的真实场景泛化能力不强。

此外,主成分分析网络(principal component analysis network, PCANet)方法通过学习无标签正脸图像特征,基于正脸和非正脸之间的映射得到统一的头姿鲁棒性描述(Chan等,2015)。但该方法及其变体受到特征提取能力的限制,可能无法捕捉到复杂数据中的高级特征,且拓扑结构缺乏灵活性。Yan等人(2019)提出深度迁移学习网络框架,能够克服相机多角度拍摄带来的表情识别中的跨域问题。Zhang等人(2018)通过端到端的生成学习框架,自动生成不同头部姿态的面部表情图像,增加了表情识别训练数据的多样性。Wang等人(2019)将

编码器、身份鉴别器和姿态鉴别器整合进一个对抗学习框架，提取身份和姿态鲁棒的特征表示。虽然这些方法一定程度上提升了表情识别模型对于头部姿态的鲁棒性，但模型复杂度高、训练过程对大量标记数据依赖性强，且方法主要注重提高模型整体识别性能，忽略了其在不同头部姿态角度下的识别能力，对模型的头部姿态鲁棒性认识不足。

针对上述问题，本文从数据集解析重构、表情识别模型框架创新和联合优化评估3个方面展开系统性研究，剖析头部姿态对表情的影响，提出一种半监督学习方法，以提升模型学习便利性和头部姿态鲁棒性，本文主要工作如下：1）首次对典型表情识别数据集AffectNet（Mollahosseini等，2019）进行详细的头部姿态解析，发现在数据集中存在类别和头部姿态分布不均衡问题，尤其是正脸样本占主导，随着头部姿态角度增大，数据集相应的样本数量呈指数趋势下降。为此，本文提出AffectNet-Yaw数据集，通过均匀采样不同类别和头部姿态区间样本，有效消除了测试数据中的样本不均衡问题，支持衡量模型在不同头部姿态区间和不同类别上的表情识别精度，是一个更公平、更具挑战性的模型对比基准数据集。2）提出一种基于双一致性约束的半监督表情识别方法（dual-consistency semi-supervised learning for facial expression recognition，DCSSL），以应对真实场景下非正脸表情数据难获取、难标注等挑战导致的模型学习难问题。该方法利用有标签正脸表情数据进行有监督学习，同时利用空间一致性约束和语义一致性约束基于无标签非正脸数据进行自监督学习，通过联合优化交叉熵损失、空间一致性约束损失和语义一致性约束损失，构建联合优化目标，确保有监督学习和自监督学习之间的平衡。这不仅有效降低了模型学习的数据采集成本，还提高了模型头部姿态鲁棒性。3）在AffectNet和AffectNet-Yaw数据集上进行了大量的实验对比和消融实验分析，验证了上述框架的有效性和各组成模块的作用。实验表明，本文提出的DCSSL半监督方法可以在无须对非正脸图像进行人工标注的前提下，结合双一致性约束有效提升模型在各个头部姿态角度区间上表情识别精度，并在更有挑战的AffectNet-Yaw数据集上取得媲美全监督算法的表情识别效果。

本文剖析AffectNet数据集头部姿态分布不均衡问题，提出AffectNet-Yaw数据集，对比现有表情识别方法在两个数据集上的性能差异，突出表现头部姿态变化对模型表情识别性能的影响；详细介绍提出的DCSSL方法及其设计原理；描述实验设置，呈现实验结果与分析，证明方法各主要模块的有效性。

1 头部姿态对表情识别的影响

给定数据集$\boldsymbol{D}=\{\boldsymbol{D}_{\text{train}},\boldsymbol{D}_{\text{test}}\}$，表情识别从训练集$\boldsymbol{D}_{\text{train}}$各类表情样本训练编码器和分类器，学会识别不同类别的表情。平均准确率A_i是衡量模型识别性能的主要指标，具体为

$$A_i=\frac{\sum_{(\boldsymbol{X},y)\in\boldsymbol{D}_i}\mathbb{I}\left(f\left(E(\boldsymbol{X})\right),y\right)}{\left|\boldsymbol{D}_i\right|}\quad \text{s.t.}\ \forall i\in(\text{train},\text{test})\tag{1}$$

式中，$\boldsymbol{X}$为人脸图像，y为人脸图像所对应的表情类别，$E(\cdot)$为图像样本的编码器，$f(\cdot)$为分类器，$\mathbb{I}(\cdot)$为指示函数，若模型能准确识别样本$\boldsymbol{X}$的表情类别，$\mathbb{I}\left(f\left(E(\boldsymbol{X})\right),y\right)=1$，否则$\mathbb{I}\left(f\left(E(\boldsymbol{X})\right),y\right)=0$，$\boldsymbol{D}_i$表示数据集类别。

1.1 头部姿态分布不均衡现象及其影响

头部姿态影响表情识别精度，但平均准确率缺少对头部姿态信息的考量，无法全面反映这种影响。为此，本节采用头部姿态估计方法（Lai等，2023）准确获取人脸表情样本的头部姿态，衡量模型在不同头部姿态角度区间内的表情识别准确率，以表征头部姿态对表情识别的影响。具体而言，本文对AffectNet数据集进行头部姿态分析，根据算法输出头部姿态偏航角信息将表情样本划归到[0, 15°)，[15°, 30°)，[30°, 45°)，[45°, 60°)，[60°, 90°]等5个区间，然后统计不同偏航角区间内的样本分布情况，并计算表情识别方法在不同区间内的表情识别准确率。本文将区间[0, 15°)视为正脸区间，将头部姿态偏航角在正脸区间内的样本称为正脸样本。

表1和表2统计数据表明，AffectNet训练集和测试集各偏航角区间内的样本数量有明显差异。具体而言，AffectNet训练集正脸样本占比高达76.88%，随着头部姿态偏航角的增大，其非正脸区间的样本数量显著减少，特别是区间[60°, 90°)中的样本量仅占样本总量的0.05%。这意味着在AffectNet训练集中，除了已知的类别分布不均衡问题外，还存在严重的头部姿态分布不均衡问题。虽

然AffectNet测试集每类包含等量的测试样本，类别分布均衡，但也存在头部姿态分布不均衡问题。特别地，AffectNet测试集中不含有偏航角在区间[60°, 90°)内的样本。头部姿态分布不均衡问题意味着，典型表情识别方法在AffectNet数据集上的表情识别训练和测试准确率更多地反映了模型在正脸情况下的表情识别能力，不能充分且准确地衡量模型的真实性能。

表1　AffectNet训练集类别和头部姿态分布

Table 1　Category and head pose distribution in AffectNet train set

表情类别	角度区间/(°)					总计
	0～15	15～30	30～45	45～60	60～90	
中性(neutral)	54 587	14 996	4 729	636	67	75 015
高兴(happy)	112 760	18 497	3 078	284	40	134 659
悲伤(sad)	18 125	5 547	1 613	139	15	25 439
惊讶(surprise)	9 743	3 267	1 005	124	4	14 143
害怕(fear)	4 609	1 391	378	47	4	6 249
厌恶(disgust)	2 798	813	220	18	3	3 852
生气(anger)	17 560	5 464	1 695	227	17	24 963
总计	220 182	49 975	12 718	1 475	150	284 500

表2　AffectNet测试集类别和头部姿态分布

Table 2　Category and head pose distribution in AffectNet test set

表情类别	角度区间/(°)					总计
	0～15	15～30	30～45	45～60	60～90	
中性(neutral)	363	120	25	2	0	510
高兴(happy)	423	66	9	2	0	500
悲伤(sad)	342	127	30	1	0	500
惊讶(surprise)	336	140	21	2	0	499
害怕(fear)	347	138	14	1	0	500
厌恶(disgust)	343	128	26	2	0	499
生气(anger)	353	119	24	4	0	500
总计	2 507	838	149	14	0	3 508

1.2　AffectNet-Yaw数据集

为了更精确、全面地评估表情识别模型的性能，本文从类别和头部姿态两个维度重新组织AffectNet数据集，提出AffectNet-Yaw数据集，其训练和测试样本分布见表3和表4。通过精心设计的数据抽样方式，AffectNet-Yaw测试集几乎包含等量的来自不同类别和不同头部姿态角区间的测试样本，消除了头部姿态分布不均衡问题，为头部姿态变化下的自然场景表情识别提供了一个更加公平和真实的评估环境。

各头部姿态区间的累进表情识别精度如图2所示，图2表明，在AffectNet-Yaw测试集上模型只有在不同类别、不同头部姿态下都表现出更好的识别效果，才能取得更高的表情识别精度。相反，模型只需准确识别AffectNet测试集正脸样本即可取得>70%的识别精度。

1.3　AffectNet-Yaw上典型方法评估

为定量分析头部姿态变化对表情识别的影响，本节保持原训练设置不变，在AffectNet-Yaw上对EfficientFace、MA-Net和POSTER(pyramid cross-fusion transformer)进行训练和测试，将测试结果与

表3　AffectNet-Yaw训练集类别和头部姿态分布

Table 3　Category and head pose distribution in AffectNet-Yaw train set

表情类别	角度区间/(°)					总计
	0～15	15～30	30～45	45～60	60～90	
中性(neutral)	54 537	15 046	4 779	686	117	75 165
高兴(happy)	112 720	15 046	3 128	334	55	134 784
悲伤(sad)	18 075	5 597	1 663	289	21	25 645
惊讶(surprise)	9 703	3 317	1 055	174	6	14 225
害怕(fear)	4 559	1 441	427	57	5	6 489
厌恶(disgust)	2 748	863	270	28	4	3 913
生气(anger)	17 510	5 514	1 745	227	21	25 017
总计	219 852	50 325	13 067	1 795	229	285 268

表4　AffectNet-Yaw测试集类别和头部姿态分布

Table 4　Category and head pose distribution in AffectNet-Yaw test set

表情类别	角度区间/(°)					总计
	0～15	15～30	30～45	45～60	60～90	
中性(neutral)	50	50	50	50	22	222
高兴(happy)	50	50	50	50	15	215
悲伤(sad)	50	50	50	50	6	206
惊讶(surprise)	50	50	50	50	2	202
害怕(fear)	50	50	50	50	1	201
厌恶(disgust)	50	50	50	10	1	161
生气(anger)	50	50	50	10	4	164
总计	350	350	350	270	51	1 371

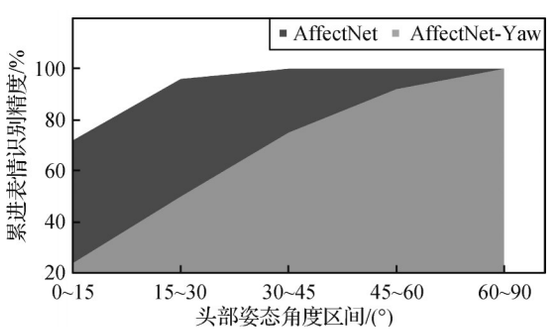


图2　各头部姿态区间的累进表情识别精度

Fig. 2　Cumulative FER accuracy across different head pose intervals

其在AffectNet数据集上的表现进行对比。如表5所示，由于AffectNet-Yaw测试集包含更多的非正脸表情样本，在AffectNet上测试表现良好的模型在AffectNew-Yaw上呈现出不同程度的性能衰退。EfficientFace、MA-Net和POSTER在AffectNet上的平均识别精度为62.97%、56%和65.85%，但在AffectNet-Yaw上却只有42.91%、54.52%和63.35%，性能下降20.06%、1.48%和2.50%。这不仅说明头部姿态对模型表情识别能力产生显著负面影响，也表明AffectNet-Yaw数据集由于具备更好的头部姿态多样性，其测试精度地能更准确反映模型的真实自然场景表情识别能力。

从表5还可看出，对EfficientFace(Wang等，2023)、MA-Net(Zhao等，2021)和POSTER(Zheng等，2023)在AffectNet测试集不同头部姿态区间内的表情识别准确率进行衡量，发现这3个方法在正脸区间(即[0，15°)区间)内的表情识别准确率最高，随着头部姿态偏航角增大，模型性能呈下降趋势。但总体上，方法的整体识别准确率与方法的正脸区间

表 5　典型的表情识别方法在 AffectNet 和 AffectNet-Yaw 数据集上的表情识别准确率

Table 5　FER accuracy of typical FER methods on AffectNet and AffectNet-Yaw datasets

/%

模型	数据集	各头部姿态角度/(°)					平均
		0～15	15～30	30～45	45～60	60～90	
EfficientFace(Wang 等,2023)	AffectNet	65.25	53.98	56.37	42.50	–	62.97
	AffectNet-Yaw	43.14	43.42	38.28	44.07	64.00	42.91
MA-Net(Zhao 等,2021)	AffectNet	59.55	49.63	53.69	50.00	–	56.00
	AffectNet-Yaw	56.28	56.85	49.14	53.33	70.00	54.52
POSTER(Zheng 等,2023)	AffectNet	68.12	61.95	63.08	42.85	–	65.85
	AffectNet-Yaw	61.14	68.28	55.42	62.22	64.00	63.35

注:“–”表示无对应结果。

表情识别精度相近,这证明了 1.1 节的论断。

2　DCSSL 表情识别方法

增加训练数据中非正脸样本数量是提高模型头部姿态鲁棒性的一种合理策略(Lopes 等,2017;Ma 等,2022;Psaroudakis 和 Kollias,2022)。然而,由于面部表情具有快速变化的特点以及由人种、性别、性格等多方面因素引起的个体情感表达差异(Cai 等,2023;Liu 等 2023a;Stoychev 和 Gunes,2022),非正脸表情数据的采集和标注存在巨大挑战。一方面,大角度头部转动会导致面部自遮挡,显著影响数据采集质量;另一方面,不同标注者对同一对象的同一表情有不同的感受,头部姿态增加了这种标注的不确定性。为解决这一问题,本文提出 DCSSL 表情识别方法,采用“有监督 +自监督”的半监督学习框架,如图 3 所示,直接从无标签非正脸表情样本学习识别侧脸表情,而无需人工进行非正脸表情标注。

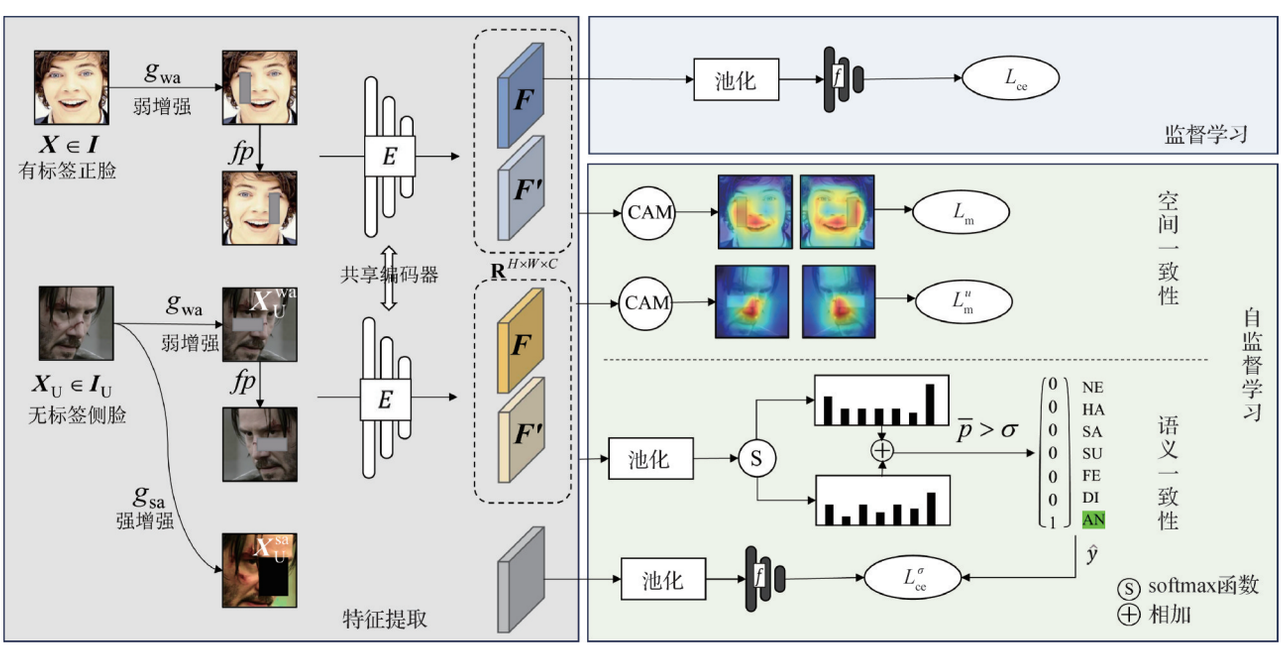


图 3　空间和语义双重一致性半监督学习框架图

Fig. 3　Illustration of the dual consistency semi-supervised learning framework

2.1　方法概述

具体地,给定表情识别模型 $f(E(\cdot))$、有标签正脸训练样本集合 $\boldsymbol{I}$ 和无标签非正脸样本集合 $\boldsymbol{I}_U$,本文 DCSSL 首先通过特征提取将样本映射到特征空间,

得到特征图；然后，在有监督学习部分，方法仅在有标签正脸样本集合$\boldsymbol{I}$上采用交叉熵损失优化模型，具体为

$$L_{\mathrm{ce}}=\frac{1}{|\boldsymbol{I}|}\sum_{(\boldsymbol{X},y)\in\boldsymbol{I}}L(f(E(\boldsymbol{X})),y) \tag{2}$$

式中，$\boldsymbol{X}$为人脸图像，y为人脸图像所对应的类别标签，$L(\cdot)$表示计算预测结果与真实标签y之间的交叉熵损失。

在自监督学习部分，方法进一步将无标签非正脸样本集合$\boldsymbol{I}_{\mathrm{U}}$纳入模型学习，利用空间一致性模块和语义一致性模块两部分共同约束模型学习过程。其中，空间一致性模块要求模型为样本（包括有标签正脸样本和无标签非正脸样本）及其翻转输出匹配的类别激活，增强其局部特征表示能力；语义一致性模块对无标签非正脸样本应用强弱非对称数据增强，预测弱增强样本及其翻转表情类别，并根据联合预测结果选择分类置信度较高的样本所对应的强增强样本，计算模型在强增强样本上的交叉熵损失，约束模型为样本强弱增强输出一致的语义预测。通过结合有标签正脸有监督学习和无标签非正脸自监督学习，本文DCSSL方法旨在从包含大量无标签非正脸数据的人脸表情样本（即$\boldsymbol{I}\cup\boldsymbol{I}_{\mathrm{U}}$）中自主学习，逐渐提高模型对非正脸表情的识别能力，增强模型对头部姿态变化的鲁棒性。空间一致性模块和语义一致性模块在上述模型半监督学习过程中发挥重要作用，下面将详述该双一致性模块的作用机理。

2.2 空间一致性模块

表情是由面部肌肉运动引起的一种情绪外化表现，捕捉面部局部信息变化是识别面部表情的重要方向（Makhmudkhujaev 等，2019；Uddin 等，2017）。本文空间一致性模块旨在增强表情识别模型的局部特征表示能力，以更准确地识别面部表情。对于样本$\boldsymbol{X}$，该模块首先应用数据弱增强算子$g_{\mathrm{wa}}(\cdot)$将样本调整为表情识别模型要求的输入大小，然后随机擦除部分图像区域添加空间扰动，再将样本翻转得到$\boldsymbol{X}'=fp\left(g_{\mathrm{wa}}(\boldsymbol{X})\right)$，其中$fp(\cdot)$是图像水平翻转函数。将样本$\boldsymbol{X}$及其翻转$\boldsymbol{X}'$输入模型，模块使用特征提取器$E$提取特征图$\boldsymbol{F}\in\mathbf{R}^{H\times W\times C}$和$\boldsymbol{F}'\in\mathbf{R}^{H\times W\times C}$，其中$H$、$W$和$C$表示高、宽和通道数，使用类别激活映射（class activation mapping，CAM）计算样本类别激活图$\boldsymbol{M},\boldsymbol{M}'\in\mathbf{R}^{H\times W\times L}$，其中$L$表示通道数，具体为

$$\begin{aligned}\boldsymbol{M}_k(i,j)&=\boldsymbol{W}_{:,k}^{\mathrm{T}}\boldsymbol{F}(i,j)\\ \boldsymbol{M}'_k(i,j)&=\boldsymbol{W}_{:,k}^{\mathrm{T}}\boldsymbol{F}'(i,j)\end{aligned} \tag{3}$$

式中，$\boldsymbol{W}\in\mathbf{R}^{C\times L}$表示分类器$f(\cdot)$权重，$\boldsymbol{W}_{:,k}^{\mathrm{T}}\in\mathbf{R}^{C}$为与第$k$类别相关的各特征通道权重的转置，$\boldsymbol{F}(i,j),\boldsymbol{F}'(i,j)\in\mathbf{R}^{C}$分别代表特征图$\boldsymbol{F}$和$\boldsymbol{F}'$上空间位置$(i,j)$处的局部特征。

考虑卷积网络局部连接特点，类别激活图$\boldsymbol{M}$，$\boldsymbol{M}'$应能反映表情变化和随机擦除对样本局部特征信息的影响。受此启发，空间一致性模块约束类别激活图$\boldsymbol{M}$，$\boldsymbol{M}'$反映这种对称干扰，增强模型的局部特征表示能力，具体为

$$\begin{aligned}L_{\mathrm{m}}&=\frac{1}{|\boldsymbol{I}|}\sum_{(\boldsymbol{X},y)\in\boldsymbol{I}}\left\|\boldsymbol{M}-fp(\boldsymbol{M}')\right\|_2\\ L_{\mathrm{m}}^{\mathrm{u}}&=\frac{1}{|\boldsymbol{I}_{\mathrm{U}}|}\sum_{\boldsymbol{X}_{\mathrm{U}}\in\boldsymbol{I}_{\mathrm{U}}}\left\|\boldsymbol{M}-fp(\boldsymbol{M}')\right\|_2\end{aligned} \tag{4}$$

式中，$\|\cdot\|_2$表示L_2范数，用于计算激活图$\boldsymbol{M}$和其翻转后的激活图$\boldsymbol{M}'$之间的平方差。

第3.3节实验表明，计算模型在有标签正脸上的空间一致性约束损失L_{m}和无标签非正脸样本上的空间一致性约束损失$L_{\mathrm{m}}^{\mathrm{u}}$是必要的，它们有助于模型利用无标签非正脸样本提升对非正脸表情的识别能力。

2.3 语义一致性模块

语义一致性模块采用自学式学习（self-taught learning）（Jaiswal 等，2021；Liu 等，2023c；Raina 等，2007）从无标签非正脸样本中选择置信度较高的样本，与有标签正脸样本一起用于模型训练。自学习式学习是一种广泛使用的半监督学习策略（Kim 等，2022；Wang 等，2013）。通过对任意无标签样本$\boldsymbol{X}_{\mathrm{U}}\in\boldsymbol{I}_{\mathrm{U}}$进行分类预测，得到类别预测结果$\hat{y}=f\left(E\left(\boldsymbol{X}_{\mathrm{U}}\right)\right)$和相应的分类置信度$p$，自学式学习基于分类置信度是否大于指定阈值$\sigma$（即$p>\sigma$）选择置信度较高的无标签样本打上伪标签，从而扩充有标签数据，增强有监督学习过程。自学式学习的每一轮（epoch）学习都需要基于更新的模型对无标签样本进行重新预测、筛选和打标，直至模型对无标签样本的预测收敛到一定的稳定状态，样本筛选结果保持不变。

不难发现，自学式学习容易受噪声数据干扰。一些被模型以较高置信度错误分类的样本将持续以较高的置信度出现在后续有监督训练过程中，给模

型学习带来负面影响。为了克服该问题，本文语义一致性模块在自学式学习过程中，采用非对称的数据筛选方式从无标签样本中选择高质量数据用于模型优化。具体地，模块对无标签样本应用两种强度不同的数据增强方式，得到数据增强样本。$\boldsymbol{X}_{\mathrm{U}}^{\mathrm{wa}}=g_{\mathrm{wa}}\left(\boldsymbol{X}_{\mathrm{U}}\right)$和$\boldsymbol{X}_{\mathrm{U}}^{\mathrm{sa}}=g_{\mathrm{sa}}\left(\boldsymbol{X}_{\mathrm{U}}\right)$，其中，$g_{\mathrm{wa}}(\cdot)$和$g_{\mathrm{sa}}(\cdot)$分别为数据弱增强算子和强增强算子；接着，模块根据弱增强样本$\boldsymbol{X}_{\mathrm{U}}^{\mathrm{wa}}$的分类结果和置信度，选择其对应的强增强样本$\boldsymbol{X}_{\mathrm{U}}^{\mathrm{sa}}$用于模型训练，具体为

$$L_{\mathrm{ce}}^{\sigma}=\frac{1}{|\boldsymbol{I}_{\mathrm{U}}^{\sigma}|}\sum_{(\boldsymbol{X}_{\mathrm{U}}^{\mathrm{sa}},\hat{y})\in\boldsymbol{I}_{\mathrm{U}}^{\sigma}}L(f(E(\boldsymbol{X}_{\mathrm{U}}^{\mathrm{sa}})),\hat{y}) \tag{5}$$

式中，$\boldsymbol{I}_{\mathrm{U}}^{\sigma}$是由分类置信度大于$\sigma$的无标签样本$\boldsymbol{X}_{\mathrm{U}}$的强增强$\boldsymbol{X}_{\mathrm{U}}^{\mathrm{sa}}$及其弱增强$\boldsymbol{X}_{\mathrm{U}}^{\mathrm{wa}}$的预测标签$\hat{y}=f\left(E\left(\boldsymbol{X}_{\mathrm{U}}^{\mathrm{wa}}\right)\right)$组成的伪标签数据集合。

语义一致性模块约束模型应对同一个样本的强弱增强表现出相同的语义。它用模型对样本弱增强的预测结果与样本强增强组成非对称的“样本—标签”对进行模型训练，实现了无标签样本选择和模型优化过程的松散耦合。在每一轮自学习过程中，上一轮被模块选择的无标签样本可以获得不同的类别预测或较小的分类置信度，减小了无标签样本分类和选择错误随模型训练逐渐扩散并最终对模型学习产生负面影响的风险。为保障样本选择的正确性，本文还对无标签样本的弱增强$\boldsymbol{X}_{\mathrm{U}}^{\mathrm{wa}}$及其镜像翻转$\boldsymbol{X}_{\mathrm{U}}^{'\mathrm{wa}}$同时进行分类预测，并联合两样本的分类预测结果和模型置信度选择分类结果相同且平均置信度（$\bar{p}=\frac{p+p'}{2}$）大于指定阈值的无标签非正脸样本供模型优化训练，如图7所示。所提出的语义一致性模块，一方面通过约束模型为非正脸样本的强弱增强输出相同的类别预测，为模型训练提供额外的正则化效果；另一方面，模块自适应地从非正脸样本中选择容易准确识别的侧脸数据用于模型训练，增强模型对非正脸表情的识别能力。

2.4 优化目标

综上，本文DCSSL方法通过结合有监督学习和自监督学习，从有标签正脸样本和无标签非正脸样本进行半监督表情识别优化，总体优化目标为

$$L=L_{\mathrm{ce}}+\lambda(L_{\mathrm{m}}+L_{\mathrm{m}}^{\mathrm{u}})+\mu L_{\mathrm{ce}}^{\sigma} \tag{6}$$

式中，μ和λ分别是空间一致性约束损失和语义一致性约束损失权重系数。

3 实 验

3.1 实验设置

本文实验在提出的AffectNet-Yaw数据集上进行。为确保对比的公平性，实验选择表情识别研究中使用最为广泛的ResNet50（He等，2016）作为模型骨干网络，使用MS-Celeb-1M（Guo等，2016）人脸识别预训练权重进行初始化，模型输入大小设置为224，批次大小设置为128。学习率和权重衰减因子分别设置为0.000 1和0.000 5，使用随机梯度下降（stochastic gradient descent）优化器（Robbins和Monro，1951）进行了200次迭代。学习率衰减算法采用了余弦退火衰减算法（Loshchilov和Hutter，2017）。空间一致性超参数λ、语义一致性超参数μ和无标签非正脸样本筛选阈值σ分别设置为5、1和0.95。这些实验超参数的选择经过精心考虑，以确保实验的准确性和可重复性。在数据预处理部分，样本弱增强仅对输入图像进行随机擦除；样本强增强使用随机增强算法（Ariz等，2019）、随机擦除算法（Berthelot等，2020）和一致性训练增强算法（consistency training augmentation）（Cubuk等，2020）等数据增强方法对输入图像进行更加强烈的扰动，应用了颜色反转、锐化、旋转、平移、随机裁剪和随机擦除等图像变换操作。模型性能评估则采用平均准确率作为评估指标。作为补充，实验还报告了不同方法在不同头部姿态区间的表情识别准确率，对方法进行更加细致的比较。

3.2 对比实验

本文实验主要对比了先进的全监督表情识别方法EfficientFace、SCN（self-cure network）（Wang等，2020b）、FDRL（feature decomposition and reconstruction learning）（Ruan等，2021）、MA-Net、DDAMFN（dual-direction attention mixed feature network）（Zhang等，2023）、EAC（erasing attention consistency）（Zhang等，2022）、DACL（deep attentive center loss）（Farzaneh和Qi，2021）、DAN（distract your attention network）（Wen等，2023）、Ada-DF（adaptive distribution fusion）（Liu等，2023b）和POSTER。在保持方法的原训练设置不变的情况下，实验仅将模型训练和测试数据更换为AffectNet-Yaw训练集和测试集，在训练集所有表情数据（包括正脸和非正脸样本）上进

行有监督训练，并在测试集上进行测试。两个基线模型baseline-0和baseline用于对比分析各表情识别方法对头部姿态变化的鲁棒性。其中，baseline-0为仅基于AffectNet-Yaw训练集正脸样本进行监督训练得到的朴素表情识别模型；baseline是在AffectNet-Yaw训练集所有样本上进行全监督训练得到的表情识别模型。它们采用与本文方法相同的骨干网络和初始化权重。

表6列举了各对比方法在AffectNet-Yaw测试集上的表情识别效果。结果表明，本文方法在不同头部姿态区间上相对于POSTER以外的其他方法均表现出显著的性能提升，取得了次优的测试效果。具体地，在区间[0，15°)，[45°，60°)上，本文方法表情识别准确率分别达到58.85%和70%，相比于DAN分别提升2%和11.12%。在区间[15°，30°)，[30°，45°)，[60°，90°)上，本文方法相对于EfficientFace、SCN、FDRL、DCAL、MA-Net、DDAMFN、EAC、DAN和Ada-DF都取得了最优或次优的识别精度。特别是在区间[45°，60°)和[60°，90°)上，本文方法相比其他对比方法性能总体提升明显，表明本文方法对大角度头部姿态下表情识别的有效性。MA-Net在各角度区间均有不错的表现，平均识别准确率达到54.52%，说明增强模型对面部局部特征的关注是提升表情识别精度和鲁棒性的行之有效的方向。本文后续消融实验部分将为此提供进一步的证据。对比基线模型baseline-0和baseline还可发现：1）仅依赖于正脸表情样本，模型在正脸以外的其他角度区间上测试准确率随头部姿态角度增大呈下降趋势。该趋势在EfficientFace和MA-Net上也有所体现。2）向训练数据中添加非正脸数据确实可有效提高模型在各头部姿态角度下的表情识别准确率，但非正脸标签样本数量不是决定因素。模型baseline相比baseline-0性能提升并不明显。相反，在等量的数据下，本文DCSSL方法通过从无标签非正脸数据和有标签正脸数据进行半监督学习，大幅提高了模型在各角度区间的表情识别准确率，表现出更好的头部姿态鲁棒性。

3.3 消融实验

3.3.1 空间一致性模块有效性讨论

如表7所示，与模型A0相比，仅在有标签正脸训练样本上施加空间一致性约束（即模型A1），模型各头部姿态角下的表情识别精度便得到大幅提升。

表6 各表情识别方法在AffectNet-Yaw测试集各角度区间的表情识别准确率比较

Table 6 FER accuracy of each model in different head pose intervals on AffectNet-Yaw test set

/%

类型	模型	各头部姿态角度/(°)					平均
		0～15	15～30	30～45	45～60	60～90	
全监督方法	baseline-0	52.00	46.57	36.00	39.62	44.00	43.79
	baseline	44.57	53.14	42.00	48.14	48.00	46.93
	EfficientFace（Wang等，2023）	43.14	43.42	38.28	44.07	64.00	42.91
	SCN（Wang等，2020b）	51.71	51.42	42.00	42.00	80.00	50.00
	FDRL（Ruan等，2021）	52.00	54.57	41.71	52.22	60.00	50.36
	DCAL（Farzaneh和Qi，2021）	52.28	54.57	44.85	51.85	**84.00**	52.04
	MA-Net（Zhao等，2021）	56.28	56.85	49.14	53.33	70.00	54.52
	DDAMFN（Zhang等，2023）	56.00	56.85	49.42	42.57	<u>76.00</u>	55.10
	EAC（Zhang等，2022）	58.00	58.00	<u>51.14</u>	58.51	62.00	56.49
	DAN（Wen等，2023）	56.85	61.42	50.28	58.88	72.00	57.29
	Ada-DF（Liu等，2023b）	58.28	60.57	49.42	57.40	72.00	57.44
	POSTER（Zheng等，2023）	**61.14**	**68.28**	**55.42**	<u>62.22</u>	64.00	**63.35**
半监督方法	DCSSL(本文)	<u>58.85</u>	<u>61.42</u>	49.71	**70.00**	74.00	<u>59.92</u>

注：加粗、下划线字体分别为各列最优、次优结果。

其中，大角度区间下的识别精度提升最为明显，在区间[45，60)表情识别精度提升 21.86%，在区间[60，90)表情识别精度提升 18%。考虑无标签非正脸样本并在实验中对其施加空间一致性约束（即模型 A2），可以进一步带来 1.1% 平均识别准确率的提升，验证了空间一致性模块的有效性。此外，由于训练数据中头部姿态角度在区间[15，30)的无标签非正脸样本较其他区间有明显数量优势，模型 A2 较 A1 在区间[15，30)表现出明显的性能提升，增益达 7.43%。这表明空间一致性模块虽然没有头部姿态感知能力，但能较好地捕捉数据中的头部姿态信息，对提升表情识别模型的头部姿态鲁棒性大有裨益。

表 7　语义一致性模块和空间一致性模块对模型表情识别准确率的影响

Table 7　Effect of semantic consistency module and spatial consistency module on model's FER accuracy

/%

模型	L_{ce}	L_m	L_m^u	L_{ce}^{σ}	各头部姿态角度/(°)					平均
					0～15	15～30	30～45	45～60	60～90	
A0	√	–	–	–	52.00	46.57	36.00	39.62	44.00	43.79
A1	√	√	–	–	55.42	56.28	**52.28**	<u>61.48</u>	62.00	56.27
A2	√	√	√	–	57.71	**63.71**	**52.28**	54.07	<u>64.00</u>	<u>57.37</u>
A3	√	√	–	√	<u>58.28</u>	59.71	47.71	59.25	60.00	56.20
A4	√	√	√	√	**58.85**	<u>61.42</u>	<u>49.71</u>	**70.00**	**74.00**	**59.92**

注：加粗、下划线字体分别为各列最优、次优结果，“A”表示消融实验，A0 与 baseline-0 等价，A4 与 DCSSL（本文）等价，“√”表示采用，“–”表示未采用。

3.3.2　语义一致性模块有效性讨论

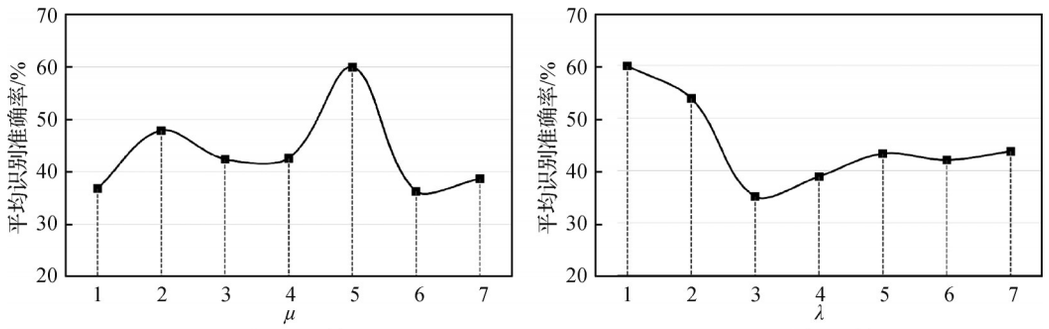
语义一致性模块并不总能带来表情识别模型性能的进一步提升。对比表 7 模型 A1 和 A3 可以看到，模型 A3 仅在区间[0，15°)和区间[15°，30°)上分别取得 2.86% 和 3.43% 的表情识别准确率提升，而在其他区间的识别准确率均有不同程度下降，平均识别准确率基本持平。这可能是因为模型在训练过程中容易对接近正脸的无标签训练样本以较高的置信度预测其正确类别，这些数据用于模型训练，提高了模型对正脸和接近正脸样本的表情识别效果，但对于提升模型在其他大角度区间的表情识别并无帮助。而模型 A4 相比于模型 A2 在大角度区间[45°，60°)和[60°，90°]上的表情识别准确率得到大幅提升，平均识别准确率提升了 2.55%。两相对比说明，语义一致性模块有效性是有条件的，仅当表情识别模型自身已经具有一定的头部姿态鲁棒性时，模块才能有效地从无标签非正脸数据中筛选更多高质量非正脸样本进行模型训练，从而进一步提升其头部姿态鲁棒性。

3.3.3　权重系数的影响

为了更好地联合有监督学习损失、空间一致性约束损失和语义一致性约束损失优化模型，本文通过一系列详尽的实验探究了式(6)中空间一致性约束损失和语义一致性约束权重系数设置对模型性能的影响。具体来说，本文采用控制变量渐近优化法进行权重设置。首先，将空间一致性约束损失权重 λ 设置为 1，采用步长为 1 的网格搜索找到最优的语义一致性权重 μ。然后，固定语义一致性权重为最优权重，在此基础上继续采用网格搜索法找到最优空间一致性约束权重。有监督学习损失作为表情识别主要优化目标，其权重始终设置为 1。对于每一种权重系数组合，本文在相同的 AffectNet-Yaw 训练数据和测试数据上进行模型训练和测试，以确保实验一致性和权重选择的可靠性。如图 4 所示，不同权重选择对模型性能有显著影响，当空间一致性约束损失的权重设置为 1 且语义一致性约束损失的权重设置为 5 时，模型在 AffectNet-Yaw 测试数据上取得最优的表情识别性能。

3.3.4　模块设计讨论

本节讨论空间一致性模块是否应用于无标签非正脸数据、语义一致性模块是否使用非对称数据增强结构以及无标签数据的空间一致性约束应用于样本弱增强或强增强等不同模块设计对模型表情识别能力和头部姿态鲁棒性的影响。具体如下：1）baseline-1 在 baseline-0 基础上对有标签正脸数据和无标签非正脸数据都应用空间一致性约束，无语义一致



(a) 语义一致性　(b) 空间一致性

图4　双一致性约束权重选择实验

Fig. 4　Experiments on weight selection of the two consistency constraints ((a) semantic consistency; (b) spatial consistency)

性约束，同表7模型A2；2）baseline-2仅对有标签正脸数据应用空间一致性约束，无语义一致性约束，但使用自学式学习直接根据模型对无标签样本的预测置信度选择对应样本用于模型训练；3）baseline-3在baseline-2基础上进一步对无标签非正脸样本应用空间一致性约束；4）baseline-4类似于本文DCSSL方法，但与本文对无标签非正脸样本的弱增强应用空间一致性约束不同，其对无标签非正脸样本的强增强应用空间一致性约束。结合baseline、baseline-0和本文DCSSL，不同模块设计下模型优化目标组成、优化损失计算，以及模型在AffectNet-Yaw上的测试结果如表8所示。其中，不同模块设计下的损失组成和计算方式如下：

1）全监督设置

baseline：

$$L_{\mathrm{ce}}=\frac{1}{\left|\boldsymbol{I}\cup\boldsymbol{I}_{\mathrm{U}}\right|}\sum_{(\boldsymbol{X},y)\in\boldsymbol{I}\cup\boldsymbol{I}_{\mathrm{U}}}L(f(E(\boldsymbol{X})),y)$$

L_{m}、$L_{\mathrm{m}}^{\mathrm{u}}$、$L_{\mathrm{ce}}^{\sigma}$无对应结果。

baseline-0：

$$L_{\mathrm{ce}}=\frac{1}{\left|\boldsymbol{I}\right|}\sum_{(\boldsymbol{X},y)\in\boldsymbol{I}}L(f(E(\boldsymbol{X})),y)$$

L_{m}、$L_{\mathrm{m}}^{\mathrm{u}}$、$L_{\mathrm{ce}}^{\sigma}$无对应结果。

2）半监督设置

baseline-1：

$$L_{\mathrm{ce}}=\frac{1}{\left|\boldsymbol{I}\right|}\sum_{(\boldsymbol{X},y)\in\boldsymbol{I}}L(f(E(\boldsymbol{X})),y)$$

$$L_{\mathrm{m}}=\frac{1}{\left|\boldsymbol{I}\right|}\sum_{(\boldsymbol{X},y)\in\boldsymbol{I}}\left\|\boldsymbol{M}-fp(\boldsymbol{M}')\right\|$$

$$L_{\mathrm{m}}^{\mathrm{u}}=\frac{1}{\left|\boldsymbol{I}_{\mathrm{U}}\right|}\sum_{\boldsymbol{X}_{\mathrm{U}}\in\boldsymbol{I}_{\mathrm{U}}}\left\|\boldsymbol{M}-fp(\boldsymbol{M}')\right\|$$

L_{ce}^{σ}无对应结果。

baseline-2：

$$L_{\mathrm{ce}}=\frac{1}{\left|\boldsymbol{I}\right|}\sum_{(\boldsymbol{X},y)\in\boldsymbol{I}}L(f(E(\boldsymbol{X})),y)$$

$$L_{\mathrm{m}}=\frac{1}{\left|\boldsymbol{I}\right|}\sum_{(\boldsymbol{X},y)\in\boldsymbol{I}}\left\|\boldsymbol{M}-fp(\boldsymbol{M}')\right\|$$

$L_{\mathrm{m}}^{\mathrm{u}}$无对应结果。

$$L_{\mathrm{ce}}^{\sigma}=\frac{1}{\left|\boldsymbol{I}_{\mathrm{U}}^{\sigma}\right|}\sum_{(\boldsymbol{X}_{\mathrm{U}},\hat{y})\in\boldsymbol{I}_{\mathrm{U}}^{\sigma}}L(f(E(\boldsymbol{X}_{\mathrm{U}})),\hat{y})$$

baseline-3：

$$L_{\mathrm{ce}}=\frac{1}{\left|\boldsymbol{I}\right|}\sum_{(\boldsymbol{X},y)\in\boldsymbol{I}}L(f(E(\boldsymbol{X})),y)$$

$$L_{\mathrm{m}}=\frac{1}{\left|\boldsymbol{I}\right|}\sum_{(\boldsymbol{X},y)\in\boldsymbol{I}}\left\|\boldsymbol{M}-fp(\boldsymbol{M}')\right\|$$

$$L_{\mathrm{m}}^{\mathrm{u}}=\frac{1}{\left|\boldsymbol{I}_{\mathrm{U}}\right|}\sum_{\boldsymbol{X}_{\mathrm{U}}\in\boldsymbol{I}_{\mathrm{U}}}\left\|\boldsymbol{M}-fp(\boldsymbol{M}')\right\|$$

$$L_{\mathrm{ce}}^{\sigma}=\frac{1}{\left|\boldsymbol{I}_{\mathrm{U}}^{\sigma}\right|}\sum_{(\boldsymbol{X}_{\mathrm{U}},\hat{y})\in\boldsymbol{I}_{\mathrm{U}}^{\sigma}}L(f(E(\boldsymbol{X}_{\mathrm{U}})),\hat{y})$$

baseline-4：

$$L_{\mathrm{ce}}=\frac{1}{\left|\boldsymbol{I}\right|}\sum_{(\boldsymbol{X},y)\in\boldsymbol{I}}L(f(E(\boldsymbol{X})),y)$$

$$L_{\mathrm{m}}=\frac{1}{\left|\boldsymbol{I}\right|}\sum_{(\boldsymbol{X},y)\in\boldsymbol{I}}\left\|\boldsymbol{M}-fp(\boldsymbol{M}')\right\|$$

$$L_{\mathrm{m}}^{\mathrm{u}}=\frac{1}{\left|\boldsymbol{I}_{\mathrm{U}}^{\mathrm{sa}}\right|}\sum_{\boldsymbol{X}_{\mathrm{U}}^{\mathrm{sa}}\in\boldsymbol{I}_{\mathrm{U}}^{\mathrm{sa}}}\left\|\boldsymbol{M}-fp(\boldsymbol{M}')\right\|$$

$$L_{\mathrm{ce}}^{\sigma}=\frac{1}{\left|\boldsymbol{I}_{\mathrm{U}}^{\sigma}\right|}\sum_{(\boldsymbol{X}_{\mathrm{U}}^{\mathrm{sa}},\hat{y})\in\boldsymbol{I}_{\mathrm{U}}^{\sigma}}L(f(E(\boldsymbol{X}_{\mathrm{U}}^{\mathrm{sa}})),\hat{y})$$

DCSSL（本文）：

$$L_{\mathrm{ce}}=\frac{1}{\left|\boldsymbol{I}\right|}\sum_{(\boldsymbol{X},y)\in\boldsymbol{I}}L(f(E(\boldsymbol{X})),y)$$

$$L_{\mathrm{m}}=\frac{1}{\left|\boldsymbol{I}\right|}\sum_{(\boldsymbol{X},y)\in\boldsymbol{I}}\left\|\boldsymbol{M}-fp(\boldsymbol{M}')\right\|$$

$$L_{\mathrm{m}}^{\mathrm{u}}=\frac{1}{\left|\boldsymbol{I}_{\mathrm{U}}\right|}\sum_{\boldsymbol{X}_{\mathrm{U}}\in\boldsymbol{I}_{\mathrm{U}}}\left\|\boldsymbol{M}-fp(\boldsymbol{M}')\right\|$$

表 8　不同模块设计下模型各角度区间表情识别准确率比较

Table 8　FER accuracy in different head pose intervals under different module designs

/%

类型	模型	各头部姿态角度/(°)					平均
		0～15	15～30	30～45	45～60	60～90	
全监督设置	baseline	44.57	53.14	42.00	48.14	48.00	46.93
	baseline-0	52.00	46.57	36.00	39.62	44.00	43.79
半监督设置	baseline-1	57.71	**63.71**	52.28	54.07	64.00	57.37
	baseline-2	**64.00**	58.28	<u>52.28</u>	57.03	60.00	<u>58.02</u>
	baseline-3	56.28	57.14	**52.85**	<u>62.22</u>	58.00	57.00
	baseline-4	51.71	48.85	40.00	51.85	**78.00**	49.27
	DCSSL(本文)	<u>58.85</u>	<u>61.42</u>	49.71	**70.00**	<u>74.00</u>	**59.92**

注：加粗、下划线字体分别为各列最优、次优结果。

$$L_{\mathrm{ce}}^{\sigma}=\frac{1}{\left|\boldsymbol{I}_{\mathrm{U}}^{\sigma}\right|}\sum_{(\boldsymbol{X}_{\mathrm{U}}^{\mathrm{sa}},\hat{y})\in\boldsymbol{I}_{\mathrm{U}}^{\sigma}}L(f(E(\boldsymbol{X}_{\mathrm{U}}^{\mathrm{sa}})),\hat{y})$$

对比 baseline、baseline-0、baseline-1 和 baseline-2 可以看到，在已知空间一致性模块能显著提高模型在各头部姿态区间上的准确率前提下，继续引入自学式学习可进一步提升模型的平均识别准确率，且在个别头部姿态区间表情识别准确率提升明显。例如，与 baseline-1 相比，baseline-2 应用自学式学习后平均识别准确率提升 0.65%，在正脸区间的识别准确率提升了 6.29%。但直接利用无标签非正脸样本进行自学式学习与空间一致性约束兼容性不强，需要细心搭配。例如，baseline-3 与 baseline-1 相比仅在区间[45°，60°)上取得 8.15% 的性能提升，在其他各区间内的表情识别精度都有所下滑，尤其在[15°，30°)和[60°，90°]区间内的表情识别准确率下滑平均超过 6%；而仅去除 baseline-3 对无标签非正脸样本的空间一致性约束(即 baseline-2)却能提高模型对小角度和大角度人脸的表情识别能力。在语义一致性约束下，空间一致性约束应用于无标签非正脸样本强增强(即 baseline-4)还是弱增强(本文方法)对模型性能也有显著影响。相比 baseline-3 和 baseline-4，本文 DCSSL 方法提出语义一致性模块，在语义一致性约束基础上，要求将无标签非正脸样本空间一致性约束应用于弱增强样本，较 baseline-3 和 baseline-4 都取得了较大提升，凸显了本文方法模块设计的优越性。

4　结　论

本文聚焦头部姿态对面部表情识别的影响，首先，数据分析确证了表情识别数据集中存在严重的头部姿态分布不均衡的问题，论证了该问题对现有表情识别性能的影响。其次，本文构建了测试数据表情类别分布和头部姿态分布都更加均衡的 AffectNet-Yaw 数据集，支撑对表情识别方法的头部姿态鲁棒性进行评估。最后，提出了一种创新的半监督表情识别方法 DCSSL，巧妙地融合了空间一致性和语义一致性约束，在统一的半监督学习框架中从有标签正脸数据和无标签非正脸数据学习，无需非正脸表情人工标注即可提高模型性能和头部姿态鲁棒性。本文 DCSSL 和 AffectNet-Yaw 数据集分别为头部姿态变化下的表情识别提供了一个有效的创新解决方案和一个高质量的评估基准。

本文算法在 AffectNet-Yaw 数据集上进行了实验验证，其平均识别精度较 MA-NET 和 EfficientFace 等全监督方法平均表情识别精度分别提升了 5.40% 和 17.01%，性能达到了领域领先水平，并且通过消融实验，验证了所提算法的效性。

然而，该方法在处理数据标签不均衡的场景时表现不佳。由于训练过程中头部姿态分布不均衡，模型更倾向于优先学习低角度数据，导致对低角度姿态的表现较好，而在高角度姿态下表现较差。未来将采用区分采样或课程学习等策略来缓解无标

签数据不平衡的问题。此外,该方法的训练效率还不高,未来将通过优化训练过程来提高效率。通过这些改进措施,可以提升模型在不同头部姿态下的性能和鲁棒性,从而更好地满足实际应用中的需求。

参考文献(References)

Ariz M, Villanueva A and Cabeza R. 2019. Robust and accurate 2D-tracking-based 3D positioning method: application to head pose estimation. Computer Vision and Image Understanding, 180: 13-22 [DOI: 10.1016/j.cviu.2019.01.002]

Berthelot D, Carlini N, Cubuk E D, Kurakin A, Sohn K, Zhang H and Raffel C. 2020. ReMixMatch: semi-supervised learning with distribution alignment and augmentation anchoring [EB/OL]. [2024-01-03]. https://arxiv.org/pdf/1911.09785.pdf

Cai J, Meng Z B, Khan A S, Li Z Y, O'Reilly J and Tong Y. 2023. Probabilistic attribute tree structured convolutional neural networks for facial expression recognition in the wild. IEEE Transactions on Affective Computing, 14(3): 1927-1941 [DOI: 10.1109/TAFFC.2022.3156920]

Chan T H, Jia K, Gao S H, Lu J W, Zeng Z N and Ma Y. 2015. PCANet: a simple deep learning baseline for image classification? IEEE Transactions on Image Processing, 24(12): 5017-5032 [DOI: 10.1109/TIP.2015.2475625]

Cubuk E D, Zoph B, Shlens J and Le Q V. 2020. Randaugment: practical automated data augmentation with a reduced search space//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 3008-3017 [DOI: 10.1109/CVPRW50498.2020.00359]

Cui X Y, He C, Zhao H K and Wang M L. 2024. Combining ViT with contrastive learning for facial expression recognition. Journal of Image and Graphics, 29(1): 123-133 (崔鑫宇, 何翀, 赵宏珂, 王美丽. 2024. 融合ViT与对比学习的面部表情识别. 中国图象图形学报, 29(1): 123-133) [DOI: 10.11834/jig.230043]

Derbali M, Jarrah M and Randhawa P. 2023. Autism spectrum disorder detection: video games based facial expression diagnosis using deep learning. International Journal of Advanced Computer Science and Applications, 14(1): 110-119 [DOI: 10.14569/IJACSA.2023.0140112]

Farzaneh A H and Qi X J. 2021. Facial expression recognition in the wild via deep attentive center loss//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2401-2410 [DOI: 10.1109/WACV48630.2021.00245]

Guo Y D, Zhang L, Hu Y X, He X D and Gao J F. 2016. MS-Celeb-1M: a dataset and benchmark for large-scale face recognition//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 87-102 [DOI: 10.1007/978-3-319-46487-9_6]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Jaiswal A, Babu A R, Zadeh M Z, Banerjee D and Makedon F. 2021. A survey on contrastive self-supervised learning. Technologies, 9(1): #2 [DOI: 10.3390/technologies9010002]

Jeong M and Ko B C. 2018. Driver's facial expression recognition in real-time for safe driving. Sensors, 18(12): #4270 [DOI: 10.3390/s18124270]

Kim S, Kim D, Cho M and Kwak S. 2022. Self-taught metric learning without labels//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 7421-7431 [DOI: 10.1109/CVPR52688.2022.00728]

Lai H J, Tang Z H and Zhang X Y. 2023. RepEPnP: weakly supervised 3D human pose estimation with EPnP algorithm//Proceedings of 2023 International Joint Conference on Neural Networks (IJCNN). Gold Coast, Australia: IEEE: 1-8 [DOI: 10.1109/IJCNN54540.2023.10191300]

Le N, Nguyen K, Tran Q, Tjiputra E, Le B and Nguyen A. 2023. Uncertainty-aware label distribution learning for facial expression recognition//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 6077-6086 [DOI: 10.1109/WACV56688.2023.00603]

Liu H W, Cai H L, Lin Q C, Zhang X W, Li X F and Xiao H. 2023a. FEDA: fine-grained emotion difference analysis for facial expression recognition. Biomedical Signal Processing and Control, 79: #104209 [DOI: 10.1016/j.bspc.2022.104209]

Liu S, Xu Y, Wan T M and Kui X Y. 2023b. A dual-branch adaptive distribution fusion framework for real-world facial expression recognition//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/ICASSP49357.2023.10097033]

Liu X, Zhang F J, Hou Z Y, Mian L, Wang Z Y, Zhang J and Tang J. 2023c. Self-supervised learning: generative or contrastive. IEEE Transactions on Knowledge and Data Engineering, 35(1): 857-876 [DOI: 10.1109/TKDE.2021.3090866]

Lopes A T, De Aguiar E, De Souza A F and Oliveira-Santos T. 2017. Facial expression recognition with Convolutional Neural Networks: coping with few data and the training sample order. Pattern Recognition, 61: 610-628 [DOI: 10.1016/j.patcog.2016.07.026]

Loshchilov I and Hutter F. 2017. Stochastic gradient descent with warm restarts [EB/OL]. [2024-01-03]. https://arxiv.org/pdf/1608.03983v3.pdf

Ma F, Li Y, Ni S G, Huang S L and Zhang L. 2022. Data augmentation for audio-visual emotion recognition with an efficient multimodal

conditional GAN. Applied Sciences, 12(1): #527 [DOI: 10.3390/app12010527]

Makhmudkhujaev F, Abdullah-Al-Wadud M, Iqbal M T B, Ryu B and Chae O. 2019. Facial expression recognition with local prominent directional pattern. Signal Processing: Image Communication, 74: 1-12 [DOI: 10.1016/j.image.2019.01.002]

Moin A, Aadil F, Ali Z and Kang D. 2023. Emotion recognition framework using multiple modalities for an effective human – computer interaction. The Journal of Supercomputing, 79(8): 9320-9349 [DOI: 10.1007/s11227-022-05026-w]

Mollahosseini A, Hasani B and Mahoor M H. 2019. AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1): 18-31 [DOI: 10.1109/TAFFC.2017.2740923]

Psaroudakis A and Kollias D. 2022. MixAugment and Mixup: augmentation methods for facial expression recognition//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans, USA: IEEE: 2366-2374 [DOI: 10.1109/CVPRW56347.2022.00264]

Raina R, Battle A, Lee H, Packer B and Ng A Y. 2007. Self-taught learning: transfer learning from unlabeled data//Proceedings of the 24th International Conference on Machine Learning. Corvallis, USA: ACM: 759-766 [DOI: 10.1145/1273496.1273592]

Rao Q Y, Qu X, Mao Q R and Zhan Y Z. 2015. Multi-pose facial expression recognition based on SURF boosting//Proceedings of 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). Xi'an, China: IEEE: 630-635 [DOI: 10.1109/ACII.2015.7344635]

Robbins H and Monro S. 1951. A stochastic approximation method. The Annals of Mathematical Statistics, 22(3): 400-407

Ruan D L, Yan Y, Lai S Q, Chai Z H, Shen C H and Wang H Z. 2021. Feature decomposition and reconstruction learning for effective facial expression recognition//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 7656-7665 [DOI: 10.1109/CVPR46437.2021.00757]

Rudovic O, Pantic M and Patras I. 2013. Coupled Gaussian processes for pose-invariant facial expression recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(6): 1357-1369 [DOI: 10.1109/TPAMI.2012.233]

Stoychev S and Gunes H. 2022. The effect of model compression on fairness in facial expression recognition [EB/OL]. [2024-01-03]. https://arxiv.org/pdf/2201.01709.pdf

Vieriu R L, Tulyakov S, Semeniuta S, Sangineto E and Sebe N. 2015. Facial expression recognition under a wide range of head poses//Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Ljubljana, Slovenia: IEEE: 1-7 [DOI: 10.1109/fg.2015.7163098]

Wang C, Wang S F and Liang G. 2019. Identity- and pose-robust facial expression recognition through adversarial feature learning//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 238-246 [DOI: 10.1145/3343031.3350872]

Wang G T, Li J, Wu Z J, Xu J H, Shen J F and Yang W K. 2023. EfficientFace: an efficient deep network with feature enhancement for accurate face detection. Multimedia Systems, 29(5): 2825-2839 [DOI: 10.1007/s00530-023-01134-6]

Wang H, Nie F P and Huang H. 2013. Robust and discriminative self-taught learning//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA: PMLR: 298-306

Wang K, Peng X J, Yang J F, Meng D B and Qiao Y. 2020a. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, 29: 4057-4069 [DOI: 10.1109/tip.2019.2956143]

Wang K, Peng X J, Yang J F, Lu S J and Qiao Y. 2020b. Suppressing uncertainties for large-scale facial expression recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6896-6905 [DOI: 10.1109/CVPR42600.2020.00693]

Wen Z Y, Lin W Z, Wang T and Xu G. 2023. Distract your attention: multi-head cross attention network for facial expression recognition. Biomimetics, 8(2): #199 [DOI: 10.3390/biomimetics8020199]

Yan K Y, Zheng W M, Zhang T, Zong Y, Tang C G, Lu C and Cui Z. 2019. Cross-domain facial expression recognition based on transductive deep transfer learning. IEEE Access, 7: 108906-108915 [DOI: 10.1109/ACCESS.2019.2930359]

Yang X P, Yang H N, Li J T and Wang S J. 2023. Simple but effective in-the-wild micro-expression spotting based on head pose segmentation//Proceedings of the 3rd Workshop on Facial Micro-Expression: Advanced Techniques for Multi-Modal Facial Expression Analysis. Ottawa, Canada: ACM: 9-16 [DOI: 10.1145/3607829.3616445]

Zhang F F, Zhang T Z, Mao Q R and Xu C S. 2018. Joint pose and expression modeling for facial expression recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3359-3368 [DOI: 10.1109/CVPR.2018.00354]

Zhang S N, Zhang Y H, Zhang Y, Wang Y F and Song Z G. 2023. A dual-direction attention mixed feature network for facial expression recognition. Electronics, 12(17): #3595 [DOI: 10.3390/electronics12173595]

Zhang Y H, Wang C R, Ling X and Deng W H. 2022. Learn from all: erasing attention consistency for noisy label facial expression recognition//Proceedings of the 17th European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer: 418-434 [DOI: 10.1007/978-3-031-19809-0_24]

Zhao M H, Dong S S, Hu J, Du S L, Shi C, Li P and Shi Z H. 2024. Attention-guided three-stream convolutional neural network for microexpression recognition. Journal of Image and Graphics, 29(1): 111-122 (赵明华, 董爽爽, 胡静, 都双丽, 石程, 李鹏,

石争浩. 2024. 注意力引导的三流卷积神经网络用于微表情识别. 中国图象图形学报, 29(1): 111-122[DOI: 10.11834/jig.230053]

Zhao Z Q, Liu Q S and Wang S M. 2021. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. IEEE Transactions on Image Processing, 30: 6544-6556[DOI: 10.1109/tip.2021.3093397]

Zheng C, Mendieta M and Chen C. 2023. POSTER: a pyramid cross-fusion transformer network for facial expression recognition//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Paris, France: IEEE: 3138-3147[DOI: 10.1109/ICCVW60793.2023.00339]

作者简介

王宇键,男,硕士研究生,主要研究方向为计算机视觉。
E-mail:idatawyj@163.com
张建勋,通信作者,男,教授,主要研究方向为计算机视觉。
E-mail: zjx@cqut.edu.cn
何军,男,博士研究生,主要研究方向为多媒体计算与分析、机器学习、弱监督学习和计算机视觉。
E-mail: hejun@idata.ah.cn
孙仁浩,男,硕士研究生,主要研究方向为计算机视觉。
E-mail: sunrh.hfut@gmail.com
刘学亮,男,教授,主要研究方向为多媒体信息处理。
E-mail: liuxueliang@hfut.edu.cn