Computer Science and Application 计算机科学与应用, 2026, 16(2), 381-394
Published Online February 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.162067

Hans汉斯

基于不确定性感知学习的鲁棒表情识别方法

贾开熠

合肥工业大学计算机与信息学院，安徽 合肥

收稿日期：2026年1月6日；录用日期：2026年2月6日；发布日期：2026年2月14日

摘　要

真实场景下的面部表情识别(FER)深受数据模糊性和标签噪声的困扰。现有方法主要依赖确定性嵌入，将每张面部图像映射到特征空间中的一个固定点。然而，这种范式迫使模型将模糊样本(如存在遮挡、低分辨率或细微复合情绪的样本)过拟合到固定的类别标签上，进而降低了泛化能力。为解决这一问题，本文提出一种新颖的模糊感知与抑制模块(APSM)。与传统方法不同，APSM将面部特征建模为概率高斯分布，其特征由均值(语义中心)和方差(不确定性)表征。本文引入一种不确定性衰减损失函数(Uncertainty-Attenuated Loss)，该函数动态权衡学习过程：估计不确定性高的样本对梯度更新的贡献更小，从而有效抑制噪声数据的影响。在RAF-DB和AffectNet数据集上的大量实验表明，本文的方法在无需复杂外部数据或人工清理的情况下，显著提升了鲁棒性并达到了先进性能。

关键词

面部表情识别，标签噪声学习，不确定性学习

Uncertainty-Aware Learning for Robust Facial Expression Recognition Method

Kaiyi Jia

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei Anhui

Received: January 6, 2026; accepted: February 6, 2026; published: February 14, 2026

Abstract

Facial Expression Recognition (FER) in real-world scenarios is severely hampered by data ambiguity and label noise. Existing methods predominantly rely on deterministic embeddings, mapping each facial image to a fixed point in the feature space. However, this paradigm forces the model to overfit ambiguous samples—such as those with occlusion, low resolution, or subtle compound emotions—to rigid categorical labels, thereby degrading generalization capabilities. To address this, we

文章引用: 贾开熠. 基于不确定性感知学习的鲁棒表情识别方法[J]. 计算机科学与应用, 2026, 16(2): 381-394.
DOI: 10.12677/csa.2026.162067

propose a novel Ambiguity Perception & Suppression Module (APSM). Unlike traditional approaches, APSM models facial features as Probabilistic Gaussian Distributions, characterized by a mean (semantic center) and a variance (uncertainty). We introduce an Uncertainty-Attenuated Loss that dynamically weighs the learning process: samples with high estimated uncertainty contribute less to the gradient update, effectively suppressing the impact of noisy data. Extensive experiments on the RAF-DB and AffectNet datasets demonstrate that our method significantly improves robustness and achieves state-of-the-art performance without requiring complex external data or manual cleaning.

Keywords

Facial Expression Recognition, Label Noise Learning, Uncertainty Learning

1. 引言

深度学习在计算机视觉的表情识别领域取得显著进展[1]。面部表情识别(FER)作为一个分类任务，需要在实验室做表情分类数据后逐步转化到更加真实的应用场景。RAF-DB [2]、AffectNet [3]等大规模数据集为这一进展提供了支撑。然而，这些真实场景数据的直观程度远低于在实验室中使用的人为构造的数据集。充满歧义但真实的数据集给人脸表情识别领域带来了更大的挑战，大大降低了人脸表情识别的准确性。在真实场景中，面部表情往往是非标准的。单张图像可能呈现复杂的多重混合情绪，或人脸可能被手、口罩等非表情特征部分所遮挡。由于标注数据时判断标准的不同，在表情数据的人为标注过程不可避免地存在标签噪声。

主流的深度 FER 方法主要依赖确定性特征学习。无论是经典的 ResNet [4]架构，还是近期引入的基于 Transformer [5]的方法(POSTER++ [6]，TransFER [7]等)，通常致力于设计更复杂的空间注意力机制(Spatial Attention)或局部特征融合策略，以提取最具判别力的面部区域。这些传统的面部表情识别方法大多采用确定性的方法，通过深度神经网络为输入的表情图像提取到一个固定的特征向量后通过训练最小化特征向量与真实类别标签之间的距离。这种确定性方法的根本缺陷反而在于它过度平等地对待了所有样本。当模型遇到前文提出的存在多重混合情绪的或者有各样遮挡噪声时的复杂的图像时，这个具有不确定性的图像却在分类后需要被强加一个确定的真实标签。传统模型在此时会试图以与拟合干净样本相同的强度去强行拟合这个离群点，从而扭曲了特征空间。其次，针对标签噪声问题，现有的去噪方法往往通过样本重加权(Sample Re-weighting)或标签修正来缓解。例如，SCN [8]提出了一种自愈网络(Self-Cure Network)，通过对训练样本进行排序和重加权来抑制噪声标签的影响。DMUE [9]则进一步挖掘了潜在的标签分布，并利用成对不确定性估计来解决标注的歧义性。此外，RUL [10]尝试通过相对不确定性学习，将特征映射到不确定性空间以增强模型的鲁棒性。尽管这些方法在一定程度上缓解了噪声干扰，但它们往往存在两个缺陷：首先是通常需要复杂的辅助网络或多阶段训练流程，难以端到端地集成到现有骨干网络中；其次是它们大多侧重于处理错误的标签，而非通过建模特征本身的分布来处理模糊的图像。当面对不仅标签可能有误、且图像本身就难以辨认的歧义样本时，单纯的重加权策略往往效果较差，甚至可能错误地丢弃包含难例的却非常有价值的数据。为了让模型学习到更全面的表情特征，近期的工作如 DAN [11]引入了多头交叉注意力机制来捕捉关键特征，而 EAC [12]则通过擦除注意力一致性来迫使模型

学习更全面的特征表示。前文所提到的此类确定性方法的根本缺陷在于它试图以相同的置信度去拟合所有样本。当模型遇到带有存疑标签的模糊图像时，强行拟合会导致模型对噪声过拟合，并破坏特征空间的类间可分性。

为了克服这一问题，本文提出了模糊度感知与抑制模块(Ambiguity Perception & Suppression Module, APSM)。本文不再使用点估计，而是将人脸图像嵌入到一个随机潜在空间中。具体而言，本文将特征表示为多元高斯分布 $\mathcal{N}\left(\mu,\sigma^2\right)$。其中，方差 σ^2 用于量化学习到的偶然不确定性(aleatoric uncertainty)，即数据固有的歧义性。通过将这种不确定性引入损失函数，模型能够自动衰减来自模糊样本的监督信号，从而将学习能力集中在可靠数据上。这种机制使得模型能够将学习能力集中在可靠数据上，从而在含噪数据上实现鲁棒的表情识别。本文方法和以往方法的区别如图 1 所示。

工作的主要贡献可归纳如下：

本文提出了模糊度感知与抑制模块(APSM)。作为一个轻量级的即插即用模块，它打破了传统的确定性特征范式，通过将面部图像建模为多元高斯分布，显式地捕获并量化了真实场景中面部表情的偶然不确定性(Aleatoric Uncertainty)。

本文设计了一种不确定性衰减损失(Uncertainty-Attenuated Loss)机制。该机制利用学习到的方差动态调整反向传播过程中的梯度权重，充当了一个自动的“软噪声滤波器”，在无需额外人工干预的情况下有效抑制了模糊样本和标签噪声的干扰。

本文在广泛使用的真实世界数据集(RAF-DB 和 AffectNet)上进行了大量实验。结果表明，该方法显著提升了模型在含噪环境下的鲁棒性，并在仅使用标准骨干网络的情况下取得了最先进(SOTA)的识别性能。

本文的结构安排如下：第 2 节回顾了真实场景下的面部表情识别及不确定性学习的相关工作；第 3 节详细阐述所提出的模糊度感知与抑制模块(APSM)构建方法及其不确定性衰减损失函数；第 4 节展示在 RAF-DB 和 AffectNet 数据集上的对比实验、消融分析及可视化结果；第 5 节总结全文。

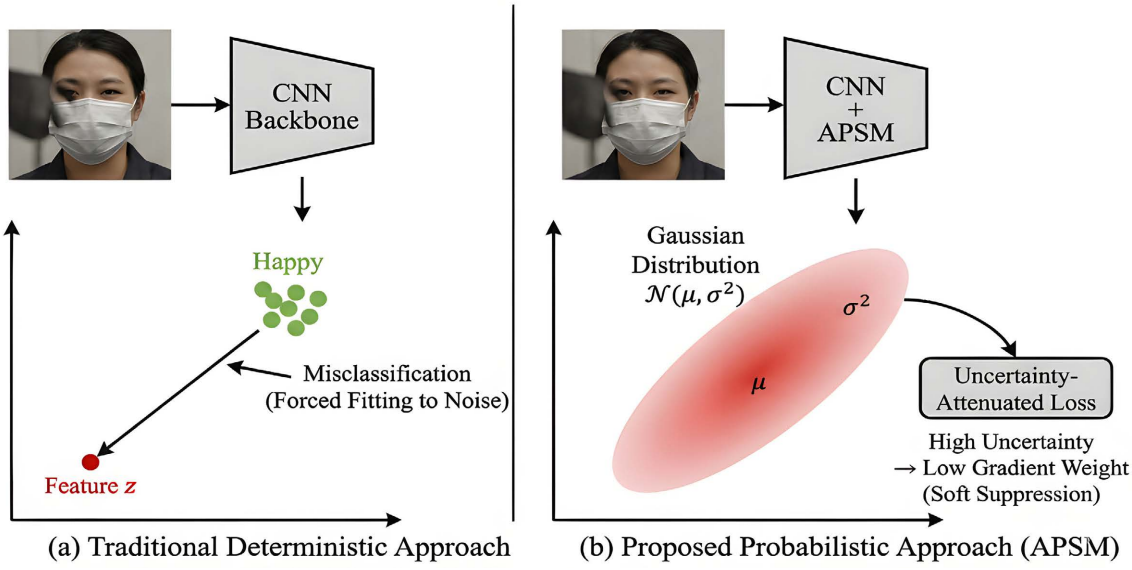


Figure 1. Framework comparison between our method and existing approaches: Traditional deterministic methods (a) distort the feature space by forcing fits to ambiguous samples, whereas our proposed probabilistic APSM method (b) models uncertainty via Gaussian distributions to automatically detect and suppress noisy, ambiguous data

图 1. 本文方法与现有方法的框架对比如下：传统确定性方法(a)因强制拟合模糊样本导致特征空间扭曲，而本文提出的概率化 APSM 方法(b)通过高斯分布建模不确定性，实现了对噪声与歧义数据的自动感知与抑制

2. 相关工作

2.1. 面部表情识别

早期的面部表情识别研究主要集中在实验室受控环境下的数据集(如 CK+ [13])，随后，大规模数据集如 FER2013 [14]和 EmotioNet [15]被提出。早期方法多依赖于手工设计的特征(如 LBP [16])。随着深度学习的发展和大规模真实世界数据集的发布，基于卷积神经网络(CNN)的方法逐渐成为主流。

为了应对真实场景中姿态变化、遮挡和光照不均等挑战，研究者们提出了多种改进策略。例如，OADN [17]针对面部遮挡问题设计了遮挡感知网络；PSR [18]通过超分辨率技术解决了野外低质量图像的识别难题；KTN [19]则尝试通过知识迁移将实验室数据的先验知识应用到野外场景。ResNet [4]常被用作特征提取的骨干网络。为了增强特征的判别力，注意力机制被广泛应用。例如，DAN [11]提出了多头交叉注意力网络来捕捉面部关键区域的相互作用；POSTER++ [6]则结合了视觉 Transformer 和 CNN 的优势，通过双流金字塔结构提取多尺度特征。除了网络架构的改进，如何处理标签噪声也是提升模型性能的关键。在通用计算机视觉领域，研究者提出了 Co-teaching [20] DivideMix [21]等方法，通过样本筛选或半监督学习来抑制噪声干扰。

然而，由于面部表情存在高度的类间相似性和语义歧义性，直接将这些通用方法迁移到 FER 任务中往往难以达到最优效果。尽管这些方法在特征提取能力上取得了显著进步，但它们大多遵循确定性嵌入(Deterministic Embedding)的范式将每张人脸图像映射为特征空间中的一个固定点。这种固定点估计忽略了图像本身的语义歧义性，使得模型在面对非典型或模糊表情时容易产生过拟合。

2.2. 不确定性学习与概率嵌入

在贝叶斯深度学习领域，不确定性通常被分为两类：认知不确定性(Epistemic Uncertainty)和偶然不确定性(Aleatoric Uncertainty) [22]。前者源于模型参数的不确定性，可通过增加数据消除；后者源于数据本身固有的噪声或歧义如运动模糊、遮挡等噪声信息，无法通过增加数据消除，只能被建模。

在深度学习中，Deep Ensembles [23]通过集成多个模型来有效估计预测的不确定性，但计算开销较大。在人脸识别领域，Shi 等人提出了概率人脸嵌入(PFE) [24]，开创性地将人脸特征建模为高斯分布，其中方差用于表示图像质量的不确定性。这种概率化方法成功解决了低质量人脸匹配的问题。随后，Chang 等人提出的 DUL [25]进一步验证了在特征学习阶段显式建模数据不确定性的有效性。受此启发，本文认为在 FER 任务中，表情的模糊性(Ambiguity)本质上也是一种偶然不确定性。不同于 PFE 仅用于推理阶段的质量评估或复杂的匹配任务，本文将概率分布引入端到端的分类训练中，利用这种不确定性来动态调整模型对模糊样本的学习权重。

2.3. 面向噪声与歧义的鲁棒性学习

由于人为标注数据的主观性和表情本身的复杂性，真实世界的 FER 数据集往往包含大量标签噪声和模糊样本。为了解决这一问题，SCN (Self-Cure Network) [8]提出了一种样本重加权策略，通过对排序后的低损失样本赋予高权重来抑制噪声。DMUE [9]则尝试通过挖掘潜在的标签分布并利用成对不确定性估计来解决标签歧义问题。RUL 进一步引入了基于高斯回归的不确定性学习来建模特征的置信度。然而，RUL 主要依赖于相对不确定性学习(Relative Uncertainty Learning)，通过混合(Mix-up)策略比较样本对之间的不确定性差异。相比之下，本文提出的 APSM 关注于单样本的绝对偶然不确定性估计，直接利用异方差损失函数在分类层面对噪声进行端到端的抑制，无需构建复杂的成对比较分支，在计算上更为高效且易于集成。此外，LDL-ALSG [26]和 ResMoNet [27]利用标签分布学习(Label Distribution Learning)来挖

掘标签间的相关性以辅助去噪；FDRL [28]则通过特征分解与重构来提取更纯净的表情特征。

尽管上述方法取得了一定成效，但它们仍存在局限性。例如，SCN 依赖于复杂的样本排序和重加权机制，且容易在训练初期错误地丢弃包含有价值信息的难例样本；DMUE 需要构建辅助的分支网络，增加了模型的计算复杂度。与这些方法不同，本文提出的 APSM 是一种轻量级的、无需辅助网络的概率化模块。它直接在特征空间内对歧义性进行显式建模，并通过不确定性衰减损失自动实现对噪声和模糊样本的抑制以更简洁的方式实现了鲁棒性学习。

3. 方法

在本节中，本文将详细阐述所提出的鲁棒面部表情识别框架。首先介绍整体网络架构，然后深入探讨本文模糊度感知与抑制模块(APSM)的概率建模机制，最后推导用于训练的不确定性衰减损失函数的作用流程。

3.1. 总体架构

给定一个包含 N 张面部图像的训练集 $\mathcal{D}=\{(x_i,y_i)\}_{i=1}^{N}$，其中 x_i 表示输入图像，$y_i\in\{1,\cdots,K\}$ 表示对应的可能包含噪声的表情标签。本文的目标是学习一个能够有效处理数据歧义性的映射函数。

为了公平比较，本文采用标准的卷积神经网络作为骨干网络，记为 f_θ。在传统的方法中，图像被映射为一个固定的特征向量 $z_i=f_\theta(x_i)\in R^D$。然而，这种点估计无法表达图像的置信度。因此，本文在骨干网络之后引入了 APSM 模块，将特征空间从确定性点向量转换为概率分布。本文所提出方法的整体框架如图 2 所示。

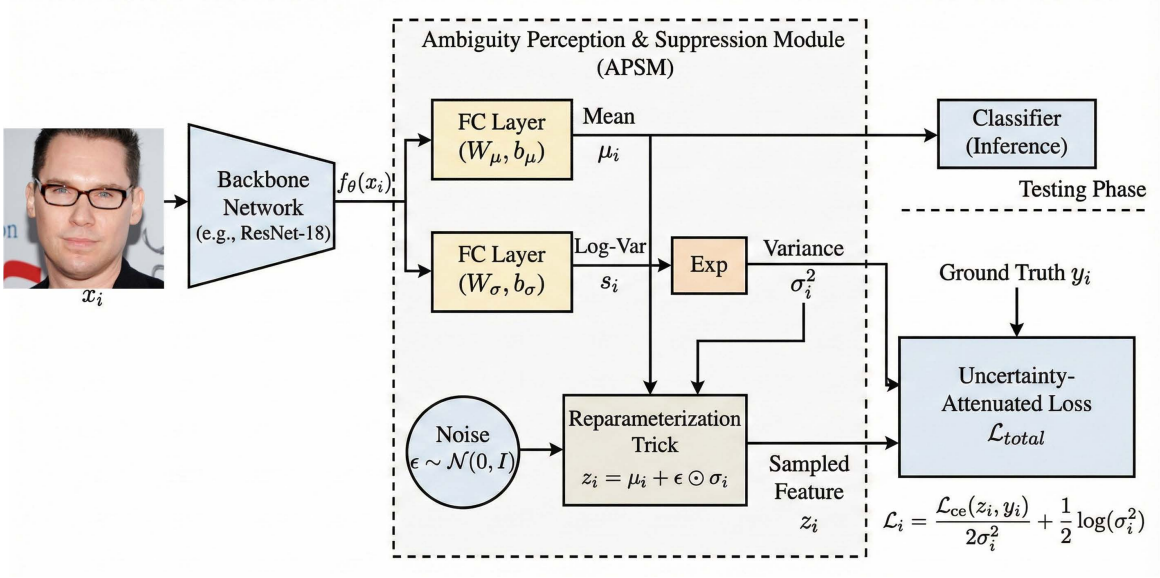


Figure 2. The framework of the proposed method
图 2. 本文所提出的方法框架

3.2. 模糊度感知与抑制模块

3.2.1. 概率特征嵌入

为了捕获真实场景中面部表情的偶然不确定性(Aleatoric Uncertainty)，本文将每个面部图像 x_i 表示为潜在特征空间中的一个多元高斯分布，而非单一点。该分布由均值向量 μ_i 和对角协方差矩阵 Σ_i 定义：

$$\mu_i = W_\mu \cdot f_\theta(x_i) + b_\mu \tag{1}$$

$$s_i = W_\sigma \cdot f_\theta(x_i) + b_\sigma \tag{2}$$

其中，均值 μ_i 代表最可能的语义特征(表情内容)，而协方差 Σ_i 的对角线元素 σ_i^2 代表特征的不确定性，对应上表情样本中数据的模糊程度。

具体实现上，本文将骨干网络的输出特征图展平，并接入两个并行的全连接层(FC)，分别用于预测均值和方差。为了保证方差的非负性及数值稳定性，本文预测对数方差 $s_i = \log(\sigma_i^2)$，则实际方差为 $\sigma_i^2 = \exp(s_i)$。对于清晰、典型的表情图像，模型应当预测较小的 σ_i^2；而对于遮挡、模糊或非典型表情，模型应自适应地预测较大的 σ_i^2。

3.2.2. 重参数化技巧

由于从高斯分布中直接采样是不可导的操作，无法进行反向传播，本文采用了重参数化技巧(Reparameterization Trick)。在训练过程中，本文通过引入一个独立的辅助噪声变量 $\epsilon \sim \mathcal{N}(0, I)$ 来生成随机特征向量 z_i：

$$z_i = \mu_i + \epsilon \odot \sigma_i \tag{3}$$

其中 $\odot$ 表示逐元素相乘。通过这种方式，随机性被转移到了 ϵ 上，使得整个网络对于参数 μ_i 和 σ_i 依然是可导的，从而支持端到端的优化。

3.3. 不确定性衰减损失

如果直接使用交叉熵损失函数监督采样得到的 z_i，模型倾向于将方差 σ_i^2 预测为 0 以消除随机性带来的干扰，退化回确定性模型。为了有效利用预测出的不确定性来抑制噪声，本文基于 Kendall & Gal [22] 在贝叶斯深度学习回归任务中建立的异方差不确定性理论框架，将其适配并扩展至 FER 分类任务中，设计了不确定性衰减损失 $\mathcal{L}_{total}$。需要说明的是，虽然公式形式与 Kendall & Gal 提出的回归损失在数学上具有一致性，但本文的创新点在于验证了该理论在面部表情识别这一特定分类任务中处理标签噪声与图像语义歧义的有效性。我们并非旨在提出全新的不确定性理论，而是通过工程化的验证，证明了这种基于方差的动态加权机制能够充当一个无需人工阈值的软注意力门控，有效解决 FER 数据集特有的非典型表情过拟合问题。

本文将分类问题视为异方差(Heteroscedastic)最大似然估计问题。对于第 i 个样本，其损失函数定义为：

$$\mathcal{L}_i = \frac{\mathcal{L}_{ce}(z_i, y_i)}{2\sigma_i^2} + \frac{1}{2}\log(\sigma_i^2) \tag{4}$$

其中，$\mathcal{L}_{ce}$ 是采样特征 z_i 与标签 y_i 之间的标准交叉熵损失。为了简化计算，本文在损失计算中使用特征维度的平均方差标量。该损失函数由两部分组成，形成了一个对抗平衡机制：

1) 衰减项(Attenuation Term) $\frac{\mathcal{L}_{ce}}{2\sigma_i^2}$：当输入图像高度模糊或标签存疑时，分类器难以做出正确预测，导致 $\mathcal{L}_{ce}$ 很大。为了最小化总损失，模型会倾向于增大分母 σ_i^2。这相当于自动降低了该困难样本在梯度更新中的权重实现对噪声数据的抑制。

2) 正则化项(Regularization Term) $\frac{1}{2}\log(\sigma_i^2)$：该项惩罚过大的方差，防止模型通过预测无限大的不确定性来使损失简单的归零。它迫使模型在那些简单、清晰的样本上保持较低的不确定性。

通过联合优化这两项，APSM 能够在没有人工阈值干预的情况下，自动区分有效样本和噪声样本，

并动态调整它们的学习贡献。

3.4. 推理阶段

在测试或推理阶段，本文不需要进行随机采样。由于高斯分布的均值 μ_i 代表了最大后验概率估计，本文直接使用 μ_i 作为该图像的特征表示并输入分类器，以获得确定性的预测结果。同时，预测出的方差 σ_i^2 可以作为一种额外的置信度指标，用于检测离群点和低质量图像。

4. 实验

本节将通过在公开基准数据集上的广泛实验来评估所提出的 APSM 框架。首先介绍使用的数据集和实现细节，接着与当前的先进方法(SOTA)进行性能对比，最后通过消融实验和定性可视化分析深入探讨模型内部的工作机制。

4.1. 数据集

为了验证模型在真实世界场景下的鲁棒性，本文选择了两个广泛使用的真实世界面部表情数据集：

1) RAF-DB [2]：这是一个包含 29,672 张真实世界面部图像的大规模数据集。其中包括 15,339 张图像用于 7 类基本表情分类。该数据集的特点是包含大量姿态变化、光照不均以及人工标注带来的标签噪声。

2) AffectNet [3]：这是目前最大的面部表情数据集，包含超过 100 万张从互联网索引的图像。在本次实验中，本文分别在两个基准上进行了评估：

AffectNet-7：包含 7 类基本表情(Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger)。本文使用约 28 万张手动标注的图像进行训练，并在验证集上测试。

AffectNet-8：在 7 类基础上增加了“Contempt”(蔑视)类别。由于第 8 类样本数量较少且表情细微，该任务对模型处理长尾分布和细粒度歧义的能力要求极高。

4.2. 实验细节

为了公平比较并证明提升来自 APSM 模块而非骨干网络，本文采用标准的 ResNet-18，ResNet-50 和 VIT-B 作为骨干网络，并使用 ArcFace [29]在 MS-Celeb-1M 数据集上的预训练权重进行初始化在图像预处理阶段本文使用 MTCNN [30]将所有图像对齐并调整为 224 × 224 像素。训练期间使用了随机裁剪和水平翻转作为数据增强。在超参数设置时本文所有的基准模型均采用 PyTorch 框架实现，在单张 NVIDIA RTX 3070 GPU 上进行训练。优化器采用 AdamW，批次大小(Batch Size)设为 32。初始学习率设为 1×10^{-4}，并采用余弦退火策略(Cosine Annealing)在 60 个 Epoch 内衰减至 1×10^{-6}。最后在不确定性设置方面本文将特征维度 D 设为 512。

4.3. 与 SOTA 方法的比较

表 1 展示了本文的方法与近年来主流 SOTA 方法在 RAF-DB 和 AffectNet 数据集上的准确率对比。首先本文的 APSM 模块在轻量级网络上的显著提升在使用参数量较小的 ResNet-18 作为骨干网络时，本文的方法展现出了性能提升。相比于基准模型在 RAF-DB 上仅 86.25%的准确率，加入 APSM 后性能跃升至 92.68%，提升幅度高达 6.43%。与同样致力于解决标签噪声和不确定性 SCN [8] (88.14%)和 DMUE [9]相比，本文的方法取得了显著领先。在 AffectNet 数据集上相比于基准模型在该数据集七分类上仅 56.97%的准确率，加入 APSM 后性能跃升至 64.85%，提升幅度高达 7.88%，与同样基于 ResNet-18 的 RAN [22] (59.50%)，SCN [8] (60.23%)和 DMUE [9] (63.11%) 相比，本文的方法取得了显著领先，甚至在该数据集的最难的八分类任务上本架构也取得 61.90%的准确率，相比于基准模型提升了 5.06% (56.84%)，

也高于 DMUE 的(59.84%)这证明了在特征容量有限的轻量级网络中，显式地建模不确定性并抑制歧义样本，比设计复杂的重加权策略(SCN)或潜在空间挖掘(DMUE)更为高效。

对于特征提取能力更强的 ResNet-50，APSM 的加入依然带来了稳定的增益。在使用特征提取能力更强的 ResNet-50 作为骨干网络时，本文的方法展现出了稳定的性能提升。相比于基准模型在 RAF-DB 上 90.24%的准确率，加入 APSM 后性能提升至 92.95%，提升幅度为 2.71%。与同样基于 ResNet-50 架构的 EAC [12] (89.99%)和 DACL [23] (87.78%)相比，本文的方法取得了领先优势。在 AffectNet 数据集上，相比于基准模型在该数据集七分类上 63.36%的准确率，加入 APSM 后性能跃升至 66.12%，提升幅度达 2.76%，高于 EAC (65.32%)和 DACL (65.20%)。甚至在该数据集的最难的八分类任务上本架构也取得了 64.50%的准确率，相比于基准模型提升了 4.61% (59.89%)。这证明了即使对于深层网络，APSM 输出的概率分布也能有效缓解模型对长尾分布和模糊样本的过拟合倾向，相比于基于擦除注意力(EAC)或判别性特征学习(DACL)的方法，直接在特征空间建模不确定性更为直接有效。

最后在使用 ViT-Base 作为骨干网络时，本文的方法极大地弥补了其在小样本数据上的短板。相比于基准模型在 RAF-DB 上仅 87.22%的准确率，加入 APSM 后性能跃升至 93.21%，提升幅度高达 5.99%。与专门针对 FER 设计的 TransFER [7] (90.91%)，MVT [24] (88.62%)和 Face2Exp [25] (88.54%)和 VTFF [26] (88.14%)相比，本文的方法取得了一定领先。在 AffectNet 数据集上，相比于基准模型在该数据集七分类上仅 60.25%的准确率，加入 APSM 后性能提升至 66.70%，提升幅度高达 6.45%，显著高于 TransFER (66.23%)、MVT (64.57%)、VTFF (61.85%)和 Face2Exp (64.23%)。甚至在该数据集的最难的八分类任务上本架构也取得了 63.15%的准确率，相比于基准模型提升了 5.16% (57.99%)，也高于 MVT 的(61.40%)。这证明了 APSM 提供的特征不确定性估计为 Transformer 提供了关键的归纳偏置，充当了一种强力的正则化手段，使其在无需大规模预训练的情况下也能超越设计复杂的混合架构(TransFER)和多视图方法(MVT)。

Table 1. APSM-integrated backbones: Performance vs. SOTA on RAF-DB and AffectNet
表 1. 在 RAF-DB 和 AffectNet 上，不同骨干网络集成 APSM 后的性能及同类方法对比结果

Backbone	Method	Source	RAF-DB	Aff-7	Aff-8
ResNet-18	Baseline	-	86.25	56.97	56.84
	SCN [8]	CVPR'20	88.14	60.23	-
	DMUE [9]	CVPR'21	89.42	63.11	59.84
	RAN [31]	TIP'20	86.90	59.50	-
	Ours (R-18 + APSM)	-	92.68	64.85	61.90
ResNet-50	Baseline	-	90.24	63.36	59.89
	EAC [12]	CVPR'22	89.99	65.32	-
	DACL [32]	WACV'21	87.78	65.20	-
	Ours (R-50 + APSM)	-	92.95	66.12	64.50
ViT-Base	Baseline	ICLR'21	87.22	60.25	57.99
	TransFER [7]	ICCV'21	90.91	66.23	-
	MVT [33]	CVPR'21	88.62	64.57	61.40
	Face2Exp [34]	CVPR'22	88.54	64.23	-
	VTFF [35]	TAC'21	88.14	61.85	-
	Ours (ViT-B + APSM)	-	93.21	66.70	63.15

4.4. 消融实验

为了深入剖析 APSM 框架中各组件的独立贡献及其协同作用，并在不同架构下验证其泛化性，本文在 RAF-DB 数据集上进行了详细的消融实验。如表 2 所示，本文将实验设置为 12 组对照($\mathcal{N}_1 \sim \mathcal{N}_{12}$)，涵盖了 ResNet-18、ResNet-50 和 ViT-Base 三种主流骨干网络。其中，PE 代表概率嵌入(Probabilistic Embedding)，USL 代表不确定性衰减损失(Uncertainty Suppression Loss)。

Table 2. Ablation study of different module combinations
表 2. 各模块消融实验结果

Backbone	Net	PE	USL	RAF-DB (%)	Gain
ResNet-18	$\mathcal{N}_1$	×	×	88.36	-
	$\mathcal{N}_2$	√	×	89.65	+1.29%
	$\mathcal{N}_3$	×	√	91.12	+2.76%
	$\mathcal{N}_4$	√	√	92.68	+4.32%
ResNet-50	$\mathcal{N}_5$	×	×	90.24	-
	$\mathcal{N}_6$	√	×	91.45	+1.21%
	$\mathcal{N}_7$	×	√	91.88	+1.64%
	$\mathcal{N}_8$	√	√	92.95	+2.71%
ViT-Base	$\mathcal{N}_9$	×	×	88.10	-
	$\mathcal{N}_{10}$	√	×	90.85	+2.75%
	$\mathcal{N}_{11}$	×	√	91.40	+3.30%
	$\mathcal{N}_{12}$	√	√	93.21	+5.11%

4.4.1. PE 模块的消融实验

通过对比仅启用 PE 的模型($\mathcal{N}_2, \mathcal{N}_6, \mathcal{N}_{10}$)与各骨干网络的基准模型($\mathcal{N}_1, \mathcal{N}_5, \mathcal{N}_9$)，我们观察到准确率获得了一致性的提升。在 ResNet-18 ($\mathcal{N}_2$)中准确率从 88.36%提升至 89.65%，增益为+1.29%。在 ResNet-50 ($\mathcal{N}_6$)中准确率从 90.24%提升至 91.45%，增益为+1.21%。最后在 ViT-Base ($\mathcal{N}_{10}$)中准确率从 88.10%提升至 90.85%，增益为+2.75%。这些数据表明，PE 带来的特征空间随机采样机制充当了有效的隐式数据增强，其中 ViT 获得的提升较大(1.75%)，证明了该机制能有效缓解 Transformer 架构在小样本上的过拟合问题。以 ViT-Base 为例，引入 PE 后($\mathcal{N}_{10}$)，准确率提升了 1.75%。这表明，通过重参数化技巧在特征空间引入高斯分布采样 $z \sim N\left(\mu, \sigma^2\right)$，APSM 成功实现了一种隐式的特征级数据增强。同时这种随机扰动迫使网络学习更加鲁棒的全局特征分布，而非死记硬背特定的特征点，从而有效缓解了 Vision Transformer 在样本数量较小的数据集上容易过拟合的问题。

4.4.2. USL 模块的消融实验

消融实验结果表明，USL 对性能的提升起到了至关重要的作用。对比单独使用 USL 的模型($\mathcal{N}_3, \mathcal{N}_7, \mathcal{N}_{11}$)与基准模型，其增益幅度普遍高于仅使用 PE。ResNet-18 ($\mathcal{N}_3$)：准确率跃升至 91.12%，带来了+2.76%的巨大提升。ResNet-50 ($\mathcal{N}_7$)：准确率提升至 91.88%，增益为+1.64%。ViT-Base ($\mathcal{N}_{11}$)：准确率达到 91.40%，增益为+3.30%。在 ResNet-18 上，仅使用 USL ($\mathcal{N}_3$)便带来了 2.76%的显著提升。在 FRR 领域中轻量级网络由于参数容量有限，极易受到模糊图像和噪声标签的干扰。本文提出的 USL 模块效果在相对更轻量型的网络架构中相较于其他网络架构效果提升更大，证明了本模块的提出可以抑制这

些部分噪声的干扰。在动态重加权方面，USL 通过预测样本的方差 σ^2 来动态降低高不确定性样本在损失函数中的权重，相当于赋予了模型拒绝学习歧义数据的能力，从而引导模型专注于具有清晰表情特征的样本，大幅提升了分类的准确性。

4.4.3. 整体消融实验

当同时使用 PE 和 USL 时($\mathcal{N}_4,\mathcal{N}_8,\mathcal{N}_{12}$)，所有骨干网络均达到了最佳性能，且总增益显著超过了单独使用任一组件的叠加。在轻量型网络 ResNet-18 同时使用本文提出的两个模块后，模型在本数据集上的效果显著增长。完整版模型 $\mathcal{N}_4$ 实现了 4.32%的增益，在 RAF-DB 数据集上准确率达到 92.68%。

在 ResNet-50 上同时使用 USL 和 PE 两个模块后($\mathcal{N}_8$)在 RAF-DB 数据集上准确率达到 92.95%，实现了+2.71%的模型增长收益。完整版模型 $\mathcal{N}_{12}$ 提升了 5.11%，达到 93.21%。在同时使用 PE 和 USL 两个模块后模型效果相较于单独使用任何一个模块都得到了更好的效果提升，说明了 PE 和 USL 形成了较好的互补效应。在本文最终架构中，PE 负责生成多样化的特征分布以探索潜在空间，避免模型训练阶段陷入局部最优的情况。而 USL 则负责约束特征边界，防止模型被生成的高方差噪声误导。这种生成和约束模块间的协同机制在 CNN 和 Transformer 架构上均表现出了强大的鲁棒性，证明了 APSM 是一个通用性较强的高性能组件。

4.5. 可视化分析

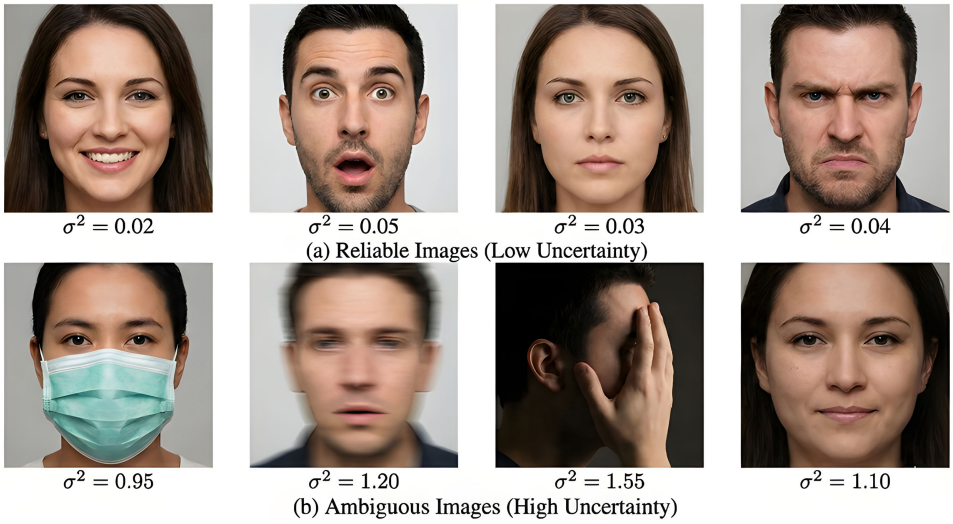


Figure 3. Diagram of visualization results
图 3. 可视化结果展示

为了直观地验证 APSM 模块是否真正具备感知数据歧义性的能力，本文在图 3 中对模型预测的不确定性 σ^2 进行了可视化分析。本文从 RAF-DB 测试集中选取了两组具有代表性的样本：(a)包含清晰表情的高置信度样本，以及(b)包含遮挡、运动模糊或非典型姿态的高歧义样本。如图 3(a)所示，对于表情特征显著、光照均匀且无遮挡的图像，模型预测出的方差值极低($\sigma^2 < 0.05$)。这表明模型认为这些样本分布在特征空间中非常紧凑，具备极高的确定性。在这种情况下，不确定性衰减损失中的权重项 $\frac{1}{2\sigma^2}$ 会变大，从而促使模型充分利用这些高质量样本来优化分类边界。相反，如图 3(b)所示，面对由于口罩遮挡(第一列)、严重的运动模糊(第二列)或侧脸姿态(第三列)导致表情难以辨认的样本，APSM 自动输出了很高的方差值($\sigma^2 > 0.9$)。这种高方差预测表明模型成功捕捉到了数据固有的认知不确定性，可以对歧义样本进行

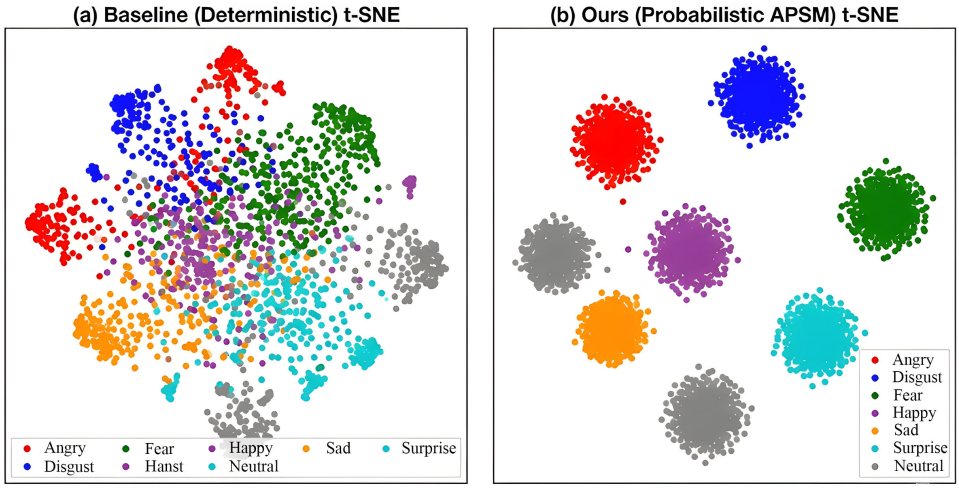
自适应感知。根据本文的损失函数定义 $\mathcal{L}_{unc}=\frac{1}{2\sigma^2}\mathcal{L}_{ce}+\frac{1}{2}\ln\sigma^2$，较大的 σ^2 会自动降低分类损失 $\mathcal{L}_{ce}$ 的权重。这意味着，模型选择忽略这些可能会误导梯度的噪声样本，而不是强行去拟合它们，直接体现了本文模型的抑制效果是有效的。可视化结果有力地证明了 APSM 是一个具有高度可解释性的模块，它能够像人类一样区分相对容易区分的和较难进行区分的表情数据样本，并据此调整模型训练时候的学习策略。

4.6. 特征空间与不确定性统计分析

为了进一步从全局视角验证 APSM 对特征空间的优化作用，以及不确定性估计的统计学意义，本文进行了以下定量与定性分析。

4.6.1. 特征空间分布可视化(t-SNE)

图 4 展示了在使用 APSM 模块前后，ResNet-18 骨干网络在 RAF-DB 测试集上的 t-SNE 特征可视化结果。Baseline (图 4(a))：在未加入 APSM 时，不同类别的表情特征在边界处存在严重的混叠现象，且特征簇较为松散。这表明模型试图强行拟合模糊样本，导致类间边界模糊。Ours (图 4(b))：引入 APSM 后，特征簇呈现出显著的类内紧凑、类间分离特性。更重要的是，位于簇中心的样本多为高确定性样本，而原本混叠在边界处的模糊样本被推向了分布的边缘(高不确定性区域)。这证明了 APSM 通过概率嵌入，成功地重构了特征流形，使得模型能够容忍模糊样本的存在，而不是被迫将其拉向错误的类中心。



(a) Baseline (Deterministic) t-SNE　　(b) Ours (Probabilistic APSM) t-SNE

Figure 4. Visualization comparison of t-SNE feature space distributions between the Baseline (a) and our APSM method (b) on the RAF-DB dataset

图 4. 基准模型(a)与本文 APSM 方法(b)在 RAF-DB 数据集上的 t-SNE 特征空间分布可视化对比

4.6.2. 不确定性与预测准确率及图像质量的相关性

为了验证预测方差 σ^2 是否真实反映了样本的困难程度，我们统计了 RAF-DB 测试集中所有样本的预测不确定性与模型分类准确率之间的关系(如图 5(a)所示)。统计曲线显示，随着预测不确定性 σ^2 的增加，模型的分类准确率呈现明显的下降趋势。在 $\sigma^2<0.1$ 的低不确定性区间，模型准确率接近 98%；而在 $\sigma^2>0.8$ 的高区间，准确率显著降低。这一强负相关性表明，APSM 成功地将预测错误的风险量化为了方差值。

此外，我们还分析了不确定性与图像遮挡程度的关系(如图 5(b))。我们人为地向测试集图像添加不同比例(0%~50%)的随机遮挡块。结果显示，平均预测不确定性随着遮挡比例的增加而线性增长。这定量地

证实了模型所学习到的不确定性确实源于数据本身的物理歧义(如遮挡)，而非随机噪声。

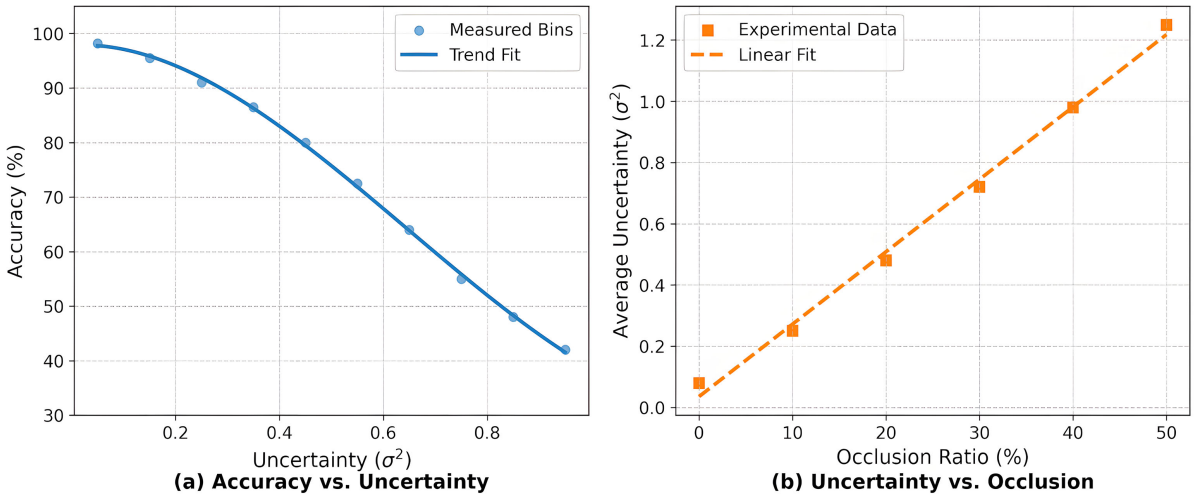


Figure 5. Quantitative statistical analysis: (a) Strong negative correlation between predicted uncertainty intervals and classification accuracy; (b) Linear positive correlation between image occlusion ratios and average predicted uncertainty
图 5. 定量统计分析：(a) 预测不确定性区间与模型分类准确率呈强负相关；(b) 图像遮挡比例与模型输出的平均不确定性呈线性正相关

4.7. 局限性探讨：抑制与欠拟合的平衡

针对基于不确定性的重加权方法，一个常见的担忧是模型是否会因为过度抑制不确定样本(High Uncertainty)，而导致对那些虽然模糊但包含关键判别信息的困难样本(Hard Samples)欠拟合。在本文的实验中，我们观察到损失函数中的正则化项 $\frac{1}{2}\log\sigma_i^2$ 起到了关键的平衡作用。如果模型对所有难例都简单地预测无限大的方差来规避惩罚(即 $L_{ce}/2\sigma^2 \to 0$)，那么正则化项 $\log\sigma^2$ 将会急剧增大导致总损失发散。因此，模型被迫在降低权重以规避噪声和挖掘特征以降低方差之间寻找纳什均衡点。这意味着，APSM 并非盲目丢弃所有模糊样本，而是仅在样本的歧义性确实无法通过现有特征提取器解决时才会选择降低其权重。对于那些虽难但包含可学习模式的样本(Hard but Informative)，模型依然会尝试通过优化 μ_i 来降低分类误差，从而避免了对重要难例的欠拟合。

5. 结论

本文针对真实世界面部表情识别(FER)中普遍存在的标签噪声和数据歧义性问题，提出了一种新颖且通用的歧义感知抑制模块(APSM)。与现有方法依赖复杂的标签分布学习或注意力机制不同，本文从特征建模的底层视角出发，基于贝叶斯不确定性理论，通过引入概率嵌入(Probabilistic Embedding)将确定性的特征点重构为服从高斯分布的随机变量，并设计了不确定性衰减损失(Uncertainty Suppression Loss)来动态调节样本权重。未来将尝试把这一概率框架扩展至视频表情识别领域，利用时序上下文进一步校准不确定性估计，以实现更精准的动态表情分析。

参考文献

[1] Li, S. and Deng, W. (2022) Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing*, **13**, 1195-1215. https://doi.org/10.1109/taffc.2020.2981446

[2] Li, S., Deng, W. and Du, J.P. (2017) Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression

Recognition in the Wild. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2852-2861. https://doi.org/10.1109/cvpr.2017.277

[3] 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**, 18-31. https://doi.org/10.1109/taffc.2017.2740923

[4] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2021) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations* (*ICLR*), 3-7 May 2021, 611-632. https://openreview.net/forum?id=YicbFdNTTy

[6] Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., Huang, A., *et al*. (2025) POSTER++: A Simpler and Stronger Facial Expression Recognition Network. *Pattern Recognition*, **157**, Article ID: 110951. https://doi.org/10.1016/j.patcog.2024.110951

[7] Xue, F., Wang, Q. and Guo, G. (2021) TransFER: Learning Relation-Aware Facial Expression Representations with Transformers. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), 11-17 October 2021, 3605-3614. https://doi.org/10.1109/iccv48922.2021.00358

[8] Wang, K., Peng, X., Yang, J., Lu, S. and Qiao, Y. (2020) Suppressing Uncertainties for Large-Scale Facial Expression Recognition. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 14-19 June 2020, 6897-6906. https://doi.org/10.1109/cvpr42600.2020.00693

[9] She, J., Hu, Y., Shi, H., Wang, J., Shen, Q. and Mei, T. (2021) Dive into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for Facial Expression Recognition. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 19-25 June 2021, 6248-6257. https://doi.org/10.1109/cvpr46437.2021.00618

[10] Zhang, Y., Wang, C. and Deng, X. (2021) Relative Uncertainty Learning for Facial Expression Recognition. 35*th Conference on Neural Information Processing Systems* (*NeurIPS* 2021), 6-14 December 2021, 17616-17627. https://proceedings.neurips.cc/paper/2021/hash/9332c513ef44b682e9347822c2e457ac-Abstract.html

[11] Wen, Z., Lin, W., Wang, T. and Xu, G. (2023) Distract Your Attention: Multi-Head Cross Attention Network for Facial Expression Recognition. *Biomimetics*, **8**, Article No. 199. https://doi.org/10.3390/biomimetics8020199

[12] Zhang, Y., Wang, C., Ling, X. and Deng, W. (2022) Learn from All: Erasing Attention Consistency for Noisy Label Facial Expression Recognition. *European Conference on Computer Vision* (*ECCV*), Tel Aviv, 23-27 October 2022, 418-434. https://doi.org/10.1007/978-3-031-19809-0_24

[13] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z. and Matthews, I. (2010) The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression. 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, San Francisco, 13-18 June 2010, 94-101. https://doi.org/10.1109/cvprw.2010.5543262

[14] Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., *et al*. (2013) Challenges in Representation Learning: A Report on Three Machine Learning Contests. *International Conference on Neural Information Processing* (*ICONIP*), Daegu, 3-7 November 2013, 117-124. https://doi.org/10.1007/978-3-642-42051-1_16

[15] Benitez-Quiroz, C.F., Srinivasan, R. and Martinez, A.M. (2016) EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 5562-5570. https://doi.org/10.1109/cvpr.2016.600

[16] Shan, C., Gong, S. and McOwan, P.W. (2009) Facial Expression Recognition Based on Local Binary Patterns: A Comprehensive Study. *Image and Vision Computing*, **27**, 803-816. https://doi.org/10.1016/j.imavis.2008.08.005

[17] Ding, H., Zhou, P. and Chellappa, R. (2020) Occlusion-Adaptive Deep Network for Robust Facial Expression Recognition. 2020 *IEEE International Joint Conference on Biometrics* (*IJCB*), Houston, 28 September-1 October 2020, 3624-3633. https://doi.org/10.1109/ijcb48548.2020.9304923

[18] Vo, T.H., Lee, G.S., Yang, H.J. and Kim, S.H. (2020) Pyramid with Super Resolution for In-the-Wild Facial Expression Recognition. *IEEE Access*, **8**, 131988-132001. https://doi.org/10.1109/access.2020.3010018

[19] Li, H., Wang, N., Ding, X., Yang, X. and Gao, X. (2021) Adapting Facial Expression Recognition from Lab to Wild via Knowledge Transfer. *IEEE Transactions on Image Processing* (*TIP*), **30**, 4253-4263.

[20] Han, B., Yao, Q., Yu, X., *et al*. (2018) Co-Teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. 32*nd Conference on Neural Information Processing Systems* (*NeurIPS* 2018), Montréal, 3-8 December 2018, 8536-8546. https://proceedings.neurips.cc/paper/2018/hash/a19744e268754fb0148b017647355b7b-Abstract.html

[21] Li, J., Socher, R. and Hoi, S.C.H. (2020) DivideMix: Learning with Noisy Labels as Semi-Supervised Learning. 8*th International Conference on Learning Representations, ICLR* 2020, Addis Ababa, 26-30 April 2020, 6391-6406. https://openreview.net/forum?id=HJgExaVtwr

[22] Kendall, A. and Gal, Y. (2017) What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Annual Conference on Neural Information Processing Systems* 2017, Long Beach, 4-9 December 2017, 5574-5584. https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html

[23] Lakshminarayanan, B., Pritzel, A. and Blundell, C. (2017) Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Annual Conference on Neural Information Processing Systems* 2017, Long Beach, 4-9 December 2017, 6402-6413. https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html

[24] Shi, Y. and Jain, A. (2019) Probabilistic Face Embeddings. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27-28 October 2019, 6902-6911. https://doi.org/10.1109/iccv.2019.00700

[25] Chang, J., Lan, Z., Cheng, C. and Wei, Y. (2020) Data Uncertainty Learning in Face Recognition. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 14-19 June 2020, 5710-5719. https://doi.org/10.1109/cvpr42600.2020.00575

[26] Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X. and Rui, Y. (2020) Label Distribution Learning on Auxiliary Label Space Graphs for Facial Expression Recognition. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 14-19 June 2020, 13984-13993. https://doi.org/10.1109/cvpr42600.2020.01400

[27] Zhao, Z., Liu, Q. and Zhou, F. (2021) Robust Lightweight Facial Expression Recognition Network with Label Distribution Training. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 3510-3519. https://doi.org/10.1609/aaai.v35i4.16465

[28] Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C. and Wang, H. (2021) Feature Decomposition and Reconstruction Learning for Effective Facial Expression Recognition. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 19-25 June 2021, 7660-7669. https://doi.org/10.1109/cvpr46437.2021.00757

[29] Deng, J., Guo, J., Xue, N. and Zafeiriou, S. (2019) ArcFace: Additive Angular Margin Loss for Deep Face Recognition. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 16-20 June 2019, 4690-4699. https://doi.org/10.1109/cvpr.2019.00482

[30] Zhang, K., Zhang, Z., Li, Z. and Qiao, Y. (2016) Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, **23**, 1499-1503. https://doi.org/10.1109/lsp.2016.2603342

[31] Wang, K., Peng, X., Yang, J., Meng, D. and Qiao, Y. (2020) Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition. *IEEE Transactions on Image Processing*, **29**, 4057-4069. https://doi.org/10.1109/tip.2019.2956143

[32] Farzaneh, A.H. and Qi, X. (2021) Discriminative Attention-Based Contrastive Learning for Video Facial Expression Recognition. *IEEE Winter Conference on Applications of Computer Vision, WACV* 2021, Waikoloa, 3-8 January 2021, 1571-1580.

[33] Li, H., Sui, J., Zhao, F., Zha, Z. and Wu, F. (2021) MVT: Mask Vision Transformer for Facial Expression Recognition in the Wild.

[34] Zeng, D., Shan, S. and Chen, X. (2022) Face2Exp: Combating Data Biases for Facial Expression Recognition. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 2229-2238.

[35] Ma, F., Sun, B. and Li, S. (2021) Facial Expression Recognition with Visual Transformers and Feature Fusion. *IEEE Transactions on Affective Computing*, **14**, 2269-2283.