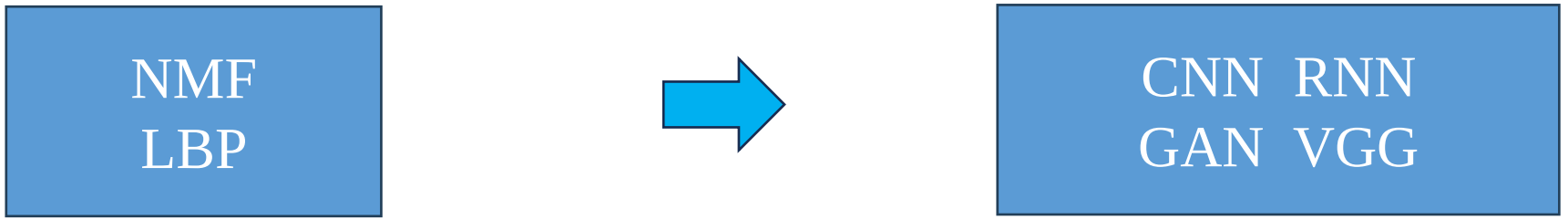
一种用于面部表情识别的双向注意混合特征网络

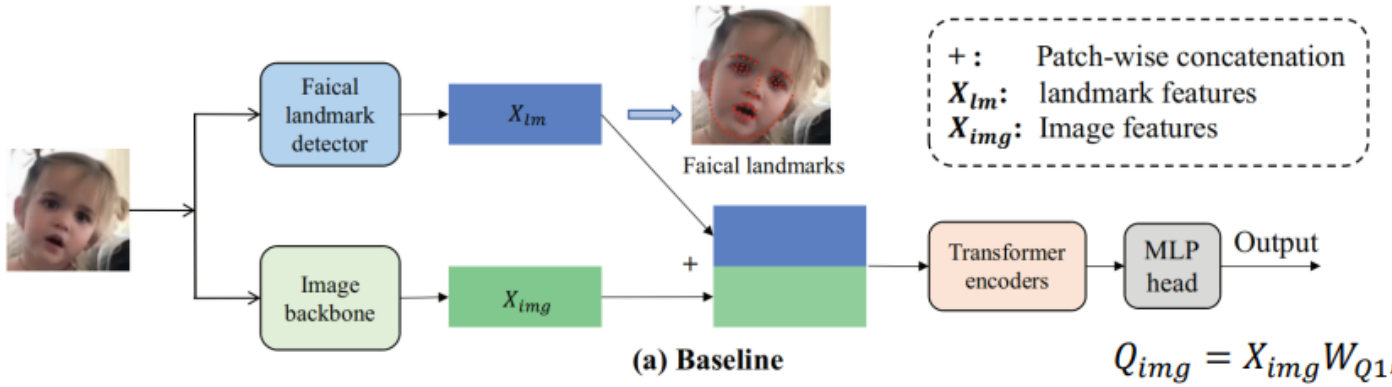
RESEARCH BACKGROUND - 面部表情学习

- 面部表情识别（Facial Expression Recognition, FER）作为一个重要的研究领域，已受到学术界数十年的关注。传统的FER方法依赖于手工特征或浅层学习技术，如非负矩阵分解（Non-Negative Matrix Factorization, NMF）、局部二值模式（Local Binary Patterns, LBP）以及稀疏学习。
- 近年来，深度学习技术彻底变革了计算机视觉领域，为FER带来了显著进展。卷积神经网络（Convolutional Neural Networks, CNNs）、递归神经网络（Recurrent Neural Networks, RNNs）以及生成对抗网络（Generative Adversarial Networks, GANs）被广泛用于应对FER的复杂挑战。



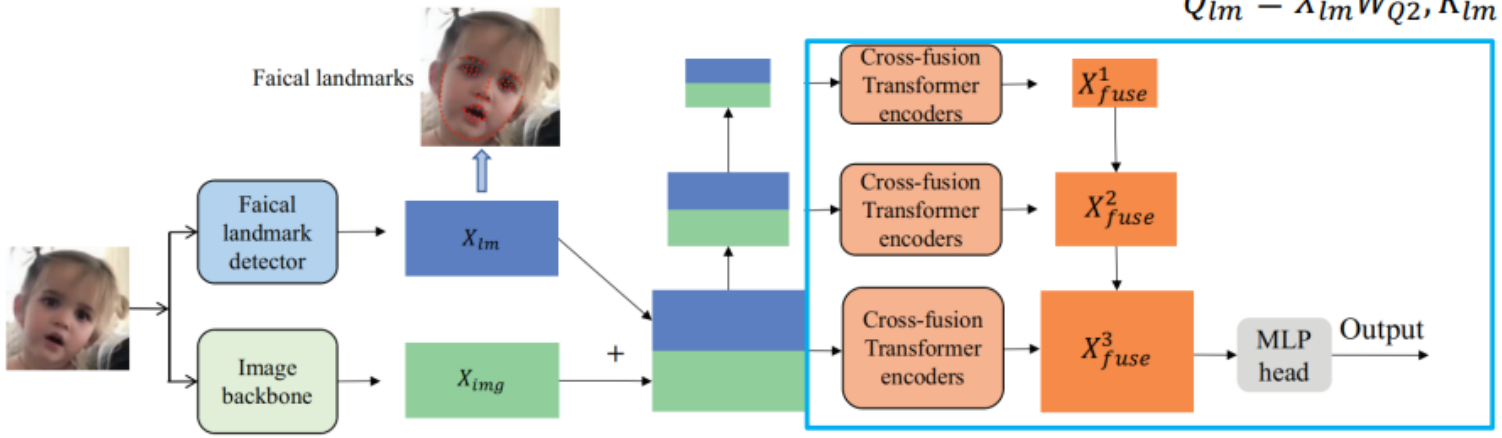
RESEARCH BACKGROUND – 注意力机制

- 近年来，Transformer[16]作为一种强大的新范式在许多任务中表现卓越，其多头注意力机制使其在多个领域超越了传统的递归神经网络（Recurrent Neural Networks, RNNs）和卷积神经网络（CNNs）。这一成功促使研究人员探索并改进基于Transformer的方法，用于解决各种视觉任务。
- 在面部表情识别（FER）领域，也有研究引入了注意力机制：



(a) Baseline

$$Q_{img} = X_{img}W_{Q1}, K_{img} = X_{img}W_{K1}, V_{img} = X_{img}W_{V1},$$
$$Q_{lm} = X_{lm}W_{Q2}, K_{lm} = X_{lm}W_{K2}, V_{lm} = X_{lm}W_{V2},$$

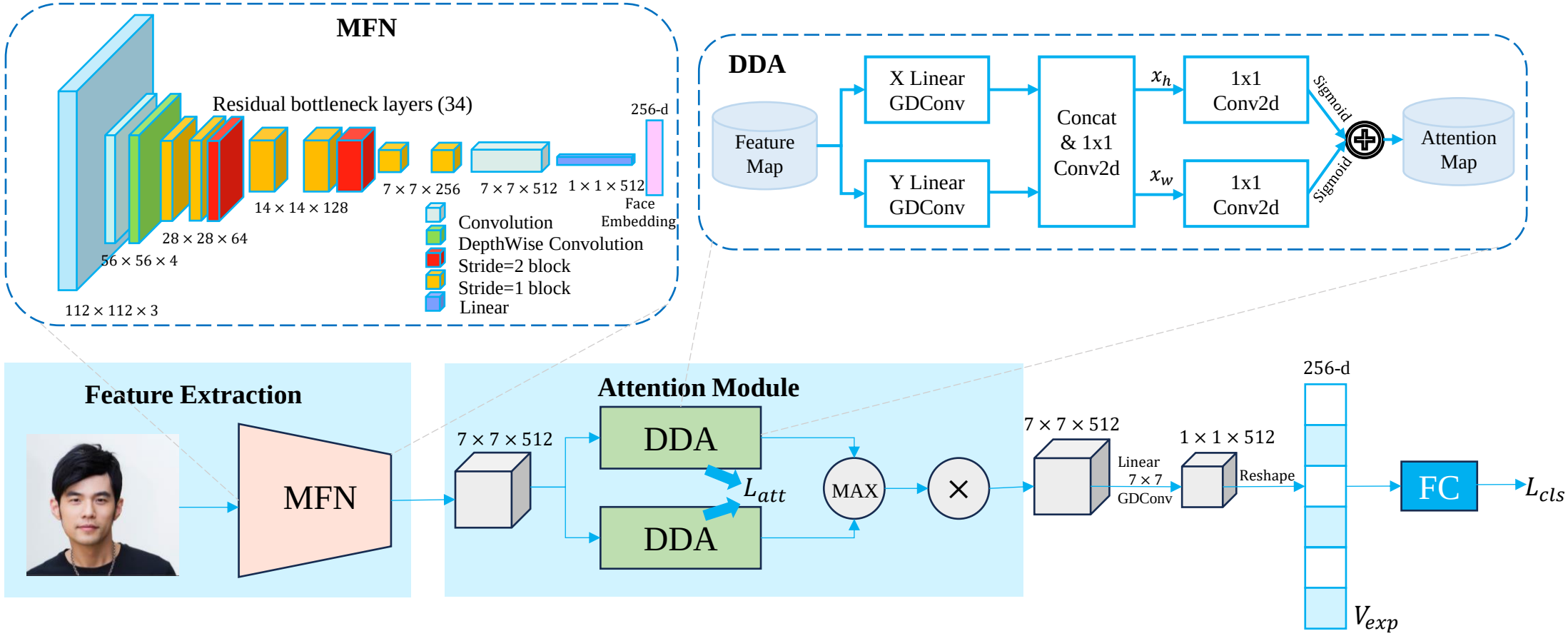


(b) POSTER

对得到的image feature和landmark feature分别点乘Q, K, V做交叉注意力机制

MAIN METHOD

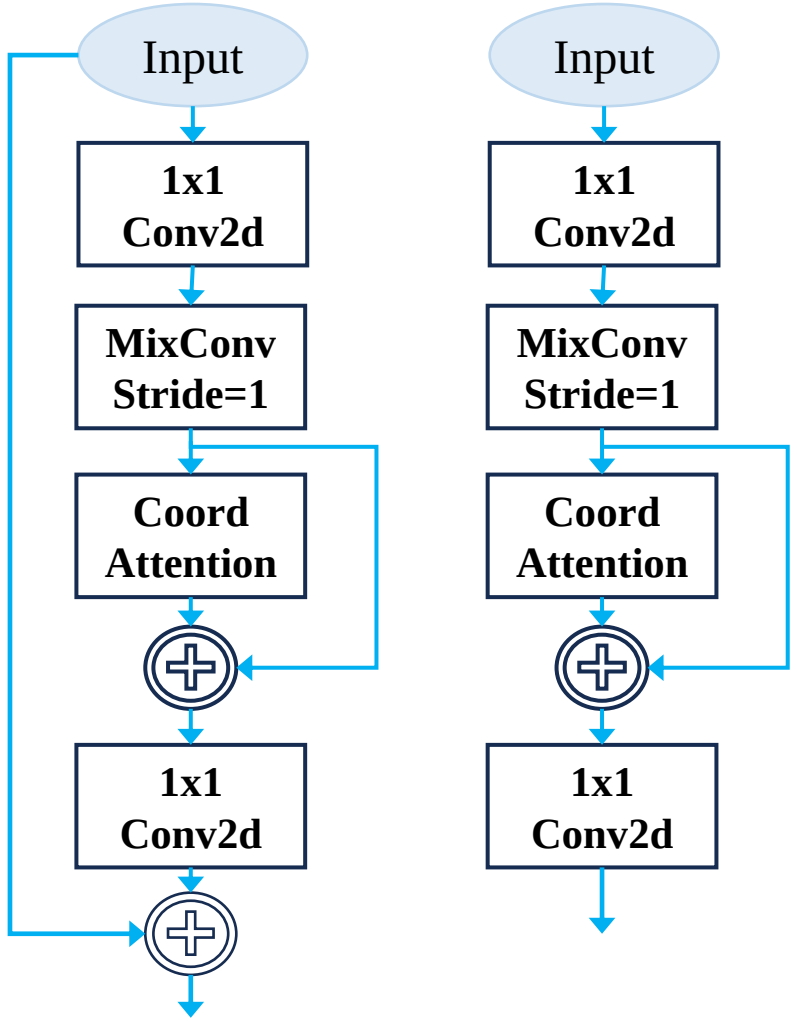
- DDAMFN的整体架构主要由两部分组成：混合特征网络（MFN）和双向注意网络（DDAN）。首先，将面部图像输入MFN，生成基础特征图作为输出。接着，通过DDAN在垂直和水平方向生成注意力图。最终，注意力图被重塑为特定维度，并通过全连接层预测图像的表情类别。



混合特征网络（MFN）

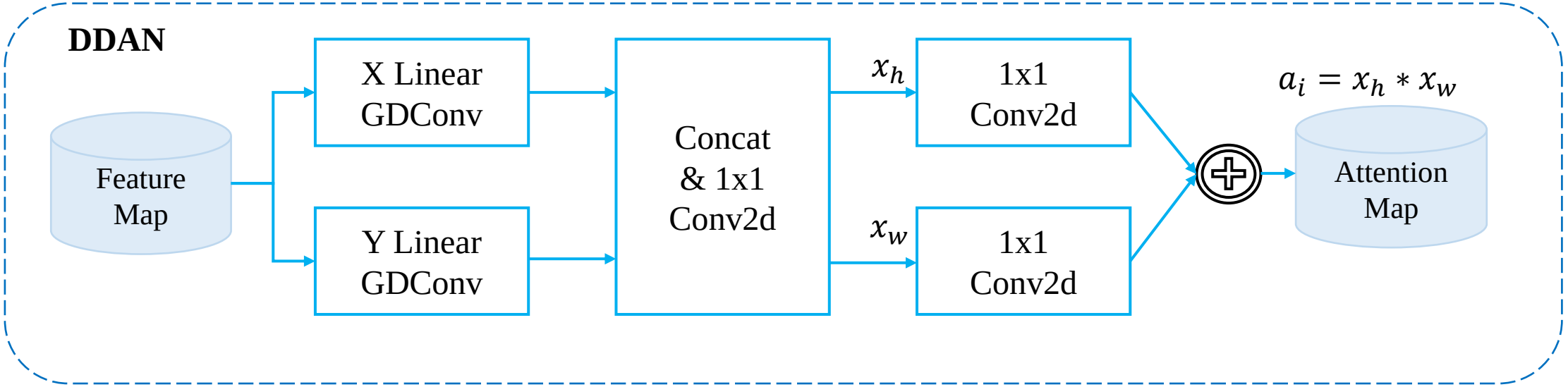
- MFN由两种主要构建模块组成：残差瓶颈块和非残差块。

- 残差瓶颈块旨在捕获复杂特征，并促进网络中的信息传播。通过引入残差连接，该模块能够缓解退化问题，并在训练过程中改善梯度流动。
- 非残差块则通过非残差连接增强模型的表示能力，使MFN能够捕获多样化且具有区分性的面部特征，从而更有效地完成FER任务。



双向注意网络（DDAN）

- DDAN由多个独立的双向注意（DDA）头组成，每个头都用于捕获网络中的长距离依赖关系。DDAN模块的基础结构采用了坐标注意力机制。
- 注意力头最初从水平方向和垂直方向生成方向感知特征图。然而，本文用线性深度卷积（GDConv）替代了传统的平均池化操作，这种修改可以在不同的空间位置学习到截然不同的重要性。



双向注意网络（DDAN）

- 为了确保每个双向注意力头专注于不同的面部区域，为DDAN模块引入了一种新颖的损失函数，称为注意力损失（Attention Loss）。
- 注意力损失通过计算不同双向注意力头生成的注意力图之间的均方误差（MSE）定义。具体来说，注意力损失是这些MSE损失之和的倒数，其数学表达式为：

$$L_{att} = \left(\sum_{i=0}^{n} \sum_{k=0,k\neq i}^{n} MSE(a_i, a_k) \right)^{-1}$$

  - n表示注意力头的数量，
  - a_i和a_k分别是两个不同头生成的注意力图。

TOTAL LOSS FUNCTION

- 在损失函数方面，训练过程中采用标准的交叉熵损失。该损失函数有效地衡量了预测类别概率与真实标签之间的差异，有助于优化模型参数。
- 整体损失函数可以表示为：

$$L = L_{cls} + \lambda_a L_{att},$$

- 其中，L_{cls}表示标准的交叉熵损失，L_{att}是注意力损失，λ_a是超参数，默认值为0.1。

EXPERIMENT 1 – RAG-DB

- **Data Pre-processing**

- 使用RetinaFace 来检测RAF-DB数据集中的面部和关键点（包括两只眼睛、鼻子和两个嘴角）。

- 所有的面部图像都被对齐并调整为标准化的112×112像素大小。

- 对于RAF-DB使用了水平翻转、随机旋转和擦除来进行数据增强，提高了DDAMFN在训练过程中的鲁棒性和泛化能力。

EXPERIMENT 1 – K-FOLD CROSS VALIDATION

在不同的交叉验证过程中，达到了平均90.635的验证水平。

数据集	RAF-DB
Fold 1	90.69
Fold 2	90.23
Fold 3	89.89
Fold 4	90.82
Fold 5	89.83
Fold 6	91.02
Fold 7	91.82
Fold 8	90.55
Fold 9	90.55
Fold 10	90.95
平均准确率 (%)	90.635

EXPERIMENT 2 – COMPARISONS

- 为了证明提出的模型在RAF-DB数据集上的有效性，与7个baseline进行了对比，对比结果如下图所示，发现DDAMFN在RAF-DB数据集上具有较好的表现。

Methods	Accuracy (%)
RAN[30]	86.90
SCN[31]	87.03
DACL[28]	87.78
MViT[27]	88.62
PSR[32]	88.98
DAN[17]	89.70
TransFER[18]	90.91
DDAMFN	90.635

EXPERIMENT3 – ABLATION STUDY

- MFN的有效性
- 为了评估MFN骨干网络的有效性，开展了系列对比实验：

方法	准 确 率 (%)	参数量 (Params)	FLOPs
MobileFaceNet	87.52	1.148 M	230.34 M
ResNet-18	87.47	16.78 M	2.6 G
ResNet-50	89.63	41.56 M	6.31 G
MFN (our backbone)	90.32	3.973 M	550.74 M
DDAMFN (our model)	91.35	4.106 M	551.22 M

- MFN在RAF-DB数据集上的准确率为90.32%，超越了其他三个骨干网络的表现。

EXPERIMENT3 – ABLATION STUDY

- DDAN的有效性
- 为了验证DDAN的有效性，进行了消融实验，评估了MFN和DDAN在RAF-DB数据集上的影响：

MFN	DDAN	RAF-DB (%)
√	√	91.35
√	-	90.32

- DDAN的引入使得模型在RAF-DB和数据集上的性能提升了1.06%.

EXPERIMENT3 – ABLATION STUDY

- 注意力头数的影响
- 为了研究DDA头数量对模型性能的影响，进行了关于DDA头数量变化的实验：

DDA头数量	RAF-DB (%)
0	90.32
1	90.67
2	91.35
3	91.11
4	91.21

- 使用两个DDA头的模型在RAF-DB数据集上取得了显著优于其他DDA头数量的表现。

THANK YOU FOR LISTENING!

汇报人：许笑颜
SA24229007