**Research on the Structure of Offline Handwritten Signature Verification Models Based on Transformer**

**(基於Transformer的離線手寫簽名驗證模型結構研究)**

**Abstract（摘要）**

離線手寫簽名驗證是生物特徵技術的應用*,* 其根據用戶提供的手寫簽名与存儲的参考簽名對比以驗證用戶身份, 被廣泛用於安全認證, 金融交易等安全領域. 學術研究中, 學者們定義了作者獨立和作者依賴任務以評估離綫手寫簽名驗證模型的泛化和判斷性能.

深度學習的視覺模型在離綫手寫簽名驗證領域中取得了顯著進展，但是模型對高像素的手寫簽名圖像特徵學習能力具有一定的缺陷, 并且注意力機制的Transformer具有訓練過程模型收斂速度慢的缺點. 為了在一定程度上解決該問題, 本工作基於Transformer提出OSVTF端到端模型結構, 在組合特徵的基礎上增加多尺度融合特徵, 并且對部分子網絡進行了加速訓練收斂的優化.

在實驗部分, OSVTF模型在BHSig-B 80/20數據集的作者獨立和作者依賴任務中EER為4.54%和2.13%, 在BHSig-H 100/60數據集的EER為3.90%和2.68%, 達到了日常生產使用的水準. 在消融實驗部分, OSVTF在CEDAR 50/5數據集的作者獨立任務中EER為4.75%, 證實加速訓練收斂優化的有效性. 整體實驗結果證實在跨數據集的場景下, OSVTF相較過去的離綫簽名驗證模型在性能上取得了一定程度的提升, 在此基礎上具有更優秀手寫簽名圖像特徵學習能力和更快的訓練收斂速度.

**關鍵詞**: 離線手寫簽名驗證; 多尺度融合; 端到端; OSVTF.

**1. Introduction（引言）**

**1.1. Research Background（研究背景）**

Biometrics technology是一种根据指纹、面部、虹膜等个体生理特征或声纹、手写签名等行为特征进行识别或验证的技术。该技术被广泛应用于安全认证、金融交易、门禁系统等安全领域**[12]**的个人身份识别或验证。在识别场景下，系统将根据用户提供的生理或行为特征，识别系统数据库中已有的用户资料。这种场景适用于指纹、虹膜识别个人身份等。其次在验证场景下，用户需要向系统提供验证的身份和特征信息，系统将根据已经存储的特征信息和用户当前提供的信息以此判断当前用户是否为声明的用户，适用于智能手机解锁、国际关口等需要声明个人身份和提供个人特征信息的场景。

手写签名是日常生活中较为重要的个体行为特征，其作为在法律、金融、行政等领域验证个人身份的主要特征，因为它在收集过程中是无法侵入的，诸多验证个人身份的技术中也将手写签名视为主要特征之一。手写签名会根据个人的书写习惯产生不同的文字风格，如楷书和草书。甚至随着时间推移，个人书写风格可能会发生改变，相隔一段时间后用户提供的签名（定义为Query）可能和之前输入系统的参考签名（定义为Reference）存在一定的差异。由此在用户手写签名验证的对比工作上会出现一定困难，比如个人签名上每个字的拐弯笔画角度、直线长度和尾部勾起都会存在差异。上述也提到个人书写习惯不同最终得到的手写签名可能略有不同，因此会有一部分人为了能够得到某个用户的身份权限，会恶意伪造手写签名以通过安全隐私检查系统，甚至通过长时间练习以达到和某用户相似风格的手写签名。由此学术界展开了对手写签名验证的一系列研究，希望能够设计出相关模型以帮助相关工作人员更高效地完成流程重复较多的手写签名验证工作，从而让更多工作人员参与更核心的工作。

在学术界中，手写签名验证根据数据收集途径不同分为离线和在线两种类型。离线手写签名的收集过程是通过用户在纸张上进行书写过程后得到；在线手写签名是采用数字化台方式收集用户的手写签名，收集的手写签名图像可能会受到设备的影响，如笔的位置、倾斜度、压力等**[2]**。除了数据收集途径不同分为不同类型的手写签名验证外，学者们定义了Writer-Independent (WI)和Writer-Dependent (WD)两种类型的分支工作以评价模型算法。WI任务指的是模型结构中输入一对手写签名图像，根据reference签名和query签名提取的特征从而判断query签名是否为伪造的。WD任务则是在WI任务输入的基础上增加作者id，根据作者id使用对应的作者分类器以判断query签名是否为伪造的。这两种分支任务虽然在最后输出的都是query签名的标签类别（genuine或forged），但是在不同程度上能够反映模型不同的性能，WI任务注重学习签名群体的全局特征，不需要针对每个用户单独学习一个作者分类器，具有新用户无需重新训练的优良拓展性，同时能够体现模型的泛化和判别能力强弱；WD任务则是额外为每个用户单独训练一个作者分类器，能够考虑到单独作者不同风格签名依然认为是真实签名的高精度性，但是需要每个用户收集足够多的样本，训练成本高，泛化能力较弱。由此WI和WD任务能够在不同程度上体现模型的整体性能，根据两种任务的评价指标来综合评价模型性能。

随着科技水平的发展，人工智能领域已经能够根据用户提供的样本图和关键词生成以假乱真的图片，正常人肉眼难以鉴别真假。因此会存在部分人为生成伪造签名地行为，从而达到以假乱真来通过安全系统的验证，造成个人隐私权、财产被侵犯、窃取等危险后果。单靠人工力量进行验证会存在一定的社交风险和判断错误，因此研究离线手写签名验证，一定程度上能够降低伪造签名通过验证的风险，通过相关算法模型的帮助，尽可能减少人工验证的错误，从而更好地保证个人隐私权和财产安全。

**1.2. Research Motivation and Importance（研究动机和重要性）**

本工作主要研究的是离线手写签名验证任务，其本质上是一种图像分类任务的衍生分支，与传统图像分类任务不同的是，离线手写签名验证任务要求输入的是reference和query一对的签名图像样本，WD任务需要额外输入一个作者id以训练作者独立的分类器。离线手写签名验证模型将根据输入的reference和query签名图像提取相关的图像特征，根据特征进行对比从而输出genuine或forged的类别标签。由此该任务具有一定的挑战性和创新的需求，设计模型或算法架构时不仅需要考虑模型提取一对图像特征的关键性，同向双流的模型架构将直接影响reference签名和query签名图像提取的特征关键性，一旦提取的特征不具备那么重要的信息，那么将导致模型的判断能力急剧下降。同时由于WI和WD任务的定义以及现实生活中的运用场景，要求了其必须具备较高的准确率和泛化能力才能防止个人隐私财产侵犯，更是提升了对模型性能的需求，由此对于离线手写签名验证模型的设计与实验是一项极具挑战的图像任务。

在离线手写签名验证模型的设计上，分为特征提取器和WD任务的作者分类器两部分。在过去学者们的研究中，大多数根据样本分布等数据集群特征以人工设计图像特征提取方法，采取传统机器学习的方法以判断签名是否为伪造的**[12]**，如采取特征之间的距离或SVM以判断签名是否伪造的。但是这种人工设计特征的方式具有特定数据集群的缺陷，它必须要求数据集的作者签名风格统一，细微的差异将作为离群值导致传统机器学习方法的判断准确率下降，但是实际生产工作中是需要拥有这些一定差异包容度的，因此这种采取人工设计提取特征的方法逐渐淘汰。学者们希望能够拥有一种更关注某个部位并且不用多次人工干涉的图像特征提取器，随着近六年深度学习的快速发展，具有共享参数的局部视野运算的卷积神经网络 (CNN)在诸如MNIST、ImageNet这类传统图像分类任务上取得了不俗的成绩，与传统机器学习方法对比CNN的准确率和泛化能力更是得到了进一步验证。由此手写签名验证领域的学者们引入了CNN作为特征提取器，并且实验证明这种方式的准确率和泛化能力相较之前的方法有了很大的突破**[11]**。

即使CNN提取图像特征的能力出众，其核心思想的卷积核运算具有一定的局部视野强化学习能力，能够关注图像的局部特征；但是对于手写签名来说不仅需要关注到字体笔画的拐角程度，还需要关注签名图像整体的风格、字体等因素，可能会出现同一作者签名位置不同的情况，因此CNN作为特征提取器还具有一定的缺陷。随着自然语言处理领域的发展，具有全局特征学习的Transformer**[36]**的出现引起了学者们的注意，随即图像领域的学者们引入了Transformer的Encoder-Decoder架构对图像分类任务进行实验**[4]**，实验证明这种方式在模型参数收敛的情况下性能要优于CNN的特征提取器，但是训练达到模型参数收敛的条件较为苛刻，原因是Transformer的注意力机制需要对整体图片进行学习，CNN则是采取了共享参数的方式学习图像局部特征，从而能够更快达到模型收敛，但是在实际使用上泛化能力较差，需要不断微调数据集以达到更高的模型性能。由此衍生出CNN作为Backbone提取多通道特征图，拉平后作为特征向量进入Transformer中进全局行注意力特征的运算**[9]**，从而达到多通道特征图在缺乏整体信息的情况下，经过Transformer的注意力机制加强整体信息的特征，实验证明这种方式可以更有效地学习图像特征，模型从而达到更优秀准确率和泛化能力。综上所述，CNN+Transformer风格的模型架构成为了近几年图像领域的主流框架，离线手写签名验证任务中也开始引入这种框架并且实验证明这种风格的模型架构具有良好的模型性能**[9]**，因此本工作提出的基于Transformer的离线手写签名验证模型同样是CNN+Transformer风格的模型架构，顺应近几年深度学习中该领域的发展趋势，并且在此基础上增加了多尺度融合特征，在满足日益增长的高清图片分辨率要求的同时，能够在多个尺度上学习签名图像字体笔画拐角等特征，弥补了经过CNN后多通道特征图在尺度方面上的不足。在最终输出分类阶段，与传统的图像分类任务不同的是，本工作提出的模型在最后输出的特征中将收集之前各个模块的特征，进行整体拼接后进行类别标签预测或模型训练，这种方式已经被证实是有效的**[38]**。研究的挑战部分在于如何有效地运用多尺度融合特征进行类别预测和模型训练，这种运用整体模型特征的训练、预测方式能够在整体上更好地训练各个部份的模型权重，模型训练到收敛的阶段将会超越以往大部份深度学习模型架构，对于相关领域而言更是提供了一种新颖的思路，能够融合卷积神经网络和注意力机制的特征方式，为后续更简洁的融合模块提供了一定的设计思路；相反这种方式将增加的设备的运算推理压力和模型训练成本，需要经过多种实验以证明这种方式是有效的。

**1.3. Research Objectives（研究目标）**

本工作中提出的Offline Signature Verification TransFormer (OSVTF)架构，整体上是基于孪生网络的思路设计的模型，如图1所示。图示

AI 生成的内容可能不正确。

图1：OSVTF模型架构流向图

孪生网络的核心思想是两个共享权重的子网络，分别接受两个输入，输出两个特征向量，随后使用距离函数计算相似度**[10]**。在离线手写签名验证的WI和WD任务中，输入reference和query签名的一对图像，因此整体架构方面参考了TransOSV **[41]**，在此基础上增加了backbone和FPN Fusion模块，对Encoder、Conv-Module和Decoder部分进行了一系列的优化和调整，更有利于各个部份特征的变化与学习，实验部分则是需要证实这些模块能够提升模型对于伪造签名的判断能力和不同场景的泛化能力。

在整体的研究计划和实验上，依旧采取WI和WD两种大方向的实验方案以整体评价模型性能和质量，在此基础上将针对CNN的多通道特征图与Transformer的注意力机制特征进行一定的可视化，需要关注这些模块在图像特征学习上是否能够更好地把握住图像的部份重要特征信息。此外，多尺度特征的融合方式有很多种，如Mask R-CNN**[13]**采取的特征金字塔网络风格（FPN-Style）**[2]**，这种融合方式包括但不限于均值累加、直接相加、拼接，因此后续会对这几种融合方式进行控制变量的模型训练以获得最佳性能的融合多尺度特征图方式。此外以往孪生网络的方式是根据距离函数计算相似度以完成相关任务，这种方式能否在增加了多尺度融合特征后依然有效需要进行一定的验证，本工作在此基础上将参考传统图像分类任务的CNN处理分类方式，增加全局平均池化层的方式作为分类器对特征进行分类标签预测，以此判断是否会优于以往的距离相似度判断。

综上所述，本工作将分为三部分研究阶段：1.初步对OSVTF模型架构进行模型周期训练，验证多尺度特征和模型的优化调整方案能否在原有架构上有所提升；2.针对FPN Fusion模块的多尺度融合方式，将在第一阶段的基础上采取控制变量法对模型进行小幅度的微调训练，筛选出最佳的多尺度特征融合方法；3.针对WD任务，在最后的分类器阶段增加全局平均池化的分类器，和以往离线手写签名验证的采取距离相似度预测的方法进行对比，从而判断两种分类器的优势和劣势，并且在WD任务中采取更优的分类器进行实验以验证是否奏效。

**2. Literature Review（文献综述）**

在Biometrics technology中，手写签名是较为重要的个人身份验证特征，其验证算法或模型的核心挑战在于签名的高类内差异（同一人签名的自然变化）与低类间差异（伪造签名与真实签名的相似性），将基于这两种方向上对模型和算法进行评估。而手写签名验证任务根据图像数据的处理方式不同，分为离线和在线验证两种。本工作中着力研究静态图像的离线手写签名验证任务，与在线验证不同的是在数字化设备上动态接受用户的签名，即笔迹数据是实时变化的，将收集从书写开始到结束过程的全部信息；离线验证则是完全通过静态的图片数据以进行用户验证，这种方式不对数字化设备设置任何限制，只需要通过表格让作者进行书写即可，因此在数据收集过程将大大减少工作量。

离线手写签名收集的图像数据，进行模型或算法验证时需要经过一定的预处理，而良好的数据预处理技术一定程度上会提高系统验证准确率。为了提高签名图像质量，学者们采取了RGB图像转换到GRAY单通道图像、平滑像素去除噪点的图像处理**[40]**。这种处理方式能够很好的减少了白色纸张上收集的手写签名图像不必要的特征，如RGB三通道图像中某些通道像素点的值为255，而具备手写笔迹的像素点的值往往很小。但是这种处理方法存在具有笔迹特征的部分的像素点的值非常小，而其他空白部分的像素点的值非常大的缺陷，因此后续学者们对这些图像进行黑白反转的处理，经过验证这种处理方式能够得到更优秀的验证准确率**[11]。**

在离线签名验证的早期阶段的1990s-2000s时期，学者们基于几何特征以分辨签名图像样本对的真伪性**[33]**，通过人工提取图像的宽高比、笔画长度等全局特征和拐点、曲率等局部特征进行匹配识别。这种方法奠定了初期离线签名验证的特征提取模板，人工提取特征后进行相似度计算。在1996年，Lee等人提出模板匹配的方式以进行离线签名验证，其通过动态时间规整（DTW）对齐签名轨迹，从而判断签名是否为伪造的**[20]**。随着统计学理论的发展，在2000s-2010s时期，出现了两种工作的流派：特征工程主导的一系列特征设计和分类器设计。在2011年，Vargas等人首次提及结合纹理工程（LBP、HOG）与结构特征（梯度方向分布），将特征工程引进离线签名验证的特征提取工作中。这些特征工程将导致签名图像特征数量急剧上升，因此需要学者们开始采取机器学习的方法以作为分类器，如 Edson等人于2005年采用隐马尔可夫模型提出离线签名验证系统**[17]**，HMM模型着重建模了签名时许的变化，即将图像每个像素当成词向量，从而更好地把握笔画之间的关系；Malik等人于2013年采用SVM作为分类器以完成更高维度特征数的手写签名验证，SVM是学术界中最常用的一种二分类任务的分类器，还能处理高特征维度，因此后续在分类器阶段大多数都是采取SVM的方式。即使传统的机器学习方法改进了早期签名验证系统的性能，但是仍然存在极大程度依赖人工提取特征的缺陷，人工设计繁多的特征提取步骤将导致各个特征部分差异过大，无法保证合并成特征向量后能够具有一定的统计性，且手工设计特征难以捕捉复杂模式，如笔压、局部纹理。而且这种方法训练得到的模型泛化能力不足，对跨数据集和跨语言场景适应性差，并且计算时间成本极高**[28]**。

随着计算机硬件的发展，离线手写签名逐渐向深度学习过渡。深度学习早期是以卷积神经网络为主的发展过程，在2012年Krizhevsky A.等人提出AlexNet**[19]**在ImageNet比赛的图像分类任务上遥遥领先使用传统机器学习算法的第二名参赛者夺得冠军，使得卷积神经网络为代表的深度学习逐渐成为图像任务的主流方法。后续学术界推出更深层次的卷积神经网络VGG**[34]**，该架构是在AlexNet的基础上增加了更多的卷积层，提取更多通道的特征图。但是这种叠加卷积层的计算方式使特征图尺寸变得更小的同时导致每个阶段特征的丢失，为了解决这种问题，在2016年K. He等人提出了残差连接的ResNet**[14]**。在每个尺寸特征图的卷积层之间，增加了一个残差连接，将原始输入模型层的特征图累加到输出的特征图。这种方式在监督学习的卷积神经网络中，经过权重共享的卷积核运算，能够一定程度上弥补原始图像特征信息，并且这种残差连接的方式成为后续深度模型保留局部特征的主流方式。卷积神经网络在图像分类任务上的优质表现，使得离线手写签名验证开始借助卷积神经网络以代替早期人工设计特征的步骤。在2017年，L. G. Hafemann等人首次使用卷积神经网络提取签名图像特征进行验证**[11]**，采取卷积部分运算输出特征图作为签名图像特征，输出阶段设计了两个全连接层共用签名图像特征以输出作者ID和签名类别。这种处理方式同时结合了WI和WD任务，同时满足两者输出并且证明这种方式显著提高了离线签名验证的准确性。这种方式及其依赖模型训练技巧，训练过程及其困难达到参数收敛结果，但是也为后续手写签名验证任务提供了使用深度学习方法的思路。在2017年，Y. Hafemann等人提出使用预训练的CNN（如VGG）提取签名图像的全局和局部特征，并将特征输入SVM进行分类，能够显著提升跨数据集的泛化能力**[11]**，并且进行了端到端的优化。类似的M. Diaz等人于2019年通过迁移学习训练CNN，结合SVM作为联合分类器，在GPDS数据集**[35]**上实现低错误率。这种CNN+SVM的方法成功地将签名验证任务模型的误判率下降到了传统机器学习无法比拟的程度，后续都是这种深度学习模型作为特征提取器和SVM作为联合分类器的方式以进行签名验证。并且根据数据集图像尺寸，S. Dey等人于2020年提出了多尺度CNN特征与SVM融合的SigNet**[32]**，采用CNN中多个卷积层输出的不同尺度特征图以作为签名图像特征，并与SVM结合优化分类边界，实验证明多尺度特征将很大程度上提升了整体模型的鲁棒性，在跨数据集上拥有了非常优秀的模型性能。但是CNN在签名图像特征的学习能力上存在难以捕捉签名图像的全局上下文信息，如笔画间的长距离依赖。

随着深度学习的自然语言处理领域发展，A. Vaswani等人于2017年提出基于注意力机制的Encoder-Decoder架构的Transformer**[3]**，注意力机制能对词向量具有全局特征感知的能力，在可变句子长度的机器翻译任务中能够更好地捕捉上下文信息，这种在机器翻译任务中体现了其优秀的全局泛化能力，在跨语言多个不同语种的翻译任务中都体现了该架构优秀的全局感知上下文能力。介于Transformer的全局特征感知能力，A. Dosovitskiy等人于2021年首次将Transformer的注意力机制用于图像分类任务中，提出不使用卷积运算的Vision Transformer (ViT)**[4]**，将输入图像进行分块平坦化和映射作为词向量以输入Transformer Encoder中进行注意力特征计算，最后使用Encoder输出的特征向量进行MLP以生成图像类别概率。这种方式将图像素映射后作为token的图像特征处理方法，在Transformer架构上得到了领先CNN的识别准确率，并且在验证、测试集上体现了Transformer强大的全局泛化能力，因此图像领域各类任务逐渐开始采取注意力机制对图像特征进行二次处理。在2022年，X. Wang等人首次将ViT进行离线签名验证，解决了上述CNN难以捕捉签名图像全局上下文信息，利用自注意力机制建模签名全局结构，提升对复杂伪造签名的区分能力**[39]**，在GPDS-960和CEDAR**[3]**数据集上相比同期模型的Equal Error Rate下降了12%。在2023年Y. He等人同样基于ViT提出TransOSV**[41]**，该架构提供了一个端到端的基于Transformer的离线签名验证系统，同时结合了全局语义信息和局部差异信息，更有效利用Transformer进行局部与全局特征建模。其引入了多层特征组合机制，将来自不同模块的特征拼接为最终的签名表示，实现对伪造签名的多角度检测；同时基于对比任务的Focal Loss**[10]**和处理样本不均衡Double-Margin Loss**[25]**设计了Focal Contrast Loss，解决上述问题的同时加强对离线签名任务公开数据集中hard samples的学习，还融合了不同模块特征的对比学习。并且不同于以往离线签名验证的孪生网络，提出了part decoder，通过对比reference和query的token级attention，定位并聚焦于伪造签名的敏感部位，将注意力机制的权重引入进decoder部分特征中，增强模型对伪造签名的部分位置感知。但是TransOSV仍存在缺陷，它采取了多层特征组合机制，从不同角度检测签名图像对特征，但是在签名图像尺度不同的情况下并不具备强大的泛化能力，且decoder部分对于token级的特征不具备良好的抽象表达能力。

综上所述，本工作将在TransOSV架构基础上，新增backbone提取多尺度特征图，采取FPN Fusion的方式对多尺度特征进行融合，以增强模型架构对多尺度特征的学习。同时为了后续模块更好地抽象表达多尺度融合特征，对Conv-Module进行了下采样的结构设计，在part decoder增加输入映射以及交叉注意力的多头调整，增强模型对图像特征的抽象表达能力，能够在经过Encoder全局上下文注意力特征学习后，加强关注伪造签名的局部敏感区域。在WD任务上，由于采取了大量的卷积输出特征图和多层特征组合机制，因此将按照传统图像分类任务提供一个全局平均池化层**[22]**以进行签名类别预测。

**3. Methods（论文主题算法等）**

**3.1. Overview**

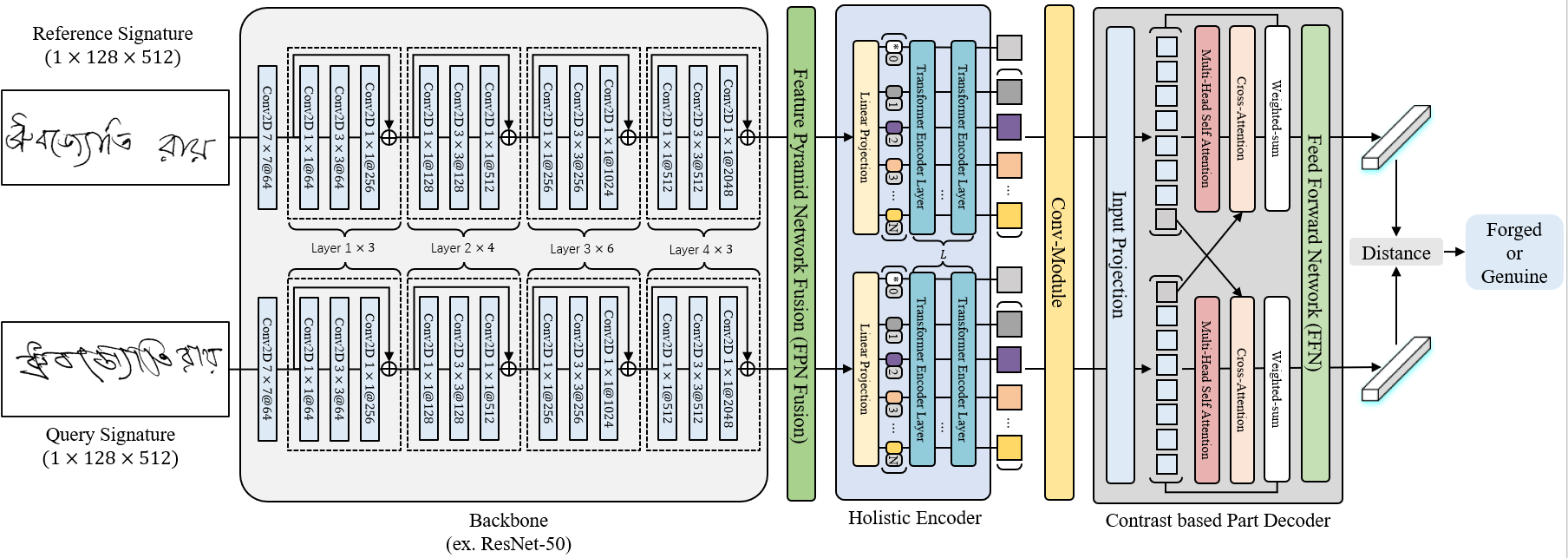
本工作将定义手写签名的标签类别forged为1和genuine为0，输入的reference签名图像定义为，query签名图像定义为。

图2: OSVTF模型架构示意图

提出的多尺度特征融合的Offline Signature Verification TransFormer (OSVTF) 模型架构如图2所示，由Backbone 、FPN Fusion 、Holistic Encoder 、Conv-Module 、Contrast based Part Decoder五个部分组成。在模型中我们将输入一对签名图像样本和，初始输入图像经过共享权重的Backbone（图1中是以ResNet-50**[14]**为例子）得到多尺度特征图集合，经过FPN Fusion模块将不同尺度特征图进行融合，输出多尺度融合特征图。随即进入共享权重的Holistic Encoder得到holistic平坦特征和patch embeddings 和。为了进一步整合特征图，将上述的patch embeddings调整形状至二维形状，经过共享权重的Convolutional Module (Conv-Module)得到它的输出，它们将直接进入Contrast based Part Decoder以获得交叉注意力的对比part平坦特征。最后预测阶段，对和进行Global Average Pooling(GAP)运算得到多尺度融合平坦特征和卷积平坦特征。将全部平坦特征进行拼接得到OSVTF的总特征向量和，将基于和进行判断是否为伪造的。接下来将深入解析Backbone、FPN Fusion、Holistic Encoder、Conv-Module和Contrast based Part Decoder结构。

**3.2. Backbone**

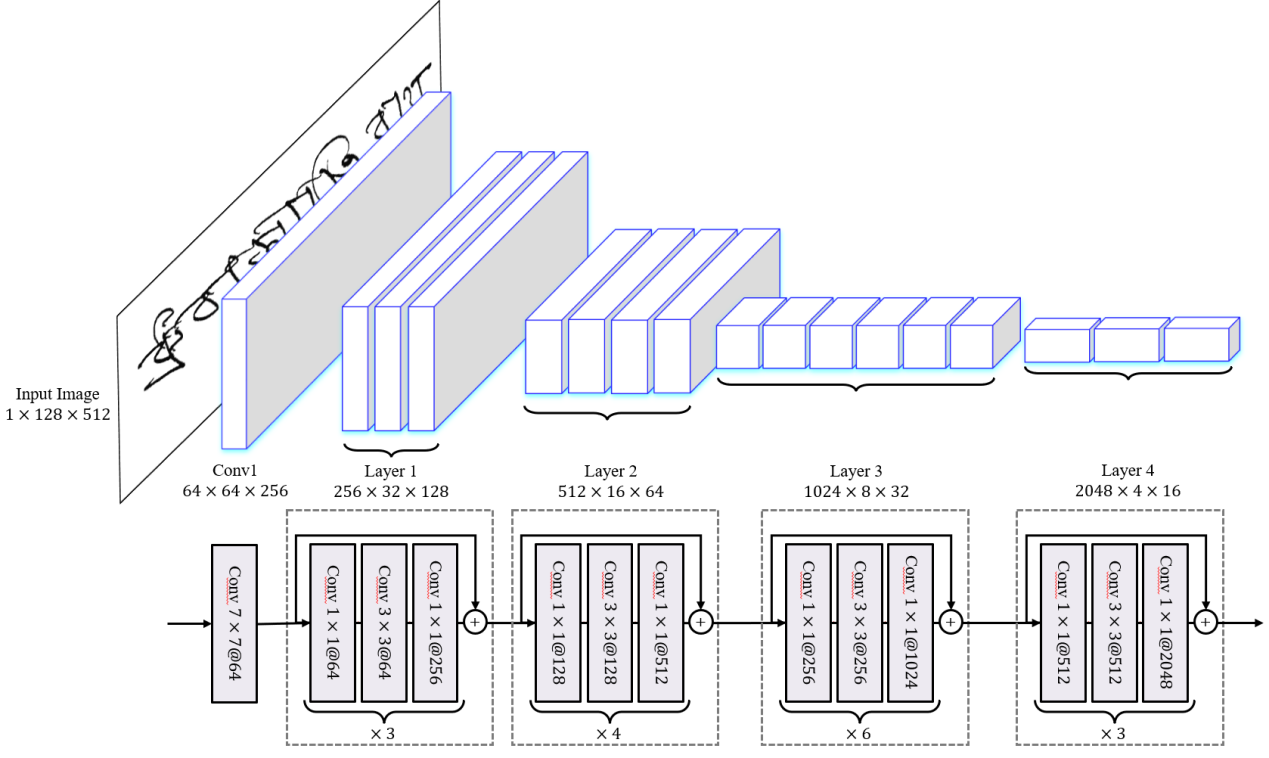


图3: Backbone (ex. ResNet-50) 架构示意图

在Backbone部分将采取舍弃输出部分中Average Global Pooling和全连接层的CNN的架构。CNN架构中大多数由输入部分的卷积层和4个定义的卷积层组成，以ResNet-50**[14]**为例如图3所示，每一层都是由若干个bottleneck块组成，单个bottleneck块由两个卷积核为的卷积2D层组成，这样安排主要有以下目的：1. 降低特征维度以达到提升计算效率的目的；2. 中间的卷积在较低维空间中提取如笔画、边缘等局部空间特征具有保持感受野的同时减少冗余，这样先减少表达通道数后再扩充至输入的表达通道能够在bottleneck中找到更好的特征学习参数，具有较好的特征抽象表达能力；3. 在实际运算当中能够搭配identity skip connection能够快速传播梯度，避免梯度消失，同时这样的设计使得中间参数量和计算负担大大降低，能够堆叠更多的层数如ResNet-101，促进更深网络的训练；4. 这种模块化设计在卷积参数上具有容易微调的优势，中间通道数可以根据环境进行调整可以控制性能-效率平衡。除此之外，ResNet系列网络引入了残差连接，解决了图像经过深度卷积运算后部分特征缺失的问题，在每个bottleneck块中引入残差连接，将输入部分的特征图与bottleneck的输出进行累加，能够一定程度上减少特征图在经过一定数量卷积层运算后特征丢失的问题。

假设卷积层输入特征图，输出特征图，卷积2D层的卷积核大小（卷积核权重），填充，步幅，偏置项。对于特征图的第个输出通道、空间位置为的像素点计算公式如下：

此外，每个bottleneck块中，除了最后一层卷积2D层后只接BatchNorm (BN)**[16]**运算，每个卷积2D层后会紧接着BN和ReLU激活函数**[29]**运算，经过BN和ReLU激活函数的空间位置为的像素点值如下：

其中，是卷积2D层输出的批次多通道特征图中第通道平面的特征统计均值和方差，是可学习标度和平移参数。由此可见BN的运算思想是对某个批次数据进行归一化，具有提升模型参数训练时收敛速度的效果。在每个bottleneck块运算中，对经过三次卷积2D运算的特征图进行一个累加残差操作，即输入bottleneck块的特征图会累加到三次卷积2D运算输出的特征图上，以此传递到下一个bottleneck块。同时，backbone每个层运算后会经过一个下采样运算，以此在多通道特征图上减少冗余特征，将采取Max Pooling 2D的方式以等价下采样操作**[14]**。上述的残差链接方式，能够在保证特征图经过卷积运算后保留一定的原始图像特征，因此ResNet系列网络最适合作为深度模型的Backbone以提取图像多通道特征图。

采取backbone主要是为了提取图像多尺度的特征图，因此在最终输出上包含了除输入卷积层外的每个层的不同尺度特征图，如下：

其中，每个尺度的多通道特征图，表示layer 1输出特征图的宽度、高度和通道数，当backbone为ResNet-50时，），每一层特征图的通道和尺寸之间都是两倍的关系。

**3.3 FPN Fusion**

为了能够有效利用backbone提取的多尺度特征图，采用Feature Pyramid Network Fusion (FPN Fusion)模块以融合不同尺度的特征图，其采用的是FPN-Style**[23]**的模型架构，如图4所示。

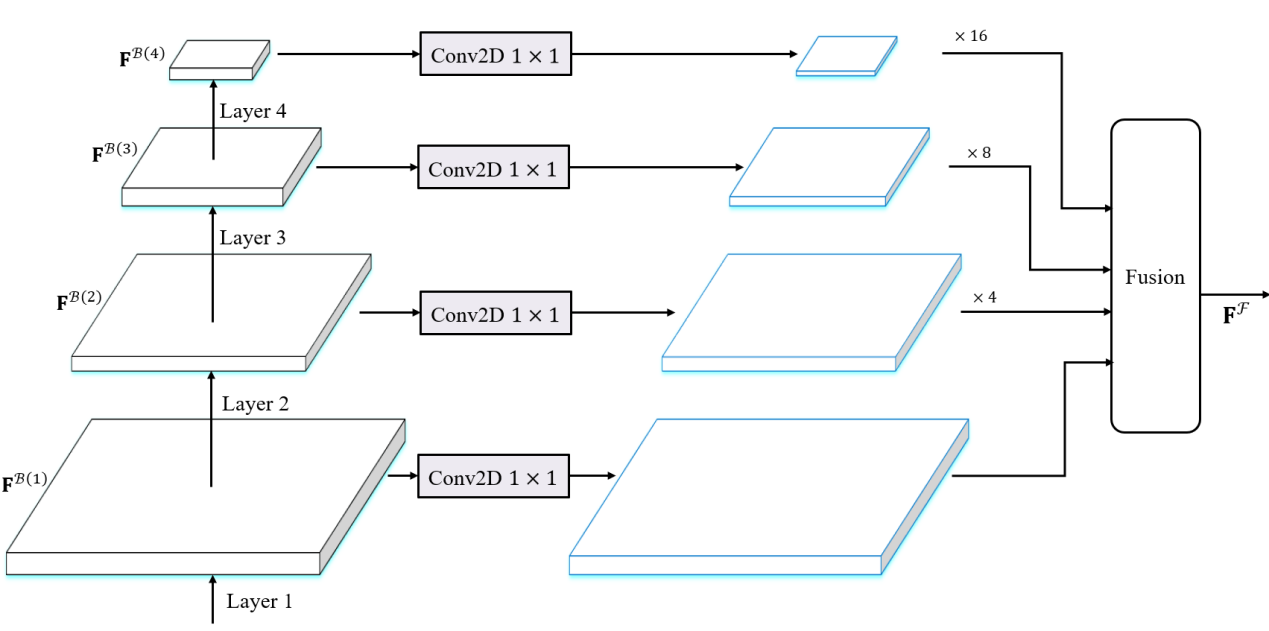


图4: FPN Fusion示意图

其中bottom-up过程为backbone推理获得不同尺度特征图的过程。在bottom-down中，对于每一层特征图将采取卷积核为的卷积2D层进行映射，将每一个尺度的不同通道数特征图映射到相同特征维度上，一定程度上达到降维的操作，从而加速模型推理速度。同时对进行上采样，增加尺寸至与相同大小，后续根据融合方式以得到多尺度融合特征图。在上采样过程中采取双线性插值**[18]**，对于第层特征图上采样输出，空间位置的值计算公式如下：

其中，表示宽度和高度方向上的小数部分，表示放大倍数，表示数值的整数部分。这种上采样方式能够很好地保留原始特征地结构，同时引入平滑过渡如图5所示。

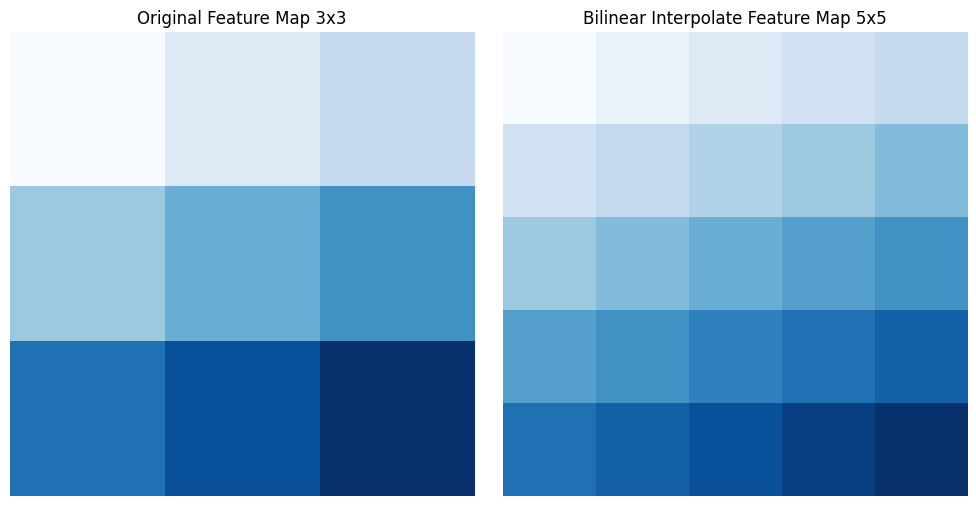


图5: 双线性插值示例

每个上采样像素点都会根据原始特征图像素点进行加权计算，不仅保留了原始特征图的结构比例，同时使得像素点之间过渡较为平滑，一定程度上能够反映不同尺度对于图像特征点的影响，比如某个笔画的特征将被放大。

在每个尺度特征图经过双线性插值地上采样后，将根据三种方式进行融合：累加、加权平均、通道拼接。第一种方式是传统的FPN-Style做法，直接对降维后的多尺度特征图直接进行累加运算，类似上述提到的残差链接。第二种方式的设计思路是为了能够保证不同场景的签名图像都能够拥有一定的学习能力，通过设定权重比例能够更注重某个尺度的特征，默认是采取的全部相等的权重。第三种方式是在通道维度上进行拼接，紧接着一个卷积核的卷积2D运算，这种方法能够最大程度保留所有尺度的特征，但是可能会导致模型计算开销过大，导致模型收敛速度变慢。

融合多尺度特征图后，最后将经过一个卷积核的卷积2D层以平滑上采样过程中带来的checkerboard artifacts，最终输出多尺度融合特征图。

**3.4. Holistic Encoder**

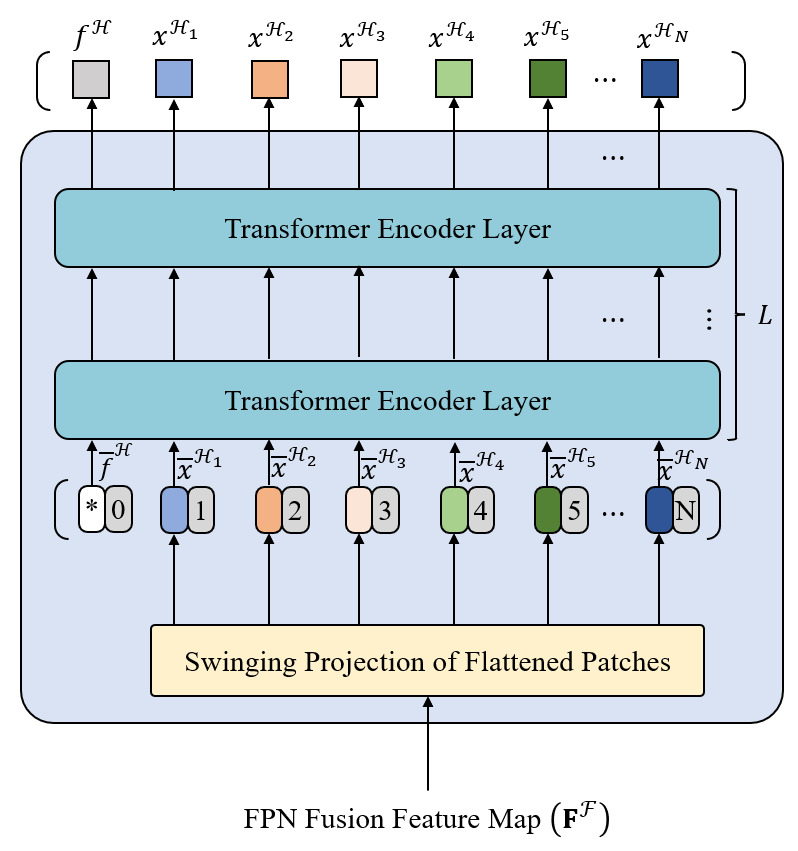


图6: Holistic Encoder结构

Holistic Encoder是参考了Vision Transformer (ViT) **[4]**设计的架构如图6。在输入阶段采取了swing embeddings**[24]**的方法代替之前的分块操作，根据设定的窗口大小和步幅，在输入的多尺度融合特征图上进行分块和映射，得到patch embeddings后进行平坦化以作为特征向量，其中表示映射特征维度数，向量长度计算如下：

在swing embeddings中拼接一个定义为class token的可学习权重，即输入Transformer Encoder Layer的特征向量为。由于对特征图进行了平坦化运算，因此需要在特征向量上累加positional embeddings**[36]**，以此使特征向量包含其在多尺度融合特征图上的位置信息。

经过swing embeddings后，特征向量进入Transformer Encoder Layer以进行注意力特征计算。其中Transformer Encoder Layer由个堆叠形成，如图4所示。

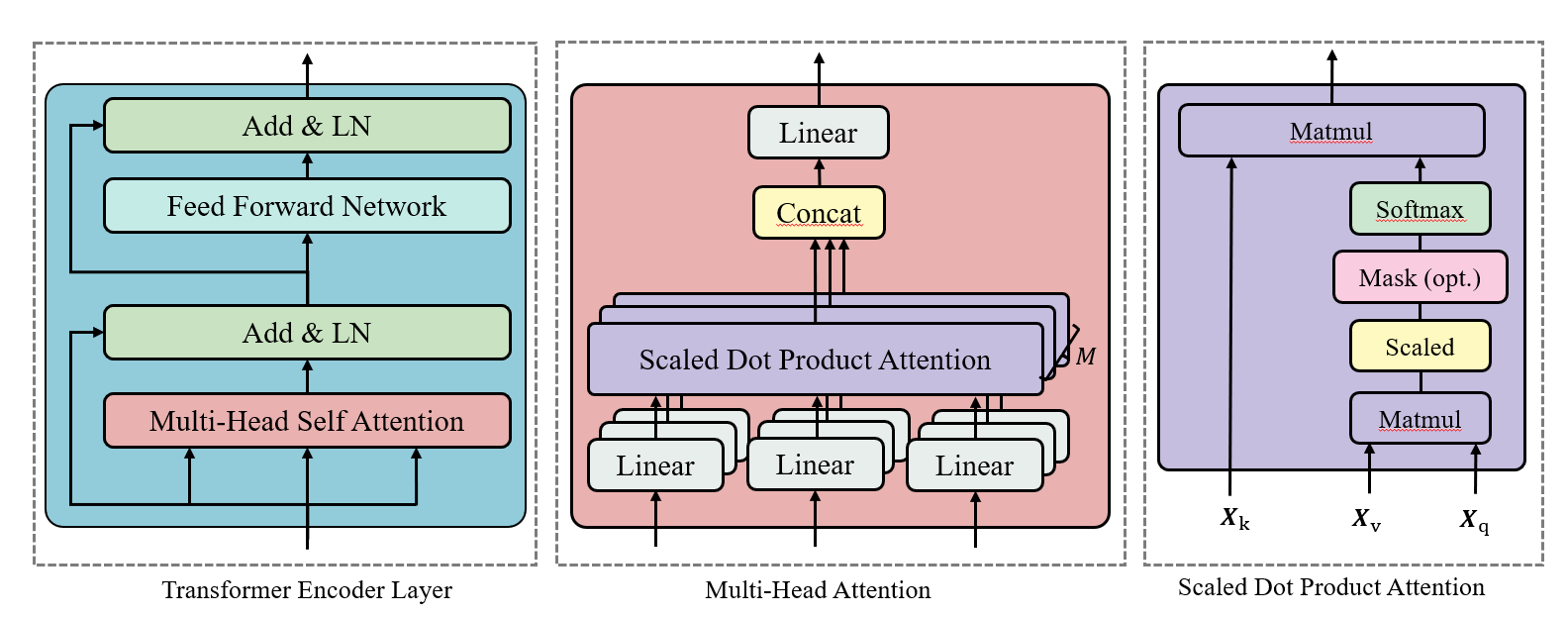


图7: Transformer Encoder Layer结构

每个Transformer Encoder Layer由一个Multi-Head Self AttentionAdd & LN和Feed Forward NetworkAdd & LN组成，其中Add&LN 表示残差链接和LayerNormalization (LN) **[1]**。输入的特征向量首先进入Multi-Head Self Attention (MHSA)以进行注意力机制计算，MHSA本质上是Multi-Head Attention (MHA)，只是在运算中和映射前的来源相同，即输入特征向量将同时进入每个头的线性映射层进行降维，从而生成查询向量、键向量、值向量，并且。其中注意力机制分为多个头计算是为了节省计算资源，将特征向量输入个头的三个线性层以降维映射，从而达到并行地注意力计算，最后进行拼接和线性层映射会输入向量相同特征维度，因此MHSA的输出向量形状与输入向量形状相同。在MHSA中每个头的注意力特征向量均采取scaled dot product attention**[36]**计算，公式如下：

其中，、、之间是软查询关系，每个查询不是选择一个具体的键-值对，而是对所有的key-value进行加权平均，权重是通过和的相似度计算而来。的每个特征token会对的每个特征token建立一个相关性分布即Softmax**[37]**的注意力权重，根据这个权重来加权融合的值。该机制的优势在于能同时关注多个相关token，使得模型能更充分建模上下文信息，比如位置的注意力权重，表示在第位置上对第个位置key的权重，这个权重的值越大，说明第位置的query对第位置的key具有很强的上下文关系，反之越小则说明关系很小。根据这个注意力权重，再次乘会将自身向量注意力权重高的地方特征将进一步放大，并且token之间将包含注意力信息，从而达到每个token拥有自身全局的信息，能够更加有效地利用重要特征，忽视掉无用特征，在手写签名图像中将更有效地关注笔画部分，忽视空白部分。由此单个Transformer Encoder Layer的MHSA输出如下。

其中，表示 MHSA输入阶段的线性映射层权重，表示MHSA中最后线性映射层权重，表示每个头、、的特征维度数，因此必须满足。

在MHSA部分进行了维数较大的矩阵乘法，为了减少梯度爆炸或消失风险，同时为了减少原始特征向量信息缺失，因此在MHSA后紧接着一个残差链接和LN层。LN层的主要作用是归一化每个位置的特征向量，从而加快模型收敛的同时提高模型稳定性。与BN的均值和方差不同，LN的均值和方差来源于当前特征维度上，计算如下：

其中，表示MHSA输出特征向量的第个特征维度。后续的归一化运算与BN层一样，整个MHSAAdd & LN输出的计算公式如下：

其中，表示上一个Transformer Encoder Layer的输出，若，则。在经过注意力计算后，引入Feed Forward Network (FFN)增强整体模型对复杂模式的建模能力**[8]**，整个FFN计算如下：

其中，表示FFN中第一个线性层的权重和偏置，表示FFN中第二个先行层的权重和偏置。GeLU**[15]**函数和ReLU函数作用相似，均是进行非线性转换的激活函数。整个FFN就是对模型进行一个非线性转换，与Attention的全局交互机制形成互补，联合构建强大的上下文感知表示。随后紧接着Add & LN补充特征信息和加快模型收敛，后续每一层的Transformer Encoder Layer均为堆叠效果，加深特征向量的注意力权重，整个Holistic Encoder的输出为。

**3.5. Conv-Module**

图像特征经过Holistic Encoder捕捉序列中元素的相对位置信息进行特征学习，但其存在绝对位置信息的学习局限性。签名验证任务需要精确到笔画位置信息（如笔顺、局部结构差异），因此在这之后为了加强部分小部分图像特征，将Holistic Encoder的patch token输出进行rearrange，对平坦化的token重新转换成2D形状以进行卷积运算，其中。本模块引入卷积操作对Transformer编码器输出的patch特征进行再处理，加强位置感知能力。在Conv-Module架构上，参考了TransOSV的Conv设计**[41]**，根据以往CNN的样式，调整了中间卷积2D层的输出通道数增加了下采样运算，如表1。

表1：Conv-Module结构信息

|  |  |  |
| --- | --- | --- |
| 层 | Kernel Size | 输出图像形状 |
| Conv 2D+ReLU |  |  |
| Conv 2D+ReLU |  |  |
| Max Pooling 2D |  |  |
| Conv 2D+ReLU |  |  |
| Conv 2D+ReLU |  |  |
| Max Pooling 2D |  |  |

与传统CNN的通道数逐渐变多不同，这里采取的是整体下采样运算思维，对patch特征进行一定的降维运算，在整合局部细节特征的基础上抽象表达特征图的局部特征，目的是为了提升模型的泛化能力和鲁棒性。在backbone和FPN Fusion的基础上增加该操作，进一步强化模型对图像局部特征的感知能力，强化对签名关键局部区域的关注，同时过滤背景干扰。在Conv-Module模块中，每个Conv2D层运算与backbone中的Conv2D层运算相同，均采取Conv2D+ReLU的组合，并且经过两个Conv2D+ReLU后会进行一个下采样运算，即Max Pooling 2D运算，对特征图进行整合和剔除非重要部分特征，加强特征表达性。

在模型训练过程，Conv-Module的输出在进入Contrast based Part Decoder的同时，将经过一个Global Average Pooling (GAP)运算，以此对Conv-Module输出的特征图进行全局取样，从而得到一个平坦的局部卷积特征，以完成训练和推理过程中对应的特征提取工作。

**3.6. Contrast based Part Decoder**

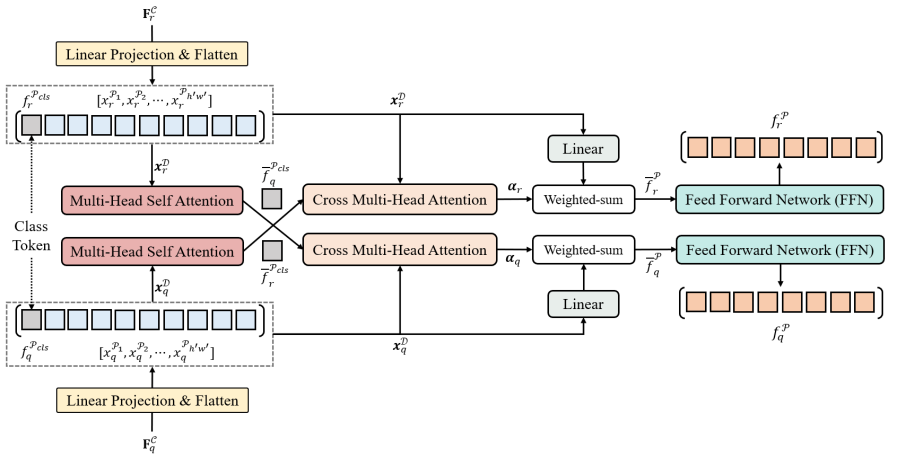
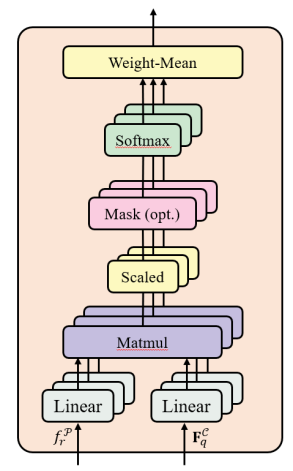
 

图8：Contrast based Part Decoder & Cross Multi-Head Attention

该部分是在TransOSV的Contrast based Part Decoder**[41]**基础上，对Cross-Attention进行了并行计算的改进，引用了Holistic Encoder中MHSA的多头注意力计算机制。与Holistic Encoder部分相似的是，在输入阶段会对reference和query的特征图映射到特征维度上，随后进行平坦化处理和增加一个可学习权重，同样定义为class token，即输入部分得到的是一对平坦化特征图向量。随后对进行MHSA运算，取出各自输出特征向量的class token 作为query矩阵交叉输入Cross Multi-Head Attention（CMHA）。CMHA的结构如图8-b所示，与MHA的区别在于只生成query矩阵和key矩阵计算其交叉注意力权重，且最后的concat替换成加权平均以处理多个头的注意力权重。对于单个CMHA的输出交叉注意力权重计算公式如下：

其中，表示输入的query向量和key向量，表示CMHA中输入部分的线性映射层权重。由此计算过程如下：

对注意力权重和经过映射的平坦特征向量进行weight-sum运算，融合生成交叉注意力权重的平坦特征。其经过FFN后得到decoder的交叉注意力平坦特征，整个计算过程如下：

这种类似与Transformer Decoder**[3]**结构的Contrast based Part Decoder，有利于生成区分真实签名和伪造签名的局部特征信息，并且强调了一对数据样本之间的特征关系，加强模型对样本中一对数据之间的特征学习。在训练过程中，将输出的交叉注意力权重进行稀疏性损失计算，迫使交叉注意力权重分布集中，避免平均化以无法关注伪造签名的敏感部分。

**3.7. Loss Function**

**3.7.1. Euclidean Distance**

OSVTF是一种类似孪生网络的模型架构，这种风格强调模型是共享权重的双流向形式，整体是为了对比两者之间的差距，因此将基于输入两者特征的相似度以设计损失函数。本工作将采取Euclidean Distance**[5]**以作为一对特征样本的相似度计算，假设输入一对特征，其两者之间距离计算公式如下。

**3.7.2. Focal Contrast Loss**

在手写签名样本数据处理阶段，会将作者的手写签名进行两者之间不会重复的配对操作，例如BHSig-B**[3]**数据集中一共包含了100位作者的手写签名图像，其中每位作者拥有24张真实签名图像（）和30张伪造签名图像（，真实签名图像作为reference签名，伪造签名和真实签名一起作为query签名，由此将生成positive样本对276个和negative样本对720个。因此每位作者一共拥有996个数据样本对，其中positive样本对和negative样本对之间的比例不接近50:50，因此positive样本和negative样本数量不均衡，属于hard samples范畴**[33]**。

早期算法、模型中用于二分类的对比样本相似度损失函数是Contrast Loss**[10]**，计算公式如下。

其中，表示边界值，即forged样本应至少保持的距离。对于的样本对，则直接作为损失值，使模型参数学习上对真实样本对的特征更相似；相反时，若reference和query的特征距离小于，则使用作为损失值，使模型对于伪造签名对的特征不小于。但是这种损失无法动态强调hard samples，容易出现模型参数过拟合问题。为了减少过拟合的风险，对genuine样本对同样引入一个边界值，减少genuine对之间不必要的收缩，即Double-Margin损失**[25]**计算公式如下。

其中，两个边界值，对于genuine样本对，惩罚其距离超过的情况；对于forged样本对，促进距离大于。虽然double margin一定程度上优化了contrast loss容易过拟合的问题，但是在两对样本数据且的情况下，通过OSVTF计算两个样本对的距离，若，那么损失函数应该给予更大的损失，但是会平等地对待这两个样本对，在TransOSV中引入了动态权重的概念，命名为Focal Contrast Loss**[41]**，计算公式如下。

其中，表示Sigmoid激活函数**[26]**，用于动态生成权重；为两个边距值；表示缩放因子，用于调节hard samples响应强度的缩放系数。当样本对的标签为genuine时，如果特征距离越大于，则当前权重值越大；当标签为forged时，如果特征距离越小于，则权重越大。由此将同时拥有动态调整训练样本的关注度、重点优化“容易混淆”的签名和提升模型泛化能力的优点，将作为OSVTF模型的主要训练损失函数之一。根据各个模块提取的平坦特征，将采取四个部分以计算focal contrast loss，整体Focal Contrast Loss计算如下。

其中，为训练超参数，用于分配各个特征的比重。

**3.7.3. Sparsity Loss**

在Contrast based Part Decoder中，对于交叉注意力计算时会出现注意力权重过于均匀分布的情况，实际上应该增强模型关注最显著的局部差异区域，由此采取交叉熵的方式以生成对比掩码的熵约束，希望交叉注意力权重能够分布稀疏，从而提升Decoder的local attention精准性**[TransOSV: Offline Signature Verification with Transformers]**，计算公式如下。

由于该部分并不是作为主要模型训练，因此引入作为该部分的超参数参与OSVTF训练中。综上所述，OSVTF的总体损失函数构成如下：

**4. Experiments**

**4.1. Experimental Setup**

本工作中将基于ViT **[4]** configuration以build OSVTF。The FPN Fusion的输出特征图的特征维度数与第一层尺度特征卷积图通道数相同。The encoder包含8个transformer encoder layer，每个encoder layer中MHSA包含12个头，768个特征，尺寸为步幅为的滑动窗口。Conv-Module部分，设置其中间卷积层的通道数为512 。 Decoder部分中，MHSA和CMHA的特征维度数、头的数量与encoder的各部分保持一致，在CMHA后weighted-sum的对patch embeddings映射的特征维度数设置为512。

在损失函数上，沿用TransOSV的部分设置和。The hyper-parameters设置为。在优化器上，所有实验均采取momentum为0.9、权重衰减为1e-4的SGD Optimizer。全部实验均在2张NVIDIA 3080 GPUs上进行并且设置总的batch size为16。

**4.2. Metrics**

在评价指标上，不同于常规的二分类任务，离线手写签名验证任务需要对negative样本进行特别关注，因此使用生物识别和签名验证任务中的标准指标**[7]**来评估OSVTF的性能，即False Rejection Rate (FRR)、False Acceptance Rate (FAR)和Equal Error Rate (EER)。在常规的二分类任务中，会使用混淆矩阵**[6]**的概念以统计样本标签和预测标签的数量，其格式如表2。

表2：离线签名验证混淆矩阵概念表

|  |  |  |
| --- | --- | --- |
| 预测标签/原始标签 | (Positive) | (Negative) |
|  | True Positive (TP) | False Negative (FN) |
|  | False Positive (FP) | True Negative (TN) |

由此可得ACC、FRR、FAR的计算公式如下：

EER的计算则是使用若干个阈值和距离进行对比，若样本对的特征距离大于阈值，则其类别为forged，否则为genuine。本工作在训练过程中将根据批次数据样本对的模型特征计算距离，根据样本对的距离最小值和最大值以设置若干个阈值，从而自动计算当前批次下每个阈值下的FAR和FRR。对每个阈值的FRR和FAR进行计算，当FAR和FRR之间的差异最小时，则同时输出当前阈值的FRR、FAR，并且计算EER如下：

由此，实验阶段将基于ACC、FRR、FAR、EER四个评价指标来综合评估模型性能。

**4.3. Dataset**

*BHSig-B and BHSig-H*. 这两个数据集均由印度理工学院古瓦哈提分校（IIT Guwahati）发布**[3]**的孟加拉语和印地语手写签名数据集。其中BHSig-B包含100个作者的Bengali手写签名，BHSig-H包含160个作者的Hindi手写签名图像。每个用户手写签名了24张真实签名和30张伪造签名。

*CEDAR*. 该数据集是由Center of Excellence for Document Analysis and Recognition开发并发布的英语手写签名数据集**[30]**。其包含了55个作者的英文手写签名。每个用户手写签名了24份真实签名和24份伪造签名。

**4.4. Results Analysis**

**4.4.1 Writer-Independent Results**

在WI任务上，OSVTF采取欧氏距离法计算两者特征之间的距离判断签名是否为伪造的，以此评估OSVTF是否优于以往模型拥有离线手写签名验证性能。初始阶段在BHSig-B和BHSig-H上进行验证，和过去的算法保持一致，对BHSig-B的100个作者按照50:50和80:20的比例对100个作者划分成训练集和测试集，对BHSig-H的160个作者按照100:60的比例对160个作者划分成训练集和测试集，得到的初步实验结果如表3。

表3: BHSig-B和BHSig-H数据集的WI签名验证性能比较

|  |  |  |  |
| --- | --- | --- | --- |
| Model | FRR | FAR | EER |
| BHSig-B dataset | | | |
| 50/50 |  |  |  |
| SigNet **[32]** | 13.89 | 13.89 | 13.89 |
| CaP **[24]** | 3.96 | 3.96 | 3.96 |
| IDN **[38]** | 5.24 | 4.12 | - |
| InceptionSVGNet **[27]** | 2.22 | 3.88 | - |
| TransOSV **[41]** | 9.90 | 9.90 | 9.90 |
| TransOSV [Ours] | 12.41 | 12.41 | 12.41 |
| OSVTF [Ours] | 10.72 | 10.85 | 10.79 |
| 80/20 |  |  |  |
| DeepHsv**[21]** | 11.92 | 11.92 | 11.92 |
| TransOSV | 3.56 | 3.56 | 3.56 |
| TransOSV [Ours] | 5.11 | 5.11 | 5.11 |
| OSVTF [Ours] | 4.47 | 4.61 | 4.54 |
| BHSig-H dataset 100/60 |  |  |  |
| SigNet **[32]** | 15.36 | 15.36 | 15.36 |
| IDN **[38]** | 4.93 | 8.99 | - |
| CaP **[24]** | 5.97 | 5.97 | 5.97 |
| InceptionSVGNet **[27]** | 3.33 | 6.38 | - |
| TransOSV **[41]** | 3.24 | 3.24 | 3.24 |
| TransOSV [Ours] | 4.86 | 4.86 | 4.86 |
| OSVTF [Ours] | 3.86 | 3.94 | 3.90 |

其中，在BHSig-B数据集中，50:50的比例下OSVTF 10.79% EER，复现的TransOSV的12.41% EER，相差大约15.01%；80:20的比例下OSVTF的 4.54% EER，复现的TransOSV 5.11% EER，相差大约12.56%。在BHSig-H的数据集中，复现的TransOSV 4.86% EER，OSVTF 3.90% EER，相差大约24.62%。在WI任务上OSVTF性能要优于TransOSV性能大约17%。在上述实验环境下OSVTF与TransOSV的性能相比略显逊色，但是在相同环境下OSVTF对比TransOSV的性能具有一定的提升，证实图像的多尺度融合特征在组合特征基础上对图像特征学习能力具有一定程度的提升。在BHSig-B 50/50数据集中，OSVTF得到的结果远不及CaP，IDN，InceptionSVGNet，但是在BHSig-B 80/20和BHSig-H 100/60数据集上超越了同类型模型，一定程度上是因为这些网络是对该比例数据集的特化网络，因此可以体现OSVTF在跨数据集场景中能够很大程度适应各个数据集图像的变化。

**4.4.2 Writer-Dependent Results**

在该部分，除了对增加多尺度融合特征的模型进行实验外，增加了深度学习传统图像分类任务的分类器（GAP分类器），以验证是否相较SVM的分类器更契合于拥有多尺度融合特征的模型。因此对BHSig-B的80:20比例和BhSig-H进行验证，结果如表4。

表4: BHSig-B和BHSig-H数据集的WD签名验证性能比较

|  |  |  |  |
| --- | --- | --- | --- |
| Model | FRR | FAR | EER |
| BHSig-B dataset 80/20 | | | |
| TransOSV [Ours] | 3.78 | 3.78 | 3.78 |
| OSVTF [SVM] | 2.74 | 2.80 | 2.77 |
| OVSTF [GAP] | 2.05 | 2.21 | 2.13 |
| BHSig-H dataset 100/60 |  |  |  |
| TransOSV [Ours] | 4.23 | 4.23 | 4.23 |
| OSVTF [SVM] | 3.17 | 3.21 | 3.19 |
| OVSTF [GAP] | 2.67 | 2.69 | 2.68 |

同样使用SVM作为分类器的情况下，在BHSig-B数据集的实验结果中，OSVTF 2.77% EER相较TransOSV的3.78% EER下降了36.46%；在BHSig-H数据集的实验结果中，OSVTF 3.17% EER相较TransOSV的4.23% EER下降了33.44%。在WD任务上OSVTF性能要优于TransOSV性能大约34.95%。在不同分类器的实验结果上，BHSig-B的采取GAP分类器的OSVTF 2.13% EER要优于采取SVM分类器的OSVTF 2.77% EER；BHSig-H的采取GAP分类器的OSVTF 2.68% EER要优于采取SVM分类器的3.19% EER。在整体WD任务上，同样采取SVM分类器的OSVTF性能上要优于TransOSV大约34.95%，采取GAP分类器时能再提升19.03%。由此可以判断对于采取了多尺度融合特征的OSVTF，采取GAP的作者分类器相较SVM能够更好地处理上述特征。

**4.5. Ablation Experiment**

上述实验结果中表明，增加了多尺度融合特征的OSVTF相较TransOSV在WI任务上性能提升了17%左右，在WD任务上性能提升了35%左右（采取GAP分类则提升45%左右）。在本部分工作，将针对多尺度融合特征、Conv-Module下采样方法、decoder多头优化和稀疏性损失以进行消融实验，以此验证各部分对于OSVTF的性能提升幅度，得到的WI实验结果如表5。

表5：WI消融实验 (EER %)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model |  |  |  |  | BHSig-B  (80:20) | BHSig-H  (100:60) | CEDAR  (50:5) |
| TransOSV |  |  |  |  | 5.11 | 4.86 | 5.94 |
| Baseline | √ |  |  |  | 4.71 | 4.39 | 5.40 |
|  |  | √ |  |  | 4.83 | 4.55 | 5.67 |
|  |  |  | √ |  | 4.92 | 4.76 | 5.61 |
|  | √ | √ | √ |  | 4.64 | 4.21 | 5.21 |
| OSVTF | √ | √ | √ | √ | 4.54 | 3.90 | 4.75 |

由此可以说明，在TransOSV的基础上增加多尺度特征图的融合特征，能够大幅度减少模型误判率。采取了下采样方式的Conv-Module（在decoder中同步增加了输入映射），增强了特征的抽象表达能力，以用于更复杂的模型运算当中。在decoder的多头优化上，增加了多个角度去关注特征向量，因此模型性能得到了一定的提升。在sparsity loss的训练方案上，稍微弱化了decoder中注意力集中分布的情况，使得更加关注签名图像的各个局部特征。

在本工作的Conv-Module和decoder部分，设计了降维→升维的映射结构，将整个Conv-Module从原本的特征维度数变化变成了，为了顺应这种变化在Encoder输入阶段引入了一个“升维”的线性映射层。为了验证“降维→升维“映射结构的可行性，在BHSig-B 50/50数据集的实验中设置偏大的学习率和20个训练周期，优化前的损失函数和评价指标变化如图9所示，优化后的损失函数和评价指标变化如图11、12所示。

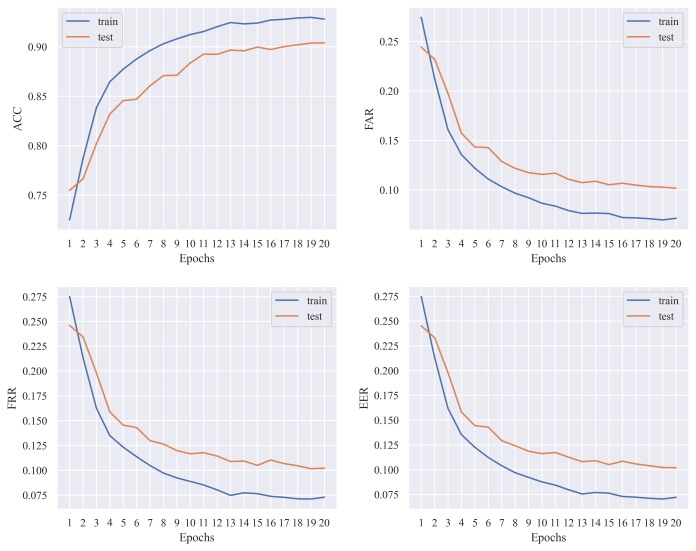
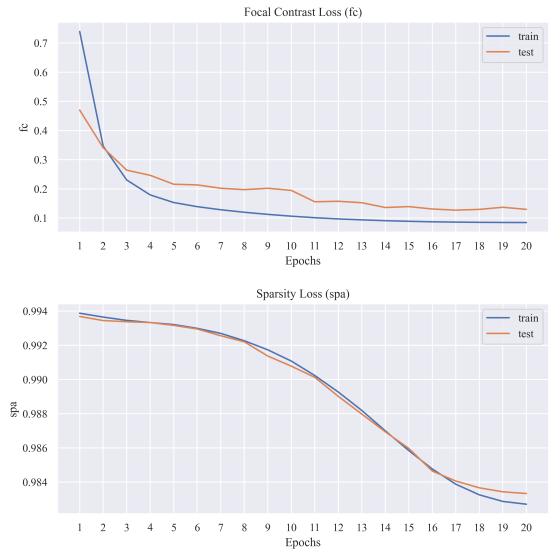
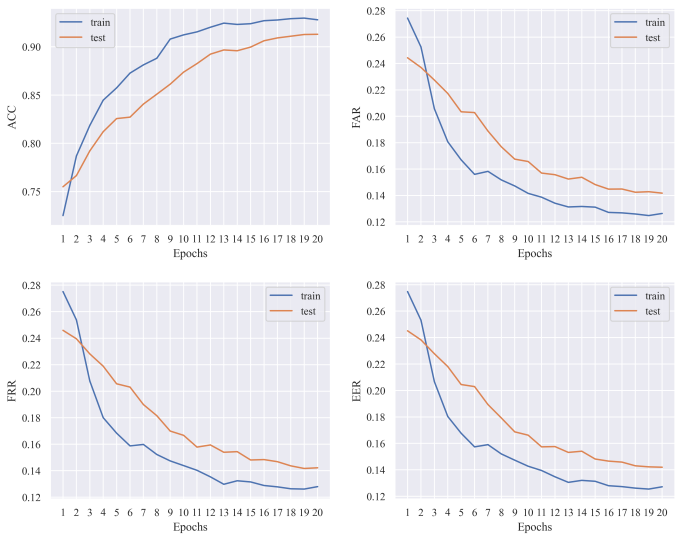
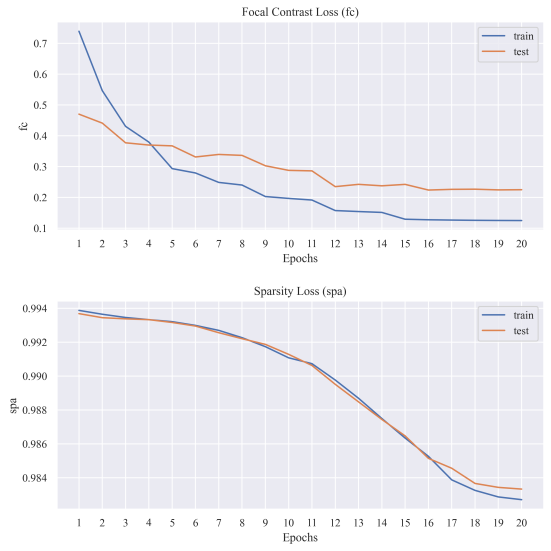


图9：previous Conv-Module loss and metrics diversification

  
图10：optimized Conv-Module loss and metrics diversification

根据结果可以发现，优化前的OSVTF在拟定实验方案中，大约在第7、8个Epoch时参数开始收敛，而优化后的OSVTF大约在第3、4个Epoch时参数开始收敛。这种下采样后再映射的特征处理方式，能够有效地加速模型训练过程的收敛速度。由于OSVTF中间网络层提取的特征数量较多，因此采取这种方式能够更快速学习伪造签名的敏感部分特征，并且结合交叉多头注意力的计算，更进一步强化了对伪造签名敏感部分的特征学习，并且这种方式能够减少一定的模型参数量，使得模型权重文件体积变小，推理速度加快。

**5. Conclusions**

在本工作中，针对离线手写签名验证的现阶段深度学习模型，基于CNN+Transformer风格提出了图像多尺度特征融合的深度学习模型架构。该模型架构是在类似孪生网络风格的TransOSV模型架构基础上进一步改进，在采取的多种不同特征组合机制的基础上，增加了CNN中多个网络层提取不同尺度特征图的特征，并使用FPN-Style的FPN Fusion模块对多尺度特征进行融合。其次，对子网络的Conv-Module和Part Decoder进行了适应多尺度融合特征的优化。为了加速增加了多尺度融合特征的模型推理速度，对Conv-Module采取了类似下采样的调整，将Holistic输出的特征向量重构成2D图像特征后，在特征维度上采取了下采样的卷积运算，通过该部分降低通道数量和图像尺寸的方法，以优化图像中无用特征的冗余。在Part Decoder中，增加一个映射运算，将Conv-Module下采样操作的图像特征平坦化后映射至高维度中，并经过改良的交叉多头注意力计算，从而在增加了特征提取子网络的模型架构中加速整体推理过程。

在实验阶段部分，总共分为在WI任务上的复现实验、WD任务上分类器的对比实验、模型消融实验三部分。在WI任务上，采取BHSig-B和BHSig-H数据集对初步构建的OSVTF进行一定epochs训练。针对BHSig-B数据集，50/50比例是为了评估模型的泛化能力，80/20比例是为了评估模型的综合学习能力。在WD任务上，着重对比GAP分类器与SVM分类器的不同。在模型实验部分，针对多尺度融合特征、Conv-Module、part decoder的优化进行了横向对比。实验结果表明，采取了多尺度融合特征后，对Conv-Module和part decoder的优化能够使得模型训练时参数更快收敛。

在横向对比上，OSVTF的性能已经优于之前的TransOSV模型。但是在模型训练过程，受限于训练环境和模型细节缺失，因此无法完美复现TransOSV模型的性能。同时在离线签名验证任务中，没有对图像进行特殊的预处理，设置的图像尺寸是略大于ImageNet的图像尺寸，并且在backbone上是采取了弱化的ResNet-34进行训练。因此在下一步工作中，将尝试设置更多GPU，针对backbone不同的ResNet系列模型，设置固定的图像尺寸和图像预处理工作，从而实验出不同规格的OSVTF模型，能够尽可能地在各个设备环境下进行工作。

**Acknowledgements (致谢)**

首先，我要衷心感谢我的导师刘新教授。教学期间她给予了我许多鼓励，让我能在研究难题疑点的时候克服困难，战胜心理懒惰，不断地挑战、超越自己。从选题报告至毕业论文期间，提供了许多研究角度和创新点，研究期间给予我许多关键性建议，细心检查论文并提出相关修改建议，这些使我更好地完善每个部分内容，不仅使我掌握了更多知识，还懂得如何处理好各种细节。

其次，我要感谢创新工程学院的授课教授们，在这两年的学习时间内，教授们倾尽所学传授知识，让我以不同视角重新构建了以前的知识框架。并且授课过程给予了我展示个人能力的舞台，让我学习到新知识的同时锻炼自己的演讲能力，以英语为重的授课方式在一定程度上锻炼了自己的英语读写能力。

最后，我要感谢我的家人在学习期间给予我的支持和帮助，每一次关键时刻都能让我冷静地去思考当时处境，从自身出发寻找解决方法，也让我懂得为人做事一定要冷静对待，听他人说话时要成为一名良好的倾听者，跟他人说话时要做一名口齿伶俐的诉说者。

**References (参考文献)**

[1]

J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer Normalization,” *arXiv preprint arXiv:1607.06450*, 2016.

[2]

D. Banerjee, K. Dasgupta, D. Ganguly, and K. Chatterjee, “A Survey of Offline Handwriting Signature Recognition,” Mar. 2019.

[3]

A. Bhatawdekar, S. Bhattacharya, R. Khatri, and R. Tiwari, “BHSig260: A Dataset for Offline Signature Verification,” *arXiv preprint arXiv:2004.07563*, 2020, [Online]. Available: <https://arxiv.org/abs/2004.07563>

[4]

A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: <https://arxiv.org/abs/2010.11929>

[5]

R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2000.

[6]

T. Fawcett, *An Introduction to ROC Analysis*, vol. 27. 2006.

[7]

M. A. Ferrer, J. B. Alonso, and C. M. Travieso, “Offline Geometric Parameters for Automatic Signature Verification Using Fixed-Point Arithmetic,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 993–997, 2005.

[8]

I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[9]

H. Li S. Zhang and M. Ouyang, “Multimodal Signature Verification via Hybrid CNN-Transformer Networks,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 11230–11239.

[10]

R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality Reduction by Learning an Invariant Mapping,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 1735–1742. doi: [10.1109/CVPR.2006.45](https://doi.org/10.1109/CVPR.2006.45).

[11]

L. G. Hafemann, R. Sabourin, and L. S. Oliveira, “Learning features for offline handwritten signature verification using deep convolutional neural networks,” *Pattern Recognition*, vol. 70, pp. 163–176, 2017, doi: <https://doi.org/10.1016/j.patcog.2017.05.012>.

[12]

L. G. Hafemann, R. Sabourin, and L. S. Oliveira, “Offline handwritten signature verification — Literature review,” in *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Nov. 2017, pp. 1–8. doi: [10.1109/IPTA.2017.8310112](https://doi.org/10.1109/IPTA.2017.8310112).

[13]

K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” *IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017, doi: [10.1109/ICCV.2017.322](https://doi.org/10.1109/ICCV.2017.322).

[14]

K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. [Online]. Available: <https://arxiv.org/abs/1512.03385>

[15]

D. Hendrycks and K. Gimpel, “Gaussian Error Linear Units (GELUs),” *arXiv preprint arXiv:1606.08415*, 2016.

[16]

S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in *Proc. 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 448–456.

[17]

E. J. R. Justino, F. Bortolozzi, and R. Sabourin, “A comparison of SVM and HMM classifiers in the off-line signature verification,” *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1377–1385, 2005, doi: <https://doi.org/10.1016/j.patrec.2004.11.015>.

[18]

R. G. Keys, *Cubic Convolution Interpolation for Digital Image Processing*, vol. 29. 1981.

[19]

A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, Curran Associates Inc., 2012, pp. 1097–1105.

[20]

L. L. Lee, T. Berger, and E. Aviczer, “Reliable on-line human signature verification systems,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 643–647, Jun. 1996, doi: [10.1109/34.506415](https://doi.org/10.1109/34.506415).

[21]

C. Li, F. Lin, Z. Wang, C. Yu, L. Yuan, and H. Wang, “Deephsv: User-independent offline signature verification using two-channel CNN,” in *International Conference on Document Analysis and Recognition (ICDAR)*, 2019.

[22]

M. Lin, Q. Chen, and S. Yan, “Network In Network,” in *Proc. International Conference on Learning Representations (ICLR)*, 2014. [Online]. Available: <https://arxiv.org/abs/1312.4400>

[23]

T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature Pyramid Networks for Object Detection,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017, doi: [10.1109/CVPR.2017.106](https://doi.org/10.1109/CVPR.2017.106).

[24]

Z. Liu *et al.*, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 10012–10022.

[25]

X. Lu, L. Huang, and F. Yin, “Cut and Compare: End-to-End Offline Signature Verification Network,” in *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 176–183. doi: [10.1109/ICPR48806.2021.9412496](https://doi.org/10.1109/ICPR48806.2021.9412496).

[26]

T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[27]

R. K. Mohapatra, K. Shaswat, and S. Kedia, “Offline handwritten signature verification using CNN inspired by inception V1 architecture,” in *2019 Fifth International Conference on Image Information Processing (ICIIP)*, 2019, pp. 263–267. doi: [10.1109/ICIIP47207.2019.8985685](https://doi.org/10.1109/ICIIP47207.2019.8985685).

[28]

Y. Muhtar, W. Kang, A. Rexit, Mahpirat, and K. Ubul, “A Survey of Offline Handwritten Signature Verification Based on Deep Learning,” in *2022 3rd International Conference on Pattern Recognition and Machine Learning (PRML)*, Jul. 2022, pp. 391–397. doi: [10.1109/PRML56267.2022.9882188](https://doi.org/10.1109/PRML56267.2022.9882188).

[29]

V. Nair and G. E. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines,” *Proc. 27th International Conference on Machine Learning (ICML)*, pp. 807–814, 2010.

[30]

S. Pankanti, S. Prabhakar, and A. K. Jain, “Cedar: A database of handwritten signatures for benchmarking signature verification systems,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2000, pp. 5–8. doi: [10.1109/ICIP.2000.895066](https://doi.org/10.1109/ICIP.2000.895066).

[31]

R. Plamondon and G. Lorette, “Automatic signature verification and writer identification—The state of the art,” *Pattern Recognition*, vol. 22, no. 2, pp. 107–131, 1989, doi: [10.1016/0031-3203(89)90059-9](https://doi.org/10.1016/0031-3203(89)90059-9).

[32]

S. Dey A. Dutta and J. I. Toledo, “Signet: A Deep Learning Architecture for Offline Signature Verification Using Multi-Scale CNN and SVM,” *IEEE Access*, vol. 8, pp. 118942–118956, 2020.

[33]

A. Shrivastava, A. Gupta, and R. Girshick, “Training Region-based Object Detectors with Online Hard Example Mining,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 761–769.

[34]

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: <https://arxiv.org/abs/1409.1556>

[35]

J. F. Vargas, M. A. Ferrer, C. M. Travieso, and J. B. Alonso, “Off-line Handwritten Signature GPDS-960 Corpus,” in *Proc. 9th International Conference on Document Analysis and Recognition (ICDAR)*, 2007, pp. 764–768. doi: [10.1109/ICDAR.2007.4376996](https://doi.org/10.1109/ICDAR.2007.4376996).

[36]

A. Vaswani *et al.*, “Attention is All You Need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008. [Online]. Available: <https://arxiv.org/abs/1706.03762>

[37]

M. Wang, S. Lu, D. Zhu, and J. Lin, “A High-Speed and Low-Complexity Architecture for Softmax Function in Deep Learning,” in *2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, Chengdu, China: IEEE, 2018, pp. 223–226.

[38]

P. Wei, H. Li, and P. Hu, “Inverse discriminative networks for handwritten signature verification,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[39]

X. Wang Y. Chen, Z. Liu and L. Jin, “Vision Transformer for Offline Handwritten Signature Verification,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 3, pp. 450–465, Jul. 2022.

[40]

M. H. M. Yusof and V. K. Madasu, “Signature Verification and Forgery Detection System,” in *Proc. Student Conference on Research and Development (SCORED 2003)*, 2003, pp. 9–14.

[41]

Y. Zhang *et al.*, “TransOSV: Offline Signature Verification with Transformers,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [Online]. Available: <https://arxiv.org/abs/2305.08615>