**Thesis Proposal**

**(論文選題報告)**

**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一個具體應用, 基於個人提供的手寫簽名以驗證身份. 日常用於安全認證, 金融交易, 存取控制系統和其他安全領域.

    離線手寫簽名驗證包含兩種類型的任務: 作者依賴和作者獨立. 第一種任務是將手寫簽名與資料庫的對應用戶手寫簽名進行對比驗證。第二種任務是僅透過提供的手寫簽名來判斷簽名是否偽造

本文研究内容包括: 1, 研究與復現卷積神經網絡與Transformer融合的模型架構，以實現結合學習圖像局部特徵和全局特徵的模型架構. 2. 提出端到端的一種結合卷積神經網絡和Transformer的離綫手寫簽名驗證模型架構OSVTF，能夠更佳全面地提取手寫簽名圖像特徵. 3. 先使用支持向量機作爲特徵分類器以鑒別手寫簽名是否為僞造的，根據調研結果給出更優秀的特徵分類器，實現更優秀性能的離綫手寫簽名驗證模型架構。

**關鍵詞**: 離線手寫簽名驗證; 端到端; Transformer; OSVTF; 支持向量機.

**1. Introduction（引言）**

**1.1. Research Background（研究背景）**

Biometrics technology是一种根据特征、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企业的安全认证、金融交易、门禁系统等安全领域**[1]**。

该技术主要应用于识别和验证场景。在第一种场景下，用户仅需提供个体或生理特征样本，生物特征识别系统根据样本在注册用户中识别提供特征的用户。而第二种场景则是在第一种场景的基础上，向系统声明自己的身份，随后系统将根据以上信息验证注册用户中是否为指定用户。

手写签名是日常生活中较为重要的个体行为特征，因为其作为在法律、金融、行政等领域验证个人身份的主要特征。同时也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。由此便衍生出手写签名验证的具体应用，根据用户提供的签名（定义为query）与系统中该用户存储的签名（定义为Reference）进行对比验证，从而鉴定声明的身份是否属实。

手写签名验证，根据收集途径不同分为离线和在线两种类型。离线手写签名验证的个人签名收集过程是通过用户在纸张上进行书写过程后得到，而在线手写签名验证采用数字化台方式收集用户的手写签名特征，因此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。由此可以看出离线手写签名的优势，可以随时随地不受设备限制地收集用户个人签名图像，以对个人身份进行验证。

研究离线手写签名验证，将更全面理解和对比过去传统机器学习方法与近十年深度学习方法的离线手写签名验证区别，从而提出具有更高效、更精准的手写签名验证模型架构。由此，安全领域可以更替为更保障的手写签名验证模型或算法，能够更有效且精准地完成每个用户地个人身份验证请求，从而达到节约机器验证的时间成本，为其他核心任务提供更多的计算机推理资源和计划时间，一定程度上能够节约验证个人身份成本。由于新的模型架构提出，会对已有的模型提供一种新的设计或改良方案，从而衍生出更多优秀的离线手写签名验证模型架构，以此推进离线手写签名验证模型或算法的研究水平。

**1.2. Motivation（动机）**

离线手写签名验证在收集数据的过程，虽然不会受到数字化台等设备因素影响，但是用户使用纸张书写个人签名的时候不能保证多次书写得到的签名完全一模一样，可能也会受到实体笔、纸张厚度等影响。因此收集数据后进行的预处理过程一定程度上会影响系统验证效果，一个良好的图像预处理技术会直接影响系统验证性能。

在过去的离线手写签名验证研究中，学者们根据样本分布以设计图像特征，采取传统机器学习的方法以验证签名是否为伪造的**[1]**。但是这种方法存在缺陷，验证可信度和准确率很大程度依赖于人工设计的特征，且对手写签名的字体风格等有限制要求，数据预处理的时间成本十分巨大。随着深度学习和卷积神经网络发展，近几年学者们逐渐采取卷积神经网络的卷积部分对手写签名进行推理，将推理得到的多通道特征图作为手写签名验证的特征向量。这种监督学习的特征提取方式取得了不错的效果**[2]**。但是仅仅通过卷积神经网络提取多通道特征图，会因为其模型架构中卷积层的共享参数卷积核运算，导致学习图像特征过程会缺失部分全局特征。基于此在图像分类任务上学者们基于Transformer提出了性能更优秀的Vision Transformer架构**[10]**。与卷积神经网络相比，实验证明Transformer**[9]**内部的注意力机制拥有更优秀的全局特征学习能力，能够在特征的重要部分累加注意力权重，能够更关注有效特征。由此期望提出一种能够弥补卷积神经网络缺少的全局特征学习能力架构，从而更好地结合卷积神经网络优秀的局部特征学习能力。

离线手写签名验证包含两种任务：Writer-Dependent (WD)和Writer-Independent (WI)。WD任务是需要Reference和Query一对手写签名样本以验证签名是否伪造，一定程度上会依赖于writer的Reference的特征。WI任务是仅仅通过Query来验证签名是否伪造，不会依赖于writer的Reference特征。过去学者们是采取支持向量机作为特征分类器，根据人工设计特征以验证手写签名是否伪造。当图像尺寸越大，特征数量越多，从而导致支持向量机的模型参数难以收敛，增加整体模型架构训练成本。由此期望提出一种能够较快收敛且效果优于支持向量机的特征分类器。

**1.3. Research Objectives（研究目标）**

针对WD和WI任务，本文主要目标如下：1. 复现OSV（Offline Signature Verification）模型; 2. 提出一种结合卷积神经网络局部特征学习能力和Transformer全局特征学习能力的OSV模型架构; 3. 提出更快收敛且效果更好地支持向量机的离线手写签名特征分类器

第一阶段，将现有较新的OSV模型进行复现，分析和总结OSV模型架构细节和特点，找出它们之间的共同点和不足，从而更好地找到优化关键点与方法。

第二阶段，在复现模型的基础上提出基于Transformer的优化方案，采取对比实验以验证模型性能比以往模型性能更优秀。在对比实验过程中，约束图像预处理和验证方法一致，从而找出适合模型架构的最终训练验证方案。

第三阶段，由于图像分类任务中的卷积神经网络在最后输出类别概率上与离线手写签名验证任务的有所区别，在提出的模型上先使用支持向量机进行实验是否能够有效使用提取的图像特征，再更替为效果更优的特征分类器。

**2. Literature Review（文献综述）**

Biometrics technology被广泛应用于各种安全领域。这类系统目的是根据生理或行为特征识别一个人**[1]**。该技术识别一个人除了通过指纹、面部、虹膜等信息，还可以通过个人手写签名以识别个人身份。

手写签名可以用于识别和验证任务。手写签名识别任务是给定一个人的手写签名，系统将根据该签名在数据库中寻找此手写签名匹配的用户。手写签名验证任务是用户提供手写签名的同时，对系统声明自己是数据库中某个用户，随后系统将其手写签名与数据库中对应的用户手写签名资料库进行对比，从而验证是否确实为此人。

手写签名验证任务根据数据分为两种类型：offline和online。Offline指的是在个人在一张纸上获取不同人的签名，然后扫描进行验证。Online指的是通过个人签名的数字化平板电脑识别签名**[3]**。前一种方式不会受到数字化平板电脑型号的影响，因此离线手写签名验证不需要在数据预处理过程进行过多的操作，就可以得到纯粹的个人书写签名。

无论离线还是在线手写签名，它们收集得到的数据均是图像数据，因此一个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。为了提高签名图像的质量，学者们均采取了RGB图像转换到GRAY单通道图像的处理，更有甚者为了去除噪点使用了平滑像素的图像处理**[4]**。

离线手写签名早期均是采取机器学习的方式，如 Edson等人利用隐马尔可夫模型提出了离线签名验证系统**[5]**。这种传统机器学习方法存在缺陷：此方法极其依赖人工设计的特征进行训练，分类的准确性与这些特征直接相关**[2]**。随着深度学习和卷积神经网络发展，2012年Krizhevsky A.等人提出AlexNet**[6]**在ImageNet比赛的图像分类任务上遥遥领先使用传统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，学术界后续推出更深层次的卷积神经网络如VGG**[7]**，具有残差连接的ResNet**[8]**。这些卷积神经网络均是有多个卷积层和线性层组成，前面部分的卷积层是使用了共享参数的卷积核遍历图像像素点的值，从而达到学习图像特征的效果。该方式提取的图像 特征与传统图像特征提取方法不同的是，卷积神经网络提取特征的方式是属于监督学习，其必须依靠数据量才能够拥有一个较好的特征提取效果。由此，卷积神经网络在图像分类任务上即使最后分类依然通过先行层输出类别概率，但是依然拥有先天优势。

上述提到卷积神经网络是使用了卷积核对图像像素进行遍历以学习图像特征，但是卷积运算存在局部视野的缺陷。虽然它能够学习图像的局部特征，但是在学习全局特征的效果上较差，导致卷积神经网络的图像分类任务在较为优秀的识别准确率上存在达到最优秀识别准确率的阻碍。随着深度学习的自然语言处理领域发展，2017年Vaswani A.等人提出用于机器翻译的Transformer模型**[9]**。该模型采取了Encoder-Decoder的机器翻译主流架构，与以往机器翻译深度学习模型不同的是，Transformer引入了注意力机制[]。在注意力机制的作用下，能够对词向量进行全局特征感知，生成预测词语时会参考上一个词语的特征信息和输入词向量的特征信息，因此可以生成长短不一句子预测。并且实验证明了Transformer相较以往的Encoder-Decoder机器翻译模型拥有更优秀的全局泛化能力。

介于Transformer的注意力全局特征感知能力，Dosovitskiy A. 等人于2021年首次将Transformer Encoder用于图像分类任务中，命名为Vision Transformer (ViT)**[10]**。ViT的内部结构完全抛弃了卷积层，将输入图像进行patch和 Embeddings操作 后，每个patch的embedding当成词向量单个token，由此图像被拉平后当作词向量进入Transformer Encoder。经过若干个Encoder Layer，最后设置一个MLP以映射类别概率，从而完成图像分类任务。实验表明这种方式的模型在图像分类任务上得到了领先卷积神经网络的识别准确率，由此也衍生出CNN + Transformer的深度学习模型架构，如图像目标检测的DETR**[11]**、图像分割的MaskFormer**[12]**。这种方法采取的是摘除最后分类的Linear层卷积神经网络，输入图像以提取多通道特征，将该特征作为Transformer的输入向量从而进行全局特征感知。CNN + Transformer的开山之作DETR在MS COCO**[13]**的验证效果虽然没有以往采取NMS等后处理操作的R-CNN**[14]**，YOLO系列网络**[15]**识别准确率高，但是该方式的优势在于不需要任何的Archer Box等目标检测先验框预设置和NMS后处理，是一种端到端的模型架构。这种方式一定程度上简化了模型推理流程，但是也增加了模型训练时间成本，需要设计更精细的训练方案以保证模型参数收敛。

Transformer在图像任务上相较CNN具有优秀的全局特征感知能力，因此离线手写签名验证领域中学者们提出了一种基于Transformer的模型架构TransOSV**[16]**。该模型采取类似Encoder-Decoder架构，输入Reference和Query均经过RGB转GRAY图像预处理后，进入以ViT Encoder作为Holistic Encoder。其次将Holistic Encoder输出进行卷积运算，最后对Reference和Query的特征进行 Contrast based Part Decoder运算。在训练过程中，TransOSV是汇总了Transformer Encoder输出class特征、卷积模块输出特征和解码器输出特征以计算训练损失完成模型训练，最后特征分类器同样采取支持向量机进行验证签名是否伪造。该架构在Decoder的Cross-attention中将Reference特征和Query特征进行了注意力计算，对Reference和Query的特征进行关联度注意力学习，能够更好地联系Reference和Query的关系。但是这种架构由于输入图像是直接patch和Embeddings进入Transformer Encoder，一定程度上会缺少图像多通道特征信息。由于ViT其图像尺寸大小是全局约束的，当图像进行resize缩小的时候一定会丢失部分关键信息。

综上所述，本文将在TransOSV**[16]**架构基础上，采取CNN + Transformer的方式，在图像进入Transformer Encoder之前先由CNN提取多通道特征图，随后进入Transformer以补足全局特征信息。此外优化Holistic Encoder架构，能够接受图像尺寸变动，从而更好地适配不同数据集样本的实验方案。

**3. Methods（论文主题算法等）**

在本章节中，首先总结使用的方法并详细介绍所提出的模型架构。深入解析所提出的包含编码器、卷积模块、解码器架构的特征提取器具体细节。随后将介绍模型参数训练过程的损失函数、优化器、作者独立分类器**[17]**等细节。最后描述是否共享特征提取器模型参数的实验、分类器种类的实验，以达到效果良好的手写签名离线验证任务。

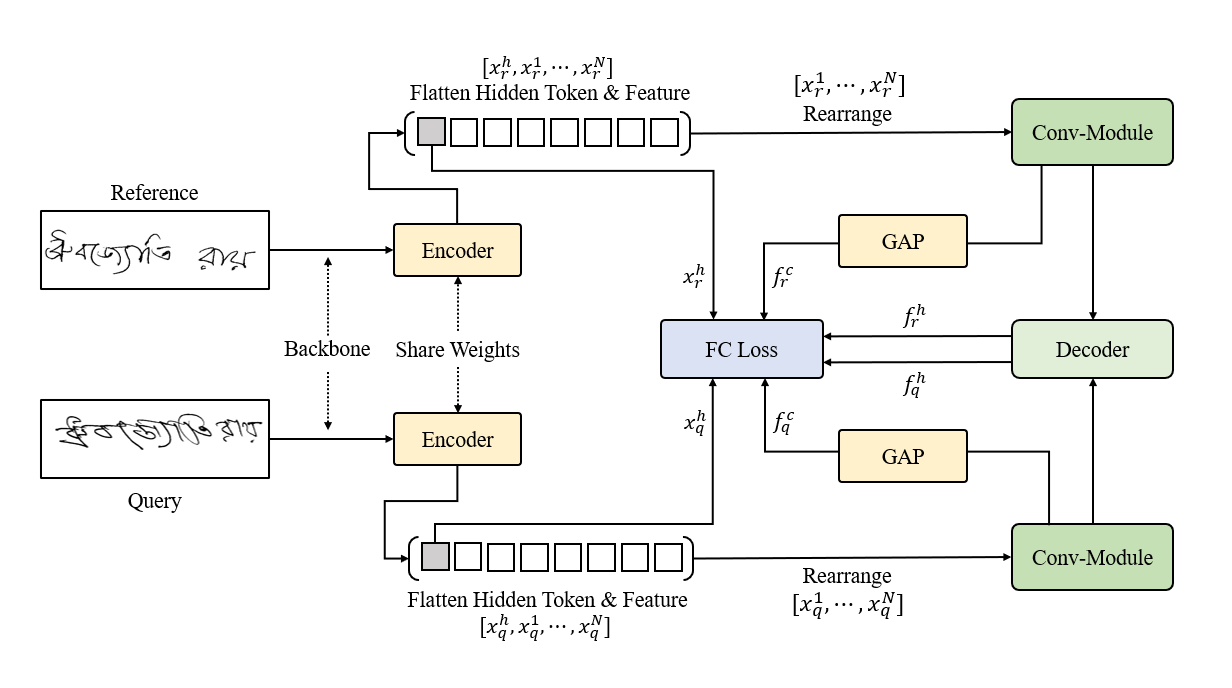
**3.1. Overview** 

图1: Offline Signature Verification TransFormer (OSVTF)

提出的OSVTF如图1所示。输入一对尺寸相同的手写签名图像 Reference和Query。起始会经过Backbone以提取多通道特征图。接着多通道特征图经过权重共享的Encoder以学习图像特征，该部分推理将提取得一对平坦特征token 和一对平坦特征图。随后平坦特征图会经过reshape到二维图像向量再进行卷积模块推理。其中卷积模块的输出将分为两个方向：第一个方向将通过Global Average Pooling层得到平坦卷积特征；另一个方向将成对进入Decoder得到平坦注意力解码特征。训练过程将收集推理过程中平坦特征token 、平坦卷积特征和平坦注意力解码特征以进行Focal Contrast Loss (FC)损失计算，从而进行模型权重的训练。接下来将深入解析Encoder、Decoder和Conv-Module结构。

**3.2. Encoder**

Encoder是沿用了ViT(Vision Transformer) **[10]**的模型架构如图2所示。

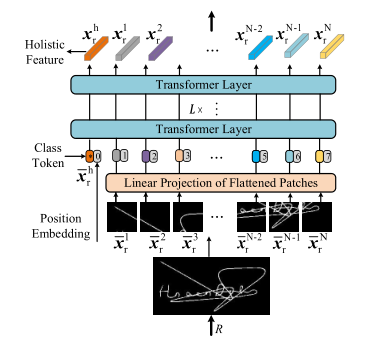


图2: Vision Transformer

以Reference（形状为）举例，其中表示图像通道数、高度和宽度。ViT首先将输入图片分成块和平坦化，其中patch number如式（1）所示。

其中表示floor function。和分别表示patch size和卷积层步数，因为ViT原文架构是采取卷积层以完成分块操作。其次，拼接一个可学习权重Class Token和patch图像累加Position Embedding，形状是。由于Transformer Encoder要求输入的是Token向量，图像patch后会丢失相对位置信息，因此累加了Position Embedding重新赋值了相对位置特征。经过上述处理后进入层Transformer Encoder Layer推理得到Encoder的平坦特征token与特征图，整个Encoder推理如式（2）。

其中表示Embeddings。表示Position Embeddings。每个Encoder结构相同，由个堆叠形成，如图3所示。

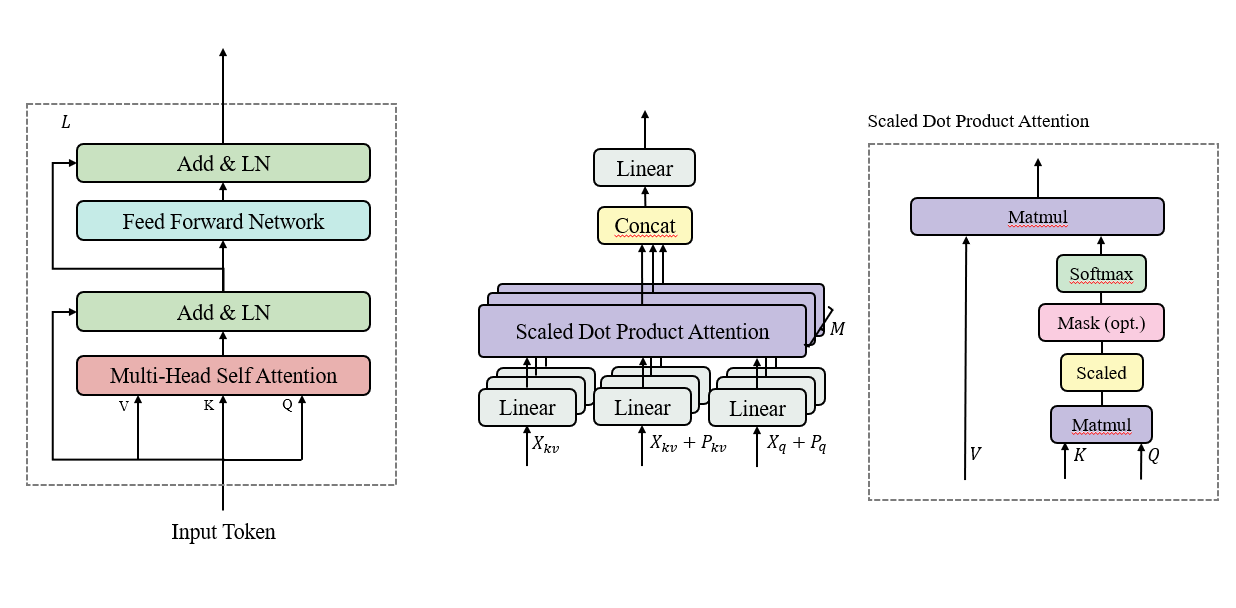


图3: Transformer Encoder

每个Encoder Layer均由一个Multi-Head Self Attention (MHSA)、Feed Forward Network (FFN) 和两个残差累加&Layer Normalization (LN) 组成。首先，输入Token进入Encoder之前，会经过三个Linear层将输入向量进行映射得到，将映射后的三个向量输入MHSA中进行注意力特征计算。Encoder中MHSA采取的和Transformer**[9]**的MHSA相同，均是使用了Scaled Dot Product Attention，具体计算公式如式（3）。

其中表示MHSA的维度数。由于Scaled Dot Production Attention在计算过程涉及两次规模较大的矩阵乘法，因此在输入过程会根据定义头的数量以生成头数量乘3的维度线性映射权重，以此将输入向量映射成较低维度数的三个矩阵。在每个头上都进行运算，最终通过一个大的线性层将注意力计算后得到的矩阵以映射到输入向量维度数。由此堆叠Transformer Encoder能够多次进行MHSA计算从而更好学习整体向量特征。

**3.3. Conv-Module**

输入图像经过Encoder后得到的均是平坦化特征向量，因此需要对其进行reshape才能够进行卷积运算。同时平坦化Token后会丢失部分二维信息，因此在reshape后经过卷积模块能够弥补部分二维信息。卷积模块的组成和卷积神经网络相似，由四个Conv2D+ReLU和两个Max Pooling 2D组成，顺序如表1。

表1：

|  |  |  |
| --- | --- | --- |
| 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 |  |  |

经过Conv-Module后得到的输出会进行两个方向的传播：第一个方向是进行Global Average Pooling (GAP) 计算和平坦化，从而得到平坦卷积特征；另一个方向是成对进入Decoder进行解码注意力计算。

**3.4. Decoder**

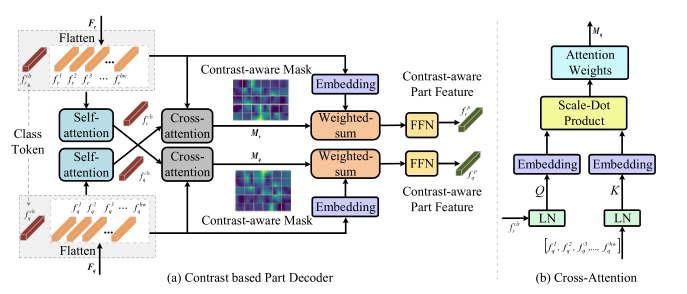


图4：

OSVTF中的Decoder与Transformer Decoder不同，它采取了两个MHSA和Cross-attention，如图4-(a)所示。与ViT输入阶段相似，在输入平坦化卷积特征之前，会初始化一对与平坦化卷积特征相同维度数的可学习参数，定义为Class Token。它的作用与ViT-Encoder的Class Token作用一致，均是定义了它在一个新的维度上学习图像是否为赝品的一个调整权重，从而使得模型能够更好地学习图像特征。

同上述MHSA一样，计算公式如式（3）。Decoder中先对插入了class token的平坦化卷积特征进行MHSA运算得到注意力机制的平坦化特征向量。随后为了能够学习reference和query之间的关系，引入了Cross-attention如图4-b，基于注意力机制定义的线性映射为Query矩阵，MHSA输出的平坦化注意力特征为Key矩阵。与MHSA相似，根据reference或query的可学习参数class token，和另外图像的平坦化特征进行注意力机制计算，一定程度上注意力机制内部的查询矩阵能够反应图像之间的关注度，由于没有输入值Value矩阵，因此Cross-attention得到注意力权重。该mask权重包含了reference和query之间各个Token（即各个像素点）的注意力权重，由此对输入Decoder的特征再经过Embeddings后与注意力权重进行广播后累加，从而得到的是一对注意力机制的平坦化注意力特征。经过FFN后得到最终的Decoder输出，这对特征后续将根据FC Loss来优化Decoder模型参数。

**3.5. Loss Function**

OSVTF的训练过程将采取了两种损失函数：Sparsity Loss, Focal Contrast Loss。

**Sparsity Loss**

由于OSV的图像中只有少数区域包含签名验证任务的判别信息，且Decoder中具有稀疏性，将采取计算交叉熵的方式以生成对比度感知掩码的多样化分布**[16]**，从而更好的训练Decoder的线性映射参数，计算公式如式（4）。

**Focal Contrast Loss**

在以往的非监督算法中，对比两个样本或特征之间的差异是采取计算distance的方式以衡量两者差异。定义为计算包括Encoder、Conv-Module和Decoder输出的一对平坦化特征向量之间差异。由此可以定义评价两个对象差异的Contrastive Loss**[18]**计算公式如式（5）。

其中，表示query的signature是forged，相反时表示赝品。为了抑制模型训练过程过拟合的问题，基于CaP**[19]**引入双倍边际损失如式(6)。

但是上述的双倍边际损失中存在缺点，当两对reference和query signature样品中query为正品，即的情况下，会得到两个此样本模型推理特征的距离。若，那么损失函数应该给予样本更大的损失/权重。但是在上述的损失函数中会平等地对待样本，因此基于Focal loss**[20]**的不平衡训练样本情况，改良得到最终的Focal Contrast Loss如式（7）。

其中表示Sigmoid function。为两个边距值。表示缩放因子。

**3.6. Dataset**

和以往学者们使用的公开数据集相同，将采取三种数据集以验证OSVTF模型的性能：BHSig-B、BHSig-H**[21]**和CEDAR**[22]**。

**BHSig-B & BHSig-H**

由印度理工学院古瓦哈提分校（IIT Guwahati）发布。其中BHSig-B是包含了100个用户的孟加拉语（Bengali）手写签名，BHSig-H是包含了160个用户的印地语（Hindi）。每个用户手写签名了24张真实签名和30张伪造签名。

**CEDAR**

由Center of Excellence for Document Analysis and Recognition开发并发布。该数据集包含了55个用户的英文手写签名。每个用户手写签名了24份真实签名和24份伪造签名。

**4. Schedule for the thesis（论文时间安排）**

**2024.06 – 2024.10**

对OSV模型进行调研，复现近几年基于CNN和Transformer的OSV模型架构。整理离线手写签名验证公开数据集，分析总结TransOSV模型的训练、推理细节，完成TransOSV架构的复现代码。

**2024.11– 2024.12**

完成本文提出的OSVTF架构实现代码。调研学术界用于一对特征样本的分类器模型架构或算法，完成这些分类器的复现代码从而在OSVTF上实现。制定OSVTF的模型训练、验证方案，在GPU服务器上进行试运行，目标是计算训练模型开销和收敛大致时间，从而租赁对应服务器进行模型训练。

**2025.01 – 2025.02**

进行以下实验：1. OSVTF的Conv-Module和GAP是否共享参数。2. OSVTF的Decoder中 Embeddings是否共享参数。3. OSVTF的注意力运算是否需要重新累加一次位置编码。4. 找出最佳模型性能的训练超参数。将根据实验结果和导师反馈完成初稿撰写，成果包括拟完成初稿和最新研究进展综述。

**2025.03 – 2025.04**

根据实验结果和导师反馈意见修改论文，解决提出的问题，准备提交最终版本论文。准备答辩材料和演示文稿。

**2024.05 - 2025.06**

完成论文的最终审核和格式调整，并成功通过答辩，最终成果是获得硕士学位并发表论文。

**5. Publication（出版物）**

* Learning Spatiotemporal Features for Video Semantic Segmentation Using 3D Convolutional Neural Networks, ISCSIC-2022, IEEE, 14 March 2023, DOI: 10.1109/ISCSIC57216.2022.00023.
* End-to-End Chinese Lip-Reading Recognition Based on Multi-modal Fusion,'' ICFTIC-2022, IEEE, 27 March 2023, DOI: 10.1109/ICFTIC57696.2022.10075247.
* Offline Signature Verification Using a 2D Attention Encoder-Decoder Network'', ICNCC-2023, ACM, 07 March 2024, DOI: 10.1145/3638837.3638880.

**References (参考文献)**

[1] 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.

[2] 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.

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

[4] N. Y. Choudhary, R. Patil, U. Bhadade, and B. M. Chaudhari, "Signature Engineering and Applied Sciences Research (IJIEASR)," *International Journal of Innovative Engineering and Applied Sciences Research*, vol. 2, no. 1, pp. [pages if available], Jan. 2013.

[5] J. Edson, R. Justino, E. Bortolozzi, and R. Sabourin, "An offline signature verification using HMM for random and skilled forgeries," in *Proc. 6th Int. Conf. Document Analysis and Recognition*, pp. 1031-1034, Sept. 2001.

[6] 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.

[7] 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].

[8] 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].

[9] A. Vaswani *et al.*, “Attention is All You Need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008. [Online].

[10] 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].

[11] N. Carion *et al.*, “End-to-End Object Detection with Transformers,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229. [Online].

[12] B. Cheng, E. Xie, H. Zhang, Y. Zhu, and Y. Qiao, “MaskFormer: Masked Image Modeling for Visual Tasks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. [Online].

[13] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755. [Online].

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580–587. [Online].

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. [Online].

[16] 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].

[17] K. Muralidharan, F., A., and A., “Learning Features for Offline Handwritten Signature Verification using Deep Convolutional Neural Networks,” in *Proceedings of the 2018 International Conference on Computing, Analytics and Networking (ICCAN)*, 2018, pp. 135–139.

[18] 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.

[19] 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.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[21] A. Bhatawdekar, S. Bhattacharya, R. Khatri, and R. Tiwari, “BHSig260: A Dataset for Offline Signature Verification,” *arXiv preprint arXiv:2004.07563*, 2020, [Online].

[22] 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.