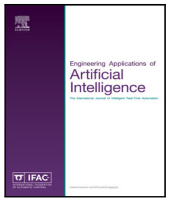
Engineering Applications of Artificial Intelligence 118 (2023) 105639

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



2C2S: A two-channel and two-stream transformer based framework for offline signature verification

Jian-Xin Ren [a], Yu-Jie Xiong [a,*], Hongjian Zhan [b,c], Bo Huang [a]

[a] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China
[b] School of Communication & Electronic Engineering, East China Normal University, Shanghai 201620, China
[c] Chongqing Institute of East China Normal University, Chongqing 401120, China

ARTICLE INFO

Keywords:
Offline signature verification
Transformer
Signature pair
Attention mechanism

ABSTRACT

Recently, with the outstanding performance of the transformer in NLP, approaches that employ the transformer to address vision problem is becoming a research focus. However, transformer-based research rarely focuses on signature verification. To fill this gap, this paper proposes a two-channel and two-stream transformer approach (2C2S) to cope with the signature verification problem. 2C2S is composed of original and central streams. The original stream receives the original signature pair as input, and the central stream receives the signature pair generated by cropping the central at the original pair as input. In order to establish the associations among feature channels, a squeeze-and-excitation operation is applied between two standard Swin Transformer blocks. Moreover, an up-sampling enhancement module directly steers the model to focus on useful information. The verification accuracy of 2C2S on SUES-SiG and several publically available datasets: CEDAR, BHSig-B, and BHSig-H, reaches 93.25%, 90.68%, 100%, and 72.22%, respectively. Extensive experiments illustrate that the proposed framework is competitive with the existing techniques for offline handwritten signature verification.

1. Introduction

Signature verification is an essential application in the field of computer vision. It is a commonly used biometric verification approach like face recognition, iris recognition, and fingerprint recognition. As it is stable, efficient, and convenient, signature verification is widely applied in administration, financial, and criminal forensic verification. Generally, the signature verification process involves two phases: registration and verification (Hafemann et al., 2016). In the registration phase, the reference library is constructed from the features of the signature sample provided by the user. In the verification phase, the system judges the identity claims for realism by comparison between reference and query signatures. Moreover, according to the data collection process, a signature verification system is classified into two categories: online (dynamic) and offline (static) (Hafemann et al. (2017b), Javidi and Jampour (2020)). The dynamic information containing abundant textural details, such as writing speed, angle, pressure, and trajectories, can be extracted from dedicated devices for an online signature verification system (Fahmy, 2010). In contrast, the offline signature samples are collected by scanning or photographing a document containing the signatures, which contributes to a missing dynamic feature. As a result, an online signature verification system is easier to achieve higher verification accuracy than offline (Hafemann et al., 2017a). However, online signature verification heavily relies on dedicated devices in the practical application, while offline has less restrictive conditions and a wide range of application scenarios. Consequently, online signature verification is not commonly used in practice. In this case, it has profound significance and application value for information security in establishing an efficient and accurate offline signature verification system (Xiong et al., 2017). This paper focuses on offline signature verification.

This paper proposes a novel two-channel and two-stream transformer-based framework (2C2S) for offline handwritten signature verification. The approach takes the modified Swin Transformer as the backbone for feature extraction, which utilizes the attention mechanism to focus on the slight stroke difference between genuine and forged. Unlike most signature verification existing methods, reference signature and query signature are two-channel of a signature pair as the input in concatenation to 2C2S. Then, the multi-scale signature features are extracted, and the relation between channels is captured after a modified Swin Transformer. Subsequently, these features are fed into the up-sampling enhancement module to help the network focus on useful information. Finally, the variants of the original input are fed into the extraction module again, and the final fused feature is distinguished as genuine or forged by a fusion and decision module. Experimental results show the superiority of the proposed approach. The main contributions of this paper are as follows:

* Corresponding author.
E-mail address: xiong@sues.edu.cn (Y. Xiong).

https://doi.org/10.1016/j.engappai.2022.105639
Received 29 July 2022; Received in revised form 5 November 2022; Accepted 10 November 2022
Available online 28 November 2022

Table 1
Advantages and limitations of different signature verification methods.

Categories	Methods	Advantages	Limitations
Traditional	Aravinda et al. (2019)	It learned global histogram features for detecting shape and texture information and is accurate for verification.	The large vectors lead to high storage costs, and the expanded descriptor enlarges the feature extraction time.
	Thakare and Deshmukh (2018)	The feature descriptor is complete and invariant to scaling, rotation, and translation.	It needs more computational time for low-power devices.
	Goon and Eng (2021)	It has stable features and high operation speed.	It needs to distinguish interest points and does not contain spatial information.
2-Channel	Li et al. (2019)	It mitigated the problem of overfitting and greatly reduces parameter space to search.	The feature extraction ability is not strong enough.
	Alvarez et al. (2016)	It employed VGG-16 as a feature extractor for signature verification, which has a simple structure and decent predictive power.	The simple structure results in the loss of locality information.
	Berkay Yilmaz and Ozturk (2018)	It was a hybrid offline signature verification method based on CNN, which adopts a low-dimension global average pooling layer to learn a distance metric.	The significant association of the reference and query signatures may be lost with the feature extraction processing.
Siamese	Xiong and Cheng (2021)	It utilized the attention module to enhance the feature extraction ability of the network. The designed multiple branches structure and contrastive pairs provide more features.	The multi-scale features are not considered, and the ability of attention is limited in structure.
	Dey et al. (2017b)	It adopted the siamese network to model a feature space, in which similar observations are placed.	It is not sufficiently deep to explore complex features, which leads to decreased precision.
	Ghosh (2021)	It used two different RNN models (LSTM and BLSTM) under the siamese framework to learn feature representations.	It is difficult to train two different models, and spends more time training a model for each writer.

1. We propose a two-channel and two-stream framework (2C2S) that is designed for offline handwritten signature verification. It is based on the transformer structure and utilizes the attention mechanism to capture stroke information efficiently.
2. We introduce a squeeze-and-excitation (SE) operation between two standard Swin Transformer blocks, which helps the framework to establish the links among feature channels.
3. We propose an up-sampling enhancement module to steer the model to focus on important features and ignore redundant information.

2. Related work

This section continues as follows: Section 2.1 analyzes various threats to signature verification. Previous related work will be discussed in Section 2.2. These signature verification methods are divided into two classes: traditional and CNN, and CNN methods are further subdivided by input. The advantages and limitations of these methods are presented in Table 1 in detail. Section 2.3 introduces several classic vision transformer and their application.

2.1. Threat to signature verification

The signature verification aims to distinguish between genuine and forged signatures to protect user assets and privates from unauthorized use. In modern societies, numerous public and private documents are required to sign by the participants to generate a legally binding effect. As a result, the signatures are able as admissible evidence in law. Forged signatures are considered critical threats to signature verification. Pummelled by huge benefits, the events of signature forgeries have emerged in an endless stream. Some skilled forgeries even can produce signatures close to the genuine through repeated practice. Currently, forgery has caused significant financial damage and posed privacy threats to society (Shaukat et al., 2020b). According to the degree of imitation, forgeries are typically divided into three categories (Hafemann et al., 2017b): random forgery, simple forgery, and skilled forgery. Random forgery signature means that the forger performs it without knowledge of any original author information. Simple forgery refers to a forger who only knows the name of the authentic author but does not train to increase proficiency (Zhou et al., 2020). Skilled forgery indicates that the forger imitates the signatures under fully grasping the signature information and characteristics of the authentic author, which is the greatest threat to signature verification. It is relatively easy to distinguish the random and simple forgery owing to the vast gap between the writing contents and characteristics compared with the original author. However, the high degree of similarity between skilled forgery and genuine signatures may be challenging for offline signature verification because it is not easy to distinguish the subtle strokes between these two signatures. Fig. 1 shows examples of signatures from four used datasets: CEDAR, BHSig-H, BHSig-B, and SUES-SiG. Each row contains two genuine signatures of the same writer and four skilled forgeries. Lines 1 to line 4 shows the samples of SUES-SiG, BHSig-B, BHSig-H, and CEDAR, respectively. It is noticed that skilled forgeries are very similar to genuine signatures.

2.2. Signature verification

For offline handwritten signature verification, the final accuracy is dependent critically on the performance of feature extraction algorithms. Conventional machine learning-based algorithms including Histogram of Oriented Gradient (HOG) (Aravinda et al., 2019), Scale-Invariant Feature Transform (SIFT) (Thakare and Deshmukh, 2018), and Speeded-Up Robust Features (SURF) (Goon and Eng, 2021) have strict restrictions on the format and content of input signature samples and rely heavily on hand-crafted features to guide the system to complete the verification task. Traditional signature verification techniques are generally subjected to a complete verification process that involves preprocessing, feature extraction, similarity measure, and verification. However, except preprocessing is valid indeed, others do not guarantee desired verification performance due to the dependence on manual feature engineering (Shaukat et al., 2022). This is primarily caused for the complicated handwriting feature being challenging to be characterized. Thus, although researchers have made tremendous efforts to develop traditional approaches, there is little improvement obtained in performance (Alam et al., 2021).

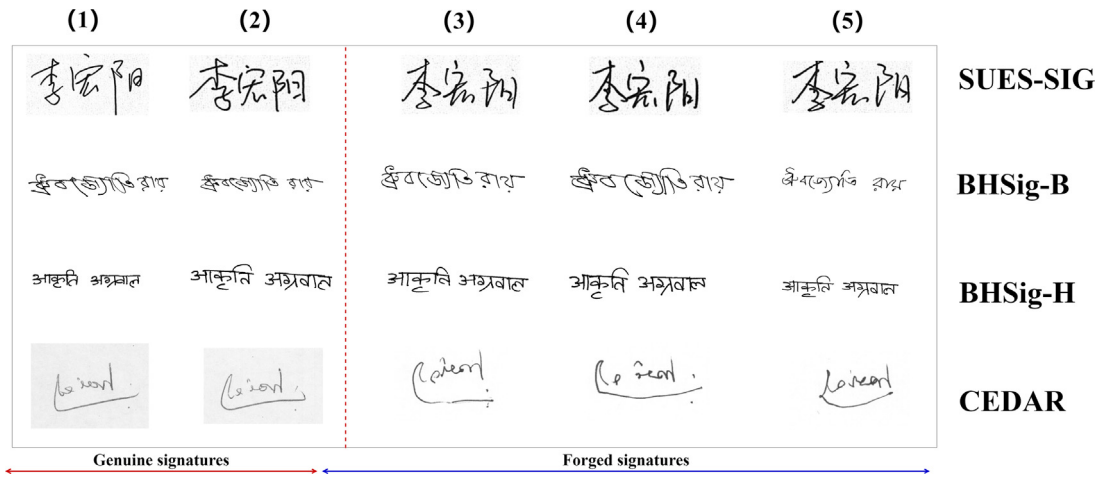


Fig. 1. Samples of SUES-SiG, BHSig-B, BHSig-H and CEDAR. Column (1) and (2) are genuine signatures and column (3), (4), and (5) are forged signatures.

Recent work attempts to learn complex handwriting featuring using Convolutional Neural Network (CNN) and achieved inspired results. Compared with traditional signature verification techniques that rely heavily on manual feature engineering, CNN is not only automatically extracts features but also more transferable and effective. Currently, the CNN methods can be divided into two classes in terms of different inputs: two-channel and siamese. Two-channel models are straightforward and efficient, but they generally lack concerns for feature channels consisting of reference and query signatures. Li et al. (2019) proposed a two-channel CNN for offline signature verification, which employs Inception-V3 architecture as the feature extractor. The network takes the signature pair as input and calculates the similarity between the signatures by the final two logits. Two dropout layers are inserted into the network to mitigate the problem of overfitting. Alvarez et al. (2016) proposed two-channel method that is a VGG-16 based structure for signature verification. Berkay Yilmaz and Ozturk (2018) proposed a two-channel hybrid offline signature verification method, which employed a low-dimension global average pooling layer to learn a distance metric. Siamese is the most prevalent approach for signature verification. However, the interaction information between reference and query signatures is not considered during the feature extraction. Xiong and Cheng (2021) proposed a multiple siamese network to cope with signature verification, which employed an effective attention module to improve performance. The network has four weight-shared branches and receives contrastive pairs as input. The out features are finally formed into four feature pairs and are classified by a voting mechanism. Dey et al. (2017b) proposed convolutional siamese network, named SigNet, for signature verification. The network is trained for learning a feature space to minimize the Euclidean distance. Ghosh (2021) proposed a siamese network for signature verification, in which one stream is the Long Short-Term Memory (LSTM) model and the other is the Bi-directional Long Short-Term Memory (BLSTM). Using two different RNN models, it is easier to learn feature representations. The specific comparisons are listed in Table 1.

Compared with the above, the proposed method has several strengths. First, it utilizes the popular transformer structure as the feature extractor, which adopts global relationship modelling to enlarge the receptive fields and achieve rich feature information. Second, a two-stream architecture is employed to fuse the multi-scale signature information, which is demonstrated as an effective way to enhance the verification performance in the Experimental section. Third, a squeeze-and-excitation (SE) is applied to equip the model with the ability to establish the links among feature channels. Fourth, an up-sampling enhancement module is proposed to guide the model to focus on important features.

2.3. Vision transformer

The transformer is a self-attention-based architecture with encoders and decoders, first proposed by Vaswani et al. (2017). In natural language processing (NLP), the transformer has emerged as a mainstream tool owing to its outstanding performance on sequence modelling and machine translation tasks. Taking inspiration from the great success of transformers in NLP, some researchers combine attention mechanisms with convolutional neural networks to enhance signature verification (Ahrabian and BabaAli, 2019; Xiong and Cheng, 2021). Even though (Cordonnier et al., 2019) demonstrated in theoretical terms that self-attention blocks could be effective and efficient, these models have tended to perform less than the current state-of-the-art (SOTA) CNN models. Recent, many works attempt to directly employ a standard transformer in vision tasks under the fewest possible modifications. Dosovitskiy et al. (2021) proposed a pioneering transformer backbone and achieved the impressive speed-accuracy trade-off on object recognition. ViT is a pure transformer structure that performed a great performance on many public datasets by splitting an image into 2D patches as input. Like BERT, the disadvantage of ViT lies in the fact that it requires pre-training with a large-scale dataset (JFT-300M comprised of 300 million images) and expensive computational costs. In light of this, Touvron et al. (2021a) present a Data-efficient Image Transformer (DeiT) to improve its suitability for training on ImageNet-1k (Deng et al., 2009). Liu et al. (2021) proposed a general-purpose transformer backbone to overcome high computational complexity problems, named Swin Transformer. It greatly reduces the computational cost with the specific mechanism of partition windows and achieved state-of-the-art results on object recognition, object detection and semantic segmentation. Besides, some transformer-based models such as Wang et al. (2021), Han et al. (2021), Xu et al. (2021), and Touvron et al. (2021b) achieve excellent results in vision tasks. Given the success of the transformer structure, it has emerged as a hot spot in computer vision.

3. Method

This section provides an architectural overview of the proposed 2-Channel and 2-Stream (2C2S) Transformer network in detail. The overall architecture of the proposed 2C2S is presented in Fig. 2. As the name implies, it consists of two streams, original and central, to facilitate processing in the spatial domain at multiple-scale. The signature pairs are the two-channel input of two streams, with the first channel always representing a reference signature and the second representing a query signature. More specifically, the original stream

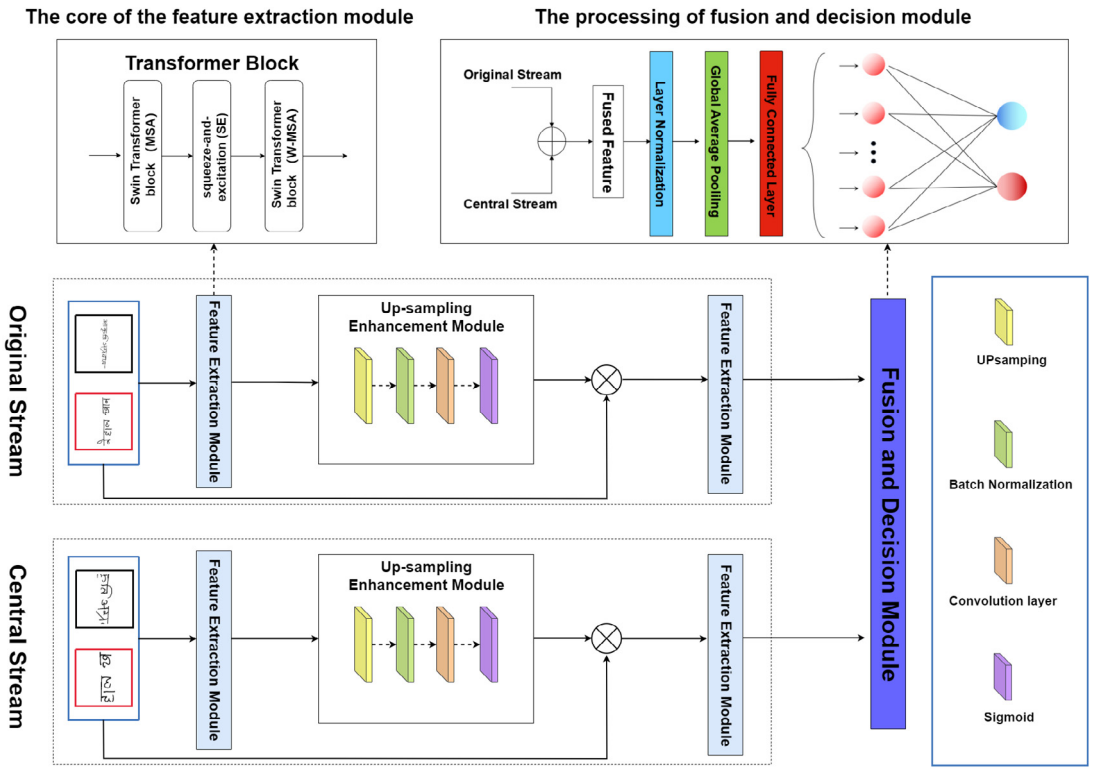


Fig. 2. The Architecture of the Proposed Method.

Algorithm 1: The pseudocode of the proposed framework.

Input: input signature pair $D = \{(R^{(n)}, Q^{(n)}\}_{n=1}^{N})$

for n=1...N **do**
  A random sample $(R^{(n)}, Q^{(n)})$ is selected;
  # **Step 1: transform the samples**;
  Gain cropped signatures $(r^{(n)}, q^{(n)})$ by centering crop the original signatures and resizing them to the original size.;
  $O \leftarrow$contact$[R^{(n)}, Q^{(n)}]$,# Concatenate the original signatures;
  $C \leftarrow$contact$[r^{(n)}, q^{(n)}]$,# Concatenate the cropped signatures;
  # **Step 2: input signature pairs to the feature extraction module(FEM)**;
  $O_p \leftarrow$FEM(O),#Original stream ;
  $C_p \leftarrow$FEM(C),#Central stream;
  # **Step 3: input features to the up-sampling enhancement module (UE)**;
  $O_u \leftarrow \mathrm{UE}(O_p)$;
  $C_u \leftarrow \mathrm{UE}(C_p)$;
  # **Step 4: input features to the feature extraction module (FEM)**;
  $O_s \leftarrow \mathrm{FEM}(O_u)$;
  $C_s \leftarrow \mathrm{FEM}(C_u)$;
  # **Step 5: the output of two streams is fed to the fuse and decision module (FD)**;
  output $\leftarrow \mathrm{FD}(O_s, C_s)$;
end

uses a two-channel signature pair (224 × 224 × 2) comprised of the reference and query signatures as input. For the central stream, it receives a signature pair that is produced by cropping the central 112 x 112 part of the original signatures as input. One reason for utilizing the two-stream architecture is that multi-scale image information is a key factor affecting the performance, which may be helpful for signature verification. Furthermore, the model is forced to place more attention on the pixels closer to the center of the input by considering the central part twice.

The overall structure of 2C2S comprises two streams, central and original, respectively. The two streams are identical except for the input. Take the original stream as a detailed example. It mainly consists of the feature extraction module and up-sampling enhancement module (UE). The input pair is first fed into the feature extraction module to generate hierarchical feature representations and establish the links among feature channels. Then, an up-sampling enhancement module is applied to generate the importance scores of the features, which helps the model more selectively focus on useful information. Subsequently, the retuned original feature map is fed into the feature extraction module for learning new feature representations. It should be noted that the Transformer block is the core of the feature extraction module, where a squeeze-and-excitation (SE) operation is applied between two standard Swin-Transformer blocks. More details will be given in the 3.1.1 Transformer Block section. The central stream is performed exactly as the original stream, and both outputs are fed into the fusion and decision module to achieve signature verification eventually. The details of the 2C2S algorithm and implementation are presented in Algorithm 1.

3.1. Extraction module

This paper adopts the modified Swin Transformer as the extraction module, where the squeeze-and-excitation (SE) is applied between the window-based multi-head self-attention (W-MSA) module and the shifted window-based multi-head self-attention (SW-MSA) module. The two modules are standard Swin Transformer blocks (Liu et al., 2021). There are marked differences among channels due to the original pair channels are composed of reference and query signatures. As a result, it is essential to employ SE to capture the interaction information between channels to increase the proportion of features in the relatively important channel. As shown in Fig. 3, the extraction module is composed of four stages.

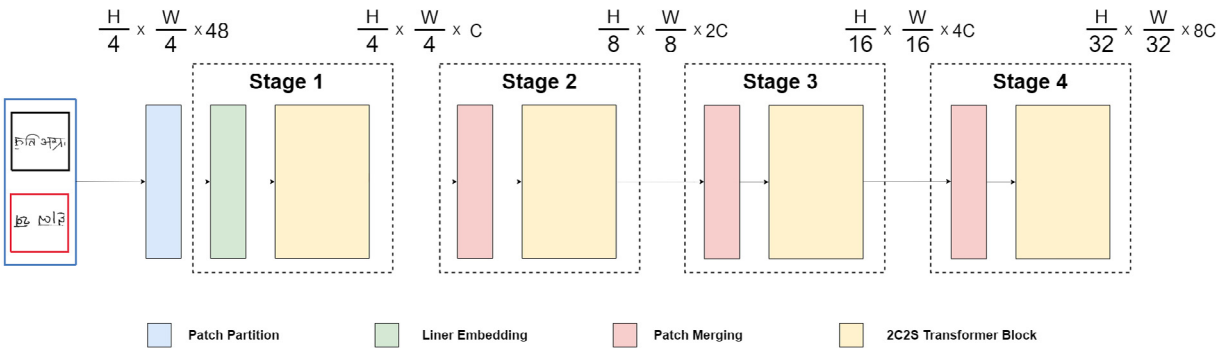


Fig. 3. The Process of the Extraction Module.

The inputs are divided into a series of patches by a patch partition. The primary role of patch partition is spatial reduction and channel expansion. In detail, assuming that the input pairs have a size of $H \times W \times 2$, its channel dimension is first extended to 48, and spatial is reduced to $H/4 \times W/4$ after the patch partition. After that, these patches are fed into Stage 1 for feature extraction. Stage 1 is composed of the linear embedding and transformer block. The effect of the linear embedding matter is to project the feature dimensions of input into arbitrary dimensions (represented by C). The prime role of the Transformer block is to learn multi-scale feature representation and establish the associations among channels. Specifically, the input patches are first projected feature dimensions into dimensions C by a linear embedding layer. Then, patch tokens are implemented with the transformer block for learning feature representation. Following a hierarchical design, the output resolution of the four stages gradually decreases from 4 to 32-stride. Stage $2 \sim 4$ are similar to stage 1 except for replacing linear embedding with patch merging. Patch merging is similar to the focus operation in Yolo (Zhu et al., 2021), which is used for down-sampling and extending dimensions while keeping the information intact. Specifically, the input patches are reduced to $2\times$, and the feature dimensions are increased to $2\times$.

3.1.1. Transformer block

The transformer block is the key component in the extraction module. As shown in Fig. 4, the whole process of the transformer block is divided into three steps. Step 1 and step 3 are standard Swin Transformer blocks, consisting of layer normalization (LN) layer, residual connection and 2-layer MLP, except for the window-based multi-head self-attention (W-MSA) module and the shifted window-based multi-head self-attention (SW-MSA). Moreover, since SW-MSA is performed on the basis of W-MSA, W-MSA and WS-MSA are present alternately, and W-MSA is always anterior to WS-MSA. Step 2 is a squeeze-and-excitation (SE) that aims to help the model capture the interaction information between channels. These relationships can be represented with the following equation:

$$\hat{F}_1 = W - MSA(LN(F)) + F \tag{1}$$

$$F_2 = SE(MLP(LN(\hat{F}_1)) + \hat{F}_1) \tag{2}$$

$$\hat{F}_3 = SW - MSA(LN(F_2)) + F_2 \tag{3}$$

$$F_3 = MLP(LN(\hat{F}_3)) + \hat{F}_3 \tag{4}$$

Where F represents input features, $\hat{F}_1$ represents the output features of the W-MSA module of the $step1$, SE represents the squeeze-and-excitation, LN represents layer normalization, and MLP represents the 2-layer perceptron.

3.1.2. Squeeze-and-excitation

The squeeze-and-excitation (SE) is first proposed by Hu et al. (2020) for object classification. It relies on aligning channel-wise feature maps to learn channel attention. As shown in Fig. 5, the shape (H, W, C) of the feature map is first squeezed to $(1, 1, C)$ by a global pooling. This processing can be expressed with the following equation:

$$Z_C = F_{sq}(f_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} f_c(i, j) \tag{5}$$

Where f_c denotes the 2-dimensional matrix of the input feature map f, c denotes the channel of f. H and W denote height and width of the f, respectively. i and j denote the ordinate and the abscissa of the f, respectively.

Then the excitation operation is applied to model the correlation between the fused feature channels for learning the weight of each channel, which consists of two fully-connected layers, a ReLu non-linearity layer, and a sigmoid activation function. Finally, a scale operation is utilized to apply these normalized weights to the original feature maps to generate the final output. This processing can be expressed with the following equation:

$$F_{ex}(Z, W) = Sigmoid(W_2 \times ReLU(W_1 \times Z)) \tag{6}$$

Where W_1 and W_2 represent the weight matrix for the fully connected layer. The r represents the scaling parameter that is used to reduce the number and complexity of the model.

The up-sampling enhancement module (UE) is designed to pay more attention to important information and ignore irrelevant ones. As shown in Fig. 2, UE consists of an up-sampling, conventional layer, batch normalization, and sigmoid activation function. UE receives the output of the extraction module as an input. For convenient computation, the output is first transformed from the two-dimensional matrix $(8C, H \times W)$ into the three-dimensional matrix $(8C, H, W)$. Then, deconvolution layers with the nearest neighbor up-sampling are applied to recover the spatial size of feature maps back to the original input. Afterward, a 1×1 convolution is employed to reduce dimensionality to the original size. In order to prevent the gradients from vanishing in the deep network, batch normalization is added prior to the sigmoid activation function. Eventually, each importance score is multiplied by the respective original feature map, where the higher the importance score, the greater the significance of the feature in the respective original input. Based on the above process, the formulation can be described as follows:

$$UE = S(BN(Conv(UP(X)))) \tag{7}$$

Where UP represents nearest neighbor up-sampling, $Conv$ represents a 1×1 convolution, BN represents batch normalization, and S represents the sigmoid activation function.

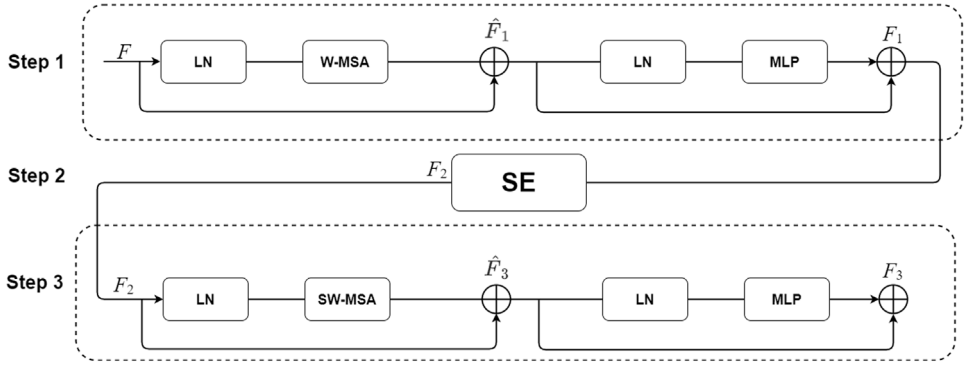


Fig. 4. The Inner Structure of Transformer Block.

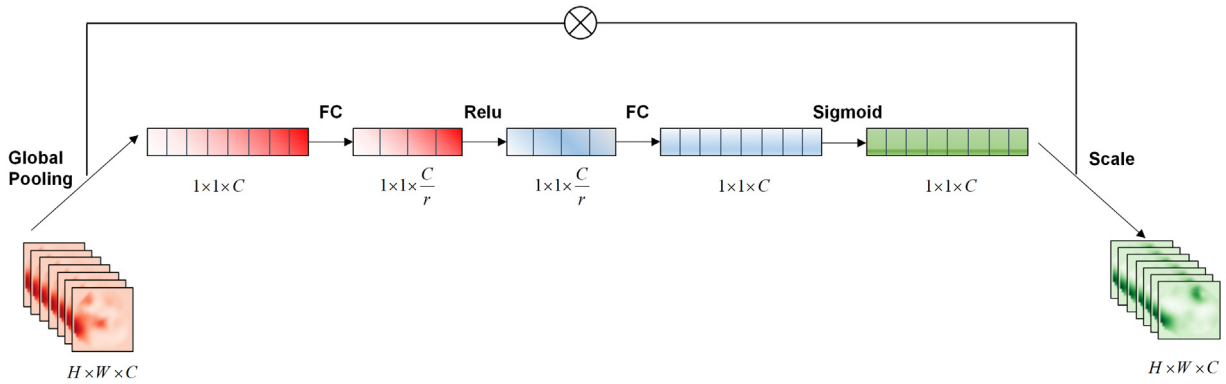


Fig. 5. Illustrations of the Squeeze-and-Exception (SE).

3.2. Fusion and decision module

The fusion and decision module consists of layer normalization, global average pooling, and a fully connected layer, which is utilized to distinguish genuine signatures and forgeries by transferring the fused features into a one-dimensional vector. As shown in Fig. 2, the fused feature is first fed into a layer normalization (LN) for the stability of the feature distribution, where the fused feature is produced by summing the output of the original stream and central. Then, a global average pooling (GAP) is introduced to reduce the redundancy of the network. Finally, a fully connected layer is employed to fuse all multi-scale features. Moreover, the cross-entropy loss is utilized as the loss function to measure the difference between the reference signature and the test. The loss function can be described as follows:

$$Loss(x, class) = -x[class] + \log(\sum_i exp(x[i])) \tag{8}$$

Where the x is the output vector of the final output layer of the network, the $class$ is the labels of the input signature pair, which ranges from $[0, 1]$.

4. Experiments

This paper train the model based on the framework of Pytorch (1. 4. 0) with NVIDIA-3070 and i7-9700k CPU. The initial learning rate is set to 0.001 and the learning rate decay is set to 0.95. The batch size is set to 4 owing to the GPU memory limitation.

4.1. Datasets and experimental protocol

To demonstrate the effectiveness of the proposed 2C2S approach, a series of experiments are conducted on four signature datasets: CEDAR, BHSig-B, BHSig-H and SUES-SiG. The CEDAR is a widely known English signature dataset, including 24 genuine and 24 forged signatures from 55 authors. All signature images are acquired at 300 dpi in gray-scale format and stored as PNG images. Moreover, the dataset provides a degree of background noise and gives an apparent tilt in some signature images. Following the previous work, 50 authors are employed to train the model, and 5 authors are used to test the model. The BHSig260 is a large Indic script signature dataset and contains the signatures of 260 individuals from BHSig-B and BHSig-H dataset, respectively. The BHSig-B is a Bengali dataset, which comprises 24 genuine and 30 skilled forgery signatures from 160 writers. The BHSig-H is a Hindi dataset that consists of 24 genuine and 30 skilled forgery signatures provided by 100 writers. Both datasets are acquired in gray-scale format with 300 dpi resolution and stored as TIFF images. Notably, all signatures are collected from writers of different educational backgrounds and ages to mimic realistic scenarios. In order to make a comparison with the existing works, 60 authors in Hindi are used to train, and the remaining is used to test. Furthermore, 50 authors in Bengali are used to train, and the remaining is used to test. SUES-SiG is a small-scale and challenging Chinese signature dataset. The dataset includes 9 genuine signatures and 9 skilled forged signatures from 20 writers, where all signature images are acquired at 300 dpi in gray-scale format and stored as PNG images.

All signatures in datasets are established pairwise because the inputs to the proposed approach are pairs of reference and query signatures. In this case, the training and testing datasets are composed of genuine-genuine (positive) and genuine-forged (negative) signature pairs. Taking the Bengali datasets as an example, each writer has $C_{24}^2 = 276$ positive and $24 \times 30 = 720$ negative samples. Moreover, in order to avoid sampling bias, the same number of negative samples as the positive samples are randomly selected (276/720). The specific division is presented as Table 2.

4.2. Evaluation metrics

The performance of the proposed method is assessed by four indices: false acceptance rate (FAR), false rejection rate (FRR), Equal Error Rate

Table 2
Details of experimental protocol on different datasets.

Dataset	Train	Test	Positive pairs	Negative pairs
CEDAR	50	5	276	276
Bengali	50	50	276	276
Hindi	100	60	276	276
SUES-SiG	15	5	36	36

Note: Columns 2 and 3 represent the number of writers used for training and testing, respectively. Columns 4 and 5 represent the number of positive and negative pairs trained per writer, respectively.

Table 3
Comparison of the proposed method with the state-of-the-art methods on four signature databases (%).

Datasets	Methods	FAR	FRR	ACC	EER
CEDAR	2-Channel-2-Logit (Li et al., 2019)	–	–	**100.00**	**0**
	MobileNetV2 (Sharvari, 2021)	–	–	96.00	–
	MSN (Xiong and Cheng, 2021)	3.18	**0**	98.40	1.63
	SigNet (Hafemann et al., 2017a)	**0**	**0**	**100.00**	**0**
	Bhunia et al. (2019)	5.01	6.12	–	–
	Sharif et al. (2020)	4.67	4.67	–	–
	2C2S (ours)	**0**	**0**	**100.00**	**0**
Bengali	2-Channel-2-Logit (Li et al., 2019)	10.44	9.37	88.08	11.92
	MobileNetV2 (Sharvari, 2021)	–	–	85.63	–
	MSN (Xiong and Cheng, 2021)	10.42	6.44	91.56	8.43
	SigNet (Hafemann et al., 2017a)	13.89	13.89	86.11	13.89
	SURDS (Chattopadhyay et al., 2022)	19.89	**5.42**	87.34	–
	Jadhav and Chavan (2018)	–	–	90.36	–
	LBP and ULBP (Pal et al., 2016)	33.82	33.82	66.18	33.82
	2C2S (ours)	**5.37**	8.11	**93.25**	**6.75**
Hindi	2-Channel-2-Logit (Li et al., 2019)	–	–	86.66	13.34
	MobileNetV2 (Sharvari, 2021)	–	–	75.00	–
	MSN (Xiong and Cheng, 2021)	17.06	**5.16**	88.88	11.31
	SigNet (Hafemann et al., 2017a)	15.36	15.36	84.64	15.36
	SURDS (Chattopadhyay et al., 2022)	12.01	8.98	89.50	–
	Pal et al. (2016))	24.47	24.47	75.53	24.47
	2C2S (ours)	**8.66**	9.98	**90.68**	**9.32**
SUES-SiG	ResNet50 (He et al., 2016)	33.89	30.56	67.78	32.23
	MSN (Xiong and Cheng, 2021)	27.22	36.11	68.33	31.67
	ViT (Dosovitskiy et al., 2021)	39.44	35.56	65.00	37.50
	2C2S (ours)	**27.22**	**28.33**	**72.22**	**27.78**

(EER), and Accuracy (ACC). FAR reflected the ratio of the number of false acceptances divided to all forged samples, and FRR reflected the ratio of the number of false rejections to all genuine samples. The EER is a common metric used to assess the performance of the biometric system, and it is employed to evaluate the equilibrium point where FRR equals FAR. Accuracy is the ratio of a number of correctly predicted samples to all predicted samples. Sepcific calculations are presented in the following equation:

$$FAR = \frac{FP}{TN + FP} \tag{9}$$

$$FRR = \frac{FN}{TP + FN} \tag{10}$$

$$ACC = \frac{TP + TN}{TN + FN + TP + FP} \tag{11}$$

Where TP (True Positive) denotes the number of genuine signatures predicted as genuine, TN (True Negative) denotes the number of forged signatures predicted as forgeries, FN (False Negative) denotes the number of genuine signatures predicted as forgeries, and FP (False Positive) denotes the number of forged signatures predicted as genuine signatures.

4.3. Comparisons to the state-of-the-art

We compare the proposed 2C2S against state-of-the-art methods on four datasets. As shown in Table 3, the performance characteristics of each method are presented. It can be seen that the proposed 2C2S significantly outperforms the other method on all datasets except CEDAR, where several solutions and the proposed method show identical performance.

Specifically, the listed methods all achieve excellent performance on the CEDAR dataset, where three approaches even obtain the accuracy of 100%. Two aspects are largely responsible for the higher performance. One reason is that the dataset is relatively simple, and the other is that most samples are used for training. For the Bengali dataset, our highest accuracy is 93.25%, which is 1.69% higher than the second-best MSN. Moreover, the SURDS achieves the lowest FRR (5.42%), but the higher FAR (19.89%) substantially affected the final performance. Compared to the listed methods, the proposed 2C2S obviously exceeds the other methods tested on the Bengali dataset. For the Hindi dataset, the proposed method provides an accuracy of 90.68%, which is the only method listed above accuracy of 90%. This indicates that the proposed 2C2S achieves better verification performance. Notably, compared with the same type of 2-channel network i.e., Li et al. (2019), the proposed method consistently outperforms it in terms of performance, including FRR, FAR, EER, and ACC on three public datasets, which confirms the superiority of the proposed approach. Moreover, a comparative experiment is conducted to evaluate the model performance on the SUES-SiG dataset. It can be seen that our method also works for complex Chinese signatures, which indicates the general suitability of our approach. Based on the results, the proposed 2C2S method outperforms all the compared methods and can effectively distinguish genuine and forged samples.

4.4. Cross datasets validation

For analyzing the generalization ability of the proposed 2C2S, a cross-language evaluation is performed on the four independent datasets. Generally, the stronger the cross-dataset performance is, the better model generalization capabilities are. As shown in Table 4,

Table 4
Cross datasets validation results (%).

Dataset	CEDAR	Bengali	Hindi	SUES-SiG
CEDAR	**100.00**	50.00	50.00	51.04
Bengali	50.80	**93.25**	74.41	52.17
Hindi	61.92	79.69	**90.68**	54.43
SUES-SiG	50.00	51.42	50.88	**72.22**

Table 5
Analysis of Two-stream structure for 2C2S (%).

Model	ACC	FAR	FRR	ERR
Single-stream	89.93	8.89	11.24	10.07
2S (OR)	91.44	**6.09**	11.54	8.55
2S (OC)	**92.04**	6.11	**9.81**	**7.96**

Note: The 2S represents a Two-stream structure. The OR represents the input of 2S consists of original and random-cropped on original signatures. The OC represents the input of 2S comprise of original and center-cropped on original signatures.

Table 6
Analysis of up-sampling enhancement module for 2C2S (%).

Model	ACC	FAR	FRR	ERR
Single-stream	89.93	8.89	11.24	10.07
Single-stream+UE	90.67	8.37	10.27	9.32
Two-stream	92.04	6.11	9.81	7.96
Two-stream+UE	**92.38**	**5.59**	**9.65**	**7.62**

Note: UE represents the up-sampling enhancement module.

the cross-language of generalization performance is presented, where the rows and columns of the table correspond to the training and testing datasets, respectively. It is evident that the best performance is achieved when the model is trained and tested on the same language datasets. However, a significant decrease in performance accuracy is observed when performed on cross-language datasets. The decreased performance is mainly attributed to the difference in languages and writing styles, which is especially evident on the CEDAR dataset. A similar result has been occurred in Wei et al. (2019) and Dey et al. (2017a). Remarkably, the tests on Bengali and Hindi datasets achieve a decent across-language performance owing to both being Indic script signature datasets.

4.5. Ablation experiments

In this section, several ablation experiments are performed on Bengali dataset to gain further insights into the contribution of each module in the proposed 2C2S. For fair comparisons, a single-stream network based Swin-Transformer is also trained as the baseline performance.

The performance of the two-stream structures is first assessed by comparing them with the baseline. For OC and OR, OC represents the original and central streams, and OR represents the original and random streams. The original stream receives the original signature pair as input. The central and random streams receive the signature pairs as input generated by cropping (at the original signatures) the central and random parts, respectively. The baseline is designed as a single-stream network, which receives a two-channel signature pair consisting of a reference signature and query signature. It is evident from Table 5 that both two-stream models have unequivocal advantages in terms of performance compared to the baseline. Generally speaking, multi-scale input has a distinct advantage in conveying feature information compared with single-scale, which provides more informative feature maps to enhance the verification performance. Moreover, the performance of the OR is inferior to the OC. The primary reason for the result is that the feature information is largely distributed on the central pixel of signature images. The results indicate that the model performance of the two-stream architecture is superior compared with a single-stream architecture. As a result, the two-stream approach can learn more effective feature representation from the multi-scale inputs.

The up-sampling enhancement module (UE) is essential in directly steering the model to focus on useful information. To validate the effectiveness of the UE, the performance of the single-stream and two-stream (original and central) models with and without the UE module is compared on the Bengali dataset. The detailed results are listed in Table 6. As can be clearly seen, the accuracy is improved from 89.93% to 90.67% and 92.04% to 92.38% in the single-stream and two-stream models with UE module, respectively. For the single-stream model, the added UE module shows a slight increase in accuracy (0.74%). In contrast, the introduction of UE for the two-stream model remain has an effect but is weaker (0.34%). These results indicate that a slight performance improvement can be achieved by using UE.

TB (Transformer block) is a critical component of the extraction module. Building upon Swin Transformer Block, a squeeze-and-excitation (SE) is introduced, which aims to capture the interaction between feature channels. The performance of the single-stream and two-stream models with and without TB on the Bengali dataset is given in Table 7. As can be observed from the results, the accuracy of the single-stream model using TB is significantly improved from 89.93% to 91.88%, which indicates that TB has a distinctly positive effect in the single-stream model. However, the performance gains in the two-stream model with TB, the boost is not apparent. The main reason for this discrepancy is that the information presented by the multi-scale input of the two-stream model has some degree of overlap with TB provides. Altogether, these results show that TB is effective.

In the following experiments, the up-sampling enhancement module (UE) and transformer block (TB) are employed as a single or combined to search for an optimal module combination on the single-stream and two-stream models (Tian et al., 2021). Table 8 presents the performance of different method combinations on the Bengali dataset. It can be seen that both single-stream and two-stream models can gain further promotion in performance by utilizing the two components. The two-stream model with TB get the lowest FAR, while it has few advantages in other respects. Of all combinations, UE+TB leds to the optimum performance improvements. Specifically, UE+TB is combined to produce a minor impact in the single-stream model, where only a slight accuracy improvement (0.09%) over TB is used alone. In contrast, the combined UE and TB can be boosted mutually in a two-stream model, where both provide better coupling than employed alone.

Table 7
Analysis of Trnasformer block for 2C2S (%).

Model	ACC	FAR	FRR	ERR
Single-stream	89.93	8.89	11.24	10.07
Single-stream + TB	91.79	5.67	10.83	8.21
Two-stream	92.04	6.11	**9.81**	7.96
Two-stream + TB	**92.43**	**4.88**	10.26	**7.57**

Note: TB represents the Trnasformer block.

Table 8
Verification performance of different combinations (%).

Model	ACC	FAR	FRR	ERR
Single-stream	89.93	8.89	11.24	10.07
Single-stream+UE	90.67	8.37	10.27	9.32
Single-stream + TB	91.79	5.67	10.83	8.21
Single-stream+UE+TB	91.88	5.48	10.76	8.12
Two-stream	92.04	6.11	9.81	7.96
Two-stream+UE	92.38	5.59	9.65	7.62
Two-stream + TB	92.43	**4.88**	10.26	7.57
2C2S (ours)	**93.24**	5.40	**8.15**	**6.75**

Note: TB represents Transformer Block, UE represents the up-sampling enhancement module.

Table 9
Results of different methods for feature fusion (%).

Model	ACC	FAR	FRR	ERR
2C2S (concat)	92.20	6.30	9.31	7.80
2C2S (sum)	**93.24**	**5.40**	**8.15**	**6.75**

Table 10
Comparing testing efficiency and accuracy on SUES-SiG for different methods.

Model	Param	FLOPs	Training speed (image/s)	Testing speed (image/s)	SUES-SiG ACC
2C2S (two-stream)	47.40M	7.91G	30	74	72.22
ViT-L/16 (single-stream)	307M	190.70G	11	54	65.00
ResNet50 (single-stream)	25.55M	3.81G	83	348	67.78

This study also explores two ways of fusing the two streams. The first strategy is concat, which concatenates both streams before the layer normalization. The second approach sums up the outputs of two streams and is referred to as the sum. As shown in Table 9, the second strategy (sum) outperforms the first strategy (concat).

In this experiment, several factors that can affect the efficiency of the proposed method are analyzed, such as trainable parameters, FLOPs, training, and testing speed. Table 10 shows the results of the comparison. ViT and ResNet50 adopt single-stream structures, and the proposed method is a two-stream structure. Normally, the processing speed of the single-stream model is almost two times faster than the two-stream model under the same conditions. Compared with the classic ViT model, the proposed method is far superior in every aspect. The parameters and FLOPs of ResNet50 are similar to the parameters in the proposed method, while it presents an advantage in both training and testing speed. At present, high computational complexity is the most common problem transformer structures are facing, and ours is no exception. Since CNN and transformer structures have different calculation methods, CNN usually provides an advantage over the transformer in the processing speed. Although some researchers have made great efforts to solve this problem, faster training speed is always achieved at the expense of accuracy. Beyond this, single-stream is also the main reason for affecting training and testing. In fact, accuracy is considered more important than learning efficiency and speed in signature verification for safety and privacy concerns. In this experiment, the proposed method achieves better accuracy than ResNet50 (4.44%). As a result, although ResNet50 has an efficiency advantage, the proposed method is more suitable for signature verification.

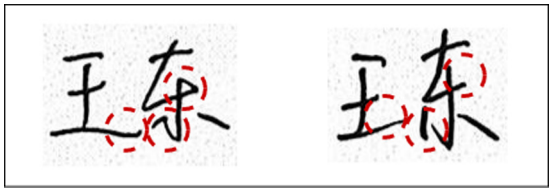


Fig. 6. Two signatures from the different writer, have been verified correctly by the proposed method.

4.6. Discussion

This paper attempts to employ the transformer structure to solve the signature verification problem. It applies an attention mechanism to capture subtle variations in strokes to distinguish genuine signatures and forgeries. Abundant experimental evidence shows the superior performance of the proposed framework over the compared methods. This section further analyzes the proposed method with several samples and discusses the time complexity of the proposed framework. A representative example correctly verified by the proposed approach is shown in Fig. 6. The two signatures are very similar in terms of strokes, shape, and orientation. Thus, it is difficult for other existing methods to verify the two signatures are from different writers. However, the style of the two signatures (red dashed circle) has minor differences. Nevertheless, the proposed method can easily distinguish these subtle differences by employing the strong attention mechanism.

Fig. 7 presents a pair of signatures verified correctly by the proposed approach. It can be seen that the right signature (red dashed circle) has an evident noise, and the second character in both signatures has

Fig. 7. Two signatures from the same writer, have been verified correctly by the proposed method.

some differences. For these reasons, existing methods misidentified the two signatures are from different writers. The proposed up-sampling enhancement module can steer the model to pay more attention to important information and ignore irrelevant ones, which is also why the proposed method distinguishes the signatures correctly.

However, although our method obtains a superior performance, there are also limitations. Here we focus on the time complexity of the proposed framework. The time complexity can be utilized to measure the time efficiency of the algorithm (Shaukat et al., 2020a). Since it is based on Swin Transformer, only additional time complexity will be discussed here. In this paper, the use of Squeeze-and-Excitation (SE) adds additional network structure, which increases the amount of time complexity. Specific time complexity is calculated as : $O(\frac{2}{r}\sum_{s=1}^{s} N_s C_s^2)$. Where r represents the reduction ratio, S represents the stage of feature extractor, N_s represents the number of Transformer blocks in this stage, C_s^2 represents the number of out channels in this stage. Moreover, the up-sampling enhancement module also results in increased complexity. The time complexity of up-sampling and the convolution is calculated as : $O(knd^2)$. Where n represents the size of the feature map, k represents the size of the convolution kernel, d and represents the dimensions of the feature map. In fact, the two-stream structure is a major contributor to the increased time complexity, which means it is at least twice as high as in single-stream.

5. Conclusions

This paper proposes a two-channel and two-stream framework (2C2S) for offline handwritten signature verification based on the transformer, which consists of original and central streams. The original stream receives the original signatures as input. The central stream receives the signatures generated by cropping the central at the original signatures as input. In this case, multi-scale signature features are extracted from both streams to characterize the writer's individuality. Since the input feature channel consists of reference and query signatures, a squeeze-and-excitation (SE) operation is applied in two standard Swin Transformer blocks for establishing the links among feature channels. Moreover, an up-sampling enhancement module is designed to help the model attend more to important features and ignore irrelevant information. The best accuracy on BHSig-B, BHSig-H, CEDAR, and SUES-SiG datasets reaches 93.25%, 90.68%, 100%, and 72.22%, respectively. Experimental results indicate that the proposed method outperforms other previous approaches and is promising for offline signature verification.

While 2C2S exhibits a high verification performance, the ability of the self-attention mechanism seems to be limited in signature verification. After all, since the self-attention mechanism is designed explicitly for single-input vision tasks, it is not perfectly adapted to pairwise verification tasks. Our future work will devote to develop a new attention mechanism that is highly applicable to pairwise verification tasks.

CRediT authorship contribution statement

Jian-Xin Ren: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **Yu-Jie Xiong:** Data curation, Writing – original draft, Funding acquisition, Supervision, Writing – review & editing. **Hongjian Zhan:** Visualization, Investigation, Supervision. **Bo Huang:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Funding

This work was supported by the National Natural Science Foundation of China (62006150); Science and Technology Commission of Shanghai Municipality (21DZ2203100); National Key Research and Development Program of China (2020AAA0109300); Shanghai Young Science and Technology Talents Sailing Program (19YF1418400).

References

Ahrabian, K., BabaAli, B., 2019. Usage of autoencoders and siamese networks for online handwritten signature verification. Neural Comput. Appl. 31 (12), 9321–9334.
Alam, T.M., Shaukat, K., Hameed, I.A., Khan, W.A., Sarwar, M.U., Iqbal, F., Luo, S., 2021. A novel framework for prognostic factors identification of malignant mesothelioma through association rule mining. Biomed. Signal Process. Control 68, 102726.
Alvarez, G., Sheffer, B., Bryant, M., 2016. Offline signature verification with convolutional neural networks. Tech. Report.
Aravinda, C., Meng, L., Uday Kumar Reddy, K., Prabhu, A., 2019. Signature recognition and verification using multiple classifiers combination of hu's and HOG features. In: 2019 International Conference on Advanced Mechatronic Systems (ICAMechS). pp. 63–68.
Berkay Yilmaz, M., Ozturk, K., 2018. Hybrid user-independent and user-dependent offline signature verification with a two-channel CNN. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.
Bhunia, A.K., Alaei, A., Roy, P.P., 2019. Signature verification approach using fusion of hybrid texture features. Neural Comput. Appl. 31 (12), 8737–8748.
Chattopadhyay, S., Manna, S., Bhattacharya, S., Pal, U., 2022. SURDS: self-supervised attention-guided reconstruction and dual triplet loss for writer independent offline signature verification. arXiv preprint arXiv:2201.10138.
Cordonnier, J.B., Loukas, A., Jaggi, M., 2019. On the relationship between self-attention and convolutional layers.. In: International Conference on Learning Representations.
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255.
Dey, S., Dutta, A., Toledo, J., Ghosh, S., Lladós, J., Pal, U., 2017a. SigNet: Convolutional siamese network for writer independent offline signature verification. Pattern Recognit. Lett..
Dey, S., Dutta, A., Toledo, J.I., Ghosh, S.K., Lladós, J., Pal, U., 2017b. Signet: Convolutional siamese network for writer independent offline signature verification. arXiv preprint arXiv:1707.02131.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations.
Fahmy, M.M., 2010. Online handwritten signature verification system based on DWT features extraction and neural network classification. Ain Shams Eng. J. 1 (1), 59–70.
Ghosh, R., 2021. A recurrent neural network based deep learning model for offline signature verification and recognition system. Expert Syst. Appl. 168, 114249.
Goon, L.W., Eng, S.K., 2021 2107, (1) 012069,
Hafemann, L.G., Sabourin, R., Oliveira, L.S., 2016. Analyzing features learned for offline signature verification using deep CNNs. In: 2016 23rd International Conference on Pattern Recognition. ICPR, IEEE, pp. 2989–2994.
Hafemann, L.G., Sabourin, R., Oliveira, L.S., 2017a. Learning features for offline handwritten signature verification using deep convolutional neural networks. Pattern Recognit. 70, 163–176.
Hafemann, L.G., Sabourin, R., Oliveira, L.S., 2017b. Offline handwritten signature verification—literature review. In: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications. IPTA, IEEE, pp. 1–8.
Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y., 2021. Transformer in transformer. Adv. Neural Inf. Process. Syst. 34, 15908–15919.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
Hu, J., Shen, L., Sun, G., 2020. Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. 42 (8), 2011–2023.

Jadhav, S.K., Chavan, M., 2018. Symbolic representation model for off-line signature verification. In: 2018 9th International Conference on Computing, Communication and Networking Technologies. IEEE, pp. 1–5.

Javidi, M., Jampour, M., 2020. A deep learning framework for text-independent writer identification. Eng. Appl. Artif. Intell. 95, 103912.

Li, C., Lin, F., Wang, Z., Yu, G., Yuan, L., Wang, H., 2019. Deephsv: User-independent offline signature verification using two-channel CNN. In: 2019 International Conference on Document Analysis and Recognition. IEEE, pp. 166–171.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE International Conference on Computer Vision. pp. 10012–10022.

Pal, S., Alaei, A., Pal, U., Blumenstein, M., 2016. Performance of an off-line signature verification method based on texture features on a large indic-script signature dataset. In: 2016 12th IAPR Workshop on Document Analysis Systems. IEEE, pp. 72–77.

Sharif, M., Khan, M.A., Faisal, M., Yasmin, M., Fernandes, S.L., 2020. A framework for offline signature verification system: Best features selection approach. Pattern Recognit. Lett. 139, 50–59.

Sharvari, K., 2021. A comparative study of transfer learning models for offline signature verification and forgery detection. J. Univ. Shanghai Sci. Technol. 23 1129–1139.

Shaukat, K., Luo, S., Chen, S., Liu, D., 2020a. Cyber threat detection using machine learning techniques: A performance evaluation perspective. In: 2020 International Conference on Cyber Warfare and Security. ICCWS, IEEE, pp. 1–6.

Shaukat, K., Luo, S., Varadharajan, V., 2022. A novel method for improving the robustness of deep learning-based malware detectors against adversarial attacks. Eng. Appl. Artif. Intell. 116, 105461.

Shaukat, K., Luo, S., Varadharajan, V., Hameed, I.A., Chen, S., Liu, D., Li, J., 2020b. Performance comparison and current challenges of using machine learning techniques in cybersecurity. Energies 13 (10), 2509.

Thakare, B.S., Deshmukh, H.R., 2018. A combined feature extraction model using SIFT and LBP for offline signature verification system. In: 2018 3rd International Conference for Convergence in Technology (I2CT). IEEE, pp. 1–7.

Tian, F., Gao, Y., Fang, Z., Fang, Y., Gu, J., Fujita, H., Hwang, J.-N., 2021. Depth estimation using a self-supervised network based on cross-layer feature fusion and the quadtree constraint. IEEE Trans. Circuits Syst. Video Technol. 1751–1766.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H., 2021a. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357.

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H., 2021b. Going deeper with image transformers. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 32–42.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Conf. Workshop Neural Inform. Process. Syst. 30 (11), 6000–6010.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: IEEE/CVF International Conference on Computer Vision. pp. 568–578.

Wei, P., Li, H., Hu, P., 2019. Inverse discriminative networks for handwritten signature verification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5764–5772.

Xiong, Y.-J., Cheng, S.-Y., 2021. Attention based multiple siamese network for offline signature verification. In: International Conference on Document Analysis and Recognition. Springer, pp. 337–349.

Xiong, Y.-J., Lu, Y., Wang, P.S., 2017. Off-line text-independent writer recognition: a survey. Int. J. Pattern Recognit. Artif. Intell. 31 (05), 1756008.

Xu, W., Xu, Y., Chang, T., Tu, Z., 2021. Co-scale conv-attentional image transformers. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 9981–9990.

Zhou, J., Sun, J., Zhang, M., Ma, Y., 2020. Dependable scheduling for real-time workflows on cyber–physical cloud systems. IEEE Trans. Ind. Inform. 17 (11), 7820–7829.

Zhu, X., Lyu, S., Wang, X., Zhao, Q., 2021. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In: IEEE/CVF International Conference on Computer Vision. pp. 2778–2788.