2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles (SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Metaverse)

SET: a squeeze-and-excitation transformer for offline signature verification

Jian-Xin Ren, Jue Chen*, and Yu-Jie Xiong*
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Shanghai 201620, China
jadeschen@sues.edu.cn and xiong@sues.edu.cn

***Abstract*—Offline handwritten signature verification, which is widely used in finance, commerce, and criminal forensic identification, plays an essential role in the fields of biometrics and document forensics. The development in deep learning has led to significant advances in signature verification over the past decade. However, it is still challenging to distinguish between skilled forgeries and genuine signatures because both are close similarities with only subtle differences in strokes. With this paper, we develop a novel squeeze-and-excitation transformer structure (named SET) for feature extraction and signature verification. SET comprises four stages and receives a two-channel signature pair consisting of reference and query signatures as input. The SET block is the core of each stage, which is utilized to enhance the feature learning ability and strengthen the association between feature channels. We evaluate the proposed SET on several public datasets (CEDAR, BHSig-B, and BHSig-H). Experimental results demonstrate that our approach outperforms existing methods.**

***Index Terms*—transformer, offline signature verification, signature pair, squeeze-and-excitation**

I. Introduction

Handwritten signature verification is simple and reliable biometrics to verify the identities of individuals. It is becoming increasingly appreciated in industry and academia owing to its efficiency, safety, and availability [1]. Typically, depending upon the data collection means, signature verification systems can be partitioned into two categories: online and offline signature verification systems [2]. In the online signature verification, the signature samples are collected while generating the dynamic handwriting trajectory. As a result, it contains position and temporal sequence information. In the offline system, the samples are obtained by scanning the signature data as static digital images. In this case, the online methods usually achieve better verification performance than offline methods [3].

The traditional method is extraordinarily burdensome and difficult to implement. It relies on the fully trained professional to perform a manual verification, and substantial time and energy are often expended in the verification process. These significantly limit the potential scope of signature verification applications. As computer vision technology evolves, some earlier works attempt to utilize machine learning-based algorithms to extract signature features, including HOG [4], SIFT [5], GMM [6]), and SURF [7]. Compared to the artificial verification approach, these algorithms enhance recognition efficiency and reduce the total amount of manual work [8]. However, these methods are extremely dependent on handcrafted feature representations that cannot extract complex stroke features well and are not appropriate for large-scale verification applications. Recent signature verification technology has been evolving along with the development of artificial intelligence [9]. Due to its high learning capability and flexibility, more researchers have utilized deep convolutional neural networks (CNN) to extract signature features. Okawa et al. [10] proposed a new feature extraction method that utilized Fisher Vector fused to KAZE features from both foreground and background offline signature images. Dey et al. [11] proposed a Siamese network with shared weights, SigNet, for signature verification. Cheng et al. proposed an attention-based model to to extract discriminative information for offline signature verification. Maergner et al. [12] proposed an approach that utilized triplet networks and state-of-the-art CNN models for signature verification.

Recently, the significant success of the transformer in natural language processing (NLP) has sparked a great deal of interest from the computer vision community. Some research has attempted to take Transformer beyond the natural language processing (NLP) field to the vision domain. The pioneering work of Dosovitskiy et al. [13] proposed the first vision transformer backbone for object classification. Touvron et al. [14] proposed a data-efficient image transformer (called DeiT) for decreasing dependence on large-scale datasets. Zheng et al. [15] proposed a SEgmentation TRansformer (SETR) for semantic segmentation tasks. Liu et al. [16] present a Swin Transformer that utilizes a spatially shifted window for modeling global and boundary features and achieved an excellent performance in vision tasks.

Inspired by the huge success of the transformer, a novel squeeze-and-excitation transformer approach (SET) is proposed for offline signature verification. The model takes a modified Swin-Transformer as the feature extractor. It adopts the hierarchical structure of CNN composed of four stages for increasing the receptive field. The input to SET is a two-channel signature pair that is composed of reference and query

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

979-8-3503-4655-8/22/$31.00 ©2022 IEEE

DOI 10.1109/SmartWorld-UIC-ATC-ScalCom-DigitalTwin-PriComp-Metaverse56740.2022.00263

signatures. The signature pair is first divided into a series of patches for converting inputs into sequence embeddings. Then these signature patches are extracted feature and learn the link between channels by the SET module. A decision layer is ultimately applied for distinguishing between genuine and forgery signatures.

II. Method

A. Overall Architecture

The network architecture of our proposed squeeze-and-excitation transformer (SET) model is illustrated in figure 1. The SET contains four stages to learn multiple-scale feature representation. The proposed approach receives the shape of a 224 x 224 x 2 signature pair that comprises reference and query signatures as input. The benefit of this two-channel input is that it can drastically reduce the parameter space and improve network speed, which allows the two-channel network a superior fit for signature verification. This paper adopts the modified Swin-B as the feature extractor. We implement the squeeze-and-excitation module (SE) into the standard Swin Transformer block because the input of SET is comprised of reference and query signatures connected in series that imply the potentially great connection between channels. Consequently, there is a need to establish a direct link between channels to capture the essential features through the SE module. As shown in figure 1, the feature extractor consists of four stages. The signature pair with a size of H × W × 2 is first to differentiate into non-overlapping patches by patch partition and fed into stage 1. Stage 1 consists of the linear embedding and the SET block. After the linear embedding, these patches are projected into arbitrary dimensions (denoted by C). Subsequently, the SET block is employed to learn feature representation and capture deep associations between channels. Following the pyramid structure, the output feature map is progressively reduced from 4 to 32-stride. Stage 2 $\sim$ 4 are almost the same as stage 1, except that the patch merging is substituted for the linear embedding. Specifically, the patch merging is employed for down-sampling and enlarging dimensions, which is similar to the focus operation in Yolo ([17]).

B. SET Block

The SET block is the core component of our approach, constructed based on the Swin Transformer blocks. As shown in figure 2, the SET block mainly involves three steps. Step 1 and step 3 are composed of a LayerNorm (LN) layer, a multi-head self-attention, a residual connection, and an MLP layer. The difference between the two steps is step 1 adopts the window-based multi-head self-attention (W-MSA) module, and step 3 adopts the shifted window-based multi-head self-attention (SW-MSA) module. The two modules facilitate information interactions, which is similar to the concept of receptive field expansion in CNN [18]. Step 2 is a squeeze-and-excitation module (SE) that aims to enhance channel links. These relationships are expressed by the following equation:

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

$$P_2 = SE(MLP(LN(\hat{P}_1)) + \hat{P}_1) \tag{2}$$

$$\hat{P}_3 = SW - MSA(LN(P_2)) + P_2 \tag{3}$$

$$P_3 = MLP(LN(\hat{P}_3)) + \hat{P}_3 \tag{4}$$

Where P denotes input feature map, $\hat{P}_1$ denotes the output feature map of W-MSA, SE denotes squeeze-and-excitation.

C. Squeeze-and-Excitation

The squeeze-and-excitation module (SE) is not new in computer vision. It is first proposed by [19], which is utilized in object classification to increase performance. As shown in figure 3, a global pooling operation is first applied to squeeze the feature map (H, W, C) into $(1, 1, C)$. It can be represented with the following equation:

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

Where f_c represents the feature matrix of the input f, c represents the channel dimension of f. H and W represent the height and width dimensions of f, respectively. i and j represent the ordinate and the abscissa of the feature matrix, respectively.

Subsequently, the association between the fused feature channels is established by the excitation operation that is composed of two fully-connected layers, a rectified linear unit (ReLU), and a logistic sigmoid function. Finally, the normalized weight and the original input are applied to a scale operation for generating the final feature representation. The process can be represented with the following equation:

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

Where W_1 and W_2 are the weight matrix for the fully connected layer. The r denotes the scaling parameter that is utilized to decrease the amount and complexity of the model. The s donates a logistic sigmoid function.

III. Experiments

A. Datasets and experimental protocol

We evaluate the validity of the proposed SET on three publicly available datasets: CEDAR [20], BHSig-B, and BHSig-H [21], as described in the following. CEDAR is an English signature dataset that contains 55 writers, and each writer has 24 genuine and 24 forged signatures, respectively. All signature samples are requested to be written on a predefined paper of 2×2 inches to achieve 1,320 genuine and 1,320 forged signature images. Moreover, the signature Image is collected by scanning at 300 dpi and is stored in PNG format. To comparison for existing works, we utilize 50 writers to train and 5 writers to test the model. BHSig-B and BHSig-H are two sub-datasets from a large Indic script signature

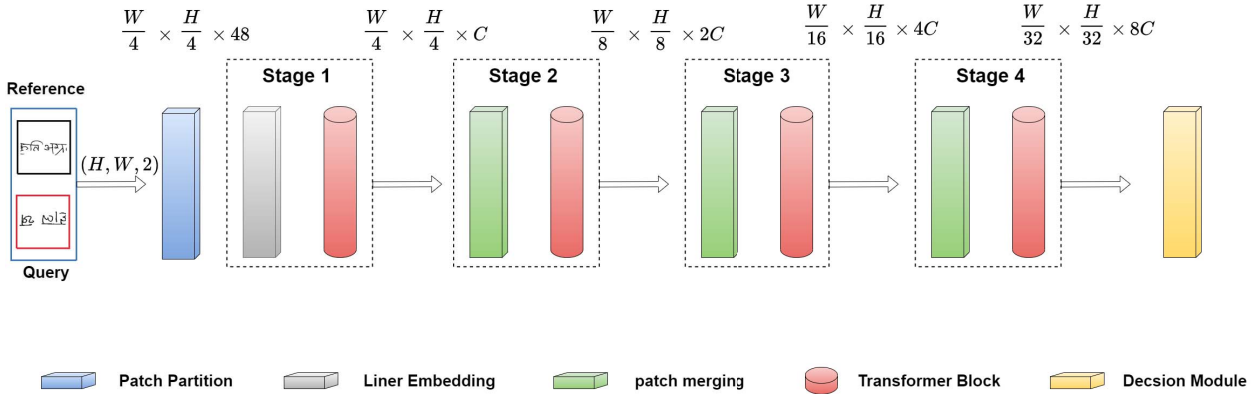

Fig. 1. The overall architecture of the proposed approach.

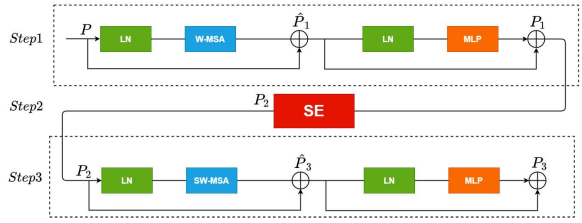

Fig. 2. The inner of SET block.

dataset, BHSig260. BHSig260 contains 260 writers, each with 24 genuine and 30 skilled forgery signatures, where 100 writers belong to BHSig-B, and 160 writers belong to BHSig-H. As a result, BHSig-B includes 2,400 genuine and 3,000 forged signatures. BHSig-H includes 2,840 genuine and 4,800 forged signatures. Moreover, both are requested to scan in gray-scale images and stored in TIFF format. Following the previous work, 100 authors in BHSig-H are used to test, and the others are used to train. Fifty authors in BHSig-H are used to test, and the others are used to train.

Since the input of the proposed SET is a signature pair, the datasets are composed of the signature pairs that consist of reference and query signatures. In this case, the positive samples consist of genuine-genuine signature pair, and the negative samples consist of genuine-forged signature pair. Taking the CEDAR dataset as an example, every writer can be divided into $C_{24}^2 = 276$ positive samples and $24 \times 24 = 576$ negative samples. To avoid imbalanced 276 positive samples are randomly selected from all positive samples. The detailed divisions are presented in Table 1.

TABLE I
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

B. Evaluation Metrics

In order to demonstrate the proposed approach, we applied four indices: Accuracy (Acc), False Rejection Rate (FRR), False Acceptance Rate (FAR), and Equal Error Rate (EER).

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

To evaluate the effectiveness of the proposed SET, we compare it against state-of-the-art methods on three datasets. The performance of these methods is provided in Table 2. It can be seen that state-of-the-art performances are achieved on all datasets. Specifically, our SET, SigNet, and 2-Channel-2-Logit obtain the highest accuracy by 100% on the CEDAR dataset. The leading cause of higher performance is that the dataset is simplistic, and most samples are used to train. On the BHSig-B dataset, the proposed method achieves the highest accuracy performance (91.79%), which shows the superior performance of the proposed method. Notably, SURDS obtains the lowest FRR, while the higher FAR influences the final results. For the BHSig-H dataset, the proposed approach achieves an accuracy of 90.06%. Compared to the other methods, it is the only method with an accuracy greater than 90%, which illustrates the excellent performance of the proposed method.

TABLE III
CROSS DATASETS VALIDATION RESULTS (%).

dataset	CEDAR	Bengali	Hindi
CEDAR	**100.00**	50.00	50.00
Bengali	50.80	**91.79**	74.41
Hindi	61.92	79.69	**90.06**

D. Cross datasets validation

We also evaluate the generalization ability of the proposed SET with cross-language testing. As a result, a cross-language experiment is performed, in which the proposed model is trained on one dataset, then tested on another dataset. Table 3 provides the cross-language of generalization accuracy, where the columns correspond to the testing dataset and the rows

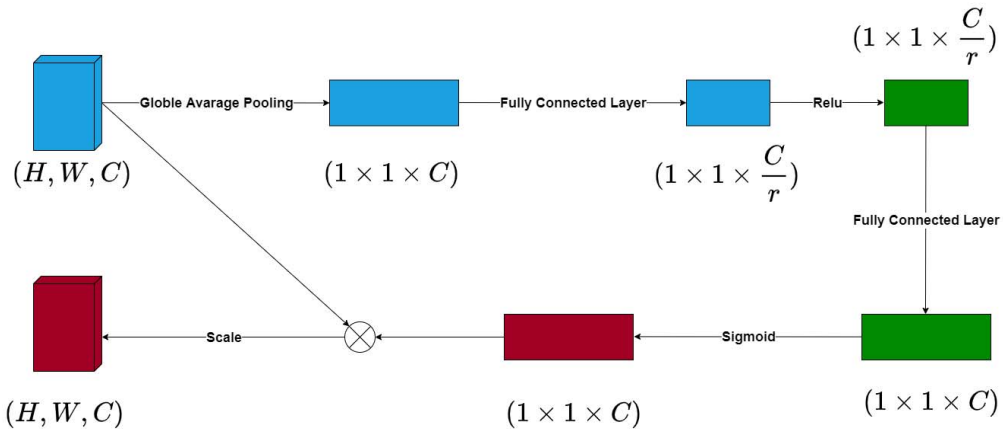

Fig. 3. The squeeze-and-excitation module.

TABLE II
COMPARISON OF THE PROPOSED METHOD WITH THE STATE-OF-THE-ART METHODS ON THREE SIGNATURE DATASETS (%).

Datasets	Methods	FAR	FRR	ACC	EER
CEDAR	2-Channel-2-Logit [22]	-	-	**100.00**	**0**
	MSN [23]	3.18	**0**	98.40	1.63
	SigNet [3]	**0**	**0**	**100.00**	**0**
	A. K. Bhunia et al. [24]	5.01	6.12	-	-
	M. Shari et al. [25]	4.67	4.67	-	-
	SET (ours)	**0**	**0**	**100.00**	**0**
Bengali	2-Channel-2-Logit [22]	10.44	9.37	88.08	11.92
	MSN [23]	10.42	6.44	91.56	8.43
	SigNet [3]	13.89	13.89	86.11	13.89
	SURDS [26]	19.89	**5.42**	87.34	-
	S. K. Jadhav et al. [27]	-	-	90.36	-
	LBP and ULBP [21]	33.82	33.82	66.18	33.82
	SET (ours)	**5.67**	10.83	**91.79**	**8.21**
Hindi	2-Channel-2-Logit [22]	-	-	86.66	13.34
	MSN [23]	17.06	**5.16**	88.88	11.31
	SigNet [3]	15.36	15.36	84.64	15.36
	SURDS [26]	12.01	8.98	89.50	-
	S. Pal et al.[21]	24.47	24.47	66.18	75.53
	SET (ours)	**8.93**	10.94	**90.06**	**9.32**

correspond to the training dataset. It can be seen that the model achieves more accuracy when it is trained and tested on the same dataset. However, the accuracy significantly declines when the model is trained and tested on different language datasets. This is primarily because of clear writing habits and language style differences.

E. Conclusions

In this paper, we propose a squeeze-and-excitation transformer (SET) to address the problem of offline signature verification. We attempt to utilize the two-stream structure with a modified Swin-Transformer to extract the multi-scale feature. The proposed hierarchical pyramid structure approach can capture local and global signature information. We also employ the squeeze-and-excitation module (SE) to construct the association between different channels, which enhances the feature extraction capability and captures abstract channel information. The proposed approach achieves performance on three public signature datasets. Experimental results indicate that the proposed SET is effective for offline signature verification. In our future work, we will seek ways to enhance the redundant information to steer the model to focus on useful information directly.

REFERENCES

[1] Y.-J. Xiong, Y. Lu, P. S. Wang, Off-line text-independent writer recognition: a survey, International Journal of Pattern Recognition and Artificial Intelligence 31 (05) (2017) 1756008.

[2] L. G. Hafemann, R. Sabourin, L. S. Oliveira, Offline handwritten signature verification—literature review, in: 2017 seventh international conference on image processing theory, tools and applications (IPTA), IEEE, 2017, pp. 1–8.

[3] L. G. Hafemann, R. Sabourin, L. S. Oliveira, Learning features for offline handwritten signature verification using deep convolutional neural networks, Pattern Recognition 70 (2017) 163–176.

[4] W. Zhou, S. Gao, L. Zhang, X. Lou, Histogram of oriented gradients feature extraction from raw bayer pattern images, IEEE Transactions on Circuits and Systems II: Express Briefs 67 (5) (2020) 946–950.

[5] M. Kasiselvanathan, V. Sangeetha, A. Kalaiselvi, Palm pattern recognition using scale invariant feature transform, International Journal of Intelligence and Sustainable Computing 1 (1) (2020) 44–52.

[6] P. An, Z. Wang, C. Zhang, Ensemble unsupervised autoencoders and gaussian mixture model for cyberattack detection, Information Processing & Management 59 (2) (2022) 102844.

[7] P. Agrawal, T. Sharma, N. K. Verma, Supervised approach for object identification using speeded up robust features, International Journal of Advanced Intelligence Paradigms 15 (2) (2020) 165–182.

[8] J. Zhou, K. Cao, X. Zhou, M. Chen, T. Wei, S. Hu, Throughput-conscious energy allocation and reliability-aware task assignment for renewable powered in-situ server systems, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 41 (3) (2021) 516–529.

[9] J. Zhou, J. Sun, M. Zhang, Y. Ma, Dependable scheduling for real-time workflows on cyber–physical cloud systems, IEEE Transactions on Industrial Informatics 17 (11) (2020) 7820–7829.

[10] S. Masoudnia, O. Mersa, B. N. Araabi, A.-H. Vahabie, M. A. Sadeghi, M. N. Ahmadabadi, Multi-representational learning for offline signature verification using multi-loss snapshot ensemble of cnns, Expert Systems with Applications 133 (2019) 317–330.

[11] S. Dey, A. Dutta, J. I. Toledo, S. K. Ghosh, J. Lladós, U. Pal, Signet: Convolutional siamese network for writer independent offline signature verification, arXiv preprint arXiv:1707.02131 (2017).

[12] P. Maergner, V. Pondenkandath, M. Alberti, M. Liwicki, K. Riesen, R. Ingold, A. Fischer, Combining graph edit distance and triplet networks for offline signature verification, Pattern Recognition Letters 125 (2019) 527–533.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale (2021).

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

[15] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 6881–6890.

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows (2021) 10012–10022.

[17] X. Zhu, S. Lyu, X. Wang, Q. Zhao, Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios, in: IEEE/CVF International Conference on Computer Vision, 2021, pp. 2778–2788.

[18] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, Z. He, A survey of visual transformers, arXiv preprint arXiv:2111.06091 (2021).

[19] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (8) (2020) 2011–2023.

[20] M. K. Kalera, S. Srihari, A. Xu, Offline signature verification and identification using distance statistics, International Journal of Pattern Recognition and Artificial Intelligence 18 (07) (2004) 1339–1360.

[21] S. Pal, A. Alaei, U. Pal, M. Blumenstein, 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, 2016, pp. 72–77.

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

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

[24] A. K. Bhunia, A. Alaei, P. P. Roy, Signature verification approach using fusion of hybrid texture features, Neural Computing and Applications 31 (12) (2019) 8737–8748.

[25] M. Sharif, M. A. Khan, M. Faisal, M. Yasmin, S. L. Fernandes, A framework for offline signature verification system: Best features selection approach, Pattern Recognition Letters 139 (2020) 50–59.

[26] S. Chattopadhyay, S. Manna, S. Bhattacharya, U. Pal, Surds: self-supervised attention-guided reconstruction and dual triplet loss for writer independent offline signature verification, arXiv preprint arXiv:2201.10138 (2022).

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