TSOSVNet: Teacher-student collaborative knowledge distillation for Online Signature Verification

chandra sekhar v
IIIT-SriCity
Andhra Pradesh, India
chandrasekhar.v@iiits.in

Avinash Gautam
BITS-Pilani,
Pilani-333031,India
avinash@pilani.bits-pilani.ac.in

Viswanath P
IIIT-SriCity,
Andhra Pradesh, India
viswanath.p.v@iiits.in

Sreeja SR
IIIT-SriCity,
Andhra Pradesh, India
sreeja.srv@iiits.in

Rama Krishna Sai G
IIT-Tirupati,
Andhra Pradesh, India
rkg@iittp.in

Abstract

Online signature verification (OSV) is a standardized personal authentication scheme with wide social acceptance in critical real-time applications include access control, m-commerce, etc. Even though the current advances in Deep learning (DL) technologies catalysed state-of-the-art frameworks for challenging domains like computer vision, speech recognition, etc., the DL-based frameworks are voluminous with huge trainable parameters and are hard to deploy in real-time systems demanding faster inference. To adopt DL into OSV for improved performance, we propose an OSV framework made up of teacher-student collaborative knowledge distillation (TSKD) technique. A heavy Transformer based teacher is trained first and the teacher knowledge is distilled into a very lightweight Convolutional Neural Network (CNN) based student. A well trained teacher network results in an efficient deep representative feature learning by the student and results in a performance improvement. In a thorough set of experiments with three popular and standard datasets, i.e., the MCYT-100, SUSIG, and SVC, TSOSVNet framework, with a CNN based student model requiring only 3266 trainable parameters results in an EER of 12.42% compared to the recent SOTA 13.38% by a model with 206277 parameters in skilled_01 category of MCYT-100 dataset. In comparison to cutting-edge CNN-based OSV models, the proposed TSOSVNet produced a state-of-the-art EER in the most of the test categories with an average of 90% lesser trainable parameters.

1. Introduction

Online Signature biometrics is efficient in user recognition, verification, and security applications. The online signature has inherent advantages, like it cannot be forgotten, cannot get lost, and is almost impossible to imitate perfectly or duplicate compared to other biometrics such as passwords, keys and ID cards [9, 23] etc,. Due to these advantages, online signature verification gained momentum to use in real-time applications and demonstrated outstanding performance in numerous computer vision activities like person-identification, transactions, agreements, etc [9, 47].

Tolosona *et al*. [42] put forward an OSV framework consisting of a Siamese based RNNs, advanced GRU's to extract the deep representative features and feed them into the DTW network. The framework resulted in better EER in skilled forgeries compared to DTW-based frameworks. In 2019, a series of OSV techniques have been proposed based on convolutional neural networks (CNNs) [3], [48], [32], [31], [34], [19], [21] and these frameworks yield lesser EER compared to the sequence models. In [3], three types of features representing a signature's, physical, frequency and statistical properties are calculated, and these features are feed into the ensembled classifier. Using a normalised distribution summation technique, the classification outcomes of seven classifiers are combined. Vorugunti *et al*. [32] designed an OSV framework using fusion of feature representations by fusing the hand-crafted and deep representative features from an Auto Encoder. The composite feature set is passed on to the Depth-wise Separable Convolutional Neural Network (DWSCNN), which can achieve an EER of 13.38% in one-shot learning. Vorugunti *et al*. [8], combined CNN and LSTM to form a composite model, in which the local deep features from CNN are input to LSTM net-

work to learn long-term dependencies of an input signature. Lai *et al*.[19] devised a model in which synthetic signatures are generated using a neuromotor with different distortion levels.To rank the synthesised signature samples, a simple one-dimensional convolutional network is employed. The framework yields an EER of 5.50% in the skilled_01 category of the MCYT-100 dataset. In an extension work, Lai *et al*.[20], proposed synthetic signature generation technique using Sigma Lognormal model by applying different distortion levels on genuine template signatures. The synthetic signatures are ranked based on average precision optimization. These samples are used to train the 1D CNN based OSV framework. The framework results in an EER of 3.84% in skilled_01 category of MCYT-100 dataset. Recently Vorugunti *et al*.[44] Analysed the influence of convolution type and order on the EER output of OSV frameworks.

Even though various deep learning based OSV frameworks focusing on CNN [32, 31, 34, 19, 3, 44], LSTM [42], combination of CNN and LSTM [8] have been proposed, the CNN suffers from inherent drawbacks like acting only on local receptive fields of the given input signature and ineffective to capture global signing patterns [43, 51]. The LSTM frameworks have fundamental pitfalls like depending on previously computed hidden states to calculate the current output restricts the efficiency of the OSV frameworks based on CNN and LSTM backbones [7, 2]. Recently, Transformer-based architectures resulted in superior performance in Natural Language Processing (NLP) [43], Speech [45], and Image processing [51]. Vision Transformer (ViT) [12] has been proposed, which overcomes the drawbacks of LSTM and CNNs [6]. Unlike the LSTM-based frameworks, Transformer architecture, enables access any part of the sequence history, regardless of the length of the sequence. This makes the Transformer potentially more befitting for learning the recurring patterns with long-term dependencies like online signatures [1].

Based on the above discussion, even though the fact that Transformer variants are powerful and best suits in effectively learning the long-term signing patterns, Transformer based OSV remains unexplored in the literature. The massive number of trainable parameters results in extremely high training and inference costs. They are unsuitable for low-power, resource-constrained devices due to their high memory and computation requirements [5, 39]. Hence, to reap the benefits of Transformer architecture in OSV and to overcome the performance-limiting bottlenecks, our contribution in this work is two-fold:

1) **C1:** We successfully apply Transformer architecture to Online Signature Verification. So that self-attention mechanism, which is critical element of learning long term spatial relationship of each signature is effectively applied to OSV.

2) **C2:** We design a teacher-student based knowledge distillation (TSKD) technique which involves the Transformer based teacher network and distilling the knowledge from the Teacher to a single convolution layer lightweight CNN based student network.

3) **C3:** We conduct a comprehensive set of experiments, to demonstrate Transformer's efficiency in addressing long-term dependencies better than sequence models like LSTM, GRU.

The remainder of the paper is organized as follows. Section 2 discuss the proposed Teacher-Student based OSV framework. Section 3 examines the experimentation and comparative analysis of the proposed TSOSVNET framework with the recent and state-of-the-art frameworks and the conclusions are drawn in Section 4.

2. Proposed Teacher-student network based Online Signature Verification framework

As shown in Fig. 1 and 3, we have developed three teacher networks 1. A CNN based teacher network comprising of two convolution blocks followed by the dense layer with 64 nodes (VanillaCNN). 2. As shown in Fig. 1, a teacher encompassing the transformer with convolution block is marked in red (TranConv). 3. As shown in Fig. 1 (a) and (b), the same Transformer architecture, the convolution block is replaced with a single dense layer consisting of 128 and 64 nodes (TranDense). As depicted in Fig. 4, the student architecture remains same for all the three teacher architectures. Further, we refer to each teacher network as VanillaCNN, TranConv, and TranDense respectively. The rationale behind selecting the three types of teacher architectures is to investigate the efficient network structures and the quality of intermediate feature embeddings and soft logits learned by the corresponding teacher networks to transfer the robust feature representations to the student network.

2.1. Input Signature format to the OSV framework

As depicted in Fig. 1 and Table 2, for each acquired raw signature from a writer, a set of statistical formulae based features, which characterize the whole signature, are computed. E.g. (standard deviation of x)/Δx, where 'x' indicates the trace of the x-coordinates of the signature. As illustrated in Table 2, for the MCYT-100 dataset, a total of 100 global features and for SVC, and SUSIG datasets 47 global features are computed.

2.2. Transformer: TranConv

To explain the workflow of the proposed architecture, we are considering a signature from the MCYT-100 dataset. As depicted in Fig. 1, a feature vector of length 100×1 representing a signature is input to the Transformer. Initially, for the feature vector linear and periodic time features are

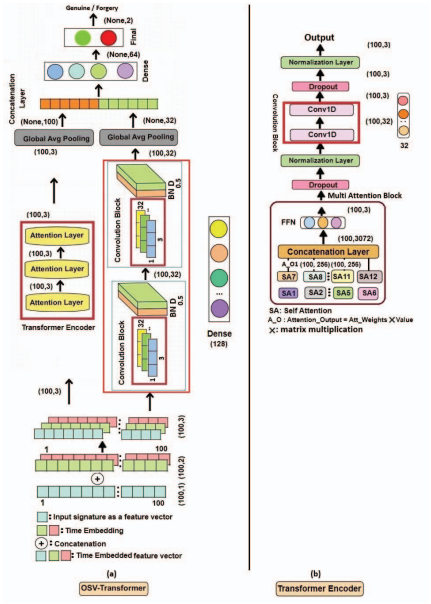

Figure 1. The illustration of the Transformer architecture and corresponding encoder architecture.

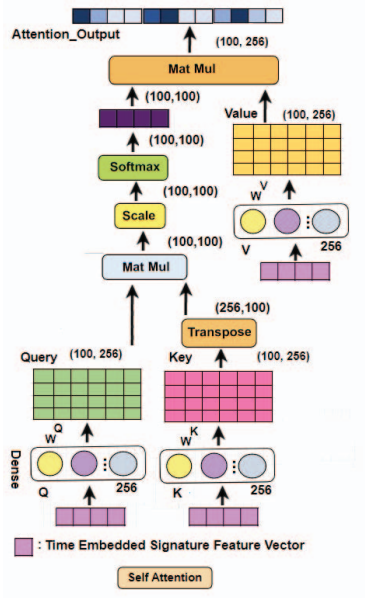

Figure 2. The self attention mechanism of the encoder architecture.

computed and concatenated with the original input feature vector which result in a batch of size 100×3. The time embedded feature vector $F = (f_1, ..., f_T)$ of dimension 100×3 is passed onto both the encoder and convolution blocks.

2.2.1 Self Attention Block

As portrayed in Fig. 1(b), the attention layer is composed of multi-head attention block, which consists of a set of self-attention blocks. We have set the number of heads to 12. The self-attention block is depicted in Fig. 2. Same set of operations are performed in each self-attention block, which is described as follows: the time-embedded signature feature vector F of size 100×3, is passed to the two

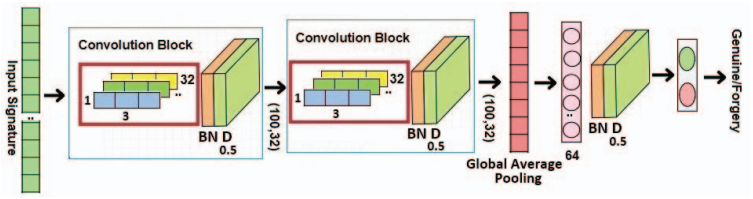

Figure 3. The architecture of the VanillaCNN teacher network composed of only convolution blocks.

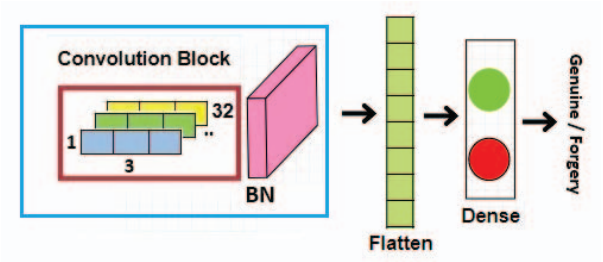

Figure 4. The architecture of the Student network of the proposed OSV framework.

dense layers named Q and K consisting of 256 neurons each. The attention weights $Query$, Key, and $Value$ are computed as $Query = F \cdot W^Q$, $Key = F \cdot W^K$ and $Value = F \cdot W^V$. W^Q, W^K and W^V represents the weight matrices learnt through the dense layers Q, K and V. The outputs of the self-attention layer are computed as follows: $Attention\_Output$ = softmax$((Query \cdot Key^T)/\sqrt{256}) \cdot Value$. The $Attention\_Output$ from each self-attention head is of dimension $(100, 256)$. As depicted in Fig. 1(b), the output from 12 attention heads is concatenated to form an attention vector of length $(100, 3072)$. The concatenated attention vector is feed into a dense layer of size 3, to output a final vector of length $(100, 3)$. The output from the multi attention block is passed to a convolution layer with 32 filters of dimension 1×1 and results in a convolved feature vector of length 100×32. The output is passed to the second convolution layer with three filters of dimension 1×1 and results in a feature vector of size 100×3. The output from the transformer encoder is 100×3.

2.2.2 Convolution Block

As illustrated in Fig. 1, the time-embedded feature vector $F = (f_1, ..., f_T)$ of dimension 100×3 is passed on to a set of convolution blocks. The first convolution block is a 1D convolution layer consisting of 32 filters, each of size 1×3 and outputs a set of deep representative feature vectors of size 100×32 and passed on to the second convolution block. The same set of operations are performed on the output of the first convolution block and outputs a feature vector of dimension 100×32.

The output from the transformer encoder block is of size 100×3, and the convolution block is of size 100×32. The outputs are passed on to the global average pooling (GAP) layers, and the resultant outcome is concatenated to form an output of size 100×32. Finally, the feature vector is passed onto the fully-connected layer based classifier for classification into Genuine and Forgery. As discussed above, in the case of TranDense, the convolution blocks (which are marked red) in Fig. 1 (a) and (b) are replaced with dense layers of size 128 and 32 nodes respectively and the same set of operations are performed.

2.3. Transformer: VanillaCNN

As depicted in Fig. 3, a VanillaCNN is a CNN with two convolution blocks. The transformer encoder network is not used. An input feature vector $F = (f_1, ..., f_T)$ of size 1×100 is given as an input to the first convolution block, which contains 32 filters, each of size 1×3, and outputs a feature map of size 100×32. The similar operations is executed in the second convolution block and outputs a representative feature map of size 100×32. Global Average Pooling is applied to the deep feature representations from the convolution block. Finally, the classifier outputs the signature class.

2.4. Student Network

As depicted in Fig. 4, we have used same student network for all three types of teacher networks, i.e, VanillaCNN, TranDense, and TranConv. The student network consists of a single convolution block of 32 filters, each of size 1×3. The feature representations from the convolution block of size 100×32 are flattened and passed onto the final softmax layer for classification. As illustrated in Table

Table 1. The number of parameters required for a teacher network and corresponding distilled student network.

Teacher Type	Teacher	Student	% of parameters reduced
Only Conv1D Layers (VanillaCNN)	208738	6658	96.81%
Transformer + Dense (TranDense)	83432	6658	92.01%
Transformer + Convolution (TranConv)	118580	3266	97.25%

1, the student network requires only 3266 parameters compared to the corresponding teacher TranConv. This results in a reduction of 97.25% of parameters. We have used a widely used knowledge distillation approach to facilitate the transfer knowledge from more significant knowledge capacity teachers to a student [13, 27, 50].

2.5. Knowledge Distillation Technique

In this section, we briefly discuss about the knowledge distillation [50, 17] procedure from the teacher network to the student network. As depicted in Fig 1 (a), the last feature of the teacher network can be expressed as $T_{f1} = T(F; W^{AttnLayer3})$, $T_{f2} = T(F; W^{Convlayer2})$. $T_F = T_{f1}||T_{f2}$ and corresponding logit as $logits\_T = C_t(Dense(T_F); W^{dense})$. $AttnLayer3$ represents the third and last attention layer of the attention block and $W^{AttnLayer3}$ represents the corresponding weights. As depicted in Fig 1, $Convlayer2$ represents the convolution layer of the second convolution block, $Dense$ represents a dense layer of the classifier C_t. Similarly, we compute the last feature of the student's network as $S_f = C_s(F; W^{Convlayer1})$. As depicted in Fig 4, the students network consists of a single convolution block and a single dense layered classifier, the features extracted by the student are equal to the logits $logits\_S$. The difference between the teacher and the student output logits is measured by the Kullback–Leibler (KL) divergence, which can be stated as : $\frac{1}{N}\sum_{i=1}^{N} L_{KL}(C_t(Dense(T(F_i; W^{AttnLayer3})||$ $T(F_i; W^{Convlayer2})), W^{dense}), C_s(F; W^{Convlayer1}))$ The following is the definition of the cross-entropy loss between the true label and the student network's predicted value: $\frac{1}{N}\sum_{i=1}^{N} L_{CE}(C_s(F; W^{Convlayer1}), y_i)$. The distiller's objective is to reduce the difference between the teacher and student predictions, as well as the difference between the student output and the true label, i.e., $argmin_{ws}\sum(\alpha * \tau^2 * L_{KL} + (1-\alpha) * L_{CE})$. ws represents the parameters of the student, α denotes the weight of KL divergence. τ indicates the distillation temperature, which is used to produce a softer probability distribution (softmax) over classes (Genuine or Forgery).

3. Experimentation and Results

This section initially provides a summary of the datasets used to evaluate the performance of our proposed TSOSVNet. As illustrated in Table 1, We evaluated our framework and compared it to the most recent SOTA OSV frameworks using three benchmark datasets (MCYT-100, SVC, and SUSIG). In this section, In line with the literature [14, 22, 24, 18], we have thoroughly appraised the proposed framework in Skilled_01 (which evaluates one-shot learning of the framework), Skilled_05, Skilled_10, Skilled_15, Skilled_20 categories.

Let 'W' indicate the number writers in the system. 'R', 'F' indicates the number of Real/Genuine and Fake/Forgery signature samples specific to each user. W=100, R=25, F=25 in MCYT-330 and MCYT-100 datasets. In evaluating the Skilled_01 category for U_i, one genuine and one forgery sample of U_i is used to train the framework. Remaining 'R-1', i.e., (25-1=24) genuine samples are used to test the framework for True Acceptance Rate (TAR). 'F-1', i.e., (25-1 =24) forgery signature samples are used to gauge the False Acceptance Rate (FAR) of the framework. Similar is the case with Skilled_05/10/15/,..,/(G-5) categories. In short Skilled_01/05/10/15/20 are represented as S_01/05/10/15/20.

As shown in Table 1, the student network reduces the number of parameters by 96.81%, 92.01%, and 97.25%, respectively for each corresponding teacher network. In case of TranConv, the student has 3266 trainable parameters which reduces the parameter count by 97.25%, followed by VanillaCNN and TranDense. To choose the best teacher architecture among the three and to appraise the best teacher network with SOTA OSV frameworks, we did an ablation study, which is depicted in Table 3.

Table 3 illustrates the EER outcome of the proposed teacher and corresponding student networks with MCYT-100 dataset. Even though, the student distilled from VanillaCNN resulted in the least EER of 11.69% and 2.05% respectively, in S_01 and S_05 categories, the student distilled from TranConv resulted in the least EER in S_10, S_15 and S_20 categories. The student distilled from TranDense is unable to result in lesser EER values in any category. Similar studies have been done for SVC and SUSIG datasets.

Based on the above discussion, we can summarize that, in all three datasets, the student distilled from TranConv resulted in the least EER compared to other teacher networks. This performance can be attributed to the following reasons :1) The larger and more robust teacher networks are expected to learn efficient, representative, and reliable knowledge. 2) The CNN efficiently captures spatial local signing patterns and the Transformer is efficient in capturing long-range correlations of signing patterns. The hybrid

Table 2. The datasets specifics used in the experimental study.

DataSet	MCYT-100	SVC	SUSIG
Total number of Writers	100	40	94
Number of features representing a signature	100	47	47
Number of genuine/ real signatures for each writer	25	20	20
Number of forgery/ fake signatures for each writer	25	20	10
Total set of real signatures	2500	800	1880
Total set of fake signatures	2500	800	940

teacher network resulted in the efficient distillation of short and long range representative features to the student network.

Fig. 5 portrays the EER recorded for each user by the three teachers, i.e., VanillaCNN, TranConv, and TranDense, and Fig. 6 shows the EER outcome by the corresponding students in Skilled_01 category of MCYT dataset. The number of users who reached the matching EER is shown by the intensity of the colour map. Figure. 5(a) depicts that the teacher network representing VanillaCNN yields 0 EER for close to 6 users. About 15 individuals obtained an EER ranging from 0.4 to 0.5, and for three users, between 95-100, the EER is 1.0. Figure. 6(a) illustrates the EER outcome by the student network distilled from VanillaCNN. Compared to Fig. 5a, Fig. 6a is more dense in the upper region. More EER is recorded between 0.20 to 0.65. This indicates the teacher network is ineffective, and the student cannot generalize the input signature classification. The same is shown by an EER of 11.69%, as depicted in the Table 3. Similar is the case with Fig. 5b and Fig. 6b. In the case of Fig. 6c, the teacher TranConv records slightly higher EER than VanillaCNN and TranDense.

Table 4 illustrates the performance appraisal of the proposed framework with the latest frameworks pertaining to MCYT-100 dataset, which is the most commonly used dataset in the OSV domain. As depicted in Table 4, we have compared the effectiveness of the proposed OSV framework with both the DL and non-DL-based OSV frameworks. The DL-based frameworks proposed in the literature till now [33], [30], [4], [49], and [46] require approximately 95101, 35423, 206277, 10000, and 580000 trainable parameters respectively. The student network resultant from TranConv requires only 3266 trainable parameters and still achieves the state-of-the-art EER outcome in Skilled_01, Skilled_10, Skilled_15 and Skilled_20 categories. Achieving an EER of 12.42% in the Skilled_01 category (one-shot learning) with only 3266 trainable parameters endorse the competence of the proposed framework.

Table 5 compares the EER results obtained using the proposed framework to those obtained using the most recent frameworks evaluated using the SVC dataset. In Skilled_01 category, the proposed framework outcomes an EER of 6.45%, which is 20% higher than the framework with SOTA EER 5.37%. But as depicted in the table, the proposed framework requires 96.56% parameters lesser than the framework with State-of-the-art EER. Table 6 illustrates the EER outcome of the proposed framework with other latest frameworks assessed with the SUSIG dataset and can make similar observations.

Fig. 7 illustrates the attentions learnt during training by the three teacher networks VanillaCNN, TranDense and TranConv respectively. The intensity or score of each feature of the time embedded input feature vector determines the importance of each feature in classifying a signature as genuine or forgery. Fig. 7(a) illustrates the attentions representing the most varying deep representative features of user id 1 learned by VanillaCNN. The model looks into the feature ids/numbers 2,4,27 (marked in thick red) are the most important representatives in classifying the signature of user id 1.

4. Conclusion and Future Work

Our contribution in this work is twofold. We have designed a hybrid CNN-based Transformer network for Online Signature Verification (OSV). TranConv architectures take merits of both CNN's and transformers to capture local and global signing patterns of the users. The inherent disadvantage of the transformers in both performance and computational cost is overcome by employing Teacher Student network-based knowledge Distillation. The TrancConv based student resulted in state-of-the-art EER in Skilled_01 category with only one convolution layer. The novel advantages like lightweight and higher classification accuracy of the proposed framework makes it competent to adopt in challenging real time applications like M-Commerce etc.

References

[1] D. Ahn, S. Kim, H. Hong, and B. Chul. Star-transformer: A spatio-temporal cross attention transformer for human action recognition. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6836—-6846, 2021.

Table 3. The EER outcome by the three Teachers and the student network in MCYT (DB1) dataset. ”S” represents the skilled categories. The 01,05,10,15,20 represents the number of signature samples used for training.

Method	S_01	S_05	S_10	S_15	S_20
Teacher: CNN based:	11.58	1.37	0.2	0.25	0.1
Student:	11.69	2.05	1.47	0.2	1.3
Teacher: Transformer based on Dense:	10.98	3.12	1.37	0.85	0.4
Student:	12.75	2.87	1.27	0.65	0.4
Teacher: Transformer based on Conv1D:	30.09	3.89	2.17	1.1	1.2
Student:	12.42	3.07	1.23	0.45	0.3

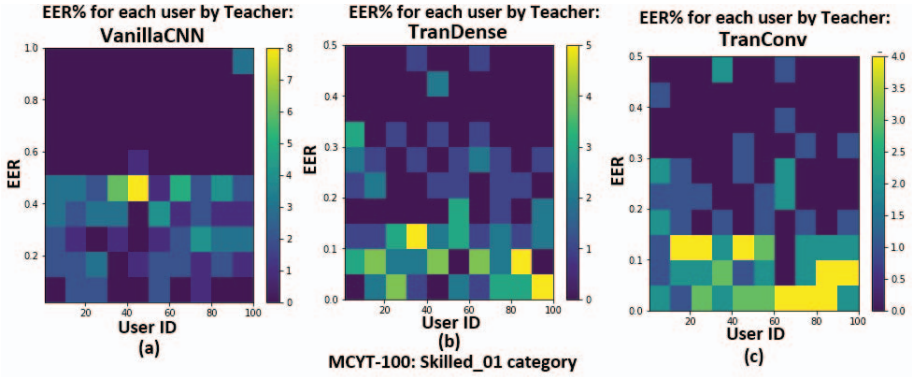

Figure 5. The histogram representing the resultant EER by three teacher networks in Skilled_01 category of MCYT-100 dataset.

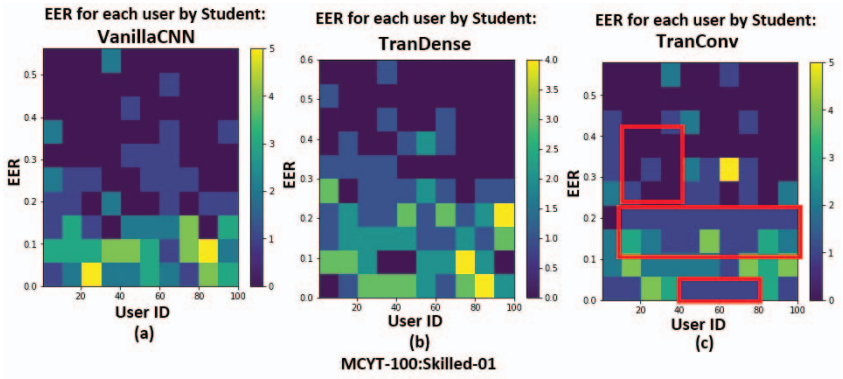

Figure 6. The histogram representing the resulting EER by corresponding student networks in Skilled_01 category of MCYT-100 dataset.

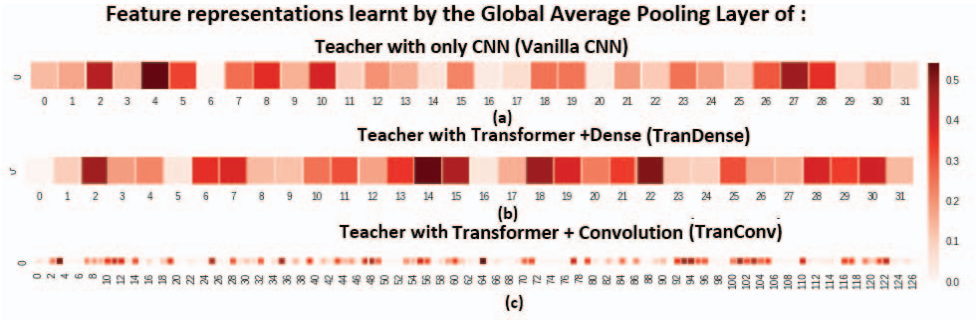

Figure 7. Illustration of the attention paid by the three types of teachers network on the outcome of GAP layer.

Table 4. Comparison of the proposed framework with the latest models on the MCYT (DB1) dataset. ”S” indicates the skilled categories. The 01,05,10,20 represents the number of signature samples used for training.

Technique	S_01	S_05	S_10	S_15	S_20	Number of Parameters
Proposed: Teacher student based OSV	**12.42***	3.07	**1.23****	**0.45****	**0.3****	3266
few shot learning (Only CNN) [33]	13.42	7.03	5.7	3.95	2.2	95101
LSTM+CNN [30]	15.57	1.88	**0.67***	0.73	**0.00***	35423
feature fusion (Only CNN) [4]	**13.38****	3.02	1.83	1.25	1.2	206277
Deep Learning + DTW [46]	-	2.40	-	-	-	5800321
several histograms [28]	-	4.02	-	-	2.72	-
VQ+DTW [35]	-	**1.55***	-	-	-	-
Stroke-Wise [10]	13.72	-	-	-	-	-
Target-Wise [10]	13.56	-	-	-	-	-
Information Divergence [41]	-	3.16	-	-	-	-
WP+BL DTW [36]	-	2.76	-	-	-	-
DTW-Normalization(F13) [11]	-	8.36	-	-	-	-
Time-series averaging+DTW [26]	-	**0.72***	-	-	-	-
Stability Modulated Dynamic Time Warping (F13) [11]	-	13.56	-	-	-	-
Curvature feature [15]	-	10.22	8.25	6.38	-	-
Torsion Feature [15]	-	9.22	7.04	5.12	-	-
Curvature + Torsion Feature [15]	-	6.05	4.23	3.10	-	-
VSA + DTW [24]	-	3.24	-	-	-	-
VSAr + DTW[24]	-	2.68	-	-	-	-
Time-series averaging+gradient boosting [25]	-	**1.28****	-	-	-	-
Down-sampling [29]	-	1.80	-	-	-	-
Two-tier ensemble [3]	-	2.84	-	-	-	-
2D Representation [49]	-	7.69	-	-	- -	10000 -
Multi scale residual[37]	-	1.38	-	-	-	-
Soft-DTW[16]	-	1.92	-	-	-	-
Gene-expression programming [40]	-	3.62	2.57	-	-	-

Table 5. Comparison of the proposed framework with the latest works on SVC dataset.

Technique	S_01	S_05	S_10	S_15	Number of Parameters
Proposed: Teacher student based OSV	6.45	1.7	0.75	0.6	3266
Few shot learning (Only CNN) [33]	**5.83****	**0.87***	**0.35****	**0.2****	95101
LSTM+CNN [30]	6.71	**1.05****	**0.00***	**0.10***	35423
Two-tier ensemble [3]	-	2.20	-	-	-
Multi scale residual[37]	-	2.33	-	-	-
Sig-2D[21]	**5.37***	-	-	-	- -
2D Representation [49]	-	5.37	-	-	10000
Gene-expression programming [40]	-	4.38	4.11	-	-
Sig-2D(without CWL)[21]	6.55	-	-	-	-

Table 6. Comparison of the proposed framework with the latest works on SUSIG dataset.

Technique	S_01	S_05	S_10	Number of Parameters
Proposed: Teacher student based OSV	**11.32****	4.85	1.0	3266
Few shot learning (Only CNN) [33]	**10.41***	**0.8****	**0.63****	95101
LSTM+CNN [30]	13.09	**1.95****	**0.20***	35423
feature fusion (Only CNN) [4]	17.96	5.17	2.07	206277
Machine Learning [38]	-	2.62	-	-

[2] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Luci, and C. Schmid. Vivit: A video vision transformer. *The International Conference on Computer Vision (ICCV)*, pages 6836—-6846, 2021.

[3] P. Bhowal, D. Banerjee, S. Malakar, and R. Sarkar. A two-tier ensemble approach for writer dependent online signature verification. *Journal of Ambient Intelligence and Humanized Computing*, 13:21—-40, 2022.

[4] V Chandra Sekhar, P Viswanath, SG Rama Krishna Sai, and M Prerana. Osvfusenet: Online signature verification by feature fusion and depthwise separable convolution based deep learning. *Neuro Computing*, 409:157–172, 2020.

[5] A. Chavan, Z. Shen, Z. Liu, Z. Liu, and E. Xing. Vision transformer slimming: Multi-dimension searching in continuous optimization space. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4921–4931, 2022.

[6] C. Chen, Q. Fan, and R. Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 357–366, 2021.

[7] P. Chu, J. Wang, and Q. You. Transmot: Spatial-temporal graph transformer for multiple object tracking. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 14870–4880, 2023.

[8] V. C.Sekhar, A.Doctor, and P.Viswanath. A light weight and hybrid deep learning model based online signature verification. *2nd Int Workshop on Machine Learning (ICDAR-WML)*, pages 53—-59, 2019.

[9] S. Dargan and M. Kumar. A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. *Expert Systems with Applications*, pages 1–22, 2020.

[10] M Diaz, A Fischer, M.A Ferrer, and R Plamondon. Dynamic signature verification system based on one real signature. *IEEE Transactions On Cybernetics*, 48:228 – 239, 2018.

[11] R Doroz, P Kudlacik, and P Porwika. Online signature verification modeled by stability oriented reference signatures. *Information Sciences*, 460:151–171, 2018.

[12] A. Dosovitskiy, L. Beyer, and A. Kolesnikov. An image is worth 16x16 words: Transformers for image recognition at scale. *The International Conference on Learning Representations (ICLR)*, pages 1–22, 2021.

[13] J. Gou, X. Xiong, and B. Yu. Channel correlation-based selective knowledge distillation. *IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS*, pages 1–12, 2022.

[14] D.S Guru and H.N. Prakash. Online signature verification and recognition: An approach based on symbolic represen-

tation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:1059–1073, 2009.

[15] L He, H Tan, and Z Huang. Online handwritten signature verification based on association of curvature and torsion feature with hausdorff distance. *springer : Multimedia Tools and Applications*, 78:253–278, 2019.

[16] J. Jiang, S. Lai, L. Jin, and Y. Zhu. Dsdtw: Local representation learning with deep soft-dtw for dynamic signature verification. *IEEE Transactions on Information Forensics and Security*, 17:2198 – 2212, 2022.

[17] Y. Kim, J. Park, Y. Jang, and M. Ali. Distilling global and local logits with densely connected relations. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6290—-6300, 2021.

[18] S Lai and L Jin. Recurrent adaptation networks for online signature verification. *IEEE Trans on Information Forensics and Security*, 14:1624–1637, 2018.

[19] S. Lai, L. Jin, L. Lin, Y. Zhu, and H. Mao. Synsig2vec: Learning representations from synthetic dynamic signatures for real-world verification. *AAAI Conference on Artificial Intelligence*, 2020.

[20] S. Lai, L. Jin, Y. Zhu, Z. Li, and Lin L. Synsig2vec: Forgery-free learning of dynamic signature representations by sigma lognormal-based synthesis and 1d cnn. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 2022.

[21] X. Liyang, W. Zhongcheng, Z. Xian, L. Yong, and W. Xinkuang. Writer-independent online signature verification based on 2d representation of time series data using triplet supervised network. *Measurement*, 80:1–28, 2022.

[22] K.S Manjunatha, S Manjunath, D.S Guru, and M.T Somashekara. Online signature verification based on writer dependent features and classifiers. *Pattern Recognition Letters*, 80:129–136, 2016.

[23] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang. Biometrics recognition using deep learning: a survey. *Artificial Intelligence Review*, pages 1–22, 2023.

[24] D Moises, A. F. Miguel, and J.Q Jose. Anthropomorphic features for on-line signatures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:2807–2819, 2019.

[25] M. Okawa. Online signature verification using single-template matching with time-series averaging and gradient boosting. *Pattern Recognition*, 102(1):1—-39, 2020.

[26] M. Okawa. Time-series averaging and local stability-weighted dynamic time warping for online signature verification. *Pattern Recognition*, 112(1):1—-39, 2020.

[27] D. Park, M. Cha, and C. Jeong. Learning student-friendly teacher networks for knowledge distillation. *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*., pages 1–12, 2021.

[28] N Sae-Bae and N Memon. Online signature verification on mobile devices. *IEEE Trans. Inf. Forensics Security*, 9:933–947, 2014.

[29] M. Saleem and B. Kovari. Online signature verification using signature down-sampling and signer-dependent sampling frequency. *Neural Computing and Applications*, 1090:1–28, 2022.

[30] V.C Sekhar, A Doctor, and P Viswanath. A light weight and hybrid deep learning model based online signature verification. In *ICDAR WML 2019 2nd International Workshop on Machine Learning*, pages 53–59, 2019.

[31] V. Sekhar, R. Sai Gorthi, and P. Viswanath. Online signature verification by few-shot separable convolution based deep learning. *15th Int Conf on Document Analysis and Recognition (ICDAR)*, pages 1125—-1129, 2019.

[32] V.Chandra Sekhar, P.Viswanath, S.S.G. Rama Krishna, and M.Prerana. Osvfusenet: Online signature verification by feature fusion and depthwise separable convolution based deep learning. *Neurocomputing*, 409(7):157–172, 2020.

[33] V.C Sekhar, R.K. Sai Gorthi, and P Viswanath. Online signature verification by few-shot separable convolution based deep learning. In *15th International Conference on Document Analysis and Recognition (ICDAR 2019)*, pages 1125–1129, 2019.

[34] V. Chandra Sekhar, P. Viswanath, M. Prerana, and S. Abhishek. Deepfuseosv: online signature verification using hybrid feature fusion and depthwise separable convolution neural network architecture. *IET Biometrics*, 9(6):259–268, 2020.

[35] A Sharma and S Sundaram. An enhanced contextual dtw based system for online signature verification using vector quantization. *Pattern Recognition Letters*, 84:22–28, 2016.

[36] A Sharma and S Sundaram. On the exploration of information from the dtw cost matrix for online signature verification. *IEEE Transactions on Cybernetics*, 48:611 – 624, 2018.

[37] Q. Shen, F. Luan, and S. Yuan. Multi-scale residual based siamese neural network for writer-independent online signature verification. *Applied Intelligence*, 52:14571—-14589, 2022.

[38] C Subhash. Verification of dynamic signature using machine learning approach. *Neural Computing and Applications*, 32:11875–11895, 2020.

[39] H. Tabani, A. Balasubramaniam, S. Marzban, E. Arani, and B. Zonooz. Improving the efficiency of transformers for resource-constrained devices. *2021 24th Euromicro Conference on Digital System Design (DSD)*, pages 449–456, 2021.

[40] H. Tan, L. He, Z. C Huang, and H. Zhan. Online signature verification based on dynamic features from gene expression programming. *Multimedia Tools and Applications*, 80:1–27, 2021.

[41] L Tang, W Kang, and Y Fang. Information divergence-based matching strategy for online signature verification. *IEEE Trans on Information Forensics and Security*, 13:861 – 873, 2018.

[42] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia. “exploring recurrent neural networks for on-line handwritten signature biometrics. *IEEE Access*, 6:5128–5138, 2018.

[43] A. Vaswani, N. Shazeer, and N. Parmar. Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*, pages 1–15, 2017.

[44] C. S. Vorugunti, B. Subramanian, A. Gautam, and V. Pulabaigari. Impact of type of convolution operation on performance of convolutional neural networks for online signature verification. *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 6290—-6300, 2022.

[45] R. Wang, J. Ao, and L. Zhou. Multi-view self-attention based transformer for speaker recognition. *The International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 6732–6736, 2022.

[46] X Wu, A Kimur, B Kenji Iwan, S Uchid, and K Kashino. Deep dynamic time warping:end-to-end local representation learning for online signature verification. In *15th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1103–1109, 2019.

[47] J. Xiang, M. Chen, P. Shan, H. Zou, S. Li, and J. Huang. Transmcgc: a recast vision transformer for small-scale image classification tasks. *Neural Computing and Applications*, pages 1–22, 2023.

[48] W. Xiaomeng, K. Akisato, I. K. Brian, U. Seiichi, and K. Kunio. Deep dynamic time warping:end-to-end local representation learning for online signature verification. *14th Int Conf on Document Analysis and Recognition (ICDAR)*, page 1103–1110, 2019.

[49] L. Xie, W. Zhong cheng, X. Zhang, Y. Li, and X. Wang. Writer-independent online signature verification based on 2d representation of time series data using triplet supervised network. *Neural Computing and Applications*, 197:21—-40, 2022.

[50] C. Xu, W. Gao, T. Li, N. Bai, G. Li, and Y. Zhang. Teacher-student collaborative knowledge distillation for image classification. *Applied Intelligence*, pages 1997—-2009, 2023.

[51] F. Yuan, Z. Zhang, and Z. Fang. An effective cnn and transformer complementary network for medical image segmentation. *Pattern Recognition*, pages 1–15, 2023.