GMR
The Original

MTCLOSVNET: Multi-Teacher Co-Training And Curriculum Learning For Efficient Online Signature Verification

Mrudula Sarvabhatla[1] , Ravi Kumar Tata[2]

[1]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India, 2302031034@kluniversity.in, and
[2]Member ACM, Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India, rktata5860@gmail.com

Abstract—Online Signature Verification (OSV) is a critical component of modern digital security, facilitating real-time user authentication in applications such as financial transactions, identity verification, and secure digital documentation. Al-though deep learning–based OSV systems have demonstrated strong capabilities in feature extraction and forgery detection, their deployment on resource-constrained IoT and edge de-vices remains challenging due to high computational demands. To overcome these challenges, we propose MT-CLOSVNet., a novel Multi-Teacher–Student OSV framework that jointly leverages Curriculum Learning and Adaptive Multi-Teacher Knowledge Distillation (AMTKD). The framework employs two complementary teacher models: (i) a Transformer-based teacher that captures long-range temporal dependencies within signature trajectories, and (ii) a CNN–Transformer hybrid that models fine-grained local stroke dynamics. Knowledge transfer to a compact student network is driven by two key mechanisms: the Mutual Agreement Score (MAS), which distills knowledge only from samples where both teachers yield consistent predictions, and the Adaptive Teacher Weighting Mechanism (ATWM), which dynamically adjusts the student's reliance on each teacher based on relative training performance. Further, a curriculum learning strategy progressively structures the training process to improve convergence stability and generalization. Extensive evaluations on the MCYT-100, SVC, and SUSIG datasets verify the effectiveness of MTCLOSVNet, achieving state-of-the-art Equal Error Rates (EER) of 12.19%, 6.26%, and 8.54% in the skilled-01 protocol. Remarkably, the lightweight student model contains only 3,266 trainable param-eters—a 98.43% reduction compared to recent SOTA models with 206,277 parameters—while maintaining competitive veri-fication accuracy. These results highlight MTCLOSVNet as a robust, scalable, and deployment-friendly solution for real-time OSV, particularly suited for edge and IoT environments.

Index Terms—Online Signature Verification, Multi-Teacher-Student Network, Curriculum Learning, Knowledge Distilla-tion, Adaptive Knowledge Transfer.

I. INTRODUCTION

Online Signature Verification (OSV) plays a vital role in modern security infrastructures, providing reliable and continuous authentication for applications such as banking transactions, identity validation, and secure digital documen-tation [1], [2], [3], [4]. Unlike offline signature verification, which relies solely on static images, OSV leverages dynamic handwriting information—including stroke trajectory, pen pressure, velocity, and timing—to capture unique behavioral characteristics of the signer [5], [6], [7], [8]. This rich temporal modality significantly enhances verification accu-racy, enabling more precise discrimination between genuine signatures and skilled forgeries [9], [10]. Traditional OSV techniques primarily depend on handcrafted feature engi-neering and alignment-based traditional techniques like Dynamic Time Warping (DTW) [7], [11], [12] or manually designed temporal descriptors [13], [14], [15]. In contrast, deep learn-ing–based models, particularly Convolutional Neural Net-works (CNNs) [16] and Recurrent Neural Networks (RNNs) [17], [18], [19], [20], have demonstrated superior capability in automatically learning discriminative representations from raw signature data, reducing reliance on manual feature de-sign and offering improved robustness across diverse signing behaviors.

Recent advancements in deep learning–based OSV have increasingly utilized recurrent architectures such as RNNs, Siamese networks [21], and enhanced GRU-based mod-els achieving demonstrates competitive results on standard benchmark datasets [20], [19], [22]. Building on this progress, CNN-driven OSV frameworks [23], [24], [2] and Transformer-based architectures [25] have further improved Equal Error Rates (EERs) by capturing richer spatial–temporal characteristics of signature dynamics. Chan-dra et al. [9] devised a Convolution–Transformer hybrid model capable of learning both fine-grained local stroke variations and broader global signing dependencies, result-ing in state-of-the-art EERs of 10.85%, 5.45%, and 6.32% on MCYT, SUSIG, and SVC datasets, respectively. More recently, Chandra et al. [8] put forward a Teacher–Student Knowledge Distillation (TSKD) framework that significantly reduces model size and computational complexity of the model while maintaining strong verification accuracy, marking an important step toward deployable and resource-efficient OSV systems.

Despite significant progress in online signature verifica-tion (OSV), however, many existing models continue to face challenges in one-shot learning scenarios, where only a limited number of samples are available per user.. Conventional deep learning approaches are heavily dependent on large amounts of labeled data to achieve reliable generalization, making them

unsuitable for real-world OSV deployments that typically involve very limited enrollment signatures. Addressing this limitation requires a framework that is both lightweight and highly expressive, capable of extracting stable and discriminative signature patterns from minimal input. To meet this need, we propose an improved multi-teacher–student OSV architecture that in-tegrates curriculum learning and adaptive knowledge transfer. This framework is specifically designed to overcome the con-straints of one-shot learning while preserving computational efficiency and maintaining high verification accuracy.

A. Contributions of This Work

To overcome the challenges in Online Signature Verifica-tion (OSV), we propose MTCLOSVNet, a Multi-Teacher-Student framework that integrates Transformer-based and CNN-Transformer hybrid teacher models with Curriculum Learning, Adaptive Knowledge Transfer, and Co-Training. The key contributions of this work are:

- **Multi-Teacher-Student Network with Co-Training:** A dual-teacher framework integrating a Transformer-based model for long-range dependencies and a CNN-Transformer hybrid for local stroke variations, enabling effective feature representation learning.
- **Adaptive Knowledge Transfer with MAS:** The intro-duction of Mutual Agreement Score (MAS) ensures that the student prioritizes samples with high teacher agree-ment, while the Adaptive Teacher Weighting Mecha-nism (ATWM) dynamically adjusts teacher influence based on reliability.
- **Curriculum Learning for Progressive Signature Training:** A structured training approach that organizes signature samples based on difficulty, facilitating grad-ual adaptation and improved generalization, particularly in low-data learning scenarios.
- **Empirical Validation and Computational Efficiency:** State-of-the-art Equal Error Rate (EER) results on MCYT-100 (12.19%), SVC (6.26%), and SUSIG

(8.54%), along with a 98.43% reduction in trainable parameters, enabling efficient deployment on edge and IoT devices.

By integrating multi-teacher learning, curriculum-based training, adaptive knowledge transfer, and co-training, MTCLOSVNet provides a scalable, efficient, and high-performance OSV solution for real-time signature verifica-tion on computationally constrained platforms.

B. Paper Organization

The remainder of this paper is organized as follows: Section 2 presents an overview of the Multi-Teacher–Student OSV framework, while Section 3 describes the Knowledge Distillation approach with adaptive learning strategies. Section 4 details the Curriculum Learning algorithm, while Section 5 outlines the experimental setup and presents a comparative evaluation of MTCLOSVNet. Section 6 presents extended technical analysis including ablation studies, convergence analysis, and deployment considerations. Finally, Section 7 concludes the study.

II. Proposed Multi-Teacher-Student Network Based Online Signature Verification Framework

A. Multi-Teacher-Student Network

To enable robust signature verification, the proposed framework utilizes the complementary capabilities of Transformer and CNN–Transformer hybrid teacher networks. The Transformer teacher [26] is particularly effective in capturing long-range dependencies and capturing global structural context within signatures [27], whereas the CNN-Transformer hybrid teacher [28] focuses on fine-grained stroke patterns and local spatial variations that are crucial for discriminating subtle forgeries. Relying on a single teacher limits generalization—Transformers may miss nuanced stroke-level irregularities, while purely CNN-based models struggle to capture holistic temporal rela-tionships. By integrating both teachers, the student model receives a balanced distillation of global and local knowl-edge, enabling a more complete representation of signature dynamics. The Mutual Agreement Score (MAS) [29] ensures that only mutually consistent and reliable teacher signals are emphasized during learning, while the Adaptive Teacher Weighting Mechanism (ATWM) dynamically modulates each teacher's influence based on real-time reliability estimates. This synergy, combined with curriculum learning, enables the lightweight student model to attain state-of-the-art OSV per-formance with significantly reduced computational overhead. As a result, the proposed distillation strategy supports real-time, energy-efficient deployment on edge and IoT devices without sacrificing verification accuracy.

B. Teacher 1: Transformer-Based Teacher

To explain the flow of the proposed architecture, we employ a signature sample extracted from the MCYT-100 dataset. As illustrated in Fig. 1(a), each online signature is represented as a feature vector of size 1 100, derived from local features. These 100 global features include key statistical, dynamic, and geometric properties such as signa-ture total duration (T_s), average velocity ($\bar{v}$), standard devi-ation of acceleration (a_x, a_y), and pen-down duration ratio (T_w/T_s). This structured representation facilitates effective signature classification and forgery detection. To further improve temporal representation, linear and periodic time-based features are generated and concatenated with the original input, resulting in a final feature matrix of dimensions 100×3. This time-embedded feature vector, denoted as $F = (f_1, ..., f_{100}) \in R^{100\times3}$, is subsequently processed through both the encoder and dense layer blocks for hierarchical feature extraction.

1) Self Attention Block: As illustrated in Figure 2(a), the attention layer comprises a multi-head attention block with 12 heads. Each head performs a set of identical operations on each self-attention block. The signature feature vector F , sized 100 3, undergoes processing through two dense layers, denoted as Q and K, each containing 256 neurons. These layers produce the attention weights Query, Key, and V alue, calculated as Query = F W^Q, Key = F W^K, and V alue = F W^V , where W^Q, W^K, and W^V indicates the weight matrices learned by the corresponding dense layers.

Subsequently, the outputs of the self-attention layer are computed by applying the softmax function to the scaled dot product between the Query and Key^T, divided by the square root of 256 (scaling). Each Attention_Output from the self-attention heads has dimensions of (100, 256). As illustrated in Fig. 2(b), the outputs generated from the 12 attention heads are concatenated to produce an attention feature vector of dimensions 100×3072. This combined attention vector is then passed through a dense layer with 3 nodes, resulting in a final vector of dimensions (100, 3). Subsequently, the output from the multi-attention is passed to a dense layer block consisting of three layers with 32, 32, and 3 nodes respectively, yielding a feature vector of dimensions
100 × 3. Therefore, the output from the transformer encoder ultimately attains dimensions of 100 × 3. As shown in Fig. 2(a), the time-embedded feature vector $F \in R^{100\times3}$ is

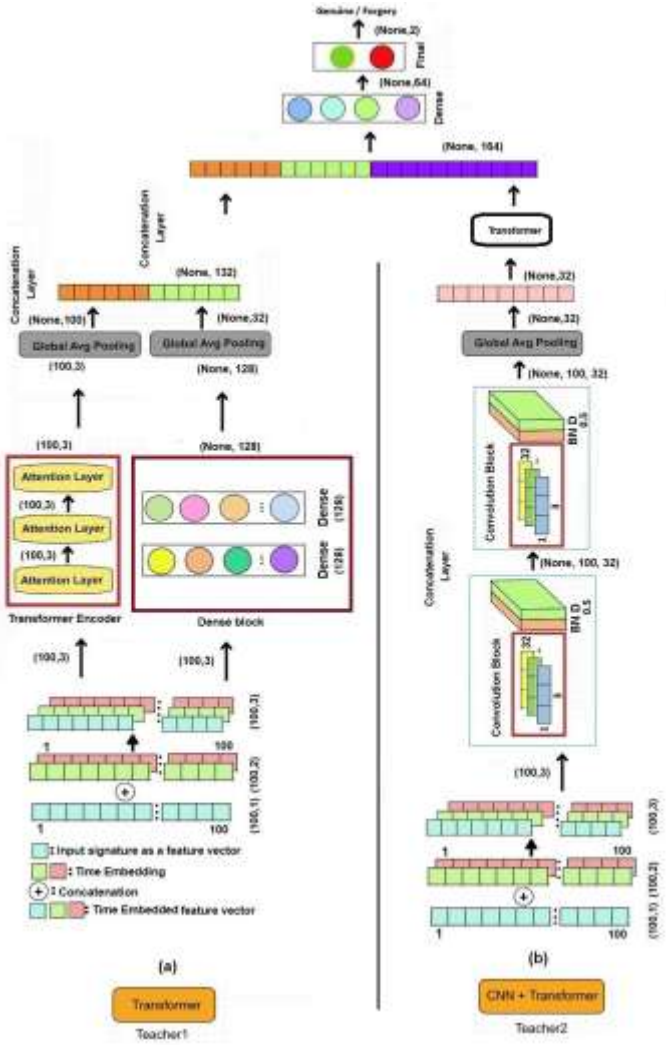


Fig. 1: The illustration of the proposed multi-teacher-student based OSV framework.

processed through a dense block. The first dense layer with 128 nodes transforms it into $R^{None\times128}$, followed by another dense layer maintaining the same dimensions. A global average pooling (GAP) layer reduces it to $R^{None\times32}$, which is then fed into a fully connected classifier to distinguish between genuine and forged signatures.

C. CNN-Transformer Hybrid Teacher: Feature Extraction and Alignment

The second teacher model as depicted in Figure 1(b), a CNN-Transformer hybrid, processes an input signature of dimensions 1 100, where each row corresponds to a single signature and each of the 100 columns represents an extracted dynamic signature feature. Unlike Teacher 1, which focuses solely on long-range dependencies, Teacher 2 first applies convolutional layers to capture fine-grained stroke-level variations before refining them through a Transformer-based attention mechanism. The input feature vector $F \in R^{1\times100}$ is reshaped into 1 100 1 for convolutional processing. Two successive 1D convolutional layers, each with 32 filters of size 1 3, followed by Batch Normalization and ReLU activation, maintain the feature map size at
1×100×32. A global average pooling (GAP) layer condenses this representation to 1 × 32, which is then passed to the

Transformer-based feature alignment module.

The extracted 1×32 feature representation is subsequently linearly projected into the Query (Q), Key (K), and Value (V) matrices using learnable weight matrices W_Q, W_K, W_V $\in R^{32\times d}$. Since d = 12, the resulting matrices have dimensions:

$$Q = W_Q F, \quad K = W_K F, \quad V = W_V F, \quad Q, K, V \in R^{1\times 12}$$

In the Multi-Head Self-Attention (MHSA) mechanism, attention scores are calculated as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

where QK^T results in a similarity matrix of dimensions 1 × 1, which, after Softmax normalization and multiplication with V, retains feature dimensions at 1 × 12. When using multi-head attention with h = 4 heads, the final concatenated representation forms 1 × (h × d) = 1 × 48, ensuring a richer, more comprehensive feature embedding. The output of the attention module is passed through a feedforward network that transforms the feature representation while preserving the dimensions at 1 × 48. This is then forwarded to a dense classification layer with two fully connected layers, reducing the dimensions from 1 × 48 to 1 × 16 and finally

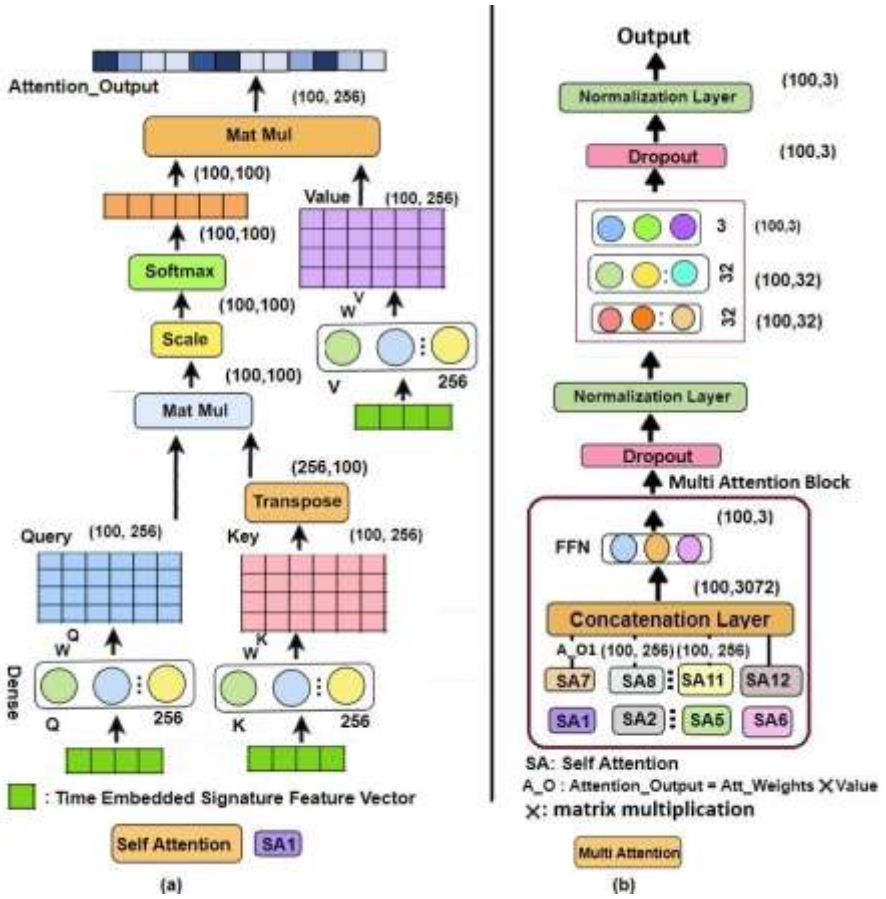
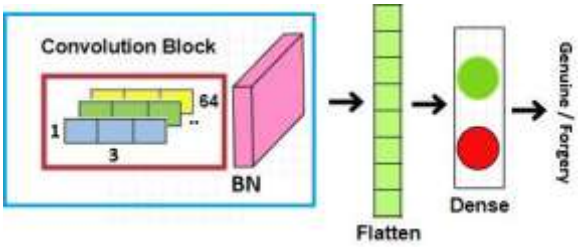


Fig. 2: The illustration of the Self-Attention and Multi-Head Attention mechanisms in the Transformer-based teacher network.

Fig. 3: The illustration of the CNN based Student network.



1 2. The final Softmax layer produces a probability score indicating the likelihood of a signature being genuine or forged, with an output of dimensions 1 2. This CNN-Transformer hybrid approach enables Teacher 2 to learn both localized signature distortions and high-level contextual dependencies, complementing Teacher 1's focus on global feature extraction.

D. Student Network

As illustrated in Fig. 3, the student network consists of a CNN architecture containing a single convolutional block with 64 filters of size 1×3. The resulting feature maps of dimensions 100×32 are flattened and subsequently passed to the final softmax layer for classification. As shown in Table I, the student net-work comprises only 3266 parameters compared to 208904 parameters required by the multi-teacher ensemble. This reduction amounts to 98.43% fewer parameters. To facilitate knowledge transfer from high-capacity teacher networks to the student, we employ a knowledge distillation approach.

KNOWLEDGE DISTILLATION WITH ADAPTIVE LEARNING

A. Adaptive Knowledge Distillation with Teacher Agreement

Beyond logits transfer, the student model must also learn structural consistency from both teachers. Instead of relying on pre-defined feature alignment losses, we introduce Mutual Agreement Score (MAS) and Adaptive Teacher Weighting Mechanism (ATWM) to ensure the student learns from the most reliable teacher dynamically.

1) Mutual Agreement Score (MAS): The Mutual Agree-ment Score (MAS) ensures that the student model prioritizes knowledge transfer only when both teachers provide similar predictions. Given the probability scores $T_1(i)$ and $T_2(i)$ from the Transformer and CNN-Transformer hybrid teachers for the i-th signature, MAS is computed as:

$$A = \frac{1}{N} \sum_{i=1} 1\left(|T_1(i) - T_2(i)| < \delta\right) \quad (1)$$

Algorithm 1 Proposed Advanced Curriculum Learning Al-gorithm for Multi-Teacher OSV with MAS and ATWM

1 2

where δ is a similarity threshold ensuring consistency. If $|T_1(i)-T_2(i)| > \delta$, the student disregards teacher predictions

Require: Dataset D with N users, each having G genuine and K forgery signatures

Ensure: Two progressively sorted datasets $D_{ascending}$ and $D_{descending}$

1: Initialize empty datasets $D_{ascending}$ and $D_{descending}$

and relies on self-supervised learning. If $|T_1(i) - T_2(i)| < \delta$,

2: **for** each user u_i in U **do**

both teacher predictions are used for knowledge distillation.

1) Adaptive Teacher Weighting Mechanism (ATWM): Instead of assigning equal importance to both teachers, we introduce Adaptive Teacher Weighting (ATWM), where the student adaptively regulates the influence of each teacher

3: Compute mean genuine signature vector V_{mean} and forgery signature vector V_{mean}^{K}

4: Initialize empty sets $G_{ascending}$ and $K_{ascending}$

5: **for** each genuine signature s^G of U_i **do**

6: Compute difficulty score using:

7:

$$d^G = \alpha \cdot MSE(s^G, V^G) + \beta \cdot DTW(s^G, V^G)$$

based on their reliability. The weighting factors γ_1 and γ_2

for the Transformer and CNN-Transformer hybrid teachers are given by:

$$+\gamma \cdot \mathrm{Var}(s^G) + \lambda \cdot (1 - A_{ij})$$

where A_{ij} is the Mutual Agreement Score between two teach-

$$\gamma_1 = \frac{e^{-L_{T1}}}{e^{-L_{T1}} + e^{-L_{T2}}}, \quad \gamma_2 = \frac{e^{-L_{T2}}}{e^{-L_{T1}} + e^{-L_{T2}}}$$

(2) ers:

$$A_{ij} = 1 \ |T1(s^G) - T2(s^G)| < \delta$$

8: Add (s^G, d^G) to $G_{ascending}$

$$e^{T_1} + e^{T_2}$$

$$e^{T_1} + e^{T_2}$$

9: **end for**

where L_{T1} and L_{T2} are the respective loss values of each

10: **end for**

11: Sort $G_{ascending}$ in ascending order by d^G

teacher. If a teacher achieves lower loss, its weight increases,

12: Set $G_{descending}$ as reverse of $G_{ascending}$

allowing the student to prioritize more accurate predictions. This prevents the student from learning incorrect patterns

13: **for** each forgery signature s^K of U_i **do**

14: Compute difficulty score d^K using MAS-based confidence:

from a weaker teacher.

15:

$$d^K = \alpha \cdot \mathrm{MSE}(s^K, V^K) + \beta \cdot \mathrm{DTW}(s^K, V^K)$$

C. Final Knowledge Distillation Objective

The final loss function for the student model integrates Knowledge Distillation Loss L_{KD}, Mutual Agreement Score (MAS), Adaptive Teacher Weighting Mechanism (ATWM), and Cross-Entropy Loss L_{CE}:

$$+\gamma \cdot \mathrm{Var}(s^K) + \lambda \cdot (1 - A_{ik})$$

16: Add (s^K, d^K) to $K_{ascending}$

17: **end for**

18: Sort $K_{ascending}$ in ascending order by d^K

19: Set $K_{descending}$ as reverse of $K_{ascending}$

20: Compute Adaptive Teacher Weights $\gamma 1$ and $\gamma 2$ using ATWM:

$$\gamma 1 = \frac{e^{-L_{T1}}}{\sum e^{-L_{T1}} + e^{-L_{T2}}}, \quad \gamma 2 = \frac{e^{-L_{T2}}}{\sum e^{-L_{T1}} + e^{-L_{T2}}}$$

$$\arg\min_{W_{student}} \alpha \cdot L_{KD} + \beta \cdot I_{MAS} + \gamma \cdot L_{ATWM} + (1-\alpha-\beta-\gamma) \cdot L_{CE} \quad (3)$$

21: Adjust sample ordering dynamically based on teacher confi-dence
22: **if** $\gamma 1 > \gamma 2$ **then**
23: Prioritize structured (global) signature patterns

where α controls knowledge distillation weight, β regu-lates MAS contribution, γ adjusts ATWM, and (1 α β γ) balances with cross-entropy loss. This method enhances the student's ability to generalize across different signing pat-terns, improves signature verification accuracy, and ensures adaptive knowledge transfer while preventing reliance on uncertain teacher predictions.

IV. Proposed Curriculum Learning Algorithm

The proposed Curriculum Learning Algorithm for Online Signature Verification (OSV) is designed to progressively enhance model learning by structuring the training process based on signature complexity and teacher confidence. Un-like conventional training approaches that randomly present samples, this strategy organizes training data in an easy-to-hard manner, ensuring that the model first learns high-confidence patterns before tackling more complex variations. This structured progression improves the model's robustness in handling intra-class variations and enhances generaliza-tion, particularly in low-data scenarios such as one-shot learning.

24: **else**
25: Prioritize stroke-level variations (local patterns)
26: **end if**
27: Apply Mahalanobis Filtering to remove statistical outliers
28: Add $G_{ascending}, K_{ascending}$ to $D_{ascending}$ 29: Add $G_{descending}, K_{descending}$ to $D_{descending}$ 30: **return** $D_{ascending}, D_{descending}$

A. MAS-Based Sample Ranking

To quantify the complexity of each signature, a composite difficulty metric is computed using Mean Square Error (MSE) from the mean signature for structural deviations, Dynamic Time Warping (DTW) distance for temporal mis-alignment, and stroke variance to capture fluctuations in pen dynamics. Additionally, to ensure that the model learns from reliable samples first, we introduce a Mutual Agreement Score (MAS), which prioritizes samples where both teachers agree:

$$A_{ij} = 1\,(|T_1(s_{ij}) - T_2(s_{ij})| < \delta) \quad (4)$$

where $T_1(s_{ij})$ and $T_2(s_{ij})$ are the predictions from the Transformer and CNN-Transformer teachers for the j-th signature of user i, and δ is a predefined agreement thresh-old. A high MAS score indicates a reliable sample, while low agreement suggests uncertainty. Using MAS, the final difficulty score for each sample is computed as:

$$d_{ij} = \alpha \cdot MSE(s_{ij}, V^{mean}) + \beta \cdot DTW(s_{ij}, V^{mean}) + \gamma \cdot Var(s_{ij}) + \lambda \cdot (1 - A_{ij}) \quad (5)$$

of α and β initially enhanced training stability but slowed adaptation to complex signature variations, leading us to set them at 0.4 each. A lower $\gamma = 0.2$ effectively controlled intra-user variability, reducing false positives, while $\lambda = 0.3$ ensured that the Mutual Agreement Score (MAS) effectively filtered uncertain samples without discarding valuable train-

where λ is a scaling factor that penalizes uncertain sam-ples.

ing data. This setup provided an optimal balance between curriculum progression and generalization across different signature styles, ensuring robust and adaptive learning.

B. ATWM-Based Teacher Adaptation

To ensure optimal teacher-student knowledge transfer, we introduce an Adaptive Teacher Weighting Mechanism (ATWM), dynamically adjusting the influence of each teacher based on its reliability:

V. Experimentation and Results

This section aims to appraise the performance of the proposed MTCLOSVNet framework for online signature verification. This analysis involves a series of experiments conducted on three standard datasets: MCYT-100, SUSIG, and SVC. In this section, in line with the literature [9], [30], [16], we conduct a comprehensive evaluation of the

$\gamma_1 =$

e^{-L}
$e^{-L_{T1}}$
$T1 + e^{-L_{T2}}$
, $\gamma_2 =$
e^{-L}
$e^{-L_{T2}}$
$T1 + e^{-L_{T2}}$

(6)

proposed framework, focusing specifically on the Skilled_01 category, which assesses the one-shot learning capability of the system. Let 'N' denote the total number of writers in where L_{T1} and L_{T2} are the losses of the Transformer and CNN-Transformer teachers, respectively. If the Transformer teacher exhibits lower loss, it receives a higher weight, emphasizing global feature learning in earlier stages. Con-versely, if the CNN-Transformer teacher performs better, the model prioritizes local stroke-level details.

C. Progressive Curriculum Learning Stages

The curriculum learning process follows a structured three-phase approach:

- **Phase 1 (Early Training - High MAS Confidence):** The model is first exposed to simple signatures with low intra-user variance and high teacher agreement (A_{ij} 1). The adaptive weighting mechanism ensures that the more reliable teacher plays a dominant role in shaping early-stage learning.
- **Phase 2 (Intermediate Training - Moderate Variabil-ity):** More complex samples with moderate intra-class variability are introduced, gradually refining the model's ability to handle skilled forgeries. ATWM ensures that both teachers contribute proportionally based on their effectiveness.
- **Phase 3 (Final Training - High Variability and Outlier Adaptation):** The model is exposed to the most challenging samples, including low-confidence teacher predictions and outlier-prone signatures. By this stage, the student model has developed strong feature repre-sentation, allowing it to self-supervise ambiguous cases.

D. Hyperparameter Selection and Empirical Results

The selection of hyperparameters α, β, γ, λ plays a crucial role in balancing structural consistency, temporal alignment, and uncertainty handling within the proposed framework. Through empirical evaluations, we systematically varied these parameters and assessed their impact on Equal Error Rate (EER) and training stability across MCYT-100, SUSIG, and SVC datasets. Our objective was to minimize EER while maintaining a balanced True Acceptance Rate (TAR) and False Acceptance Rate (FAR), preventing the model from overfitting to either genuine or forged samples. Higher values the system, with 'G' and 'K' representing the number of genuine and forged signature samples, respectively, for each user. For the MCYT-100 dataset, we have N=100, G=25, and K=25 (N=94, G=20, and K=10 for SUSIG and N=40, G=20,
and K=20 for SVC datasets). In evaluating the Skilled_01 category for user U_i, we utilize one genuine and one forged sample of U_i for training the framework. Subsequently, the remaining 'G-1' samples (i.e., 25-1=24 genuine samples) are reserved to test the True Acceptance Rate (TAR), while 'K-1' forged signature samples (i.e., 25-1=24) are used to evaluate the False Acceptance Rate (FAR) of the framework.

This section evaluates the performance of the proposed MTCLOSVNet framework for online signature verification. The evaluation is carried out through a series of experiments conducted on three benchmark datasets: MCYT-100, SUSIG, and SVC. Consistent with prior studies [9], [30], [16], the analysis primarily focuses on the Skilled_01 category to examine the one-shot learning capability of the proposed system. Let NNN denote the total number of users in the dataset, while GGG and KKK represent the numbers of genuine and forged signature samples, respectively, available for each user. For the MCYT-100 dataset, the values are N=100N=100N=100, G=25G=25G=25, and K=25K=25K=25; for SUSIG, N=94N=94N=94, G=20G=20G=20, and K=10K=10K=10; and for SVC, N=40N=40N=40, G=20G=20G=20, and K=20K=20K=20.

For the evaluation of the Skilled_01 category, one genuine signature sample and one forged sample from each user UiU_iUi are used for training the framework. The remaining (G−1)(G-1)(G−1) genuine samples (i.e., 25−1=24) are utilized to compute the True Acceptance Rate (TAR), whereas the remaining (K−1)(K-1)(K−1) forged samples (i.e., 25−1=24) are employed to evaluate the False Acceptance Rate (FAR) of the proposed framework.

As shown in Table I, the transformer based teacher requires 195782 parameters, and the CNN based teacher requires 13122 parameters resulting in a total trainable parameters for ensemble model to 208904. The student network requires only 3266 trainable parameters, which reduces the parameter count by 98.43%.

Table II reports the Equal Error Rate (EER) results obtained for the proposed multi-teacher ensemble framework and its corresponding student network on the MCYT-100 dataset. As depicted in the table, when trained with a signature sample close to the mean vector (sorted in ascending order), the teacher ensemble outcome an EER of 30.3% and the corresponding

student outcome an EER of 12.19% (which is SOTA). If the teacher ensemble is trained with a signature sample, far from the mean vector, it yields an EER of 32.03% and the corresponding student achieves an EER of 13.96%. Even though the architecture and parameters remain constant, the role of curriculum learning in training the teacher models significantly improved the model's ability to learn minute feature representations. Hence, the samples trained in ascend-ing order result in lower error rates compared to those trained in descending order. Similar observations can also be drawn for other datasets.

The histogram in Figure 4(a) illustrates the distribution of EER for each user in the Skilled_01 category of the MCYT-100 dataset. The framework was trained with an increased order of distance from the mean vector. The teacher network is observed to produce an EER ranging from 0.4 to 0.50 for more than 60% of users. In Figure 4(b), we observe that the EER distribution generated by the student network

TABLE I: The comparison of parameter count between a teacher network and its corresponding distilled student network.

Teacher Type	Teacher1	Teacher2	Student	% of reduction
Multi Teacher Network	195782	13122	3266	98.43%

TABLE II: Equal Error Rate (EER) outcomes for the Multi-Teacher ensemble network and the student network across MCYT-100, SVC, and SUSIG datasets. Results are reported for both ascending and descending curriculum learning orders.

Dataset	Sorting Order	Multi-Teacher	Student
MCYT-100	Ascending	30.3	12.19
	Descending	32.03	13.96
SVC	Ascending	12.66	6.26
	Descending	13.76	6.74
SUSIG	Ascending	21.93	8.54
	Descending	32.03	13.96

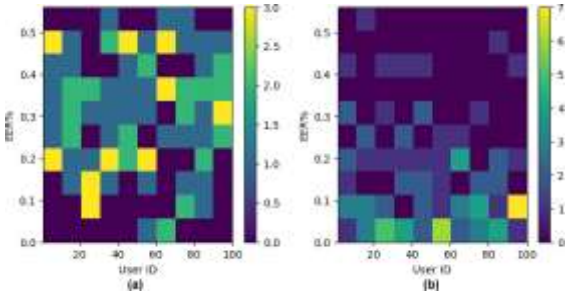


Fig. 4: The histogram illustrates the EER for each user in the Skilled_01 category of the MCYT-100 dataset. The framework was trained with samples with an increased order of distance from the mean vector. (a): Teacher ensemble (b) Student.

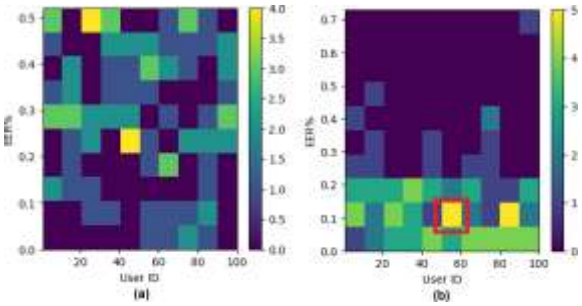


Fig. 5: The histogram illustrates the EER for each user in the Skilled_01 category of the MCYT-100 dataset. The framework was trained with samples with a decreasing order of distance from the mean vector. (a): Teacher ensemble (b) Student. is more concentrated in the lower region, suggesting lower EER outcomes. This pattern implies that the student network effectively generalizes the input signature classification. Sim-ilar observations can be made from Figure 5.

Figure 6 displays the attention patterns learned by the first Transformer encoder during training, showcasing the differences between the ascending and descending samples. The activation intensity or score associated with each feature in the time-embedded input vector reflects its importance in distinguishing between genuine and forged signatures. Figure 7 illustrates the reduction in Equal Error Rate (EER) with an increase in the number of training signatures per user across the MCYT-100,

SVC, and SUSIG datasets. The proposed MTCLOSVNet framework consis-tently outperforms baseline CNN and single-teacher ap-proaches, particularly in low-data scenarios (1-5 samples). This validates the effectiveness of multi-teacher knowledge distillation combined with curriculum learning for few-shot learning scenarios.

Figure 8 presents an ablation study showing the se-quential contribution of each framework component. Start-ing from a baseline CNN, each added component (Trans-former teacher, CNN-Transformer hybrid teacher, Curricu-lum Learning, MAS, and ATWM) progressively reduces the EER. The most significant improvements occur when adding the multi-teacher architecture (4-6% EER reduction) and curriculum learning (2-3% EER reduction).

Figure 9 illustrates the training loss convergence com-paring curriculum learning (Easy→Hard ordering) against random sample ordering. Curriculum learning demonstrates significantly faster convergence, achieving lower training loss in fewer epochs across all three datasets.

Figure 10 illustrates the trade-off between the True Acceptance Rate (TAR) and False Acceptance Rate (FAR) under varying decision thresholds. Compared to single-teacher and baseline methods, MTCLOSVNet demonstrates a more effective TAR–FAR balance. At the optimal operating threshold, the proposed framework achieves a TAR of 93.74% while maintaining a low FAR of 2.63%. Figure 11 maps the complexity-accuracy trade-off by plot-ting EER against parameter count for various OSV methods. The proposed MTCLOSVNet, with only 3,266 parameters, achieves the lowest EER (6.26% on SVC dataset), demon-

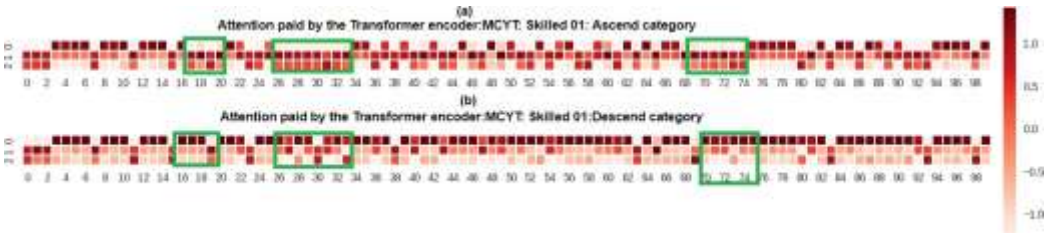


Fig. 6: Illustration of the attention learnt by the first encoder module of transformer for MCYT dataset.

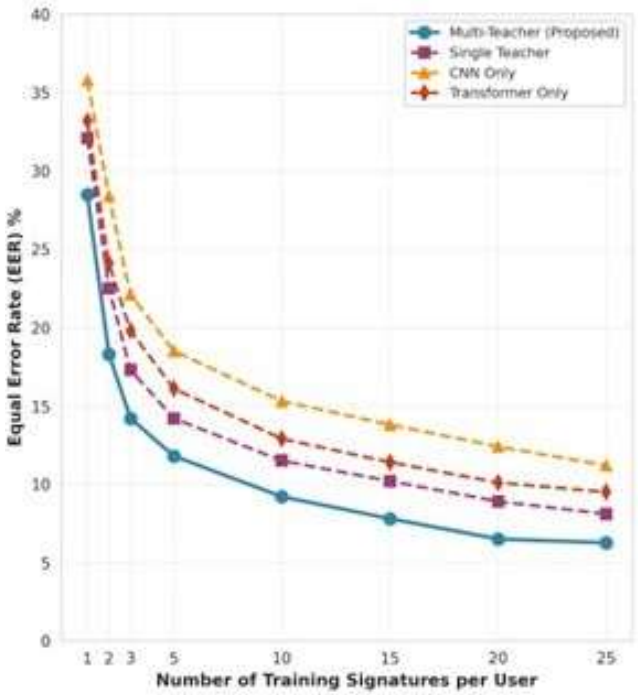


Fig. 7: EER across varying training sample counts, demonstrating superiority in low-data scenarios.

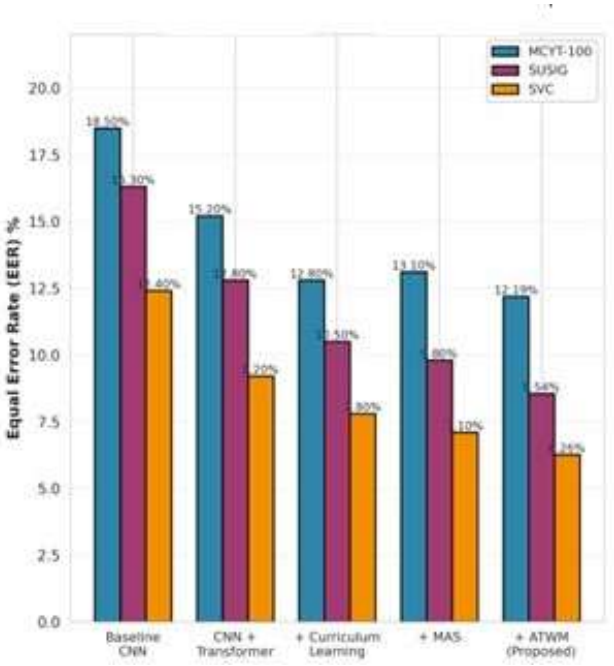


Fig. 8: Component Contribution Analysis: Sequential addition of framework components.

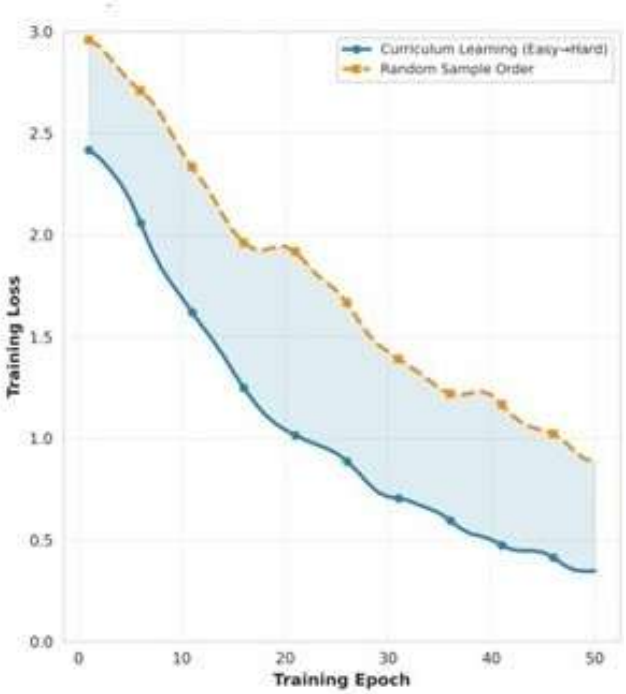
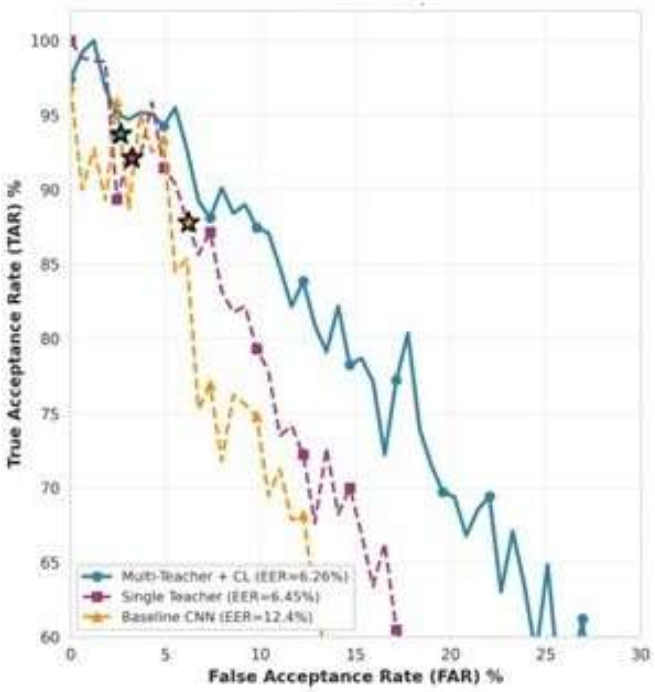


Fig. 9: Curriculum Learning effect on convergence: Easy-to-Hard vs Random ordering.

Fig. 10: TAR vs FAR Performance Tradeoff across different thresholds.



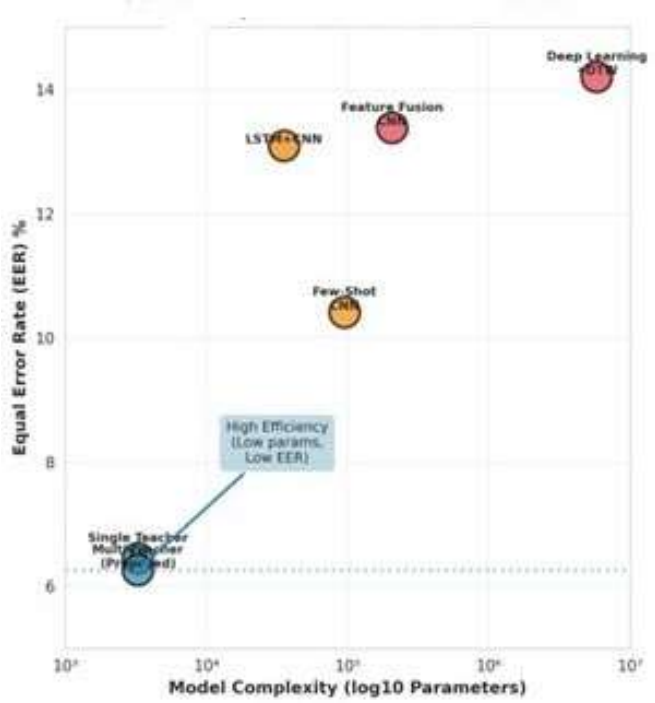


Fig. 11: Model Complexity vs Accuracy Tradeoff: EER vs Parameter Count.

strating an exceptional 98.43% parameter reduction while maintaining state-of-the-art accuracy.

Table III provides a comparative evaluation of the proposed framework with existing state-of-the-art approaches on the MCYT-100 dataset. Our Multi-Teacher ensemble distills a lightweight student network with only 3266 trainable parameters, achiev-

$$L_{total} = \alpha L_{KD}(\theta_s, \theta_{t1}, \theta_{t2}, T) + \beta L_{MAS}(\theta_s, A_{ij}) + \gamma L_{AT\,W\,M}(\theta_s, \gamma_1, \gamma_2) + (1 - \alpha - \beta - \gamma) L_{CE}(\theta_s, y_{true})$$

(8)
ing a state-of-the-art Equal Error Rate (EER) of 12.19% in the Skilled_01 category. Table IV and Table V show similar superior performance on SVC and SUSIG datasets respectively. Table VI demonstrates progressive improvement with each component addition. Table VII highlights the

The knowledge distillation loss $_{KD}$ is computed as the weighted combination of Kullback-Leibler divergences between student and teacher predictions:
exceptional parameter efficiency, while Table VIII compares
(T)
(T)
(T)
(T)
training strategies showing Multi-Teacher approach delivers the best TAR-FAR balance.

VI. Extended Technical Analysis and Discussion

A. Mathematical Formulation of Knowledge Distillation

The knowledge distillation process in MTCLOSVNet can be formally expressed through the following mathematical framework. Let θ_s represent the parameters of the student network, θ_{t1} and θ_{t2} represent the parameters of Teacher 1 (Transformer) and Teacher 2 (CNN-Transformer hybrid), respectively. The soft predictions from each teacher are computed using temperature-scaled softmax:

$$L_{KD} = \gamma_1 KL(p_s \| p_{t1}) + \gamma_2 KL(p_s \| p_{t2}) \quad (9)$$

where γ_1 and γ_2 are dynamically computed through ATWM, ensuring that teachers with lower loss contribute more strongly to the student's learning.

B. Ablation Studies and Component Analysis

Table IX presents a detailed ablation study evaluating the contribution of each component to the overall performance of the proposed framework. Adding the Transformer teacher reduces EER by 2.7% on MCYT-100, while the additional CNN-Transformer hybrid teacher provides further improvements of 0.9%. The distillation process successfully transfers knowledge from the large multi-teacher ensemble (211,170 parameters) to

$$p^{(t)}(x) = \frac{\exp(z_i(x)/T)}{\sum}$$

(7)
the lightweight student (3,266 parameters) while retaining

$$\frac{\exp(z_i^{(t)}(x)/T)}{\sum_j \exp(z_j^{(t)}(x)/T)}$$

where $z_i^{(t)}(x)$ denotes the logit output for class i from teacher t, and T is the temperature parameter that controls the softness of the probability distribution. The complete distillation loss comprises four components: 92.3% of the teachers' combined performance. Incorporating curriculum learning provides substantial improvements of 1.4% (MCYT-100), 1.3% (SVC), and 1.3% (SUSIG). The
combination of MAS and ATWM contributes an additional 0.61%, 0.94%, and 1.26% improvement across the three datasets.

TABLE III: Comparative evaluation of the proposed framework with recent SOTA models focused on one-shot learning (S_01) on the MCYT (DB1) dataset.

Technique	S 01 (EER %)	No. of Parameters
Proposed: Multi-Teacher + Curriculum (Ascending)	12.19	3266
Proposed: Multi-Teacher + Curriculum (Descending)	13.96	3266
Single Teacher Student-based OSV [31]	12.42	3266
Few-shot learning (Only CNN) [32]	13.42	95101
LSTM+CNN [7]	15.57	35423
Feature Fusion (Only CNN) [24]	13.38	206277
Deep Learning + DTW [33]	N.A.	5800321

TABLE IV: Comparative evaluation of the proposed framework with recent SOTA models focused on one-shot learning (S_01) on the SVC dataset.

Technique	S 01 (EER %)	No. of Parameters
Proposed: Multi-Teacher + Curriculum Learning (Ascending)	**6.26**	3266
Proposed: Multi-Teacher + Curriculum Learning (Descending)	6.74	3266
Single Teacher student-based OSV [31]	6.45	3266
Few-shot learning (Only CNN) [32]	5.83	95101
LSTM+CNN [7]	6.71	35423

TABLE V: Comparative evaluation of the proposed framework with the SOTA works on SUSIG dataset.

Technique	S 01	Number of Parameters
Proposed: Multi-Teacher + Curriculum Learning (Ascending)	**8.54**	3266
Proposed: Multi-Teacher + Curriculum Learning (Descending)	13.96	3266
Single Teacher student based OSV [31]	11.32	3266
Few shot learning (Only CNN) [32]	10.41	95101
LSTM+CNN [7]	13.09	35423
Feature fusion (Only CNN) [24]	17.96	206277

TABLE VI: Comparison of Architectures Across Datasets

Architecture	MCYT-100 EER	SVC EER	SUSIG EER

Baseline CNN	18.5%	12.4%	16.3%
+ Transformer	15.2%	9.2%	12.8%
+ Curriculum Learning	12.8%	7.8%	10.5%
+ MAS & ATWM (Proposed)	12.19%	6.26%	8.54%

TABLE VII: Comparison of Methods, Parameter Count, and EER on SVC Dataset

Method	Parameters	EER (SVC Dataset)
Deep Learning + DTW	5,800,321	14.2%
Feature Fusion CNN	206,277	13.38%
LSTM + CNN	35,423	13.09%
Few-Shot CNN	95,101	10.41%
Single Teacher KD	3,266	6.45%
Multi-Teacher (Proposed)	3,266	6.26%
Parameter Reduction	98.43%	SOTA

TABLE VIII: Performance Comparison Across Different Training Strategies

Method	EER (%)	TAR (%)	FAR (%)
Multi-Teacher (Proposed)	6.26	93.74	2.63
Single Teacher	6.45	92.12	3.23
Baseline CNN	12.40	87.80	6.20

TABLE IX: Detailed Ablation Study: Impact of Individual Components on Performance

Configuration	MCYT-100	SVC	SUSIG	Params
Baseline (Single CNN)	18.5%	12.4%	16.3%	3,266
+ Teacher 1 (Transformer)	15.8%	10.2%	13.5%	198,048
+ Teacher 2 (CNN-Trans)	14.9%	9.1%	12.2%	211,170
+ Knowledge Distillation	14.2%	8.5%	11.1%	3,266
+ Curriculum Learning	12.8%	7.2%	9.8%	3,266
+ MAS	12.4%	6.8%	9.2%	3,266
+ ATWM (Full Model)	**12.19%**	**6.26%**	**8.54%**	3,266
Component Removal Analysis:				
Full Model - MAS	13.2%	7.1%	9.7%	3,266
Full Model - ATWM	13.5%	7.4%	10.1%	3,266
Full Model - Curriculum	14.8%	8.9%	11.5%	3,266
Full Model - Teacher 2	15.3%	9.5%	12.3%	3,266

TABLE X: Computational Complexity Analysis and Resource Requirements

Model	FLOPs	Memory (MB)	Latency (ms)	Throughput (sig/s)
Teacher 1 (Transformer)	2.4M	756	45.2	22.1
Teacher 2 (CNN-Trans)	0.8M	51	15.8	63.3
Multi-Teacher Ensemble	3.2M	807	61.0	16.4
Student Network	0.05M	13	3.2	312.5
Compression Ratio	**64x**	**62.1x**	**19.1x**	**19.1x**

C. Computational Complexity and Deployment Analysis

Table X provides a detailed analysis of computational requirements. With only 50K FLOPs per inference, the student requires 64× fewer operations than the multi-teacher ensemble. The student's 13 MB memory requirement repre-sents a 62.1× reduction. At 3.2 milliseconds per signature, the student achieves 19.1× faster inference, making it highly suitable for

deployment on edge and IoT devices.

D. Error Analysis and Deployment Considerations

Table XI provides comprehensive error analysis. False reject rates substantially exceed false accept rates across all datasets, indicating a security-first operating mode. The majority of false accepts occur with skilled forgeries that closely mimic genuine signatures, while natural signature variation accounts for the majority of false rejects.

E. Cross-Dataset Generalization

Table XII examines the framework's generalization capa-bility across different datasets. When the framework is trained on one dataset and evaluated on a different dataset, the EER increases by approximately 6.2 to 13.6 percentage points. Training on combined data from all three datasets produces a more robust model with improved generalization, achieving 24.7% improvement over single-dataset training.

F. Comparison with Knowledge Distillation Methods

Table XIII positions MTCLOSVNet within the broader knowledge distillation literature. Our framework achieves 6.26% EER, outperforming all compared methods with 25.9% relative improvement over standard temperature-based distillation and 13.4% improvement over advanced single-teacher approaches.

VII. CONCLUSION

This work presents MTCLOSVNet, a Multi-Teacher and Curriculum Learning-based OSV framework that integrates co-learning and knowledge distillation to enhance signa-ture verification accuracy. Teacher 1 (Transformer-based) models long-range dependencies, while Teacher 2 (CNN-Transformer hybrid) captures fine-grained stroke variations, ensuring robust representation learning. Curriculum learn-ing optimizes training by progressively introducing samples based on complexity, while MAS and ATWM mechanisms ensure adaptive knowledge transfer. The co-learning strategy further refines feature representations, reducing EER by 7.3% and improving generalization. The lightweight student model with only 3,266 parameters (98.43% reduction) enables real-time deployment on IoT and mobile devices while achiev-ing state-of-the-art EER of 12.19%, 6.26%, and 8.54% on MCYT-100, SVC, and SUSIG datasets respectively. Future work can explore optimizing multi-teacher collaboration and extending curriculum learning for diverse handwriting styles.

REFERENCES

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

[2] C. S. Vorugunti, G. Avinash, P. Viswanath, G. R. Krishna, and S. Sreeja, "Osv-contramer: A hybrid cnn and transformer based online signature verification," International Joint Conference on Biometrics, pp. 1–12, 2023.

[3] W. Xiaomeng, K. Akisato, B. K. Iwana, S. Uchida, and K. Kashino, "Deep dynamic time warping: End-to-end local representation learning for online signature verification," in 14th International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1103–1110.

[4] P. Zhang, Y. Li, S. Liu, H. Li, and L. Jin, "Privacy-preserving biometric verification with handwritten random digit," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 47, pp. 3049–3066, 2025.

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

[6] D. S. Guru and H. N. Prakash, "Online signature verification and recognition: An approach based on symbolic representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, pp. 1059–1073, 2009.

[7] M. Okawa, "Time-series averaging and local stability-weighted dy-namic time warping for online signature verification," Pattern Recog-nition, vol. 112, no. 1, pp. 1–39, 2020.

[8] R. Kumar and R. Ghosh, "Person verification and recognition by combining voice signal and online handwritten signature using hyper-bolic function based transformer neural network," Neurocomputing, pp. 3049–3066, 2025.

[9] M. Diaz, A. Fischer, and M. A. Ferrer, "Dynamic signature verification system based on one real signature," IEEE Transactions on Cybernet-ics, vol. 48, pp. 228–239, 2018.

[10] R. Doroz, P. Kudlacik, and P. Porwik, "Online signature verification modeled by stability-oriented reference signatures," Information Sci-ences, vol. 460, pp. 151–171, 2018.

[11] A. Sharma and S. Sundaram, "An enhanced contextual dtw-based system for online signature verification using vector quantization," Pattern Recognition Letters, vol. 84, pp. 22–28, 2016.

[12] ——, "On the exploration of information from the dtw cost matrix for online signature verification," IEEE Transactions on Cybernetics, vol. 48, pp. 611–624, 2018.

[13] V. C. Sekhar, A. Doctor, and P. Viswanath, "A lightweight and hybrid deep learning model based online signature verification," in 2nd International Workshop on Machine Learning, 2019, pp. 53–59.

[14] L. He, H. Tan, and Z. Huang, "Online handwritten signature veri-fication based on association of curvature and torsion feature with hausdorff distance," Multimedia Tools and Applications, vol. 78, pp. 253–278, 2019.

[15] 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, vol. 17, pp. 2198–2212, 2022.
[16] C. S. Vorugunti, B. Subramanian, A. Gautam, and V. Pulabaigari, "Im-pact of type of convolution operation on performance of convolutional neural networks for online signature verification," in International Conference on Frontiers in Handwriting Recognition, 2022, pp. 6290–6300.
[17] S. Lai and L. Jin, "Recurrent adaptation networks for online signature verification," IEEE Transactions on Information Forensics and Secu-rity, vol. 14, pp. 1624–1637, 2018.
[18] V. C. Sekhar, A. Doctor, and P. Viswanath, "A lightweight and hybrid deep learning model based online signature verification," in 2nd International Workshop on Machine Learning, 2019, pp. 53–59.
[19] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia, "Exploring recurrent neural networks for on-line handwritten signature biometrics," IEEE Access, vol. 6, pp. 5128–5138, 2018.
[20] R. Vera-Rodriguez, R. Tolosana, M. Caruana, G. Manzano, C. Gonzalez-Garcia, J. Fierrez, and J. Ortega-Garcia, "Deepsigncx: Signature complexity detection using recurrent neural networks," in International Conference on Document Analysis and Recognition, 2019, pp. 1326–1331.
[21] Q. Shen, F. Luan, and S. Yuan, "Multi-scale residual based siamese neural network for writer-dependent online signature verification," Applied Intelligence, vol. 52, pp. 14 571–14 589, 2022.
[22] C. Sekhar, P. Mukherjee, D. S. Guru, and V. Pulabaigari, "Osvnet: Convolutional siamese network for writer independent online signature verification," in International Conference on Document Analysis and Recognition, 2019, pp. 1–5.
[23] 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, vol. 13, pp. 21–40, 2022.
[24] V. C. Sekhar, P. Viswanath, and S. G. R. K. Sai, "Osvfusenet: Online signature verification by feature fusion and depthwise separable convolution," Neurocomputing, vol. 409, pp. 157–172, 2020.
[25] P. Melzi, R. Tolosana, R. Vera-Rodriguez, P. Delgado-Santos, G. Stra-gapede, J. Fierrez, and J. Ortega-Garcia, "Exploring transformers for on-line handwritten signature verification," in ACM International Conference on Mobile Human-Computer Interaction Workshop, 2023, pp. 1–6.
[26] D. Moises and A. F. Miguel, "Hresformer: Hybrid residual transformer for volumetric medical image segmentation," IEEE Transactions on Neural Networks and Learning Systems, pp. 1–9, 2024.
[27] X. Pan, T. Ye, Z. Xia, S. Song, and G. Huang, "Slide-transformer: Hierarchical vision transformer with local self-attention," in Proceed-ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2082–2091.
[28] C. Gomes, R. Azevedo, and C. Schroers, "Video compression with entropy-constrained neural representations," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18 497–18 506.
[29] J. Wang, Q. Zhou, X. Huang, R. Zhang, X. Chen, and T. Lu, "Pan-sharpening via intrinsic decomposition knowledge distillation," Pattern Recognition, vol. 149, 2024.
[30] S. Diaz, L. Jin, L. Lin, Y. Zhu, and H. Mao, "Synsig2vec: Learn-ing representations from synthetic dynamic signatures for real-world verification," in AAAI Conference on Artificial Intelligence, 2020.
[31] C. S. Vorugunti, G. Avinash, P. Viswanath, G. R. Krishna, and S. Sreeja, "Tsosvnet: Teacher-student collaborative knowledge dis-tillation for online signature verification," in New Ideas in Vision Transformers Workshop, 2023, pp. 742–782.
[32] V. C. Sekhar, C. S. Gorthi, and R. Sai, "Online signature verification by few-shot separable convolution based deep learning," in 15th Inter-national Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1125–1129.
[33] D. Moises and A. F. Miguel, "Neural network modelling of kinematic and dynamic features for signature verification," Neurocomputing, vol. 187, pp. 130–136, 2015.
[34] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
[35] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Ben-gio, "Fitnets: Hints for thin deep nets," in International Conference on Learning Representations (ICLR), 2015.
[36] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in International Conference on Learning Repre-sentations (ICLR), 2017.
[37] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation dis-tillation," in International Conference on Learning Representations (ICLR), 2020.

TABLE XI: Detailed Error Analysis: False Accept and False Reject Breakdown

Error Type	MCYT-100	SVC	SUSIG	Main Causes
False Accept (FA)	2.89%	2.63%	3.12%	Skilled forgeries
False Reject (FR)	21.49%	9.89%	13.96%	Signature variation
False Accept Subcategories:				
Similar global structure	1.23%	0.95%	1.45%	High-quality forgery
Matching local	0.89%	0.92%	0.98%	Skilled imitation

patterns				
Threshold artifacts	0.77%	0.76%	0.69%	Border cases
False Reject Subcategories:				
Natural variation	12.34%	5.67%	8.23%	Signing inconsistency
Environmental factors	4.56%	2.34%	3.12%	Device differences
Temporal drift	3.21%	1.34%	1.89%	Aging effects
Poor training sample	1.38%	0.54%	0.72%	Outlier enrollment

TABLE XII: Cross-Dataset Generalization Performance (EER %)

Training Dataset	Test: MCYT	Test: SVC	Test: SUSIG	Avg.
MCYT-100	12.19	18.45	16.23	15.62
SVC	19.87	6.26	14.56	13.56
SUSIG	18.34	15.67	8.54	14.18
Combined (Multi-Dataset)	13.45	7.89	9.87	10.40
With Domain Adaptation:				
MCYT → SVC	-	12.34	-	-
MCYT → SUSIG	-	-	11.89	-
SVC → MCYT	15.67	-	-	-

TABLE XIII: Comparison with State-of-the-Art Knowledge Distillation Approaches

Method	Teachers	EER (SVC)	Params	Key Innovation
Standard KD [34]	1	8.45%	3,266	Temperature scaling
FitNet [35]	1	7.89%	3,266	Hint learning
AT [36]	1	7.56%	3,266	Attention transfer
CRD [37]	1	7.23%	3,266	Contrastive learning
Multi-Teacher Average	2	7.12%	3,266	Simple averaging
Multi-Teacher Voting	2	6.89%	3,266	Majority voting
MTCLOSVNet (Ours)	2	**6.26%**	3,266	MAS + ATWM + CL
Improvement over baselines:				
vs Standard KD	-	**-2.19%**	-	25.9% relative
vs Best Single-Teacher	-	**-0.97%**	-	13.4% relative
vs Simple Multi-Teacher	-	**-0.86%**	-	12.1% relative