

# DetailSemNet: Elevating Signature Verification through Detail-Semantic Integration

Meng-Cheng Shih<sup>1</sup>, Tsai-Ling Huang<sup>1</sup>, Yu-Heng Shih<sup>1</sup>, Hong-Han Shuai<sup>1</sup>, Hsuan-Tung Liu<sup>2</sup>, Yi-Ren Yeh<sup>3</sup>, and Ching-Chun Huang<sup>1</sup>

<sup>1</sup> National Yang Ming Chiao Tung University, Taiwan

{mcshih.ee11, christina.ii12, ra890927.cs12, hhshuai, chingchun}@nycu.edu.tw

<sup>2</sup> E.SUN Financial Holding Co., Ltd, Taiwan ahare-18342@esunbank.com.tw

<sup>3</sup> National Kaohsiung Normal University, Taiwan yryeh@nknu.edu.tw

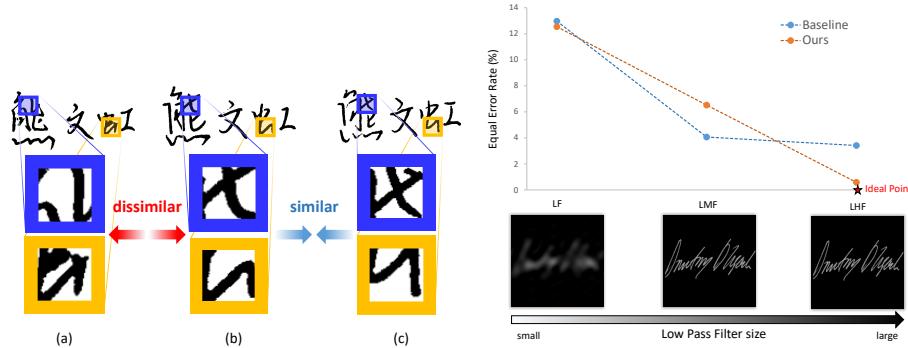
[https://github.com/nycu-acm/DetailSemNet\\_OSV](https://github.com/nycu-acm/DetailSemNet_OSV)

**Abstract.** Offline signature verification (OSV) is a frequently utilized technology in forensics. This paper proposes a new model, **DetailSemNet**, for OSV. Unlike previous methods that rely on holistic features for pair comparisons, our approach underscores the significance of fine-grained differences for robust OSV. We propose to match local structures between two signature images, significantly boosting verification accuracy. Furthermore, we observe that without specific architectural modifications, transformer-based backbones might naturally obscure local details, adversely impacting OSV performance. To address this, we introduce a **Detail-Semantics Integrator**, leveraging feature disentanglement and re-entanglement. This integrator is specifically designed to enhance intricate details while simultaneously expanding discriminative semantics, thereby augmenting the efficacy of local structural matching. We evaluate our method against leading benchmarks in offline signature verification. Our model consistently outperforms recent methods, achieving state-of-the-art results with clear margins. The emphasis on local structure matching not only improves performance but also enhances the model's interpretability, supporting our findings. Additionally, our model demonstrates remarkable generalization capabilities in cross-dataset testing scenarios. The combination of generalizability and interpretability significantly bolsters the potential of **DetailSemNet** for real-world applications.

**Keywords:** Offline Signature Verification · Feature Disentanglement · Local Matching

## 1 Introduction

Handwritten offline signature verification (OSV) is a pivotal biometric technology, especially in sectors like banking and commerce. The core goal of this technology is to authenticate a signature by comparing it against a known original. This comparison involves analyzing a test signature alongside a reference signature, allowing the system to determine whether the test image is a forgery



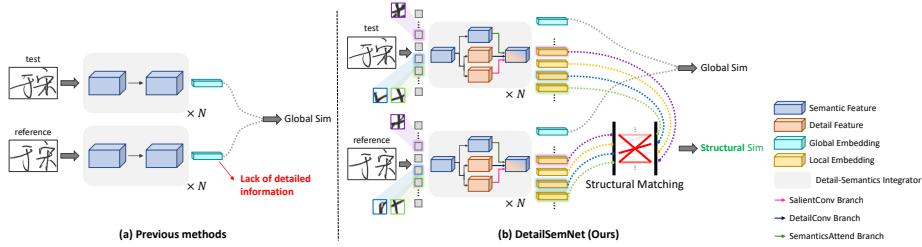
**Fig. 1:** Three samples from the ChiSig dataset. Signature (a) originates from a different individual than signatures (b) and (c). At first glance, these signatures appear remarkably similar when viewed holistically. However, detailed analysis at the patch level reveals distinct differences between them, which are frequently overlooked in previous methodologies.

**Fig. 2:** We employ filters to extract both semantic pattern (low-frequency) and stroke structure and style detail (high-frequency) for improved verification. Leveraging high-frequency data enhances performance, unlike the baseline transformer model, which solely relies on low-frequency patterns and does not benefit from high-frequency features.

or genuine. Such critical assessments of authenticity are vital in maintaining security and trust in various applications [40].

Signature verification is challenging due to several factors. First, people have unique ways of signing. Second, there's often not enough detailed information about how each signature stroke is made. Finally, sophisticated forgeries can be hard to distinguish from genuine signatures. The essence of signature verification lies in comparing the similarity of the subtle stylistic characteristics concealed within the reference signature and the testing one rather than focusing on the specific contents of the signatures [40].

Traditional approaches for OSV heavily rely on manual handcrafted feature engineering [7, 13]. In recent years, numerous deep-learning methods have been proposed [40]. These deep-learning methods have demonstrated significant advancements in verification performance when compared to traditional handcrafted features. However, some key issues still need to be well addressed by previous deep-learning methods. Despite the importance of the reasoning process that considers the similarity between global features from the holistic signature image. They lack the incorporation of structural comparison among local patch features to measure similarity. A global representation destroys image structures and leads to the loss of local information. Local features (stroke structure and style) offer discriminative and transferable information for offline signature verification (Fig. 1). Hence, a desirable metric-based OSV should possess the ca-



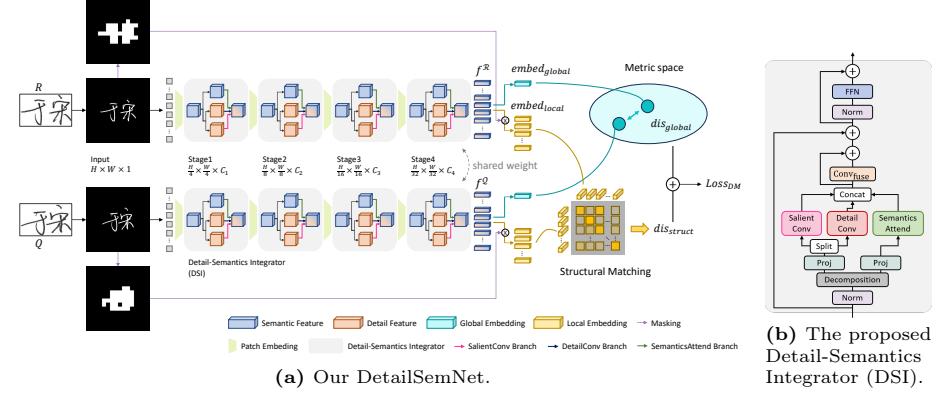
**Fig. 3:** Conventional OSV method vs. Our proposed method: The left figure shows the traditional approach, lacking detailed feature information and relying solely on global similarity for comparison. On the right, our method, called **DetailSemNet**, employs the **Detail-Semantics Integrator** to divide features into Semantic and Detail components. The Semantic component acquires contextual information through the SemanticsAttend Branch, while the Detail component is processed via the SalientConv and DetailConv Branches. Integrating these outputs yields feature representations containing both detailed and semantic information. Additionally, the model utilizes Structural Matching techniques to emphasize detailed information alongside global similarity.

pability to leverage local discriminative representations for metric learning while minimizing the influence originating from irrelevant regions.

To address the issues mentioned above, we propose a new model **DetailSemNet** for offline signature verification. In this model, **Structural Matching** is proposed to align local patch tokens. This mechanism enhances the model’s ability to capture local discriminative features, thereby significantly improving its identification capabilities. While integrating **Structural Matching** directly into DetailSemNet has been observed to strengthen performance, we noted a crucial limitation in attention/transformer-based models, where they often lose detailed information during the token feature extraction process. Our preliminary analysis, illustrated in Fig. 2, supports this observation. Traditional transformer-based models primarily focus on low-frequency patterns, neglecting the high-frequency information crucial for distinguishing between similar signatures. Consequently, when tested with images rich in high-frequency details, the Equal Error Rate (EER) performance shows negligible improvement. This observation suggests a performance improvement gap that deserves further exploration.

To this end, as shown in Fig. 3, we deliberately designed multi-branch networks to extract the Detail and Semantics components and handle them separately during the feature extraction process. This approach allows us to retain more detailed information, resulting in the model exhibiting improved discriminative capabilities. Compared with the conventional transformer-based method, as shown in Fig. 2, our approach can well use high-frequency information to enhance system performance. Below, we summarize our contributions.

1. Our method introduces **Structural Matching**, a novel technique designed to optimize the matching of local embeddings. Combined with global distance



**Fig. 4:** In our **DetailSemNet**, the **Detail-Semantics Integrator** splits features into two components: Semantic and Detail. The Semantic component is sent to the **SemanticsAttend Branch** to gather context information, while the Detail component goes through the **SalientConv Branch** and **DetailConv Branch**. The outputs of these three branches are then fused, creating features that incorporate both detailed and semantic information. In addition to global similarity, the model also performs structural matching on the features, allowing it to pay more attention to detailed information.

measures, it enables a more comprehensive assessment of similarity in offline signature verification, enhancing the model’s accuracy and reliability.

2. We propose the **Detail-Semantics Integrator**, an innovative network that ensures the detail and Semantics features are extracted. This integrator significantly improves the model’s feature extraction capabilities and achieves a finer understanding of the signature data, making it particularly well-suited for the OSV task.
3. Our proposed method exhibits superior performance compared to existing methods. This superiority is evident both in single-dataset testing scenarios and cross-dataset evaluations. Such performance highlights the model’s generalization ability.

## 2 Related works

### 2.1 Offline Signature Verification (OSV)

Offline signature verification has been a subject of study for many years, as evidenced by Dey et al. [6]. More recently, a shift has been observed where deep-learning methods surpass manually-designed feature methods in performance [7, 13]. For instance, Wei et al. [40] improved feature extraction for verification using gray-inverted images and attention modules. Li et al. [16] innovatively integrated sequential representations into static signature images, creating a unified framework for offline verification. Further, Li et al. [17] leveraged an

adversarial network to enhance signature verification capabilities. The use of a dual-channel Network to measure image dissimilarity was introduced by Li et al. [15]. Lately, Lu et al. [24] proposed a network which assesses pairs of images using a cycling method to make observations. Additionally, Li et al. [18, 19] marked a significant milestone by being the first to introduce a transformer framework in this domain.

Distinguishing our work from these prior efforts, we present a novel approach to offline signature verification by directly utilizing local patch tokens to learn a distance metric and explicitly find the patch matching between an image pair, thereby harnessing their potential to enhance the verification process.

## 2.2 Local Matching

Developing an effective Local Matching method for OSV is challenging. The Hungarian algorithm [12] efficiently solves the Assignment Problem but may not suit signature verification due to its strict one-to-one assignment constraint. Signatures often vary and do not align perfectly. Various distance metrics measure similarity between sets: the Hausdorff distance (HD) [10] is sensitive to outliers, the Chamfer distance (CD) [41] sums squared distances between nearest neighbors, and the Earth Mover’s distance (EMD) [41, 45] finds the least costly way to align sets. While EMD is more detail-sensitive than CD, it is also computationally complex [22], requiring a balance between efficiency and capturing the nuances of signature variations.

## 2.3 Backbone Designs and Their Properties

The field of visual data processing has evolved from Convolutional Neural Networks (CNNs) to Vision Transformer models (ViTs). However, recent studies, like Raghu et al. [31] and Bai et al. [2], have explored the strengths and limitations of ViTs compared to CNNs. A key finding is that while ViTs are adept at global information aggregation through self-attention, they might not leverage local image structures as effectively as CNNs and reduce the ability to model local fine-grained details [20, 25, 30, 36].

To address this, new ViT architectures have been introduced. PVT [38] adopts a pyramid-like structure for diverse resolution feature maps, enhancing its utility in dense prediction tasks. Swin [23] focuses on local window self-attention. DAT [42] incorporates a deformable attention mechanism, offering flexible adaptability for local feature representation. BiFormer [49] uses a Bi-Level Routing Attention mechanism, focusing on semantically rich regions for fine-grained attention. These advancements underscore the importance of detailed feature extraction and semantic learning, which are vital for Offline Signature Verification (OSV) tasks. Inspired by these insights, we proposed **DetailSemNet** to combine the advantages of CNNs and ViTs, making it particularly effective for OSV.

### 3 Method

Our model, as illustrated in Fig. 4a, processes a pair of input images,  $R$  (Reference) and  $Q$  (Query), for signature verification. These images undergo preprocessing to convert them into binary images of dimensions  $H \times W \times 1$  coupled with a foreground map, which is then used as inputs for the model. The images are segmented into patches (i.e., tokenized) and fed into the feature extraction backbone, which consists of four stages. Each stage involves a Patch Embedding layer followed by several layers of the proposed **Detail-Semantics Integrator** (DSI) module (see Sec. 3.2 and Fig. 4b). The backbone’s output comprises two sets of token features,  $f^R$  and  $f^Q$ , which are utilized to compute two types of distances,  $dis_{global}$  and  $dis_{struct}$ . First, as shown in Eq. (1), we calculate the global  $L_2$  distance  $dis_{global}$  using the global embeddings  $embed_{global}^R$  and  $embed_{global}^Q$ , derived by averaging the feature set  $f^R$  and set  $f^Q$  accordingly. Second, our Structural Matching technique is applied to ascertain a local structural distance  $dis_{struct}$ , computed from the local embeddings  $embed_{local}^R$  and  $embed_{local}^Q$  as outlined in Eq. (5). The process for calculating  $dis_{struct}$  is detailed in Sec. 3.1.

$$dis_{global} = L_2(embed_{global}^R, embed_{global}^Q) \quad (1)$$

In the evaluation phase, we utilize the combined distance, defined in Eq. (2), as the similarity measurement of  $R$  and  $Q$  and make a final verification decision by thresholding.

$$dis = \lambda_0 \times dis_{global} + dis_{struct}, \quad (2)$$

where  $\lambda_0$  is a hyperparameter, used to adjust the weighting between the two distances.

#### 3.1 Local Structural Matching

To implement Structural Matching, we focus on utilizing features from image regions containing drawn strokes for local similarity measurement and filtering out the background tokens according to the foreground map. This step is particularly beneficial due to the sparsity of signature images, where some local patches are uninformative. In addition, we mask out the background tokens after our feature extraction backbone to reduce the information loss.

To extract the foreground map, we first resize the input image from  $H \times W \times 1$  to  $h \times w$  to match the size of the extracted token feature sets (i.e.,  $f^R$  and  $f^Q$ ). Following this, we apply global thresholding to binarize the resized image and achieve the foreground map,  $Mask$ , which delineates the relevant regions containing strokes. It’s important to note that a more sophisticated text region segmentation technique can be employed when the input images have cluttered backgrounds. Later, this  $Mask$  is applied to filter out irrelevant tokens from the entire token set, leaving us with significant tokens. These tokens are subsequently processed through a linear layer to generate the local embeddings  $embed_{local}$ , which are crucial for calculating the local structural distance  $dis_{struct}$ . Specifically, we define the corresponding masks of  $R$  and  $Q$  as  $Mask^R$  and  $Mask^Q$ .

The local embeddings of  $R$  and  $Q$  and their structural distance  $dis_{struct}$  can be calculated by

$$embed_{local}^R = linear(Mask^R(f^R)), \quad (3)$$

$$embed_{local}^Q = linear(Mask^Q(f^Q)), \text{ and} \quad (4)$$

$$dis_{struct} = SM(embed_{local}^R, embed_{local}^Q). \quad (5)$$

In Eq. (5), to effectively match local structures, or tokens, we developed the function  $SM$  to quantify the similarity between two sets of local embeddings,  $embed_{local}^R = \{r_0, r_1, \dots, r_{N-1}\}$  and  $embed_{local}^Q = \{q_0, q_1, \dots, q_{M-1}\}$ , where  $N$  and  $M$  represent the number of embeddings remaining after masking. To determine the local distance  $dis_{struct}$ , we initially calculate the cosine distance between pairs of local embeddings by

$$d_{ij} = 1 - \frac{r_i^T q_j}{\|r_i\| \|q_j\|}, \quad 0 \leq i \leq N-1 \text{ and } 0 \leq j \leq M-1. \quad (6)$$

This calculation forms the basis of the ground distance matrix  $D = (d_{ij}) \in \mathbb{R}^{N \times M}$ , which is used for token matching.

The matching relationship between the tokens within the two local embedding sets is essential to obtain  $dis_{struct}$ . To represent the matching relationship, we introduce a flow matrix  $F = [f_{ij}] \in \mathbb{R}^{N \times M}$ , where each element  $f_{ij}$  indicates the ratio of the matching assignment from  $r_i$  to  $q_j$ . Note that  $r_i$  can match to multiple  $q_j$ . Once the optimal flow matrix  $F^* = [f_{ij}^*]$  (i.e., representing the ideal matching relationships) is determined, we can then compute a more accurate similarity measure. As defined in Eq. (7), this is achieved by performing an element-wise multiplication of the ground distance matrix  $D = (d_{ij})$  with the optimal flow matrix  $F^* = [f_{ij}^*]$ , followed by summing up these products and normalizing the result. The outcome of this process is a refined similarity measure that more accurately reflects the proper correspondence between the two sets of local tokens.

$$dis_{struct} = \frac{\sum_{i=0}^{N-1} \sum_{j=0}^{M-1} d_{ij} f_{ij}^*}{\sum_{i=0}^{N-1} \sum_{j=0}^{M-1} f_{ij}^*}. \quad (7)$$

Here, we conceptualize the problem of determining the optimal matching flow,  $F^* = [f_{ij}^*]$ , as the process of minimizing the Earth Mover's Distance (EMD) [34], whose details would be provided in the supplementary. The EMD framework allows us to assign different weights to each local embedding, given the total weights of all tokens sum up to 1. In this context, the weight of each local embedding represents its significance in the matching process. Our experiments find that a simple uniform weight can also achieve good results. However, if we have the prior information, we can adjust the weights for further improvement. Next, to discover the optimal matching flow, we reformulate the EMD optimization problem as a linear programming challenge and utilize the Sinkhorn algorithm [4] for problem-solving. The Sinkhorn algorithm smoothens

the EMD calculation through entropic regularization, making it possible to solve this linear programming problem effectively. Once the optimal matching flow,  $F^* = [f_{ij}^*]$ , has been calculated, we can calculate  $dis_{struct}$  by Eq. (7).

### 3.2 Detail-Semantics Integrator

Although transformer-based models have been widely adopted, several studies have discussed how the Multi-head Self-Attention module (MSA) indiscriminately suppresses high-frequency signals, leading to significant information loss. These discussions include approaches such as [2, 28, 37, 48]. This is not advantageous for tasks like OSV. Addressing this, as depicted in Fig. 4b, we have developed the **Detail-Semantics Integrator** (DSI), a novel feature enhancement technique tailored for feature maps. The DSI begins by decomposing the input feature  $X$  into two parts: the Semantic Feature  $Sem[X]$  and Detailed Feature  $Det[X]$ . To maintain computational efficiency, we compute  $Sem[X]$  by performing local average pooling on  $X$ , which effectively captures lower-frequency parts. Conversely,  $Det[X]$ , representing higher-frequency parts, is computed by  $X - Sem[X]$ . Later, after a 1-by-1 projection layer,  $Sem[X]_{proj}$  undergoes processing in the SemanticsAttend Branch  $SemAtt$ , using an attention-based module to extract semantic features  $Y_{Sem}$ . On the other hand,  $Det[X]_{proj}$  is processed through a convolution-based module  $Conv$  to extract fine-grain details  $Y_{Det}$  from local features. The steps are as follows:

$$Y_{Sem} = SemAtt(Sem[X]_{proj}) \text{ and} \quad (8)$$

$$Y_{Det} = Conv(Det[X]_{proj}). \quad (9)$$

Moreover, the convolution-based module is strategically divided into the *Salient-Conv* branch and the *DetailConv Branch*. The *SalientConv* branch integrates both maximum filter and convolution layers, where the maximum filter is particularly effective in retaining salient features and helps to highlight the prominent aspects of the signature images. Conversely, the *DetailConv* branch, consisting of two successive convolution layers, is tailored to draw out finer details. Unlike attention mechanisms, convolutions excel at detecting intricate high-frequency details, making them ideal for processing the  $Det[.]$  part of the input. For feature decomposition, the projected detailed features  $Det[X]_{proj}$  are divided equally along the channel direction, with each part directed to either *SalientConv* or *DetailConv* for specialized processing. The detailed steps are as follows:

$$Det1[X], Det2[X] = split(Det[X]_{proj}). \quad (10)$$

$$Y_{Det1} = SalientConv(Det1[X]). \quad (11)$$

$$Y_{Det2} = DetailConv(Det2[X]). \quad (12)$$

In the final stage of DSI (Fig. 4b), the outputs from the *SemAtt*, *SalientConv*, and *DetailConv* branches are concatenated along the channel dimension to form a unified feature representation, followed by a Residual Convolution layer (*Conv<sub>fuse</sub>*). This layer further integrates the concatenated features to ensure a seamless blend of detailed and semantic information. Moreover, consistent with the architecture of most transformer-based models, DSI incorporates a Feed-Forward Network (FFN) and Layer Normalization (LN) to enhance the model’s processing capabilities. In Sec. 4.5, we demonstrate through experiments that this particular design of DetailSemNet is highly effective for OSV tasks.

### 3.3 Loss Function

The output of our signature verification model is  $dis$  (Eq. (2)), which denotes the distance between two input signature images. To train the model, we employ a double-margin contrastive loss [8, 24], defined as follows:

$$Loss_{DM} = y\{max(0, dis - m)\}^2 + (1 - y)\{max(0, n - dis)\}^2. \quad (13)$$

Here, the supervised label  $y$  is assigned a value of 1 for positive (genuine-genuine) signature pairs and 0 for negative (genuine-forged) signature pairs. The parameters  $n$  and  $m$  represent the margin values used in the loss calculation. Importantly,  $m$  is constrained to be less than  $n$  to ensure the loss function behaves as intended.

## 4 Experiments

We evaluated our approach using four challenging datasets, each representing a different language. These include the CEDAR Dataset [11] (English), BHSig-B Dataset [27] (Bengali), BHSig-H Dataset [27] (Hindi), and ChiSig Dataset<sup>4</sup> [44] (Chinese). To further test the robustness of our model, we conducted cross-language experiments. These involved training the model on a dataset and subsequently testing it on a dataset in a different language. Our model was trained on an NVIDIA GeForce RTX 2080Ti GPU. The initial model pre-trained weights were derived from the ImageNet1K dataset [5].

We performed comprehensive comparisons with several existing methods [16, 18, 40] using a variety of metrics. These metrics include the False Rejection Rate (FRR), False Acceptance Rate (FAR), Equal Error Rate (EER), Area Under the Curve (AUC), and Accuracy (Acc). Furthermore, we utilize the Equal Error Rate (EER) to identify the point where FRR and FAR are equal. This equilibrium point informs the threshold used to calculate Accuracy (Acc) and facilitates the verification decision-making process.

---

<sup>4</sup> The latest ChiSig Dataset, unlike the others, has not been previously utilized for Offline Signature Verification (OSV) tasks. Therefore, it was specifically used in our ablation study to underscore the efficacy of our method.

**Table 1:** Signature verification comparison on BHSig-H(%) and BHSig-B(%).

Method	BHSig-H				BHSig-B			
	FAR	FRR	Acc ↑	EER ↓	FAR	FRR	Acc ↑	EER ↓
SigNet [6]	15.36	15.36	84.64	15.36	13.89	13.89	86.11	13.89
IDN [40]	4.93	8.99	93.04	6.96	4.12	5.24	95.32	4.68
DeepHsv [15]	-	-	86.66	13.34	-	-	88.08	11.92
SDINet [16]	6.24	3.77	95.00	5.11	3.30	7.86	94.42	5.39
CaC [24]	5.97	5.97	94.03	5.97	3.96	3.96	96.04	3.96
AVN [17]	5.46	5.91	94.32	5.65	7.33	5.07	93.80	6.14
TransOSV [18, 19]	3.39	<b>3.39</b>	96.61	3.39	9.95	9.95	90.05	9.95
2C2S [33]	8.66	5.16	90.68	9.32	5.37	8.11	93.25	6.75
SURDS [3]	12.01	8.98	89.50	-	19.89	5.42	87.34	-
MA-SCN [46]	5.73	4.86	94.99	5.32	9.96	5.85	92.86	8.18
SigGCN [32]	12.96	11.27	87.88	12.17	4.06	<b>3.95</b>	95.99	4.00
Co-Tuplet [9]	6.76	6.56	-	6.68	5.93	6.20	-	6.12
HybridFE [43]	11.74	11.74	88.26	11.74	8.36	8.36	91.64	8.36
SPD Manifold [50]	-	-	-	15.60	-	-	-	11.10
<b>Ours</b>	<b>1.07</b>	3.59	<b>98.24</b>	<b>2.07</b>	<b>0.95</b>	4.04	<b>98.19</b>	<b>2.11</b>

#### 4.1 Results on BHSig-B and BHSig-H Datasets

The BHSig260 Dataset [27] includes both the BHSig-B and BHSig-H Datasets. The BHSig-B Dataset contains signatures from 100 individuals from Bengal, with each individual contributing 24 genuine signatures and 30 forgeries. For our model’s training, we utilized the signatures from 50 of these individuals, while the signatures from the remaining 50 were reserved for testing purposes. In contrast, the BHSig-H Dataset consists of signatures from 160 individuals, with each providing 24 genuine signatures and 30 forged signatures. In this case, our training involved signatures from 100 individuals, with the remaining 60 individuals’ signatures set aside for the testing phase.

Our mode performance was compared with several conventional approaches, including SigNet [6], IDN [40], DeepHsv [15], SDINet [16], CaC [24], AVN [17], TransOSV [18,19], 2C2S [33], SURDS [3], MA-SSN [46], SigGCN [32], HybridFE [43], and SPD Manifold [50]. The evaluation results are detailed in Tab. 1.

Regarding the BHSig-B Dataset, our method demonstrated impressive results, achieving an Accuracy (Acc) of 98.19%, a False Acceptance Rate (FAR) of 0.95%, and a False Rejection Rate (FRR) of 4.04%. Additionally, the model recorded an Equal Error Rate (EER) of 2.11%. Compared to the best available result from other methods, our approach exhibits a significant performance gain of 1.85%. Turning to the BHSig-H Dataset, our method continued its strong performance, achieving an Accuracy of 98.24%, an FAR of 1.07%, and an FRR of 3.59%. The EER for this dataset was an impressive 2.07%. Compared with the leading comparative methods, our approach marks a notable performance improvement of 1.32%.

## 4.2 Results on CEDAR Dataset

Within the CEDAR signature dataset [11], each individual is represented by 24 genuine and 24 forged signatures, all written in English. Aligning with methodologies from previous studies, we used the signatures of 50 individuals to train our model and reserved the signatures of the remaining 5 individuals for testing. In this setup, a positive sample is created by pairing a reference signature with a genuine signature, while a negative sample is formed by pairing a reference signature with a forged signature. Thus, each signatory contributes 276 positive and 276 negative pairs for verification.

Method	FAR	FRR	Acc ↑	EER ↓
SigNet [6]	-	-	-	4.63
IDN [40]	5.87	2.17	-	3.62
SDINet [16]	3.42	0.73	-	1.75
CaC [24]	4.34	4.34	95.66	4.43
AVN [17]	3.26	4.42	96.16	3.77
MA-SCN [46]	19.21	18.35	80.75	18.92
MLDF [1]	-	-	-	5.00
Co-Tuplet [9]	3.33	3.55	-	3.51
HybridFE [43]	9.95	9.95	90.05	9.95
SPD Manifold [50]	-	-	-	8.53
<b>Ours</b>	<b>0.36</b>	<b>0.58</b>	<b>99.53</b>	<b>0.58</b>

We conducted a comparative analysis of our model with other conventional approaches such as SigNet [6], IDN [40], SDINet [16], CaC [24], AVN [17], MA-SCN [46], MLFD [1], Co-Tuplet [9], HybridFE [43], and SPD Manifold [50]. The results of these comparative analyses on the CEDAR Dataset are detailed in Tab. 2. Here, our method demonstrated exemplary performance, achieving an accuracy of 99.53%, a False Acceptance Rate (FAR) of 0.36%, and a False Rejection Rate (FRR) of 0.58%. Additionally, the model attained an Equal Error Rate (EER) of 0.58%. Compared to the best results from other methods, our approach appeared as a top-performing model.

## 4.3 Verification on Cross-Dataset Scenario

To evaluate the generalization capabilities of our model, we train the model on a dataset and directly test the model on other datasets with different languages. The effectiveness of our model in this regard is demonstrated by the test results presented in Tab. 3. Compared to other methods, our approach consistently outperforms them, indicating its adaptability and robustness across different languages without model finetune.

## 4.4 Ablation Studies

We conducted ablation experiments to evaluate the impact of each module introduced in our proposed method, including Structural Matching (SM), Detail-Conv Branch (DCB) and SalientConv Branch (SCB). These tests were performed across four different datasets, and the results are presented in Tab. 4. Through

**Table 3:** The zero-shot cross-lingual OSV task (cross-dataset) testing results. (EER%)

Train	BHSig-H			BHSig-B			CEDAR		
Test	BHSig-B	CEDAR	BHSig-H	CEDAR	BHSig-H	BHSig-B	BHSig-H	BHSig-B	
SigNet [6]	39.35	40.43	35.43	50.00	44.39	35.85			
IDN [40]	25.88	50.00	25.70	50.00	49.64	49.99			
CaC [24]	14.66	29.49	30.41	33.71	39.08	38.07			
SURDS [3]	27.74	-	32.99	-	-	-			
TransOSV [18, 19]	18.66	-	17.17	-	-	-			
<b>Ours</b>	<b>7.46</b>	<b>14.05</b>	<b>15.91</b>	<b>7.32</b>	<b>16.35</b>	<b>8.40</b>			

**Table 4:** Ablation study of Structural Matching (SM), DetailConv Branch (DCB), and SalientConv Branch (SCB). The table presents the results using the abbreviations M, D, and S in that order. Four different datasets are tested.

M D S	BHSig-H			BHSig-B			CEDAR			ChiSig		
	EER	AUC	Acc									
x x x	4.70	0.991	95.82	3.37	0.995	97.14	3.41	0.994	96.81	12.47	0.947	88.68
✓ x x	4.67	0.992	95.88	3.29	0.995	97.22	1.99	0.998	98.08	10.69	0.964	89.82
x ✓ x	2.72	0.997	97.70	2.51	0.997	97.44	1.45	0.999	98.77	8.91	0.972	91.73
x x ✓	2.87	0.997	97.69	2.66	0.997	97.68	1.59	0.998	98.41	8.65	0.977	92.62
x ✓ ✓	2.62	0.997	97.80	2.50	0.997	97.89	1.74	0.999	98.59	7.00	<b>0.985</b>	93.89
✓ ✓ x	2.51	0.997	97.87	2.19	<b>0.998</b>	98.03	1.09	0.999	98.95	8.65	0.977	91.35
✓ x ✓	2.72	0.997	97.74	2.19	<b>0.998</b>	98.15	2.10	0.998	98.19	6.36	0.983	93.89
✓ ✓ ✓	<b>2.07</b>	<b>0.998</b>	<b>98.24</b>	<b>2.11</b>	<b>0.998</b>	<b>98.19</b>	<b>0.58</b>	<b>1.000</b>	<b>99.53</b>	<b>5.85</b>	<b>0.985</b>	<b>94.40</b>

these experiments, we observed a consistent trend of progressive performance improvement with the sequential incorporation of each proposed modification.

To further understand the impact of our proposed Structural Matching, we experimented with its integration at various stages within the model. The outcomes of these tests are detailed in Tab. 5. Notably, the results demonstrate that integrating Structural Matching towards the end of the process yields the most favorable performance.

#### 4.5 Comparison of Different Backbones

We also evaluate the performance impact of various transformer backbones on our model. We conducted training under identical conditions using different transformer architectures, including PVT [38], Swin [23], SPACH [47], DAT [42], and BiFormer [49], for comparison. The results, illustrating how each backbone influenced the model’s performance, are presented in Tab. 6. Compared with other backbones, our model exhibited better performance, underscoring the effectiveness of our design.

**Table 5:** Apply Structural Matching to different stages on the BHSig-H.

stage	EER(%) ↓
3	3.47
3 & 4	3.05
<b>4 (Ours)</b>	<b>2.09</b>

**Table 6:** Comparison of Different Backbones. We test our results on the BHSig-H and BHSig-B. All the models are under identical conditions.

Dataset	BHSig-H		BHSig-B	
	EER	Acc	ERR	Acc
PVT [38]	4.62	96.06	2.72	97.55
Swin [23]	4.24	96.15	10.27	91.01
SPACH [47]	3.71	96.67	3.30	96.90
DAT [42]	4.94	95.73	20.09	82.77
BiFormer [49]	4.38	96.10	8.66	92.38
<b>Ours</b>	<b>2.07</b>	<b>98.24</b>	<b>2.11</b>	<b>98.19</b>

#### 4.6 Comparing with Verification Tasks

Verification Tasks like OSV, Re-ID, and face verification aim to assess image similarity but also face distinct technique challenges. *E.g.*, a Re-ID task contends with pose variations [21], occlusions [35], or non-discriminative appearance issues [26]. In contrast, face verification challenges stem from resolution differences [14], aging appearance [39], and wearing accessories [29]. Our method highlights the need for OSV tasks to discern fine-grained differences in signature pairs, a challenge that is less prominent in other tasks. We applied state-of-the-art Re-ID models to the OSV task, and the results are shown in Tab. 7.

#### 4.7 Visualize Matching Results

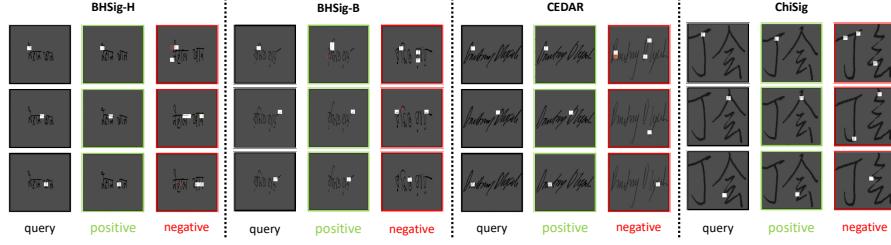
Visualizing matching flows during inference adds interpretability to our model, as shown in Fig. 5. The reference image is placed on the left. It queries a specific patch and then demonstrates how it corresponds to a positive image in the middle and a negative image on the right. Positive pairs typically show correct patch correspondences, while negative pairs struggle with mismatches. We also visualized the impact of Structural Matching in Fig. 6, showing that models without it have difficulty achieving precise patch matches.

#### 4.8 Impact of High-Frequency Information on Performance

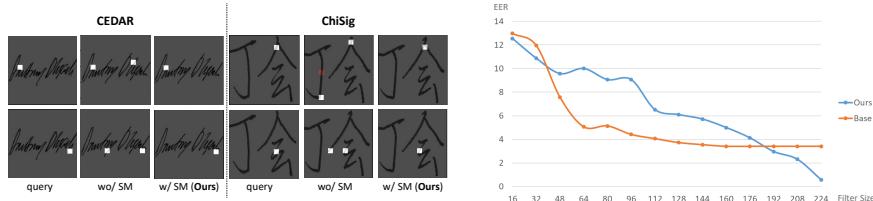
We use low-pass filters to create testing images with varying high-frequency (HF) content, evaluating both our model and a conventional transformer-based method. The results in Fig. 7 show the impact of HF details on performance, with the X-axis indicating the amount of HF information. Higher HF content generally reduces EERs. For the transformer-based model, EER plateaus at 3.41%

**Table 7:** Comparison with Re-ID models on BHSig-H.

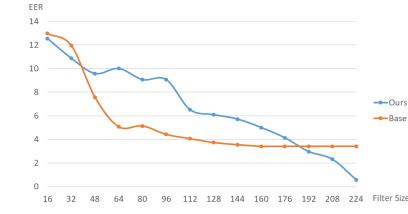
Method	FAR	FRR	Acc ↑	EER ↓
BPB [26]	2.73	7.25	96.02	4.42
PAT [35]	4.95	13.32	92.73	7.85
<b>Ours</b>	<b>1.07</b>	<b>3.59</b>	<b>98.24</b>	<b>2.07</b>



**Fig. 5:** Illustrating the matching results of our model on signature pairs. The sample pairs are selected from the four datasets. Our model demonstrates correct matching results when tested on positive pairs; whereas, when tested on negative pairs, it exhibits matching at incorrect positions, sometimes even at multiple locations.



**Fig. 6:** This figure illustrates the matching results of our model on signature pairs with or without our Structural Matching (SM). The results demonstrate how our Structural Matching improves the model's ability to capture detailed features.



**Fig. 7:** The X-axis shows the amount of high-frequency information in the testing images, with higher values indicating greater inclusion. This graph illustrates that our approach efficiently leverages high-frequency details to reduce EER.

beyond a HF threshold of 164, showing a saturation point. In contrast, our model significantly reduces EER from 6.52% to 0.58%, as HF content increases from 112 to 224.

## 5 Conclusion

In this paper, we introduce **DetailSemNet**, a novel model for Offline Signature Verification (OSV) that emphasizes local patch features in Structural Matching, a shift from traditional holistic approaches. **DetailSemNet** also incorporates the Detail-Semantics Integrator (DSI) to enhance structural matching, effectively capturing detailed and semantic aspects. Our results demonstrate that **DetailSemNet** outperforms existing methods in both single-dataset and cross-dataset scenarios, highlighting its strong generalization capability and potential for real-world application. These findings indicate the effectiveness of combining the DSI module with Structural Matching in OSV models, positioning **DetailSemNet** as a significant advancement in forensic technology.

## Acknowledgements

This work was supported in part by E.SUN Financial Holding, and financially supported in part (project number: 112UA10019) by the Co-creation Platform of the Industry Academia Innovation School, NYCU, under the framework of the National Key Fields Industry-University Cooperation and Skilled Personnel Training Act, from the Ministry of Education (MOE) and industry partners in Taiwan. It also supported in part by the National Science and Technology Council, Taiwan, under Grant NSTC-112-2221-E-A49-089-MY3, Grant NSTC-110-2221-E-A49-066-MY3, Grant NSTC-111- 2634-F-A49-010, Grant NSTC-112-2425-H-A49-001-, and in part by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and the Ministry of Education (MOE), Taiwan. (Corresponding author: Ching-Chun Huang.)

## References

1. Arab, N., Nemmour, H., Chibani, Y.: A new synthetic feature generation scheme based on artificial immune systems for robust offline signature verification. *Expert Syst. Appl.* **213**(PC) (mar 2023). <https://doi.org/10.1016/j.eswa.2022.119306>
2. Bai, J., Yuan, L., Xia, S.T., Yan, S., Li, Z., Liu, W.: Improving vision transformers by revisiting high-frequency components. In: European Conference on Computer Vision. pp. 1–18 (2022), [https://doi.org/10.1007/978-3-031-20053-3\\_1](https://doi.org/10.1007/978-3-031-20053-3_1)
3. Chattopadhyay, S., Manna, S., Bhattacharya, S., Pal, U.: Surds: Self-supervised attention-guided reconstruction and dual triplet loss for writer independent offline signature verification. 2022 26th International Conference on Pattern Recognition (ICPR) pp. 1600–1606 (2022), <https://api.semanticscholar.org/CorpusID:246275631>
4. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 26. Curran Associates, Inc. (2013), [https://proceedings.neurips.cc/paper\\_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). <https://doi.org/10.1109/CVPR.2009.5206848>
6. Dey, S., Dutta, A., Toledo, J.I., Ghosh, S.K., Llados, J., Pal, U.: Signet: Convolutional siamese network for writer independent offline signature verification. arXiv preprint arXiv:1707.02131 (2017)
7. Dutta, A., Pal, U., Lladós, J.: Compact correlated features for writer independent signature verification. In: 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 3422–3427 (2016). <https://doi.org/10.1109/ICPR.2016.7900163>
8. Hao, J., Dong, J., Wang, W., Tan, T.: Deepfirearm: Learning discriminative feature representation for fine-grained firearm retrieval. 2018 24th International Conference on Pattern Recognition (ICPR) pp. 3335–3340 (2018), <https://api.semanticscholar.org/CorpusID:47010593>

9. Huang, F.H., Lu, H.M.: Multiscale feature learning using co-triplet loss for offline handwritten signature verification. arXiv preprint arXiv:2308.00428 (2023)
10. Huttenlocher, D., Klanderman, G., Rucklidge, W.: Comparing images using the hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence **15**(9), 850–863 (1993). <https://doi.org/10.1109/34.232073>
11. Kalera, M., Xu, A.: Offline signature verification and identification using distance statistics. IJPRAI **18**, 1339–1360 (11 2004). <https://doi.org/10.1142/S0218001404003630>
12. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly **2**(1-2), 83–97 (1955). <https://doi.org/https://doi.org/10.1002/nav.3800020109>, <https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109>
13. Kumar, R., Sharma, J., Chanda, B.: Writer-independent off-line signature verification using surroundedness feature. Pattern Recognition Letters **33**(3), 301–308 (2012). <https://doi.org/https://doi.org/10.1016/j.patrec.2011.10.009>, <https://www.sciencedirect.com/science/article/pii/S0167865511003552>
14. Kuo, C., Tsai, Y.T., Shuai, H.H., Yeh, Y.r., Huang, C.C.: Towards understanding cross resolution feature matching for surveillance face recognition. In: Proceedings of the 30th ACM International Conference on Multimedia. p. 6706–6716. MM '22, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3503161.3548402>
15. Li, C., Lin, F., Wang, Z., Yu, G., Yuan, L., Wang, H.: Deepshv: User-independent offline signature verification using two-channel cnn. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 166–171 (2019). <https://doi.org/10.1109/ICDAR.2019.00035>
16. Li, H., Wei, P., Hu, P.: Static-dynamic interaction networks for offline signature verification. In: AAAI Conference on Artificial Intelligence (2021), <https://api.semanticscholar.org/CorpusID:235306077>
17. Li, H., Wei, P., Hu, P.: Avn: An adversarial variation network model for handwritten signature verification. IEEE Transactions on Multimedia **24**, 594–608 (2022). <https://doi.org/10.1109/TMM.2021.3056217>
18. Li, H., Wei, P., Ma, Z., Li, C., Zheng, N.: Offline signature verification with transformers. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2022). <https://doi.org/10.1109/ICME52920.2022.9859886>
19. Li, H., Wei, P., Ma, Z., Li, C., Zheng, N.: Transosv: Offline signature verification with transformers. Pattern Recognition **145**, 109882 (2024). <https://doi.org/https://doi.org/10.1016/j.patcog.2023.109882>, <https://www.sciencedirect.com/science/article/pii/S0031320323005800>
20. Li, K., Yu, R., Wang, Z., ming Yuan, L., Song, G., Chen, J.: Locality guidance for improving vision transformers on tiny datasets. In: European Conference on Computer Vision (2022), <https://api.semanticscholar.org/CorpusID:250699338>
21. Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., Hu, J.: Pose transferrable person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4099–4108 (2018). <https://doi.org/10.1109/CVPR.2018.00431>
22. Liu, M., Sheng, L., Yang, S., Shao, J., Hu, S.M.: Morphing and sampling network for dense point cloud completion. arXiv preprint arXiv:1912.00280 (2019)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)

24. Lu, X., Huang, L., Yin, F.: Cut and compare: End-to-end offline signature verification network. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 3589–3596 (2021). <https://doi.org/10.1109/ICPR48806.2021.9412377>
25. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16317–16326 (June 2021)
26. Ni, H., Li, Y., Gao, L., Shen, H.T., Song, J.: Part-aware transformer for generalizable person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11280–11289 (2023)
27. Pal, S., Alaei, A., Pal, U., Blumenstein, M.: Performance of an off-line signature verification method based on texture features on a large indic-script signature dataset. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS). pp. 72–77 (2016). <https://doi.org/10.1109/DAS.2016.48>
28. Park, N., Kim, S.: How do vision transformers work? In: International Conference on Learning Representations (2022), <https://openreview.net/forum?id=D78Go4hVcx0>
29. Phan, H., Nguyen, A.: Deepface-emd: Re-ranking using patch-wise earth mover's distance improves out-of-distribution face identification. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2022). <https://doi.org/10.1109/cvpr52688.2022.01962>, <http://dx.doi.org/10.1109/CVPR52688.2022.01962>
30. Pinto, F., Torr, P.H.S., Dokania, P.K.: An impartial take to the cnn vs transformer robustness contest. In: European Conference on Computer Vision (2022), <https://api.semanticscholar.org/CorpusID:251040759>
31. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021), <https://openreview.net/forum?id=G18FHfMVTZu>
32. Ren, C., Zhang, J., Wang, H., Shen, S.: Vision graph convolutional network for writer-independent offline signature verification. 2023 International Joint Conference on Neural Networks (IJCNN) pp. 1–7 (2023), <https://api.semanticscholar.org/CorpusID:260387734>
33. Ren, J.X., Xiong, Y.J., Zhan, H., Huang, B.: 2c2s: A two-channel and two-stream transformer based framework for offline signature verification. Engineering Applications of Artificial Intelligence **118**, 105639 (2023). <https://doi.org/https://doi.org/10.1016/j.engappai.2022.105639>, <https://www.sciencedirect.com/science/article/pii/S0952197622006297>
34. Rubner, Y., Tomasi, C., Guibas, L.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision **40**, 99–121 (11 2000). <https://doi.org/10.1023/A:1026543900054>
35. Somers, V., Vleeschouwer, C.D., Alahi, A.: Body part-based representation learning for occluded person re-identification. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE (Jan 2023). <https://doi.org/10.1109/wacv56688.2023.00166>, <http://dx.doi.org/10.1109/WACV56688.2023.00166>
36. Wang, H., Wu, X., Huang, Z., Xing, E.P.: High-frequency component helps explain the generalization of convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8681–8691 (2020). <https://doi.org/10.1109/CVPR42600.2020.00871>

37. Wang, P., Zheng, W., Chen, T., Wang, Z.: Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. In: International Conference on Learning Representations (2022), <https://openreview.net/forum?id=0476oWmiNNp>
38. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 548–558 (2021). <https://doi.org/10.1109/ICCV48922.2021.00061>
39. Wang, X., Zhou, Y., Kong, D., Currey, J., Li, D., Zhou, J.: Unleash the black magic in age: A multi-task deep neural network approach for cross-age face verification. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). pp. 596–603 (2017). <https://doi.org/10.1109/FG.2017.75>
40. Wei, P., Li, H., Hu, P.: Inverse discriminative networks for handwritten signature verification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5757–5765 (2019). <https://doi.org/10.1109/CVPR.2019.00591>
41. Wu, T., Pan, L., Zhang, J., WANG, T., Liu, Z., Lin, D.: Density-aware chamfer distance as a comprehensive metric for point cloud completion. In: In Advances in Neural Information Processing Systems (NeurIPS), 2021 (2021)
42. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4794–4803 (June 2022)
43. Xiong, T., Zhang, X.: Hybrid feature extraction based deep learning model for offline signature verification. 2023 6th International Conference on Software Engineering and Computer Science (CSECS) pp. 1–6 (2023), <https://api.semanticscholar.org/CorpusID:267703979>
44. Yan, K., Zhang, Y., Tang, H., Ren, C., Zhang, J., Wang, G., Wang, H.: Signature detection, restoration, and verification: A novel chinese document signature forgery detection benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 5163–5172 (June 2022)
45. Zhang, C., Cai, Y., Lin, G., Shen, C.: Deepemd: Differentiable earth mover’s distance for few-shot learning. IEEE Transactions on Pattern Analysis and Machine Intelligence p. 1–17 (2022). <https://doi.org/10.1109/tpami.2022.3217373>, <http://dx.doi.org/10.1109/TPAMI.2022.3217373>
46. Zhang, X., Wu, Z., Xie, L., Li, Y., Li, F., Zhang, J.: Multi-path siamese convolution network for offline handwritten signature verification. In: Proceedings of the 2022 8th International Conference on Computing and Data Engineering. p. 51–58. ICCDE ’22, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3512850.3512854>
47. Zhao, Y., Wang, G., Tang, C., Luo, C., Zeng, W., Zha, Z.J.: A battle of network structures: An empirical study of cnn, transformer, and mlp. arXiv preprint arXiv:2108.13002 (2021)
48. Zheng, B., Zhou, D.W., Ye, H.J., chuan Zhan, D.: Preserving locality in vision transformers for class incremental learning. 2023 IEEE International Conference on Multimedia and Expo (ICME) pp. 1157–1162 (2023), <https://api.semanticscholar.org/CorpusID:258170025>
49. Zhu, L., Wang, X., Ke, Z., Zhang, W., Lau, R.: Biformer: Vision transformer with bi-level routing attention. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

50. Zois, E., Tsourounis, D., Kalivas, D.: Similarity distance learning on spd manifold for writer independent offline signature verification. *IEEE Transactions on Information Forensics and Security* **PP**, 1342 – 1356 (11 2023). <https://doi.org/10.1109/TIFS.2023.3333681>