Joint Identity Verification and Pose Alignment for Partial Fingerprints

Xiongjun Guan, Zhiyu Pan, Jianjiang Feng, *Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

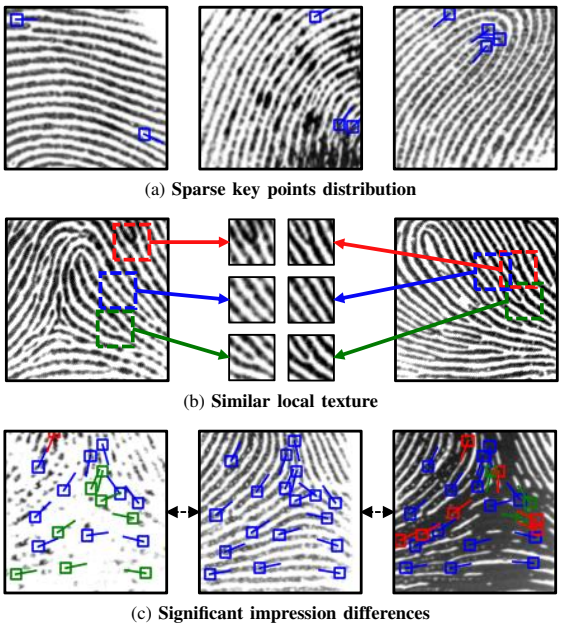
***Abstract*—Currently, portable electronic devices are becoming more and more popular. For lightweight considerations, their fingerprint recognition modules usually use limited-size sensors. However, partial fingerprints have few matchable features, especially when there are differences in finger pressing posture or image quality, which makes partial fingerprint verification challenging. Most existing methods regard fingerprint position rectification and identity verification as independent tasks, ignoring the coupling relationship between them — relative pose estimation typically relies on paired features as anchors, and authentication accuracy tends to improve with more precise pose alignment. In this paper, we propose a novel framework for joint identity verification and pose alignment of partial fingerprint pairs, aiming to leverage their inherent correlation to improve each other. To achieve this, we present a multi-task CNN (Convolutional Neural Network)-Transformer hybrid network, and design a pre-training task to enhance the feature extraction capability. Experiments on multiple public datasets (NIST SD14, FVC2002 DB1_A & DB3_A, FVC2004 DB1_A & DB2_A, FVC2006 DB1_A) and an in-house dataset demonstrate that our method achieves state-of-the-art performance in both partial fingerprint verification and relative pose estimation, while being more efficient than previous methods. Code is available at: https://github.com/XiongjunGuan/JIPNet.**

***Index Terms*—Fingerprint recognition, partial fingerprint, fingerprint verification, fingerprint pose estimation, transformer.**

I. Introduction

FINGERPRINTS are unique patterns composed of ridges and valleys on the finger surface. Due to its easy collection, high stability and strong recognizability, this biological characteristic exhibits significant application value. As early as the nineteenth century, researchers had conducted systematic studies on fingerprints [1], [2]. The reliability of fingerprint recognition gained official recognition in the twentieth century and found extensive application within the judicial sphere [2]. With the advancement of sensors and algorithms, fingerprint recognition technology has been swiftly applied in civilian and commercial fields, such as portable devices, financial services, and access control systems, providing people with convenience and security [3].

Fueled by consumer demand and technological progress, the requirements towards integration, miniaturization, and portability has become prominent within the consumer electronics sector. This trend compels manufacturers to consistently reduce the size of sensors to accommodate progressively compact device architectures. Fingerprint sensors, currently one of the most popular biometric modules, have also introduced various miniaturized solutions for various applications [3]. However, the size of fingerprint sensors is inevitably limited, leading to a significant reduction in the information available for matching [4]–[6]. Fig. 1 shows three representative challenging scenarios in partial fingerprint recognition, which can be even more difficult on fingers with large posture differences or poor quality (such as wear and wrinkles), resulting in a sharp decline in performance of feature extraction and matching [7]. This not only reduces the user experience, but also poses great security risks [8].



(a) **Sparse key points distribution**

(b) **Similar local texture**

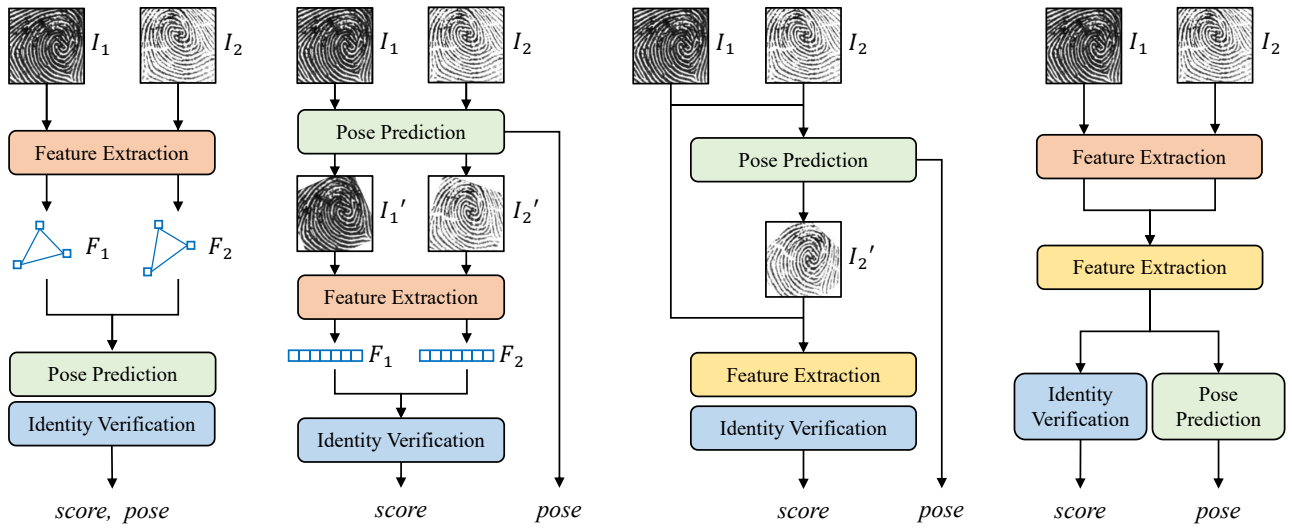
(c) **Significant impression differences**

Fig. 1. Partial fingerprint matching faces three challenges: sparse minutiae distribution, similar local texture and significant modal difference. This figure shows corresponding representative examples: (a) Three weakly-textured fingerprints that lack appropriate landmark feature points. (b) Local ridge patterns of fingerprints from different fingers have high similarity. (c) Fingerprint features in different skin conditions (in this case from left to right are dry, normal and wet respectively) may be missed (green) or incorrectly extracted (red). Visualized feature points are minutiae extracted by VeriFinger [9].

A natural idea for partial fingerprint matching is to transfer existing solutions for rolled or plain fingerprint recognition. Nevertheless, conventional algorithms usually rely on a sufficient number of key points to establish and compare spatial structural relationships [10], [11]. While certain studies use neural networks to extract fixed-length region descriptors [12]–[15], consequently reducing the dependence on point features, these approaches require either large effective area

This work was supported in part by the National Natural Science Foundation of China under Grant 62376132 and 62321005. (*Corresponding author: Jianjiang Feng*.)

The authors are with Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: gxj21@mails.tsinghua.edu.cn; pzy20@mails.tsinghua.edu.cn; jfeng@tsinghua.edu.cn; jzhou@tsinghua.edu.cn).



(a) **Key points based [10]** (b) **Fixed-length region representation based [12]** (c) **Pairwise interrelated feature based [7]** (d) **Joint estimation based (ours)**

Fig. 2. Frameworks of different fingerprint matching algorithms. The process passed by parallel arrows indicates that the corresponding modules share weights and their functions can be executed independently. *Feature Extraction* in red and gold extract independent and interrelated features from paired data respectively. *Pose Predictor* in (a)(c)(d) and (b) estimate relative and absolute pose respectively.

or exceedingly precise alignment. Considering this, some researchers simultaneously input fingerprint images and extract interrelated features to fully utilize complete information [7], [16], which effectively improves the matching performance on partial fingerprints. However, to the best of our knowledge, almost all existing fingerprint matching methods (compared in Fig. 2) regard pose rectification and identity verification as independent tasks, ignoring the coupling relationship between them. Specifically, the estimation of relative pose typically relies on paired regional features as anchors, and the accuracy of authentication generally increases with enhanced precision in pose alignment. This is similar to a pictorial jigsaw puzzle where decision-makers need to consider whether (are they neighboring image fragments?) and how (what's the reassemble solution?) any two pieces fit together [17], [18]. It can be considered as a mixed process, as the relative alignment relationship is confirmed (one of the eight lateral positions for a regular puzzle) simultaneously when two pieces are determined to be adjacent. On the other hand, an input piece can be classified as outside (mismatched) when it cannot be properly connected to the reference piece at any position. The comprehensive consideration of identity verification and pose alignment can be transferred to fingerprints, which is feasible because previous works [7], [19] have demonstrated the complementarity and compatibility between above two tasks. In addition, the paradigm of joint decision can better balance the robustness and speed of overall process, rather than performing them individually.

In this paper, we propose a multi-task CNN-Transformer hybrid network which **J**ointly performs **I**dentity verification and **P**ose alignment for partial fingerprints, called **JIPNet**. Paired images are input instead of their simplified features to provide more complete information, enabling more flexible and precise analysis. The intrinsic connection between feature correspondence and position correlation is exploited, aiming to promote each other to achieve higher performance. Unlike previous independent two-stage approaches, the estimation of identity and pose are jointly executed in our proposed framework to leverage their complementarity. Naturally, the efficiency is also improved because these two tasks are compactly integrated. Inspired by works on natural images [20]–[23], we utilize a shared weight convolution layers to extract regional features and transformers to capture both local and global correlations. Besides, a fingerprint enhancement pre-training task is specially designed to enhance the compatibility with different image modalities and various texture patterns.

Extensive experiments are conducted on multiple datasets, including several public datasets (*NIST SD14* [24], *FVC2002 DB1_A & DB3_A* [25], *FVC2004 DB1_A & DB2_A* [26], *FVC2006 DB1_A* [27]) and an in-house dataset. Experimental results show that the proposed algorithm outperforms previous state-of-the-art methods in both partial fingerprint verification and pose estimation, and also demonstrates high efficiency.

The main contributions of our paper can be summarized as:

- We propose a novel approach for partial fingerprint recognition by jointly estimating the authentication probability and relative pose instead of the previous independent stages.
- A CNN-Transformer hybrid network structure is presented to aggregate their respective advantages in local feature extraction and global information interaction.
- A lightweight pre-training task on fingerprint enhancement is specifically designed to further improve the generalization ability of our proposed network.
- Extensive experiments are conducted on diverse datasets to evaluate representative state-of-the-art algorithms, which demonstrates the superiority of our proposed method.

The paper is organized as follows: Section II reviews the related works. Section III introduces the proposed partial

fingerprint recognition algorithm. Section IV describes the details and usage of datasets. Section V presents the experimental results and discussions. Finally, we draw conclusions in Section VI.

II. Related Work

According to different feature representation forms, existing one-to-one fingerprint matching algorithms can be mainly divided into three categories: key points based, fixed-length region descriptor based, and pairwise interrelated feature based. In the first two types of algorithms highly generalized features are stored and compared here (in other words, each fingerprint undergoes feature extraction only once) to enable efficient searches in large-scaled galleries. On the other hand, pairwise interrelated feature based methods could use richer information (features need to be re-extracted for each match) and is more suitable for confirmation scenarios with high precision requirements and few gallery / reference fingerprints, such as unlocking a mobile phone. Given the multitude of fingerprint matching algorithms, this rough classification may not comprehensively encompass all works. Nonetheless, we still endeavor to summarize them to the best of our knowledge. Fig. 2 shows the corresponding schematic diagrams of these three types of matching algorithms and ours.

A. Key Points Based

Minutiae are the most popular features used for fingerprint matching [3]. Researchers define minutiae descriptors using the attributes of minutiae themselves [10] and auxiliary information (such as orientation [28], period map [29], ridge [11] etc.). Considering the lack of reliable minutiae in latent fingerprints, Cao *et al.* [30] uniformly sample images into small patches and used a network to extract fixed length vector to describe virtual minutiae. However, as explained in Fig. 1, these features are not discriminative enough in partial fingerprints. On the other hand, some studies introduce universal key points in computer vision field into fingerprints [6], [31]. Features extracted in this way are more densely distributed, but at the same time more sensitive to unreliable image details and therefore easily confused. There are also algorithms that exploit Level-3 features from high-resolution images [32], [33], which are relatively expensive to deploy in most commercial applications. After getting key point sets, correspondence between them will be established based on the similarity of features and spatial structures, and then relative poses and matching scores are calculated.

B. Fixed-length Region Representation Based

Early studies use hand-crafted statistics to characterize fingerprint regions and convert them into fixed length vectors [34]–[36]. This feature form significantly improves the speed of large database indexing at the expense of reduced accuracy, and is usually used for rough searching. Engelsma *et al.* [12] propose an end-to-end network to extract fixed-length global representations of minutiae and textures, which greatly improves the performance of fixed-length representation methods. Inspired by this, researchers have conducted more exploration and improvement of fixed-length representation methods based on deep learning [14], [15], [37], [38]. In addition, Gu *et al.* [13] conduct dense sampling on complete fingerprints and compare the fixed length descriptors of each patch, ultimately making a comprehensive decision. These algorithms usually require pose rectification in advance, such as adaptively adjusting images through a spatial transformation layer [12], [15], rectifying absolute pose of fingerprints through a separate network [38], or exploiting the positioning relationships implicit in minutiae [14]. The error of pose estimation is tolerated to a certain extent because the extracted features are a generalization of regional characteristics. However, current methods may struggle to address partial fingerprints, as the pose estimation on them is often unstable.

C. Pairwise Interrelated Feature Based

Compared with key points or fixed-length descriptors, gray-level features of pixels contain raw information about local textures. Traditional algorithms use image correlation for fingerprint verification [39], [40], which can be affected by skin distortion and has limited discriminating ability. Dense fingerprint rectification and registration can reduce the impact of skin distortion on image correlation [41], [42], but simple correlation cannot separate genuine and impostor matches very well. With the development of deep learning, researchers apply neural networks to directly extract and verify interrelated features from two input images, significantly improving the verification performance [7], [43]–[45]. Specifically, He *et al.* [7] use multi-rotation and multi-size cropped image pairs as combined inputs to further assist the network in understanding scenarios with large spatial transformations or small overlapping areas. It should be noted that a significant difference between this type of method and methods based on fixed-length region descriptor is that the relative pose of a certain fingerprint is rectified instead of the respective absolute pose. Features of such methods need to be compared with precise spatial correspondence and thus have stricter requirements for alignment. Obviously, the verification performance will significantly decline with inaccurate pose estimation [7].

III. Method

In this paper, we propose JIPNet, a CNN-Transformer hybrid network, to jointly estimate the authentication probability and relative pose of partial fingerprint pairs. Fig. 3 gives the complete flowchart of our proposed algorithm. The overall process can be divided into four phases, which are indicated by different color bars in schematic: independent feature extraction (red), interrelated feature extraction (gold), identity verification (blue), and pose prediction (green). For an input fingerprint pair, our algorithm first extracts the respective features in parallel using convolutional blocks with shared weights (in Stem and Stages 1, 2). These two sets of features are then concatenated and fed to transformer blocks in Stage 3 (along with some extra microprocessing) for sufficient information interaction both locally and globally. Finally, respective

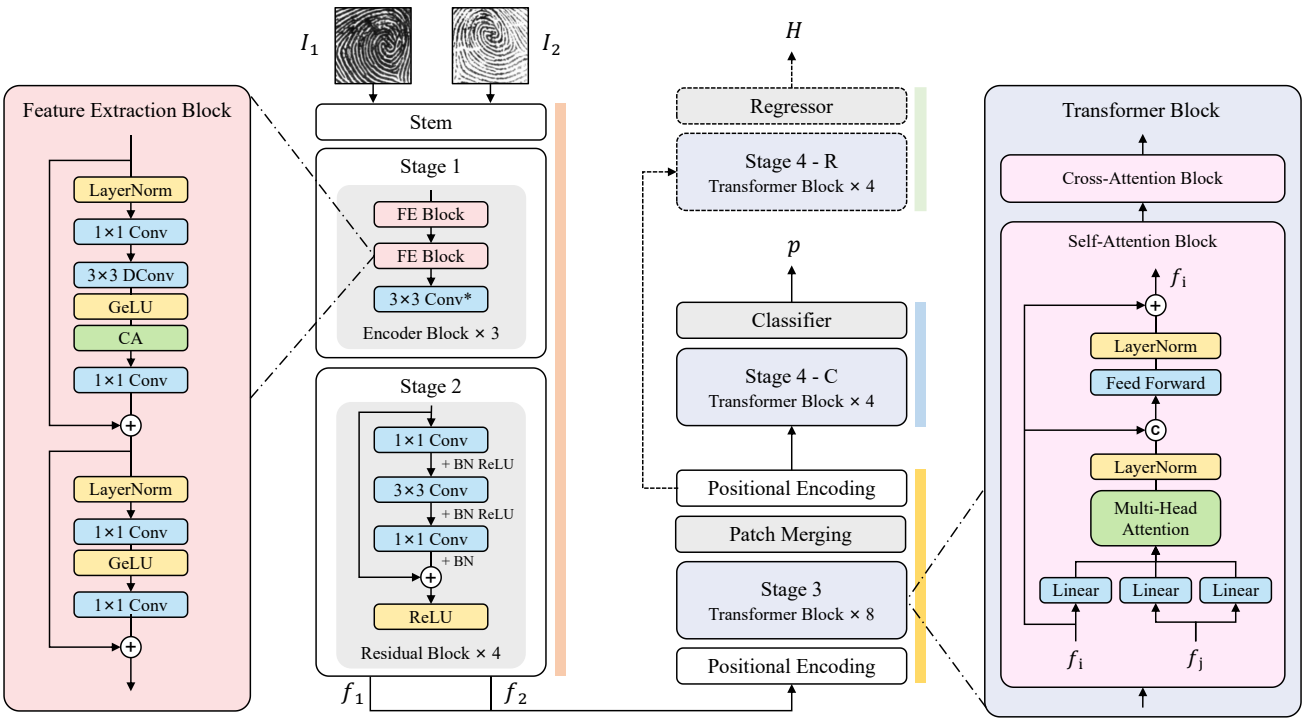


Fig. 3. An overview of JIPNet. Paired fingerprint patches with the same shape are input, specifically 160×160, 120×120, or 96×96 in this paper. The outputs 'p' and 'H' of respective task heads correspond to the classification probability (whether the input fingerprints come from the same finger) and rigid transformation parameters (relative translation and rotation) respectively. Detailed structure is shown in Table I. Bars are presented on the left to indicate each phase, where the color definition refers to Fig. 2. The process passed by parallel arrows indicates that the corresponding modules share weights and their functions can be executed in parallel. * represents the number of channels are doubled after the corresponding convolution. Dotted parts are only used to assist training and could be pruned in practical verification tasks.

TABLE I
DETAILED CONFIGURATIONS OF JIPNET. A PAIR OF 160×160 FINGERPRINT PATCHES ARE INPUTTED AS AN EXAMPLE.

Phase	Layer	Operator	Output shape
Independent Feature Extraction	Stem	3×3 Conv	$32, 160 \times 160$
	Stage 1	Encoder Block	$64, 80 \times 80$
		Encoder Block	$128, 40 \times 40$
		Encoder Block	$256, 20 \times 20$
	Stage 2	Residual Block $\times 4$	$264, 20 \times 20$
Interrelated Feature Extraction	Positional Encoding		$264, 400$
	Stage 3	Transformer Block $\times 8$	$264, 400$
	Patch Merging		$384, 100$
	Positional Encoding		$384, 100$
Identity Verification	Stage 4 - C	Transformer Block $\times 4$	$384, 100$
		Concatenate	$768, 100$
	Classifier	Linear Projection	$1, 1$
		Sigmoid	$1, 1$
Pose Prediction	Stage 4 - R	Transformer Block $\times 4$	$384, 100$
		Concatenate	$768, 100$
	Regressor	Linear Projection	$4, 1$

output layers utilize the aggregated information to predict classification probabilities for authentication (Stage 4 - C and Classifier) and regression values for relative poses (Stage 4 - R and Regressor). It should be noted that these two tasks are performed in an integrated manner instead of independently as previous methods because we hope to utilize the mutually beneficial coupling relationship between each other, inspired by works on jigsaw puzzles [17], [18]. In particular, pose-related regression layers could be pruned to improve efficiency since only identity identification is necessary in most applications.

A. Independent Feature Extraction

Previous studies have proven that early stacked convolution blocks can significantly improve the stability and performance of subsequent vision transformers [23], [46]. Similarly, we also use convolution blocks to parallelly extract independent features from each input partial fingerprint in the early stages of proposed network.

Specifically, the feature extraction module sequentially contains: (1) a 3×3 convolutional stem with stride 1 (output 32 channels); (2) stacked encoder blocks in Stage 2 for preliminary extraction of robust features, which consisting of two Feature Extraction (FE) blocks and a convolutional layer with stride 2 for downsampling; (3) stacked residual blocks in Stage 3 to further enhance the expressive ability of neural network while mitigating gradient problems in deep layers. Among them, FE block is introduced from a simple but efficient baseline for image restoration tasks [47], which is formally similar to the combination of Mobile Convolution

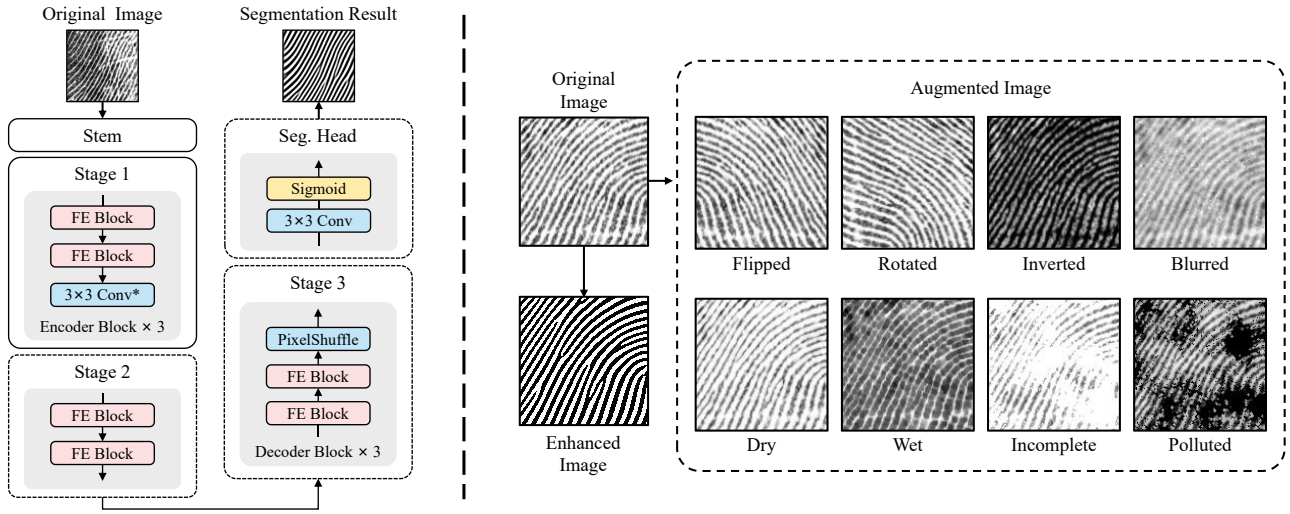


Fig. 4. Illustration of fingerprint enhancement task for pre-training. The network architecture is shown on the left, where solid and dotted boxes indicate the pre-trained parameters will (or not) be loaded into corresponding modules in JIPNet. Detailed structure is shown in Table II. The right subfigure gives representative examples of inputs (original & augmented image) and targets (enhanced image) generation.

TABLE II
DETAILED CONFIGURATIONS OF THE FINGERPRINT ENHANCEMENT NETWORK IN PRE-TRAINING TASK. A PAIR OF 160×160 FINGERPRINT PATCHES ARE INPUTTED AS AN EXAMPLE.

Layer	Operator	Output shape
Stem	3×3 Conv	$32, 160 \times 160$
Stage 1	Encoder Block	$64, 80 \times 80$
	Encoder Block	$128, 40 \times 40$
	Encoder Block	$256, 20 \times 20$
Stage 2	FE Block $\times 2$	$256, 20 \times 20$
Stage 3	Decoder Block	$128, 40 \times 40$
	Decoder Block	$64, 80 \times 80$
	Decoder Block	$32, 160 \times 160$
Seg. Head	3×3 Conv	$1, 160 \times 160$
	Sigmoid	$1, 160 \times 160$

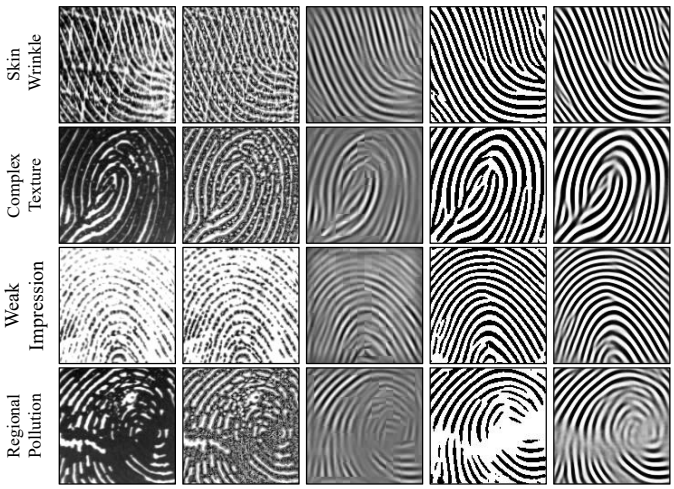


Fig. 5. Comparison of representation enhancement methods on four low-quality fingerprints. Each column from left to right is original image and corresponding enhancement results of CLAHE [54], FingerNet [55], VeriFinger [9] and our proposed method.

(MBConv) block [48] and Multilayer Perceptron (MLP) block [49], to improve the robustness of feature encoding in different image qualities. The details of corresponding architecture is presented on the left side of Fig. 3, where 'DConv' and 'CA' are depthwise convolution [48] and channel attention [50] respectively.

Furthermore, we design a pre-training task and corresponding data augmentation strategies to help the network better extract and reconstruct the essential features of fingerprints. As shown in Fig. 4, the pre-training network consists only of FE block and convolution without extra complex components, arranged in a U-shape for fingerprint enhancement follows [51], [52]. For data preparation of inputs, we randomly crop small patches (128×128 in this paper) from high-quality rolled fingerprints as prototypes to ensure sufficient texture patterns at each position while accelerating the training process. Corresponding cases after augmentation are given on the right of Fig. 4, including:

- Mirror flip, rotation (by 90, 180 or 270 degree) and grayscale inversion to increase data diversity;
- Gaussian blur and noise with different cores and variances to simulate irregular sensor noise;
- Dilation and erosion with structural elements of different sizes to simulate dry and wet fingers respectively;
- White blobs with random parameters which is added or subtracted to simulate fingerprint-specific noise in incomplete or contaminated scenarios [53].

These augmentation strategies are randomly selected and combined to generate the final input image. On the other hand, the corresponding binary fingerprint calculated by VeriFinger SDK 12.0 [9] is used as the enhancement target. Parameters of Stem and Stage 1 are then loaded into corresponding modules of JIPNet.

Fig. 5 shows representative examples of several typical enhancement methods. It can be seen that our approach is

smoother than methods based on grayscale mapping [54] or filtering [55]. In addition, our approach also shows certain robust performance when regional information is missing or wrong. This qualitative comparison strongly proves that our pre-training strategy could help the network understand more essential and complete features, and also avoids additional cumbersome enhancement steps. Further quantitative results are given in ablation experiments below.

B. Interrelated Feature Extraction

The hybrid strategy of using early CNN and subsequent transformer to aggregate local details and global information has achieved great success in multiple fields [20]–[22]. Inspired by these, we utilize transformers to capture long-range dependencies within and across paired features, thereby ensuring sufficient information interaction. The paired features f_1 and f_2 are first biased with a sinusoidal positional encoding term [20] as follows:

$$\begin{aligned} PE_{(x,y,4i)} &= \sin\left(x/10000^{4i/d}\right) \ , \\ PE_{(x,y,4i+1)} &= \cos\left(x/10000^{4i/d}\right) \ , \\ PE_{(x,y,4i+2)} &= \sin\left(y/10000^{4i/d}\right) \ , \\ PE_{(x,y,4i+3)} &= \cos\left(y/10000^{4i/d}\right) \ , \end{aligned} \tag{1}$$

where x and y denote the 2D position, i and d denote the current and total dimensions. Features with added absolute positional information are then spatially flattened into 1D sequences and fed into stacked transformers with interleaved self- and cross- attention. It should be noted that the next two attention blocks have the same structure, including residual-connected Multi-Head Attention (MHA) and Feed-Forward Network (FFN) [49], and only the data streams are differentiated by their respective expected effects. Specifically, let $\mathrm{SA}\left(f_i, f_j\right)$ (Self-Attention) and $\mathrm{CA}\left(f_i, f_j\right)$ (Cross-Attention) represent the corresponding sub-blocks in Fig. 3, the complete process of transformer block is defined as:

$$\begin{aligned} f_1^{\mathrm{S}} = \mathrm{SA}\left(f_1, f_1\right) \ , \ f_2^{\mathrm{S}} = \mathrm{SA}\left(f_2, f_2\right) \ , \\ f_1^{\mathrm{C}} = \mathrm{CA}\left(f_1^{\mathrm{S}}, f_2^{\mathrm{S}}\right) \ , \ f_2^{\mathrm{C}} = \mathrm{CA}\left(f_2^{\mathrm{S}}, f_1^{\mathrm{S}}\right) \ . \end{aligned} \tag{2}$$

The linear projection of f_i and f_j in MHA from left to right are $Query$, Key and $Value$ matrices, which are then used for dot-product attention formally as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \ , \tag{3}$$

where d_k is the Key dimention and the number of heads is fixed to 6 at this stage. In order to further condense features while reducing information loss, a patch merging layer [56] is subsequently incorporated for downsampling. Finally, the positional encoding is applied again in the same way of Equation 1.

C. Identification Verification & Pose Prediction

Similar to the previous phase, features are separately fed into stacked identical transformer blocks with head of 12 to prepare for the corresponding task head. We define identity verification as a binary classification task, that is, judging whether the input image pair comes from a same finger. Let $\mathcal{F}$, BN, Swish and $\mathrm{AvgPool}$ denote fully connected layer, batch normalization, swish activation and adaptive average pooling respectively, the classification probability p is calculated as:

$$\begin{aligned} \mathrm{MLP}[\cdot] &= \{\mathcal{F}, \mathrm{BN}, \mathrm{Swish} \ , \mathcal{F}, \mathrm{BN}, \mathrm{Swish}\} \ , \\ \mathrm{Cla}[\cdot] &= \{\mathcal{F}, \mathrm{AvgPool}, \mathrm{MLP}\} \ , \\ p &= \mathrm{Sigmoid}(\mathrm{Cla}[f]) \ , \end{aligned} \tag{4}$$

where f is the concatenated by the last f_1^{C} and f_2^{C} in channel dimension, operator $\{\ldots\}$ means executing the contained functions sequentially from left to right. On the other hand, we treat pose prediction as a regression task and predicting the rigid transformation parameters H as:

$$\begin{aligned} \mathrm{Reg}[\cdot] &= \{\mathcal{F}, \mathrm{AvgPool}, \mathrm{MLP}\} \ , \\ H &= \mathrm{Reg}[f] \ . \end{aligned} \tag{5}$$

Following previous work [7], we represent the rotation relationship based on relative values of sine and cosine to circumvent the constraints on numerical range of angles, and directly describe the translation in terms of horizontal and vertical displacements. With the prediction target of $H = [r_{\mathrm{c}}, r_{\mathrm{s}}, t_{\mathrm{x}}, t_{\mathrm{y}}]$, the rigid alignment relationship of paired fingerprints can be modeled as:

$$\begin{aligned} &\cos\theta = r_{\mathrm{c}}/\sqrt{r_{\mathrm{c}}^2 + r_{\mathrm{s}}^2} \ , \ \sin\theta = r_{\mathrm{s}}/\sqrt{r_{\mathrm{c}}^2 + r_{\mathrm{s}}^2} \ , \\ &\begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & t_{\mathrm{x}} \\ \sin\theta & \cos\theta & t_{\mathrm{y}} \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} u_2 \\ v_2 \\ 1 \end{bmatrix} \ , \end{aligned} \tag{6}$$

where θ is the relative rotation angle, (u, v) represents any corresponding position in images 1 and 2.

D. Loss Function

We optimize our network parameters through a comprehensive objective function consisting of two items. The total loss can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cla}} + \lambda \cdot \mathcal{L}_{\mathrm{reg}} \ , \tag{7}$$

where $\mathcal{L}_{\mathrm{cla}}$ and $\mathcal{L}_{\mathrm{reg}}$ are the identity classification term and pose regression term, respectively. The trade-off parameter λ is set to 0.002 in this paper. Additionally, an extra image segmentation loss $\mathcal{L}_{\mathrm{seg}}$ is proposed for fingerprint enhancement, which is applied only during pre-training.

1) Identity Verification Loss: For classification of identity verification, we adapt focal loss [57] to help our model better focus on difficult samples:

$$\mathcal{L}_{\mathrm{cla}} = \begin{cases} -\alpha\left(1-p\right)^{\gamma}\log p, & \text{if } y = 1 \\ -\left(1-\alpha\right)p^{\gamma}\log(1-p), & \text{if } y = 0 \end{cases} \ , \tag{8}$$

where p represents the predicted probability, y represents the 0-1 label of ground truth. Balance factors α and γ are set to 0.2 and 2.0 empirically.

2) Pose Alignment Loss: For regression of relate pose, weighted mean square error is utilzed for supervision:

$$\begin{aligned}\mathcal{L}_{\mathrm{reg}} = \omega \cdot \left[(\cos\theta^* - \cos\theta)^2 + (\sin\theta^* - \sin\theta)^2\right] \\ + (1-\omega) \cdot \left[\left(t_{\mathrm{x}}^* - t_{\mathrm{x}}\right)^2 + \left(t_{\mathrm{y}}^* - t_{\mathrm{y}}\right)^2\right]\end{aligned} \tag{9}$$

where the superscript * indicates the ground truth, otherwise the estimated result. The relative angle θ and pixel level displacement t is converted from the originally defined output form H and Equation 6. Same as [7], we fix ω to 0.99 to balance the optimization of rotation and translation.

3) Image Segmentation Loss: We enhance the input fingerprints in form of binary segmentation, where the grayscale differences on modalities are minimized to make the texture pattern of ridges more prominent. A simplified focal loss [57] is defined as the corresponding metric:

$$\begin{aligned}\mathcal{L}_{\mathrm{seg}} = -\frac{1}{|M|}\sum_{M}\left(1-q\right)^{\gamma}\log\left(q\right), \\ q = y \cdot p + (1-y)\cdot(1-p),\end{aligned} \tag{10}$$

where p is the probability of foreground, y is the ground truth of corresponding binary image obtained by VeriFinger [9], and M is the image mask. The hyperparameter γ is fixed to 2.0 in experiments.

IV. Dataset Description

Comprehensive evaluations are conducted on extensive public datasets, including the representative public benchmarks *NIST SD14* [24], *FVC2002 DB1_A & DB3_A* [25], *FVC2004 DB1_A & DB2_A* [26], and *FVC2006 DB2_A* [27]. In addition, we also use a private partial fingerprint dataset *THU Small* to further examine the performance in capacitive smartphone situations. Table III presents a detailed description about the composition and usage of these datasets. In particular, we merge several fingerprint databases into a larger dataset, called the *Hybrid Database (Hybrid DB)* , to make the implementation of various experiments more convenient and sufficient. On the other hand, we further evaluate the performance of cross-domain generalization on other separate datasets. Rolled fingerprint is utilized because it covers a more complete area at different pressing poses than plain fingerprints. Image examples of different datasets are shown in Fig. 6. It can be seen that datasets selected in our experiments contains diverse fingerprint impressions with various modalities (normal, dry, wet, polluted, wrinkled, etc.).

A. Parital Fingerprint Pair Simulation

Considering the lack of publicly available datasets of partial fingerprints, image patches of specified sizes are synthesized from rolled or plain datasets, similar to previous works [7], [16]. For any two fingerprints from the same finger, we first use VeriFinger [9] to align them based on matching minutiae. Alignment in this way is accurate in most cases because original image pairs generally have large enough overlap areas of high quality. Subsequently, we calculated the common mask of aligned fingerprints and randomly select the patch center of one partial fingerprint from it. Taking this center as the zero point, another patch center is uniformly sampled in polar coordinates with a random angle ranging from -180° to 180° and random radius ranging from 0 pixel to $100\,/\,70\,/\,20$ pixels. Finally, full fingerprints are cropped into partial images of 160×160, 128×128 and 96×96 with a random relative rotation from $[-180^\circ, 180^\circ]$, aiming to comprehensively evaluate the performance under different partial fingerprint sizes. The parameters of rigid translation and rotation are converted to ground truth H according to Equation 6. It should be noted that each pair of partial fingerprints are cropped from different original images to avoid possible information leakage caused by the same grayscale relationship or distortion pattern.

B. Matching Protocols

In order to balance the number of genuine / impostor matches and different fingerprint types / modalities, we implemented specific matching protocols during training and testing. For the *FVC* series, each fingerprint is combined with all mated fingerprints (with order) as genuine match pairs, and the first image of each pair is combined with a random impression of another finger as a new impostor match pair. Taking *FVC2002 DB1_A* as an example, a total of $8 \times 7 \times 100 \times 2 = 11,200$ image pairs are combined in this way. This protocol is also applied to *NIST SD14* with a slightly different that symmetric matches are avoided, considering the large number of rolled images. All combined pairs of full fingerprints are then randomly cropped to simulate several partial fingerprint scenes, with the number of executions set to 4 (NIST SD14, FVC2004 DB1_A & DB2_A, FVC2006 DB2_A) for training / testing and 1 (*FVC2002 DB1_A & DB3_A*) for testing. As mentioned above, some representative datasets are merged as *Hybrid DB* and subsequently divided, secifically $430,768$ for training and $22,672$ for testing, which are distinguished by suffixes '_A' and '_B' in the following experiments. The selection of genuine matches in real partial fingerprint dataset *THU Small* also follows this principle, while pairing the first impression of every finger with each other as impostor pairs. All detailed usage information is listed in corresponding footnotes of Table III.

V. Experiments

In this section, several representative state-of-the-art fingerprint algorithms are compared with our proposed approach, including:

- **A-KAZE** [6], a simple but comparative partial fingerprint recognition algorithm based on commonly used key points;
- **VeriFinger** SDK 12.0 [9], a widely used, top performing commercial software mainly based on minutiae;
- **DeepPrint** [12], a global fixed-length representation extracted by deep network, which is highly influential;
- **DesNet** [13], a descriptor network mainly designed for latent fingerprint, used to extract localized deep descriptors of densely sampled patches;
- **AFR-Net** [15], an attention-driven fingerprint recognition network that extracts complementary global representations through ViT and ResNet embeddings.

TABLE III
ALL FINGERPRINT DATASETS USED IN EXPERIMENTS.

Type	Dataset	Scanners	Description	Usage	Genuine pairs	Impostor pairs
Rolled	NIST SD14[a] [24]	Inking	27,000 fingers × 2 impressions	train & test[c]	108,000	108,000
Plain	THU Small[b]	Capacitive	100 fingers × 8 impressions	test	5,600	9,900
	FVC2002 DB1_A [25]	Optical	100 fingers × 8 impressions	test	5,600	5,600
	FVC2002 DB3_A [25]	Capacitive	100 fingers × 8 impressions	test	5,600	5,600
	FVC2004 DB1_A [26]	Optical	100 fingers × 8 impressions	train & test[c]	22,400	22,400
	FVC2004 DB2_A [26]	Optical	100 fingers × 8 impressions	train & test[c]	22,400	22,400
	FVC2006 DB2_A [27]	Optical	140 fingers × 12 impressions	train & test[c]	73,920	73,920

[a] NIST SD14 was publicly available but later removed from public domain by NIST.
[b] Corresponding dataset is private. Fingerprints are entered by volunteers in any comfortable pressing pose, with no special declaration or device restrictions. That is, there are quite a few genuine matching pairs with little or no overlap.
[c] Datasets are merged and proportionally divided for training (95 %, called Hybrid DB_A) and testing (5 %, called Hybrid DB_B). Identities between these two subsets are completely isolated.

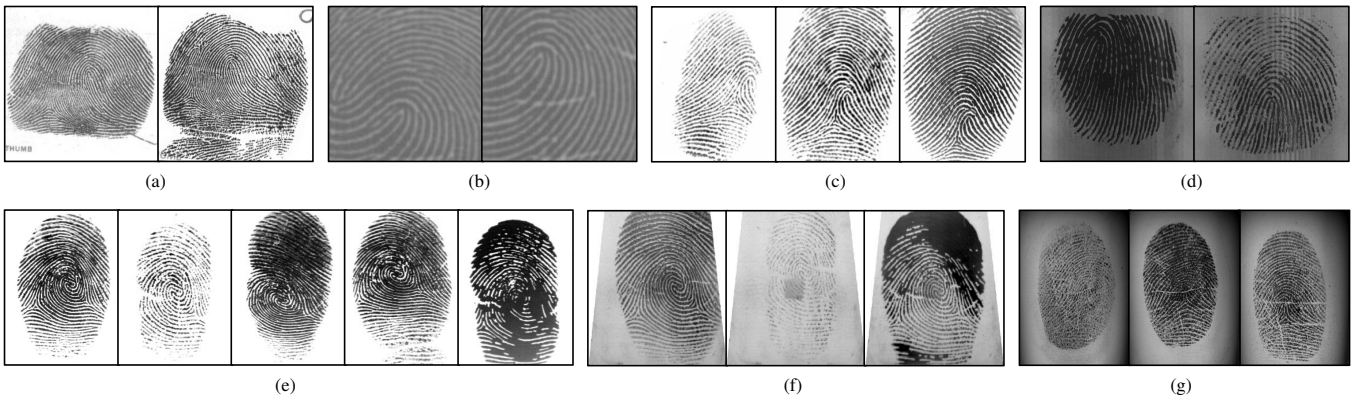


Fig. 6. Image examples from different fingerprint datasets (a) NIST SD14 [24], (b) THU Small, (c) FVC2002 DB1_A [25], (d) FVC2002 DB3_A [25], (e) FVC2004 DB1_A [26], (f) FVC2004 DB2_A [26], (g) FVC2006 DB2_A [27].

- **PFVNet** [7], a state-of-the-art partial fingerprint verification algorithm recently proposed, which introduces multi-level features fusion and local self-attention mechanism.

In particular, since the effective information of partial fingerprints is significantly reduced compared to full-size fingerprints, the original pose rectification pre-step adopted in DeepPrint [12], DesNet [13] and AFR-Net [15] cannot be performed as expected. Therefore, in this paper we use AlignNet, part of PFVNet [7], to estimate and correct the relative pose of image pairs as substitute, which is marked by subscript '*'.

We conducted extensive experiments to thoroughly evaluate the performance of above algorithms under partial fingerprints at different sizes, which in terms of score distribution, matching performance, alignment accuracy, visual analysis and efficiency. In addition, ablation studies are demonstrated to validate the effectiveness of corresponding modules and strategies proposed in this paper.

A. Implementation Details

All partial fingerprint training processes are performed on the generated dataset *Hybrid DB_A* with an initial learning rate of $1\mathrm{e}{-3}$ (end of $1\mathrm{e}{-6}$), cosine annealing scheduler, default AdamW optimizer and batch size of 128 until convergence (about 12 epochs). The pre-training task proposed in Section III-A is trained for 200 epochs using a subset of $5,000$ rolled images with corresponding random augmentation. On the other hand, we completely reimplemented fixed-length representation based methods [12], [13] on full-size fingerprints in corresponding manner. *Hybrid DB_B*, which has the same source as training set, is used to evaluate the performance under same data distribution. Besides, scenarios of other image sizes and datasets are directly tested without fine-tuning to reflect the cross-domain adaptability.

B. Score Distribution

We first present the score distribution of genuine and impostor matches on *Hybrid DB_B* with image size 160×160 to qualitatively demonstrate the discriminative capabilities of different algorithms. Corresponding curves are shown in Fig. 7, where the scores of A-KAZE [6], VeriFinger [9], and DesNet [13] are linearly mapped to appropriate ranges that approximate $[0, 1]$ for intuitive comparison. A more separated pairwise distribution (i.e., fewer overlapping areas) indicates that the corresponding algorithm can better distinguish whether a certain fingerprint pair comes from the same finger.

Moreover, three common indicators are further computed and reported in Table IV to quantitatively measure the differences between genuine and impostor probability distributions.

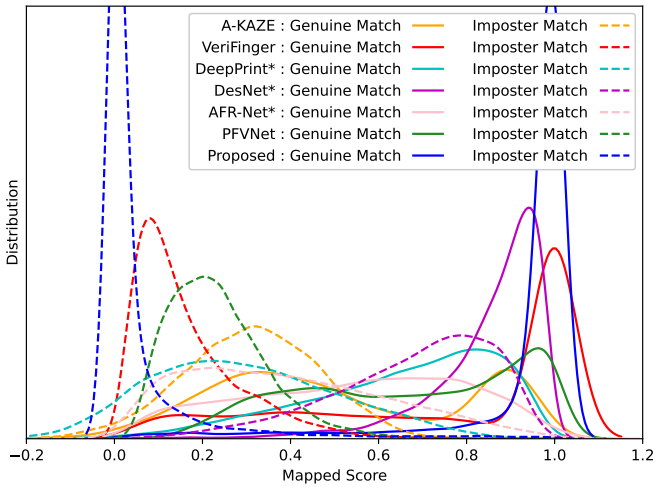


Fig. 7. Probability density distributions of genuine and impostor matching scores on Hybrid DB_B with image size 160 × 160. Scales on vertical axis are hidden because we are more concerned with the relative values.

TABLE IV
PROBABILITY DISTRIBUTION DISTANCE BETWEEN GENUINE AND IMPOSTOR SCORES ON HYBRID DB_B WITH IMAGE SIZE 160 × 160.

Method	JS	EMD	MMD
A-KAZE [6]	0.42	0.20	0.72
VeriFinger [9]	0.70	0.59	3.21
DeepPrint* [12]	0.59	0.36	2.03
DesNet* [13]	0.47	0.17	1.10
AFR-Net* [15]	0.26	0.15	0.43
PFVNet [7]	0.73	0.43	3.01
Proposed	**0.89**	**0.87**	**6.36**

Among them, the Jensen–Shannon (JS) divergence is defined as:

$$\begin{aligned} \mathrm{KL}\left(p\|q\right) &= \int p(x)\log\frac{p(x)}{q(x)}dx\ , \\ \mathrm{JS}\left(p\|q\right) &= \frac{1}{2}\mathrm{KL}\left(p\|\frac{p+q}{2}\right)+\frac{1}{2}\mathrm{KL}\left(q\|\frac{p+q}{2}\right)\ , \end{aligned} \tag{11}$$

where p and q correspond to any two distributions. Let $\Pi(p,q)$ represent all possible joint distributions of p and q, F represent the functions of sample space, the Earth-Mover's Distance (EMD) and Maximum Mean Discrepancy (MMD) are calculated by

$$\mathrm{EMD}\left(p,q\right) = \inf_{\gamma\sim\Pi(p,q)} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|]\ , \tag{12}$$

and

$$\mathrm{MMD}\left(F,p,q\right) = \max_{f\in F}\left(E_{x\sim p}[f(x)]-E_{y\sim q}[f(y)]\right)\ . \tag{13}$$

Similarly, the larger these values, the farther between corresponding distribution pairs, which normally means a better identity verification performance.

According to the results in Fig. 7 and Table IV, our method outperforms all other methods, while VeriFinger [9] and PFVNet [7] each have sub-optimal performance in some parts. In particular, the distinguishing ability of DesNet* [13] is not ideal because it is designed based on the prior assumption of high-precision alignment, which is difficult to satisfy in current partial fingerprint situations. Another noteworthy phenomenon is that the discriminative ability of AFR-Net [15] is not satisfactory in this small image patch scenario, despite its advanced performance on rolled/plain fingerprint matching. We believe that its attention driven architecture effectively increases the ability to express local textures. However, this high sensitivity may actually limit robustness when there is limited available information (small effective fingerprint area) or excessive interference (limited overlap ratio).

C. Matching Performance

In this subsection, four classic indicators are adopted to comprehensively evaluate methods from the perspective of different deployment scenarios, including:

- Accuracy (ACC) under optimal thresholds, to intuitively evaluate the correctness of model prediction;
- Area Under Curve (AUC) based on Receiver Operating Characteristic (ROC) curves, to approximately evaluate the overall classification performance;
- True Accept Rate (TAR) under specified False Accept Rate (FAR) of 1e-3, which is often used to measure the security performance of biometric solutions.
- Equal Error Rate (EER), where both false alarm and impostor pass errors are equal.

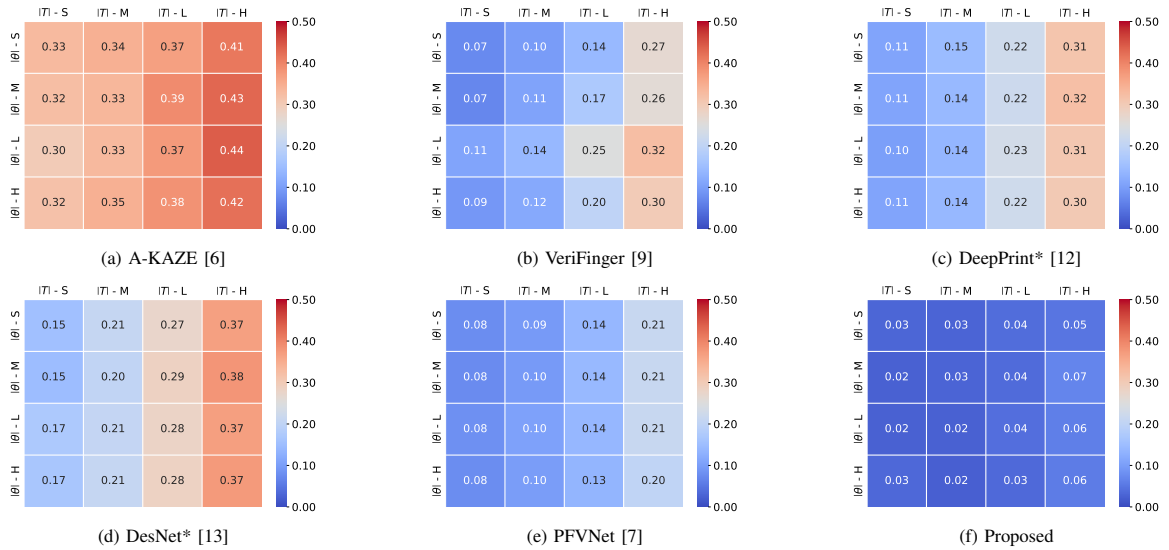
Table V presents the corresponding matching performance of state-of-the-art algorithms and our proposed approach on several representative datasets of fingerprints with different sizes. Furthermore, we group the genuine matches on *Hybrid DB_B* according to their relative rigid transformation relationship and compare these algorithms under different rotations and translations using EER. Fig. 8 shows the corresponding comparisons of representative methods for each type of matching framework. Additionally, the Detection Error Tradeoff (DET) curves on all test databases are calculated in Fig. 9 to provide more complete information for analysis.

These encouraging experimental results strongly demonstrates the superiority of our proposed JIPNet, which outperforms previous advanced methods in almost all cases. PFVNet [7] also shows remarkable performance, especially in indicators of ACC and AUC. However, it can be observed that the relative performance deteriorates markedly when training and testing data cross domains. Region representation based methods [12], [13], [15] are inferior to others, which is reasonable because the highly generalized form of fixed-length descriptors sacrifices many potential details for faster search speed. On the other hand, emphasizing certain texture features located in non overlapping areas may also lead to incorrect mismatches. That is to say, genuine matches with very small overlapping areas and impostor matches with very similar local patterns lack sufficient stable discrimination under such regional representations, which is also reflected in Fig. 7. Specifically, although localized descriptors with deeper dimensions are designed in DesNet [13], their effectiveness is limited, as stated in Section V-B. On the other hand, the performance of all algorithms drops significantly as the

TABLE V
MATCHING PERFORMANCE OF STATE-OF-THE-ART ALGORITHMS AND OUR APPROACH ON FOUR REPRESENTATIVE DATASETS OF PARTIAL FINGERPRINT WITH THREE DIFFERENT SIZES.

Size	Sketch	Method	Hybrid DB_B				FVC2002 DB1_A				FVC2002 DB3_A				THU Small			
			ACC	AUC	TAR	EER	ACC	AUC	TAR	EER	ACC	AUC	TAR	EER	ACC	AUC	TAR	EER
160 × 160	Fig. 2a	A-KAZE [6]	0.67	0.70	0.28	0.37	0.82	0.86	0.56	0.22	0.73	0.77	0.36	0.31	0.89	0.87	0.68	0.21
		VeriFinger [9]	0.85	0.88	0.56	0.18	0.95	0.97	0.85	0.07	0.90	0.93	**0.76**	0.12	0.92	0.91	**0.76**	0.15
	Fig. 2b	DeepPrint* [12]	0.79	0.87	0.16	0.21	0.84	0.92	0.32	0.16	0.81	0.89	0.16	0.19	0.82	0.86	0.26	0.22
		DesNet* [13]	0.74	0.81	0.05	0.27	0.79	0.87	0.16	0.21	0.76	0.84	0.08	0.25	0.77	0.78	0.20	0.29
		AFRNet* [15]	0.64	0.67	0.01	0.37	0.71	0.77	0.03	0.29	0.65	0.70	0.01	0.35	0.65	0.62	0.01	0.40
	Fig. 2c	PFVNet [7]	0.86	0.94	0.47	0.14	0.90	0.96	0.61	0.10	0.90	0.96	0.39	0.11	0.84	0.87	0.47	0.22
	Fig. 2d	Proposed	**0.96**	**0.99**	**0.71**	**0.04**	**0.99**	**0.99**	**0.93**	**0.01**	**0.97**	**0.99**	0.67	**0.03**	**0.92**	**0.97**	0.70	**0.10**
128 × 128	Fig. 2a	A-KAZE [6]	0.60	0.63	0.22	0.39	0.69	0.73	0.45	0.27	0.63	0.67	0.20	0.34	0.79	0.74	0.40	0.34
		VeriFinger [9]	0.72	0.74	0.39	0.27	0.82	0.85	0.71	0.13	0.77	0.79	0.55	0.20	0.83	0.80	0.51	0.28
	Fig. 2b	DeepPrint* [12]	0.67	0.73	0.05	0.29	0.69	0.75	0.14	0.26	0.69	0.75	0.09	0.27	0.70	0.69	0.05	0.37
		DesNet* [13]	0.64	0.69	0.03	0.32	0.66	0.71	0.05	0.29	0.64	0.69	0.04	0.32	0.70	0.67	0.06	0.38
		AFRNet* [15]	0.54	0.54	0.01	0.45	0.56	0.57	0.01	0.40	0.54	0.54	0.01	0.43	0.63	0.52	0.01	0.49
	Fig. 2c	PFVNet [7]	0.73	0.81	0.16	0.24	0.71	0.79	0.18	0.25	0.73	0.82	0.16	0.24	0.73	0.74	0.14	0.33
	Fig. 2d	Proposed	**0.91**	**0.97**	**0.63**	**0.05**	**0.94**	**0.99**	**0.89**	**0.02**	**0.91**	**0.97**	**0.60**	**0.05**	**0.88**	**0.93**	**0.53**	**0.15**
96 × 96	Fig. 2a	A-KAZE [6]	0.52	0.52	0.07	0.49	0.54	0.54	0.26	0.45	0.52	0.52	0.07	0.48	0.66	0.53	0.06	0.47
		VeriFinger [9]	0.59	0.60	0.25	0.31	0.65	0.66	0.53	0.20	0.62	0.63	0.26	0.24	0.74	0.67	0.25	0.37
	Fig. 2b	DeepPrint* [12]	0.61	0.64	0.05	0.30	0.59	0.61	0.01	0.30	0.61	0.64	0.08	0.27	0.64	0.56	0.02	0.45
		DesNet* [13]	0.58	0.60	0.02	0.35	0.57	0.60	0.02	0.32	0.56	0.58	0.06	0.32	0.65	0.58	0.02	0.45
		AFRNet* [15]	0.51	0.51	0.01	0.48	0.51	0.50	0.01	0.48	0.50	0.50	0.01	0.48	0.44	0.51	0.01	0.49
	Fig. 2c	PFVNet [7]	0.64	0.71	0.09	0.30	0.60	0.65	0.10	0.31	0.62	0.69	0.14	0.28	0.66	0.64	0.02	0.41
	Fig. 2d	Proposed	**0.79**	**0.87**	**0.48**	**0.08**	**0.81**	**0.89**	**0.77**	**0.04**	**0.79**	**0.87**	**0.63**	**0.07**	**0.79**	**0.82**	**0.28**	**0.26**

TAR represents TAR @ FAR = 1e-3.



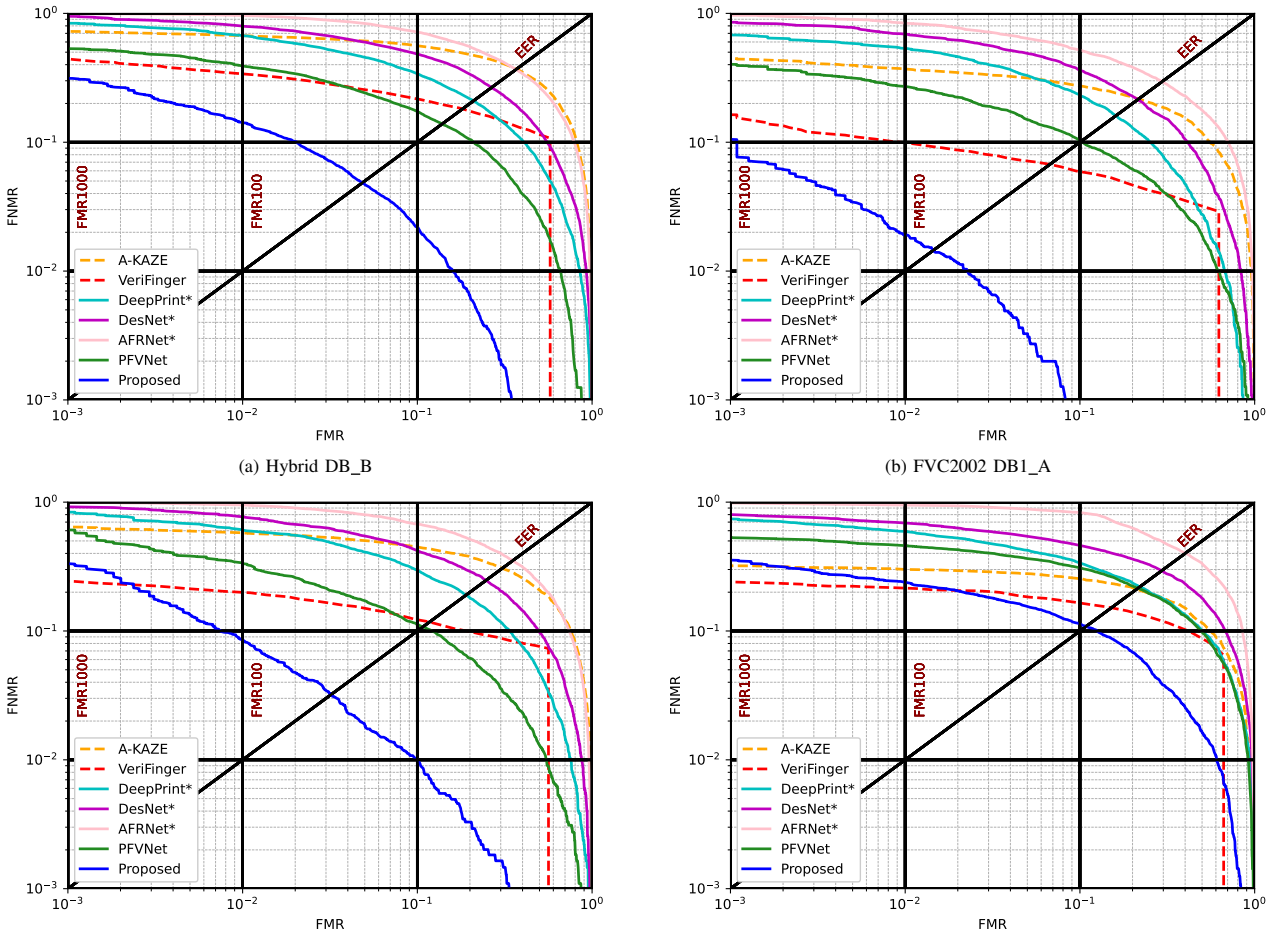
(a) A-KAZE [6] (b) VeriFinger [9] (c) DeepPrint* [12]

(d) DesNet* [13] (e) PFVNet [7] (f) Proposed

Fig. 8. Equal Error Rate matrix under different degrees of relative rigid transformation on Hybrid DB_B with image size 160 × 160. Genuine matches are divided into 16 subsets based on the value of relative rotation θ and translation T, which in terms of Small, Medium, Large, and Huge segmented evenly by numerical range. Impostor matches are not modified again.

image size decreases or distance increases, but our scheme still leads and exhibits attractive robustness and stability. It is worth noting that the impact of rotation is not evident in Fig. 8, because key points based matching score do not require pose rectification and the prediction error in other algorithms remains at a similar level across different relative angles (confirmed in the following experiments).

As analyzed above, these compared methods focus on different aspects of fingerprint attributes, exhibiting differentiated trends in score distribution and matching advantages. This



(a) Hybrid DB_B (b) FVC2002 DB1_A

(c) FVC2002 DB3_A (d) THU Small

Fig. 9. DET curves of state-of-the-art algorithms and our approach on partial fingerprints of 160 × 160. Solid and dotted lines represent deep learning methods and traditional methods respectively.

TABLE VI
MATCHING PERFORMANCE OF DIFFERENT SCORE LEVEL FUSION STRATEGIES WITH IMAGE SIZE 160 × 160.

Dataset	**VeriFinger** [9] + DeepPrint* [12]	**PFVNet** [7] + VeriFinger [9]	**PFVNet** [7] + DeepPrint* [12]	**Proposed** + VeriFinger [9]	**Proposed** + DeepPrint* [12]	**Proposed** + PFVNet [7]
Hybrid DB_B	0.57 ↑0.01	0.66 ↑0.19	0.47 –	**0.76** ↑0.05	0.74 ↑0.03	0.71 –
FVC2002 DB1_A	0.86 ↑0.01	0.87 ↑0.26	0.62 ↑0.01	0.94 ↑0.01	0.92 ↓0.01	**0.95** ↑0.02
FVC2002 DB3_A	0.76 –	0.77 ↑0.38	0.49 ↑0.10	**0.87** ↑0.20	0.68 ↑0.01	0.73 ↑0.06
THU Small	0.76 –	0.77 ↑0.30	0.47 –	**0.80** ↑0.10	0.66 ↓0.04	0.66 ↓0.04

TAR @ FAR = 1e-3 is reported.
Scores of bold method are used as the baselines of corresponding columns. '+' indicates another type of method for fusion.

phenomenon motivates us to conduct multiple fusion strategies to further observe their complementarity. Specifically, scores from two sets are linearly weighted and summed with the best classification accuracy. As represented in Table VI, most fusions effectively improves the corresponding performance, especially the introduction of minutiae information (VeriFinger) into other methods that rely on texture features, while our method still leads the way.

D. Alignment Accuracy

The original intention of introducing the pose estimation sub stage in our network is to help it better understand the spatial contrast relationship and thus facilitate partial fingerprint verification. Nevertheless, we still assess the alignment accuracy as it is indeed a specific aspect that can be effectively compared, allowing us to gauge their capability to rectify relative positions and potential influence on subsequent matching. Similar to [7], we express the estimation error in the form of

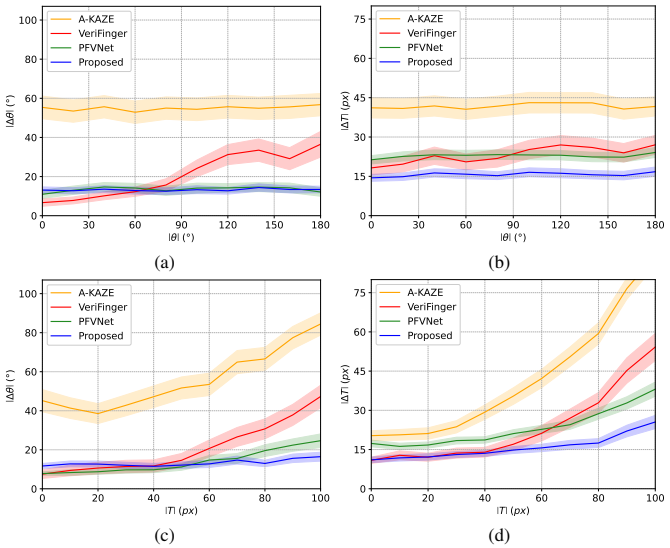


Fig. 10. Estimation errors of alignment on Hybrid DB_B in image size 160×160. (a) Rotation errors and (b) Translation errors under varying $|\theta|$. (c) Rotation errors and (d) Translation errors under varying $|T|$. Lines and corresponding light-colored bands reflect the mean and standard deviation.

translation and rotation as:

$$\begin{aligned} |\Delta\theta| &= |\theta^* - \theta| \,, \\ |\Delta T| &= \sqrt{(t_{\mathrm{x}}^* - t_{\mathrm{x}})^2 + (t_{\mathrm{y}}^* - t_{\mathrm{y}})^2} \,, \end{aligned} \tag{14}$$

where symbols are defined the same as in Equation 9. As shown in Fig. 10, most methods can provide effective predictions within an acceptable range, while our approach has a slight edge. Besides, it can be seen that estimation errors of both rotation and translation increase significantly with distance and remaining roughly stable for angle. One convincing explanation is that distance greatly alters the overlap area, which drastically affects the internal feature comparison and subsequent decision-making process.

E. Ablation Study

Extensive ablation experiments are conducted to examine the effects of different modules and strategies introduced in our network. EER and $\mathcal{L}_{\mathrm{reg}}$ (defined in Equation 9) are used to measure the performance of matching and alignment, respectively. As shown in Table VII, overall, the CNN based approach outperforms Vision transformer (ViT) based approach, while the CNN-Transformer hybrid series achieves the best performance. We attribute it to the respective advantages of convolutional and transformer blocks in extracting local features and establishing global connections. In addition, the introduction of parallel relative alignment estimation head achieves great success in EER metric across all three types of network architectures, which proves that there is indeed an exploitable promotion relationship between pose prediction and identity verification. Furthermore, the joint estimation strategy also shows certain positive effect on pose prediction task when applied in networks containing ViT decoders. At the same time, the additional alignment head can be pruned during deployment so that the matching speed will not be affected.

TABLE VII
ABLATION STUDY OF THE PROPOSED NETWORK WITH DIFFERENT MODULES AND STRATEGIES ON FVC2002 DB1_A WITH IMAGE SIZE 160×160.

Encoder *		Decoder *		+ ID Head *	+ Pose Head *	EER	$\mathcal{L}_{\mathrm{reg}}$
CNN	ViT	CNN	ViT				
✓		✓		✓		0.25	-
✓		✓			✓	-	18.2
✓		✓		✓	✓	0.21	20.1
	✓		✓	✓		0.47	-
	✓		✓		✓	-	27.2
	✓		✓	✓	✓	0.29	26.7
✓			✓	✓		0.17	-
✓			✓		✓	-	10.6
✓			✓	✓	✓	0.07	9.2
✓§			✓	✓	✓	**0.01**	**1.1**

§ indicates the corresponding module has been pretrained through the process shown in Figure 4.
* each represents specific stages in proposed network, specifically *Encoder* for Stage 1 & 2, *Decoder* for Stage 3, *ID Head* for Stage 4 - C, *Pose Head* for Stage 4 - R.

Finally, the comparison between last two rows demonstrated that our proposed lightweight pre-training task can further improve the performance at a small extra training time cost. A convincing explanation is that it ensures the extracted features contain complete pattern information of fingerprint ridges, while also performing a certain degree of denoising. In this way, the inputs for subsequent two modules, namely identity verification and pose prediction, is transferred from the image level to the feature level, which assists the network to better focus on key features and perform joint optimization more seamlessly.

F. Visual Analysis

In order to provide more specific and intuitive comparison with interpretability, we employ occlusion sensitivity [16] to approximate the contribution of distinct local areas to the overall matching judgment. Fig. 11 shows four representative visualization results. Query fingerprint in images pairs is centered and search fingerprint is rigidly aligned based on the ground truth as background. Subsequently, a 32×32 mask is slid and applied to search fingerprints and the differences between scores before and after occlusion are calculated, which are normalized and overlaid as the heatmap of search fingerprint. Considering that there is no reference relationship between the matching scores directly obtained (or mapped) by different algorithms, we convert these scores into a comparable probability form and attached them in corresponding visualization results. Specifically, given the overall distributions (shown in Fig. 7), the verification accuracy can be expressed as:

$$\begin{aligned} P\left(y_1|s\right) &= \frac{P\left(s|y_1\right)P\left(y_1\right)}{P\left(s|y_1\right)P\left(y_1\right) + P\left(s|y_0\right)P\left(y_0\right)} \\ &= \frac{f_1\left(s\right)P\left(y_1\right)}{f_1\left(s\right)P\left(y_1\right) + f_0\left(s\right)P\left(y_0\right)} \,, \end{aligned} \tag{15}$$

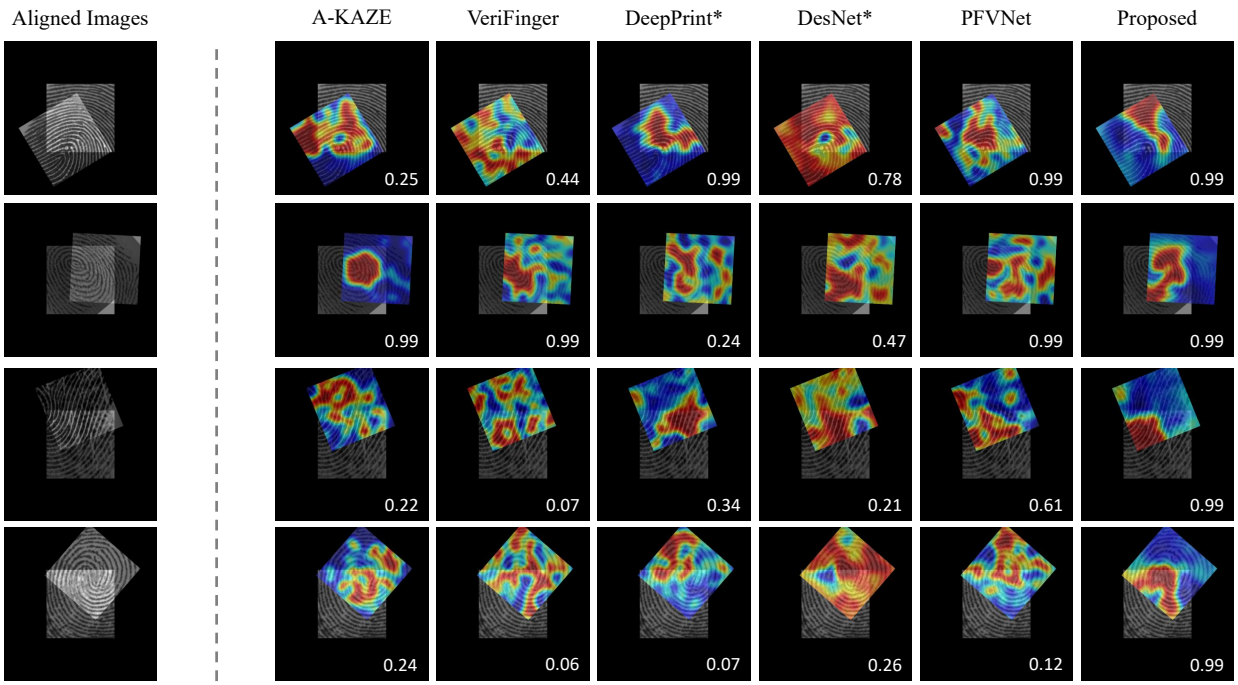


Fig. 11. Representative visualization results of genuine matches based on occlusion sensitivity [16]. Red areas indicate a large influence on matching judgment, while blue areas indicate the opposite. Numbers on the bottom right are the accuracy of corresponding methods when they are verified.

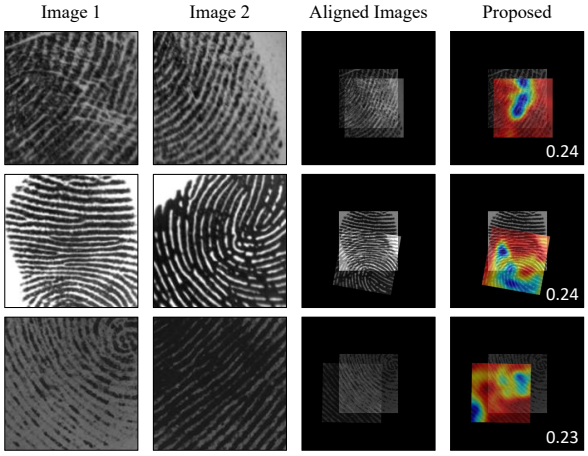


Fig. 12. Challenging cases where our approach fails. The way of visualization is the same as Fig. 11.

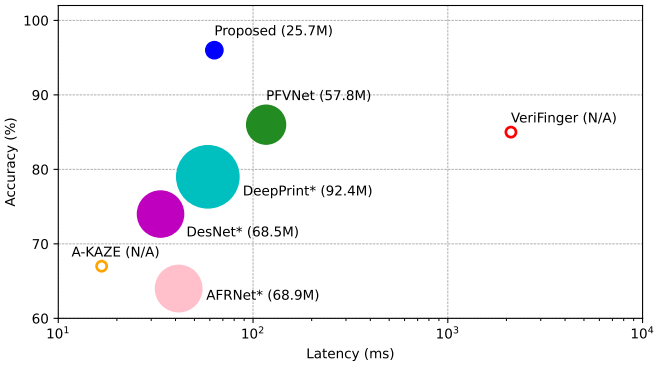


Fig. 13. Comparisons of the trade-off between performance and efficiency on Hybrid DB_B in image size 160×160. The size of solid points represents the model size of deep learning methods, while hollow points represent non learning algorithms.

where s is the genuine matching score, P and f represent the probability and probability density, y represents the identity to be verified (1 is the same, 0 is different). It can be seen that algorithms based on key points [6], [9] and region representation [12], [13] expose their respective shortcomings when there is a lack of sufficient features in overlapping areas (row 1) and significant differences in ridge texture of non overlapping areas (row 2). PFVNet [7] has clear advantages in both cases, even with minimal overlap (row 3). However, it fails in the last example with confusion in rotation. At the same time, our method demonstrates optimal stability and robustness in focusing on overlapping regions and gives correct decisions with the highest accuracy.

Meanwhile, some failure cases in Fig. 12 show that current JIPNet still needs to be improved in some extreme scenarios, such as: (i) sever image defacement caused by incorrect collection conditions (row 1); (ii) significant misalignment of ridges caused by finger distortion (row 2); (iii) extremely weak texture, which refers to almost no minutiae or changes in orientation, as well as large modal differences (row 3).

G. Efficiency

In this subsection, we present a comparison of deployment efficiency across different algorithms. As shown in Fig 13, our proposed method achieves a new state-of-the-art trade-off between matching performance, speed, and size. This clearly highlights our streamlined joint estimation framework of pose and identity, as well as the appropriate and efficient network design. The calculated time covers a complete process from inputting an original image pair to outputting the final

estimation result. All algorithms except for VeriFinger [9] (encapsulated SDK in C/C++) are implemented in Python and the batch size is set to 1, which are deployed on a single NVIDIA GeForce RTX 3090 GPU with an Intel Xeon E5-2680 v4 CPU @ 2.4 GHz.

VI. Conclusion

In this paper, we propose a joint framework of identity verification and pose alignment for partial fingerprints. A novel CNN-Transformer hybrid network, named JIPNet, is presented to combine their advantages in feature extraction and information interaction, promoting the attention and utilization of valuable information. In addition, a lightweight pre-training task is designed to improve the representation ability of feature encoder by reconstructing enhanced images. Comprehensive experiments on multiple databases demonstrate the effectiveness and superiority of our proposed method. Future studies will focus on further improving the performance in more challenging scenarios such as extreme low quality and small overlap, and extending our proposed method to one-to-many matching tasks.

References

[1] F. Galton, *Finger prints*. Cosimo Classics, 1892, no. 57490-57492.

[2] H. C. Lee, R. E. Gaensslen, and R. Ramotowski, *Lee and Gaensslen's advances in fingerprint technology (3rd Edition)*. CRC Press, 2012.

[3] D. Maltoni, D. Maio, A. K. Jain, and J. Feng, *Handbook of Fingerprint Recognition (3rd Edition)*. Springer International Publishing, 2022.

[4] C. I. Watson, C. Watson, and C. Wilson, *Effect of image size and compression on one-to-one fingerprint matching*. US Department of Commerce, National Institute of Standards and Technology, 2005.

[5] B. Fernandez-Saavedra, R. Sanchez-Reillo, R. Ros-Gomez, and J. Liu-Jimenez, "Small fingerprint scanners used in mobile devices: the impact on biometric performance," *IET Biometrics*, vol. 5, no. 1, pp. 28–36, 2016.

[6] S. Mathur, A. Vjay, J. Shah, S. Das, and A. Malla, "Methodology for partial fingerprint enrollment and authentication on mobile devices," in *2016 International Conference on Biometrics (ICB)*, 2016, pp. 1–8.

[7] Z. He, J. Zhang, L. Pang, and E. Liu, "PFVNet: A partial fingerprint verification network learned from large fingerprint matching," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3706–3719, 2022.

[8] A. Roy, N. Memon, and A. Ross, "MasterPrint: Exploring the vulnerability of partial fingerprint-based authentication systems," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 9, pp. 2013–2025, 2017.

[9] VeriFinger SDK 12.0. Accessed: Apr. 11, 2024. [Online]. Available: https://www.neurotechnology.com/verifinger.html

[10] R. Cappelli, M. Ferrara, and D. Maltoni, "Minutia Cylinder-Code: A new representation and matching technique for fingerprint recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2128–2141, 2010.

[11] H. Choi, K. Choi, and J. Kim, "Fingerprint matching incorporating ridge features with minutiae," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 2, pp. 338–345, 2011.

[12] J. J. Engelsma, K. Cao, and A. K. Jain, "Learning a fixed-length fingerprint representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 1981–1997, 2021.

[13] S. Gu, J. Feng, J. Lu, and J. Zhou, "Latent fingerprint registration via matching densely sampled points," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1231–1244, 2021.

[14] S. Wu, B. Liu, Z. Wang, Z. Jia, and J. Feng, "Minutiae-awarely learning fingerprint representation for fingerprint indexing," in *2022 IEEE International Joint Conference on Biometrics (IJCB)*, 2022, pp. 1–8.

[15] S. A. Grosz and A. K. Jain, "AFR-Net: Attention-driven fingerprint recognition network," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 6, no. 1, pp. 30–42, 2024.

[16] S. Chen, Z. Guo, X. Li, and D. Yang, "Query2set: Single-to-multiple partial fingerprint recognition based on attention mechanism," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1243–1253, 2022.

[17] M.-M. Paumard, D. Picard, and H. Tabia, "Deepzzle: Solving visual jigsaw puzzles with deep learning and shortest path optimization," *IEEE Transactions on Image Processing*, vol. 29, pp. 3569–3581, 2020.

[18] S. Markaki and C. Panagiotakis, "Jigsaw puzzle solving techniques and applications: a survey," *The Visual Computer*, vol. 39, no. 10, pp. 4405–4421, 2023.

[19] Y. Duan, J. Feng, J. Lu, and J. Zhou, "Estimating fingerprint pose via dense voting," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2493–2507, 2023.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 213–229.

[21] W. Jiang, E. Trulls, J. Hosang, A. Tagliasacchi, and K. M. Yi, "COTR: Correspondence transformer for matching across images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6207–6217.

[22] Z. Shen, J. Sun, Y. Wang, X. He, H. Bao, and X. Zhou, "Semi-dense feature matching with transformers and its applications in multiple-view geometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7726–7738, 2023.

[23] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2023.

[24] NIST Special Database 14. Accessed: Apr. 11, 2024. [Online]. Available: https://www.nist.gov/srd/nist-special-database-14

[25] FVC 2002: the Second International Competition for Fingerprint Verification Algorithms. Accessed: Apr. 11, 2024. [Online]. Available: http://bias.csr.unibo.it/fvc2002/default.asp

[26] FVC 2004: the Third International Fingerprint Verification Competition. Accessed: Apr. 11, 2024. [Online]. Available: http://bias.csr.unibo.it/fvc2004/default.asp

[27] FVC2006: the Fourth International Fingerprint Verification Competition. Accessed: Apr. 11, 2024. [Online]. Available: http://bias.csr.unibo.it/fvc2006/default.asp

[28] F. Chen, J. Zhou, and C. Yang, "Reconstructing orientation field from fingerprint minutiae to improve minutiae-matching accuracy," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1665–1670, 2009.

[29] J. Feng, "Combining minutiae descriptors for fingerprint matching," *Pattern Recognition*, vol. 41, no. 1, pp. 342–352, 2008.

[30] K. Cao, D.-L. Nguyen, C. Tymoszek, and A. K. Jain, "End-to-end latent fingerprint search," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 880–894, 2019.

[31] M. Yamazaki, D. Li, T. Isshiki, and H. Kunieda, "SIFT-based algorithm for fingerprint authentication on smartphone," in *2015 6th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*, 2015, pp. 1–5.

[32] R. F. S. Teixeira and N. J. Leite, "A new framework for quality assessment of high-resolution fingerprint images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 10, pp. 1905–1917, 2017.

[33] F. Zhang, S. Xin, and J. Feng, "Combining global and minutia deep features for partial high-resolution fingerprint matching," *Pattern Recognition Letters*, vol. 119, pp. 139–147, 2019.

[34] A. K. Jain, S. Prabhakar, L. Hong, and S. Pankanti, "Fingercode: a filterbank for fingerprint representation and matching," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 2. IEEE, 1999, pp. 187–193.

[35] L. Nanni and A. Lumini, "Local binary patterns for a hybrid fingerprint matcher," *Pattern recognition*, vol. 41, no. 11, pp. 3461–3466, 2008.

[36] R. Cappelli, "Fast and accurate fingerprint indexing based on ridge orientation and frequency," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 6, pp. 1511–1521, 2011.

[37] S. A. Grosz, J. J. Engelsma, E. Liu, and A. K. Jain, "C2CL: Contact to contactless fingerprint matching," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 196–210, 2022.

[38] Y. Duan, Z. Pan, J. Feng, and J. Zhou, "Fingerprint matching with localized deep representation," *arXiv preprint arXiv:2311.18576*, 2023.

[39] K. Nandakumar and A. K. Jain, "Local correlation-based fingerprint matching." in *ICVGIP*, 2004, pp. 503–508.

[40] A. Lindoso, L. Entrena, J. Liu-Jimenez, and E. San Millan, "Increasing security with correlation-based fingerprint matching," in *2007 41st Annual IEEE International Carnahan Conference on Security Technology*. IEEE, 2007, pp. 37–43.
[41] X. Guan, Y. Duan, J. Feng, and J. Zhou, "Regression of dense distortion field from a single fingerprint image," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 4377–4390, 2023.
[42] X. Guan, J. Feng, and J. Zhou, "Phase-aggregated dual-branch network for efficient fingerprint dense registration," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 5712–5724, 2024.
[43] F. Zeng, S. Hu, and K. Xiao, "Research on partial fingerprint recognition algorithm based on deep learning," *Neural Computing and Applications*, vol. 31, no. 9, pp. 4789–4798, 2019.
[44] B. Bakhshi and H. Veisi, "End to end fingerprint verification based on convolutional neural network," in *2019 27th Iranian Conference on Electrical Engineering (ICEE)*, 2019, pp. 1994–1998.
[45] Y. Liu, B. Zhou, C. Han, T. Guo, and J. Qin, "A novel method based on deep learning for aligned fingerprints matching," *Applied Intelligence*, vol. 50, pp. 397–416, 2020.
[46] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *Advances in Neural Information Processing Systems*, vol. 34, pp. 30 392–30 400, 2021.
[47] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 17–33.
[48] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4510–4520.
[49] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.
[50] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.
[51] S. A. Grosz and A. K. Jain, "Latent fingerprint recognition: Fusion of local and global embeddings," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 5691–5705, 2023.
[52] Y. Zhu, X. Yin, and J. Hu, "FingerGAN: A constrained fingerprint generation scheme for latent fingerprint enhancement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8358–8371, 2023.
[53] R. Cappelli, D. Maio, and D. Maltoni, "An improved noise model for the generation of synthetic fingerprints," in *ICARCV 2004 8th Control, Automation, Robotics and Vision Conference, 2004.*, vol. 2. IEEE, 2004, pp. 1250–1255.
[54] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, 1994, pp. 474–485.
[55] Y. Tang, F. Gao, J. Feng, and Y. Liu, "FingerNet: An unified deep network for fingerprint minutiae extraction," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 108–116.
[56] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.
[57] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

Xiongjun Guan received the B.S. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2021, where he is currently pursuing the Ph.D. degree under the supervision of Prof. Jianjiang Feng with the Department of Automation. His research interests include fingerprint recognition, computer vision and pattern recognition.

Zhiyu Pan received his Bachelor of Engineering (BEng) degree in Electronic Science and Technology from Beijing Institute of Technology, China, in 2020. He is currently pursuing a Ph.D. degree in the Department of Automation at Tsinghua University. His research interests include biometrics, human action analysis, and computer vision. Specifically, his current work focuses on fingerprint recognition, multi-modal learning, and related areas.

Jianjiang Feng received the B.Eng. and Ph.D. degrees from the School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, China, in 2000 and 2007, respectively. From 2008 to 2009, he was a Post-Doctoral Researcher with the PRIP Laboratory, Michigan State University. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing. His research interests include fingerprint recognition and computer vision.

Jie Zhou received the B.S. and M.S. degrees from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1995. From 1995 to 1997, he served as a Post-Doctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a Full Professor with the Department of Automation, Tsinghua University. His research interests include computer vision, pattern recognition, and image processing. In recent years, he has authored more than 300 papers in peer-reviewed journals and conferences. Among them, more than 100 papers have been published in top journals and conferences such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, and CVPR. He is an associate editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence and two other journals. He received the National Outstanding Youth Foundation of China Award. He is a Fellow of the IAPR and a senior member of the IEEE.