UniRotNet: Mask-Only Canonical Alignment for Contactless Palm ROI Localization

Amani Almaktari[1], Mohammed Bokash[2], Vikas T. Humbe[3]
[1,2,3]*School of Technology, Swami Ramanand Teerth Marathwada University Sub-Centre, Latur, Maharashtra, India*
doi.org/10.64643/IJIRT210748

***Abstract*—Reliable palm region extraction is a critical prerequisite for contactless palm biometrics, where pose variation and background clutter can severely degrade matching accuracy. Many high-performing palmprint and palm-vein recognition methods particularly those based on deep convolutional features and texture coding rely on geometrically normalized inputs and therefore depend on a stable preprocessing front-end. This paper presents a mask-guided three-stage pipeline tailored to this scenario: (i) a U-Net segments the hand from the background, (ii) a lightweight network (UniRotNet) enforces a canonical upright pose by estimating in-plane rotation from the predicted binary mask alone, and (iii) lightweight CNN regressors localize a square palm region of interest (ROI) from the aligned mask. By operating on binary masks in Stages II and III, the design reduces the dependence of alignment and ROI localization on raw appearance. Experiments on the CASIA Multi-Spectral Palmprint database, using subject-disjoint splits over three random seeds, yield high segmentation accuracy (IoU = 0.9911 ± 0.0011), precise mask-only rotation correction (MAE = 3.004 ± 0.567°), and accurate ROI localization on aligned masks (mIoU = 0.909 ± 0.002) with negligible outside-of-hand leakage. Compared with an end-to-end regressor trained on raw images with axis-aligned box targets (Image-PalmROI), the mask-guided regressors give tighter ROIs (mIoU = 0.905–0.909 vs 0.641) and, on images aligned with the recorded hand angle, lower downstream matching error (4.11% vs 9.66% EER with a fine-tuned ResNet-18 matcher), indicating that mask-guided canonical alignment provides a stable front-end for contactless palm biometrics on this dataset.**

***Index Terms*—Canonical alignment, Hand segmentation, Palm biometrics, ROI localization, Rotation correction.**

I. INTRODUCTION

A. Background and Motivation

Contactless palm biometrics has gained increasing attention as a practical modality for palmprint and palm-vein recognition, driven by usability and hygiene requirements in deployments where touch-based sensors are undesirable or impractical [1], [2]. Contactless capture is more user-friendly than contact-based acquisition and avoids device-surface contamination, but it introduces substantially larger variability in hand pose, scale, orientation, and imaging conditions [3], [4]. This variability is especially pronounced in unconstrained environments, where the hand may appear under background clutter, illumination changes, multi-band acquisition settings, and partial occlusions caused by sleeves or wrist accessories [5], [6], [7]. Such factors directly affect preprocessing stability and, in particular, region-of-interest (ROI) extraction the step that determines whether subsequent feature extraction and matching operate on geometrically comparable palm regions [1], [5], [8]. When the ROI is geometrically unstable across samples, the extracted biometric representation becomes less discriminative even if the feature extractor itself is powerful [9], [10].

The necessity of explicit preprocessing steps such as rotation correction and ROI extraction depends on the choice of feature extraction technique. A subset of existing contactless palmprint methods adopts rotation-invariant feature representations that embed geometric invariance directly into the descriptor, bypassing the need for explicit pose normalization. For example, Mohamed-Abdul-Cader et al. [11] proposed a bispectral feature-based approach that achieves rotation and scale invariance in contactless palmprint recognition without an explicit alignment step. Descriptors based on histograms of oriented lines [12] and binary orientation co-occurrence coding [13] have similarly been reported to tolerate small rotations and translations. However, many high-performing contactless palmprint and palm-vein recognition

systems particularly those based on deep convolutional features, texture coding, or appearance-sensitive representations rely on stable geometric preprocessing to achieve reliable matching [5], [14], [15], [16]. For such methods, the quality of the preprocessing pipeline directly determines recognition performance. This paper targets the latter scenario, and the proposed mask-guided pipeline is designed to serve as a robust and efficient front-end for appearance-dependent recognition systems.

B. Challenges in Contactless Palm Biometrics

Despite significant progress in palm biometrics, robust ROI extraction in contactless settings remains challenging for several reasons. First, pose variation and in-plane rotation cause the same hand to appear at arbitrary orientations, making ROI localization inconsistent across capture sessions. Second, background clutter and illumination changes degrade segmentation quality and introduce spurious edges that mislead geometric heuristics. Third, spectral variation in multi-band acquisition alters appearance statistics, making appearance-driven preprocessing less reliable across channels. Fourth, occlusions and boundary artifacts near the wrist (sleeves, bright edges, partial hand truncation) confuse classical orientation estimators and shift the predicted ROI center. Finally, end-to-end joint regression of the ROI from the raw image is theoretically appealing but introduces its own challenges under unconstrained pose; we revisit this empirically in Section V-D with a like-for-like end-to-end baseline (Image-PalmROI). These challenges collectively suggest that explicit, geometry-driven intermediate representations such as binary masks can reduce sensitivity to appearance variability and improve downstream processing stability.

C. Motivation for Mask-Guided Design

The central design choice in this paper is to separate appearance-driven processing (segmentation) from geometry-driven processing (alignment and ROI localization). Once a binary hand mask is available, all subsequent operations rotation estimation and ROI regression can be performed on the mask alone, without access to raw image intensities. This separation has two main advantages: the mask representation is inherently invariant to illumination changes and spectral-band shifts that are common sources of variability in multi-spectral palm acquisition, and the mask boundary encodes hand-shape cues (wrist direction, fingertip positions) that directly support orientation estimation. The proposed pipeline demonstrates that this separation yields a stable and efficient preprocessing front-end suitable for downstream appearance-sensitive recognition systems.

D. Contributions

We summarize our main contributions as follows:

- Three-stage mask-guided pipeline.

We propose a pipeline consisting of Stage I hand segmentation, Stage II mask-guided canonical alignment (rotation correction via UniRotNet), and Stage III palm ROI localization from aligned masks. The pipeline is designed to serve as an efficient and robust preprocessing front-end for appearance-sensitive contactless palmprint and palm-vein recognition systems.

- Segmentation model design and evaluation.

We design a robust hand segmentation module and systematically evaluate a U-Net [17] framework with two encoder backbones (ResNet34 [18] and MobileNetV2 [19]) under a unified training and evaluation protocol, highlighting the accuracy–efficiency trade-off for practical deployment.

- UniRotNet.

We introduce a lightweight rotation estimation network that operates on predicted binary masks and supports configurable training modes (classification, regression on cos θ / sin θ, and circular-domain fusion), and we analyze the angular-error and bin-accuracy trade-offs across modes.

- Efficient ROI localizers.

We develop two lightweight ROI regression models (Fast-PalmROI and Lite-PalmROI) and evaluate them under subject-disjoint splits across three random seeds against multiple baselines.

- Comprehensive evaluation.

We provide quantitative and qualitative analyses for each stage, a head-to-head comparison against an end-to-end ROI regressor (Image-PalmROI) that operates directly on raw images, a downstream palmprint-matching evaluation against ground-truth and

predicted ROIs using both a classical matcher (CompCode) and a learned matcher (fine-tuned ResNet-18), and a complexity and runtime analysis.

The remainder of this paper is organized as follows. Section II reviews related work on hand segmentation, rotation-invariant feature extraction, geometric normalization, and unconstrained palm ROI extraction, motivating the proposed design. Section III presents the three-stage pipeline. Section IV describes the experimental setup and evaluation metrics. Section V reports results and discussion. Section VI presents conclusions and future work.

II. RELATED WORK

Recent work on contactless palm biometrics emphasizes robust front-end preprocessing hand segmentation, geometric normalization, and ROI extraction as a prerequisite for reliable matching under unconstrained acquisition [5], [14], [15]. We organize the related literature into four themes (1) rotation-invariant versus normalization-dependent feature extraction, (2) hand segmentation as a robust front-end, (3) ROI extraction under unconstrained conditions, and (4) alignment and geometric normalization and close with a positioning statement situating the present work.

A. Feature Extraction: Invariant vs. Normalization-Dependent Approaches

A fundamental distinction in contactless palmprint recognition concerns whether the feature extractor handles geometric variability internally or relies on an explicit preprocessing pipeline. Methods in the first category embed rotation or scale invariance directly into the descriptor, whereas methods in the second category assume that input images have been geometrically normalized in advance.

Among invariant approaches, Mohamed-Abdul-Cader et al. [11] proposed a bispectral feature method that achieves rotation and scale invariance for contactless palmprint recognition without explicit ROI extraction or pose correction, demonstrating that statistical spectral features can maintain discriminability across orientations. Histograms of oriented lines [12] and binary orientation co-occurrence coding [13] have similarly been reported to tolerate small rotations and translations through their orientation-histogram and multi-orientation encoding strategies. Deep representations that incorporate learned geometric invariance for example via spatial transformer modules or strong rotation augmentation during training have also been explored as a route to pose-robust embeddings.

A second body of work, however, depends on stable geometric input. Line-based coding methods such as PalmCode [20] and CompCode [21], competitive-code variants, and many deep feature extractors trained on aligned data lose spatial correspondence between gallery and probe under even moderate rotational misalignment, with measurable matching degradation. The BPFNet framework [22] explicitly formulates ROI localization and alignment as prerequisites for bimodal palm matching, reinforcing that preprocessing quality directly affects recognition accuracy. The present paper targets this normalization-dependent scenario.

B. Hand Segmentation as a Robust Front-End

Segmenting the hand region to suppress background clutter is a widely adopted first step in contactless palm pipelines. Ito et al. [15] proposed HandSegNet, a CNN-based segmentation method tailored for contactless palmprint recognition, showing that reliable hand masks can be obtained even in challenging scenes. Qiao et al. [16] extended this direction with HSANet, which segments full palmar regions across diverse backgrounds by combining low-level details with high-level semantic cues. Yalcin et al. [23] reported that modern backbones (U-Net with ViT or ResNet34) outperform traditional auto-encoder approaches for palmprint segmentation. These advanced architectures offer high accuracy at substantial computational cost, motivating our use of an efficient MobileNetV2 backbone in Stage I.

C. ROI Extraction Under Unconstrained Conditions

ROI extraction is a core requirement because localization errors propagate directly to feature extraction and matching. Lin et al. [5] combined YOLOv5-lite for coarse palm localization with an improved U-Net for keypoint-based ROI construction, achieving strong accuracy at compact size. Luo and Zhong [14] presented an adaptive ROI extraction pipeline targeting variations in palm size, posture, illumination, and background. Dai et al. [1] used an improved lightweight HRNet to detect palm keypoints under complex backgrounds. Wu and Leng [24]

presented a video palmprint system with assisted placement guidance, and Fan et al. [7] introduced touchless palm sensing with dynamic registration. Closely related noncontact palm-vein ROI extraction work [1], [8] reinforces the importance of robust localization in cluttered environments. In contrast to keypoint-regression methods, our Stage III directly regresses a square ROI parameterized by center and side length from an aligned binary mask, providing a simpler and lighter regression target that avoids the ambiguity of individual keypoint labeling.

D. Alignment and Geometric Normalization

Rotation and pose normalization remain essential in touchless acquisition because the hand may appear at arbitrary in-plane orientation. Brown and Bradshaw [25] showed that explicit alignment improves robustness under contactless variability by stabilizing geometric consistency before recognition.

Qiao et al. [16] argued that precise normalization and consistent region extraction are critical for reliable touchless hand-based biometrics. BPFNet [22] formulates ROI localization and alignment jointly via bounding-box regression and translation estimation, showing that geometric alignment integrated with ROI prediction supports touchless palm matching. Classical geometric normalization based on principal component analysis or finger-valley keypoints is susceptible to the 180° orientation ambiguity and to boundary artifacts from occlusions, motivating learning-based orientation estimation that incorporates semantic shape cues.

Cross-device [26] and cross-dataset [27], [28] palmprint methods further illustrate the broader need for preprocessing strategies that reduce dependence on dataset-specific appearance statistics.

E. Positioning of This Work

Most existing pipelines either extract the ROI directly from appearance images [5], [25] or incorporate alignment in an appearance-dependent manner. The proposed pipeline occupies a distinct design point: only Stage I is appearance-driven, while Stages II and III operate exclusively on binary masks, which makes alignment and ROI localization robust to illumination changes, spectral-band shifts, and background variation by design.

Unlike end-to-end methods that jointly optimize all tasks, the staged design allows each module to be independently validated and replaced, and the explicit rotation correction in Stage II resolves the dominant 180° flip ambiguity inherent to principal-axis estimators. We empirically validate this design choice in Section V-D by training a structurally identical end-to-end regressor (Image-PalmROI) that operates directly on raw images and showing that the cascaded mask-guided design produces tighter ROIs and stronger downstream matching.

III. PROPOSED MASK-GUIDED THREE-STAGE PIPELINE

A. Overview

We propose a mask-guided pipeline that converts an input hand image into a canonical palm representation through three stages: (i) hand segmentation, (ii) mask-guided rotation correction (UniRotNet), and (iii) palm ROI localization from the aligned mask (PalmROINet).

The pipeline is specifically designed to support normalization-dependent downstream feature extraction methods by providing a geometrically stable palm ROI. Given an input image I, Stage I predicts a binary hand mask M.

In Stage II, UniRotNet estimates the in-plane rotation correction angle $\hat{\theta}$ from M and produces an aligned mask by rotation normalization: $M_{align} = \mathcal{R}(M, -\hat{\theta})$ where $\mathcal{R}(\cdot)$ denotes a 2D in-plane rotation about the mask canvas center (implemented with nearest-neighbor interpolation to preserve binary edges).

In Stage III, PalmROINet (instantiated as either Fast-PalmROI or Lite-PalmROI) predicts the square palm ROI parameters $B = (c_x, c_y, s)$ from M_{align}, where (c_x, c_y) is the ROI center, and s is the side length. Finally, the ROI is extracted from the original image I using the predicted parameters, yielding the final normalized ROI. Fig. 1 summarizes the full pipeline.

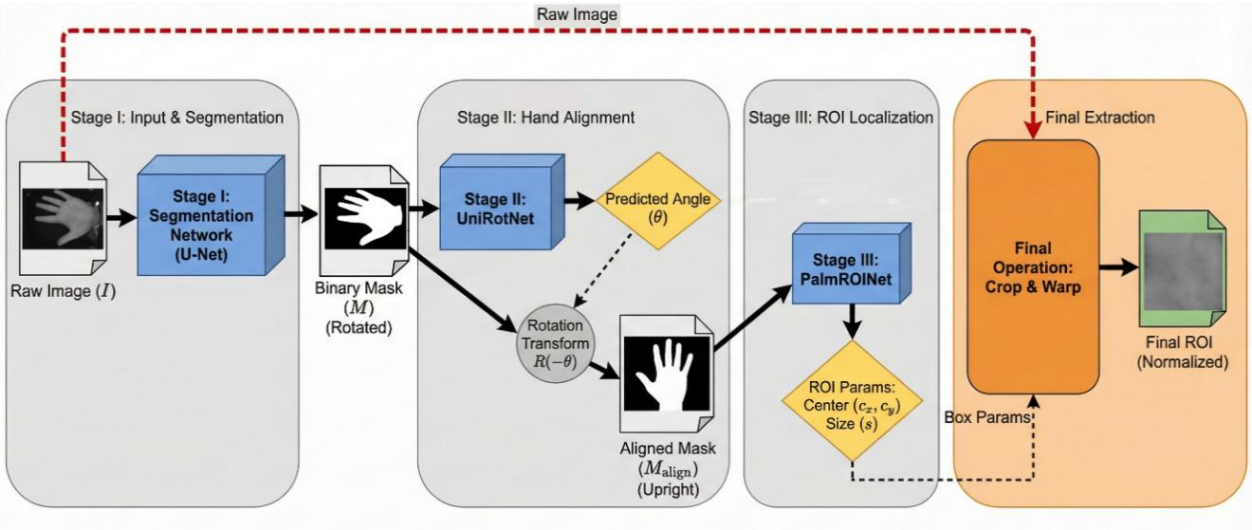


Fig. 1. Overview of the proposed mask-guided three-stage pipeline. Stage I segments the hand from the input image I to produce a binary mask M. Stage II (UniRotNet) estimates the correction angle $\hat{\theta}$ and generates an upright aligned mask M_{align}. Stage III (PalmROINet; implemented as either Fast-PalmROI or Lite-PalmROI) regresses the square palm ROI parameters $B = (c_x, c_y, s)$, which are used to extract the final normalized ROI crop from I. Solid arrows denote data flow, while dashed arrows denote parameter flow.

B. Stage I: Hand Segmentation

1) Model Architecture

Stage I performs pixel-wise hand segmentation using a U-Net encoder–decoder network with skip connections to combine high-level semantic cues with fine-grained spatial details. The encoder extracts multi-level feature representations from low-level edge/texture patterns to higher-level structural cues. The decoder reconstructs a high-resolution prediction map via upsampling while fusing encoder features through skip connections to sharpen boundaries and recover spatial detail (Fig. 2).

We evaluate two encoder backbones within the same U-Net framework: ResNet34, a higher-capacity reference encoder that provides a strong segmentation baseline, and MobileNetV2, a lightweight and compute-efficient backbone. Both encoders are initialized with ImageNet [29] pretrained weights.

The input to Stage I is a hand image $I \in \mathbb{R}^{H\times W\times 3}$ (grayscale images are replicated to three channels to match ImageNet-pretrained encoders). The output is a predicted binary mask $M \in \{0,1\}^{H\times W}$. The network outputs a single-channel logit map $L \in \mathbb{R}^{H\times W}$ converted to probabilities via sigmoid $P = \sigma(L)$, and binarized: $M(x, y) = 1[P(x,y) \geq \tau]$, with $\tau = 0.5$.

2) Loss Function and Optimization

We use a combined Binary Cross-Entropy (BCE) and Dice loss: $\mathcal{L} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}$. The network is optimized using AdamW [30] with initial learning rate 3×10^{-4} and weight decay 1×10^{-4}, following a warmup-cosine schedule with 200 warmup steps. Automatic Mixed Precision (AMP) is used for training efficiency.

3) Training Protocol

Models are trained with batch size 8 for up to 100 epochs, with early stopping (patience = 15) based on validation Dice score. Training uses mild augmentation: horizontal flipping ($p = 0.5$), brightness and contrast jitter of ±10% ($p = 0.2$), and shifting by ±2%, scaling by ±10% and rotation by ±10° ($p = 0.3$). The best checkpoint is selected per validation Dice. Experiments are repeated over three random seeds {42, 2025, 1337}, and results are reported as mean ± std.

4) Evaluation Metrics

Segmentation performance is evaluated per test image by comparing the predicted binary mask $\hat{M}$ with the ground-truth mask M, from which we derive pixel-wise counts TP, FP, FN, TN. We report standard overlap measures, including Dice and Intersection-over-Union (IoU):

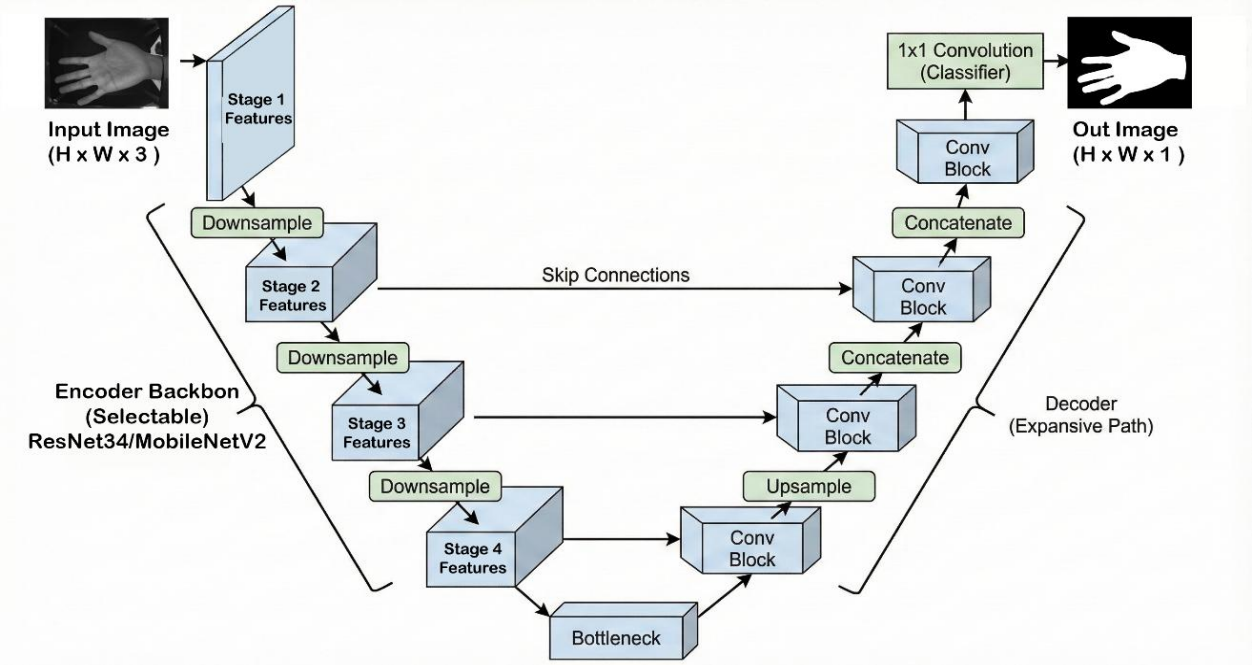


Fig. 2. U-Net segmentation architecture with a swappable encoder backbone (ResNet34 or MobileNetV2). The encoder extracts multi-scale features via downsampling, which are fused with decoder features through skip connections; the decoder upsamples to produce a single-channel output (logits) converted to a binary mask via sigmoid and a fixed threshold.

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}, \text{IoU} = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}}. \quad (1)$$

To characterize over- versus under-segmentation behavior, we also report pixel-level classification metrics (Precision, Recall, Specificity, and Accuracy) derived from TP, FP, FN, TN. Beyond region overlap, we quantify boundary quality using Boundary-F1, which is particularly relevant for the downstream geometry-driven alignment and ROI localization stages.

Boundaries are extracted via erosion-based differencing, $B(M) = \mathbb{I}[M \neq \text{Erode}(M)]$, and matched using a tolerance τ: a boundary pixel is considered correct if it lies within τ pixels of the counterpart boundary (and vice versa), implemented via distance transforms. The final boundary score is computed as the harmonic mean of boundary precision and boundary recall:

$$\text{Boundary-F1} = \frac{2\,\text{BPrec}\,\text{BRec}}{\text{BPrec}+\text{BRec}}. \quad (2)$$

In our setup, τ is set proportional to the image diagonal, $\tau = r\sqrt{H^2 + W^2}$ with $r = 0.01$ (yielding $\tau \approx 9.6$px for $H = 576, W = 768$). The corresponding performance tables and qualitative examples are provided in Section V-A.

C. Stage II: Mask-Guided Canonical Alignment (UniRotNet)

1) Overview and Motivation

After obtaining a binary hand mask from Stage I, Stage II rotates this mask to a canonical orientation. This canonical alignment is achieved by predicting a rotation correction angle so that all hands exhibit a consistent pose with the fingers pointing upward. This reduces rotation-related variation and makes subsequent steps particularly ROI localization more stable and accurate, thereby improving the reliability of feature extraction. Importantly, this stage uses only the predicted masks and does not require the original images. By relying on hand geometry rather than appearance cues (e.g., texture or intensity), the alignment becomes less sensitive to changes in illumination, background clutter, or spectral band.

2) Rotation Estimation Network (UniRotNet)

UniRotNet as shown in *Fig. 3* uses a ResNet18 backbone adapted to single-channel input (the first convolution is replaced by a new, randomly initialized single-channel convolution; the remaining layers are initialized from ImageNet). A configurable prediction head is attached depending on the training mode.

Classification mode (cls): The angle space [0°, 360°) is discretized into C = 72 bins (5° per bin). The network outputs logits $z \in \mathbb{R}^C$, and the predicted bin $b = argmax_c(z_c)$. The predicted angle is θ_{cls}= (b + 0.5) × (360/C). Training uses cross-entropy with label smoothing ε = 0.05.

Regression mode (reg): The network predicts a 2D vector v = (ĉ, ŝ) corresponding to (cos θ, sin θ), avoiding the 0/360° discontinuity. The predicted angle is θ_{reg}=wrap (atan2(ŝ, ĉ)). Training uses SmoothL1 (Huber) [31] loss against the target unit vector [cos θ, sin θ].

Fusion mode (both): Regression and classification outputs are combined in the circular domain by converting each to a unit vector and taking a weighted sum: $v_{fuse} = \frac{\alpha \cdot v_{reg} + (1-\alpha) \cdot v_{cls}}{|\alpha \cdot v_{reg} + (1-\alpha) \cdot v_{cls}|}$, with α = 0.5. The final angle is θ_{fuse} = wrap (atan2($v_{\text{fuse},y}, v_{\text{fuse},x}$)). This strategy ensures stable predictions near the 0°/360° boundary.

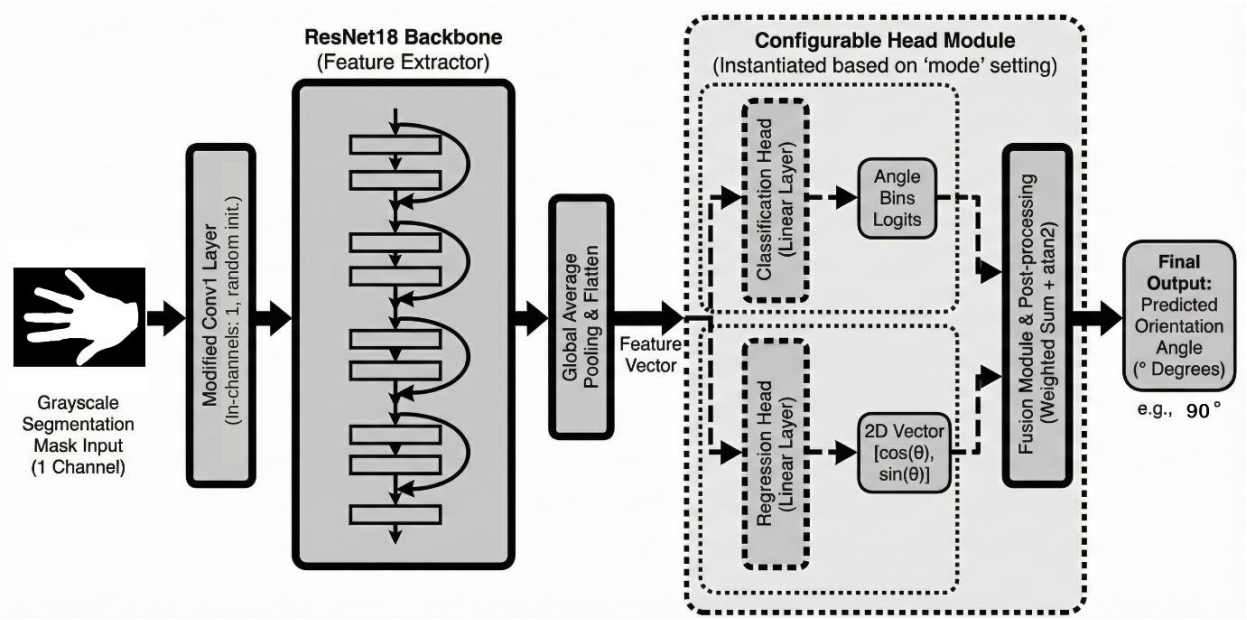


Fig. 3. Architecture of the proposed rotation estimation network (UniRotNet). A ResNet18 backbone extracts features from a single-channel hand mask, followed by a configurable head module instantiated as a classification head, regression head, or both depending on the training mode.

3) Loss Function

Depending on the selected mode, UniRotNet optimizes a classification loss, a circular regression loss, or both.

In classification mode, we quantize θ ∈ [0, 360) into a bin index b ∈ {0, …, C − 1} and minimize cross-entropy with label smoothing ε = 0.05, i.e., $\mathcal{L}_{cls}$ = CE(z,b), where $z \in \mathbb{R}^C$ denotes the classification logits.

In regression mode, the network predicts a 2D vector v = (ĉ, ŝ), which is ℓ2-normalized and supervised against the target unit vector [cos θ, sin θ] using SmoothL1 loss, i.e., $\mathcal{L}_{reg}$= SmoothL1(norm(v), [cos θ, sin θ]). When both heads are enabled, the total objective is the unweighted sum of the two losses: $\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{reg}$.

4) Alignment Transform (Rotation Normalization)

Given the predicted angle $\hat{\theta}$, we generate an aligned mask by undoing the estimated rotation:

$$M_{align} = \mathcal{R}(M, -\hat{\theta}) \quad (3)$$

where $\mathcal{R}(\cdot)$ denotes 2D rotation around the image center. In implementation, we use an affine rotation matrix computed around the mask canvas center and apply warpAffine with nearest-neighbor interpolation to preserve binary edges. After warping, we re-binarize the result to maintain a clean mask.

5) Training Protocol

UniRotNet is trained on masks predicted by a U-Net with a ResNet34 encoder that was trained on the training subjects of the same seed, using subject-disjoint splits repeated across three random seeds; at test time it receives the masks of the deployed Stage I

model (MobileNetV2 encoder). We use AdamW with learning rate 3×10^{-4} and weight decay 1×10^{-4}. The learning rate follows a warmup-cosine schedule with 200 warmup steps and is updated at every optimization step. Models are trained without data augmentation for up to 100 epochs with batch size 32 and early stopping (patience = 15) based on validation MAE (computed using circular error). Automatic Mixed Precision (AMP) is enabled to improve training efficiency. The corresponding performance tables and qualitative examples are provided in Section V-B.

D. Stage III: Palm ROI Localization

1) ROI Parameterization and Preprocessing

The palm ROI is parameterized by a square defined by center coordinates (cx_{px}, cy_{px}) and side length s_{px}. Ground-truth ROI labels were created by one of the authors with a custom annotation tool: the centre and side of the square were set with sliders on the upright hand mask in several rounds; in the later rounds the sliders started from the proposal of an earlier ROI regressor, squares extending outside the hand mask were corrected by hand, and every label was checked visually.

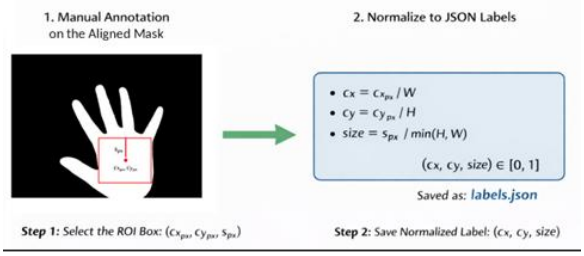


Fig. 4. Palm ROI definition on an aligned binary hand mask.

All masks are letterbox-normalized to 256×256 using nearest-neighbor interpolation and symmetric padding, providing a compact fixed-size input (Fig. 4). Ground-truth labels are mapped into letterbox-normalized coordinates for training, and predictions are inverted back to original image space for evaluation.

2) ROI Localization Models

Fast-PalmROI is a compact CNN built from four convolutional blocks followed by a small MLP head with sigmoid output, designed for low latency from a single-channel 256×256 input (Fig. 5). Lite-PalmROI adapts MobileNetV2 (single-channel input, regression head with sigmoid) to improve representation capacity while remaining computationally efficient (Fig. 6). Both models are trained using SmoothL1 (Huber) regression loss in letterbox-normalized coordinate space.

3) Training Protocol

Both regressors are trained using AdamW (learning rate 1×10^{-3}, weight decay 1×10^{-4}) with cosine annealing LR schedule (T_{max} = 50) for 50 epochs with batch size 128. Data augmentation includes random horizontal flipping (p = 0.5) with consistent target update. Best checkpoint is selected per minimum validation SmoothL1 loss. For all seeds, the training, validation and test inputs of both regressors (Table V) are the masks predicted by the Stage I model of seed 1337, rotated upright with the ground-truth angle. No mask of a validation or test subject of a seed is used for training, although 12 of the 15 test subjects of seed 42 and 13 of seed 2025 were training subjects of that Stage I model; evaluating with the masks of each seed's own Stage I model instead changed the mIoU of Lite-PalmROI by less than 0.001.

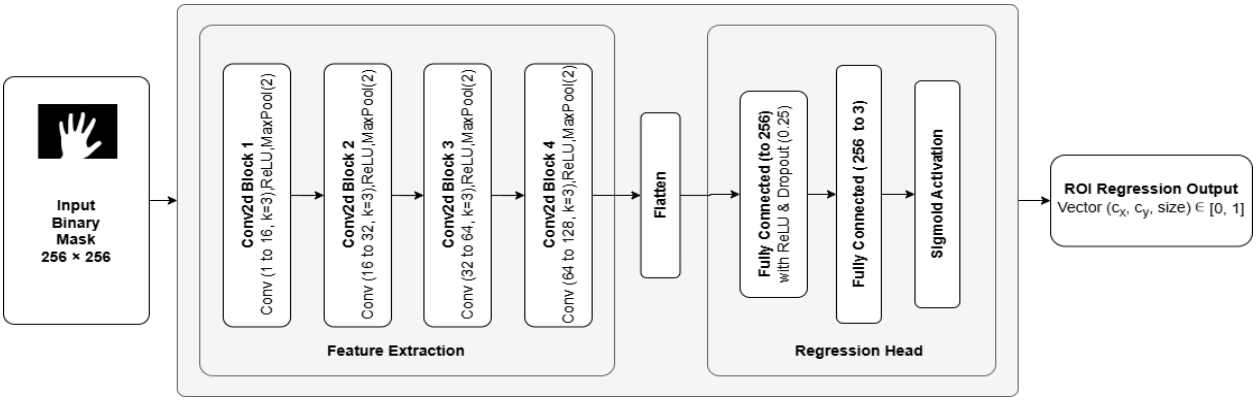


Fig. 5. Schematic overview of the Fast-PalmROI regression network configuration.

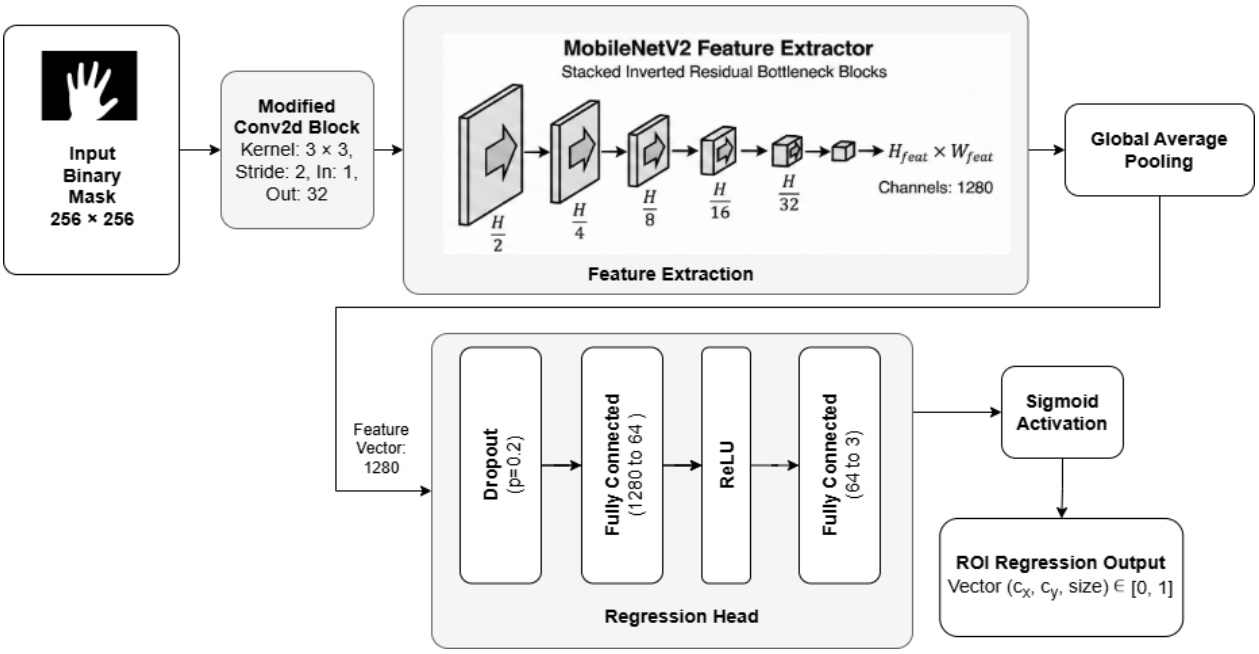


Fig. 6. Block diagram of the Lite-PalmROI lightweight palm ROI detection network.

4) Baselines

Two geometric baselines are evaluated: (1) Fixed-Mean ROI, which uses the training set mean ROI parameters as a constant prediction for all test samples; and (2) Mask-BBox (Shrink) ROI, which computes the tight foreground bounding box, forms a square side, applies shrinkage (α = 0.45) and a vertical shift (β = 0.08 × s_{px}) to bias toward the palm, and normalizes the result. The shrinkage factor α and vertical shift β were selected on the validation split by maximizing mean IoU against the GT ROI, then frozen for all test-set evaluations.

IV. EXPERIMENTAL SETUP

A. Dataset and Protocol

All experiments are conducted on the CASIA Multi-Spectral Palmprint Database [32], which contains palm images acquired from 100 subjects with both left and right hands, across two acquisition sessions. In each session, images are acquired under six spectral conditions (White, 460, 630, 700, 850, 940 nm), with three images per spectrum, yielding 100 × 2 × 2 × 6 × 3 = 7,200 samples. Stages I–II operate on a standardized 576×768 canvas. Stage III applies an additional 256×256 letterbox normalization.

To ensure identity-level generalization, we adopt subject-disjoint splits and repeat all experiments over three random seeds (42, 2025, 1337). For each seed, subjects are partitioned into 70/15/15% for train/validation/test, respectively. All samples of a subject (across hands and spectra) are assigned exclusively to one split, yielding 5,040 training samples and 1,080 samples for each of validation and test per seed. At test time, Stages II and III operate on the binary masks predicted by Stage I (MobileNetV2 encoder, chosen for its accuracy–efficiency trade-off); the masks used to train Stages II and III are given in their training protocols (Sections III-C and III-D). All reported results are computed on the held-out test split for each seed, reported as mean ± sample standard deviation over three seeds.

B. Implementation

All models are implemented in PyTorch and trained on an NVIDIA GeForce RTX 3080 Ti Laptop GPU (16 GB), with Automatic Mixed Precision (AMP) used to accelerate training. Across all stages, a consistent AdamW optimizer is used. Stages I and II employ a warmup-cosine learning rate schedule (200 warmup steps, cosine decay, per-step update). Stage III uses cosine annealing LR (T_{max} = 50, per-epoch update). Stages I and II use early stopping, and in all stages the best checkpoint is selected according to stage-specific validation criteria (Dice for segmentation; MAE for rotation; SmoothL1 loss for ROI regression). Stage-

specific hyperparameters are summarized in each corresponding subsection.

C. Evaluation Metrics

We report evaluation metrics tailored to each stage (Table I). Stage I is evaluated using overlap (IoU, Dice), boundary quality (Boundary-F1 with tolerance $\tau = r \times \sqrt{H^2 + W^2}$, r = 0.01), and pixel-wise metrics (Precision, Recall). Stage II is evaluated using circular error statistics: MAE (degrees), Median, p90, p95, and tolerance-based bin accuracies (Acc@exact-bin, Acc@±1 bin). Stage III is evaluated using ROI overlap (mIoU), threshold accuracies (Acc@0.5, Acc@0.7), localization errors (center error in pixels, size error in pixels), and leakage (outside fraction: fraction of predicted ROI area outside the hand mask). All metrics are computed on the test split and reported as mean ± std over seed-level means.

TABLE I. Summary of evaluation metrics by stage.

Stage	Metrics
Hand Segmentation	Overlap: IoU, Dice. Boundary: Boundary-F1 (with tolerance τ). Pixel-wise: Precision, Recall (optionally Specificity, Accuracy).
Mask-Guided Canonical Alignment	Circular error $e(\hat{\theta}, \theta)$: MAE, Median, p90, p95. Bin-based: Acc@exact-bin, Acc@$\pm$1bin. Optional: tolerance curve $P(e \leq t)$vs. t(degrees).
Palm ROI Localization	ROI overlap: mIoU (predicted vs. GT ROI squares). Threshold accuracy: Acc@0.5, Acc@0.7. Localization errors: center error (px), size error (px). Leakage: outside fraction (ROI area outside hand mask).
Complexity/Runtime	Params, GMACs, latency.

D. Downstream Matchers

Two palmprint matchers are used to evaluate the discriminative quality of the localized ROIs (results in Section V-D). The first is a classical Competitive Code (CompCode) implementation following Kong & Zhang's parameterization [21]: a real Gabor filter bank with six orientations ($\theta_k = k\pi/6$, k = 0,…,5), radial frequency $\omega = 0.0916$, kernel size 35 × 35, and σ derived from the half-bandwidth formula. Each 128 × 128 ROI is encoded as a per-pixel orientation-index map; matching uses normalized Hamming distance over a ±2-pixel translation overlay (the minimum is taken over the 25 candidate shifts), which absorbs sub-pixel ROI misalignment without requiring per-image registration.

The second matcher is a ResNet-18 [18] backbone, ImageNet-pretrained and fine-tuned per seed on the 170 train+val identities at 850 nm (70 + 15) subjects × 2 hands = 170 identities; subject-disjoint from the 15 × 2 = 30 test identities). Fine-tuning uses plain softmax cross-entropy for 50 epochs with AdamW (initial learning rate 3×10^{-4}, weight decay 5×10^{-4}), a cosine learning-rate schedule with a 3-epoch linear warmup, batch size 128, and mild augmentation (random rotation of ±10°, translation of ±4%, scaling of ±5%, brightness and contrast jitter of 0.2, and random erasing with p = 0.25). At test time the classifier head is discarded; the 512-dimensional penultimate feature vector is L2-normalized and matched by cosine distance.

For both matchers, scoring on the test split follows a cross-session, single-band protocol at 850 nm (full protocol described in Section V-D): the per-probe distance to a gallery identity is the minimum across that identity's three templates, Rank-1 accuracy is the top-1 retrieval rate over probes, and Equal Error Rate (EER) is computed by sweeping a threshold over the genuine and impostor distance distributions.

V. RESULTS AND DISCUSSION

This section reports stage-wise results for the proposed three-stage pipeline under subject-disjoint evaluation repeated over three random seeds $\{42, 2025, 1337\}$. We first evaluate Stage I hand segmentation using overlap and boundary metrics and compare encoder backbones in terms of accuracy and efficiency. Stage II mask-guided canonical alignment (UniRotNet) is then assessed using circular-error statistics and bin-based accuracies. Stage III palm ROI localization is evaluated with overlap- and error-based measures, including a head-to-head comparison against an end-to-end ROI regressor (Image-PalmROI) that operates on raw images without an alignment stage. Finally, we

report the downstream palmprint-matching performance attainable from each ROI source using both a classical (CompCode) and a learned (fine-tuned ResNet-18) matcher, with a closing complexity and end-to-end discussion. Tabulated results are complemented throughout by distribution plots and qualitative panels that expose stability and failure modes.

A. Stage I: Hand Segmentation

Both encoder backbones achieve near-saturated overlap and boundary accuracy (Table II). MobileNetV2 attains IoU = 0.9911 ± 0.0011 and Boundary-F1 = 0.9981 ± 0.0007, while ResNet34 attains IoU = 0.9909 ± 0.0018 and Boundary-F1 = 0.9976 ± 0.0014; Precision and Recall are similarly high for both backbones, indicating balanced behavior without systematic over- or under-segmentation. Per-image distributions of Dice, IoU, and Boundary-F1 are tightly concentrated near the upper bound with limited seed sensitivity (Fig. 7, Fig. 8); the cross-encoder Dice comparison in Fig. 9 confirms that scores cluster near 1.0 with only a small lower tail corresponding to harder samples.

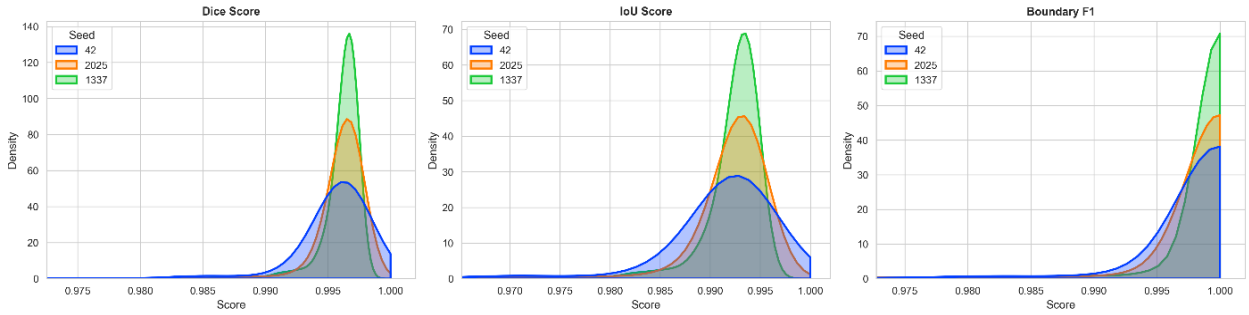


Fig. 7. Detailed Performance Analysis - MobileNetV2.

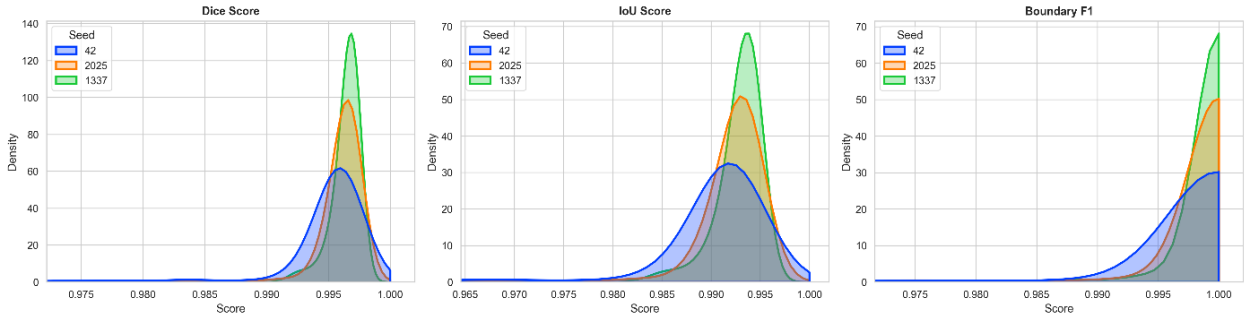


Fig. 8. Detailed Performance Analysis - ResNet34.

Beyond accuracy, MobileNetV2 offers a clear efficiency advantage (Table III): it reduces parameters from 24.44 M to 6.63 M and compute from 52.99 to 22.92 GMACs, while improving CPU latency from 333.5 to 204.7 ms/img.

GPU latency is essentially tied (8.30 vs. 8.74 ms/img), suggesting that at this input resolution runtime is dominated by memory traffic and decoder-side operations rather than FLOPs.

Taken together, MobileNetV2 offers the better accuracy–efficiency trade-off for Stage I and produces masks of sufficient quality to support the geometry-driven processing in Stages II–III.

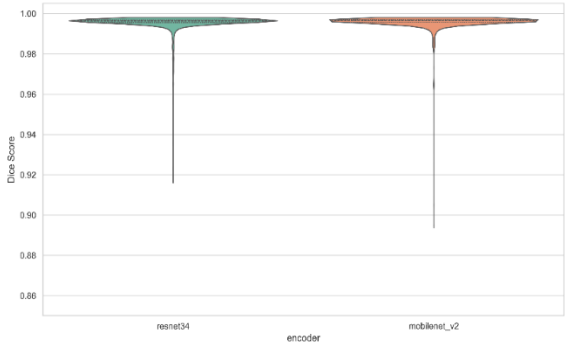


Fig. 9. Comparative Dice Score Distribution.

TABLE II. Hand segmentation performance on the test set aggregated over three seeds (mean ± std).

Encoder	IoU	Dice	Boundary-F1	Precision	Recall
MobileNetV2	0.9911 ± 0.0011	0.9955 ± 0.0006	0.9981 ± 0.0007	0.9950 ± 0.0012	0.9961 ± 0.0003
ResNet34	0.9909 ± 0.0018	0.9954 ± 0.0009	0.9976 ± 0.0014	0.9956 ± 0.0007	0.9952 ± 0.0012

TABLE III. Computational complexity and inference efficiency of the segmentation at 576× 768 resolution.

Encoder	Params (M)	Size (MB)	GMACs	CPU (ms/img)	CPU FPS	GPU (ms/img)	GPU FPS	GPU Mem (MB)
MobileNetV2	6.63	25.3	22.92	204.7	4.89	8.30	120.5	193.8
ResNet34	24.44	93.2	52.99	333.5	3.00	8.74	114.5	313.7

Qualitative analysis (Fig. 10) shows best-case predictions matching the ground truth with minimal boundary discrepancy; failure cases concentrate around the wrist/arm region or under occlusions such as sleeves, where false positives may leak into the background or boundaries lose crispness patterns consistent with the small lower tail of per-image IoU.

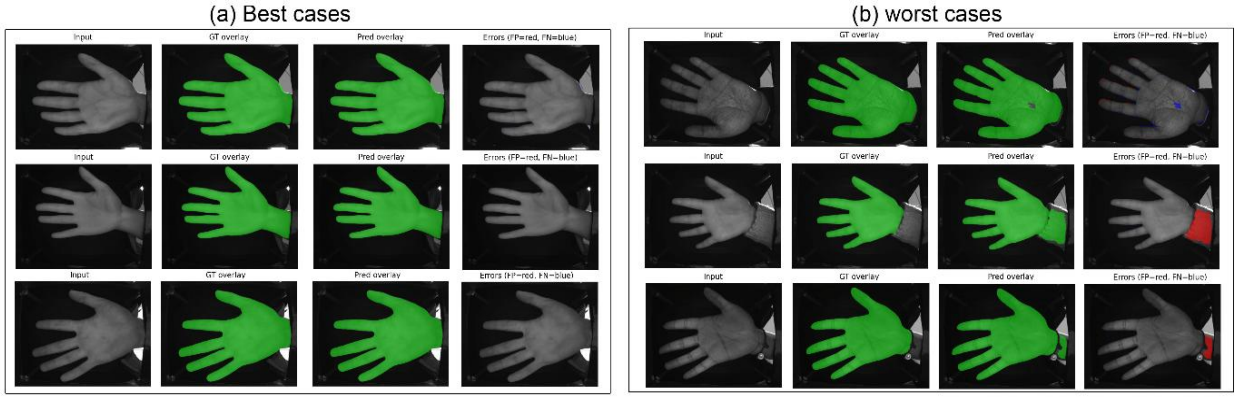


Fig. 10. Qualitative segmentation examples on the test set. (a) Best-performing cases and (b) worst-performing cases. Columns show the input image, GT overlay, prediction overlay, and error map (FP in red, FN in blue).

B. Stage II: Mask-Guided Canonical Alignment (UniRotNet)

Table IV reports mask-only rotation correction across the three training modes. Regression yields the lowest continuous error (MAE = 3.004 ± 0.567°, p95 = 8.552 ± 2.569°), classification yields the highest tolerance-based correctness (Acc@±1 bin = 0.924 ± 0.021), and fusion sits between the two. The tolerance–accuracy curves in Fig. 11 cumulative accuracy P (e ($\hat{\theta}$, θ) ≤ t) versus angular tolerance t (degrees), averaged over three seeds with shaded inter-seed variability give the same picture in distribution form.

TABLE IV. Rotation correction performance across the three modes on the test set (mean ± std over three seeds).

Mode	MAE (deg) ↓	Median (deg)	p90 (deg) ↓	p95 (deg) ↓	Acc@exact-bin↑	Acc@±1bin↑
reg	3.004 ± 0.567	2.235 ± 0.197	6.609 ± 1.833	8.552 ± 2.569	0.473 ± 0.044	0.907 ± 0.050
both	3.085 ± 0.289	2.326 ± 0.052	6.720 ± 1.085	8.689 ± 2.576	0.488 ± 0.073	0.912 ± 0.050
cls	3.581 ± 0.306	2.500 ± 0.000	7.167 ± 0.577	8.500 ± 1.000	0.525 ± 0.025	0.924 ± 0.021

The mode-wise differences in Table IV are small in absolute terms all three configurations sit within a ~0.6° MAE band but reflect different inductive biases: regression directly minimizes continuous angular error, classification trades that for slightly higher tolerance-based accuracy through bin smoothing, and fusion stabilizes predictions near the 0/360° wrap-around. We adopt regression as the headline mode for the remainder of the paper, since its lowest MAE most directly benefits the Stage III ROI regression. Importantly, the residual ~3°

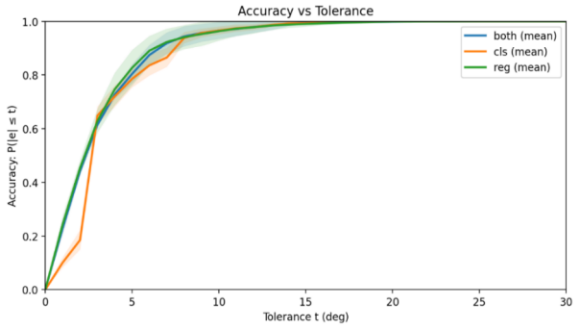


Fig. 11. Tolerance–accuracy curves for Stage II rotation correction. The plot reports cumulative accuracy $P\left(e\left(\hat{\theta},\theta\right)\leq t\right)$versus angular tolerance t(degrees), averaged over three seeds; shaded regions indicate variability across seeds.

MAE is small enough to leave the predicted square ROI essentially unchanged in shape while resolving the 180° wrist–fingertip ambiguity that affects unsupervised principal-axis methods.

Qualitative analysis in Fig. 12 illustrates best- and worst-case predictions. Best-case samples achieve correct upright canonicalization with negligible angular error.

Failure cases primarily occur for near-symmetric silhouettes (e.g., widely spread fingers producing weak directionality) or ambiguous masks where the wrist–fingertip distinction is subtle; nevertheless, the learned model remains stable in these regimes, with no observed 180° flips on the test split.

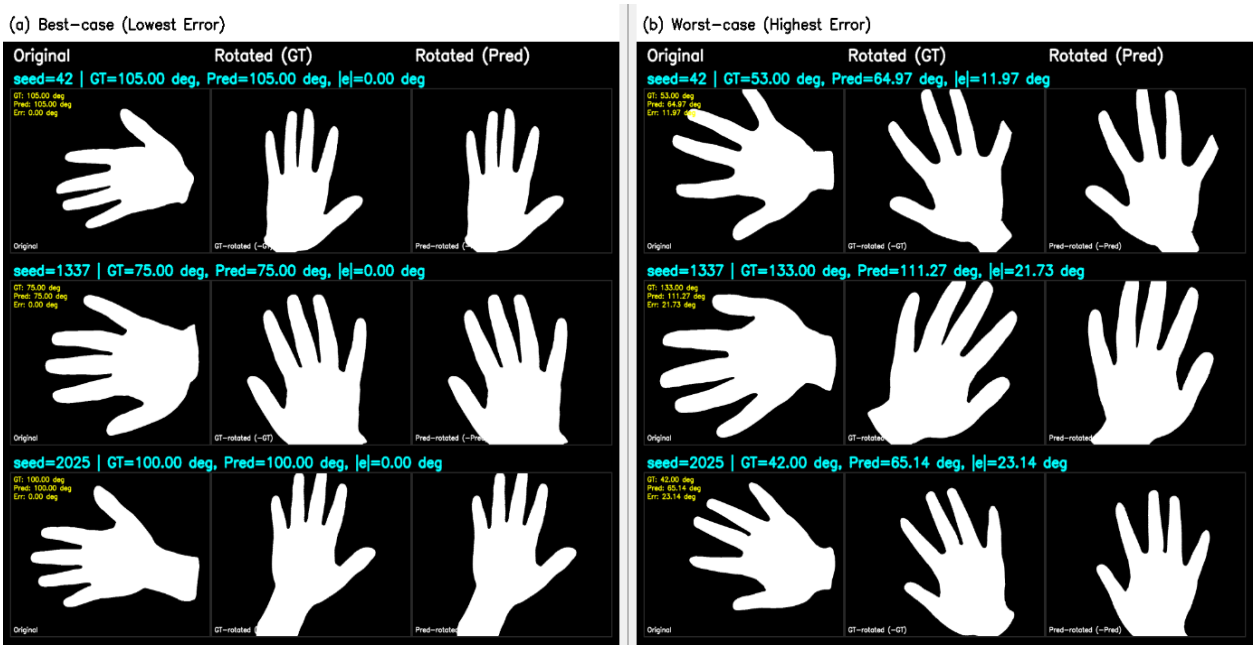


Fig. 12. Qualitative robustness analysis of the proposed UniRotNet across random seeds. (a) Best-case predictions: Samples achieving the lowest angular error, demonstrating precise canonical alignment where the fingers are correctly oriented upwards. (b) Worst-case predictions: Samples exhibiting the highest angular error.

C. Stage III: Palm ROI Localization

Table V reports palm ROI localization performance averaged over three seeds. Both proposed mask-guided regressors achieve strong accuracy with negligible outside-of-hand leakage: Fast-PalmROI reaches mIoU = 0.909 ± 0.002 and Lite-PalmROI 0.905 ± 0.004, with center and size errors near 7 px and 5 px respectively, and outside fractions on the order of 10^{-4}. The predicted square ROI therefore almost always remains within the hand mask.

To test whether the cascaded mask-guided design is necessary, we trained Image-PalmROI: a structurally identical regressor (MobileNetV2 backbone with a 3-element ROI head) that takes the raw RGB image as input and predicts ROI coordinates directly, with no intermediate hand segmentation or canonical-pose alignment. Image-PalmROI uses the same subject-disjoint splits and seeds, with axis-aligned bounding boxes of the rotated ground-truth ROI as raw-frame targets a standard convention for axis-aligned

regressors operating on tilted hand poses. For evaluation, raw-frame predictions are projected back to the aligned frame by inverse rotation through the predicted UniRotNet angle, comparing only the geometric (center and side) parameters without re-applying the bounding-box step.

Table V shows that Image-PalmROI roughly localizes the palm center (center error = 10.9 ± 1.3 px) and almost always overlaps the hand at IoU ≥ 0.5 (Acc@0.5 = 0.982 ± 0.016), but the predicted side length is poorly calibrated (size error = 54.3 ± 4.8 px, an order of magnitude worse than the mask-guided regressors), causing mIoU to drop to 0.641 ± 0.019 and Acc@0.7 to collapse to 0.204 ± 0.065. Part of this size error follows from the target definition: the axis-aligned box of a tilted square is larger than the square itself, so a regressor trained on such boxes tends to over-estimate the side. The cascaded design avoids this by absorbing pose variability in Stage II, leaving Stage III to predict ROIs in a canonical frame where both center and size are well constrained. The comparison is therefore specific to this configuration (axis-aligned box targets) and does not establish that end-to-end localization is inferior in general. Table V also evaluates the mask-guided regressors in isolation, on masks aligned with the ground-truth angle, whereas Image-PalmROI is projected with the predicted UniRotNet angle. The two purely geometric baselines (Fixed-Mean ROI and Mask-BBox shrink) underperform all three learning-based methods on every metric, confirming that learning-based ROI regression is necessary in any frame.

TABLE V. Quantitative comparison of palm ROI localization performance. Results are reported as mean ± standard deviation over three random seeds. The two proposed mask-guided regressors (Fast-PalmROI, Lite-PalmROI) outperform the end-to-end Image-PalmROI baseline across all overlap and error metrics; geometric heuristics (Fixed-Mean, Mask-BBox) underperform all learning-based methods.

Method	mIoU↑	Acc@0.5↑	Acc@0.7↑	Center Err (px)↓	Size Err (px)↓	Outside Frac↓
Fast-PalmROI	0.909 ± 0.002	0.999 ± 0.001	0.990 ± 0.009	7.0 ± 0.3	5.2 ± 0.5	3.67e-04 ± 1.82e-04
Lite-PalmROI	0.905 ± 0.004	1.000 ± 0.000	0.994 ± 0.003	7.5 ±0.1	5.2 ± 0.6	1.95e-04 ± 6.37e-05
Image-PalmROI	0.641 ± 0.019	0.982 ± 0.016	0.204 ± 0.065	10.9 ± 1.3	54.3 ± 4.8	5.78e-02 ± 6.10e-03
Fixed-Mean ROI	0.505 ± 0.028	0.504 ± 0.098	0.091 ± 0.058	62.0 ± 4.0	17.2 ± 0.5	1.02e-01 ± 1.93e-03
Mask-BBox (Shrink) ROI	0.360 ± 0.014	0.132 ± 0.042	0.003 ± 0.002	91.5 ± 4.3	26.9 ± 2.3	1.79e-01 ± 1.89e-03

Both ROI localizers are lightweight, with complementary accuracy–speed profiles (Table VI): Lite-PalmROI is the smaller model (2.31 M parameters), while Fast-PalmROI achieves the lowest GPU latency (0.92 ms/img) at higher parameter count. The consistently small outside fractions indicate that ROI predictions are well-constrained by the aligned mask geometry, supporting stable downstream feature extraction.

Remaining errors are largely attributable to variations in hand size and wrist extent, and to imperfect mask boundaries near the forearm that can shift the apparent palm center.

Robustness plots in Fig. 13 confirm that the learning-based approaches dominate the baselines not only on average but also in the tails.

Qualitative examples (Fig. 14) show that failures concentrate on unusual wrist occlusion or extreme finger spread, where the palm region is less clearly defined; even there, the learned regressors remain stable with minimal leakage outside the hand.

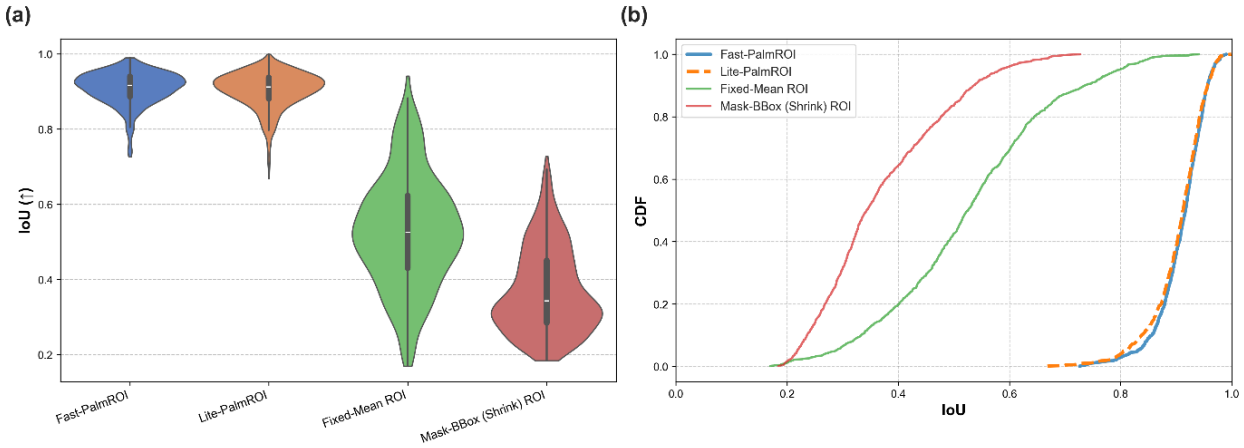


Fig. 13. Per-sample robustness analysis on the test set. (a) Violin plots illustrating the density distribution of IoU scores. (b) Cumulative Distribution Function (CDF) of IoU scores. Both plots demonstrate that Fast-PalmROI and Lite-PalmROI significantly outperform the baselines in terms of stability and median accuracy.

TABLE VI. Complexity and runtime analysis. Performance is measured on a single input (256×256, batch size 1). Fast-PalmROI demonstrates significantly lower latency on GPU, while Lite-PalmROI offers the most compact model size.

Method	Params (M) ↓	Size (MB, FP32) ↓	GMACs↓	CPU (ms/img) ↓	GPU (ms/img) ↓	Peak GPU Mem (MB) ↓
Lite-PalmROI	2.31	8.8	0.417	28.00	6.78	23.9
Fast-PalmROI	8.49	32.4	0.244	7.77	0.92	67.1

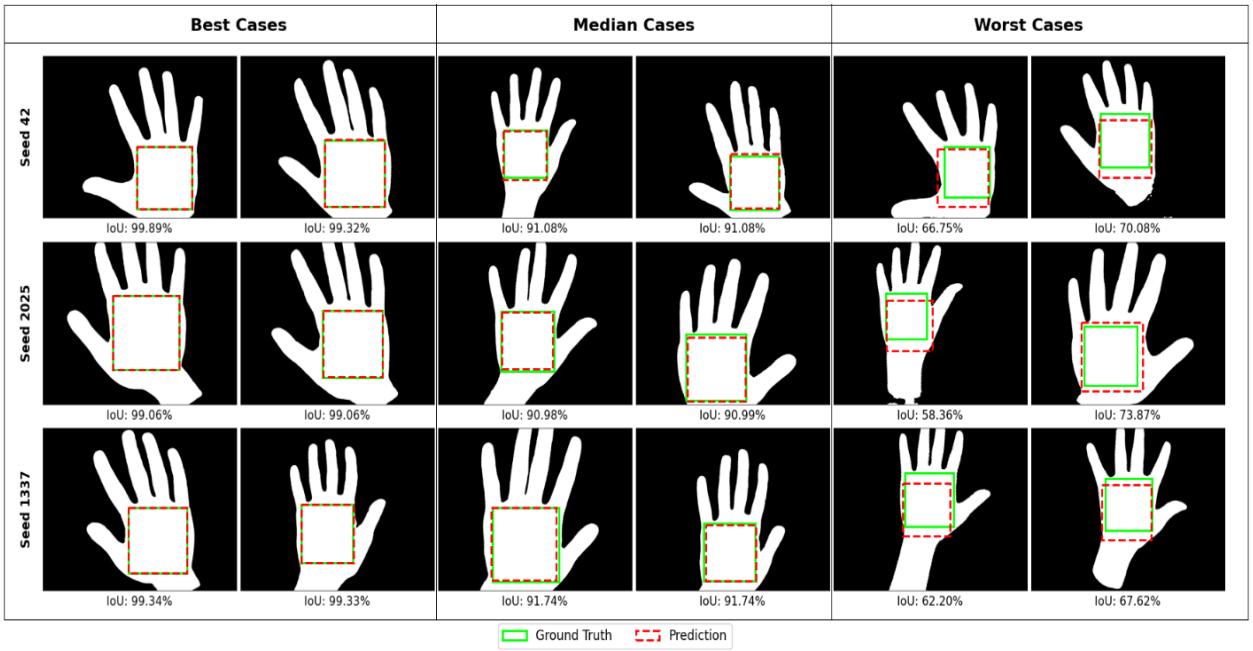


Fig. 14. Visual comparison of ROI localization performance under different initialization seeds. Rows correspond to Lite-PalmROI models trained with seeds 42, 2025, and 1337, respectively. Columns illustrate the model's consistency by showing samples ranging from best-case to worst-case scenarios. Note the consistency of predictions across different seeds even in challenging scenarios (worst cases).

D. Downstream Palmprint Matching Evaluation

ROI localization metrics measure how well a method recovers the GT geometric region, but biometric pipelines are ultimately judged by recognition performance. We therefore measure how well a downstream matcher (described in Section IV-D) performs on ROIs produced by each method, using the CASIA-MS test split with a standard cross-session, single-band protocol: the 30 test identities (15 subjects × 2 hands) provide six 850-nm samples per identity, with samples 01–03 (session 1) used as gallery templates and 04–06 (session 2) as probes. Each probe is scored against all 30 gallery identities (min over the identity's three templates), yielding 90 genuine and 2,610 impostor probe-identity comparisons per (manifest, matcher). The GT, Lite-PalmROI and Fast-PalmROI crops are taken from images aligned with the recorded hand angle, so this experiment isolates the effect of Stage III localization from the angle error of Stage II. We evaluate at 850 nm because the near-infrared band provides the highest contrast for palm-line and vein structure on CASIA-MS and is the most widely reported band in the literature.

Image-PalmROI's predictions live in the raw image frame, rotated by approximately 90° relative to the aligned-frame crops the matcher expects. To avoid conflating ROI-quality differences with frame mismatch and without giving Image-PalmROI access to the per-image UniRotNet angle, which would defeat the end-to-end baseline we apply a single constant 90° rotation to all Image-PalmROI crops at matching time. This approximates the dominant pose alignment with a generic rotation any deployed pipeline could apply, while leaving the residual per-image pose variation that no constant rotation can absorb. The remaining gap between Image-PalmROI and the mask-guided methods therefore measures the value of the per-image alignment stage directly.

Table VII reports EER and Rank-1 averaged over the three seeds. The mask-guided regressors achieve matching performance comparable to GT-localized ROIs on both matchers, within the seed-to-seed variation: Lite-PalmROI obtains EER = 4.11 ± 1.26% and Rank-1 = 90.00 ± 1.11% with the ResNet-18 matcher, against GT's 5.18 ± 1.70% / 91.85 ± 1.70%; Fast-PalmROI is similarly close (4.87 ± 0.60% / 91.48 ± 3.57%). Image-PalmROI degrades substantially: EER rises to 9.66 ± 3.34% (≈ 2× higher) and Rank-1 falls to 82.96 ± 5.25%, a ~9-point drop. The CompCode results show the same trend with a wider spread classical Gabor-orientation coding is more sensitive to residual pose misalignment than deep features and Image-PalmROI's CompCode Rank-1 (58.89 ± 5.09%) trails the mask-guided methods by 12 points. Both observations are consistent with Table V: Image-PalmROI's approximate but loosely sized ROIs partially recover identity through deep features, but classical matchers and tight Rank-1 retrieval require the consistent geometry that only the cascaded mask-guided design supplies.

TABLE VII. Downstream palmprint matching on the CASIA-MS test split (30 identities, 850 nm). EER and Rank-1 are reported as mean ± standard deviation over three random seeds, with three gallery templates (session 1) and three probes (session 2) per identity. The fine-tuned ResNet-18 matcher is trained per seed on the 170 train+val identities (subject-disjoint from the 30 test identities). Lower EER and higher Rank-1 are better.

ROI Source	CompCode EER (%) ↓	CompCode Rank-1 (%) ↑	ResNet-18 EER (%) ↓	ResNet-18 Rank-1 (%) ↑
GT	15.95 ± 5.66	73.33 ± 12.37	5.18 ± 1.70	91.85 ± 1.70
Lite-PalmROI	18.28 ± 6.05	71.11 ± 8.39	4.11 ± 1.26	90.00 ± 1.11
Fast-PalmROI	18.07 ± 5.21	71.48 ± 5.48	4.87 ± 0.60	91.48 ± 3.57
Image-PalmROI	21.56 ± 1.85	58.89 ± 5.09	9.66 ± 3.34	82.96 ± 5.25

E. End-to-End Discussion

The combined evidence from Tables V and VII supports the central design choice of this paper: separating appearance-driven hand segmentation (Stage I) from geometry-driven alignment and ROI localization (Stages II–III), with both later stages operating on binary masks. Stage I produces high-quality masks with crisp boundaries, Stage II resolves the 180° wrist–fingertip ambiguity and reduces residual angular error to ~3°, and Stage III then predicts ROI parameters in a canonical frame where both center and size are well constrained, yielding

mIoU ≈ 0.91. Within the compared configurations, the head-to-head comparison against Image-PalmROI shows that the cascade is not redundant: an end-to-end regressor trained on raw images with axis-aligned box targets recovers approximate palm centers but not tight, consistently sized ROIs, and its matching error is roughly twice that of the mask-guided regressors even with a constant 90° pose correction at evaluation time (the mask-guided crops use the recorded hand angle; Section V-D).

Whether this advantage holds for the complete pipeline with predicted angles, against end-to-end localizers trained with oriented targets, and on images from other sensors remains to be tested.

F. Limitations

Despite the strong results, several limitations remain. First, highly symmetric hand silhouettes (e.g., widely spread fingers forming near-radial symmetry) reduce directional cues and can make rotation correction less reliable in rare cases.

Second, the pipeline depends on the quality of the Stage I mask; boundary leakage near the forearm or wrist typically caused by sleeves, clothing edges, or partial occlusions can perturb the estimated rotation and shift the predicted ROI center. Third, while operating on binary masks improves robustness to illumination and spectral changes, it also removes appearance cues that could help disambiguate challenging cases such as subtle wrist contours. Fourth, all experiments use a single dataset and acquisition setup (CASIA-MS); generalization to other sensors, backgrounds and capture conditions is not evaluated here.

Fifth, the end-to-end comparison uses a single raw-image regressor with axis-aligned box targets; stronger end-to-end localizers were not tested. Sixth, the matching evaluation in Section V-D uses 30 subject-disjoint test identities (90 probes, 2,610 impostor probe-identity comparisons per (manifest, matcher)); EER estimates therefore carry seed-level variability of roughly ±1–2%, and while reporting trends are stable across the three seeds, absolute numbers should be interpreted with this in mind. Finally, the matching evaluation uses a single near-infrared band (850 nm); per-band generalization across the remaining CASIA-MS spectra is left to future work.

VI. CONCLUSION AND FUTURE WORK

This paper presented a mask-guided three-stage pipeline for contactless palm biometrics that converts an input hand image into a stable canonical palm ROI, serving as a robust preprocessing front-end for appearance-sensitive recognition systems.

Stage I (U-Net with a MobileNetV2 encoder) achieves near-saturated segmentation performance (IoU = 0.9911 ± 0.0011) at a favorable accuracy–efficiency trade-off.

Stage II (UniRotNet) achieves precise mask-only rotation correction (MAE = 3.0°) and resolves the 180° wrist–fingertip flip that geometric principal-axis estimators cannot recover from a binary mask alone.

Stage III localizes a square palm ROI with strong overlap (mIoU = 0.909) and negligible outside-of-hand leakage, outperforming an end-to-end raw-image regressor with axis-aligned box targets (Image-PalmROI, mIoU = 0.641) and heuristic geometric baselines. Downstream palmprint matching on the CASIA-MS test split confirms the same ordering: a fine-tuned ResNet-18 matcher applied to the proposed ROIs reaches EER = 4.11–4.87%, comparable to GT-localized ROIs (5.18%) on images aligned with the recorded hand angle, while the same matcher applied to Image-PalmROI degrades to 9.66% EER. Together, these results show that mask-only canonicalization provides a stable intermediate representation that supports reliable downstream feature extraction across the illumination bands of CASIA-MS.

Future work will pursue four directions. First, the three stages can be integrated into a single end-to-end multi-task network with a shared backbone and task-specific heads, enabling joint training with uncertainty-aware error propagation across stages while simplifying deployment. Second, rotation-focused augmentation can be strengthened by sampling correction angles uniformly over [0°, 360°] and including stronger occlusion and shape perturbations to improve robustness in edge cases. Third, the pipeline will be evaluated without re-training on other contactless palm-vein datasets and sensors, such as VERA Palmvein and SCUT-PV, to assess cross-dataset and cross-sensor generalization, and it will be compared with end-to-end keypoint detectors trained with oriented targets and the same geometric augmentation.

Finally, temporal stabilization for video streams accumulating mask, rotation, and ROI predictions across frames would extend the pipeline from single-image to streaming acquisition.

Statements and Declarations

Funding
The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests
The authors have no relevant financial or non-financial interests to disclose.

Data Availability
The data used in this study were obtained from the CASIA Multi-Spectral Palmprint Image Database. Access to the dataset is subject to the data owner's terms and is not publicly redistributable. The dataset can be requested from the official provider via their designated access portal. https://www.idealtest.org/
The ROI labels, trained models and code are available from the corresponding author on reasonable request.

CRediT authorship contribution statement
Amani Almaktari: Conceptualization, Methodology, Software, Investigation, Formal analysis, Data curation, Visualization, Writing – original draft. Mohammed Bokash: Software, Investigation, Data curation, Writing – review and editing. Vikas T. Humbe: Supervision, Conceptualization, Validation, Writing – review and editing, Project administration.

REFERENCES

[1] F. Dai, Z. Wang, X. Zou, R. Zhang, and X. Deng, "Noncontact palm vein ROI extraction based on improved lightweight HRNet in complex backgrounds," *IET Biometrics*, vol. 2024, Art. no. 4924184, 2024, doi: 10.1049/2024/4924184.

[2] A. M. M. Chowdhury, M. J. A. Khondkar, and M. H. Imtiaz, "Advancements in synthetic generation of contactless palmprint biometrics using StyleGAN models," *Journal of Cybersecurity and Privacy*, vol. 4, pp. 663–677, 2024, doi: 10.3390/jcp4030032.

[3] D. Zhong, X. Du, and K. Zhong, "Decade progress of palmprint recognition: A brief survey," *Neurocomputing*, vol. 328, pp. 16–28, 2019, doi: 10.1016/j.neucom.2018.03.081.

[4] M. Dvořák, M. Drahanský, and W. H. Abdulla, "On the fly biometric identification system using hand-geometry," *IET Biometrics*, vol. 10, pp. 315–325, 2021, doi: 10.1049/bme2.12024.

[5] C. Lin, Y. Chen, X. Zou, X. Deng, F. Dai, J. You, and J. Xiao, "An unconstrained palmprint region of interest extraction method based on lightweight networks," PLOS ONE, vol. 19, Art. no. e0307822, 2024, doi: 10.1371/journal.pone.0307822.

[6] Y. Liu and A. Kumar, "Contactless palmprint identification using deeply learned residual features," IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 2, pp. 172–181, 2020, doi: 10.1109/TBIOM.2020.2967073.

[7] D. Fan, X. Liang, C. Zhang, J. Chen, B. Wu, W. Jia, and D. Zhang, "Smart touchless palm sensing via palm adjustment and dynamic registration," Nature Communications, vol. 16, Art. no. 2912, 2025, doi: 10.1038/s41467-025-58213-7.

[8] W. Kang, Y. Liu, Q. Wu, and X. Yue, "Contact-free palm-vein recognition based on local invariant features," PLOS ONE, vol. 9, Art. no. e97548, 2014, doi: 10.1371/journal.pone.0097548.

[9] S. Zhao, L. Fei, and J. Wen, "Multiview-learning-based generic palmprint recognition: A literature review," Mathematics, vol. 11, Art. no. 1261, 2023, doi: 10.3390/math11051261.

[10] L. Zhang, Z. Cheng, Y. Shen, and D. Wang, "Palmprint and palmvein recognition based on DCNN and a new large-scale contactless palmvein dataset," Symmetry, vol. 10, Art. no. 78, 2018, doi: 10.3390/sym10040078.

[11] A.-J. Mohamed-Abdul-Cader, J. Banks, K. N. Thanh, and V. Chandran, "Rotation and scale invariant bispectral feature based recognition of contactless palmprints," in Proc. Int. Conf. Image and Vision Computing New Zealand (IVCNZ), 2019, pp. 1–7.

[12] W. Jia, R.-X. Hu, Y.-K. Lei, Y. Zhao, and J. Gui, "Histogram of oriented lines for palmprint recognition," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 44, pp. 385–395, 2013.

[13] Z. Guo, D. Zhang, L. Zhang, and W. Zuo, "Palmprint verification using binary orientation

co-occurrence vector," Pattern Recognition Letters, vol. 30, pp. 1219–1227, 2009.

[14] K. Luo and D. Zhong, "Robust and adaptive region of interest extraction for unconstrained palmprint recognition," Journal of Electronic Imaging, vol. 30, Art. no. 033005, 2021, doi: 10.1117/1.JEI.30.3.033005.

[15] K. Ito, Y. Suzuki, H. Kawai, T. Aoki, M. Fujio, Y. Kaga, and K. Takahashi, "HandSegNet: Hand segmentation using convolutional neural network for contactless palmprint recognition," IET Biometrics, vol. 11, pp. 109–123, 2022, doi: 10.1049/bme2.12058.

[16] Y. Qiao, W. Kang, D. Luo, and J. Huang, "Normalized-full-palmar-hand: Toward more accurate hand-based multimodal biometrics," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 47, pp. 6715–6730, 2025, doi: 10.1109/TPAMI.2025.3564514.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention MICCAI 2015, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.

[20] D. Zhang, W.-K. Kong, J. You, and M. Wong, "Online palmprint identification," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, pp. 1041–1050, 2003.

[21] A.-K. Kong and D. Zhang, "Competitive coding scheme for palmprint verification," in Proc. 17th Int. Conf. Pattern Recognition (ICPR), 2004, pp. 520–523.

[22] Z. Li, X. Liang, D. Fan, J. Li, and D. Zhang, "BPFNet: A unified framework for bimodal palmprint alignment and fusion," in Neural Information Processing, T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, and A. N. Hidayanto, Eds. Cham, Switzerland: Springer, 2021, pp. 28–36.

[23] K. Yalcin, R. O. Dogan, H. Dogan, and O. Bingol, "High-accuracy hand segmentation for palmprint biometric systems improving auto-encoder techniques," in Proc. 8th Int. Symp. Innovative Approaches in Smart Technologies (ISAS), 2024, pp. 1–4, doi: 10.1109/ISAS64331.2024.10845515.

[24] T. Wu and L. Leng, "Video palmprint recognition system based on modified double-line-single-point assisted placement," Journal of Multimedia Information System, vol. 8, pp. 23–30, 2021, doi: 10.33851/JMIS.2021.8.1.23.

[25] D. Brown and K. Bradshaw, "Deep palmprint recognition with alignment and augmentation of limited training samples," SN Computer Science, vol. 3, Art. no. 11, 2021, doi: 10.1007/s42979-021-00859-3.

[26] L. Shen, Y. Zhang, K. Zhao, R. Zhang, and W. Shen, "Distribution alignment for cross-device palmprint recognition," Pattern Recognition, vol. 132, Art. no. 108942, 2022, doi: 10.1016/j.patcog.2022.108942.

[27] H. Shao, P. Li, and D. Zhong, "Generating stylized features for single-source cross-dataset palmprint recognition with unseen target dataset," *IEEE Transactions on Image Processing*, vol. 33, pp. 4911–4922, 2024, doi: 10.1109/TIP.2024.3451933.

[28] H. Shao, Y. Zou, C. Liu, Q. Guo, and D. Zhong, "Learning to generalize unseen dataset for cross-dataset palmprint recognition," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 3788–3799, 2024, doi: 10.1109/TIFS.2024.3371257.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv* preprint arXiv:1711.05101, 2017, doi: 10.48550/arXiv.1711.05101.

[31] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical*

Statistics, vol. 35, pp. 73–101, 1964, doi: 10.1214/aoms/1177703732.

[32] Y. Hao, Z. Sun, T. Tan, and C. Ren, "Multispectral palm image fusion for accurate contact-free palmprint recognition," in *Proc. 15th IEEE Int. Conf. Image Processing (ICIP)*, 2008, pp. 281–284, doi: 10.1109/ICIP.2008.4711746.