Ridge Scale Aligned Diffusion for Identity Preserving and Style Controllable Fingerprint Synthesis

Anonymous CVPR submission

Paper ID *****

Abstract

Training robust fingerprint recognizers is constrained by privacy regulations and limited intra-class diversity in public datasets. We propose a controllable fingerprint synthesis framework built on Stable Diffusion with ControlNet and a Multi-IP-Adapter derived from IP-Adapter to generate structure-consistent, style-diverse fingerprints conditioned on a content image. Our design supports two modes: (i) augmenting a real identity with diverse sensor styles, and (ii) privacy-motivated synthesis using virtual identities sampled from a DDPM prior to reduce direct dependence on real identity data. To enable fine control, we introduce ridge-scale normalization and dual-mask spatial injection to better separate ridge regions from background during generation. Experiments demonstrate high visual fidelity, consistent retention of content-guided structural cues under style transfer, and improved downstream recognition. In particular, joint training with real and our synthetic data boosts ViT TAR@FAR=0.1% from 80.08% to 91.28%, exceeding joint-training results obtained with FPGAN-Control and PrintsGAN.

1. Introduction

Deep fingerprint recognition models require large-scale data with sufficient *intra-class* variation (multiple impressions per identity under different sensors, pressure, and capture conditions). However, collecting and sharing real fingerprints is heavily restricted due to privacy constraints: biometric leakage is irreversible, and regulations often prevent large-scale acquisition and release. As a result, public datasets are limited in both scale and diversity, leading to overfitting and poor cross-sensor generalization.

Generative models offer a promising alternative by synthesizing realistic fingerprints to expand diversity without additional data collection, including GANs [1] and diffusion models such as DDPM [2]. Prior GAN-based fingerprint synthesis methods can generate high-quality images [3, 4], but structure-consistent augmentation remains challenging due to limited controllability and identity duplication at scale [3]. Diffusion-based fingerprint synthesis has recently shown strong ridge-detail modeling [5–9], yet existing approaches still lack fine-grained control over (i) cross-sensor style transfer, (ii) global shape variation (e.g., plain vs. rolled), and (iii) explicit separation of ridge foreground from background.

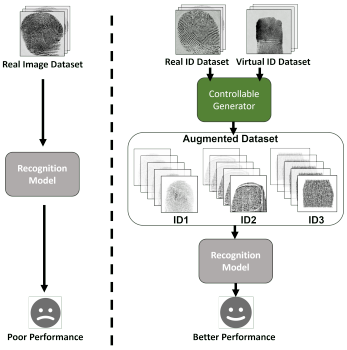


Figure 1. Illustration of structure-consistent augmentation: synthetic fingerprints preserve identity while introducing diverse styles, addressing limited intra-class diversity.

We propose a controllable fingerprint synthesis framework based on Stable Diffusion [10], combining ControlNet [11] for structure-consistent generation and a *Multi-IP-Adapter* derived from IP-Adapter [12] for disentangled style/shape/text control. The framework supports (1) **augmentation mode** that generates diverse impressions from a real identity, and (2) **privacy-motivated mode** that uses DDPM-sampled virtual identities as control inputs [2]. To stabilize cross-style fusion, we further introduce ridge-scale normalization (ridge frequency estimation follows standard orientation-field based formulations [13]) and dual-mask spatial injection.

As illustrated in Fig. 1, our framework produces multiple diverse impressions for each identity, improving intra-class diversity for robust recognition training.

Contributions. (1) A unified diffusion-based pipeline

for **structure-consistent**, style-diverse fingerprint synthesis with real-identity augmentation and **privacy-motivated** virtual-identity generation modes. (2) Multi-domain conditioning (style/acquisition-type/text) with decoupled cross-attention and dual-mask injection for spatial control. (3) Ridge-scale normalization to align ridge frequency between the content and the **target** style. (4) Extensive evaluation showing **consistent retention of content-guided structural cues** and **substantial** recognition gains (e.g., ViT TAR@FAR=0.1%: 80.08% → 91.28% with joint training).

2. Related Work

2.1. Fingerprint Synthesis

Classical simulators (e.g., SFinGe) generate fingerprints via handcrafted ridge models and noise, but are limited in realism and controllability [14]. GAN-based methods improve visual quality for fingerprint synthesis [15–18], yet often suffer from unstable training and limited coverage at scale; structure-consistent augmentation remains challenging due to issues such as identity duplication and constrained controllability [3, 4]. Recent diffusion-based approaches show superior ridge detail modeling and have been explored for fingerprint synthesis and augmentation [5–9], but many lack explicit mechanisms for cross-sensor style control, global shape variation, and foreground/background disentanglement.

2.2. Controllable Generation and Style Transfer

Diffusion models enable high-fidelity synthesis [2, 10], but text-only prompting is often insufficient for preserving precise fingerprint structure. ControlNet injects structural conditions (e.g., edges) to anchor spatial layout [11], while IP-Adapter transfers reference-image priors via decoupled attention [12]. Related controllable diffusion plugins such as T2I-Adapter further demonstrate that lightweight adapters can provide strong controllability while keeping the diffusion backbone frozen [19]. Fingerprint generation imposes stricter constraints than generic style transfer: small geometric distortions can corrupt ridge topology and minutiae. Our method combines ControlNet-based structural anchoring [11] with multi-domain IP-Adapter conditioning [12], augmented by ridge-scale normalization (based on standard ridge/orientation analysis [13]) and dual-mask injection to preserve ridge integrity under strong style changes.

3. Proposed Method

Our proposed framework is built on Stable Diffusion [10] and is designed to achieve structure-consistent, style diversity, and spatially controllable synthesis (Fig. 2). The diffusion process is guided by ControlNet [11], which enforces structural consistency using a Canny edge representation of the input fingerprint. To flexibly control visual appearance, a *Multi-IP-Adapter* derived from IP-Adapter [12] injects style, shape, and text features through decoupled cross-attention, ensuring style diversity while maintaining content-guided structural cues. Spatial controllability is further enhanced with a dual-mask strategy and ridge-scale normalization, which separately condition the fingerprint and background regions and align ridge widths for seamless fusion. We estimate ridge width by patch-wise ridge spacing using an orientation-field based approach [13]; details are provided in the supplementary material. All input and style images are preprocessed by isotropic resizing and white padding to 1024×1024, ensuring consistent spatial alignment.

The overall generation algorithm is presented in Algorithm 1.

Algorithm 1: Controllable Structure-Consistent Fingerprint Synthesis Pipeline

Input: Content fingerprint image C (real or virtual), style image S, text prompt T, fingerprint mask M_{fg}, background mask M_{bg}

Output: Synthetic fingerprint image $\hat{I}$

Step 1: Ridge-Scale Normalization;
Compute average ridge widths w_C and w_S; resize and pad C with ratio $\gamma = \frac{w_S}{w_C}$.

Step 2: Structure-Consistent Conditioning;
Extract a structural representation E_C from C using a Canny edge detector; condition Stable Diffusion with ControlNet on E_C.

Step 3: Multi-Domain Feature Extraction;
Extract style features F_{sty} from S (style encoder); extract shape features F_{sh} (shape encoder); encode text prompt to F_{txt} (text encoder).

Step 4: Decoupled Cross-Attention Fusion;
Fuse features via Eq. (1).

Step 5: Spatially-Guided Injection;
Apply dual-mask injection via Eq. (2).

Step 6: Image Synthesis;
Run the denoising U-Net conditioned on $\{F_{\text{sty}}, F_{\text{sh}}, F_{\text{txt}}\}$ and ControlNet guidance to output $\hat{I}$.

3.1. Structure-Consistent Conditioning

To ensure Structure-Consistent Conditioning, we employ ControlNet [11] with Canny edge detection as a strong structural constraint during the diffusion process. The Canny-processed control image captures ridge flow and topology with thin boundaries that align well with ControlNet's conditioning. In our implementation, Canny thresholds are set to 50/200 with an aperture size of 3.

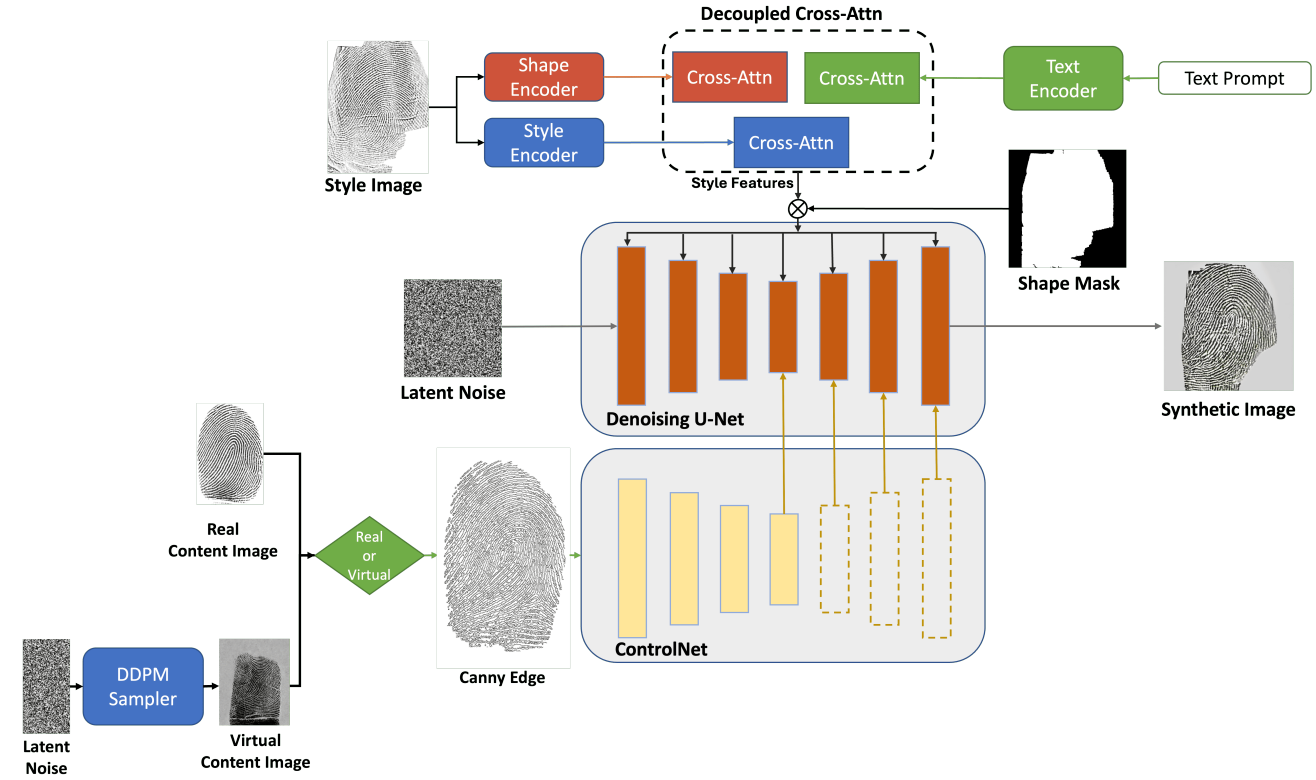


Figure 2. Overview of our proposed method. The lower part preserves content-guided ridge structure via ControlNet and U-Net conditioned on real or virtual content images. The upper part injects style, shape, and text cues through decoupled cross-attention, guided by a shape mask, to generate structurally faithful and style-controllable fingerprints.

Our framework supports two complementary input modes. In the *augmentation mode*, real fingerprint images are directly used as ControlNet inputs. In the *privacy-motivated mode*, we employ a DDPM-based sampler [2] to synthesize virtual identities, which are then used as control inputs for ControlNet. The DDPM sampler is trained from scratch on 5,000 samples from NIST SD 302 [20].

3.2. Multi-IP-Adapter for Fingerprint Style Generation

The IP-Adapter [12] injects visual priors from reference images or text into diffusion models. We extend it to disentangle *style*, *shape*, and *text* control for fingerprint synthesis (Fig. 2).

3.2.1. Style Control

We adopt a CLIP-based image encoder (e.g., OpenCLIP-ViT-H-14 [21]) as the style encoder without additional fine-tuning.

3.2.2. Text-Guided Style Control

Text prompts are encoded by a CLIP text encoder [21]. Among all components, only the text-guided style adapter is fine-tuned on NIST SD 302 [20] (5,000 samples, 100 epochs, lr=1×10^{-4}, AdamW, batch size=8 on a single NVIDIA A6000 GPU).

3.2.3. Shape Control

To explicitly control global geometry (e.g., plain vs. rolled), we extract *shape* features from the style image using a CLIP image encoder and inject them as a separate conditioning stream, preventing the ControlNet structure from dominating shape and enabling type-specific synthesis.

3.2.4. Decoupled Cross-Attention

We fuse text, style, and shape conditions through independent cross-attention branches:

$$\begin{aligned} Z_{\text{new}} = &\ \text{Attn}(Q, K^t, V^t) \\ &+ \lambda_1\, \text{Attn}(Q, K^{\text{sty}}, V^{\text{sty}}) \\ &+ \lambda_2\, \text{Attn}(Q, K^{\text{sh}}, V^{\text{sh}}). \end{aligned} \tag{1}$$

We set $\lambda_1 = \lambda_2 = 1.0$ for all experiments.

3.3. Spatially-Guided Feature Injection via Foreground/Background Masking

To enhance spatial controllability, we use a fingerprint mask M_{fg} and a background mask $M_{\text{bg}} = 1 - M_{\text{fg}}$ (mask generation details are provided in the supplementary material).

Given style features F_{sty} and a background embedding F_{bg}, we inject:

$$F_{\text{out}} = M_{\text{fg}} \odot F_{\text{sty}} + M_{\text{bg}} \odot F_{\text{bg}}. \quad (2)$$

4. Experimental Results

4.1. Visual Evaluation

We conducted a qualitative visual evaluation to demonstrate the style control and generalization ability of our fingerprint synthesis framework. In total, we generated fingerprints under 11 distinct style conditions, covering three acquisition types: rolled, slap, and plain. Eight styles correspond to sensors seen during training (Fig. 3), all from the NIST SD 302 dataset: MorphoWave_Desktop (roll), ANDI_N2N (roll), LIVETOUCH_QUATTRO (roll), S120 (roll), Undisclosed (roll; as labeled in the original dataset), Crossmatch_L_SCAN_1000PX (slap), Crossmatch_L_SCAN_1000PX (roll), and ANDI_N2N (plain). The remaining three styles represent unseen sensor domains from the FVC 2004 dataset (Fig. 3)—Crossmatch V300 (plain), U.are.U 4000 (plain), and FingerChip FCD4B14CB (plain)—which are not included in training, enabling evaluation of cross-domain generalization.

For each style, we display three images: the first two are synthetic fingerprints generated by our method, while the third is a real fingerprint sample acquired from the corresponding sensor. This side-by-side presentation enables direct visual assessment of realism, style fidelity, and structural consistency.

The generated images exhibit diverse visual characteristics and faithfully reproduce differences in ridge contrast, background noise, and image tone across sensors and acquisition types. For styles seen during training, our model closely matches the appearance of real samples and mimics sensor-specific traits. For unseen styles from FVC 2004 [22], our model transfers stylistic attributes (e.g., ridge sharpness and background artifacts) without exposure during training, demonstrating encouraging cross-domain generalization. Notably, synthesized fingerprints preserve ridge-flow from the control image while adapting to target style cues.

4.2. Matching Analysis between Synthetic and Real Content Images

To assess structural consistency after style transfer, we perform matching between each generated fingerprint and its corresponding input content image. As shown in Fig. 4 and Fig. 5, the left image in each subfigure is the input content image, while the right image is the synthetic output. The top row shows styles seen during training, and the bottom row shows unseen styles.

For matching, we adopt LightGlue [23] and XFeat [24], two deep learning-based feature matching methods that leverage dense keypoints and descriptors across the full image. Unlike conventional fingerprint matchers relying on explicit minutiae detection, these methods capture more global and nuanced structural information, making them suitable for evaluating similarity when local textures may shift under style transfer while global ridge-flow patterns remain consistent

The results in Fig. 4 and Fig. 5 show substantial correspondences between content and synthetic images for both seen and unseen styles. Although LightGlue typically yields more matches due to denser representations, both matchers consistently identify meaningful structural alignment, suggesting our method retains structure-level ridge-flow cues under significant style variation.

Table 1. NFIQ 2.0 score (↑) [25] comparison between synthetic and real images.

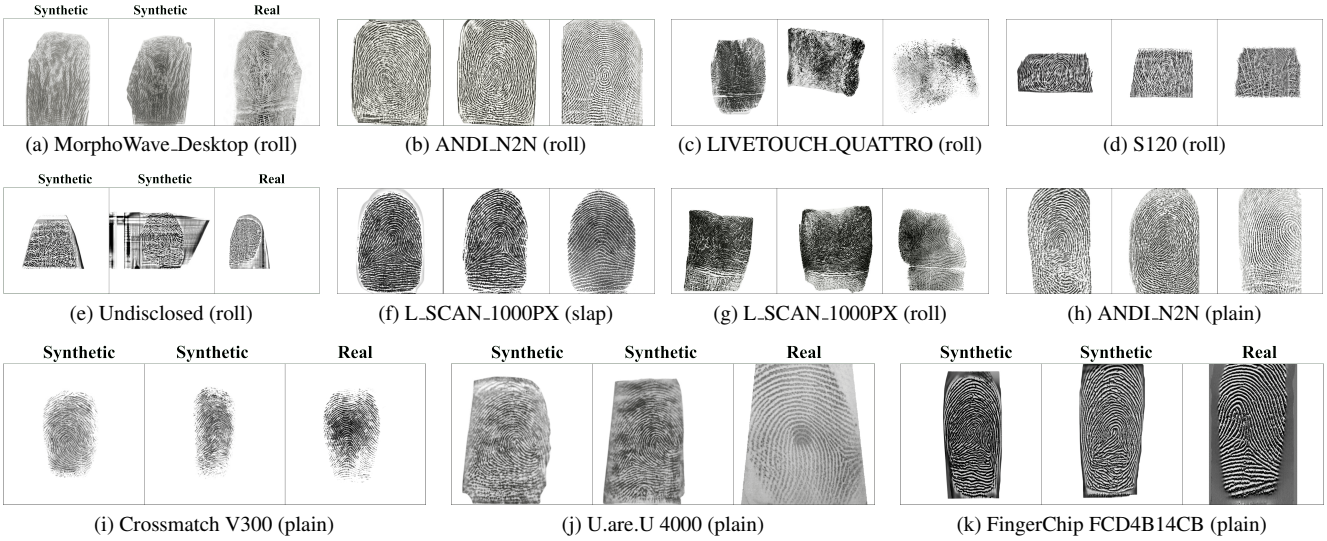
Sensor (Type)	Synthetic $\mu \pm \sigma$	Real $\mu \pm \sigma$	$\Delta\mu$	p-value
MorphoWave_Desktop (roll)	30.79 ± 9.96	30.70 ± 12.65	$+0.09$	0.945
ANDI_N2N (roll)	40.79 ± 7.45	39.65 ± 5.96	$+1.14$	0.292
LIVETOUCH_QUATTRO (roll)	26.89 ± 15.11	31.67 ± 20.29	-4.78	0.236
S120 (roll)	31.51 ± 10.87	43.38 ± 8.61	-11.87	3.44×10^{-11}
Undisclosed (roll)	27.40 ± 14.55	39.09 ± 14.38	-11.69	3.40×10^{-11}
Crossmatch_L_SCAN_1000PX (roll)	37.71 ± 12.34	48.11 ± 20.43	-10.40	4.59×10^{-4}
Crossmatch_L_SCAN_1000PX (slap)	37.02 ± 9.27	40.30 ± 6.37	-3.28	5.20×10^{-13}
ANDI_N2N (plain)	41.61 ± 8.57	40.44 ± 5.17	$+1.17$	0.154

Note: $\Delta\mu = \mu_{\text{syn}} - \mu_{\text{real}}$. Higher NFIQ 2.0 scores indicate better image quality.

4.3. Assessment of Synthetic Fingerprint Quality across Sensors

To quantitatively evaluate synthetic fingerprint quality, we compare NFIQ 2.0 [25] and NIQE [26] between real and synthetic images across eight sensor styles. For each sensor, we generate synthetic images using real fingerprints as style references under identical style conditions. We evaluate 2,000 real and 2,000 synthetic images per sensor. Tables 1 and 2 report the mean and standard deviation of both metrics. We additionally perform Welch's two-sample t-test per sensor to assess whether differences between real and synthetic scores are statistically significant. Figure 6 visualizes NFIQ 2.0 score distributions (blue: real; red: synthetic).

As shown in Table 1, NFIQ 2.0 scores of synthetic fingerprints are generally comparable to real images. For most sensors (e.g., MorphoWave_Desktop (roll) and ANDI_N2N (plain)), mean scores are nearly identical, indicating preserved biometric quality and structural fidelity. However, for certain sensors including S120 (roll) and Cross-

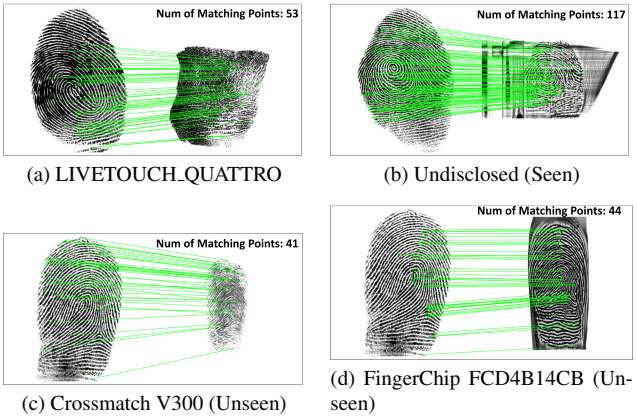


(a) MorphoWave_Desktop (roll) (b) ANDI_N2N (roll) (c) LIVETOUCH_QUATTRO (roll) (d) S120 (roll)

(e) Undisclosed (roll) (f) L_SCAN_1000PX (slap) (g) L_SCAN_1000PX (roll) (h) ANDI_N2N (plain)

(i) Crossmatch V300 (plain) (j) U.are.U 4000 (plain) (k) FingerChip FCD4B14CB (plain)

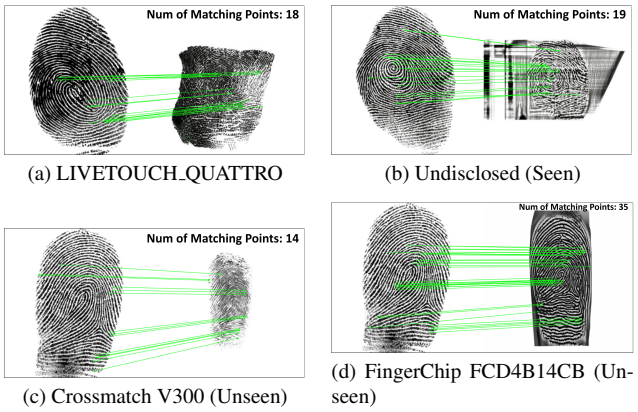
Figure 3. Visual evaluation of generated and real fingerprints. The top two rows show sensors seen during training (NIST SD 302), and the bottom row shows unseen sensors (FVC 2004). For each sensor panel, images (left→right) are two synthetic results and one real image from the same sensor style.



(a) LIVETOUCH_QUATTRO (b) Undisclosed (Seen)

(c) Crossmatch V300 (Unseen) (d) FingerChip FCD4B14CB (Unseen)

Figure 4. Matching results between the input content image (left) and the synthetic fingerprint (right) using LightGlue [23]. Green lines indicate matched feature points. Averaged over 50 test images, the number of matched points is 241.4 for seen styles and 138.1 for unseen styles.



(a) LIVETOUCH_QUATTRO (b) Undisclosed (Seen)

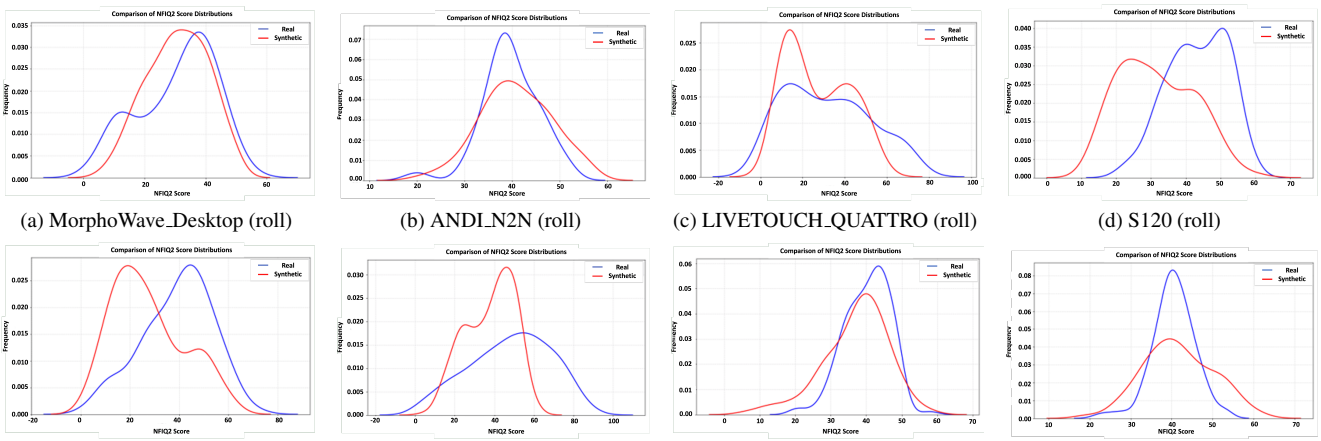
(c) Crossmatch V300 (Unseen) (d) FingerChip FCD4B14CB (Unseen)

Figure 5. Matching results between the input content image (left) and the synthetic fingerprint (right) using XFeat [24]. Green lines indicate matched feature points. Averaged over 50 test images, the number of matched points is 81.5 for seen styles and 77.1 for unseen styles.

match_L_SCAN_1000PX (roll/slap), the mean difference ($\Delta\mu < 0$) is statistically significant, suggesting slightly lower NFIQ 2.0 for synthetic images. This gap likely reflects sensor-specific statistics (e.g., ridge frequency, microcontrast, or device noise) that are not fully captured, rather than a general degradation in generative quality.

In contrast, Table 2 shows that synthetic images often achieve lower NIQE (a proxy for perceptual cleanliness under natural-image statistics) than real fingerprints for most sensors. For example, MorphoWave_Desktop (roll) and Undisclosed (roll) show cleaner ridge textures and reduced background noise. Even for Crossmatch_L_SCAN_1000PX (slap) and ANDI_N2N (plain), NIQE remains comparable, suggesting that style transfer does not introduce noticeable perceptual degradation.

While both NFIQ 2.0 and NIQE are useful quality metrics, they capture different aspects: NFIQ 2.0 assesses biometric usability (e.g., ridge clarity and minutiae extractability), whereas NIQE measures perceptual naturalness via deviations from natural-image statistics. Consequently, some sensors can exhibit divergent trends (e.g., S120 (roll) having lower NFIQ 2.0 but better NIQE), indicating that the



(a) MorphoWave_Desktop (roll) (b) ANDI_N2N (roll) (c) LIVETOUCH_QUATTRO (roll) (d) S120 (roll)

(e) Undisclosed (roll) (f) L_SCAN_1000PX (roll) (g) L_SCAN_1000PX (slap) (h) ANDI_N2N (plain)

Figure 6. Distributions of NFIQ 2.0 scores across eight sensor styles (real: blue; synthetic: red).

Table 2. NIQE score (↓) [26] comparison between synthetic and real images.

Sensor (Type)	Synthetic $\mu \pm \sigma$	Real $\mu \pm \sigma$	$\Delta\mu$	p-value
MorphoWave_Desktop (roll)	5.02 ± 0.67	6.47 ± 0.35	-1.45	1.48×10^{-61}
ANDI_N2N (roll)	7.00 ± 0.96	7.24 ± 0.59	-0.24	0.0579
LIVETOUCH_QUATTRO (roll)	6.16 ± 1.01	12.45 ± 2.24	-6.29	1.00×10^{-75}
S120 (roll)	8.08 ± 0.91	9.57 ± 0.88	-1.49	4.64×10^{-35}
Undisclosed (roll)	8.33 ± 1.51	7.22 ± 2.54	$+1.11$	1.15×10^{-5}
Crossmatch_L_SCAN_1000PX (roll)	6.68 ± 1.15	12.12 ± 0.79	-5.44	1.30×10^{-3}
Crossmatch_L_SCAN_1000PX (slap)	7.06 ± 1.17	7.43 ± 3.05	-0.37	3.50×10^{-94}
ANDI_N2N (plain)	7.57 ± 1.24	7.54 ± 0.53	$+0.03$	0.7566

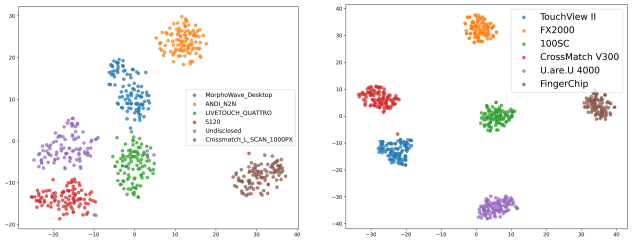
Note: $\Delta\mu = \mu_{\text{syn}} - \mu_{\text{real}}$. Lower NIQE values indicate better perceptual quality.

image may look perceptually clean while being slightly less optimal for biometric quality.

4.4. t-SNE Analysis of Synthetic Sensor Styles

To examine whether our model captures sensor-specific style information, we visualize style embeddings of synthetic fingerprints using t-SNE. Embeddings are extracted from the style encoder of our model (not an external recognition backbone), so the visualization reflects the model's internal representation of sensor style. Figure 7 shows seen and unseen sensor styles.

For Fig. 7a, we select synthetic images from six sensors (MorphoWave_Desktop, ANDI_N2N, LIVETOUCH_QUATTRO, S120, Undisclosed, Crossmatch_L_SCAN_1000PX), removing samples of different fingerprint types from the same sensor to isolate



(a) Seen (b) Unseen

Figure 7. t-SNE visualization of style-encoder embeddings for synthetic fingerprints. (a) Seen sensor styles form compact, well-separated clusters. (b) Unseen sensor styles also form distinct clusters, suggesting the learned style representation extends to novel domains.

sensor-style effects. The resulting clusters are compact and well separated, indicating that our method captures sensor-specific style characteristics.

For Fig. 7b, we visualize sensors not present in training (TouchView II, FX2000, 100SC, Crossmatch V300, U.are.U 4000, FingerChip). Unseen styles also form distinct clusters, suggesting encouraging cross-domain generalization to novel sensor domains.

4.5. Training Recognition Model with Synthetic Data

To evaluate the effectiveness of our generative pipeline for augmenting fingerprint recognition, we conduct experiments using ResNet18/34/50/101 [27] and ViT-B/16 [28] backbones, all pretrained on ImageNet-1K [29]. We consider multiple training scenarios with different dataset compositions.

For real-data training, we use NIST SD 302 (Nail-to-Nail, N2N) [20], containing 1,600 real identities with 10

Table 3. TAR@FAR=0.1% and TAR@FAR=0.01% on different training datasets and architectures.

	TAR@FAR=0.1%					TAR@FAR=0.01%				
Training Dataset	**ResNet18**	**ResNet34**	**ResNet50**	**ResNet101**	**ViT**	**ResNet18**	**ResNet34**	**ResNet50**	**ResNet101**	**ViT**
NIST SD 302 (real) [20]	0.6946	0.7035	0.7132	**0.7457**	**0.8008**	0.4988	0.4741	0.5003	0.5039	**0.5698**
FPGAN-Control [4]	0.7181	0.6682	0.6626	0.6193	0.7875	0.4523	0.4725	0.4833	0.4927	0.5438
PrintsGAN [3]	0.5421	0.5031	0.4561	0.5122	0.5523	0.3951	0.3635	0.3485	0.4201	0.4524
Ours	**0.7251**	**0.7298**	**0.7251**	0.7271	0.7531	**0.5320**	**0.6059**	**0.5477**	**0.5303**	0.5639
NIST SD 302 + FPGAN-Control	0.8093	0.8514	**0.8957**	0.8719	0.8935	0.7251	0.7112	0.7478	0.7251	0.7369
NIST SD 302 + PrintsGAN	0.7299	0.7215	0.7337	0.7793	0.7688	0.5474	0.5747	0.6228	0.6329	0.6701
NIST SD 302 + Ours	**0.8575**	**0.8930**	0.8938	**0.9009**	**0.9128**	**0.7850**	**0.7857**	**0.8056**	**0.8169**	**0.8211**

impressions each. For synthetic-only training, FPGAN-Control [4] and our method each provide 2,000 synthetic identities with 10 impressions per identity, while PrintsGAN [3] provides 1,500 identities with 10 impressions per identity.

In joint training (last three rows of Table 3), the real portion consists of 16,000 images from NIST SD 302 (1,600 identities $\times$ 10 impressions). The synthetic portion includes 20,000 images from FPGAN-Control (2,000 identities $\times$ 10), 15,000 images from PrintsGAN (1,500 identities $\times$ 10), and 20,000 images from our method (2,000 identities $\times$ 10). For our method, we generate 1,000 identities by augmenting real identities (from N2N) and 1,000 identities by sampling virtual identities from a pretrained DDPM; each identity is generated with 10 style variations.

All models are evaluated on a held-out set of 400 real identities from NIST SD 302. We report TAR at FAR=0.1% and FAR=0.01%.

Table 3 shows that our method achieves strong performance across architectures. Models trained on our synthetic data outperform those trained on other synthetic datasets (FPGAN-Control and PrintsGAN). The gains are more pronounced in joint training: NIST SD 302 + Ours consistently yields the best results. For example, with ResNet101, NIST SD 302 + Ours reaches TAR=0.9009 at FAR=0.1%, and with ViT it achieves TAR=0.8211 at FAR=0.01%. We attribute these improvements to the diversity and fidelity of our synthetic data. Unlike methods that solely sample new identities from latent space, our approach combines (i) real identity augmentation to model intra-class variation and (ii) virtual identity generation to expand identity coverage, achieving strong robustness and generalization.

4.6. Computational Complexity Analysis

We compare the computational cost of our diffusion-based pipeline with the GAN-based FPGAN-Control [4]. On an NVIDIA A6000 GPU, our method requires 12.8 seconds per fingerprint image on average, whereas FPGAN-Control requires 0.05 seconds. This difference is expected: GANs generate images in a single forward pass, while diffusion models rely on iterative denoising. Nevertheless, our approach produces finer ridge structures and more complete details, and we observe improved downstream recognition in our experiments. Since fingerprint generation is primarily used for offline data augmentation rather than real-time deployment, the additional cost is acceptable given the improved fidelity and downstream utility.

5. Ablation Study

We perform a progressive ablation study from S0 to S5 to quantify the effect of each design choice. As shown in Table 4, the text-only baseline (S0) provides limited structural consistency, with LightGlue mean matches of 20.01. Adding ControlNet (S1) yields a clear improvement in structural consistency, increasing mean matches to 34.05 while also affecting the perceptual proxies (IS/IL).

We report three complementary metrics: (i) LightGlue mean matches as a proxy for structure-level alignment between the generated image and its paired content input; (ii) Inception Score (IS) as a proxy for perceptual diversity/quality; and (iii) intra-cluster LPIPS distance (IL) to quantify perceptual variation among samples within the same experimental setting (higher IL indicates larger perceptual diversity). Since IS/LPIPS are derived from natural-image models, we use them as complementary proxies rather than fingerprint-specific quality metrics.

Enabling decoupled cross-attention (S2) further strengthens structure-level alignment (40.74 matches), indicating that separating the conditioning streams helps preserve ridge topology under style transfer. With mask injection (S3), mean matches increases to 42.41 and IL decreases slightly (0.5440), suggesting reduced perceptual variation under explicit foreground/background separation. When introducing the style condition (S4), IS remains comparable (3.170) and IL increases to 0.5996 (greater perceptual variation), while mean matches decreases to 40.02, implying stronger style cues can trade off against strict structure-level alignment. Finally, the full model (S5) recovers structure-level alignment (42.52 matches) while keeping IS/IL within a comparable range, showing that combining style and shape conditions can restore structural

Table 4. Progressive ablation from S0 to S5. Results are reported using LightGlue. IS: Inception Score; IL: intra-cluster LPIPS distance (higher indicates larger perceptual diversity).

Metric	S0	S1	S2	S3	S4	S5
IS	2.841	3.678	2.931	3.151	3.170	3.144
IL	0.6795	0.6866	0.5374	0.5440	0.5996	0.5502
LightGlue mean_matches	20.01	34.05	40.735	42.405	40.015	42.52

Table 5. Sensitivity analysis under the full setting S5. IS: Inception Score; IL: intra-cluster LPIPS distance (higher indicates larger perceptual diversity).

Metric	exp11	exp12	exp13	exp14	exp15	exp17	exp18
IS	3.231	2.824	3.285	3.200	3.562	2.7338	2.7109
IL	0.5969	0.5376	0.5962	0.5617	0.5932	0.5409	0.5173
LightGlue mean_matches	38.045	44.275	41.6	40.035	40.305	46.885	59.365

fidelity without sacrificing controllability.

We additionally conduct sensitivity analysis under the full setting S5 (Table 5). shows that stronger guidance increases mean matches, but tends to reduce IS and IL (i.e., lower perceptual diversity), suggesting that overly strong structural constraints may limit style expressiveness. For decoupled cross-attention weights, symmetric settings ($\lambda_1 = \lambda_2$) produce similar IS/IL with moderate differences in matching (exp13–exp14). The asymmetric setting (exp15, $\lambda_1 = 1.5$, $\lambda_2 = 0.5$) achieves the highest IS (3.562) while keeping IL comparable to other settings, indicating that emphasizing style over shape can improve perceptual diversity without a notable loss in structure consistency.

6. Limitation & Future Work

Although our method achieves high-quality fingerprint synthesis and strong structure-consistent, several limitations remain. The diffusion-based generation process is computationally intensive due to iterative denoising steps. Moreover, the framework currently depends on predefined spatial masks to guide region-specific feature injection, which may limit its adaptability under varying sensor conditions. As potential future improvements, we plan to incorporate more efficient diffusion samplers such as DPM-Solver or Latent Consistency Models (LCM) to reduce sampling steps and inference time. Additionally, exploring automatic or data-driven mask estimation could enhance flexibility and scalability for practical deployment.

7. Conclusion

In this paper, we present a fingerprint image synthesis framework that addresses data scarcity under privacy constraints in fingerprint recognition. By integrating Stable Diffusion, ControlNet, and Multi-IP-Adapter modules, our method generates **structure-consistent** and style-diverse fingerprint images, supporting data augmentation and **privacy-motivated** training by reducing direct dependence on real identity data. Experiments show that the proposed synthetic fingerprints improve downstream recognition performance, achieving up to an **11 percentage-point** gain in TAR@FAR=0.1% under joint training compared to the real-only setting.

Overall, this work provides an effective approach to expand fingerprint datasets while reducing direct reliance on real identity data during augmentation. In future work, we plan to improve generation efficiency and further strengthen cross-sensor generalization, extending the framework to other biometric modalities such as palmprint and vein recognition. We believe this research supports diffusion-based synthesis toward practical, **privacy-conscious** biometric applications.

An anonymized code repository is available at: `https://anonymous.4open.science/r/Privacy-and-Anonymity_Privacy-metrics-1F31/`

References

[1] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2, 3

[3] Joshua James Engelsma, Steven Grosz, and Anil K Jain. Printsgan: Synthetic fingerprint generator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6111–6124, 2022. 1, 2, 7

[4] Alon Shoshan, Nadav Bhonker, Emanuel Ben Baruch, Ori Nizan, Igor Kviatkovsky, Joshua Engelsma, Manoj Aggarwal, and Gérard Medioni. Fpgan-control: A controllable fingerprint generator for training with synthetic data. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6067–6076, 2024. 1, 2, 7

[5] Kejian Li and Xiao Yang. Diffusion probabilistic model based end-to-end latent fingerprint synthesis. In *2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML)*, pages 343–349. IEEE, 2023. 1, 2

[6] Freddie Grabovski, Lior Yasur, Yaniv Hacmon, Lior Nisimov, and Stav Nimrod. Diffinger: Advancing synthetic fingerprint generation through denoising diffusion probabilistic models. *arXiv preprint arXiv:2405.04538*, 2024.

[7] Jingqiao Wang, Zicheng Zhang, and Congying Han. Exploring latent fingerprint synthesis with diffusion probabilistic models. In *Proceedings of the Future Technologies Conference*, pages 78–89. Springer, 2024.

[8] Mao-Hsiu Hsu, Yung-Ching Hsu, and Ching-Te Chiu. Inpainting diffusion synthetic and data augment with feature keypoints for tiny partial fingerprints. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 7(3):396–409, 2024.

[9] Weizhong Tang, Diego Figueroa, Donglin Liu, Kerstin Johnsson, and Alexandros Sopasakis. Enhancing fingerprint image synthesis with gans, diffusion models, and style transfer techniques. *arXiv preprint arXiv:2403.13916*, 2024. 1, 2

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2

[11] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 1, 2

[12] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 2, 3

[13] Lin Hong, Yifei Wan, and Anil Jain. Fingerprint image enhancement: algorithm and performance evaluation. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):777–789, 1998. 1, 2

[14] Raffaele Cappelli, A Erol, D Maio, and D Maltoni. Synthetic fingerprint-image generation. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 3, pages 471–474. IEEE, 2000. 2

[15] Yanming Zhu, Xuefei Yin, and Jiankun Hu. Fingergan: A constrained fingerprint generation scheme for latent fingerprint enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8358–8371, 2023. 2

[16] Ataher Sams, Homaira Huda Shomee, and SM Mahbubur Rahman. Hq-fingan: High-quality synthetic fingerprint generation using gans. *Circuits, Systems, and Signal Processing*, 41(11):6354–6369, 2022.

[17] Rafael Bouzaglo and Yosi Keller. Synthesis and reconstruction of fingerprints using generative adversarial networks. *arXiv preprint arXiv:2201.06164*, 2022.

[18] Ogban-Asuquo Ugot, Chika Yinka-Banjo, and Sanjay Misra. Biometric fingerprint generation using generative adversarial networks. In *Artificial Intelligence for Cyber Security: Methods, Issues and Possible Horizons or Opportunities*, pages 51–83. Springer, 2021. 2

[19] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024. 2

[20] Gregory P Fiumara, Patricia A Flanagan, John D Grantham, Kenneth Ko, Karen Marshall, Matthew Schwarz, Elham Tabassi, Bryan Woodgate, and Christopher Boehnen. *Nist special database 302: Nail to nail fingerprint challenge*. Gregory P. Fiumara, Patricia A. Flanagan, John D. Grantham, Kenneth Ko . . . , 2019. 3, 6, 7

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3

[22] Dario Maio, Davide Maltoni, Raffaele Cappelli, Jim L Wayman, and Anil K Jain. Fvc2004: Third fingerprint verification competition. In *International conference on biometric authentication*, pages 1–7. Springer, 2004. 4

[23] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 17627–17638, 2023. 4, 5

[24] Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R Nascimento. Xfeat: Accelerated features for lightweight image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2682–2691, 2024. 4, 5

[25] Elham Tabassi, Martin Olsen, Oliver Bausinger, Christoph Busch, Andrew Figlarz, Gregory Fiumara, Olaf Henniger, Johannes Merkle, Timo Ruhland, Christopher Schiel, et al. Nfiq 2 nist fingerprint image quality. 2021. 4

[26] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 4, 6

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6