Learnable Query-Enhanced Pose Transformation

Yi-Zhen Wang, Hong-Han Shuai
National Yang Ming Chiao Tung University
yzwang.ii12, hhshuai@nycu.edu.tw

Abstract

Pose-Guided Person Image Synthesis (PGPIS) aims to transfer a person from a source image to a target pose (e.g., skeleton) while preserving their original appearance. Although existing methods can produce high-quality results at first glance, they often suffer from noticeable distortions in fine details. We identify the root cause of these issues as the heavy reliance on pre-trained encoders for extracting visual features from the source image. To address this, we propose a novel Query Enhancement Network composed of two key components: the Query-based Feature Fusion Transformer (QFFT) and Pose-Masked Attention (PMA). The QFFT uses learnable queries to fuses multi-scale features from high to low resolution extracted by the backbone encoder, thereby significantly enhancing the realism of texture details in the generated images. To better capture the relationship between pose information and visual features from the source image, we introduce PMA that uses the pose skeleton as a mask to guide the attention mechanism to focus on the pose regions. Our method produces high-quality, visually coherent results and outperforms existing approaches on standard evaluation metrics, including FID, SSIM, and LPIPS, demonstrating its effectiveness on the DeepFashion dataset.

1. Introduction

Pose-guided person image synthesis (PGPIS) aims to transfer the appearance of a source person onto a specified target pose, facilitating various applications such as virtual reality and [14]. Additionally, images generated by PGPIS can serve as valuable augmented data for downstream tasks, including person re-identification and fashion Image classification [6, 36, 41]. Despite its considerable promise, PGPIS is challenging due to substantial discrepancies between source and target poses. These differences often hinder the accurate preservation of a subject's identity and detailed clothing features. Furthermore, achieving visual consistency while maintaining image realism poses another major difficulty.

To address these challenges, early methods predominantly relied on Generative Adversarial Networks (GANs) [4, 7, 11, 13, 18, 20, 21, 25, 29, 30, 34, 37, 40]. Recently, diffusion models [9, 31] have emerged as robust alternatives, capable of generating high-quality images by progressively refining noisy inputs through iterative denoising steps. Numerous diffusion-based techniques have thus been developed for PGPIS [1, 8, 12, 17, 33, 35], leveraging various strategies to preserve subject details and image realism. For instance, some methods utilize disentangled guidance [1, 8] or structured denoising processes [26, 39] to enhance visual consistency, while others employ auxiliary feature extractors to maintain fine-grained characteristics of the source image [12, 33]. Among these, CLIP has gained particular prominence due to its training on large-scale image–text paired datasets. The resulting image embeddings effectively capture rich visual content and stylistic attributes, offering notable advantages in both semantic alignment and visual coherence. PCDMs [33] leverage multi-stage diffusion inference along with auxiliary models

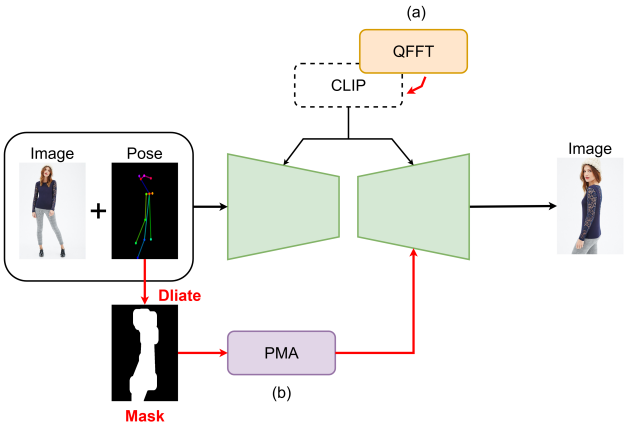

Figure 1. A simplified pipeline of our method: black lines indicate existing components, while red lines represent our proposed additions. (a) We replace the original CLIP encoder with Query-based Feature Fusion Transformer. (b) We additionally introduce a pose mask generated from the pose skeleton to help the model enhance its perception of fine details.

such as CLIP [28] and DINOv2 for detailed feature extraction. Another representative approach, MCLD [12], selectively extracts critical regions (e.g., face and clothing) instead of uniformly processing the entire image.

While leveraging auxiliary vision-language models like CLIP [28] has demonstrated effectiveness in preserving semantic coherence, such approaches also come with notable limitations. First, CLIP is trained using a contrastive loss function that maximizes similarity between matched image-text pairs while minimizing similarity between mismatched pairs. Although powerful in capturing general semantic correspondences, this training approach tends to sacrifice sensitivity to fine-grained visual distinctions. As a result, methods relying solely on CLIP may fail to adequately distinguish subtle but visually significant differences in textures or shapes. Second, CLIP's Vision Transformer (ViT) encoder accepts input images at a fixed resolution (up to 384^2 pixels), inherently limiting the model's ability to process higher-resolution images. Consequently, these methods may underperform in scenarios requiring precise spatial recognition or detailed visual synthesis (e.g., logos on clothing). Thus, a more flexible feature extraction approach that is sensitive to fine-grained details and capable of handling higher-resolution inputs remains desirable.

To tackle these limitations, we propose a novel Query Enhancement Network for PGPIS. In contrast to conventional image-to-image paradigms in Stable Diffusion, which typically rely on pretrained CLIP models to extract image features as conditions, our approach replaces CLIP with a Query-based Feature Fusion Transformer (QFFT), as illustrated in Fig. 1(a). This transformer employs a set of learnable query vectors, which are randomly initialized and iteratively refined through cross-attention layers to progressively extract fine-grained, identity-preserving features across multiple levels. To further guide the model in focusing on pose-relevant regions, we introduce a Pose-Aware Attention Mask derived from the target pose skeleton, shown in Fig. 1(b). This mask is injected into a specialized attention mechanism that dynamically modulates the focus on critical areas such as joints and limb boundaries, enabling more precise synthesis of structural details.

Our contributions can be summarized as follows:

- We propose the Query-based Feature Fusion Transformer (QFFT) to replace the pre-trained CLIP encoder, enabling more effective fusion of multi-level features. This enhances both identity preservation and visual realism.
- We design a Pose-Aware Attention Mask integrated into ControlNet to selectively emphasize pose-critical regions. This targeted attention mechanism improves the synthesis of intricate visual details and spatial accuracy.
- Experimental results show that the proposed method achieves state-of-the-art results on the DeepFashion [15] dataset, surpassing the state-of-the-art methods by +0.2% in SSIM and +1.7% in LPIPS, alongside consistent qualitative improvements.

2. Related Work

Pose-guided person image synthesis (PGPIS) has evolved substantially from its GAN-based origins. Ma *et al.* [20] first introduced the task using GANs, prompting diverse approaches to model pose transfer via affine transformations [34], flow-based warping [11, 29], and semantic region decomposition [37]. Despite these efforts, early models struggled with texture distortions and poor semantic consistency across transformed poses. UV mapping techniques [7] attempted to establish spatial correspondences between pose and appearance but remained limited in preserving fine-grained visual details. With the advent of diffusion models [9, 31], PGPIS experienced significant improvements in synthesis quality. PIDM [1] pioneered the use of pixel-level diffusion conditioned on 2D poses, while latent diffusion improved computational efficiency. Further advancements include PoCoLD [8], which leverages 3D pose conditioning for better structural modeling, and MCLD [12], which disentangles body regions to enhance identity preservation.

To improve semantic alignment, recent methods integrate auxiliary vision-language models like CLIP [28], as seen in PCDMs [33]. These approaches enrich feature representations but suffer from CLIP's fixed input resolution and its contrastive training objective, which limits sensitivity to fine-grained texture differences. In contrast, the proposed Query Enhancement Network addresses these challenges by replacing fixed encoders with the Query-based Feature Fusion Transformer (QFFT) and introducing Pose-Masked Attention (PMA). This combination enables adaptive multi-resolution feature integration and pose-guided attention, leading to more accurate structural rendering and enhanced texture realism in synthesized person images.

3. Method

3.1. Preliminary

Stable Diffusion. Our model builds on Stable Diffusion [31], a text-to-image latent diffusion model comprising two main components: a Variational Autoencoder (VAE) [5] and a U-Net based denoising network [32]. The VAE encodes images from pixel space into compact latent representations, and the UNet predicts the noise added to these latent variables. The training procedure follows the Denoising Diffusion Probabilistic Model (DDPM) framework [9], which defines a forward diffusion process that gradually corrupts the data and a backward process that learns to recover the original signal.

In the forward process, or diffusion process, Gaussian noise is progressively added to the latent variable x_0 over

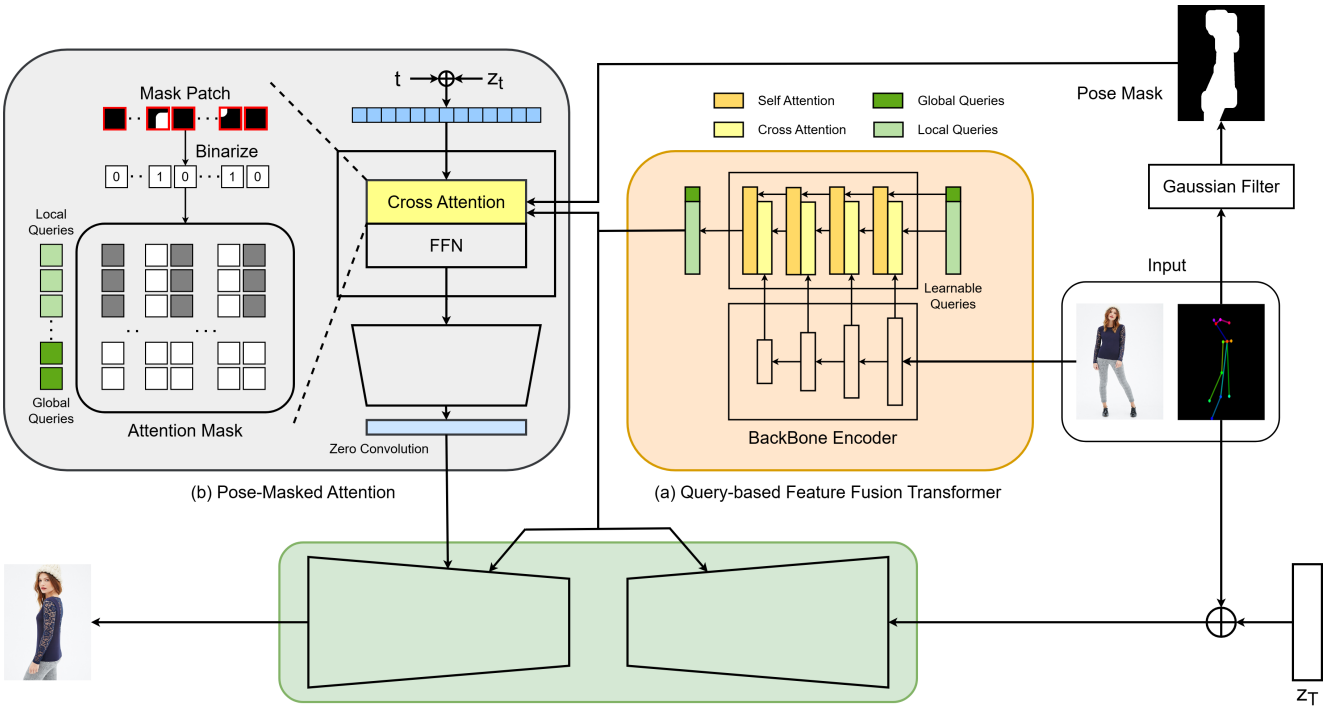

Figure 2. The overall pipeline of our Query Enhancement Network, which is built upon Stable Diffusion. For clarity and to highlight the key components of the proposed framework, the encoder and decoder of the VAE are intentionally omitted here. Additionally, we introduce a mask generated from the skeleton.(a) Through our Query Feature Fusion Transformer, the learnable query progressively integrates features of different resolutions. (b) The mask is introduced into the transformer to enhance texture detail.

$T = 1000$ steps, where $t \in [1, 1000]$:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{1}$$

where $\bar{\alpha}_t$ is derived from a fixed variance schedule, and ϵ is sampled from a standard Gaussian distribution.

The backward process, or denoising process, trains the UNet $\epsilon_\theta(x_t, t, c)$ to predict the noise ϵ from the noisy latent x_t, conditioned on t and an additional embedding c. The objective is to recover the clean latent from its noisy counterpart. The training loss is formulated as:

$$L_{\text{mse}} = \mathbb{E}_{x_0,c,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2\right]. \tag{2}$$

ControlNet. To enable controllability in the image generation process, ControlNet [38] extends Stable Diffusion by introducing external conditional inputs. The key idea is to augment the original U-Net architecture with a learnable and lightweight control branch. Specifically, ControlNet duplicates the U-Net's downsampling blocks and middle block, inserting specialized layers known as "zero convolutions" into these copies. The outputs from the control branch are then injected into the corresponding skip connections of the original U-Net as residuals, effectively balancing conditional guidance with generative consistency.

3.2. Query Enhancement Network

Overview. Our proposed framework, illustrated in Fig. 2, enhances Stable Diffusion for pose-guided person image synthesis by introducing two key modules: the **Query-based Feature Fusion Transformer (QFFT)** and **Pose-Masked Attention (PMA)**. Given training inputs $\mathcal{I} = x_s, x_g, x_{sp}, x_{tp}$, where x_s is the source image, x_g is the ground truth, x_{sp} is the source pose, and x_{tp} is the target pose, the goal is to generate a person image that preserves the identity and appearance of x_s while matching the structure of x_{tp}. The source image x_s is first encoded by a multi-scale backbone, yielding hierarchical features $F_s = [F_1, F_2, F_3, F_4]$. To prevent the direct use of F_s from introducing redundant or misaligned information, we design the Query-based Feature Fusion Transformer (QFFT). QFFT employs a set of learnable queries that progressively fuse features from different scales through a cross-attention followed by self-attention mechanism. Specifically, local queries capture fine-grained textures, while global queries encode holistic appearance and structural information, producing a unified, identity-preserving representation to condition the diffusion process.

To further enhance control over human structure, we integrate Pose-Masked Attention (PMA) into the ControlNet

branch of Stable Diffusion. PMA leverages a binary mask derived from the target pose x_{tp} to guide the attention mechanism, focusing it on body-relevant regions in the latent space. This spatial prior ensures accurate synthesis of limb positions and joint relations, while preserving global consistency through unrestricted global queries. The training objective combines a noise prediction loss $\mathcal{L}_{\text{mse}}$ that supervises the recovery of the clean latent from the ground truth x_g and an auxiliary source-to-source reconstruction loss $\mathcal{L}_{\text{rec}}$ conditioned on the source pose x_{sp} that encourages detailed appearance preservation and pose alignment. Together, QFFT and PMA enhance the controllability and fidelity of the generated images, effectively addressing challenges in pose-guided person image synthesis.

Query-based Feature Fusion Transformer. In contrast to prior methods that rely on CLIP-based embeddings for conditioning, we propose the Query-based Feature Fusion Transformer (QFFT) to directly fuse multi-scale features F_s extracted from the source image. While the idea of using learnable queries is common in recent architectures such as Q-Former [10], our design explicitly separates the queries into local queries (Q_l) and global queries (Q_g), each serving distinct roles in the fusion process. The local queries are dedicated to capturing salient fine-grained features, such as textures and material details, through cross-attention with the hierarchical features F_s. In contrast, the global queries are designed to aggregate holistic appearance and structural information across the entire image by attending to all local queries via self-attention.

Formally, in the cross-attention stage, each local query in Q_l interacts sequentially with the feature maps F_i where $i \in 1, 2, 3, 4$, acting as queries Q, while F_i serve as keys K and values V. This step enables the local queries to selectively gather useful regional details across resolutions. Subsequently, the local and global queries are concatenated and processed together via self-attention, where the global queries capture the overarching semantics and reinforce the local queries by contextualizing them within the global structure of the person. This two-stage attention strategy, which first refines local details and then aggregates global information, yields a unified and compact representation that retains both critical identity features and pose-relevant structures. With this design, QFFT provides the diffusion model with conditioning features that balance fine-grained texture fidelity with overall structural coherence.

Pose-Masked Attention. To better capture the relationship between pose information and source image features in the latent space z_t, we propose the Pose-Masked Attention (PMA) module. PMA is designed to enhance attention between human-centric regions in z_t and the learnable queries, ensuring that pose guidance is explicitly injected into the feature interaction process. Concretely, we first divide z_t into L non-overlapping patches of size $p \times p$, and project them into query embeddings $Q = W_q z_t$. Simultaneously, the learnable feature queries Q_L are projected into keys $K = W_q Q_L$ and values $V = W_v Q_L$. The attention is computed as:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{3}$$

where d_k is the dimension of the key vectors, used to normalize the dot product and ensure numerical stability. To focus attention on pose-relevant regions, PMA further introduces a spatial mask M that modulates the attention computation according to the target pose:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right)V. \tag{4}$$

The mask M is constructed from a binary pose mask m_p, which is generated by binarizing the target pose x_{tp} and applying Gaussian blur to dilate the human body regions. To align with the latent patch structure, m_p is partitioned into L patches corresponding to z_t, forming $m'_p = m_p^{(1)}, \ldots, m_p^{(L)}$, where $m_p^{(l)} = 1$ if any pixel within the l-th patch lies within the dilated pose region, and 0 otherwise. This produces a spatial prior that emphasizes patches containing human body information.

The set of learnable queries is defined as $\mathcal{Q} = \{q^{(1)}, \ldots, q^{(N)}\}$, encompassing both global and local queries, each with distinct functional roles. To preserve the global queries' capacity to capture holistic semantics, we explicitly exempt them from masking constraints, allowing them to attend freely to all latent patches. Let $M \in \mathbb{R}^{L \times N}$ denote the resulting attention mask matrix, where M_{ij} controls the interaction between the i-th latent patch and the j-th query; here, i indexes the latent patches of z_t and j indexes the queries in $\mathcal{Q}$. The mask M is formally defined as:

$$M_{ij} = \begin{cases} 0 & \text{if } m_p^{(i)} = 1, \\ 0 & \text{if } q^{(j)} \text{ is global query}, \\ -\infty & \text{otherwise.} \end{cases} \tag{5}$$

This masking strategy ensures that local queries attend only to pose-relevant patches, effectively guiding the model to focus on human body details. In contrast, global queries remain unrestricted to maintain a coherent understanding of the overall structure. By disentangling the attention pathways for local and global queries, PMA strikes a balance between detailed pose fidelity and global semantic consistency in the generated images.

3.3. Training Objective

The training objective of our model is designed to ensure both faithful reconstruction of the target image and robust alignment between pose and appearance features. The initial latent representation is obtained by encoding the ground truth image x_g using the VAE encoder, denoted as $z_0 = \mathcal{E}(x_g)$.

To guide the model in synthesizing the target pose x_{tp} conditioned on the source image x_s, we formulate the primary objective as a noise prediction loss based on the denoising diffusion process:

$$L_{\text{mse}} = \mathbb{E}_{z_0, x_s, x_{tp}, \epsilon, t}\left[\|\epsilon - \epsilon_\theta(z_t, t, x_s, x_{tp})\|_2^2\right], \quad (6)$$

where ϵ is the sampled Gaussian noise, t is the diffusion timestep, and ϵ_θ is the model's predicted noise given the current noisy latent z_t, timestep t, the source image x_s, and the target pose x_{tp}. This term supervises the model to denoise the latent representation toward the correct target pose.

Inspired by the dual-task learning strategy in DPTN [40], we further introduce an auxiliary source-to-source reconstruction loss, where the model is tasked with reconstructing the source image under the guidance of the source pose x_{sp}. This auxiliary objective reinforces the model's understanding of identity and appearance consistency:

$$L_{\text{rec}} = \mathbb{E}_{z_0, x_s, x_{sp}, \epsilon, t}\left[\|\epsilon - \epsilon_\theta(z_t, t, x_s, x_{sp})\|_2^2\right]. \quad (7)$$

Finally, the total loss is a simple summation of the two terms, balancing the learning of target pose generation and source appearance preservation:

$$L_{\text{overall}} = L_{\text{mse}} + L_{\text{rec}}. \quad (8)$$

This joint optimization not only stabilizes training but also ensures that the model maintains fidelity in both pose transfer and identity preservation.

4. Experiments

4.1. Setup

Datasets. We conduct experiments on the DeepFashion [15] In-shop Clothes Retrieval Benchmark, which contains 52,712 high-resolution person images. We use OpenPose [2] to extract 18 joint keypoints, constructing skeleton-based pose representations. Following the data split defined by GFLA [29], the dataset is partitioned into 101,966 training pairs and 8,570 testing pairs, where each pair depicts the same person in different poses. This setup provides a diverse and challenging benchmark for pose-guided person image synthesis.

Method	Venue	FID↓	LPIPS↓	SSIM↑	PSNR↑
Evaluate on 256×176 resolution					
PATN [44]	CVPR 19'	20.72	0.253	0.671	-
ADGAN [24]	CVPR 20'	14.54	0.225	0.673	-
GFLA [29]	CVPR 20'	9.82	0.187	0.708	-
PISE [37]	CVPR 21'	11.51	0.224	0.653	-
SPGNet† [19]	CVPR 21'	16.18	0.225	0.696	17.22
DPTN† [40]	CVPR 22'	17.41	0.209	0.697	17.81
NTED† [30]	CVPR 22'	8.51	0.177	0.715	17.74
PIDM† [1]	CVPR 23'	6.36	0.167	0.731	-
PoCoLD [8]	ICCV 23'	8.06	0.164	0.731	-
CFLD‡ [17]	CVPR 24'	6.80	0.157	0.734	18.23
MCLD‡ [12]	CVPR 25'	6.69	0.148	0.751	18.84
Ours	-	6.71	0.150	0.747	18.37
Evaluate on 512×352 resolution					
CoCosNet2 [42]	CVPR 21'	13.12	0.226	0.723	-
NTED † [30]	CVPR 22'	7.64	0.199	0.735	17.39
freqHPT [23]	CVPRW 23'	6.55	0.202	0.745	-
WaveIPT [22]	ICCV 23'	4.82	0.242	0.741	-
CFLD‡ [17]	CVPR 24'	7.14	0.181	0.747	17.65
MCLD‡ [12]	CVPR 25'	7.07	0.175	0.755	18.21
Ours	-	7.04	0.172	0.757	18.01

Table 1. Qualitative comparison with state-of-the-art methods. †We follow the data splitting protocol from GFLA [29], and the corresponding scores are evaluated and reported using the method from CFLD [17], as it adopts the same validation set partitioning strategy. ‡Results are obtained in their paper, since splitting protocol are identical.

Metrics. We evaluate the quality of generated images using four objective metrics: Fréchet Inception Distance (FID), *Learned Perceptual Image Patch Similarity* (LPIPS), *Structural Similarity Index Measure* (SSIM), and *Peak Signal-to-Noise Ratio* (PSNR). FID and LPIPS are feature-based metrics computed from deep neural networks, with FID measuring the Wasserstein-2 distance between feature distributions extracted by a pre-trained Inception-v3 model, and LPIPS quantifying perceptual similarity in the learned feature space. SSIM and PSNR are pixel-level metrics that assess structural and signal fidelity between the generated and ground truth images.

Implementation Details. Our model is implemented based on Stable Diffusion [31] (version 1.5), using PyTorch [27] and the HuggingFace Diffusers library. Both source and target images are resized to 512×512. The Query-based Feature Fusion Transformer adopts the Swin-B [16] encoder pretrained on ImageNet [3].

4.2. Quantitative Comparison

We compare our method against 12 advanced approaches, covering GAN-based, flow-based, attention-based, and diffusion-based methods. Evaluations are conducted at two resolutions, 256×176 and 512×352. As shown in Tab. 1, our method consistently achieves state-of-the-art perfor-

Figure 3. Qualitative comparison on the DeepFashion dataset. From left to right: source image, target pose, ground truth, DPTN, PIDM, PCDMs, MCLD, and our method. Our approach produces superior image quality, more accurately reconstructing clothing patterns and text, while effectively reducing distortion and deformation.

mance across multiple metrics at the higher resolution of 512×352. At 256×176, our performance is slightly lower than MCLD, which may be attributed to MCLD's use of 3D pose information that offers more explicit structural constraints. Furthermore, our FID score is marginally higher than that of PIDM, possibly due to PIDM's weighting strategy, which aligns closely with the dataset distribution but tends to introduce overfitting.

4.3. Qualitative Comparison

Fig. 3 presents a qualitative comparison of our method against recent state-of-the-art approaches on the DeepFashion dataset. From left to right, each column shows the source image, target pose, ground truth, DPTN[40], PIDM [1], PCDMs [33], MCLD [12], and our method. Across a variety of poses, identities, and clothing styles, our method consistently yields visually superior results.

We make three key observations from these comparisons. First, while both GAN-based and diffusion-based baselines have made progress, most methods still struggle with accurately preserving fine clothing textures such as patterns, fabrics, and text. For example, other methods often generate distorted or blurred patterns when the clothing involves intricate designs. In contrast, our method, aided by the Query-based Feature Fusion Transformer, better captures and reconstructs these high-frequency details, leading to clearer and more faithful appearance synthesis.

Second, when large pose transformations are involved, prior methods frequently introduce artifacts such as disproportionate limbs or structural misalignment. Our method mitigates these issues, maintaining both the integrity of the human body structure and the fidelity of clothing at-

Method	PIDM [1]	PCDMs [33]	MCLD [12]	Ours
Preferences	12.8%	18.7%	22.5%	46.0%

Table 2. User Study about the preferences of generated images towards ground truths.

Method	Queries Type	PMA	LPIPS↓	SSIM↑
Evaluate on 256×176 resolution				
B1	Local		0.161	0.724
B2	Local	Eq. (9)	0.154	0.735
B3	Local+Global		0.158	0.731
B4	Local+Global	Eq. (9)	0.151	0.741
Ours	Local+Global	Eq. (5)	0.150	0.747

Table 3. Qualitative comparison for ablation studies. This study investigates the effects of different learnable query types in QFFT and various masking mechanisms in PMA.

tributes. This improvement is mainly attributed to the Pose-Masked Attention mechanism, which guides attention towards human-centric regions, allowing the model to synthesize complex pose changes with greater precision.

Lastly, in scenarios with occlusions or partially missing information in the source image, our model demonstrates a stronger ability to infer plausible completions, resulting in more natural and coherent outputs. This is particularly evident in cases where parts of the body or clothing are not visible in the source image. Overall, these qualitative results highlight our method's ability to balance pose accuracy, identity preservation, and fine-grained appearance consistency better than existing alternatives.

4.4. User Study

The above quantitative and qualitative results demonstrate that the method we proposed has significant advantages. However, since the quality of PGPIS remains subjective, the evaluation results may vary between individuals. To address this, we conducted a user study using the Jab metric [1, 43] to further validate the effectiveness of our approach. In this study, participants were asked to select the image most similar to the ground truth from a set of candidates, based on three criteria: texture quality, texture preservation, and identity preservation. A total of 50 participants were recruited, each completing 25 questions. The results of the study are presented in Tab. 2. Compared to other methods, our approach achieves a score as high as 46.0%, surpassing the second-best method by 23.5%. This demonstrates that our method performs exceptionally well in preserving identity and texture based on objective metrics.

4.5. Ablation Study

We conduct ablation studies to evaluate the contributions of the Query-based Feature Fusion Transformer (QFFT) and

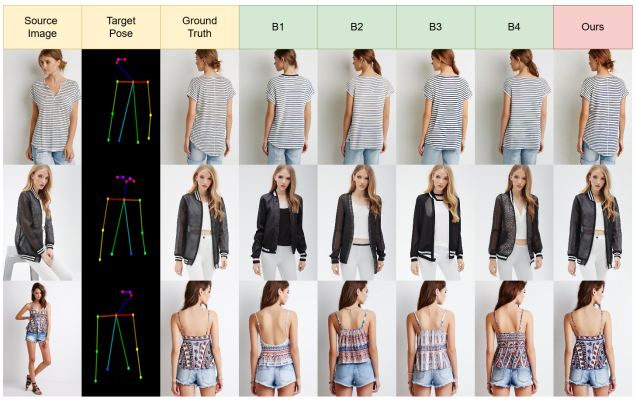

Figure 4. Qualitative ablation results. We evaluate the effectiveness of the learnable query and mask, with the corresponding settings detailed in Tab. 3. Our approach integrates both high-level semantic features and low-level visual cues to achieve more precise results.

the Pose-Masked Attention (PMA) module. The quantitative results are presented in Tab. 3, with corresponding qualitative examples in Fig. 4. As introduced in Sec. 3, QFFT employs both local and global learnable queries for multi-scale feature extraction. To isolate their effects, we systematically ablate the query types and observe the resulting changes in performance.

Additionally, we examine the impact of different masking strategies in PMA. Specifically, we propose a simplified attention mask defined by:

$$M_{ij} = \begin{cases} 0 & \text{if } m_p^{(i)} = 1, \\ -\infty & \text{otherwise.} \end{cases} \tag{9}$$

Tab.3 summarizes the effects of query types and masking strategies in PMA. We compare three configurations: disabling masking, applying the simplified mask in Eq. 9 to local queries, and applying the full mask in Eq. 5 to both local and global queries. Models B1 and B2 use only local queries, with B2 introducing the simplified mask, which already yields noticeable improvements in LPIPS and SSIM by guiding attention towards pose-relevant regions. B3 and B4 further incorporate global queries, with B4 applying the simplified mask to local queries. While global queries help, the gains are less substantial compared to the impact of masking, suggesting that local queries interacting with multi-scale features already encode partial global context. Our complete model, combining both query types with the full PMA masking strategy, achieves the best performance, demonstrating the synergy between query design and pose-aware attention for enhancing both perceptual quality and structural fidelity.

Figure 5. Qualitative results of Our Models for image appearance editing. The generated images preserve the identity and pose of the source image while drawing on the clothing style from the reference image. For each example, the first row shows the reference image, and the second row presents the corresponding generated image.

4.6. Image Appearance Editing

Our method leverages the flexibility and controllability inherent in diffusion models, enabling straightforward and effective appearance editing without the need for any additional training. Specifically, given a source image and a reference image and a binary mask m that indicating the region in the source image to be edited. Our approach seamlessly integrates information from both images during the sampling process. The noise prediction at each diffusion timestep is decomposed into

$$\epsilon'(t) = m \cdot \epsilon_t + (1 - m) \cdot z^{\text{src}},$$

where $y_z^{src} = \sqrt{\bar{a}_t} y^{src} + \sqrt{1 - \bar{a}_t}\epsilon$ as defined in Eq. (1). This formulation allows the model to selectively apply changes only within the masked region, while preserving the original content outside of it. As a result, our method generates realistic and detailed textures in the edited areas, closely aligning with the appearance cues from the reference image. At the same time, it maintains the overall visual integrity and coherence of the source image, ensuring a natural and seamless composition. The effectiveness of our approach is demonstrated through various editing examples, which are illustrated in Fig. 5.

5. Limitation and Future Work

Although the Query Enhancement Network can generate high-quality and realistic images on the DeepFashion dataset, its performance significantly declines when applied to in-the-wild datasets, which often contain complex backgrounds and previously unseen visual patterns. This performance gap underscores the model's limited generalization capability when exposed to data that deviates from the structured and relatively clean training distribution. To address this, our future research will focus on enhancing the model's adaptability and generative performance on more diverse and unconstrained datasets. In addition, we plan to explore the potential of extending this method to other challenging downstream tasks that commonly face performance bottlenecks due to insufficient or low-quality training data, such as person re-identification. We aim to improve data augmentation and cross-domain alignment in these contexts. Ultimately, our goal is to strengthen the model's generalization ability and robustness across diverse real-world applications, thereby providing more practical and impactful generative technology support for other domains.

6. Conclusion

In this paper, we propose a novel Query Enhancement Network designed for the Pose-Guided Person Image Synthesis (PGPIS) task, with a particular focus on improving the generation of fine-grained visual details. The core of our approach comprises two key components: the Query-based Feature Fusion Transformer (QFFT) and the Pose-Masked Attention (PMA) module. QFFT is designed to effectively fuse multi-resolution visual features by a query-driven mechanism, enabling the network to retain both global structure and subtle appearance cues. Meanwhile, PMA selectively attends to regions of interest based on pose guidance, allowing the model to better preserve identity information and spatial consistency during synthesis. Together, these modules form a cohesive framework that significantly enhances the quality of generated person images. Extensive experiments conducted on Deepfashion dataset demonstrate that our method achieves superior performance compared to existing state-of-the-art approaches, as measured by both qualitatively and quantitatively.

Acknowledgments

This work is partially supported by the National Science and Technology Council (NSTC), Taiwan (Grants: NSTC-112-2221-E-A49-059-MY3 and NSTC-112-2221-E-A49-094-MY3).

References

[1] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *CVPR*, 2023. 1, 2, 5, 6, 7

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 5

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[4] Patrick Esser and Ekaterina Sutter. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018. 1

[5] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2

[6] Beatriz Quintino Ferreira, Joao P. Costeira, Ricardo G. Sousa, Liang-Yan Gui, and Joao P. Gomes. Pose guided attention for multi-label fashion image classification. In *ICCV*, 2019. 1

[7] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *CVPR*, 2019. 1, 2

[8] Xiao Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, and Tao Xiang. Controllable person image synthesis with posec-onstrained latent diffusion. In *ICCV*, 2023. 1, 2, 5

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 4

[11] Yining Li, Chen Huang, , and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *CVPR*, 2019. 1, 2

[12] Jiaqi Liu, Jichao Zhang, Paolo Rota, and Nicu Sebe. Multi-focal conditioned latent diffusion for person image synthesis. In *CVPR*, 2025. 1, 2, 5, 6, 7

[13] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *ICCV*, 2019. 1

[14] Yuzhao Liu, Yuhan Liu, Shihui Xu, Kelvin Cheng, Soh Masuko, and Jiro Tanaka. Comparing vr-and ar-based try-on systems using personalized avatars. In *MDPI*, 2020. 1

[15] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 2, 5

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5

[17] Yanzuo Lu, Manlin Zhang, Andy J Ma1, Xiaohua Xie, and Jianhuang Lai. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In *CVPR*, 2024. 1, 5

[18] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *CVPR*, 2021. 1

[19] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *CVPR*, 2021. 5

[20] Liqian Ma, Xu Jia, Qianru Sun, T. Tuytelaars B. Schiele, and L. Gool. Pose guided person image generation. In *NeurIPS*, 2017. 1, 2

[21] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018. 1

[22] Liyuan Ma1, Tingwei Gao, Haitian Jiang, Haibin Shen, and Kejie Huang. Waveipt: Joint attention and flow alignment in the wavelet domain for pose transfer. In *ICCV*, 2023. 5

[23] Liyuan Ma1, Tingwei Gao, Haibin Shen, and Kejie Huang. Freqhpt: Frequency-aware attention and flow fusion for human pose transfer. In *CVPRW*, 2023. 5

[24] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *CVPR*, 2020. 5

[25] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *CVPR*, 2020. 1

[26] Chong Mou, Xintao Wang, Liangbin Xie, Jing Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 1

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and Gretchen. Learning transferable visual models from natural language supervision. In *PMLR*, 2021. 2

[29] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *CVPR*, 2020. 1, 2, 5

[30] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable. In *CVPR*, 2022. 1, 5

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Omme. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 5

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[33] Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Yang Wei. Advancing pose-guided image synthesis with progressive conditional diffusion models. In *ICLR*, 2024. 1, 2, 6, 7

[34] Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, , and Nicu Sebe. Deformable gans for pose-based human image generation. In *CVPR*, 2019. 1, 2

[35] Chang D. Yoo Trung X. Pham, Zhang Kang. Cross-view masked diffusion transformers for person image synthesis. In *ICML*, 2024. 1

[36] Junkun Yuan, Xinyu Zhang, Hao Zhou, Jian Wang, Zhongwei Qiu, Zhiyin Shao, Shaofeng Zhang, Kun Kuang Sifan Long, Kun Yao, Junyu Han, Errui Ding, Lanfen Lin, Fei Wu, and Jingdong Wang. Hap: Structure-aware masked image modeling for human-centric perception. In *NeurIPS*, 2024. 1

[37] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *CVPR*, 2021. 1, 2, 5

[38] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3

[39] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1

[40] Pengze Zhang, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *CVPR*, 2022. 1, 5, 6

[41] Tengteng Huang Zhen Zhu, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *CVPR*, 2019. 1

[42] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnetv2: Full-resolution correspondence learning for image translation. In *CVPR*, 2021. 5

[43] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. In *ECCV*, 2022. 7

[44] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *CVPR*, 2019. 5