WaveIPT: Joint Attention and Flow Alignment in the Wavelet domain for Pose Transfer

Liyuan Ma[1,2*] Tingwei Gao[1*] Haitian Jiang[2] Haibin Shen[2] Kejie Huang[2†]
[1]Alibaba Group
[2]Zhejiang University, China
{mlyarthur, jianghaitian, shen_hb, huangkejie}@zju.edu.cn, tingwei.gtw@alibaba-inc.com

Abstract

Human pose transfer aims to synthesis a new image of the source person in a target pose. Among the various existing methods, attention and flow have emerged as two of the most popular and effective approaches. Attention excels at preserving the semantic structure of the source image, which is more reflected in the low-frequency domain. Contrastively, flow is better at retaining fine-grained texture details in the high-frequency domain. To leverage the advantages of both attention and flow simultaneously, this paper proposes Wavelet-aware Image-based Pose Transfer (WaveIPT) as a novel approach to fuse the attention and flow in the wavelet domain. To improve the fusion effect and avoid interference from irrelevant information across different frequencies, WaveIPT first applies Intra-scale Local Correlation (ILC) to adaptively fuse attention and flow in the same scale according to their strengths in low and high-frequency domains. Subsequently, WaveIPT employs Inter-scale Feature Interaction (IFI) to explore inter-scale frequency features, facilitating effective information transfer across different scales. Furthermore, we introduce Progressive Flow Regularization (PFR), an effective method that alleviates the challenges of flow estimation under large pose differences. The experiments on the DeepFashion dataset demonstrate that WaveIPT achieves a new state-of-the-art in terms of both FID and LPIPS, with improvements of 4.97% and 3.89%, respectively.

1. Introduction

Human pose transfer refers to the task of transforming a given person into the target pose, which has extensive applications in movie editing, online shopping, virtual reality, etc. However, achieving both texture-preserving and realistic huaman pose transfer remains a challenge, especially when there is a significant pose difference or the person image contains complex texture.

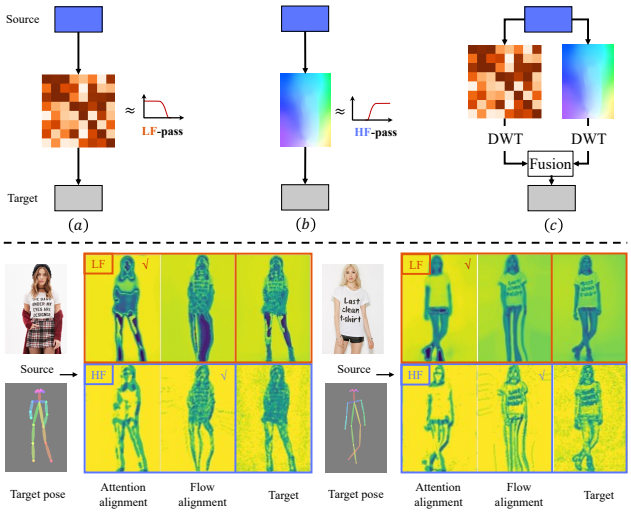

Figure 1. Illustrative comparison of (a) attention alignment-based [27, 43], (b) flow alignment-based [29, 13], and the proposed (c) wavelet-aware fusion method. The lower part shows the different frequency distributions of the source texture aligned by attention and flow. The low frequency (LF) semantic structure (the leg part) can be better recovered by attention, whereas the high frequency (HF) component from flow alignment is closer to the real target (the cloth detail).

Previous research has employed various methods to interact with the source texture and target pose, including style modulation [24, 42], deformable convolution [37], affine transformation [47], flow [29, 17, 13, 4, 49, 34, 48, 22], attention [43, 27, 33, 12, 14, 50, 44], etc. Among these methods, flow and attention have proven to be the most effective for texture alignment at the pixel or feature level.

Flow-based methods [48, 17, 34, 22] utilize appearance flows to deform and allocate the local texture information to the target positions based on the target pose. This approach is conducive to retaining high-frequency details, as

*Equal Contribution. † Corresponding author. This work was done when Liyuan Ma was an intern at Alibaba Group.

it gathers local texture information to calculate the output of target positions. However, due to the instability and difficulty of flow estimation under large pose differences, the reliability of image semantics cannot be guaranteed, especially when dealing with large pose differences, and thus leads to a deterioration in visual quality. Attention-based methods [43, 27, 14, 50] utilize various variants of cross-attention to aggregate source texture information through weighted summation. The global receptive field of the attention mechanism enables it to capture reasonable human semantic and low-frequency structure. However, it is observed that the aggregating process disrupts the relative spatial relationship, leading to a deficiency in local details in the generated images. Most prevailing methods [48, 34, 22, 43, 27, 14] solely rely on either flow or attention mechanisms for texture alignment without fully exploiting their complementary advantages. Ren *et al.* [28] propose a mask-based method to integrate flow and attention in the spatial domain. In their approach, the differences that exist in frequency domain are not taken into consideration, leading to a compromised fusion effect. Ma *et al.* [21] attempt to fuse flow and attention in the frequency domain. However, their method applies an implicitly fusion strategy, thereby failing to effectively exploit the complementarity in the frequency domain.

Our analysis, as depicted in Figure 1, highlights the differences in frequency distribution between attention and flow alignment, with flow being capable of retaining high-frequency details while attention better preserves low-frequency structures [1, 25, 32]. This observation has motivated us to propose WaveIPT to incorporate frequency information from both attention and flow adaptively. In contrast to previous wavelet-based approaches that primarily focus on improving feature representation by enhancing convolution or modified up-/down-sampling operations,our method is specially designed to address the feature fusion problem under uneven frequency distribution for human pose transfer. WaveIPT aligns the features using attention and flow mechanisms, followed by transforming the features into the wavelet domain using Discrete Wavelet Transform (DWT). The aligned features are then fused using Intra-scale Local Correlation (ILC) and Inter-scale Feature Interaction (IFI) modules. ILC supplements different frequency bands within specific scale by leveraging the advantages of flow and attention in preserving high- and low-frequency information. Additionally, the IFI mechanism is devised to promote efficient transmission between scales by adaptively updating the current-scale features onto the previous-scale features. Furthermore, convolutions with larger dilated rates are utilized to process sparse low-frequency features, leading to a more comprehensive and contextually rich feature representation.

We introduce Progressive Flow Regularization (PFR), which leverages an intermediate pose to narrow the gap between source and target poses during flow estimation training. By approximating the flow from the source to the target through the source-to-intermediate and intermediate-to-target flows, we are able to effectively reduce the pose difference and enhance the performance of flow estimation.

The main contributions of this paper are as follows:

- This paper proposes WaveIPT, a novel method that combines attention and flow with dedicated designed ILC and IFI. This approach is capable of effectively aligning the source texture in the wavelet domain and enables realistic pose transfer results.
- This work introduces PFR as an effective approach to address the challenge of flow estimation under large pose variation.
- The proposed WaveIPT achieves state-of-the-art results on the DeepFashion dataset, demonstrating the effectiveness of our method.

2. Related Works

2.1. Wavelet-related Methods

Wavelet domain analysis has emerged as an effective method in the computer vision field, including high-level discriminant tasks [39], image restoration task [15, 52, 16, 3] and image generation task [41, 38, 40, 38, 26]. [39] combines the wavelet transformation and self-attention into transformer designs to prevent information dropping in the conventional down-sampling operations. Many image restoration works explore reconstructing the finer details from a frequency decomposition perspective. Typically, SDWNet [52] proposed a wavelet reconstruction to better recover clear high-frequency details that are crucial for image deblurring. Moreover, [38] facilitated the generation both low- and high-frequency information by performing skip connections for different frequency elements.

However, no prior work has focused on applying wavelet domain analysis to pose transfer. Our paper proposes a novel wavelet-aware fusion module that enables locally realistic and globally reasonable image synthesis. Compared to other wavelet-based methods, our approach focuses on leveraging wavelet transformation to enhance texture alignment and improve pose transfer performance. We believe that our proposed method can contribute to the advancement of wavelet domain analysis in the field of pose transfer.

2.2. Pose Transfer

With the rapid development of image generation technology, human pose transfer has developed for several years. This task was first proposed by [23] to deal with the person image synthesis in different poses. Several works [42, 24]

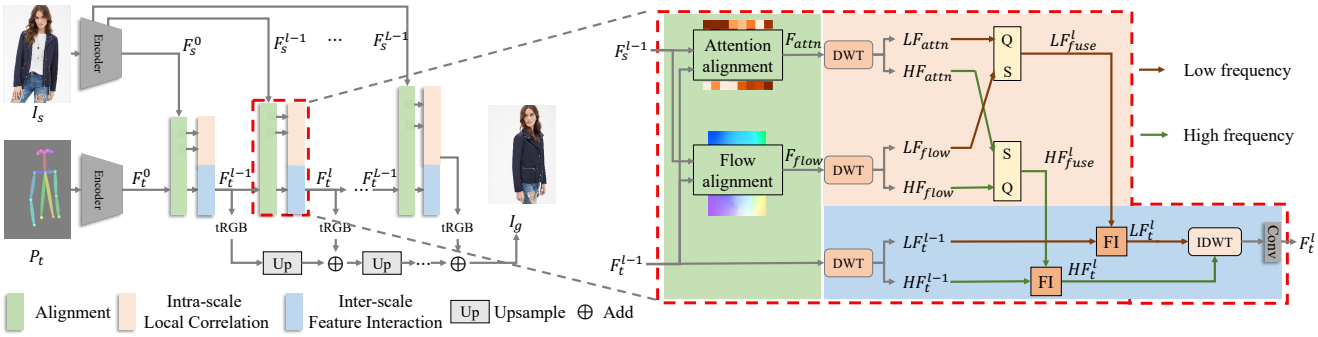

Figure 2. The framework of the proposed WaveIPT network. The source human image I_s and target pose P_t are encoded to extract L-levels source features $\{F_s^l\}_{l=0,\dots,L-1}$ and target feature vector F_t^0. The source features are deformed through Attention and Flow alignment modules. The aligned features are further decomposed into Low Frequency (LF) and High Frequency (HF) components with Discrete Wavelet Transform (DWT). Then Intra-scale Local Correlation (ILC) module fuses low- and high-frequency features adaptively within the scale and Inter-scale Feature Interaction (IFI) integrates frequency features across different scales. The frequency features are converted back to the spatial domain via Inverse Discrete Wavelet Transform (IDWT) and used to synthesize the final output image I_g. The detailed illustration of ILC and IFI can be found in Figure 3.

borrowed the idea from style transfer, which extracts style information from the specific semantic areas of person image and redistributes it into the corresponding semantic region. However, such methods commonly suffer from losing existent details in the original person image. Deformable convolution is utilized in [37] to enhance spatial alignment, but it has a restricted sampling offset that only allows for modeling subtle motion relationships. To explicitly model the geometry deformation along with pose change, flow-based approaches [29, 17, 13, 4, 49, 34, 48, 22, 36] were proposed to explore the spatial mapping relationship. [29, 22] estimated flow in an unsupervised manner. [17, 13] adopted the 3D human mesh model to acquire the 3D flow with vertices mapping. The flow-based models are capable of generating realistic texture, however, these methods fail to extract precise motion, resulting in obvious artifacts. Compared with the flow deformation method, the attention-guided methods [43, 27, 33, 12, 14, 50, 44] indicate their capability in rendering satisfactory structures and semantics undergoing dramatic pose changes. Although [28] confirms the complementary advantages between the flow and attention to generate accurate human semantics and realistic textures, it fuses the flow and attention-warped features in the spatial domain and features are mixed without distinction for different frequencies, which leads to suboptimal fusion effect. [21] employs adaptive masks to implicitly merge features of differing frequencies in flow and attention, inadvertently failling to fully leverage the complementary advantages across distinct frequency bands. In contrast, the approach presented in this paper makes explicit use of the inherent strengths demonstrated by flow and attention mechanisms in high and low frequency in the wavelet domain, which improves the quality of person image generation effectively.

3. Proposed Method

Preliminary. We first revisit Flow and Attention alignment operations which are commonly used to reassemble textures in pose transfer task and then review the basic concepts of Discrete Wavelet Transform.

Flow alignment spatially deforms the source appearance according to the target pose by calculating a point-wise 2D deformation field, which correlates the target position with local source candidates. We follow [8, 6, 5] to estimate the appearance flow in a coarse-to-fine manner. Formally, given the encoded features F_s and F_t from the source human image and target pose, the source feature aligned by flow is calculated as follows:

$$\begin{aligned} F_{flow} &= GS(FE(F_s, F_t, J_s, J_t), F_s) \\ &= GS(W_{s\to t}, F_s), \end{aligned} \tag{1}$$

where $W_{s\to t}$ is the estimated appearance flow, $GS(f, y)$ denotes the grid sampling function [10] that deforms the y with flow f, and FE means the flow estimation network. J_s and J_t represent source and target joint coordinates.

Attention alignment can be regarded as the dense correlation matrix, which predicts each position with the weighted summation of the whole source values. The double attention [2, 27, 46] is chosen as the attention alignment calculation backbone for its efficient implementation, which avoids the quadratic complexity by splitting the attentive operation into gathering and distribution. Specifically, the attention alignment operation warps the source feature F_s with the attention matrix $\mathcal{A}_{s\to t} \in \mathbb{R}^{hw\times hw}$ as follows.

$$\begin{aligned} F_{attn} &= \mathcal{A}_{s\to t}F_s = \mathcal{A}_d\mathcal{A}_g F_s \\ &= [\text{SoftMax}(\mathcal{T}_d F_t)]^T[\text{SoftMax}(\mathcal{T}_g F_s)]F_s \end{aligned} \tag{2}$$

where $\mathcal{T}_d \in R^{k\times c}$ and $\mathcal{T}_g \in R^{k\times c}$ are the corresponding convolution filters to calculate the gathering and distribution

matrixes $\mathcal{A}_d \in R^{hw\times k}$ and $\mathcal{A}_g \in R^{k\times hw}$. k is the number of convolution channels.

Discrete Wavelet Transform is capable of decomposing spatial features into multi-level wavelet sub-bands. We utilize the Haar wavelet in our implementation. Specifically, the Haar wavelet filter with low-pass filter $(1/\sqrt{2}, -1/\sqrt{2})$ and high-pass filter $(1/\sqrt{2}, 1/\sqrt{2})$ is applied in DWT to extract wavelet frequency components *LL*, *LH*, *HL*, *HH*. We use LF and HF to represent the low-frequency component *LL* and high-frequency components *LL*, *LH*, *HL*, *HH*.

3.1. Framework

The framework of WaveIPT is shown in 2. Given the source human image I_s and target pose P_t, WaveIPT extracts source and target features $\{F_s^l\}_{l=0,\dots,L-1}$ and F_t^0 and reassemble the source texture by attention and flow alignment. To fuse aligned features in different frequency bands, WaveIPT transforms the aligned source and target features into wavelet domain, where Intra-scale Local Correlation is designed to fuse frequency features adaptively in the same scale (see Sec. 3.2) and Inter-scale Feature Interaction is adopted to facilitate the information interaction across scales (see Sec. 3.3). Then, the frequency features are converted back into the spatial domain via Inverse Discrete Wavelet Transform (IDWT). The whole process can be trained end-to-end. Moreover, the Progressive Flow Regularization (see Sec. 3.4) is introduced to promote the training of flow. After aligning and fusing source features processed by flow and attention, we are capable of synthesizing the target images by hierarchically predicting RGB images. The final target image I_g is obtained by summing up all the RGB images at different levels.

3.2. Intra-scale Local Correlation

Owing to the unique strengths of attention and flow in low and high frequencies, it is reasonable for us to improve frequency feature representation by augmenting low-frequency and high-frequency features from attention and flow with inferior frequency information. Therefore, Local Correlation is designed to utilize supplementary input (S) to compensate for query (Q) input. For example, LF_{flow} can act as the S to augment LF_{attn} (Q) in the low-frequency band, while HF_{attn} (S) is used to enhance the high-frequency details in HF_{flow} (Q). Besides, we note that the corresponding features between LF_{flow}/HF_{flow} and LF_{attn}/HF_{attn} after the spatial alignment by flow and attention are now located within the same local region. Therefore, we correlate merely local areas to avoid interference caused by large receptive fields.

As shown in Figure 3, given the query input $Q \in \mathbb{R}^{hw\times c}$ as the superior input and the supplement information $S \in \mathbb{R}^{hw\times c}$ as the inferior one, whose rows are c-dimensional vectors and hw is the multiplication of feature map height h and width w. The Local Correlation weight $\mathbf{A}_i(k)$ and reshaped supplement input $\mathbf{S}_i(k)$ are calculated as follows.

$$\mathbf{A}_i(k) = \begin{bmatrix} Q^i {S^1}^T \\ \vdots \\ Q^i {S^j}^T \\ \vdots \\ Q^i {S^k}^T \end{bmatrix}^T \qquad (3)$$
$$\mathbf{S}_i(k) = \begin{bmatrix} {S^1}^T & \dots & {S^j}^T & \dots & {S^k}^T \end{bmatrix},$$

where $j \in \mathcal{N}_i$ denotes the spatial location in the neighborhood of i. We define the $\mathcal{N}_i$ as a corresponding square area with a hyperparameter radius of 3 which contains k neighboring values. Then the Local Correlation output value for the i-th token is then calculated as:

$$LC_i(k) = \text{SoftMax}(\mathbf{A}_i(k))\mathbf{S}_i(k)^T \qquad (4)$$

The ILC operation is performed repeatedly by traversing all locations in the feature map. Then the intra-scale fusion of frequency features is conducted as follows.

$$\begin{aligned} LF_{fuse}^l &= LC(Q = LF_{attn}, S = LF_{flow}) \\ HF_{fuse}^l &= LC(Q = HF_{flow}, S = HF_{attn}) \end{aligned} \qquad (5)$$

3.3. Inter-scale Feature Interaction

Upon obtaining the fused frequency features at the current scale LF_{fuse}^l/HF_{fuse}^l, the Feature Interaction module aims to investigate inter-scale relationship between features across different scales, thereby necessitating the extraction of more informative content from the current scale to enhance the previous scale features LF_t^{l-1}/HF_t^{l-1}.

To this end, we first compute the difference between them and utilize the Sigmoid function to regress the spatial weights. The spatial weights are capable of selecting the most distinguished and informative content by multiplication with current scale features. Subsequently, The selected information is then added to the previous scale input to the enhanced output. Considering the sparsity of low-frequency features, we adopt convolution layers with a larger dilated rate for low-frequency features to extract more effective contextual information. Formally, the Feature Interaction process is conducted as follows

$$\begin{aligned} LF_t^l =& \text{Conv}_l(LF_t^{l-1} + LF_{fuse}^l \times \text{Sigmoid} \\ & (LF_{fuse}^l - LF_t^{l-1})) \\ HF_t^l =& \text{Conv}_h(HF_t^{l-1} + HF_{fuse}^l \times \text{Sigmoid} \\ & (HF_{fuse}^l - HF_t^{l-1})), \end{aligned} \qquad (6)$$

where Conv_l and Conv_h denote the convolution layers with dilated rates of 3 and 1. Further detailed descriptions can be found in Figure 3.

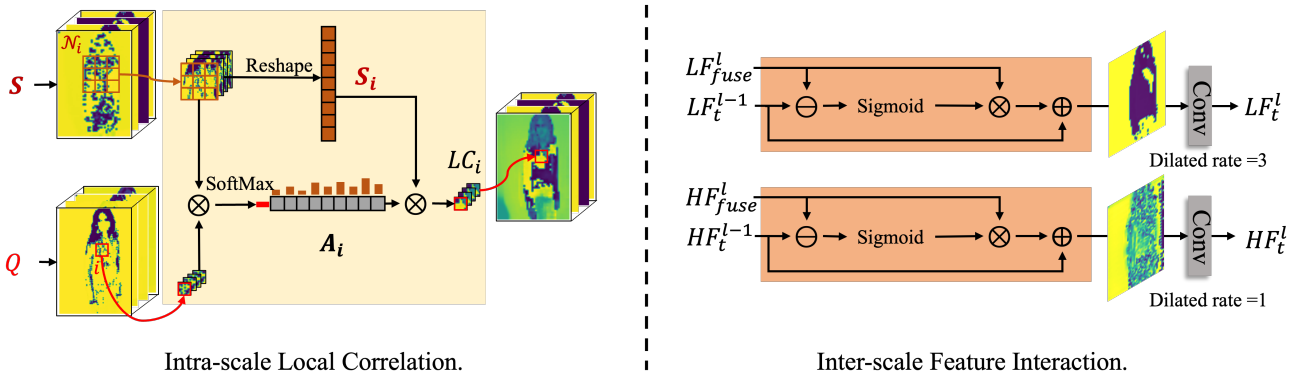

Figure 3. The illustration of Intra-scale Local Correlation and Inter-scale Feature Interaction modules. We use Q and S to represent that query input Q is augmented by supplementary input S.

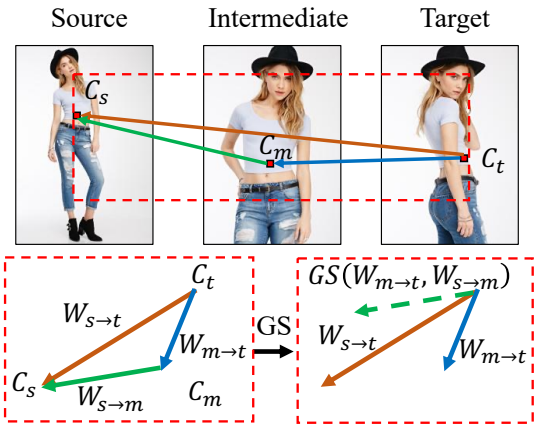

Figure 4. Illustration of Progressive Flow Regularization (PFR). The abdominal point in images under source, intermediate, and target poses is presented for example, which moves among different poses. We use orange, green, and blue arrows to denote the $W_{s\to t}$, $W_{s\to m}$, and $W_{m\to t}$ flows, respectively. We use the green dotted arrow to represent that $W_{s\to m}$ is warped by $W_{m\to t}$ with Grid Sample (GS) function.

3.4. Progressive Flow Regularization

The performance of the flow estimation suffers when the source and target poses are significantly different. To alleviate the difficulty of flow estimation brought by the pose discrepancy, we propose Progressive Flow Regularization, which employs the intermediate pose to bridge the gap between the source and target poses during training. For example, the flow $W_{s\to t}$ between the front and back poses can be constrained by the combination of front-to-side $W_{s\to m}$ and side-to-back $W_{m\to t}$ flows, which utilize the side pose as the intermediate pose to promote the flow estimation.

Specifically, we use flow estimation network to predict the $W_{s\to t}$, $W_{s\to m}$, and $W_{m\to t}$ with source pose P_s, intermediate pose P_m, and target pose P_t. Then we utilize progressive flows $W_{s\to m}$ and $W_{m\to t}$ to approximate $W_{s\to t}$. For a specific feature, we denote its feature position under source, intermediate, and target poses as $C_s = (x_s, y_s)$, $C_m = (x_m, y_m)$ and $C_t = (x_t, y_t)$. The flow vector at these positions can be calculated as follows.

$$\begin{aligned} W_{s\to t}(x_t, y_t) &= (x_s - x_t, y_s - y_t) \\ W_{s\to m}(x_m, y_m) &= (x_s - x_m, y_s - y_m) \\ W_{m\to t}(x_t, y_t) &= (x_m - x_t, y_m - y_t) \end{aligned} \tag{7}$$

When all the flows are predicted correctly, the summation of $W_{s\to m}(x_m, y_m)$ and $W_{m\to t}(x_t, y_t)$ should be equal to the $W_{s\to t}(x_t, y_t)$, which supplies a reasonable constraint for consistent flow estimation among various poses. As shown in Figure 4, to enable the calculation of the flow constraint, which requires that $W_{s\to m}(x_m, y_m)$ and $W_{m\to t}(x_t, y_t)$ are defined in the same coordinate system, we utilize Grid Sample function to resample $W_{s\to m}(x_m, y_m)$ at C_m to the same position C_t by $W_{m\to t}(x_t, y_t)$. Then Progressive Flow Regularization loss can be generalized into all positions and defined as follows.

$$\mathcal{L}_{prog} = ||GS(W_{m\to t}, W_{s\to m}) + W_{m\to t} - W_{s\to t}||_2. \tag{8}$$

It is worth noting that during the early stages of training, the flow estimation performance may be insufficient, which can lead to uncertainty regarding the effectiveness of the regularization. Thus we introduce the progressive flow loss into training after several training epochs.

3.5. Loss Functions

Except for the aforementioned progressive flow regularization, this paper further applies several loss functions to train the whole network in an end-to-end manner.

Alignment Loss $\mathcal{L}_{align}$. The alignment loss constrains the network for estimating accurate flow map and attention matrix for deformation. We apply l_1 loss to penalize the difference between downsampled source image $I_s^{\downarrow}$ and corre-

Figure 5. Qualitative comparison results with several state-of-the-art methods on DeepFashion dataset of 256×176 and 512×352 resolutions. Please zoom in for a better view.

sponding target image $I_t{}^{\downarrow}$.

$$\begin{aligned}\mathcal{L}_{align} =& ||\mathcal{S}(W_{s\to t}, I_s{}^{\downarrow}) - I_t{}^{\downarrow}||_1 + \sum_i ||\nabla(W_{s\to t})_i|| \\ &+ ||\mathcal{A}_{s\to t} I_s{}^{\downarrow} - I_t{}^{\downarrow}||_1\end{aligned} \tag{9}$$

where $||\nabla(W_{s\to t})_i||$ represents the generalized charbonnier loss function [31], which is used to ensure the smoothness of the flow.

Perceptual Loss $\mathcal{L}_{perc}$. The perceptual loss [11] encourages the generated image to be perceptually consistent with the target image at the feature level:

$$\mathcal{L}_{perc} = \sum_i ||\phi_i(I_g) - \phi_i(I_t)||_1, \tag{10}$$

where ϕ_i denotes the ith layer of the pre-trained VGG [30] network and I_g denotes the generated image.

Adversarial Loss $\mathcal{L}_{adv}$. We adopt adversarial loss to shorten the distribution distance between generated result and the real images, which thus promotes the high fidelity of the generated images.

$$\begin{aligned}\mathcal{L}_{adv} =& \mathbb{E}_{I_s, I_t, P_t}[\log(1 - D(G(I_s, P_t), P_t))] \\ &+ \mathbb{E}_{I_t, P_t}[\log(D(I_t, P_t))],\end{aligned} \tag{11}$$

where G and D represent the image generator and discriminator, respectively.

SSIM Loss $\mathcal{L}_{ssim}$. The ssim loss is employed to improve the pixel similarity between the generated images and the ground truth images:

$$\mathcal{L}_{ssim} = ||\text{SSIM}(I_g) - \text{SSIM}(I_t)||_1. \tag{12}$$

Thus the overall objective function is defined as $\mathcal{L}_{total} = \lambda_{align}\mathcal{L}_{align} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{ssim}\mathcal{L}_{ssim} + \lambda_{prog}\mathcal{L}_{prog}$, where λ_{align}, λ_{perc}, λ_{adv}, λ_{ssim} and λ_{prog} are empirically set at 15, 2, 1.5, 10 and 10, respectively.

4. Experiments

4.1. Experimental Settings

Datasets settings. Following the previous related work, we conduct experiments on the DeepFashion dataset [18]. The original resolution of DeepFashion is 1101×750, and most previous methods used the resolution of 256×172 resolution [43, 50, 20, 42, 29, 24] and 512×352 [27, 51].To ensure a fair comparison, we follow the training and evaluation setting of [43] for the 256×176 training data. For the 512×352 dataset, the data setting is aligned with the state-of-the-art [27].

Performance metrics. We evaluate the model performance with several widely used metrics, including 1) *Structure Similarity Index* (SSIM) [35] which reports the pixel-level similarity; 2) *Fréchet Inception Distance* (FID) [9] which measures the distribution distance between the synthesized and real images, reflecting the realism of the generated images; 3) *Learned Perceptual Image Patch Similarity* (LPIPS) [45] which calculates the perceptual distance between the synthetic and real images with pre-trained VGG

and AlexNet. Due to the superior network capacity, VGG provides a more precise evaluation compared to AlexNet; 4) *Reid Score* [43] that adopts the re-identification model [7] to test whether the generated query image can be matched with the corresponding gallery real image. The higher Top-k represents that the generation results better maintain the source appearance and thus can be identified effectively.

Implementation details. Our experiment is conducted with Nvidia A100 GPU and Pytorch framework. We adopt the Adam optimizer [19] and the learning rates for 256×176 and 512×352 are 2×10^{-3} and 1×10^{-3}, respectively. The learning rate remains unchanged throughout the training process. The whole training process takes 200 epochs and the batch size is set to 32.

4.2. Comparison with state-of-the-art Methods

4.2.1 Quantitative Comparison

In Table 1, we compare our proposed method with state-of-the-art models on the DeepFashion dataset. Our model outperforms other models in all metrics for the resolution of 256×176, demonstrating its significant advantage in generating high-quality images for low resolutions. For 512×352 resolution, our model has also reached a state-of-the-art level, while achieving comparable results to the best models in terms of LPIPS. This indicates that our model can also address high-quality image generation at high resolutions. It is worth noting that our model has a significant lead over other models in terms of FID, demonstrating that it can generate more realistic results. Additionally, we used Reid Score to evaluate texture consistency between generated and real images, and our method outperformed other competitors in terms of Topk scores, indicating effective texture preservation and reliability.

User Study. In order to evaluate the performance of the generated results from a human subjective perspective, we also perform the user study by randomly sampling 1,500 generated images for comparison, which largely exceeds the number of samples used in other methods (e.g. 33 samples for CASD [50] and 55 samples for SPGNet [20]). The sampled images are then evaluated by 57 professional and experienced employees, who assess their realism and consistency. The Topk-1 and Topk-3 mean that the result of the corresponding method is selected as the most realistic one or three among all others. Our model achieves the highest Topk-1 and Topk-3 scores in the user study, indicating that it generated more realistic and consistent images than other methods. The results demonstrate that our model outperforms existing methods in terms of generating high-quality images that satisfy human subjective perceptions.

Figure 6. Qualitative result of our ablation study.

4.2.2 Qualitative Comparison

We show the qualitative comparison results on Figure 5 including 256×176 and 512×352 resolutions. Among these competitors, the style modulation-based method ADGAN [24] fails to model complex spatial texture distribution, which loses the spatial information in style vector extraction. The flow-based method GFLA [29] has decreased performance when confronting large pose discrepancy since it fails to deform texture reasonably (see 2nd and 3rd rows in the left of Figure 5). CASD [50], DPTN [43], NTED [27] and CoCosNet2 [51] introduce attention mechanism to reassemble the source person's texture according to the target pose. However, the attention operation tends to smooth the detailed texture and hinder the reoccurrence of the specific image pattern (see the logos and dress pattern in the 2nd row of Figure 5, etc.). In contrast, our method takes the advantage of both attention and flow to generate plausible human texture under large pose change (see 4th and 5th rows in the left of Figure 5) and retain the texture details faithfully (see the garment preservation result in the right of Figure 5).

4.3. Ablation Study

Several ablation studies over the DeepFashion dataset are performed to evaluate the efficacy of ILC, IFI, and PFR. We train multiple variant models with different configurations. All the variants models share the same basic architecture of encoder-decoder networks, but differ in the way

		SSIM↑	FID↓	LPIPS↓		Reid Score (%)↑		
				AlexNet	VGG	Topk-1	Topk-5	Topk-10
256 × 176	ADGAN [24]	0.6721	14.4580	0.2283	0.2557	81.46	91.97	95.65
	GFLA [29]	0.7677	10.8429	0.2258	0.2765	90.84	96.64	98.11
	PISE [42]	0.7682	11.5144	0.2080	0.2498	90.09	96.35	98.02
	SPGNEet [20]	0.7758	12.7027	0.2102	0.2443	94.43	98.23	99.04
	CASD [50]	0.7248	11.3732	0.2157	0.2645	93.09	98.35	99.12
	NTED [27]	0.7715	9.2876	0.2019	0.2564	97.34	99.39	99.74
	DPTN [43]	0.7782	11.4664	0.1957	0.2459	97.69	99.35	99.63
	FreqHPT [21]	0.7800	8.9072	0.1977	0.2369	98.72	99.65	99.83
	Ours	**0.7801**	**8.8259**	**0.1955**	**0.2348**	**99.05**	**99.96**	**99.99**
512 × 352	CocosNet2 [51]	0.7236	13.3250	0.2265	0.2735	87.84	91.71	94.72
	NTED [27]	0.7376	7.7821	**0.1980**	0.2472	98.41	99.24	99.71
	FreqHPT [21]	**0.7456**	6.5522	0.2026	0.2471	98.48	**99.93**	**99.93**
	Ours	0.7416	**4.8201**	0.1983	**0.2423**	**99.00**	99.56	99.81

Table 1. Quantitative comparison results with several state-of-the-art methods on DeepFashion dataset of 256 × 176 and 512 × 352 resolutions. The best and the second-ranked results are **bold** and underlined, respectively

	SSIM↑	FID↓	LPIPS ↓		Reid Score(%)↑		
			AlexNet	VGG	Topk-1	Topk-5	Topk-10
Non-wavelet Fusion	0.7781	9.1026	0.2023	0.2412	98.32	98.92	98.93
Vanilla Wavelet Fusion	0.7785	9.0264	0.2012	0.2391	98.83	99.25	99.83
Wavelet Fusion with ILC	0.7799	8.8861	0.1967	0.2356	98.93	99.24	99.87
Wavelet Fusion with ILC+IFI	0.7800	8.9797	0.1961	**0.2347**	98.76	99.89	99.94
Wavelet Fusion with ILC+IFI +PFR (Ours)	**0.7801**	**8.8259**	**0.1955**	0.2348	**99.05**	**99.96**	**99.99**

Table 2. Ablation study of our method on DeepFashion dataset.

	ADGAN	GFLA	PISE	SPGNet	CASD	NTED	DPTN	**Ours**
Topk-1(%)↑	9.7	10.8	3.7	3.6	23.2	12.1	9.7	**27.2**
Topk-3(%)↑	21.0	38.8	21.6	21.6	45.0	59.1	24.8	**68.2**

Table 3. User study of our method on DeepFashion dataset.

they fuse and process features. The details of these variant models are presented below:

- Non-wavelet Fusion. This model follows the spatial fusion strategy proposed in [28], which predicts the mask in the spatial domain to fuse features from attention and flow.
- Vanilla Wavelet Fusion. This model utilizes the mask to fuse low and high-frequency features separately in the wavelet domain. Unlike Non-wavelet Fusion, it predicts masks in the wavelet domain to fuse features located in different wavelet frequency bands, rather than directly fusing spatial features.
- Wavelet Fusion with ILC. In contrast to the mask-based fusion strategy employed by Non-wavelet Fusion and Vanilla Wavelet Fusion, we use ILC module to integrate frequency features from attention and flow.
- Wavelet Fusion with ILC+IFI. We further incorporate IFI to improve the fusion process. This variant model is trained without PFR.
- Wavelet Fusion with ILC+IFI+PFR (Ours). We use PFR in the training, making it our full model.

From the quantitative result in Table 2, our model outperforms other variant models in most of the metrics. By processing features in the wavelet domain, Vanilla Wavelet Fusion model can generate more realistic human images and achieves better FID and LPIPS scores than Non-wavelet Fusion strategy, which is caused by the ignorance of feature differences in the spatial fusion process. However, rely-

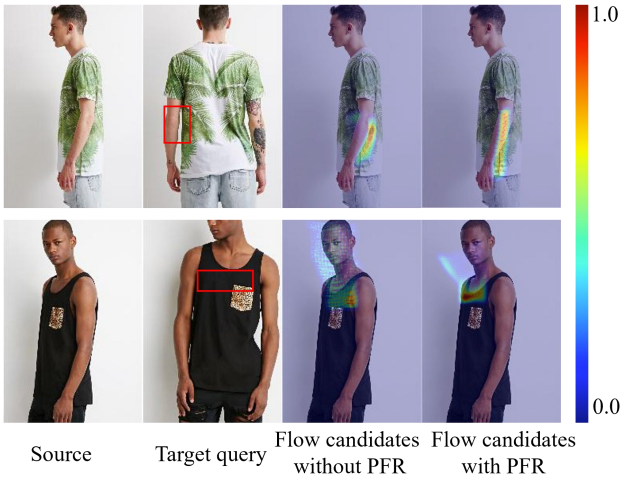

Figure 7. Visualization comparison of different flow estimation results including with (Wavelet Fusion ILC+IFI+PFR) and without PFR (Wavelet Fusion ILC+IFI) scenarios. The red rectangles denote the target query points which correlate the source candidates with the flow.

ing solely on the mask to fuse features limits its ability to capture contextual information. Our designed ILC module more effectively captures local contextual information during the fusion process, resulting in further improvement in metrics. The IFI module facilitates the transmission of frequency features from coarse to fine across different scales, which promotes the recovery of texture features and thus leads to an improvement in the generated quality.

The qualitative results are presented in Figure 6. We can see that compared with other variants, our full model obviously outperforms them in learning accurate spatial transformation and generating photo-realistic appearance images. Especially, PFR improves the reliability of the deformation between varying poses, and the body texture is better aligned as shown in the 1st and 2nd rows of Figure 6. Besides, the flow visualization results in Figure 7 corroborate our claim that progressive flow regularization can promote accurate spatial transformation in pose transfer (the target query area attains source candidates with more precise locations). More analysis about ILC and dilated rate is provided in supplementary materials.

5. Conclusion

In this paper, we present a novel wavelet-aware attention and flow fusion framework for human pose transfer. Observing the complementarity and difference between attention and flow in frequency domain, we propose WaveIPT to fuse and refine features warped by the attention and flow in a wavelet-aware manner, which effectively fuses the features from different frequency bands. Extensive experiments demonstrate the effectiveness of our proposed method compared to other state-of-the-art approaches.

Acknowledgements.
This work was supported by National Natural Science Foundation of China (Grant No. 62274142) and Hangzhou Major Technology Innovation Project of Artificial Intelligence (Grant No. 2022AIZD0060).

References

[1] Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. *arXiv preprint arXiv:2006.13318*, 2020. 2

[2] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A2-nets: Double attention networks. In *Neural Information Processing Systems*, 2018. 3

[3] Linhui Dai, Xiaohong Liu, Chengqi Li, and Jun Chen. Awnet: Attentive wavelet network for image isp. In *ECCV Workshops*, 2020. 2

[4] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. *ArXiv*, abs/1810.11610, 2018. 1, 3

[5] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8481–8489, 2021. 3

[6] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10471–10480, 2019. 3

[7] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *ArXiv*, abs/2006.02631, 2020. 7

[8] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3460–3469, 2022. 3

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 6

[10] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*, 2015. 3

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 6

[12] Kun Li, Jinsong Zhang, Yebin Liu, Yu-Kun Lai, and Qionghai Dai. Pona: Pose-guided non-local attention for human pose transfer. *IEEE Transactions on Image Processing*, 29:9584–9599, 2020. 1, 3

[13] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019. 1, 3

[14] Hongyu Liu, Xintong Han, Chengbin Jin, Lihui Qian, Huawei Wei, Zhe L. Lin, Faqiang Wang, Haoye Dong, Yibing Song, Jia Xu, and Qifeng Chen. Human motionformer: Transferring human motions with vision transformers. *ArXiv*, abs/2302.11306, 2023. 1, 2, 3

[15] Lin Liu, Jianzhuang Liu, Shanxin Yuan, Gregory G. Slabaugh, Ale*vs*. Leonardis, Wen gang Zhou, and Qi Tian. Wavelet-based dual-branch network for image demoireing. *ArXiv*, abs/2007.07173, 2020. 2

[16] Pengju Liu, Hongzhi Zhang, K. Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 886–88609, 2018. 2

[17] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5904–5913, 2019. 1, 3

[18] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 6

[19] Jonathan L Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? *Advances in neural information processing systems*, 27:1601–1609, 2014. 7

[20] Zheng Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10801–10810, 2021. 6, 7, 8

[21] Liyuan Ma, Tingwei Gao, Haibin Shen, and Kejie Huang. Freqhpt: Frequency-aware attention and flow fusion for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3490–3495, 2023. 2, 3, 8

[22] Liyuan Ma, Kejie Huang, Dongxu Wei, Zhao-Yan Ming, and Haibin Shen. Fda-gan: Flow-based dual attention gan for human pose transfer. *IEEE Transactions on Multimedia*, pages 1–1, 2021. 1, 2, 3

[23] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *arXiv preprint arXiv:1705.09368*, 2017. 2

[24] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020. 1, 2, 6, 7, 8

[25] Hoang Nt and Takanori Maehara. Revisiting graph neural networks: All we have is low-pass filters. *arXiv preprint arXiv:1905.09550*, 2019. 2

[26] Jorge Núñez, Xavier Otazu, Octavi Fors, Albert Prades, Vicenç Palà, and Román Arbiol. Multiresolution-based image fusion with additive wavelet decomposition. *IEEE Trans. Geosci. Remote. Sens.*, 37:1204–1211, 1999. 2

[27] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H. Li. Neural texture extraction and distribution for controllable person image synthesis. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13525–13534, 2022. 1, 2, 3, 6, 7, 8

[28] Yurui Ren, Yubo Wu, Thomas H Li, Shan Liu, and Ge Li. Combining attention with flow for person image synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3737–3745, 2021. 2, 3, 8

[29] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020. 1, 3, 6, 7, 8

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015. 6

[31] Deqing Sun, Stefan Roth, and Michael J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106:115–137, 2013. 6

[32] Chuanming Tang, Xiao Wang, Yuanchao Bai, Zhe Wu, Jianlin Zhang, and Yongmei Huang. Learning spatial-frequency transformer for visual object tracking. *arXiv preprint arXiv:2208.08829*, 2022. 2

[33] Hao Tang, Song Bai, Li Zhang, Philip H. S. Torr, and N. Sebe. Xinggan for person image generation. *ArXiv*, abs/2007.09278, 2020. 1, 3

[34] Jilin Tang, Yi Yuan, Tianjia Shao, Yong Liu, Mengmeng Wang, and Kun Zhou. Structure-aware person image generation with pose decomposition and semantic correlation. In *AAAI Conference on Artificial Intelligence*, 2021. 1, 2, 3

[35] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[36] Dongxu Wei, Kejie Huang, Liyuan Ma, Jiashen Hua, Baisheng Lai, and Haibin Shen. Oaw-gan: occlusion-aware warping gan for unified human video synthesis. *Applied Intelligence*, 53(1):616–633, 2023. 3

[37] Dongxu Wei, Xiaowei Xu, Haibin Shen, and Kejie Huang. C2f-fwn: Coarse-to-fine flow warping network for spatial-temporal consistent motion transfer. In *AAAI*, 2021. 1, 3

[38] Mengping Yang, Zhe Wang, Ziqiu Chi, and Wenyi Feng. Wavegan: Frequency-aware gan for high-fidelity few-shot image generation. In *European Conference on Computer Vision*, 2022. 2

[39] Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *European Conference on Computer Vision*, 2022. 2

[40] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9035–9044, 2019. 2

[41] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. Wavefill:

A wavelet-based generation network for image inpainting. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14094–14103, 2021. 2

[42] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. *arXiv preprint arXiv:2103.04023*, 2021. 1, 2, 6, 8

[43] Peng Zhang, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7703–7712, 2022. 1, 2, 3, 6, 7, 8

[44] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 1, 3

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[46] Yulun Zhang, Kai Li, Kunpeng Li, and Yun Fu. Mr image super-resolution with squeeze and excitation reasoning attention network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13425–13434, 2021. 3

[47] Zhimeng Zhang and Yu Ding. Adaptive affine transformation: A simple and effective operation for spatial misaligned image generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1167–1176, 2022. 1

[48] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3647–3656, 2022. 1, 2, 3

[49] Haitian Zheng, Lele Chen, Chenliang Xu, and Jiebo Luo. Unsupervised texture preserving flow for pose guided synthesis. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, PP, 2020. 1, 3

[50] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 161–178, Cham, 2022. Springer Nature Switzerland. 1, 2, 3, 6, 7, 8

[51] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11465–11475, 2021. 6, 7, 8

[52] Wenbin Zou, Mingchao Jiang, Yunchen Zhang, Liang Chen, Zhiyong Lu, and Yi Wu. Sdwnet: A straight dilated network with wavelet transformation for image deblurring. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1895–1904, 2021. 2