FreqHPT: Frequency-aware attention and flow fusion for Human Pose Transfer

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

Abstract

Human pose transfer is a challenging task that synthesizes images in various target poses while preserving the original appearance. This is typically achieved through aligning the source texture and supplementing it to the target pose. However, most of previous alignment methods only rely on either attention or flow, thereby failing to fully leverage distinctive strengths of these two methods. Moreover, the receptive field of these methods is generally limited in supplementation, resulting in the lack of global texture consistency. To address this issue, observing that attention and flow exhibit distinct characteristics in terms of their frequency distribution, Frequency-aware Human Pose Transfer (FreqHPT) is proposed in this paper. FreqHPT investigates the complementarity between attention and flow from the frequency perspective for improving texture-preserving pose transfer. To this end, FreqHPT first transforms the features from attention and flow into the wavelet domain and then fuses them over multi-frequency bands in an adaptive manner. Subsequently, FreqHPT globally refines the fused features in the Fourier space for texture supplement, enhancing the overall semantic consistency. Extensive experiments on the DeepFashion dataset demonstrate the superiority of FreqHPT in generating texture-preserving and realistic pose transfer images.

1. Introduction

Human pose transfer has garnered significant attention due to its potential applications in the fields of fashion and design, such as virtual try-on, art design and online fashion shopping [24, 37]. However, accurately and effectively modeling the texture transformation during the pose transfer process remains a significant challenge.

In order to achieve texture-preserving human pose transfer, previous works employ various spatial alignment methods, including deformable convolution [32], affine transformation [41], flow [3, 16, 18, 21, 26, 43] and attention [15, 17, 24, 28, 37], and then supplement the aligned source information along with the target to obtain generated images. Among these methods, flow and attention have emerged as the most effective and widely used approaches.

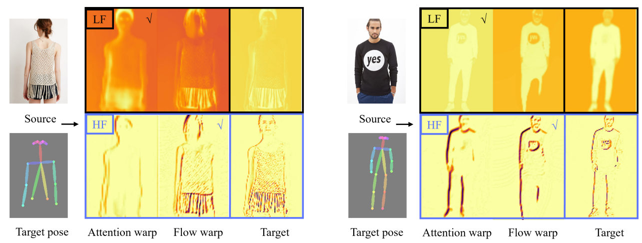

Figure 1. The illustration of our observation that flow-based and attention-based methods exhibit complementary differences in frequency distributions. This observation indicates that attention can better recover the low-frequency (LF) structure (overall human structure), whereas the high-frequency (HF) components from flow warping are closer to the target (cloth texture details).

Flow-based [38, 44, 45] and attention-based [29, 31, 42] methods have distinctive strengths and limitations in addressing human pose transfer, which is reflected by frequency domain. As illustrated in Figure 1, flow operations identify the point-wise correspondences between source and target. These methods place greater emphasis on capturing rich high-frequency texture details, but lack a guarantee for modeling overall structure. In contrast, attention operations calculate the weighted summation of all source values for each target position. These methods prioritize low-frequency semantic consistency, but may exhibit relatively weaker performance in preserving local texture and details. Therefore, relying only on flow or attention alone is insufficient to achieve texture-preserving human pose transfer. Besides, previous methods supplement source texture to the target under restricted receptive fields without awareness of their image-wide correlation, which may bring inconsistent global human semantics.

*Equal Contribution.
†Corresponding Author.

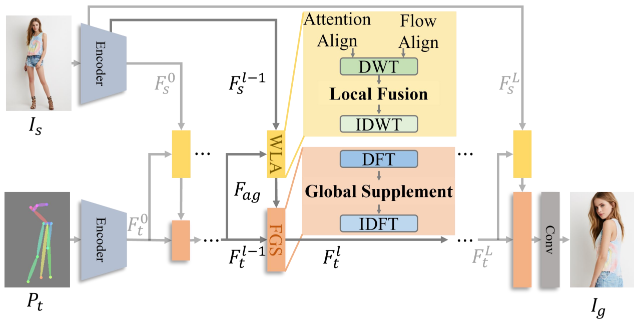

Figure 2. The overview of the proposed FreqHPT. Given the source image I_s and target pose P_t, WLA utilizes the attention and flow to deform the source features, which are fused into F_{ag} locally in the wavelet space. Then the fused features are globally supplemented by FGS in the Fourier space and decoded to generate the desired image I_g.

Based on the above analysis, we propose Frequency-aware Human Pose Transfer (FreqHPT) for texture-preseving human pose transfer, which leverages the complementarity between attention and flow in a frequency-aware manner. To boost the alignment effect, FreqHPT applies Wavelet Local Fusion (WLA) to align the texture with attention and flow by leveraging their advantages in different frequency bands. WLA decomposes the aligned texture with wavelet transform [23] and fuses the aligned features in low- and high-frequencies distinguishable and locally, which avoids the interruption between different frequencies and improves the quality of the target texture. To further supplement the aligned features, FreqHPT employs Fourier Global Supplement (FGS) to supplement the global information from aligned source to the target. FGS adopts Fourier transform [4] to extract Fourier features characterizing the global texture information and adaptively supplements the source information to the target through the interaction between Fourier features. The global Fourier frequency information complements the local information in the wavelet domain while enhancing the feature representation and model capability.

The main contributions of this paper are as follows: (1) A novel human pose transfer approach, dubbed FreqHPT, is introduced, which effectively investigates the complementarity between flow and attention in frequency domain. FreqHPT involves texture alignment achieved through the fusion of flow and attention in the wavelet domain and texture supplementation in the Fourier domain. (2) Our method surpasses existing state-of-the-art approaches both qualitatively and quantitatively, which is demonstrated through extensive experiments on the benchmark.

2. Related Works

Frequency Domain Analysis. Frequency domain analysis has shown to be effective for various computer vision tasks, such as image inpainting [35] and image generation [34]. For instance, the wavelet transform has achieved great success in these tasks by decomposing the signal into different frequency bands. Fourier transformation allows global manipulation and analysis of Fourier features, and its effectiveness in capturing global receptive fields has been demonstrated [2]. Furthermore, Fourier analysis has been applied in video generation [33] and image restoration [5,13]. However, there has been no prior work focusing on the application of frequency domain analysis in pose transfer.

Human Pose Transfer. Pose transfer aims to synthesize human images in different poses. Previous work has used style manipulation [22,36] and flow-based [16,18,26] approaches to handle this task, but they suffer from losing spatial texture details or failing to generate reasonable global semantics. Attention-based methods [24,38,45] have shown their capability in rendering global semantic structures. [25] combined the advantages of attention and flow in the spatial domain, while our method performs fusion in the frequency domain, which better captures the characteristic of frequency distribution from attention and flow.

3. Method

The framework overview is shown in Figure 2. FreqHPT utilizes WLA to align features along with Discrete Wavelet Transform (DWT) and Inverse Discrete Wavelet Transform (IDWT), and applies FGS to supplement them along with Discrete Fourier Transform (DFT) and Inverse Discrete Fourier Transform (IDFT).

3.1. Wavelet Local Alignment

As illustrated in Figure 3, WLA first warps the source feature F_s into F_{attn} and F_{flow} by attention and flow. Flow spatially deforms the source appearance into a target pose by calculating a point-wise 2D deformation field, which correlates the target position with the specific source candidates. We follow [6, 8, 10] to estimate the flow in a coarse to fine manner. Attention predicts the value of the target position with the weighted summation of the whole source values. We choose the efficient double attention [1, 24, 40] as our attention calculation backbone which splits the attention calculation into gathering and distribution. The aligned features from attention and flow contain both low-frequency semantic structures and high-frequency sharp details. Therefore, it is reasonable to fuse them adaptively across different frequency bands in the wavelet domain.

Specifically, we first decompose the spatial feature into different frequency elements with DWT. Haar wavelet filter with a low-pass filter $(1/\sqrt{2}, -1/\sqrt{2})$ and high-pass filter $(1/\sqrt{2}, 1/\sqrt{2})$ is applied to extract low-frequency LF (*LL*) and high-frequency HF components (*LH*, *HL*, *HH*). Subsequently, the mask prediction network estimates the masks with global style modulation techniques. It takes

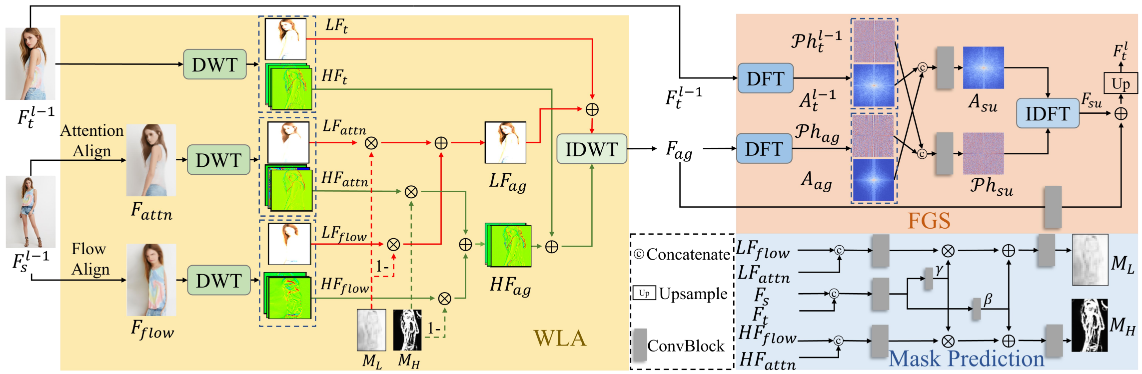

Figure 3. The detailed structure of WLA and FGS. Given the aligned features F_{attn} and F_{flow} from source feature F_s^{l-1} by attention and flow along with target feature F_t^{l-1}, WLA decomposes them into low- and high-frequency components LF and HF, which are fused by low- and high-frequency masks M_L and M_H calculated with Mask Prediction. FGS then applies DFT to transform F_t^{l-1} and F_{ag} into the Fourier features including amplitude components A_t^{l-1}/ A_{ag} and phase components $\mathcal{P}h_t^{l-1}$/$\mathcal{P}h_{ag}$, which are convolved with each other to refine feature representation globally. Finally, F_t^l is obtained by summation with convolved F_{ag}, followed by upsampling.

$LF_{flow/attn}$, $HF_{flow/attn}$, F_s, and F_t as inputs and outputs the soft masks M_L and M_H in the low- and high-frequency bands. Finally, aligned source features F_{ag} are acquired by fusing F_{attn} and F_{flow}, which is computed as the weighted summation of F_{attn} and F_{flow} with corresponding masks M_L and M_H in low- and high-frequencies.

Due to the independent processing for each frequency band, we can emphasize the semantical structures of the attention-aligned features in the low-frequency domain and retain the sharp details of the flow-aligned features in the high-frequency effectively. Finally, the wavelet features are converted back into the F_{ag} in spatial space with IDWT.

3.2. Fourier Global Supplement

Although WLF performs source texture alignment and fusion locally in the wavelet domain, it lacks consideration of the global information exchange between source and target. To address this limitation, FGS turns to supplement the F_{ag} to the target globally in the Fourier domain. Specifically, as shown in Figure 3, FGS transforms aligned source feature F_{ag} and target feature F_t with DFT and updates the global Fourier features containing Fourier amplitude and phase from $F_{ag} \in \mathbb{R}^{h\times w}$ and $F_t \in \mathbb{R}^{h\times w}$, respectively. The updated amplitude value reflecting the frequency intensity is expressed as:

$$\mathcal{A}_{su} = \mathcal{C}\{\sqrt{(\sum_{x,y} F_t cos(-2\pi\mathcal{M}))^2 + (\sum_{x,y} F_t sin(-2\pi\mathcal{M}))^2}, \sqrt{(\sum_{x,y} F_{ag} cos(-2\pi\mathcal{M}))^2 + (\sum_{x,y} F_{ag} sin(-2\pi\mathcal{M}))^2}\}, \tag{1}$$

where $\mathcal{C}\{\cdot,\cdot\}$ represents the concatenation and convolution operations, as well as $\mathcal{M}$ denotes the result value located in (u, v), which equals $\frac{x}{h}u + \frac{y}{w}v$ calculated in (x, y) coordinate of the spatial input. Besides, the refined phase value characterizing the global structure is expressed as:

$$\mathcal{P}h_{su} = \mathcal{C}\{\varepsilon(\frac{\sum_{x,y} F_t sin(-2\pi\mathcal{M})}{\sum_{x,y} F_t cos(-2\pi\mathcal{M})}), \varepsilon(\frac{\sum_{x,y} F_{ag} sin(-2\pi\mathcal{M})}{\sum_{x,y} F_{ag} cos(-2\pi\mathcal{M})})\}, \tag{2}$$

where ε denotes the $\arctan$ function, IDFT then turns the processed Fourier features back into the spatial feature:

$$F_{su} = \text{IDFT}(\mathcal{A}_{su}, \mathcal{P}h_{su}). \tag{3}$$

By globally reasoning in the Fourier domain and supplementing the aligned features F_{ag}, FGS achieves an enhanced representation F_{su}. Then the next-level feature F_t^l is calculated by summation with F_{ag} and upsampling.

3.3. Loss Functions

Deformation loss $\mathcal{L}_{defor}$ encourages accurate deformation using attention matrix $A_{s\to t}$ and flow $W_{s\to t}$. We deform the downsampled source image and calculate the l_1 distance with corresponding target image $I_t^{\downarrow}$ as $\mathcal{L}_{defor} = ||\mathcal{RS}(W_{s\to t}, I_s^{\downarrow}) - I_t^{\downarrow}||_1 + \sum_i ||\nabla(W_{s\to t})_i|| + ||\mathcal{A}_{s\to t}I_s^{\downarrow} - I_t^{\downarrow}||_1$, where $\mathcal{RS}$ is the resample function [12] and $||\nabla(W_{s\to t})_i||$ [27] enables the smoothness of the flow. Multiple widely used image-to-image loss functions such as Perceptual loss $\mathcal{L}_{perc}$ [14], Adversarial loss $\mathcal{L}_{adv}$ [7], and SSIM loss $\mathcal{L}_{adv}$ [30] are also applied for network training.

4. Experiments

Dataset and metrics. The experiments are conducted on the DeepFashion dataset [19],following the setting of [37] in 256×176 resolution and [24] in 512×352 resolution. We utilize SSIM [30], LPIPS [39] and FID [11] to measure the

Method	SSIM↑	FID↓	LPIPS↓		Reid Score (%)↑	
			AlexNet	VGG	Topk-1	MAP
ADGAN [22]	0.6721	14.4580	0.2283	0.2557	81.46	80.26
GFLA [26]	0.7677	10.8429	0.2258	0.2765	90.84	87.56
PISE [36]	0.7682	11.5144	0.2080	0.2498	90.09	87.22
SPGNet [20]	0.7758	12.7027	0.2102	0.2443	94.43	91.60
CASD [44]	0.7248	11.3732	0.2157	0.2645	93.09	91.20
NTED [24]	0.7715	9.2876	0.2019	0.2564	97.34	94.73
DPTN [37]	0.7782	11.4664	**0.1957**	0.2459	97.69	95.04
Ours	**0.7800**	**8.9072**	0.1977	**0.2369**	**98.72**	**95.93**
CocosNet2 [45]	0.7236	13.3250	0.2265	0.2735	87.84	86.75
NTED [24]	0.7376	7.7821	**0.1980**	0.2472	99.71	97.33
Ours	**0.7456**	**6.5522**	0.2026	**0.2471**	**98.48**	**97.80**

Table 1. Quantitative comparison results on DeepFashion. (Upper and lower rows are for 256×176 and 512×352 resolutions). The first and the second best are bold and underlined.

Method	SSIM↑	FID↓	LPIPS ↓		Reid Score(%)↑	
			AlexNet	VGG	Topk-1	MAP
w/o WLA	0.7769	11.5039	0.2246	0.2860	97.08	94.96
w/o FGS	0.7778	9.5311	0.2026	0.2400	98.77	96.67
Ours	**0.7800**	**8.9072**	**0.1977**	**0.2369**	**98.83**	**96.94**

Table 2. Ablation study on the DeepFashion dataset. *w/o* WLA setting adopts the similar fusion strategy with [25].

Figure 4. Qualitative results of FreqHPT on DeepFashion dataset.

image quality. Reid Score [37] further measures the Top-1 and Mean Average Precision (MAP) by re-identification model [9], which tests whether the generated query image can be matched with the corresponding real gallery image.

Quantitative comparison. According to Table 1, our method achieves the best and second-best results of all metrics in both low- and high-resolution datasets, which verifies the superiority of our proposed FreqHPT in rendering high-quality images and maintaining detailed texture.

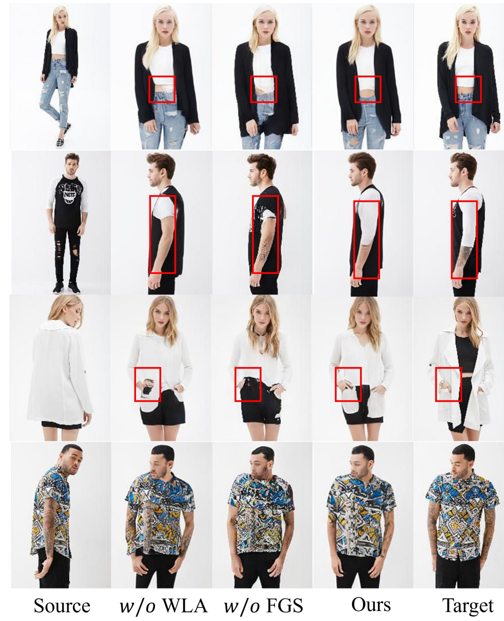

Figure 5. Visual comparison results of the ablation study.

Qualitative comparison. As depicted in Figure 4, FreqHPT is capable of facilitating the reoccurrence of source texture patterns (see 4th to 6th rows) and improving semantical consistency (see 1st row), owing to the proposed local texture alignment in WLA and global supplement in FGS.

Ablation study. We train several variant models to explore the contributions of proposed designs by removing the specific part from FreqHPT. As shown in Figure 5, the absence of WLA leads to local artifacts (e.g., distorted hand shape in the 3rd row) due to the lack of accurate local alignment. Additionally, the model without FGS fails to generate global consistent textures (e.g., the shoulder area in the 2nd row). FreqHPT generates both local and global high-quality textures. Quantitative results in Table 2 further verify the effectiveness of each component in FreqHPT.

5. Conclusion

This paper presents a novel human pose transfer framework, which effectively leverages the complementarity of attention and flow in a frequency-aware manner. The proposed method deforms the features by fusing attention and flow in the wavelet domain and then supplements the fused features to the target in the Fourier domain. Extensive experiments demonstrate that the integration of flow and attention from a frequency perspective can significantly improve texture-preserving human pose transfer.

Acknowledgement. This work was supported by Alibaba Group through Alibaba Research Intern and National Natural Science Foundation of China (Grant No. 62274142).

References

[1] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A2-Nets: Double Attention Networks. In *Advances in Neural Information Processing Systems*, 2018. 2

[2] Lu Chi, Borui Jiang, and Yadong Mu. Fast Fourier Convolution. In *Advances in Neural Information Processing Systems*, pages 4479–4488, 2020. 2

[3] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-Gated Warping-GAN for Pose-Guided Person Image Synthesis. In *Advances in Neural Information Processing Systems*, 2018. 1

[4] Matteo Frigo and Steven G Johnson. FFTW: An adaptive software architecture for the FFT. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1381–1384. IEEE, 1998. 2

[5] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier Space Losses for Efficient Perceptual Image Super-Resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2340–2349, 2021. 2

[6] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-Free Virtual Try-on via Distilling Appearance Flows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8481–8489, 2021. 2

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, 2014. 3

[8] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. ClothFlow: A Flow-Based Model for Clothed Person Generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10471–10480, 2019. 2

[9] 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. 4

[10] Sen He, Yi-Zhe Song, and Tao Xiang. Style-Based Global Appearance Flow for Virtual Try-On. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2022. 2

[11] 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. 3

[12] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems*, 2015. 3

[13] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal Frequency Loss for Image Reconstruction and Synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 13899–13909, 2021. 2

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Proceedings of the European Conference on Computer Vision*, pages 694–711. Springer, 2016. 3

[15] 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

[16] Yining Li, Chen Huang, and Chen Change Loy. Dense Intrinsic Appearance Flow for Human Pose Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019. 1, 2

[17] 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. In *The International Conference on Learning Representations*, 2023. 1

[18] 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 International Conference on Computer Vision*, pages 5904–5913, 2019. 1, 2

[19] 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. 3

[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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10801–10810, 2021. 4

[21] 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*, 25:930–941, 2021. 1

[22] 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 Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020. 2, 4

[23] Valérie Perrier, Thierry Philipovitch, and Claude Basdevant. Wavelet spectra compared to Fourier spectra. *Journal of mathematical physics*, 36(3):1506–1519, 1995. 2

[24] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural Texture Extraction and Distribution for Controllable Person Image Synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13535–13544, 2022. 1, 2, 3, 4

[25] Yurui Ren, Yubo Wu, Thomas H Li, Shan Liu, and Ge Li. Combining Attention with Flow for Person Image Synthesis. In *Proceedings of the ACM International Conference on Multimedia*, pages 3737–3745, 2021. 2, 4

[26] 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 Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020. 1, 2, 4

[27] Deqing Sun, Stefan Roth, and Michael J. Black. A Quantitative Analysis of Current Practices in Optical Flow Estima-

tion and the Principles Behind Them. *International Journal of Computer Vision*, 106:115–137, 2013. 3

[28] Hao Tang, Song Bai, Li Zhang, Philip H. S. Torr, and N. Sebe. XingGAN for Person Image Generation. In *Proceedings of the European Conference on Computer Vision*, pages 717–734, 2020. 1

[29] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 1

[30] 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. 3

[31] 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. 1

[32] Dongxu Wei, Xiaowei Xu, Haibin Shen, and Kejie Huang. C2F-FWN: Coarse-to-Fine Flow Warping Network for Spatial-Temporal Consistent Motion Transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2852–2860, 2021. 1

[33] Guang Yang, Wu Liu, Xinchen Liu, Xiaoyan Gu, Juan Cao, and Jintao Li. Delving into the Frequency: Temporally Consistent Human Motion Transfer in the Fourier Space. In *Proceedings of the ACM International Conference on Multimedia*, pages 1156–1166, 2022. 2

[34] Mengping Yang, Zhe Wang, Ziqiu Chi, and Wenyi Feng. WaveGAN: Frequency-Aware GAN for High-Fidelity Few-Shot Image Generation. In *Proceedings of the European Conference on Computer Vision*, pages 1–17. Springer, 2022. 2

[35] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. WaveFill: A Wavelet-based Generation Network for Image Inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 14094–14103, 2021. 2

[36] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. PISE: Person Image Synthesis and Editing With Decoupled GAN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7982–7990, 2021. 2, 4

[37] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring Dual-Task Correlation for Pose Guided Person Image Generation . In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7713–7722, 2022. 1, 3, 4

[38] 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 Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020. 1, 2

[39] 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. 3

[40] 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 Conference on Computer Vision and Pattern Recognition*, pages 13425–13434, 2021. 2

[41] Zhimeng Zhang and Yu Ding. Adaptive Affine Transformation: A Simple and Effective Operation for Spatial Misaligned Image Generation. In *Proceedings of the ACM International Conference on Multimedia*, pages 1167–1176, 2022. 1

[42] Jian Zhao and Hui Zhang. Thin-Plate Spline Motion Model for Image Animation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022. 1

[43] Haitian Zheng, Lele Chen, Chenliang Xu, and Jiebo Luo. Unsupervised Texture Preserving Flow for Pose Guided Synthesis. *IEEE Transactions on Image Processing*, PP, 2020. 1

[44] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross Attention Based Style Distribution for Controllable Person Image Synthesis . In *Proceedings of the European Conference on Computer Vision*, pages 161–178. Springer, 2022. 1, 4

[45] 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 Conference on Computer Vision and Pattern Recognition*, pages 11465–11475, 2021. 1, 2, 4