Pose-aware Disentangled Multiscale Transformer for Pose Guided Person Image Generation

1st Kei Shibasaki
Electrical and Information Engineering (of Keio University)
Kanagawa, Japan, shibasaki@tkhm.elec.keio.ac.jp

2nd Masaaki Ikehara
Electrical and Information Engineering (of Keio University)
Kanagawa, Japan, ikehara@tkhm.elec.keio.ac.jp

***Abstract*—Pose Guided Person Image Generation (PGPIG) is the task that transforms the pose of a person's image from the source image, its pose information, and the target pose information. Most existing PGPIG methods require additional pose information or tasks, which limits their application. Moreover, they use CNNs, which can only extract features from neighboring pixels and cannot consider the consistency of the entire image. This paper proposes a PGPIG network solving these problems by using a module containing Axial Transformers with large receptive field. The proposed method disentangles the PGPIG task into two subtasks: "rough pose transformation" and "detailed texture generation". In the former task, low-resolution feature maps are transformed by blocks containing Axial Transformer. The latter task uses a CNN network with Adaptive Instance Normalization. Experiments show that the proposed method has competitive performance with other state-of-the-art methods. Furthermore, despite achieving excellent performance, the proposed network has a significantly fewer parameters than existing methods.**

***Index Terms*—Machine Learning, Deep Learning, Pose Guided Person Image Generation, Pose Transfer, Transformers**

I. Introduction

Pose Guided Person Image Generation (PGPIG) is the task that transforms the pose of a person image from the source image, its pose information and the target pose information. PGPIG has been applied to human images/videos generation and person re-identification [6], [7], [20], [27].

PGPIG has two problems: it needs to deal with the significant pose change, and it should be consistent with the entire image. To deal with the former problem, existing methods use additional parsing maps with semantic person information for training [17], [18], [28] or set up additional tasks to capture useful features [17], [21], [28], [29]. However, preparing additional data is laborious and setting up additional task increases training time and makes hyper-parameters more difficult to determine.

Existing networks use CNNs that can extract features only from neighboring pixels. Therefore, extracting enough features between the source and target information is difficult, and they cannot take into account overall image consistency.

In this paper, we propose a network to solve these problems of existing methods. The architecture of the proposed network is shown in Fig. 1. The proposed method uses Axial Transformer [10] with large receptive field. Therefore, it does not require additional data or tasks to deal with significant pose change and can also take into account the consistency of the entire image. Since nothing additional is required, the proposed method can achieve a very small number of parameters compared to existing state-of-the-art methods. In the proposed method, PGPIG is disentangled into two subtasks, "rough pose transformation" and "detailed texture generation", which are processed by different modules. The problems of significant pose changes and overall image consistency are integrated into the "rough pose transformation". The "rough pose transformation" is performed with an Axial Transformer Transformation Block (ATTB) (Fig. 2 (a)). The "detailed texture generation" is transformed by the CNN Transformation Block (CTB) (Fig. 2 (b)).

Transformer-based modules are suitable for solving PGPIG's problems because they have wide receptive fields. However, the transformer-based module is often computationally costly when calculating large feature maps. While Axial Transformer Transformation Block is the transformer-based module, it can reduce the computational cost while keeping wide receptive field. ATTB is applied to only low-resolution feature maps, which further reduces the computational cost.

In PGPIG, "detailed texture generation" does not need a wide receptive field because texture patterns do not have long-range dependencies. Therefore, we address this subtask by transforming high-resolution feature maps with CTB using CNNs as feature extractors.

Experiments show that the proposed method has competitive performance both quantitatively and qualitatively compared to existing state-of-the-art PGPIG methods. The main contributions of this paper are as follows:

- This paper proposes a network that solves the PGPIG problems of significant pose transformation and overall image consistency without adding extra data and task by using the module including Axial Transformer. The proposed method disentangles PGPIG into "rough pose transformation" and "detailed texture generation". The former is processed by Axial Transformer Transformation Block (ATTB) and the latter is processed by CNN Transformation Block (CTB).
- Experimental results show that the proposed method is competitive with the existing state-of-the-art methods. Furthermore, the number of parameters of the proposed network is tiny (8.83M), which is only 7.54% of the size of SPGNet [17], one of the state-of-the-art methods.

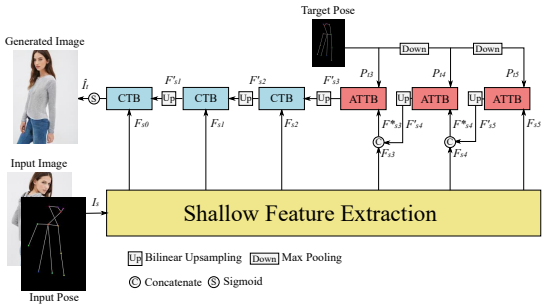


Fig. 1. Overall structure of the proposed network.

II. Related Works

A. Pose Guided Person Image Generation (PGPIG)

Tang et al. [25] proposed a method that cross pose information and appearance information. However, vanilla CNN-based modules cannot address significant pose transformations. Li et al. [14] and Ren et al. [21] proposed methods that estimate the optical flow and warp the input image to generate images. However, due to significant pose changes or occlusions, the optical flow estimation may fail, resulting in low-quality images. In addition, the optical flow estimation is often trained separately from the main body of the network. Therefore, it requires time-consuming hyperparameter tuning and increases the training time. Ma et al. [18], Li et al. [17], and Zhang et al. [28] proposed a network using additional person parsing maps to improve the quality of PGPIG. Their method first estimates the parsing map and then generates the final image. However, preparing the parsing maps is laborious, and the parsing maps themselves are often unreliable. Zhang et al. [29] proposed a method which adopts Siamese Network [13]. It makes the network efficiently capture the texture information by learning the source-to-source identity mapping as an auxiliary task. However, the auxiliary task increases the number of hyper-parameters, increasing the learning difficulty.

B. Axial Attention, Axial Transformer

Axial Attention [10] is one of the attention methods for multidimensional tensors, and Axial Transformer [10] is a Transformer that employs Axial Attention for Multi Head Attention. Transformer-based networks [5], [15], [19] have the drawbacks that the computational cost increases rapidly as the input image resolution. Axial Transformer effectively reduces the computational cost without narrowing the receptive field in the vertical and horizontal directions. Specifically, for a tensor whose input size is $R \times R \times C$, the computational cost in standard attention is $O(R^4C)$, whereas it is $O(2R^3C)$ in Axial Attention [10].

C. Conditional Positional Encoding

Since Transformer-based modules cannot capture the positional relationships of feature maps, a module that embeds the positional relationships into the feature maps is needed. Modules which embed positional relationships into a feature map is called positional encoding [26]. Conditional Positional Encoding (CPE) [4] is one of the positional encoding methods for multidimensional tensors. Unlike standard positional encoding methods [22], [26], CPE can take into account absolute positional information which is helpful for networks [4], [23].

III. Proposed Method

The entire proposed network is shown in Fig. 1. Many existing methods require additional data and tasks. In contrast, the proposed method requires no additional data or tasks.

The proposed network combines the source person image and its pose information (denoted as I_s) as input and extracts multi-scale shallow features with the Shallow Feature Extraction module. The proposed method tackles PGPIG by disentangling it into "rough pose transformation" and "detailed texture generation" tasks. Among the extracted features, low-resolution feature maps with global features are transformed by Axial Transformer Transformation Block (ATTB) for "rough pose transformation". High-resolution features are transformed by the CNN Transformation Block (CTB) for "detailed texture generation". We denote the part with the same resolution as F_{si} in Fig. 1 as level $L = i$ The kernel sizes of Conv2d in Shallow Feature Extraction and CTB are 7 (at $L = 0, 1$), 5 (at $L = 2$) and 7 (at $L = 3, 4$).

A. Shallow Feature Extraction

Shallow Feature Extraction, shown in Fig. 2 (a), is a module for extracting shallow features from the input information I_s. Shallow Feature Extraction is a multi-scale module with output $F_{si}(i = 0, 1, 2, 3, 4, 5)$. When the size of I_s is $H \times W$, the size of F_{si} is $\frac{H}{2^i} \times \frac{W}{2^i}$. The number of channels and depth of each layer are 128, 2 for $L = 0, 1$ and 64, 4 for $L = 2, 3, 4$.

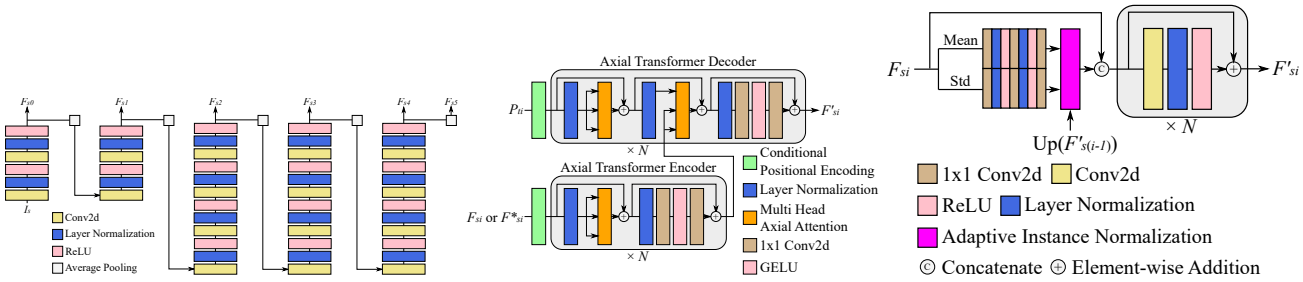
B. Axial Transformer Transformation Block (ATTB)

The Axial Transformer Transformation Block (ATTB) is the module shown in Fig. 2 (b). ATTB consists of an N-layer Axial Transformer Encoder-Decoder block. The input to the encoder is the feature F_{si} extracted from the Shallow Feature Extraction, or F^*_{si}. F^*_{si} is the concatenation of F_{si} and $F'_{s(i-1)}$ which is the output from ATTB at the lower resolution. The input to the decoder is the target pose information P_{ti}. Conditional Positional Encoding (CPE) is used for positional encoding on each input. The number of channels and the value of N in the encoder and decoder are 64 and $N = 2$ for $L = 3$, 128 and $N = 2$ for $L = 4$, and 128 and $N = 4$ for $L = 5$.

The resolution of P_{ti} is $\frac{H}{2^i} \times \frac{W}{2^i}$ if the resolution of the input image is $H \times W$. The proposed method does not use the full-resolution pose information but instead downsamples the pose information by Max Pooling and then feeds it into the network. The pose information inputted to ATTB is P_{t3}, P_{t4}, and P_{t5}. Each resolutions of pose information are $\frac{H}{2^3} \times \frac{W}{2^3}$, $\frac{H}{2^4} \times \frac{W}{2^4}$, and $\frac{H}{2^5} \times \frac{W}{2^5}$. ATTB processes only low-resolution feature maps to reduce the computational cost.

C. CNN Transformation Block (CTB)

The CNN Transformation Block (CTB) is the module shown in Fig. 2 (c). We adopt AdaIN for taking advantage of the fact



(a) Shallow Feature Extraction (b) Axial Transformer Transformation Block (ATTB) (c) CNN Transformation Block (CTB)

Fig. 2. Details of each module of the proposed network.

that textures rarely change in the PGPIG task. CTB process only high-resolution feature maps to address "detailed texture generation." The content input to AdaIN is the $F'_{s(i-1)}$ and the style input is the output of a three-layer network that processes the mean and standard deviation of F_{si}. For the number of channels and the value of N in CTB are 64 and $N = 4$ for $L = 0$, 1 and 64 and $N = 6$ for $L = 2$.

D. Loss Function

The loss functions of the proposed method are (1) and (2). L_{adv} is the adversarial loss, same objective as WGAN [1]. A standard 6-layer CNN network was used for the discriminator. L_{l1} is the mean absolute error between the generated image and ground truth. L_{perc} is the Perceptual Loss [11] with VGG19 [24] as the feature extractor. L_{style} is the Style Loss [11] with VGG19 [24] as the feature extractor. gp is the Gradient Penalty [8], which is employed to stabilize the learning. $\lambda_i (i = 1, 2, 3, 4, 5)$ is a hyper-parameter to adjust the loss ratio.

$$L_G = \lambda_1 L_{adv} + \lambda_2 L_{l1} + \lambda_3 L_{perc} + \lambda_4 L_{style} \quad (1)$$

$$L_D = \lambda_1 L_{adv} + \lambda_5 \mathrm{gp} \quad (2)$$

IV. Experiment

A. Set Up

Datasets: Deep Fashion [16] and Market-1501 [31] are used for quantitative and qualitative comparisons. These datasets are commonly used in PGPIG tasks [17], [21], [25], [29]. Deep Fashion is a dataset consisting of 52,712 high-quality images of clothed persons with clean white backgrounds. The resolution of the images is 256×176. Market-1501 is a dataset of 263,632 images of people with various viewpoints, illumination and backgrounds. The resolution of the images is 128×64. The human pose information (18 joint keypoints) are extracted from OpenPose [3]. For a fair comparison, we split the datasets as [21].

Metrics: Following existing methods [21], [25], we adopt Learned Perceptual Image Patch Similarity (LPIPS) [30], Inception Score (IS) [2] and Fréchet Inception Distance (FID) [9] as evaluation metrics.

Hyper Parameters: We set $\lambda_1 = 1, \lambda_2 = 2.5, \lambda_3 = 0.25, \lambda_4 = 250, \lambda_5 = 10$. The batch size is 8, and the number of training steps is 500,000 for Deep Fashion and 100,000 for Market-1501. In both datasets, the learning rate is 1.0×10^{-4} for both the generator and the discriminator. Learning rate decay is used in the training phase, with the learning rate multiplied by 0.1 for 250,000 and 400,000 steps in Deep Fashion and 50,000 and 80,000 steps in Market-1501. The optimizer is Adam [12], and we set $\beta_1 = 0.5$ and $\beta_2 = 0.999$.

TABLE I
QUANTITATIVE COMPARISON OF THE PROPOSED METHOD WITH SEVERAL STATE-OF-THE-ART METHODS. THE BEST VALUES ARE IN BOLD, THE SECOND BEST VALUES ARE UNDERLINED.

	DeepFashion			Market-1501			
	LPIPS↓	IS↑	FID↓	LPIPS↓	IS↑	FID↓	#Params
XingGAN [25]	0.2929	2.878	44.808	0.3059	**3.201**	37.510	44.84M
GFLA [21]	<u>0.1869</u>	2.856	**7.332**	0.2815	2.849	<u>28.042</u>	14.04M
MUST [18]	0.2467	**2.971**	17.220	-	-	-	51.45M
SPGNet [17]	0.2109	2.711	11.964	<u>0.2777</u>	2.942	30.520	117.13M
PISE [28]	0.2084	2.815	9.905	-	-	-	64.00M
DPTN [29]	0.1966	2.867	9.683	**0.2711**	2.965	28.678	<u>9.79M</u>
Ours	**0.1849**	<u>2.944</u>	<u>8.034</u>	0.2939	<u>3.091</u>	**23.307**	**8.83M**

B. Quantitative Comparison

We compare the proposed method with several state-of-the-art methods [17], [18], [21], [25], [28], [29]. The results of the quantitative comparison of generated images and model size are shown in Tab. I. This table shows that the proposed method is competitive with the state-of-the-art methods, with the first or second best performance in five of the six evaluation metrics. Most of the existing methods in Tab. I require additional parsing maps (MUST, SPGNet, PISE) or additional task setup and training (GFLA, SPGNet, DPTN). However, the proposed method does not use these additional elements and still achieves excellent performance. The proposed method is convenient and easy for application since it is free from laborious data preparation and tuning many hyperparameters.

The number of parameters in the proposed network is 8.83M, less than any state-of-the-art methods listed in Tab. I. This is 7.54% of SPGNet (117.13M). The proposed network is a tiny network, yet it achieves excellent performance.

C. Qualitative Comparison

Fig. 3 shows a qualitative comparison of the proposed method with several state-of-the-art methods. The left side of Fig. 3 shows the Deep Fashion validation images, and the right shows the Market-1501 validation images. Although

Fig. 3. Qualitative comparison between the proposed method and several state-of-the-art methods. The left side is an image generated by Deep Fashion and the right side is an image generated by Market-1501.

TABLE II
VALUES OF THE EVALUATION METRICS IN DEEP FASHION WHEN THE AXIAL TRANSFORMER PART OF ATTB IS REPLACED BY CNN AND SWIN TRANSFORMER (SWINT). THE VALUES WITH THE BEST ACCURACY ARE IN BOLD.

	LPIPS↓	IS↑	FID↓
CNN	0.2020	2.846	8.935
SwinT	0.2187	2.814	9.541
ATTB	**0.1849**	**2.944**	**8.034**

XingGAN produces blurred images, the proposed method, which also requires no additional data or tasks, produces images of the competitive quality as state-of-the-art methods that require additional data or tasks.

In the Deep Fashion image in the fourth row of Fig. 3, all methods except the proposed method and DPTN produce ambiguous images, whether short pants or long pants. Some of these methods generate unnatural textures in the fifth row of Fig. 3. This is because these methods mainly use CNNs which can extract features only from neighboring pixels. This makes it difficult to consider the consistency of the entire image. In contrast, the proposed method, which uses Transformer-based modules with a wide receptive field, can generate images by taking into account the consistency of the entire image. For the Market-1501 images, the proposed method produces images competitive with other state-of-the-art methods.

Fig. 3 also shows that the proposed method can address significant pose changes as well as the existing methods.

D. Ablation Study

ATTB as a feature extractor: Tab. II shows that the performance is quantitatively better when the ATTB is used than when other feature extractors are used. The filter operation of the CNN and the Window Attention of the Swin Transformer narrow the receptive field in the vertical and horizontal directions and thus cannot take into account the consistency of the entire image. In contrast, ATTB does not narrow the receptive field in the vertical and horizontal directions, which has good effects for the networks.

V. CONCLUSION

This paper proposes a simple but powerful network for Pose Guided Person Image Generation (PGPIG). The proposed method can address significant transformations without requiring additional data or tasks and overall image consistency problem. TThis makes the network significantly lighter. Our network achieves competitive performance while it has only 7.54% number of parameters compared to existing state-of-the art network. The proposed network is designed based on the idea that PGPIG can be divided into "rough pose transformation" and "detailed texture generation". Each subtask is processed by Axial Transformer Transformation Block (ATTB) with Encoder-Decoder structure and CNN Transformation Block (CTB).

REFERENCES

[1] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 214–223.

[2] S. Barratt and R. Sharma, "A note on the inception score," *arXiv preprint arXiv:1801.01973*, 2018.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[4] X. Chu, Z. Tian, B. Zhang, X. Wang, X. Wei, H. Xia, and C. Shen, "Conditional positional encodings for vision transformers," *arXiv preprint arXiv:2102.10882*, 2021.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[6] Z. Feng, J. Lai, and X. Xie, "Learning view-specific deep networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3472–3483, 2018.

[7] A. Grigorev, A. Sevastopolsky, A. Vakhitov, and V. Lempitsky, "Coordinate-based texture inpainting for pose-guided human image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 135–12 144.

[8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[10] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," *arXiv preprint arXiv:1912.12180*, 2019.

[11] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[13] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015, p. 0.

[14] Y. Li, C. Huang, and C. C. Loy, "Dense intrinsic appearance flow for human pose transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3693–3702.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[16] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1096–1104.

[17] Z. Lv, X. Li, X. Li, F. Li, T. Lin, D. He, and W. Zuo, "Learning semantic person image generation by region-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 806–10 815.

[18] T. Ma, B. Peng, W. Wang, and J. Dong, "Must-gan: Multi-level statistics transfer for self-driven person image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 622–13 631.

[19] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International conference on machine learning*. PMLR, 2018, pp. 4055–4064.

[20] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 650–667.

[21] Y. Ren, X. Yu, J. Chen, T. H. Li, and G. Li, "Deep image spatial transformation for person image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7690–7699.

[22] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[23] K. Shibasaki, S. Fukuzaki, and M. Ikehara, "4k real time image to image translation network with transformers," *IEEE Access*, vol. 10, pp. 73 057–73 067, 2022.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] H. Tang, S. Bai, L. Zhang, P. H. Torr, and N. Sebe, "Xinggan for person image generation," in *European Conference on Computer Vision*. Springer, 2020, pp. 717–734.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[27] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin, "Pose guided human video generation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 201–216.

[28] J. Zhang, K. Li, Y.-K. Lai, and J. Yang, "Pise: Person image synthesis and editing with decoupled gan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7982–7990.

[29] P. Zhang, L. Yang, J.-H. Lai, and X. Xie, "Exploring dual-task correlation for pose guided person image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 7713–7722.

[30] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[31] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.