A Robust Pose Transformational GAN for Pose Guided Person Image Synthesis

Arnab Karmakar(✉) and Deepak Mishra

Indian Institute of Space Science and Technology,
Thiruvananthapuram 695547, Kerala, India
arnabkarmakar.001@gmail.com, deepak.mishra@iist.ac.in

Abstract. Generating photorealistic images of human subjects in any unseen pose have crucial applications in generating a complete appearance model of the subject. However, from a computer vision perspective, this task becomes significantly challenging due to the inability of modelling the data distribution conditioned on pose. Existing works use a complicated pose transformation model with various additional features such as foreground segmentation, human body parsing etc. to achieve robustness that leads to computational overhead. In this work, we propose a simple yet effective pose transformation GAN by utilizing the Residual Learning method without any additional feature learning to generate a given human image in any arbitrary pose. Using effective data augmentation techniques and cleverly tuning the model, we achieve robustness in terms of illumination, occlusion, distortion and scale. We present a detailed study, both qualitative and quantitative, to demonstrate the superiority of our model over the existing methods on two large datasets.

Keywords: Generative adversarial networks · Pose transformation · Image synthesis · Pose guided person image generation

1 Introduction

Given an image of a person, a pose transformation model aims to reconstruct the person's appearance in another pose. While for humans it is very easy to imagine how a person would appear in a different body pose, it has been a difficult problem in computer vision to generate photorealistic images conditioned only on pose; given a single 2D image of the human subject. The idea of pose transformation can help construct a viewpoint invariant representation. This has several interesting applications in 3D reconstruction, movie making, motion prediction or human computer interaction etc.

The task of pose transformation given a single image and a desired pose, is achieved by any machine learning model in basically two steps: (1) learning the significant visual features of the person-of-interest along with the background from the given image, and (2) imposing the desired pose on the person-of-interest, generate a photorealistic image while preserving the previously learned features. Generative Adversarial Networks (GAN) [8] have been widely popular in this

R. V. Babu et al. (Eds.): NCVPRIPG 2019, CCIS 1249, pp. 89–99, 2020.
https://doi.org/10.1007/978-981-15-8697-2_8

field due to its sharp image generation capability. While the majority of successful pose transformation models use different variation of GANs as their primary component, they give little importance to efficient data augmentation and utilization of inherent CNN features to achieve robustness. Recent developments in this field have been targeted to develop complex deep neural network models with the use of multiple external features such as human body parsing [4], semantic segmentation [1,4], spatial transformation [1,20] etc. Although this is helpful in some scenarios, there is accuracy issues and computational overhead due to each intermediate step that affects the final result.

In this work, we aim to develop an improved end-to-end model for pose transformation given only the input image and the desired pose, and without any other external features. We make use of the Residual learning strategy [9] in our GAN architecture by incorporating a number of residual blocks. We achieve robustness in terms of occlusion, scale, illumination and distortion by using efficient data augmentation techniques and utilizing inherent CNN features. Our results in two large datasets, a low-resolution person re-identification dataset Market-1501 [23] and high-resolution fashion dataset DeepFashion [13] have been demonstrated. Our contributions are two folds: First, we develop an improved pose transformation model to synthesize photorealistic images of a person in any desired pose, given a single instance of the person's image, and without any external features. Second, we achieve robustness in terms of occlusion, scale and illumination by efficient data augmentation techniques and utilizing inherent CNN features.

2 Related Work

There has been a lot of research in the field of generative image modelling using deep learning techniques. One line of work follow the idea of Variational Autoencoders (VAE) [5] which uses the reparameterization trick to maximize the lower bound of data likelihood [11]. VAEs have been popular for its image interpolation capability, but the generated images lack sharpness and high frequency details. GAN [8] models make use of adversarial training for generating images from random noise. Most works in pose guided person image generation make use of GANs because of its capability to produce fine details.

Amongst the large number of successful GAN architectures, many were developed upon the DCGAN [17] model that combines Convolutional Neural Network (CNN) with GANs. Pix2pix [10] proposed a conditional adversarial network (CGAN) for image-to-image translation by learning the mapping from condition image to target image. Yan et al. [22] explored this idea for pose conditioned video generation, where the human images are generated based on skeleton poses. GANs with different variations of U-Net [18] have been extensively used for pose guided image generation. The PG^2 [14] model proposes a 2-step process with a U-Net-like network to generate an initial coarse image of the person conditioned on the target pose and then refines the result based upon the difference map. Balakrishnan et al. [1] uses separate foreground and background synthesis using a spatial transformer network and U-Net based generator. Ma et al. [15] uses pose sampling using a GAN coupled with an encoder-decoder model. Dong et al.

[4] produces state-of-the-art results in pose driven human image generation and uses human body parsing as an additional attribute for Warping-GAN rendering. These additional attribute learning generates an overhead in computational capability and affects the final results. Other significant works for pose transfer in the field of person re-identification [16] [7] mostly deals with low resolution images and a complex training procedure. In this work, we propose a simplified end-to-end model for pose transformation without using additional feature learning at any stage.

3 Methodology

The proposed pt-GAN architecture is depicted in Fig. 1.

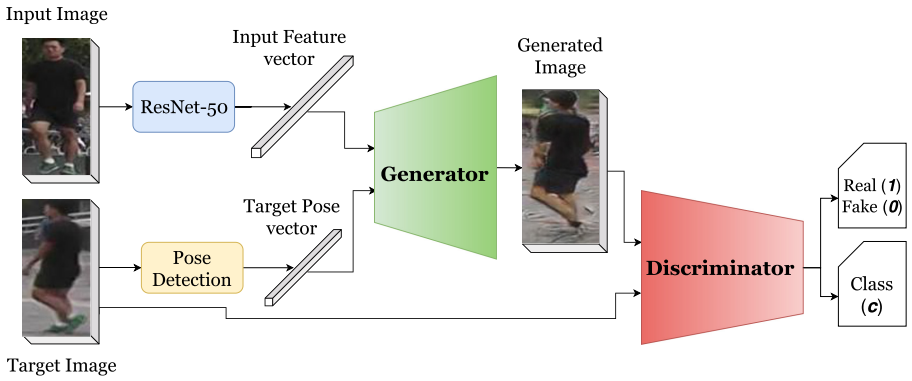

Fig. 1. Proposed architecture of the pose transformational GAN (pt-GAN). The idea is to transform the given person image to the desired pose. The additional classification branch of the Discrminator helps the Generator's learning to produce realistic images.

3.1 Pose Estimation

The image generation is conditioned on an input image and the target pose represented by a pose vector. In order to get the encoded pose vector, we use off-the-shelf pose detection algorithm OpenPose [2], which is trained without using either of the datasets deployed in this work. Given an input person image I_i, the pose estimation network OpenPose produces a pose vector P_i, which localizes and detects 25 anatomical key-points.

3.2 Generator

The Image generator (G_P) aims at producing the same person's images under different poses. Particularly, given an input person image I_i and a desired pose P_j, the generator aims to synthesize a new person image I_{P_j}, which contains the

same identity but with a different pose defined by P_j. The image vector obtained using a pretrained ResNet-50 [9] model (ImageNet [3]), and the pose vector are concatenated and fed to the generator. The architecture is depicted in Fig. 2.

The generator consists of multiple Convolution and Transposed Convolution layers. The key element of the proposed Generator is the residual blocks. Each residual block performs downsampling using convolution followed by upsampling using transposed convolution and then re-using the input by addition (Fig. 3(b)). The motivation is to take advantage of Residual Learning ($y = F(x) + x$) that can be used to pass invariable information (e.g. clothing color, texture, background) from the bottom layers to the higher layers and change the variable information (pose) to synthesize more realistic images, achieving pose transformation at the same time.

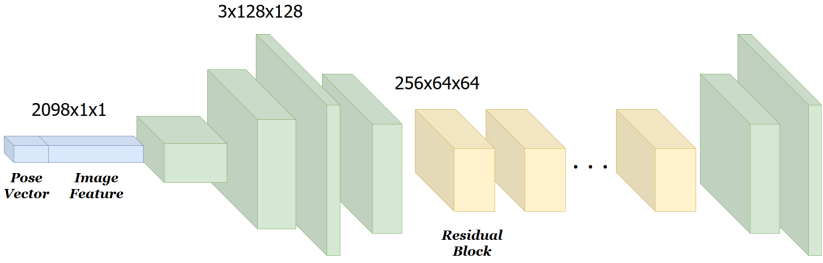

Fig. 2. Architecture of the Generator Network. The Generator network consists of 9 residual blocks, which helps the GAN to preserve low level features (clothing, texture), while transforming high level features (pose) of the subject.

3.3 Discriminator

In our implementation, the Discriminator (D_P) predicts the class label for the image along with the binary classification of determing whether the image is real or generated. Studies [12] show that incorporating classification loss in discriminator along with the real/fake loss, in turn increases the generator's capability to produce sharp images with high details. The Discriminator consists of stacked Conv-ReLU-Pool layer and the final fully connected layer has been modified to incorporate both binary loss and classification loss (Fig. 3(a)).

3.4 Data Augmentation

1. **Image Interpolation:** The input images have been resized to 256×256 before passing through ResNet. Market-1501 images (128×64) are resized to 256×128, and zero-padded to make 256×256. The images in DeepFashion are of the desired dimension by default.
2. **Random Erasing** [6]**:** Random erasing is helpful in achieving robustness against occlusion. A random patch of the input image is given random values while the reconstruction is expected to be perfect. Thus, the GAN learns to reconstruct (and remove) the occluded regions in the generated images.

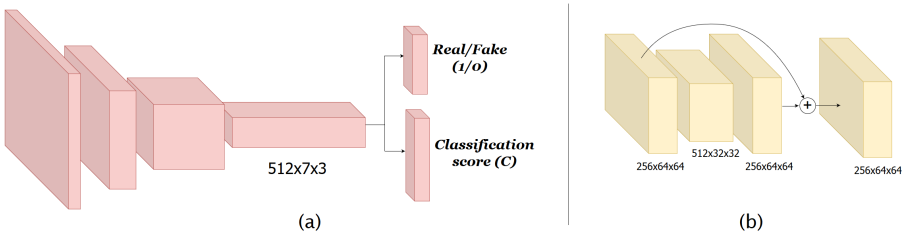

Fig. 3. (a) Architecture of the discriminator of pt-GAN. A classification task is added with the real/fake prediction. This simultaneously helps the Generator to produce more realistic images. **(b)** Architecture of the Residual Blocks used in the Generator. The Residual Learning strategy preserves low level features (color, texture) and learns high level features (pose) simultaneously.

3. **Random Crop:** The input image is randomly cropped and upscaled to the input dimension (256×256) to augment the cases where the human detection is inaccurate or only the partial body is visible.
4. **Jitter:** We use random jitter in terms of brightness, Contrast, Hue and Saturation (random jitter to each channel) to augment the effects of illumination variations.
5. **Random Horizontal Flip:** Inspecting the dataset, it is seen that most human subjects has both left-right profile images. Hence flipping the image left-right is a good choice for image augmentation.
6. **Random distortion:** We have incorporated random distortion with a grid size of 10, to compensate the distortion in the generated image as well as enforce our model to learn important features of the input image even in the presence of non-idealities.

Fig. 4. The data augmentation techniques used in this work: (a) Original image, (b) Random erasing, (c) Random crop, (d) Random distortion, (e)–(g) Random jitter: (e) Brightness, (f) Contrast, (g) Saturation; (h) Random Flip

The CNN by itself enforces scale invariance through max-pooling and convolution layers. Thereafter we claim to have achieved invariance from distortion, occlusion, illumination and scale. A demonstration of the data augmentation techniques is shown in Fig. 4.

4 Experiments

4.1 Datasets

DeepFashion: The DeepFashion (In-shop Clothes Retrieval Benchmark) dataset [13] consists of 52,712 in-shop clothes images, and 200,000 cross-pose/scale pairs. The images are of 256×256 resolution. We follow the standard split adopted by [14] to construct the training set of 146,680 pairs each composed of two images of the same person but different poses.

Market-1501: We also show our results on the re-identification dataset Market-1501 [23] containing 32,668 images of 1,501 persons. The images vary highly in pose, illumination, viewpoint and background in this dataset, which makes the person image generation task more challenging. Images have size 128×64. Again, we follow [14] to construct the training set of 439,420 pairs, each composed of two images of the same person but different poses.

4.2 Implementation and Training

For image descriptor generation, We have used a pretrained ResNet-50 network whose weights were not updated during the training of the generator and discrimiator. The input image and the target image are of the same class with different poses. The reconstruction loss (MSE) is incorporated with the negative discriminator loss to update the Generator. In our implementation, we have used 9 Residual blocks sequentially in the generator architecture. The discriminator is trained on the combined loss (binary crossentropy and categorical crossentropy).

The architecture of the proposed model is described in detail in Sect. 3. For training the Generator as well as Discriminator we have used Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The initial learning rate was set to 0.0002 with a decay factor 10 at every 20 epoch. A batch size of 32 is taken as standard.

5 Results and Discussion

5.1 Qualitative Results

We demonstrate a series of results in high resolution fashion dataset DeepFashion [13] as well as a low resolution re-identification dataset Market-1501 [23]. In both the datasets, by visual inspection, we can say that our model performs good reconstruction and is able to learn invariable information like the colour and texture of clothing, characteristics of make/female attributes such as hair

and face while successfully performing image transformation into the desired pose. The results on DeepFashion is better due to good details and simple background, whereas the low resolution affects the quality of the generated images in Market-1501. The results are demonstrated in Fig. 5.

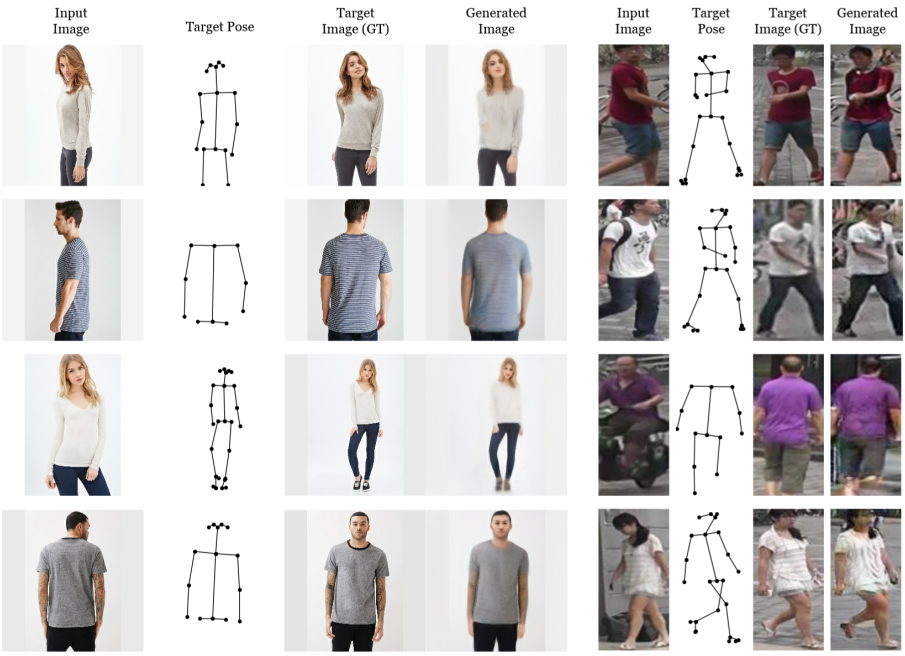

Fig. 5. Qualitative results on DeepFashion and Market-1501 datasets. The proposed model is able to reproduce good details, and also learn invariable information like the colour and texture of clothing, characteristics of make/female attributes such as hair and face while successfully perform image transformation into the desired pose.

5.2 Quantitative Results

We use two popular measures of GAN performance, namely Structural Similarity (SSIM) [21] and Inception Score (IS) [19] for verifying the performance of our model. We compare our work with the already existing methods based on SSIM and IS scores on both DeepFashion and Market-1501 datasets in Table 1. Our model achieves the best IS score in Market-1501 dataset while achieving second best results in SSIM score in both the datasets. However, the deviation from the state-of-the-art is ∼1.5% in these cases which can be overcome through rigorous testing and hyperparameter tuning. We also inspect the improvement incorporated by data augmentation as seen in Table 1. The proposed augmentation methods give an average improvement of ∼9%. This essentially strengthens our argument that a significant boost in performance can be gained by exploring effective training schemes, without changing the model parameters or loss function.

Table 1. Comparative study with existing methods in DeepFashion and Market-1501 datasets. The best and second best results are denoted in red and blue respectively.

	DeepFashion		Market-1501	
Model	SSIM	IS	SSIM	IS
pix2pix [10]	0.692	3.249	0.183	2.678
PG2 [14]	0.762	3.090	0.253	3.460
DSCF [20]	0.761	3.351	0.290	3.185
BodyROI7 [15]	0.614	3.228	0.099	3.483
Dong et al. [4]	0.793	3.314	0.356	3.409
Ours w/o augmentation	0.713	3.006	0.268	3.425
Ours (full)	0.781	3.238	0.302	3.488

5.3 Failure Cases

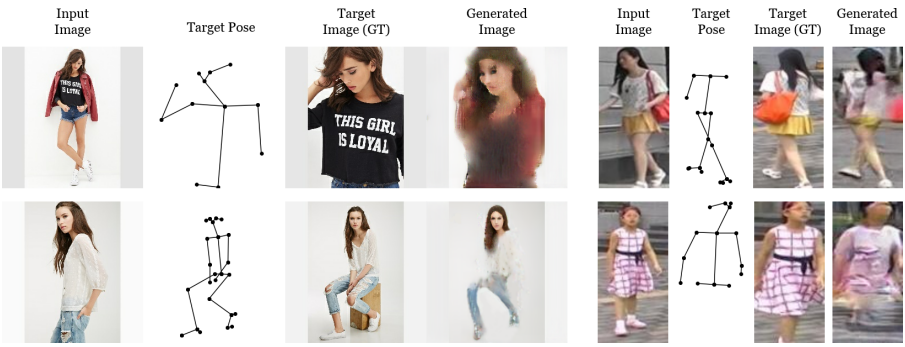

Fig. 6. Failure Cases in our pt-GAN model. If the input contains fine details (text, stripes) or the target pose is incomplete then the reconstruction is poor. The external attribute (handbag) learning is of limited success.

We analyse some of our failure cases in both the datasets to understand the shortcoming of our model. As seen in Fig. 6, the text in clothing as well as very fine patterns of clothing (stripes, dots) are not modelled properly. The external attribute features (e.g. the handbag in Fig. 6) are not learned properly as it is difficult to map external attributes to the output image when conditioned only on pose. The accuracy is also dependent on the completeness of the target pose. Finally, there is some limitation in cases where a rare complex pose is presented which has scarce training data. In Market-1501, the reconstruction of faces is not very good due to poor resolution.

5.4 Further Analysis

Along with the quantitative and qualitative results, we demonstrate a special case to show the improvement caused by data augmentation methods. As seen in Fig. 7, the occlusion in the input image is partially carried forward when the data augmentation methods are not used. With data augmentation the generated image is better in quality and the artifacts generated in the edges are less.

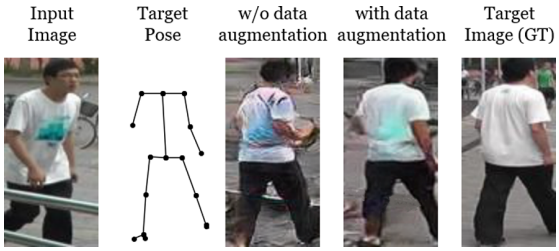

Fig. 7. Occlusion invariance using the proposed model. The occlusion is partially carried forward when data augmentation methods are not used. With data augmentation, the resultant image is free from the artifacts.

6 Conclusion

In this work, we proposed an improved end-to-end pose transformation model to synthesize photorealistic images of a given person in any desired pose without any external feature learning. We make use of the residual learning strategy with effective data augmentation techniques to achieve robustness in terms of occlusion, scale, illumination and distortion. For future work, we plan to achieve better results by utilising feature transport from the source image and conditioning the discriminator on both source image and target pose, alongwith using a perceptual (content) loss for reconstruction.

References

1. Balakrishnan, G., Zhao, A., Dalca, A.V., Durand, F., Guttag, J.: Synthesizing images of humans in unseen poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8340–8348 (2018)
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)

4. Dong, H., Liang, X., Gong, K., Lai, H., Zhu, J., Yin, J.: Soft-gated warping-GAN for pose-guided person image synthesis. In: Advances in Neural Information Processing Systems, pp. 474–484 (2018)
5. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
6. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)
7. Ge, Y., et al.: FD-GAN: pose-guided feature distilling GAN for robust person re-identification. In: Advances in Neural Information Processing Systems, pp. 1222–1233 (2018)
8. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
11. Karmakar, A., Mishra, D., Tej, A.: Stellar cluster detection using GMM with deep variational autoencoder. In: 2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS), pp. 122–126. IEEE (2018)
12. Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., Hu, J.: Pose transferrable person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4099–4108 (2018)
13. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1096–1104 (2016)
14. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: Advances in Neural Information Processing Systems, pp. 406–416 (2017)
15. Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., Fritz, M.: Disentangled person image generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 99–108 (2018)
16. Qian, X., et al.: Pose-normalized image generation for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 650–667 (2018)
17. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
19. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, pp. 2234–2242 (2016)
20. Siarohin, A., Sangineto, E., Lathuilière, S., Sebe, N.: Deformable GANs for pose-based human image generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3408–3416 (2018)

21. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
22. Yan, Y., Xu, J., Ni, B., Zhang, W., Yang, X.: Skeleton-aided articulated motion generation. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 199–207. ACM (2017)
23. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124 (2015)