TOWARDS USING CLOTHES STYLE TRANSFER FOR SCENARIO-AWARE PERSON VIDEO GENERATION

Jingning Xu[1,2], Benlai Tang[2], Mingjie Wang[3], Siyuan Bian[2,4], Wenyi Guo[2], Xiang Yin[2], Zejun Ma[2],

[1]School of Software Engineering, Tongji University, Shanghai, China
[2]ByteDance AI Lab
[3]School of Computer Science, University of Guelph, ON, Canada
[4]Department of Computer Science and Engineering, Shanghai Jiao Tong University, China

ABSTRACT

Clothes style transfer for person video generation is a challenging task, due to drastic variations of intra-person appearance and video scenarios. To tackle this problem, most recent AdaIN-based architectures are proposed to extract clothes and scenario features for generation. However, these approaches suffer from being short of fine-grained details and are prone to distort the origin person. To further improve the generation performance, we propose a novel framework with disentangled multi-branch encoders and a shared decoder. Moreover, to pursue the strong video spatio-temporal consistency, an inner-frame discriminator is delicately designed with input being cross-frame difference. Besides, the proposed framework possesses the property of scenario adaptation. Extensive experiments on the TEDXPeople benchmark demonstrate the superiority of our method over state-of-the-art approaches in terms of image quality and video coherence.

Index Terms— Clothes Style Transfer, Person Video, Scenario-aware, Disentangled Encoders, Video Consistency

1. INTRODUCTION

Inspired by the huge success of Deep Neural Networks, during this decade, the realm of *Clothes Style Transfer* has attracted intensive attention from the research community, which aims to generate realistic person scenes satisfying the demands of diverse attributes [1, 2, 3]. Due to alluring application prospects in controllable person manipulation, virtual try-on clothes texture editing, etc., lots of efforts have been made to improve its performance [2, 3, 4, 5]. More importantly, video format is a finer carrier for delivering visual experience than static images in many tasks (e.g. person pose animation [4]). However, the video-based person generation has not been investigated in depth yet by far [6].

Existing approaches for clothes style transfer can be roughly divided into two classes: warping-based [2, 3, 7, 8, 9, 10] and image generation-based models [1, 4, 5]. The former type of methods benefit from preserving informative and detailed features of the clothes. Zhang et al. [3] attempt to estimate a thin plate spline (TPS) transformation for warping the clothing items and refine the coarse results by using a composition mask. Meanwhile, and Yu et al. [2] devise a body segmentation map prediction module to delineate body parts and warped clothing regions. Despite the constant progress achieved by them, it is still difficult to directly transform the spatially misaligned body parts into the desired shapes due to the non-rigid nature of human body. Hence, these approaches are unable to attain satisfactory performance. To mitigate this issue, several approaches [7, 8, 9, 10] probe into the paradigm of transferring clothes from one person to another by estimating a complicated 3D human mesh and warping the textures to cater for the body topology. Nevertheless, these approaches fail to capture the sophisticated interplay of the intrinsic shape and appearance, and thus results in unrealistic synthesis with deformed textures.

Profiting from the invention of GANs [11, 12, 13], image generation-based approaches have prevailed in image synthesis task, such as [1, 4, 5, 14, 15, 16, 17, 18]. A core driving force behind these methods is to impose the pose guidance on the image generation process. Hence, the photo-realistic people images with arbitrary poses are rendered, whereas the clothes styles are controlled simultaneously by utilizing various valuable human attributes. Early models [14, 15, 16, 17] only focus on keeping the posture and identity without providing user manipulation of human attributes, e.g. head, pants and upper clothes, whereas several recent works [1, 4, 5] treat the clothes as a type of texture style and try to encode it. Especially, thanks to the high flexibility and strong generalization ability to other clothes styles [5, 19], the *de facto standard of* Adaptive Instance Normalization (AdaIN) and its variants [20] have witnessed remarkable improvements in extracting appearance features [1, 5]. While obtaining acceptable performance, they have obvious drawbacks: they can often only perform coarse-grained feature extraction, and cannot restore fine-grained details well, which easily causes blurry clothes and the shortage of identity information. In the field of video-based generation, existing methods formulate this problem as a generation process of image sequence and produce results in a frame-by-frame manner. For example, resorting to a set of
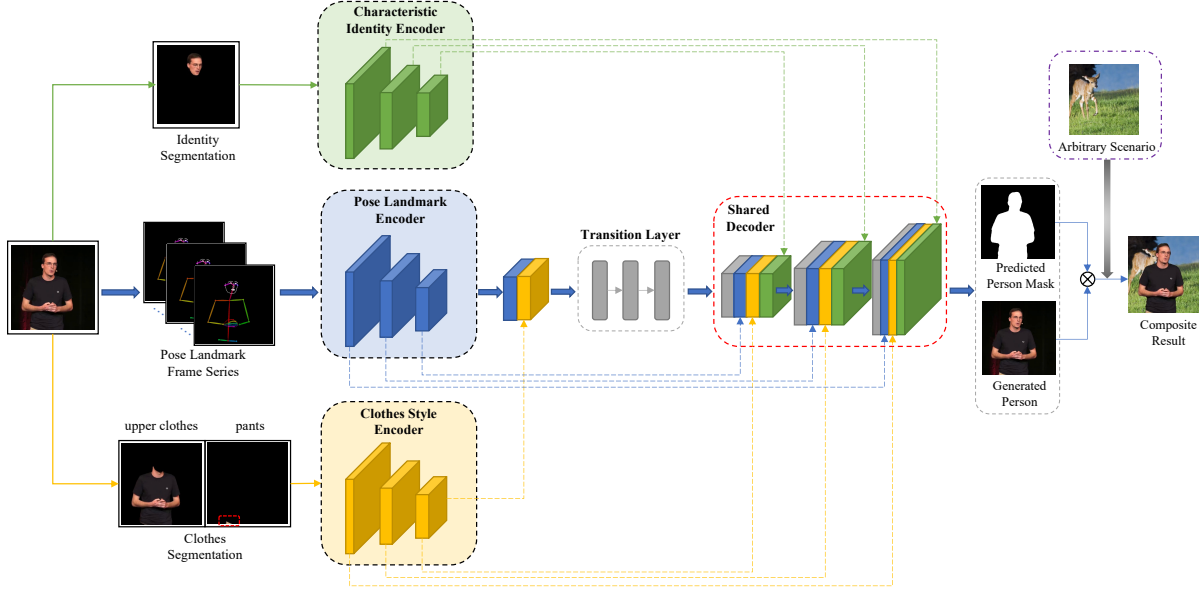
Fig. 1: The overview of the proposed framework towards using clothes style transfer for scenario-aware video person generation, which is composed of three disentangled encoders (including Pose Landmark Encoder, Characteristic Identity Encoder and Clothes Style Encoder), and a cascaded Shared Decoder. First, these three-path encoders collaboratively represent diverse sources disentangled from the original raw input. Then the separable features are aggregated and fed into the high-level decoder to generate the final output.

sparse trajectories, studies [21] and [22] leverage zeroth-order and first-order Taylor expansions such that the complex transformations are approximated. However, they rarely consider the video spatio-temporal coherence and thus affects the consistency of movements. Albeit the relatively smooth videos, a sequential generator proposed by Wang et al. [13] is solely applied in tasks of single-person pose transfer.

Such limitations and drawbacks drive our exploration of clothes style transfer for person video generation to improve the quality and consistency of the generated video. In this paper, we propose a novel framework for person video generation towards using clothes style transfer. It consists of four components: *Pose Landmark Encoder, Characteristic Identity Encoder, Clothes Style Encoder* and a *Shared Decoder*, as illustrated in Figure 1. In specific, we formulate the task of feature learning as three disentangled sub-tasks to avoid the learning conflicts existing in the single-column structures. The disentangled multi-branch scheme is inspired by the natural insight that different salient areas correspond to varying degrees of responses to multiple tasks [23]. Furthermore, to guarantee the spatio-temporal consistency in generated videos and implicitly capture the local frames dependencies, we delicately devise a variant of discriminator, named as *inner-frame discriminator*, which takes the cross-frame difference as input. Besides, it is worth noting that background/scenarios replacement has also been a new trend for video person generation since it has higher values of privacy and entertainment [24]. Unfortunately, existing

approaches [1, 5] are scenario-agnostic. Motivated by this observation, we design a strategy to expand our framework's capability of being scenario-aware and improve the flexibility of video generation. To our knowledge, this method provides the first investigation into spatio-temporal consistency and scenario awareness simultaneously for person video generation with clothes style transfer. Extensive experiments are conducted on TEDXPeople benchmark [25] and demonstrate that our proposed method outperforms existing popular approaches in terms of image quality, similarity to realistic images and video smoothness, while embracing the merit of scenario awareness for matting.

2. METHODOLOGY

2.1. Disentangled Multi-branch Encoders

To prevent the model from suffering from learning conflicts often occurring in single-column structure and capture fine-grained cues for the downstream decoder, we disentangle different spatial information from the main input and construct a multi-branch structure, which takes the pose landmark images I_p^t, characteristic identity images I_i^t and clothes style images I_c as input. Note the pose landmark images I_p^t are obtained using a commonly-used human pose estimation method proposed by [26]. Following previous works [1, 4, 5], we build our **pose landmark encoder** upon a three-layer down-sampling network to encode the pose frame sequence. The **characteristic identity encoder** aims to boost the identity information and get rid of possible dis-

tortion, especially in the face region. Specifically, the identity images I_i^t are disentangled by leveraging a human segmentation method in [27]. Unlike the pose landmark encoder with demands of extracting more abundant and sophisticated features, the identity encoder plays a role of providing weak cues for identity information, and thus we narrow the network width to half of the pose encoder while reducing the number of parameters. For the **clothes style encoder**, the same segmentor [27] is also employed to generate the clothes input images I_c, including upper clothes and pants (see red-dotted region in Figure 1). Besides, during the training phrase, a stochastic regularizer of geometric transformation is adopted to strengthen the model's generalization ability to arbitrary clothes input. In addition, to further increase model's capacity of learning representations, the combination of spatial and channel attention mechanisms [28] is involved in each layer of encoders to adaptively emphasize the informative filters or regions by softly recalibrating corresponding units.

2.2. Shared Decoder with Scenario Awareness

After flowing through multi-branch encoders in parallel, the disentangled sources are converted into the fine-grained feature representations. Then, these features are integrated and further processed in the high-level shared decoder. It is worth noting that skip connections are used here to fuse multi-level features while avoiding the issue of gradient vanishing [29]. To let the framework possess higher flexibility of arbitrarily changing scenarios, we design an auxiliary output branch to predict the scenario mask α. The final composition result can be formulated as $I_{composite} = \alpha I_{out} + (1 - \alpha)I_{bg}$, where I_{out} and α denote the generated frame and the alpha matte for new scenarios, respectively, whereas I_{bg} represents the arbitrary scenario image.

2.3. Inner-frame Discriminator

In general, two kinds of discriminators (D_P and D_I) are prevailing in current algorithms for pose-guided person image generation, such as [5, 17, 30]. In detail, D_P is leveraged to guarantee the consistency between the pose of generated image $\hat{I}_s^t$, and the target pose I_P^t, while D_I concentrates on the quality and reality of the generated images. To make full use of merits of both discriminators, both of them are adopted in our framework.

As described in Section 1, spatio-temporal consistency of generated videos is short of being considered in existing approaches. To alleviate this issues, we devise a ***inner-frame discriminator*** D_{if} in parallel with D_P and D_I, which takes the cross-frame difference as the input and contributes to better video smoothness. In other words, it operates as an strong constraint imposed on adjacent frames to suppress suddenly drastic changes. Formally, given two adjacent frames $\hat{I}_s^t$ and $\hat{I}_s^{t+1}$ in the same video, the inner-frame discriminator can be denoted as: $D_{if}(\hat{I}_s^{t+1} - \hat{I}_s^t)$. Besides, the sequence discriminator D_t in [4, 13] is also incorporated into our method to enhance the visual quality of the generated videos since
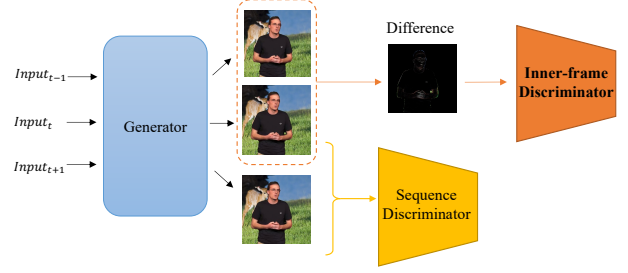

Fig. 2: The overview of the proposed inner-frame discriminator for spatio-temporal consistency in generated videos, which is accompanied by the sequence discriminator.

it enables the framework to implicitly consider the inner-relationships among frames. Figure 2 delineates the designed scheme of these two discriminators.

2.4. Objective Function

The entire training objective is to supervise the optimization of our proposed framework for the sake of high-quality person videos generation conforming to the input clothes style as well as the identity. Additionally, the posture of the generated person should be compatible with that of input pose images. The elaborate loss function of the whole network is formulated as:

$$L = \lambda_{adv}L_{adv} + \lambda_r L_r + \lambda_s L_s + \lambda_v L_v + \lambda_l L_l \quad (1)$$

where $\lambda_{adv}, \lambda_r, \lambda_s, \lambda_v, \lambda_l$ are the weights which are used to control the influence of different loss items. L_{adv} indicates the adversarial loss implemented by the LSGAN [31] loss. Following [32], the reconstruction loss L_r provides the supervision to impel the generated frames to be close to the ground truths following [32] and [33] while style loss L_s aims to maintain style coherence. More importantly, the video loss L_v is achieved by discriminators D_t and D_{if} illustrated in Section 2 to facilitate the video quality, especially for spatio-temporal coherence. Finally, it is worth noting that the truthfulness degree and identity preservation of generated images heavily rely on the synthesis quality of a family of key positions, such as head and hands parts. Therefore, to aid in improving comprehensive generation quality, we introduce a type of local characteristic discriminator which takes the head/hands segmentation results as input and computes the local characteristic loss term L_l.

3. EXPERIMENTS

3.1. Dataset and Implementation Details

We evaluate our method on the TEDXPeople dataset provided in [25]. This dataset contains 48,188 videos of TED and TED-X talks. We randomly select 5,000 videos from the entire dataset with a resolution of 256×256. 85% of the dataset is used as training data. The rest is used to validate the performance of clothes style transfer during the inference phase.
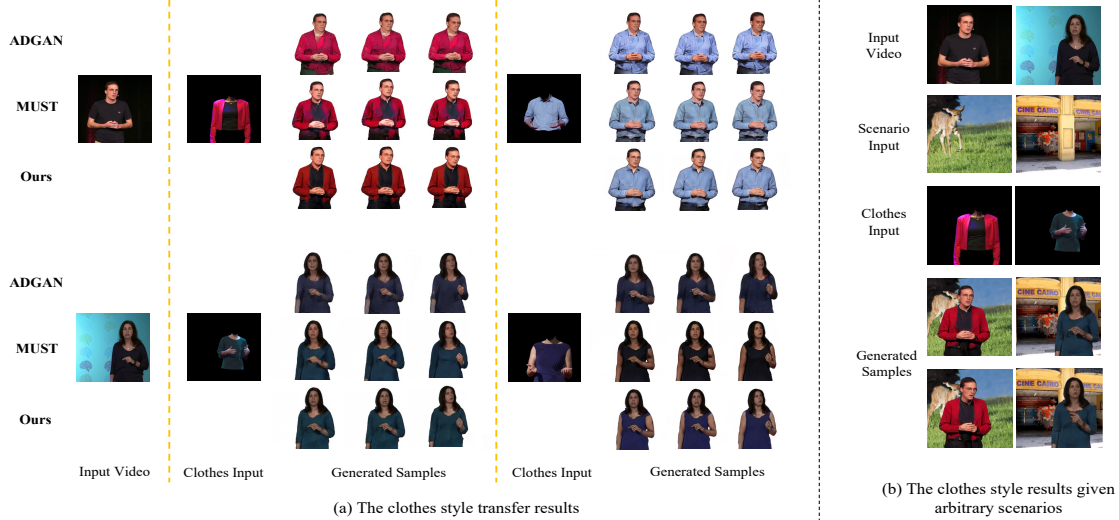
(a) The clothes style transfer results

(b) The clothes style results given arbitrary scenarios

Fig. 3: The visualization examples generated by our method on TEDXPeople dataset.

Besides, images of ImageNet dataset [34] are used as random scenario inputs during training process. The weights for the loss terms are set to $\lambda_{adv} = 1, \lambda_r = 5, \lambda_s = 50, \lambda_v = 1.5, \lambda_l = 2$. We adopt Adam optimizer [35] with the momentum set to 0.5 to train our model for around 100k iterations.

3.2. Comparison with SOTA Models

In our experiments, Structured similarity (SSIM) [36] and Peak Signal-to-Noise Ratio (PSNR) are employed to measure the similarity between the generated images and the ground truths, whereas Fréchet Inception Distance (FID) [37] is used to assess the realism and consistency of generated images [1, 4]. We evaluate the spatio-temporal consistency of the generated videos via the Fréchet Video Distance (FVD) [38]. We compare our method with two existing popular approaches MUST [1] and ADGAN[5]. The quantitative results are demonstrated in Table 1 while the visualization examples are also provided in Figure 3(a). Both quantitative results and visualization examples demonstrate the superiority of our proposed framework. Meanwhile, we further validate the scenario awareness of our framework and several example results are given in Figure 3(b). As it can be observed, the identity of the generated videos remain the same as the source persons, and our approach can manipulate clothes styles and background scenarios flexibly. Video demos are available at https://github.com/XSimba123/demos-of-csf-sa/

3.3. Ablation Study

To better understand each component in our model, we conduct ablation studies on TEDXPeople benchmark (see Table 2), which demonstrates that the multiple encoders structure (I) and attention module (II) have a great contribution to visual quality. The FVD metric shows that two spatio-temporal constraints (III) brings a significant improvement to

Method	SSIM ↑	PSNR ↑	FID ↓	FVD ↓
ADGAN [5]	0.787	19.0	19.1	0.732
MUST [1]	0.794	19.3	11.7	0.804
Our approach	**0.841**	**23.9**	**11.3**	**0.261**

Table 1: Comparison with SOTA models on TEDXPeople. The proposed method consistently surpasses the existing approaches on all metrics.

video consistency. And local characteristic discriminator (IV) also gives a boost to the similarity metrics.

I	II	III	IV	SSIM ↑	PSNR ↑	FID ↓	FVD ↓
				0.801	20.5	16.7	0.643
✓				0.804	21.5	14.6	0.498
✓	✓			0.818	21.8	12.8	0.427
✓	✓	✓		0.826	22.3	12.9	0.282
✓	✓	✓	✓	**0.841**	**23.9**	**11.3**	**0.261**

Table 2: Ablation on disentangled multi-branch encoders (I), attention mechanism (II), spatio-temporal constraints (III) and local characteristic loss (IV).

4. CONCLUSION

In this work, we propose an approach to tackle scenario-aware clothes style transfer for person videos, which is shown to outperform baseline and recent approaches. This is because our model uses a multi-encoder framework to handle different types of features. And an inner-frame discriminator is proposed to greatly enhance the video temporal coherence. Moreover, we design a framework to support scenario-aware video generation. Experimental results on TEDXPeople dataset confirmed the effectiveness and robustness of our proposed framework.

5. REFERENCES

[1] T. Ma, B. Peng, W. Wang, and J. Dong, "Must-gan: Multi-level statistics transfer for self-driven person image generation," in *CVPR*, 2021.

[2] R. Yu, X. Wang, and X. Xie, "Vtnfp: An image-based virtual try-on network with body and clothing feature preservation," in *ICCV*, 2019.

[3] X Han, Z. Wu, W. Zhe, and R. Yu, "Viton: An image-based virtual try-on network," in *CVPR*, 2018.

[4] Y. Ren, G. Li, S. Liu, and T. H. Li, "Deep spatial transformation for pose-guided person image generation and animation," *IEEE Transactions on Image Processing*, vol. 29, pp. 8622–8635, 2020.

[5] Y. Men, Y. Mao, Y. Jiang, W. Y. Ma, and Z. Lian, "Controllable person image synthesis with attribute-decomposed gan," in *CVPR*, 2020.

[6] P. Zablotskaia, A. Siarohin, B. Zhao, and L. Sigal, "Dwnet: Dense warp-based network for pose-guided human video generation," *arXiv:1910.09139*, 2019.

[7] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo, "Towards photo-realistic virtual try-on by adaptively generating-preserving image content," in *CVPR*, 2020.

[8] M. Zanfir, A. I. Popa, A. Zanfir, and C. Sminchisescu, "Human appearance transfer," in *CVPR*, 2018.

[9] Gerard Pons-Moll, Sergi Pujades, H. U. Sonny, and Michael J. Black, "Clothcap: seamless 4d clothing capture and retargeting," *Acm Transactions on Graphics*, vol. 36, no. 4CD, pp. 1–15, 2017.

[10] S. Yang, T. Ambert, Z. Pan, K. Wang, L. Yu, T. Berg, and M. Lin, "Detailed garment recovery from a single-view image," *arXiv:1608.01250*, 2016.

[11] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[12] T. C. Wang, M. Y. Liu, J. Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *CVPR*, 2018.

[13] T. C. Wang, M. Y. Liu, A. Tao, G. Liu, and J Kautz., "Few-shot video-to-video synthesis," *arXiv:1910.12713*, 2019.

[14] C. Chan, S. Ginosar, T. Zhou, and Alexei A Efros, "Everybody dance now," *rippleffect studio limited*, 2018.

[15] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, "Deformable gans for pose-based human image generation," in *CVPR*, 2018.

[16] P. Esser and E. Sutter, "A variational u-net for conditional appearance and shape generation," in *CVPR*, 2018.

[17] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *CVPR*, 2019.

[18] T. Xie, L. Liao, C. Bi, B. Tang, X. Yin, J. Yang, M. Wang, J. Yao, Y. Zhang, and Z. Ma, "Towards realistic visual dubbing with heterogeneous sources," in *ACM MM*, 2021.

[19] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019.

[20] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017.

[21] A. Siarohin, S. Lathuiliere, S. Tulyakov, E. Ricci, and N. Sebe, "Animating arbitrary objects via deep motion transfer," in *CVPR*, 2019.

[22] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[23] G. Song, Y Liu, and X. Wang, "Revisiting the sibling head in object detector," in *CVPR*, 2020.

[24] S. Lin, A. Ryabtsev, S. Sengupta, B. Curless, and I. Kemelmacher-Shlizerman, "Real-time high-resolution background matting," in *CVPR*, 2021.

[25] A. Grigorev, K. Iskakov, A. Ianina, R. Bashirov, I. Zakharkin, A. Vakhitov, and V. Lempitsky, "Stylepeople: A generative model of fullbody human avatars," in *CVPR*, 2021.

[26] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *PAMI*, 2019.

[27] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018.

[28] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[30] A. Pumarola, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "Unsupervised person image synthesis in arbitrary poses," in *CVPR*, 2018.

[31] X. Mao, Q. Li, H. Xie, Ryk Lau, and S. P. Smolley, "Least squares generative adversarial networks," in *ICCV*, 2017.

[32] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *Springer, Cham*, 2016.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[35] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[36] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans Image Process*, vol. 13, no. 4, 2004.

[37] 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," *Advances in neural information processing systems*, vol. 30, 2017.

[38] T. Unterthiner, S Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv:1812.01717*, 2018.