DISENTANGLED HUMAN ACTION VIDEO GENERATION VIA DECOUPLED LEARNING

Lingbo Yang[1]*, Zhenghui Zhao*[2]*, Shiqi Wang*[3]*, Shanshe Wang*[1]*, Siwei Ma*[1]*, Wen Gao*[1]

[1]Institute of Digital Media, School of EECS, Peking University, Beijing 100871, China
[2]LMAM, School of Mathematical Sciences, Peking University, Beijing 100871, China
[3]Department of Computer Science, City University of Hong Kong
{lingbo,zhzhao,sswang,swma,wgao}@pku.edu.cn, shiqwang@cityu.edu.hk

ABSTRACT

Recently there has been remarkable progress in synthesizing realistic human action videos by directly learning to translate pose heatmaps/stick figures to video frames in an end-to-end fashion. However, such models are not suitable for fashion-related applications that typically require flexible manipulations of visual attributes, such as the color of clothes. In this paper, we propose a disentangled human video generation framework conditioned on both the pose sequence and encoded color attributes. We aim to learn an encoder that captures the manifold structure of latent color space and a generator that fully utilizes the encoded color attributes to produce diversely-colored human action videos. To this end, we design a two-stage decoupled learning approach that uses a pre-trained color-aware encoder to guide the disentangled learning of the generator. Furthermore, a color augmentation approach is applied on raw video clips to better shape the distribution of samples in the latent color space. Comprehensive experimental results demonstrate the efficacy of our proposed methods.

Index Terms— Pose-guided video generation, decoupled learning, generative adversarial networks (GANs), feature disentanglement

1. INTRODUCTION

Human action video generation has attracted increasing attention due to its great potential in many application scenarios, such as creative movie making, interactive fashion design, and dataset augmentation for video analysis tasks. Existing works typically tackle the problem with deep generative models conditioned on representations of human action dynamics, such as semantic labels [1], natural languages [2], and human body segmentation[3]. Among various forms of representations, skeleton-based human pose has several advantages. First, the skeleton can be directly manipulated at each joint to accurately model complex human actions. Second, the skeleton is independent of human appearance, allowing the motion dynamics and visual attributes to be modeled in a disentangled manner. Third, the skeleton can be accurately estimated with state-of-the-art pose-estimation networks. As a result, pose-guided human action generation has become the mainstream, and remarkable progress has been made so far.

As the crucial foundation of video generation, pose-guided human image generation problem has attracted numerous attention. Ma *et al.* first investigated the problem in [4] and proposed a two-stage approach to synthesize images of a person under novel poses, where the person's appearance is conditioned with another input image. In [5], Ma *et al.* further disentangle the foreground, background and pose features, allowing each part of the generated image to be modeled separately. Other works focus on warping the textures from source image to target image based on pose correspondences. Guha *et al.* [6] proposed a warping-based framework that segments the human body into rigid parts before using a piece-wise affine transformation to warp the foreground texture of each part to target locations. Siarohin *et al.* [7] proposed to transfer convolutional feature maps from different body parts through deformable skip-connections. Although significant advances have been made in generating a single image under the new pose, it is often non-trivial to extend such works to video generation, mainly due to the intrinsic complexity of human action dynamics.

Recently, there has been remarkable breakthrough in generating high-resolution human dancing videos by modeling the task as an image-to-image (I2I) translation problem [8], where the generator learns to directly translate pose heatmaps/stick figures to video frames through an end-to-end framework. Wang *et al.* [3] proposed a video-to-video synthesis framework to generate realistic dancing videos from concatenated pose features including both skeletons and body masks. Chan *et al.* [9] introduced a post-processing residual network to further enhance facial details of generated video

This work was supported in part by the National Natural Science Foundation of China (6163200161571017)National Basic Research Program of China (973 Program, 2015CB351800), and High-performance Computing Platform of Peking University, which are gratefully acknowledged.

frames. However, most I2I-based human video generation methods fail to condition on human visual attributes explicitly, and can only replicate the content of training samples. In consequence, these methods are generally not suitable for real-world applications where flexible manipulations of human visual attributes are required, such as fashion design or interactive clothes editing.

Another closely related area is the conditional image generation where a deep encoder is introduced to extract latent codes that guide the generator to map a given input to diverse output images. Bao *et al.* [10] proposed cVAE-GAN for fine-grained image generation by learning the connection from latent encoding space to real image space, and used the VAE learning objective to prevent mode collapse. Zhu *et. al.* further proposed BicycleGAN [11], where a cycle consistency loss is introduced to encourage the mapping from latent codes to output images to be invertible. However, it can be speculated that extending the end-to-end training approach for video generation could still suffer from mode-collapse, mainly due to the lack of diversity in visual attributes of video datasets. In particular, the latent code distribution of training video frames would be highly sparse and concentrated, since different frames of the same person have similar visual appearance, leading to closely clustered latent codes. In consequence, it would be difficult for the encoder to properly capture the underlying structure of latent color space through such sparse samples, and the discriminative power of latent codes would be severely impaired. This forces the generator to rely on only pose-related information to synthesize the correct visual detail of the desired person, making the whole framework collapse and degenerate back to the “one-to-one” case.

In this paper, we propose a two-stage decoupled learning approach for disentangled human action video generation, where the pose and color attributes can be independently controlled. During the first stage, we train the encoder on individual frames with the goal of learning to capture the manifold structure of latent color space. During the second stage, the pre-trained encoder is used to provide color-aware attribute codes to guide the generator to map input pose feature map to realistic frames with diverse colors. To better facilitate the training and prevent mode-collapse, we introduce a color augmentation pre-processing step to better “spread out” the training samples in the latent color space, which greatly helps the learning process. Experimental results have demonstrated the efficacy of our proposed scheme.

The rest of the paper is organized as follows. In section 2, we introduce the pose-guided feature map estimation pipeline. The two-stage training approach of video generator is detailed in section 3. Section 4 shows our experimental results to demonstrate the effectiveness of our proposed framework. Finally, we conclude this paper in section 5.

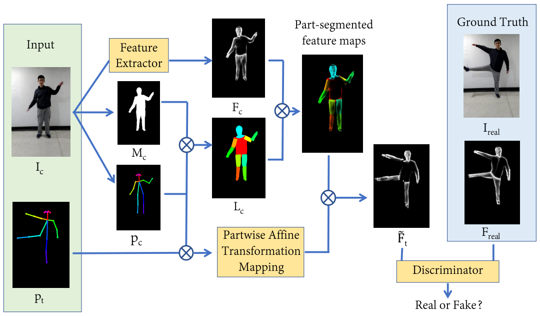


Fig. 1. The diagram of the proposed pose-guided feature map estimation pipeline.

2. POSE-GUIDED FEATURE MAP ESTIMATION

In this section, we detail our feature map estimation pipeline that calculates the pose-dependent feature maps $\mathbf{F}(\mathbf{P}, I_c) = \{F_i\}_{i=1}^N$ conditioned on the target pose sequences $\mathbf{P} = \{p_i\}_{i=1}^N$ and the input content image I_c. The whole pipeline consists of three steps: feature extraction, body parts segmentation and feature warping, as shown in Fig. 1.

2.1. Feature extraction

We use a combination of pose stick figures and deep edge maps as our pose-dependent feature maps to capture both the body configurations and fine-grained visual details of the content image. We adopt Openpose [12] to compute the coordinates of $n = 18$ human body joints, and render the corresponding pose stick figure by plotting body joints and drawing lines between joints according to human skeleton structure. For boundary map estimation, we use a pretrained RCF edge-detector [13] to obtain a gray-scale edge map. The final combined feature map F contains 4 channels (3 for RGB pose stick figures and 1 for gray-scale edge maps) and has the same size as the input image.

2.2. Body parts segmentation

The human body can be decomposed into several rigid body parts, allowing the mapping between feature maps to be modeled with a set of 2D affine transformations between each pair of body parts. We follow the convention in [7] to mark human body with 10 labels: head, torso, left/right upper/lower arm/leg. We first estimate the full body mask M_c with a pretrained network [14]. To get accurate labels $L_c = \{L_c^i\}_{i=1}^{10}$ for each part, we use simple geometric shapes to roughly locate the body parts and remove any pixels outside M_c. A colorized body part label maps is illustrated in Fig. 1.

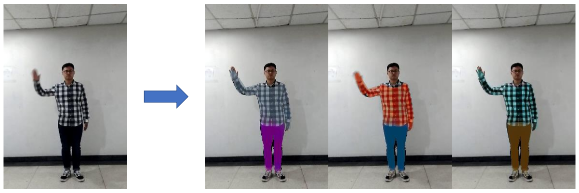


Fig. 2. The original frame and different colorized frames.

2.3. Feature map estimation

Given the extracted feature map F_c of the input image I_c, and the target pose p_t, our goal is to obtain a good estimation of the feature map F_t under p_t. The estimation consists of two steps: part-wise warping and global restoration. First, we calculate an optimal affine transformation T_i for each body part L_c^i with the objective of minimizing the least square distance between corresponding body joints from the source pose to the target pose. Then we warp the source feature map F_c onto the new pose p_t in a part-wise manner and merge the warped feature map parts with max-pooling as in [7]:

$$\tilde{F}_t = \max_{i=1:10}(T_i(F_c \odot L_c^i)) \tag{1}$$

where $\odot$ denotes entry-wise multiplication. To fill small cracks and smooth out any gray level inconsistency at boundaries between body parts, the warped feature map $\tilde{F}_t$ is further refined with an image restoration network. We use the ground truth feature maps F_{real} calculated from real video frames to train the network in an adversarial manner. Experiment results show that the post-process step leads to slightly better results.

3. TWO-STAGE DECOUPLED LEARNING FOR CONDITIONAL VIDEO GENERATION

In this section, we describe the proposed decoupled learning scheme for our conditional video generation framework. The two major modules, color-aware attribute encoder $E_c : I_c \to z_c$, and conditional generator $G : (\mathbf{F}, z_c) \to \mathbf{V}$, are trained separately in two stages. Concretely, we focus on training the encoder on color-augmented video frames to capture the structure of the latent color space in stage I, and use the trained encoder to guide the conditional video generator in stage II. We introduce the details of each training stage below.

3.1. Stage I: Color-aware person attribute encoding

Color augmentation. We adopt a pretrained colorization network [15] to randomly change the color of the person's clothes in the LAB color space, where the global colormap is determined from the AB values at user-designated anchor points. In our implementation, we assign different colors for the upper body and the lower body by controlling the color of body joints inside the respective regions. In addition, we utilize the extracted mask M to keep background contents intact. Fig. 2 shows an example of the original frame and color augmented frames. As can be observed, the colorized frames share the same structural textures as the original frame, allowing us to conveniently use the feature map of the original frame F_c for all corresponding colorized frames.

Encoder training. We train the color-aware attribute encoder with a standard BicycleGAN model [11] that contains two networks: the conditional generator $G : (A, z) \to B$ and the image encoder $E : B \to z$ (here we adopt the notations in [11]). The forward pass contains three steps: image encoding $z = E(B)$, conditional image generation $\tilde{B} = G(A, z)$ and latent regression $\tilde{z} = E(\tilde{B})$. Note that in step 3 we evaluate the latent code on the reconstructed image $\tilde{B}$ instead of the image generated from another random z as in the original BicycleGAN implementation, where the generator needs to run twice for each batch. To control the upper and lower body color independently, we train two separate encoders E_{up} and E_{down} with region-specific color augmentations. When training E_{up} we only colorize the upper body regions, and vice versa. The length of corresponding latent codes z_{up} and z_{down} are set to 16 and 8 respectively, as upper body clothes contain more complex textures (e.g. checkerboard shirt in Fig. 2). The parameter updating rule exactly follows the implementation in BicycleGAN.

3.2. Stage II: Conditional video generation with disentangled features

We use the pretrained color-aware attribute encoders from Stage I to guide the training of the video generator G that conditions on estimated feature maps F and encoded attribute vectors $z_c = [z_{up}, z_{down}]$. Although the U-net generator is commonly used for I2I tasks, the skip connections in U-net could carry unwanted noise and artifacts in estimated feature maps into generated frames. Therefore, we choose the coarse-to-fine residual generator in pix2pixHD [16] to prevent artifacts. The latent code z_c is tiled into a 3-tensor and concatenated with the input feature maps at both coarse and fine stages, where the number of input channels for the first convolution layer is also adjusted accordingly. To ensure temporal consistency, we adopt the training scheme in [9]. For each forward computation, the generator synthesizes two consecutive frames using the previous frame as reference: $v_{sync}^1 = G(F^1, z_c; v_0)$ and $v_{sync}^2 = G(F^2, z_c; v_{sync}^1)$, where v_0 is a pure-black image. The discriminator tries to distinguish between real sequence $V_{gt} = [v_{gt}^1, v_{gt}^2]$ and fake sequence $V_{sync} = [v_{sync}^1, v_{sync}^2]$. In our experiments, we use the conditional discriminator that also takes the feature maps $[F^1, F^2]$ as input. During back propagation, we fix both encoders and update the generator G and discriminator D in an adversarial manner to optimize the following objective:

$$\min_G(\max_D \mathcal{L}_{temp} + \lambda_{FM}\mathcal{L}_{FM} + \lambda_{VGG}\mathcal{L}_{VGG}) \tag{2}$$

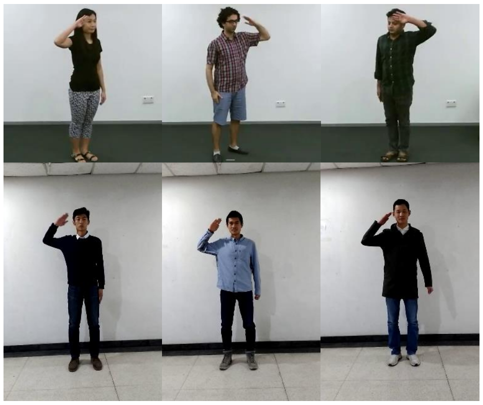


Fig. 3. Snapshots of the “salute” action in NTURGB-D (upper row) and our dataset (bottom row). In comparison, our dataset exhibits better cross-person action consistency.

Here $\mathcal{L}_{temp}$ is the temporal smoothing loss proposed in [9], $\mathcal{L}_{FM}$ is the discriminator feature matching loss in [16], and $\mathcal{L}_{VGG}$ is the perceptual loss [17] that calculates the distance between feature maps extracted with a pretrained VGG-19 network [18]. Both weights are set to 10 in our experiments.

4. EXPERIMENTAL RESULTS

4.1. Video collection

Most existing works in video generation and frame prediction typically resort to action recognition datasets like UCF-101 [19] or 3D sensing datasets like NTURGB-D [20]. However, in order to ensure the robustness of algorithms, such datasets usually contain cluttered background, varying lighting conditions and frequent limb occlusions, which pose great challenges to human video generation tasks. In order to well reflect the performance of the disentangled learning in our approach, we collected a new human action dataset containing 208 video clips of 13 actors performing 16 different actions each, including waving hands, saluting, side-stepping, etc. Each clip contains around 100 frames. The actions are carefully choreographed to avoid heavy limb occlusions and keep the body movements of the same action synchronized across different actors, as illustrated in Fig. 3.

4.2. Implementation Details

Network architecture For the training in Stage I, we use the “U-128” generator network and the “Res256” encoder from the official implementation of BicycleGAN [11]; For Stage II, we use a pix2pixHD [16] generator with 4 global downsample layers and 9 global cascaded residual blocks. The discriminator D is composed of three sub-networks with input downsampling factors $1, 2, 4$, respectively.

Dataset pre-processing We randomly select 10 persons for our experiments. For each person, we use 10 actions for training and the rest for testing. All video frames are center-cropped and resized to 384×256 before feeding into the network. The color augmentation is performed for all video clips at the beginning of both training stages. In Stage I each frame is independently colorized while in stage II all frames within the same video clip are painted with the same color to maintain temporal consistency.

Metrics	MS-SSIM ↑	LPIPS ↓
Deformable GAN	0.8975	0.0706
Proposed	**0.9146**	**0.0590**

Table 1. Objective evaluation score on testing video frames. Up arrow means higher scores are preferred, and vice versa.

Parameter updating We update the model parameters with Adam optimizer [21] for both stages, with initial learning rate $lr = 0.0002$ and $\beta_1, \beta_2 = (0.5, 0.999)$. For the training in stage I, we fix the learning rate for the first 100 epochs and linearly decay it to 0 in another 100 epochs. For the training in stage II, we empirically set the number of total training epochs to 30 to ensure convergence.

4.3. Reconstruction Results

We compared the per-frame reconstruction performance of our proposed framework with state-of-the-art pose-guided person image generation network Deformable-GAN [7]. During the testing, we select the first frame of each video clip to be the input image. We adopt two commonly used objective evaluation metrics, MS-SSIM [22] and LPIPS [23] to account for both signal-level statistical fidelity and semantic-level perceptual quality. As deformable-GAN does not explicitly warp the background contents, we masked out the background in reconstructed frames for fair comparison. Table 1 shows the evaluation results of both works. Our proposed framework outperforms the Deformable-GAN baseline in both metrics, demonstrating the ability of our network to generate high-quality video frames.

4.4. Random sampling and interpolation on latent codes

To demonstrate the capability of our video generation framework to manipulate the color of clothes, we perform random sampling and interpolation on the latent space and observe the color variation of the output videos. Given a target pose sequence $P = \{p_i\}_{i=1}^{N}$ of length N, we first sample two random points z_c^1, z_c^2 from the latent color space and generate a linearly interpolated sequence $\tilde{\mathbf{z}}_c = \{\tilde{z}_c^i\}_{i=1}^{N}$ where $\tilde{z}_c^i = (1 - i/N)z_c^1 + (i/N)z_c^2$ is the latent code for the i-th frame. Then we generate the corresponding video clip $\mathbf{V}_{interp} = G(P, \tilde{\mathbf{z}}_c)$ with respect to both pose sequence P and latent codes $\tilde{\mathbf{z}}_c$). Fig. 4 shows two generated video clips with interpolated latent code sequences. As can be observed,

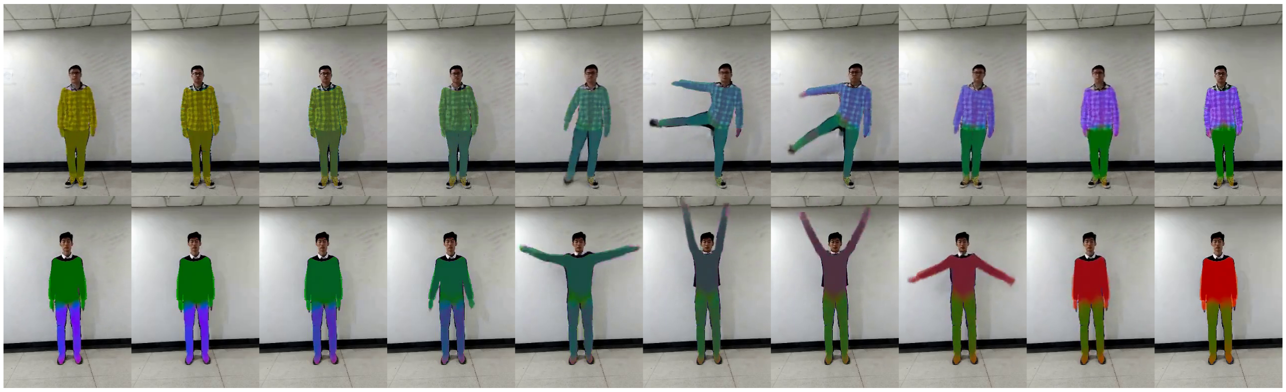
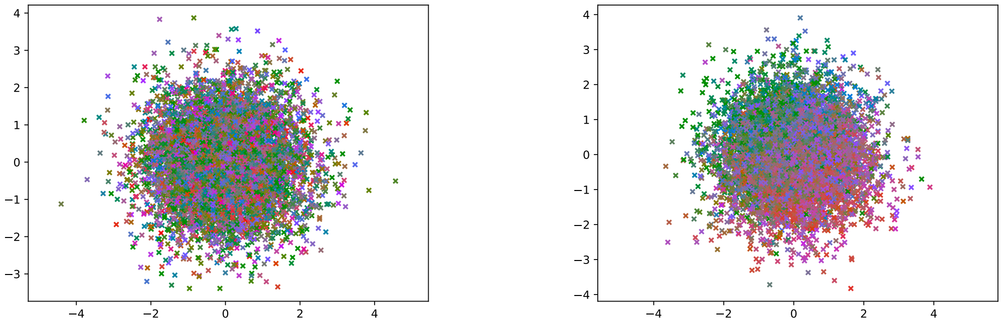


Fig. 4. Video generation results with respect to linearly interpolated latent code sequences $\tilde{z}_c$ between two randomly sampled points. 10 keyframes are showcased for each video sequence.



(a) End-to-end approach

(b) The proposed decoupled learning approach

Fig. 5. Visualizations of the sample distribution in encoded latent color space modeled trained with different approaches.

our framework is capable of continuously changing the color of clothes in generated frames by manipulating latent codes while still preserve the fine-grained texture (like the checkerboard pattern on the shirt in top row). This also indicates that the generator has properly utilized the encoded color attributes instead of just memorizing the training samples.

4.5. Visualization of the latent color space

To demonstrate the importance of decoupled training, we visualize the latent space encoded with our decoupled learning approach and the end-to-end baseline, where the encoder is directly trained from scratch following the procedure in Stage II. Here we only show the results for upper body encoders. We randomly choose 7000 colorized video frames in the training dataset, and keep record of the AB component values we designated for each frame. We visualize the latent codes on two-dimensional plane using PCA. Each point is colored in LAB space with pre-recorded AB values used for color augmentation, and the L component is set to be the average over all upper body pixels in the uncolored gray-scale frame. As illustrated in Fig. 5, the decoupled encoder successfully captures the intrinsic manifold structure of the latent color space, where a clear spectrum from cool colors (blue, green) to warm colors (red, purple) can be observed along the top-left to bottom-right diagonal. On the contrary, even with clip-wise color augmentation, the end-to-end encoder still fails to capture any meaningful structure in the latent space.

5. CONCLUSION

In this paper, we propose an effective two-stage decoupled learning approach for disentangled human action video generation task, where traditional end-to-end learning approaches would suffer from the mode-collapse problem. With carefully designed color augmentation preprocessing step, our framework can well capture the intrinsic structure of the latent code space, and learn to generate diverse human action videos where the pose feature maps and the color attributes are utilized in an disentangled manner. Experimental results have demonstrated the superiority of our decoupled learning scheme over end-to-end learning schemes, and the capability of our disentangled generation framework to produce diverse human action videos where the color of clothes can be flexibly manipulated through latent codes.

6. REFERENCES

[1] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba, "Generating videos with scene dynamics," in *Advances In Neural Information Processing Systems*, 2016, pp. 613–621.

[2] Yitong Li, Martin Renqiang Min, Dinghan Shen, David E. Carlson, and Lawrence Carin, "Video generation from text," *CoRR*, vol. abs/1710.00421, 2017.

[3] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "Video-to-video synthesis," *arXiv preprint arXiv:1808.06601*, 2018.

[4] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool, "Pose guided person image generation," in *Advances in Neural Information Processing Systems*, 2017, pp. 406–416.

[5] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz, "Disentangled person image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 99–108.

[6] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag, "Synthesizing images of humans in unseen poses," *arXiv preprint arXiv:1804.07739*, 2018.

[7] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe, "Deformable gans for pose-based human image generation," in *CVPR 2018-Computer Vision and Pattern Recognition*, 2018.

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[9] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros, "Everybody dance now," *arXiv preprint arXiv:1808.07371*, 2018.

[10] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua, "CVAE-GAN: fine-grained image generation through asymmetric training," *CoRR*, vol. abs/1703.10155, 2017.

[11] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman, "Toward multimodal image-to-image translation," in *Advances in Neural Information Processing Systems*, 2017.

[12] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[13] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Kai Wang, and Xiang Bai, "Richer convolutional features for edge detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[14] Natalia Neverova, Rıza Alp Güler, and Iasonas Kokkinos, "Dense pose transfer," *arXiv preprint arXiv:1809.01995*, 2018.

[15] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros, "Real-time user-guided image colorization with learned deep priors," *CoRR*, vol. abs/1705.02999, 2017.

[16] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.

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

[19] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[20] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[21] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Ieee, 2003, vol. 2, pp. 1398–1402.

[23] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.