Robust Pose Transfer with Dynamic Details using Neural Video Rendering

Yang-Tian Sun, Hao-Zhi Huang, Xuan Wang, Yu-Kun Lai, Wei Liu, and Lin Gao*

Abstract—Pose transfer of human videos aims to generate a high-fidelity video of a target person imitating actions of a source person. A few studies have made great progress either through image translation with deep latent features or neural rendering with explicit 3D features. However, both of them rely on large amounts of training data to generate realistic results, and the performance degrades on more accessible Internet videos due to insufficient training frames. In this paper, we demonstrate that the dynamic details can be preserved even when trained from short monocular videos. Overall, we propose a neural video rendering framework coupled with an image-translation-based dynamic details generation network (D^2G-Net), which fully utilizes both the stability of explicit 3D features and the capacity of learning components. To be specific, a novel hybrid texture representation is presented to encode both the static and pose-varying appearance characteristics, which is then mapped to the image space and rendered as a detail-rich frame in the neural rendering stage. Through extensive comparisons, we demonstrate that our neural human video renderer is capable of achieving both clearer dynamic details and more robust performance even on accessible short videos with only 2k∼4k frames.

Index Terms—Human Video Synthesis, Pose Transfer, Dynamic Details Generation, Deep Generative Model, Neural Rendering

1 Introduction

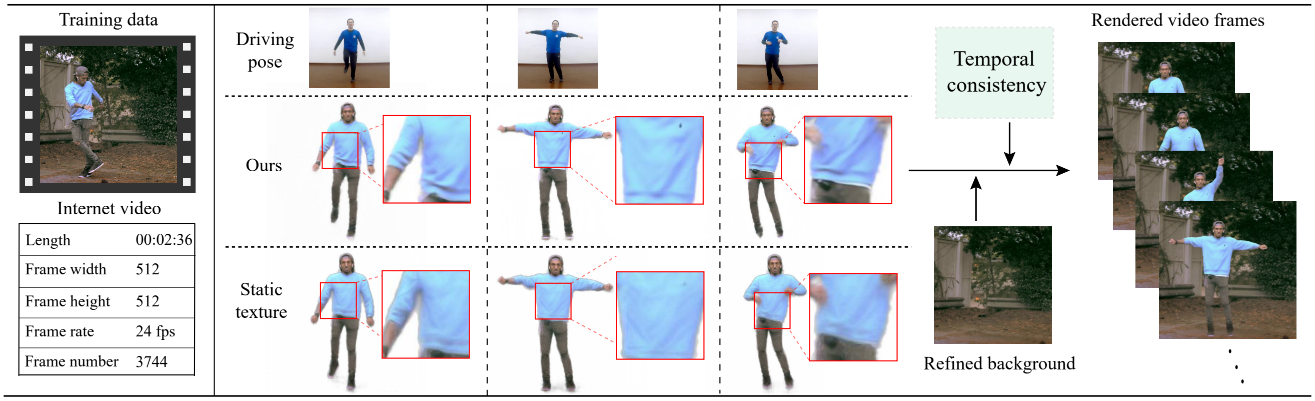
Recently, great progress has been achieved by applying neural networks to video synthesis, especially the human motion transfer task, which aims to transfer the action of a source person depicted in a video to a target one. The most essential challenges for this problem include but are not limited to synthesizing a detail-rich target video and imitating a variety of human motions, which can differ significantly from the training set.

Most existing approaches fall into one of the two categories: image-to-image translation methods [1], [2], [3], [4], [5], [6] which map pose labels to person images, and neural rendering based techniques that learn explicit 3D representations coupled with the traditional graphics rendering pipeline. Despite the powerful representational capacity of deep features that can help generate detail-rich results, image-to-image translation methods rely on black-box 2D generative networks and often introduce obvious artifacts for poses with a large deviation from the training samples. Hence, a great number of training frames are needed to cover as many poses as possible, e.g., more than 10k frames are used for each subject in [1]. For neural rendering based techniques, the explicit 3D representations are introduced and hence the generative models become more stable. However, they either learn a static texture representation during the training stage, e.g., [7], [8], which is temporally invariant and often leads to blurry results, or rely on finely reconstructed 3D models [9], which are difficult to obtain without large amounts of multi-view data. In general, all of these methods degrade when training data is limited to a common monocular video, which is more accessible in our daily life, e.g., short videos from the Internet.

To alleviate the aforementioned problems, we propose a novel generative framework that seamlessly couples image translation components and neural rendering, to gain the benefits of both and works well even with only a short monocular video. We ease the training difficulty and reduce the data dependency by presenting a novel learnable texture representation together with a corresponding pose-aware *dynamic details generation network* (D^2G-Net) embedded in the neural renderer. Specifically, rather than using a traditional static RGB-channel image, we represent the texture with a hybrid feature map, which encodes the static RGB colors explicitly and the dynamic details of human appearance implicitly. Therefore, the hybrid texture serves as a higher level description of human appearance compared with classic texture images. Meanwhile, we also predict the texture coordinates of each pixel in each frame directly for more flexible UV mapping, which sidesteps the difficulty of fine 3D reconstruction. By this means, the learned hybrid texture can be adaptively mapped to the 2D screen space in a differentiable way without the traditional rasterization rendering pipeline, which is then fed into D^2G-Net for dynamic detail generation. Since the human appearance details are pose-varying, the generation process of D^2G-Net is also conditioned on the current pose. Hence, the mapped feature can be rendered into a foreground human figure with clear details *dynamically*. Moreover, to obtain the complete background image, we propose a refinement strategy by integrating the background information from all the video frames with different foreground-background

* *Corresponding Author is Lin Gao (gaolin@ict.ac.cn).*

- *Y.-T. Sun and L. Gao are with the Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, and also with the University of Chinese Academy of Sciences, Beijing, China.*
  E-mail:sunyangtian@ict.ac.cn, gaolin@ict.ac.cn
- *H.-Z. Huang is with the Xverse.*
  E-mail: huanghz08@gmail.com
- *X. Wang and W. Liu are with the Tencent.*
  E-mail: xwang.cv@gmail.com, wl2223@columbia.edu
- *Y.-K. Lai is with the School of Computer Science and Informatics, Cardiff University, Wales, UK.*
  E-mail:LaiY4@cardiff.ac.uk



Length	00:02:36
Frame width	512
Frame height	512
Frame rate	24 fps
Frame number	3744

Fig. 1. With only an accessible Internet video as training data, our neural rendering framework is able to synthesize temporally coherent person images from given pose sequences, while keeping rich details compared with traditional rendering results using static textures.

occlusions due to different poses. Finally, we composite the predicted foreground and background together to obtain the final rendered result. Note that simply splicing the individually generated frames to the video introduces flickers and jitters. To deal with such artifacts, we incorporate the temporal loss [10] to synthesize temporally consistent videos. Here, we summarize the technical contributions as follows:

- A novel end-to-end neural rendering framework for human video generation with dynamic details using accessible monocular video as training data.
- A learnable hybrid texture representation to encode static appearance explicitly and high-frequency pose-varying details implicitly.
- A pose-aware dynamic detail generation network (D^{2}G-Net) which renders the learned texture feature into high-fidelity human images with pose-varying details.

2 Related Work

Video-based human motion transfer has been extensively studied over the recent decades due to its ability for fast video content production. The early approaches accomplished this task by manipulating existing video footage, e.g., Video Rewrite [11]. Benefiting from the rapid development of deep learning, both Generative Adversarial Networks (GANs) [12], [13], [14] and neural rendering techniques [7], [9], [15], [16], [17] are applied to the motion transfer task. Here we summarize major advances in the two branches of research, followed by an overview of the methods dedicated to video synthesis.

Image-translation based Approaches. Image-to-image translation was proposed in Pix2pix [18], where a conditional GAN with a U-net architecture [19] was used to transfer an image from one domain to another. Pix2pixHD [20] introduced a multi-scale generator to the Pix2pix pipeline for high-resolution image synthesis. Such architectures and training methods were applied to the video-based pose transfer task and achieve visually pleasing results [1], [2]. However, the fully-supervised learning can easily cause overfitting, leading to poor results when the desired poses are quite different from the training set. There are also other approaches with elaborately-designed pipelines based on the image translation architecture for more general pose-guided person image synthesis, e.g., [3], [4], [5], [6], [8], [21], [22], [23], [24], [25]. Esser *et al.* [24] decoupled the appearance from the shape-guided image synthesis network, enriching the control over the image-translation network. [3], [4], [6], [21], [23] assisted the pose-guided generation with delicate warping strategies on images or feature maps. However, these methods focus on the synthesis of general human images with low resolution, while ours aims to generate high-fidelity person-specific video.

Human Image Synthesis based on Neural Rendering. Some approaches performed human body synthesis by employing an explicit 3D representation and neural rendering. Textured Neural Avatars [7] learned a full-body neural avatar by estimating an explicit texture map and mapping the input pose to a UV-coordinate image. As *static* texture maps are used, most high-frequency details are consequently lost in the synthesized results. Liu *et al.* [9] also proposed to predict dynamic texture depending on poses. However, they employed accurate 3D reconstructions captured by dedicated devices, which are not practical in most application scenarios. In this paper, a novel hybrid texture representation with a dynamic detail generation network is proposed for the realistic and detail-rich person image generation even without high-accuracy 3D reconstructions.

Video Synthesis. Early efforts on video synthesis attempted to improve continuity in the time domain via a recurrent neural network [26], [27]. Wang *et al.* [2] appended an optical flow prediction module and a video discriminator to the off-the-shelf image translation network [20]. Here we introduce the temporal loss from [10] to enforce the network to synthesize temporally consistent results inherently without additional network modules.

3 Method

3.1 Overview

Our aim is to generate a new video of the target person imitating the specific movements using only a short monocular video as training data, while keeping high fidelity and temporal consistency. To achieve this, we propose a novel neural video rendering framework containing both a hybrid texture representation and an embedded dynamic details generation network, as illustrated in Fig. 2. Specifically, given extracted pose labels, we first predict the texture coordinates of each pixel through a UV generator. Meanwhile, we initialize the texture with a learnable feature map. Afterwards, based on the predicted UV coordinates the

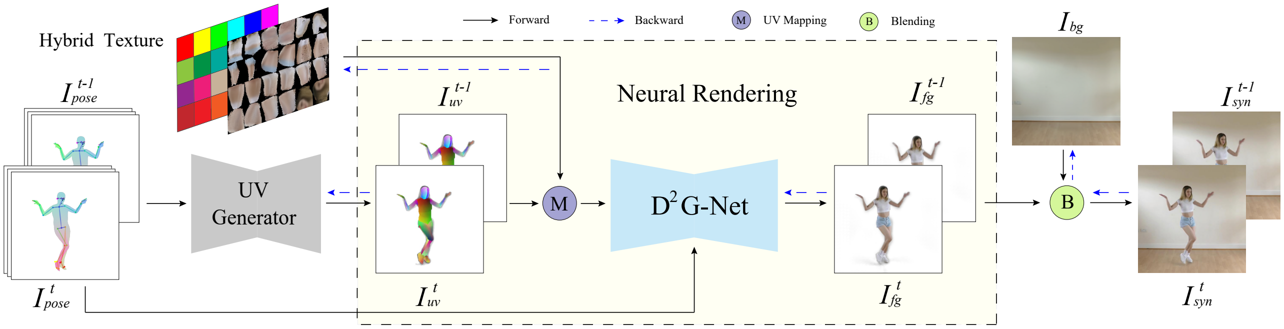


Fig. 2. The pipeline of our training process. Pose labels I^{t-1}_{pose} and I^{t}_{pose} are extracted from the temporal context of adjacent video frames. The UV generator takes pose labels as input and predicts the UV coordinates of each pixel, i.e., I_{uv}. Afterwards, the hybrid texture is mapped to the screen space according to the predicted UV coordinates and then translated to human foreground images with dynamic details. Meanwhile, background image is refined during the training process and final synthesized images are obtained through a combination of foreground and background images. Note that adjacent frames are trained as a whole for the implicit learning of temporal coherence in the neural video renderer.

texture feature is mapped to the screen space, which is then fed into the pose-conditioned dynamic details generation network for detail-rich human foreground image rendering. Moreover, to complete the generated frame we propose to combine the background refinement into the end-to-end training framework, by which means the information of frames with different foreground-background occlusions can be integrated effectively.

3.2 Foreground Image Generation

In the classical graphics rendering pipeline, the texture maps are always static RGB images and not conducive to the characterization of high-frequency detailed signals, while the geometric deformations are responsible for dynamic details. DTL [9] first proposes to learn a dynamic texture to characterize the high-frequency signals of human appearance. However, its dynamic texture learning is under the supervision of unwrapped video frames, which relies on the fine-grained character model reconstructed from multi-view training data. Here, we propose a novel neural rendering framework to enrich the dynamic fidelity details with a hybrid texture representation and the corresponding dynamic detail generation network (denoted as “D^{2}G-Net“). The former contains both explicit static color and implicit details features of human appearance, while the latter is responsible for varying signals visualization dynamically in the neural rendering process. Here we start introducing our pipeline from the representation of pose labels, followed by the constitution of hybrid texture representation. Then we expound the neural rendering process, which includes the UV mapping as well as the dynamic details generation from texture features.

Pose Labels Representation. Our pose labels contain both 2D and 3D features. On the one hand, the 2D pose label is a skeleton image based on keypoints extracted by OpenPose [28]. On the other hand, the 3D label is the projection image of a 3D human mesh, which is obtained using a video-based reconstruction approach [29]. Each pixel of the 3D labels contains 3-channel Laplacian features [30], which are intrinsic characterization of 3D geometry and capture 3D body shapes meanwhile. We adopt 2D keypoints because they are relatively more accurate and easy for tracking, while 3D information can cope with pose ambiguity and self-occlusion. The 2D and 3D labels are concatenated into a 6-channel image to represent the current pose. Considering the effect of pose trajectory on dynamic details generation, we expand the current pose label by introducing its temporal context, i.e., we form the pose label as the concatenation of the current frame pose and last two frames. Such a trajectory-based pose label is conducive to more realistic dynamic characteristics generation due to the consideration of velocity and acceleration, as we report in the supplementary material.

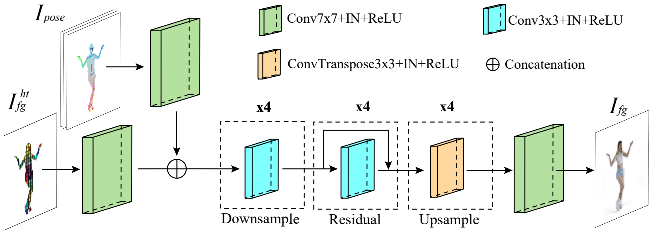


Fig. 3. The illustration of D^{2}G-Net. $I^{\text{ht}}_{\text{fg}}$ is obtained by mapping the hybrid texture to the screen space. The translation network aims to translate the high-dimensional feature to RGB color with dynamic details under the current pose. Note that the pose label is conditioned on the feature level to guide the translation process.

Hybrid Texture Representation and Initialization. To alleviate the decrease of fidelity caused by static texture map, we introduce hybrid texture representation, which is a high-dimensional feature map consisting of both explicit RGB color and implicit detail characteristics. The hybrid texture enriches the generated details by 1) encoding the appearance details implicitly and compensating the explicit RGB map and 2) regularizing the UV generator during the joint learning with D^{2}G-Net. Technically, the texture formation follows the DensePose system [31], which unwraps the human body into $N (N = 24)$ patches and provides a mapping from each pixel to a certain patch of the human surface. Therefore, the hybrid texture is also composited of N parts (denoted as T_i, where $i = 1, \cdots, N$), corresponding to N parts of human body. The hybrid texture contains 18 channels, of which the first three are initialized as RGB colors by unwrapping each frame to the texture space according to DensePose, while the others are initialized as zero. We choose the DensePose instead of the reconstructed 3D model for texture initialization due to its better alignment with frames and clearer results. We provide a more detailed explanation of our design choices in the supplementary material.

Neural Rendering. In this stage, we map the hybrid texture to the screen space differentiably by predicting the UV coordinate through a UV generator. The mapped features are then fed into D^{2}G-Net for detail-rich human foreground generation.

UV Generation. The traditional rendering pipeline relies on explicit 3D models for texture mapping. However, it is difficult to obtain fine 3D human models from only a monocular video. Moreover, the rasterization blocks the backward gradient since its indifferentiable characteristic, which makes the end-to-end training impossible. In our pipeline, we resolve this problem by predicting texture coordinates (UV) of each pixel in each frame directly with a UV generator. To be specific, the UV generator takes current pose labels as input, and outputs the UV coordinates and part probabilities for each pixel in the video frame. The part probabilities $P_i (i = 0, 1, \cdots, N)$ have the size $H \times W$, representing the probability of each pixel belonging to N parts $(P_1, \cdots, P_N)$ or background (P_0), while the coordinates $C_i (i = 1, \cdots, N)$ have the size of $H \times W \times 2$, indicating the UV coordinates of each pixel in the corresponding part. Here H and W are the height and width of video frames, respectively. Then the human foreground of *hybrid texture* $I_{\text{fg}}^{\text{ht}}$ can be obtained by

$$I_{\text{fg}}^{\text{ht}} = \sum_{i=1}^{N} P_i \cdot \phi(T_i, C_i), \tag{1}$$

where ϕ is a function that maps the hybrid texture T_i to the screen space according to UV coordinates C_i. We extract the static component, i.e., RGB channels from $I_{\text{fg}}^{\text{ht}}$ as $\widetilde{I}_{\text{fg}}$, which is a human foreground image without details and serves for subsequent regular loss calculation (Eq. 8). By this means we avoid explicit 3D shape modeling and provide a way for end-to-end training, which enables us to find the best hybrid texture representation and corresponding D^{2}G-Net.

Dynamic details generation. In this part, we aim to visualize the implicit details contained in texture features by translating the human foreground of hybrid texture, i.e., $I_{\text{fg}}^{\text{ht}}$, into detail-rich human images, denoted by I_{fg}. Since the same texture feature should be interpreted to different appearance characteristics under different poses (e.g., wrinkles on the belly when bending down while flat clothes when standing erectly), we introduce the pose label as guidance during the dynamic details generation process. We adopt the pix2pixHD [20] generator as the backbone of our D^{2}G-Net, which has achieved excellent results on the image-translation tasks. The pose label I_{pose} is conditioned by concatenating with $I_{\text{fg}}^{\text{ht}}$ at the intermediate level, as illustrated in Fig. 3. The functionality of D^{2}G-Net can be formatted as

$$I_{\text{fg}} = \text{D}^2\text{G}(I_{\text{fg}}^{\text{ht}}, I_{\text{pose}}). \tag{2}$$

Note that although $I_{\text{fg}}^{\text{ht}}$ also contains pose information, the varying features increase the difficulty of convergence during training. We perform an ablation study of pose condition in Sec. 4.3.

3.3 Background Refinement and Combination

We will complete the generated detail-rich human foreground into a final video frame in this subsection. Since we focus on the human video generation with static background, we optimize the background image during the training process, which utilizes the information from all frames. In order to start from a reasonable initial state, we initialize the background image with a state-of-the-art inpainting network.

We obtain the coarse initial background through frame-by-frame human body deduction and inpainting. Specifically, we segment the foreground (human) based on a U-net [19] from the background for each frame in the training video, and then average the inpainted results obtained using the image inpainting approach [32].

During the training stage, the initial background image is updated according to the backpropagated gradient. By this means the information of all frames is aggregated and the inpainted artifacts are eliminated effectively, as illustrated in the supplementary material. Since P_0 denotes the probability of each pixel belonging to the background, we can combine the foreground and background by

$$I_{\text{syn}} = I_{\text{fg}} \odot (1 - P_0) + I_{\text{bg}} \odot P_0, \tag{3}$$

$$\widetilde{I}_{\text{syn}} = \widetilde{I}_{\text{fg}} \odot (1 - P_0) + I_{\text{bg}} \odot P_0, \tag{4}$$

where I_{syn} is the final synthesized frame, $\widetilde{I}_{\text{syn}}$ is the synthesized frame of the static component, i.e., RGB channels from the hybrid texture, and $\odot$ represents element-wise multiplication.

3.4 Temporal Consistency

The video produced by processing each frame individually does not look realistic and natural enough because of the inevitable flickering and jittering artifacts, especially when the dynamic details are presented. To address this problem, we introduce the temporal loss [10] into the human video generation task, which is defined as the L_1 loss between the generated frame I_{syn}^{t} at time t and the warped version of the generated frame I_{syn}^{t-1} at time $t - 1$. The detailed definition and effect of the temporal loss are presented in the supplementary material.

3.5 Full Objective

We denote the UV generator and D^{2}G-Net as G_{uv} and $G_{\text{D}^2\text{G}}$, respectively. Note that G_{uv} is first pretrained on frames of a large corpus of different characters, which can be used for the end-to-end training of other characters. The pretraining helps improve the model robustness, and the finetuning process enables the generated UV map to better fit the specific character's body shape, as we demonstrate in the supplementary material.

The pretraining objective is given by minimize the following loss

$$\mathcal{L}_{\text{uv}}(P, \hat{P}, C, \hat{C}) = \mathcal{L}_{CE}(P, \hat{P}) + ||C - \hat{C}||_1, \tag{5}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss. P and C are the predicted probabilities and UV coordinates of the UV generator, while $\hat{P}$ and $\hat{C}$ are the ground truths. Note that $\hat{P}$ and $\hat{C}$ can be obtained through existing 3D priors, e.g., rasterization of rigged 3D human models. In practice, we use the results of DensePose directly.

Our whole training network can be trained in an end-to-end manner. Let I_{pose}, I_{bg} I_{real} and T be the pose label, background image, ground truth image, and hybrid texture, respectively. The overall objective is formulated as:

$$\begin{aligned}\min_{\substack{G_{\mathrm{D^2G}},\, G_{\mathrm{uv}}\\ I_{\mathrm{bg}},\, T}} \; &((\max_{D} \sum_{i\in[t-1,t]} \mathcal{L}_{\mathrm{GAN}}(I^i_{\mathrm{pose}}, I^i_{\mathrm{syn}}, I^i_{\mathrm{real}}))\\ &+ \sum_{i\in[t-1,t]} \mathcal{L}_{\mathrm{supervised}}(I^i_{\mathrm{syn}}, I^i_{\mathrm{real}})\\ &+ \sum_{i\in[t-1,t]} \mathcal{L}_{\mathrm{supervised}}(\widetilde{I}^i_{\mathrm{syn}}, I^i_{\mathrm{real}})\\ &+ \lambda_{\mathrm{temp}}\mathcal{L}_{\mathrm{temp}}(I^t_{\mathrm{syn}}, I^{t-1}_{\mathrm{syn}})), \end{aligned} \tag{6}$$

where

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}}(I_{\mathrm{pose}}, I_{\mathrm{syn}}, I_{\mathrm{real}}) = &\,\mathbb{E}[\log D(I_{\mathrm{pose}}, I_{\mathrm{real}})]\\ &+ \mathbb{E}[1 - \log D(I_{\mathrm{pose}}, I_{\mathrm{syn}})]. \end{aligned} \tag{7}$$

$\mathcal{L}_{\mathrm{supervised}}$ consists of a perceptual loss and an L_2 loss, which has the following form:

$$\begin{aligned}\mathcal{L}_{\mathrm{supervised}}(I_{\mathrm{syn}}, I_{\mathrm{real}}) = &\lambda_f ||\mathrm{VGG}(I_{\mathrm{syn}}) - \mathrm{VGG}(I_{\mathrm{real}})||_1\\ &+\lambda_{l_2}||I_{\mathrm{syn}} - I_{\mathrm{real}}||_2. \end{aligned} \tag{8}$$

The perceptual loss regularizes the generated result I_{syn} to be closer to the ground truth I_{real} in the VGG-19 [33] feature space, while L_2 loss does similar restraints at the pixel level. $\mathcal{L}_{\mathrm{temp}}$ is the temporal loss term, and its detailed definition is given in the supplementary material.

Regular loss term. Note that $\mathcal{L}_{\mathrm{supervised}}(\widetilde{I}^i_{\mathrm{syn}}, I^i_{\mathrm{real}})$ is the *regular loss term* calculated from the static component from the hybrid texture. We add this term to stabilize the UV generator and avoid the unreasonable local optimum due to the concurrent update policy of G_{uv} and $G_{\mathrm{D^2G}}$. We visualize the effect of the regular loss in the supplementary material.

4 Experiments

We perform sufficient comparisons to demonstrate the advantage of our approach, which preserves dynamic texture details while only relies on a single accessible monocular video for training. We compare our approach with the state-of-the-art methods under the same data setting. We also verify the importance of each key component in our framework, including the image translation components, pose label condition in the D^{2}G-Net, the regular loss term and the temporal loss. Moreover, benefiting from the learned precise mask of human foreground, our approach is capable of replacing the background in the synthesized video frames. One example is presented in Fig. 8.

4.1 Datasets and Metrics

We conduct massive experiments on the dataset consisting of selected videos from iPER dataset [21], several online videos and one own video, each of which lasts about 3 minutes with 25 FPS (2∼4k valid frames). To ensure the data diversity, there are multiple clothes and actions. The performance of the generative model is evaluated under two conditions depending on whether the source and target characters are the same (self-transfer) or different (cross-transfer). For the former, the Structural Similarity Index Measure (SSIM) [34] and Peak Signal-to-Noise Ratio (PSNR) are used to indicate the quality of generated results; for the latter, we use the Fréchet Inception Distance (FID) [35] with features extracted by InceptionV3 [36] network similar with [7], because there are no ground truth.

However, the above metrics can only reflect the quality of generated results from the holistic perspective. To better measure the robustness of the generative model, we propose a new metric which focuses on poses with a large deviation from the training sample (denoted as "*challenging pose*"). To be more specific, for each pose i in the validation set, we compute its nearest neighbor distance d_i with training samples. Here the distance between two poses is characterized by Euler distance of 2D keypoints. Assuming the poses with larger value of d_i are more challenging, M validation samples ($M = 10$ in our experiments) with largest d_i are selected to form the challenging poses. We evaluate the model performance under these poses with SSIM and PSNR, which is denoted as "R-SSIM"(Robust SSIM) and "R-PSNR"(Robust-PSNR) for clarity.

In addition to these metrics calculated on single frames, we introduce the metric *temporal error* E_{temp} from [10] to evaluate the temporal coherence of generated image sequences. We elaborate the definition of E_{temp} in the supplementary material.

4.2 Comparison with State of the Arts (SOTAs)

We compare our method against existing state-of-the-art methods Vid2vid [2] (V2V), Everybody Dance Now [1] (EDN) and Liquid Warping GAN [21] (LWG), using official implementations. We also compare with existing neural rendering methods Texture Neural Avatar [7] (TNA) and Dynamic Texture Learning [9] (DTL) based on our re-implementation with the same experimental settings as original, since no source code is available. Note that for TNA, we add our refined background for fair comparison. As for DTL, we use the off-the-shelf UV parameterization (DensePose [31]) for frame unwrapping since the accurate 3D reconstruction which relies on multi-view capture devices is not available. We show those comparison results with SOTA in Table 1. Moreover, since LWG model is not limited to a particular person, we only list the results for reference, but it does not participate in comparison.

TABLE 1
Quantitative comparison with SOTA methods. Our approach can synthesize more reasonable results balancing image quality and robustness. Ours also outperforms the others in temporal coherence. The numbers in italics are for reference only, not for comparison.

Method	LWG [21]	EDN [1]	V2V [2]	TNA [7]	DTL [9]	Ours
SSIM ↑	*0.821*	0.911	0.924	0.925	0.924	**0.933**
R-SSIM ↑	*0.823*	0.906	0.915	0.918	0.917	**0.930**
PSNR ↑	*32.58*	37.10	36.60	36.80	37.2	**37.44**
R-PSNR ↑	*32.44*	36.41	35.91	36.62	36.50	**36.99**
Temporal ↓	*0.78*	0.61	0.48	0.46	0.51	**0.41**
FID ↓	*71.20*	58.86	57.04	55.68	56.74	**53.49**

Comparison with direct image translation approaches. The direct image-translation based approaches [1], [2] suffer from a lack of explicit 3D representation, which leads to poor generalization ability especially on short training videos, as illustrated in Fig. 4. There is also a more significant drop of quantitative score for the "*challenging pose*", as shown in "R-SSIM"and"R-PSNR" in Table 1. We further demonstrate the dependence of such methods on the training data length by evaluating their performance under different numbers of training frames in the supplementary material.

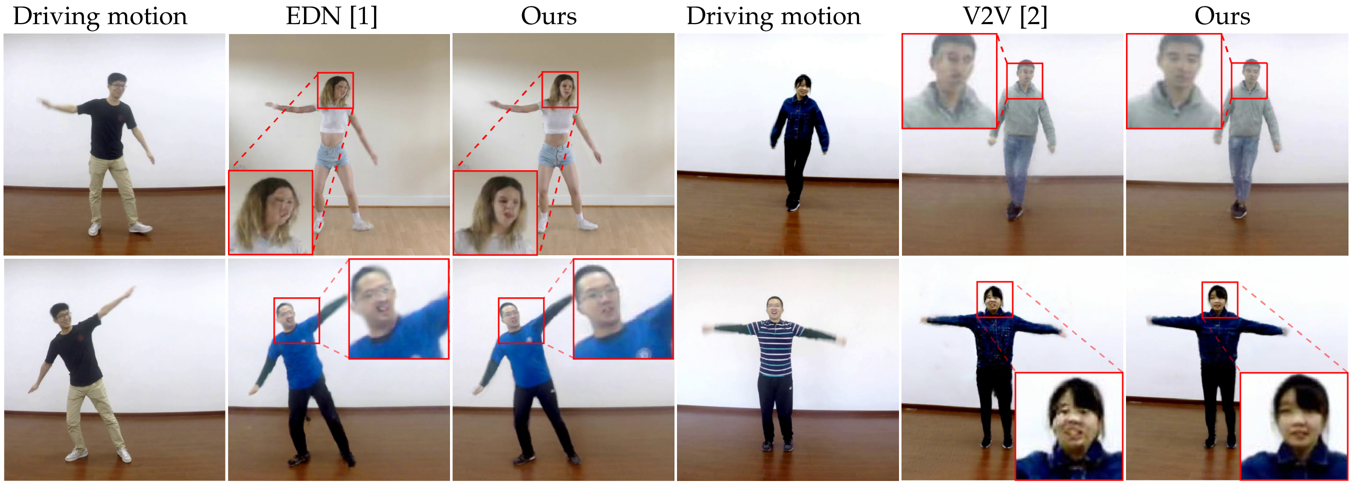


Fig. 4. Comparison with image-translation based approaches. We compare our approach with EDN [1] and V2V [2]. Our approach could produce more realistic imitating results due to the robust 3D representation, even when trained only with a short monocular video.

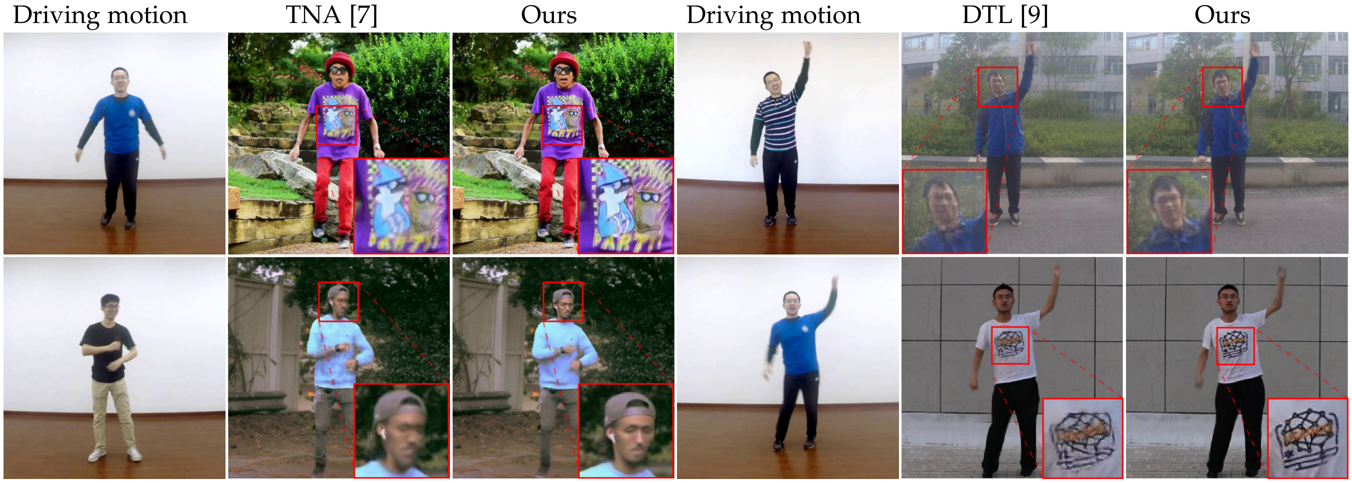


Fig. 5. Compared with other neural rendering baselines, our approach produces much clearer details due to the embedded image-translation components, i.e., the hybrid texture representation and D^{2}G-Net.

Comparison with neural rendering approaches. Compared with other neural rendering approaches (TNA [7] and DTL [9]), our method produces the results with richer details and higher fidelity under the same data configuration, as illustrated in Fig. 5. We also visualize a pose-aware dynamic details comparison in Fig. 6. Although TNA can still produce reasonable results when the training data is limited to monocular video, its static texture map goes against to the dynamic details generation under different poses. By contrast our approach achieves better performance by interpreting the high-dimensional texture feature to details dynamically. As for DTL, the dynamic characteristics are learned explicitly under the supervision of back projected frames according to a finely reconstructed 3D model. Due to the inevitable reconstruction errors with monocular training video, the performance degrades significantly. In contrast, our approach models the dynamic details with implicit image-translation components, which can be learned automatically with only the supervision on final rendering images. Such novel representation and end-to-end training framework make our approach outperform DTL.

Overall, our method synthesizes high quality results with rich high-frequency details, and is more robust to the challenging inputs compared with existing SOTA methods.

User study. We conduct a user study to measure the human perceptual quality for pose transfer results and our results are considered more realistic than baseline approaches. These comparisons are given in the supplementary material.

4.3 Ablation Study

We evaluate the role of each component in our pipeline via convictive ablation studies. Specifically, we investigate how the image translation components and pose label condition work in the generation of pose-aware high-frequency details, as reported in Fig. 7 and Table 2. We also demonstrate the importance of regular loss and temporal loss in the supplementary material.

Effect of image translation component. The ability of our approach to characterize high-frequency signals is largely benefited from the image-translation component embedded in the neural rendering framework, i.e. the D^{2}G-Net and the novel hybrid texture representation. We have performed experiments without the D^{2}G network (“w/o D^{2}G“) and with common RGB texture representation (“w/o HybridTex“) to demonstrate their effect respectively, as illustrated in the 2nd and 3rd columns of Fig. 7. Without the D^{2}G-Net, the pipeline actually degrades to the static texture and the dynamic details cannot be well exhibited. When the hybrid texture is removed, the rendered high-frequency details become diminished and unrealistic, which proves the

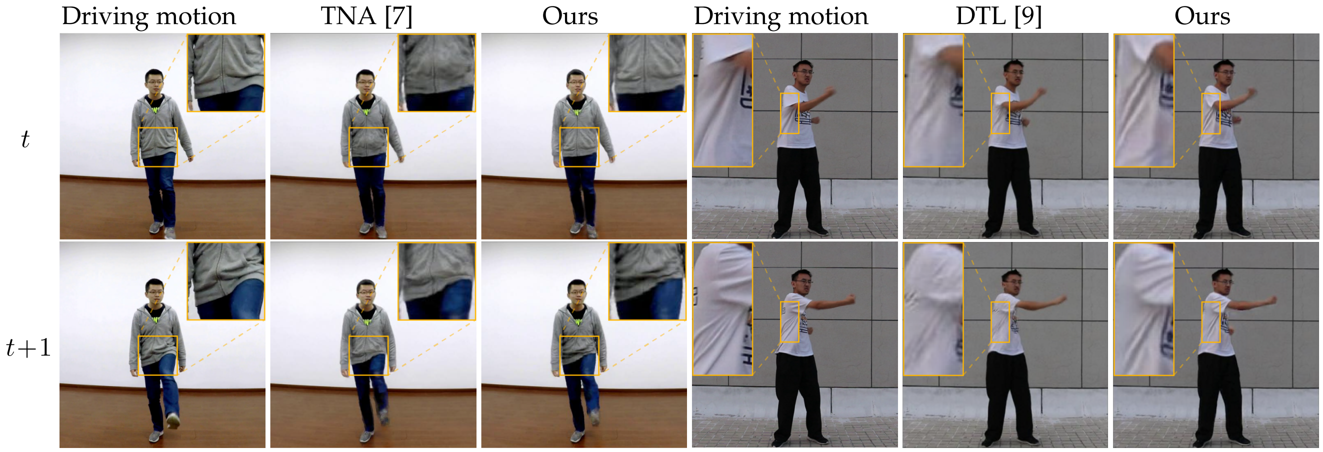


Fig. 6. Comparison of pose-aware dynamic details. The two rows correspond to consecutive frames of t and $t+1$, respectively. Compared with TNA [7] and DTL [9], our method better delineates details for different poses. We show the results of self-transfer for the reference of ground truth. More pose transfer results will be exhibited in the supplementary material.

TABLE 2
The ablation study for the effect of the image translation component. Our full pipeline scores best than other variants, which demonstrates the effect of image translation components and pose label condition.

Method	w/o D^{2}G	w/o HybridTex	w/o Pose cond	Ours
SSIM ↑	0.913	0.921	0.918	**0.933**
PSNR ↑	36.35	36.87	36.62	**37.44**

significance of the high-level appearance representation. We further verify the effect of the hybrid texture by masking out the extra channels of the learned texture in our whole pipeline, i.e., the extra channels are set to zero (“MaskTex”). The result is reported in the 5th column of Fig. 7 and it can be seen that the high-frequency signals are weakened due to the absence of the hybrid representation.

Relevance of pose label condition. To characterize the dynamic details of human surface from hybrid texture under the varying poses, we condition the D^{2}G-Net with current pose labels. Note that although there exists implicit pose information in the human foreground of texture feature, the varying feature during training process would disturb the converge of the translation network. We report the result without pose label condition (denoted as “w/o Pose cond“) in Fig. 7.

5 Limitations and Discussion

Although our approach is able to generate coherent videos with high fidelity, there are still several limitations. First, we have to train a new model for each specific person, which greatly limits the applicable scenarios of our method. Second, since our pipeline is based on existing pose tracking approaches, tracking errors would also affect the quality of generated results. This could be alleviated by the accuracy improvement of pose estimation. Third, current approach is limited to the training video with a static background. It will further improve the flexibility by extending applicable scenes to having dynamic background in the future work.

6 Conclusion

We have proposed a new approach for pose transfer on the human video synthesis task. By embedding the learnable hybrid texture into the neural rendering pipeline with the dynamic details generation mechanism, our approach is able to generate high-fidelity human video frames in an end-to-end manner. In this means, our approach avoids the dependency for abundant training data and serves as a more accessible approach. Furthermore, the pre-training strategy and additional spatial-temporal losses together help regularize the network. Overall, this approach outperforms SOTAs in both robustness and fidelity with training data from short monocular videos. We would provide the source code of PyTorch and Jittor [37] in the future. Jittor is a fully just-in-time (JIT) compiled deep learning framework.

Acknowledgment

This work was supported by the Beijing Municipal Natural Science Foundation for Distinguished Young Scholars (No. JQ21013), the National Natural Science Foundation of China (No. 62061136007 and No. 61872440), Royal Society Newton Advanced Fellowship (No. NAF\R2\192151), Tencent AI Lab Rhino-Bird Focused Research Program and the Youth Innovation Promotion Association CAS.

References

[1] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, “Everybody dance now,” in *IEEE International Conference on Computer Vision*, 2019, pp. 5932–5941.

[2] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, “Video-to-video synthesis,” in *Advances in Neural Information Processing Systems*, 2018, p. 1152–1164.

[3] Y. Li, C. Huang, and C. C. Loy, “Dense intrinsic appearance flow for human pose transfer,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3688–3697.

[4] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “Animating arbitrary objects via deep motion transfer,” in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2372–2381.

[5] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, “Pose guided person image generation,” in *Advances in Neural Information Processing Systems*, 2017, pp. 405–415.

[6] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “First order motion model for image animation,” in *Advances in Neural Information Processing Systems*, December 2019, p. 7137–7147.

[7] A. Shysheya, E. Zakharov, K.-A. Aliev, R. Bashirov, E. Burkov, K. Iskakov, A. Ivakhnenko, Y. Malkov, I. Pasechnik, D. Ulyanov, A. Vakhitov, and V. S. Lempitsky, “Textured neural avatars,” 2019, pp. 2382–2392.

[8] N. Neverova, R. A. Güler, and I. Kokkinos, “Dense pose transfer,” in *European Conference on Computer Vision*, 2018, pp. 128–143.

[9] L. Liu, W. Xu, M. Habermann, M. ZollhÃ¶fer, F. Bernard, H. Kim, W. Wang, and C. Theobalt, “Neural human video rendering by learning dynamic textures and rendering-to-video translation,” *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, pp. 1–1, 05 2020.

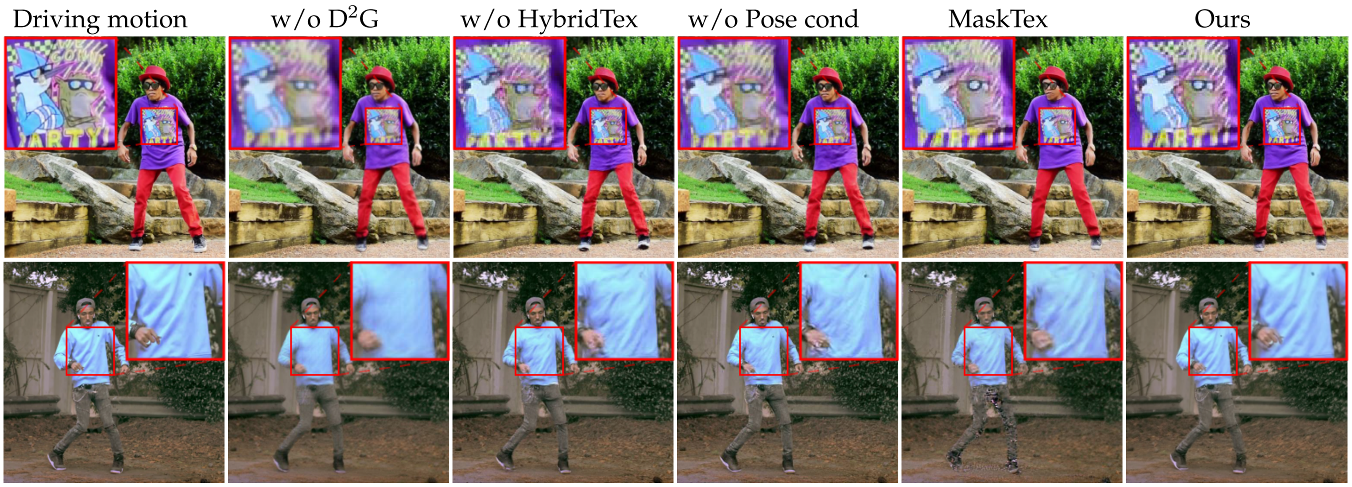


Fig. 7. We visualize the effect of key components in our approach. Compared with other variants, our approach generates clearer high-frequency details (1st row) and pose-varying wrinkles (2nd row).

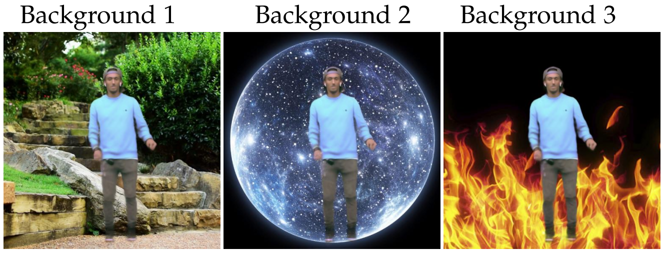


Fig. 8. Examples of the same generated person with different backgrounds. Our approach is able to generate videos with arbitrary novel backgrounds by replacing the refined background image.

[10] H. Huang, H. Wang, W. Luo, L. Ma, W. Jiang, X. Zhu, Z. Li, and W. Liu, "Real-time neural style transfer for videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7044–7052.

[11] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: driving visual speech with audio," in *Proceedings of the annual conference on Computer graphics and interactive techniques*, 1997, p. 353–360.

[12] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *ArXiv*, vol. abs/1411.1784, 2014.

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 2242–2251.

[14] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International Conference on Machine Learning*, 2017, p. 1857–1865.

[15] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering," *ACM Transactions on Graphics (TOG)*, vol. 38, pp. 1 – 12, 2019.

[16] S. Lombardi, J. M. Saragih, T. Simon, and Y. Sheikh, "Deep appearance models for face rendering," *ACM Transactions on Graphics (TOG)*, vol. 37, pp. 1 – 13, 2018.

[17] L. Liu, W. Xu, M. Zollhoefer, H. Kim, F. Bernard, M. Habermann, W. Wang, and C. Theobalt, "Neural rendering and reenactment of human actor videos," *ACM Transactions on Graphics (TOG)*, vol. 38, pp. 1 – 14, 2019.

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2017, pp. 5967–5976.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2015, pp. 234–241.

[20] 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," 2018, pp. 8798–8807.

[21] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, "Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in *International Conference on Computer Vision*, 2019, pp. 5903–5912.

[22] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. V. Guttag, "Synthesizing images of humans in unseen poses," 2018, pp. 8340–8348.

[23] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable GANs for pose-based human image generation," 2018, pp. 3408–3416.

[24] P. Esser, E. Sutter, and B. Ommer, "A variational U-Net for conditional appearance and shape generation," 2018, pp. 8857–8866.

[25] L. Ma, Q. Sun, S. Georgoulis, L. V. Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 99–108, 2018.

[26] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Advances in Neural Information Processing Systems*, 2016, p. 1857–1865.

[27] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1526–1535, 2018.

[28] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1302–1310, 2019.

[29] D. Xiang, H. Joo, and Y. Sheikh, "Monocular total capture: Posing face, body, and hands in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 957–10 966.

[30] Y. Sun, Q. Fu, Y. Jiang, Z. Liu, Y. Lai, H. Fu, and L. Gao, "Human motion transfer with 3D constraints and detail enhancement," *CoRR*, vol. abs/2003.13510, 2020. [Online]. Available: https://arxiv.org/abs/2003.13510

[31] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7297–7306.

[32] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *International Conference on Computer Vision*, 2019, pp. 4470–4479.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.

[34] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, 2004.

[35] 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," in *Advances in Neural Information Processing Systems*, 2017, p. 6629–6640.

[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[37] S.-M. Hu, D. Liang, G.-Y. Yang, G.-W. Yang, and W.-Y. Zhou, "Jittor: a novel deep learning framework with meta-operators and unified graph execution," *Science China Information Sciences*, vol. 63, no. 222103, pp. 1–21, 2020.