Realistic Human Image Animation with Dynamic Camera Effects

Minseok Kang, Jihyun Kim, ChangHee Yang, Kyeongbo Kong, Suk-Ju Kang, *Member, IEEE*

***Abstract*—Human image animation transfers a sequence of pose guidance to a character in an image to generate a dynamic video. In this paper, we enhance the realism of human image animation by introducing viewpoint-conditioned synthesis under moderate camera motion, enabling dynamic camera effects in addition to articuated human motion. However, simply employing off-the-shelf human image animation or view synthesis models in a cascaded manner leads to artifacts, because body motion and viewpoint transformation are handled independently wwithout preserving consistent depth ordering and occlusion. As such, we propose a method which jointly learns to perform both tasks within a shared 3D representation. Specifically, our framework obtains an inpainted background image and isolates the human figure from animated frames. Most importantly, we propose a background depth reconstruction module (BDRM), composed of a depth inpainter and a depth scaler, to obtain an aligned background depth. With the depth information, we convert the background and the foreground into 3D point clouds which are then rendered from target views. For the joint training, we suggest a T-shaped training method which resolves the time-consuming and costly process of collecting animated videos for each different view. In this paper, we prove the effectiveness of our background depth reconstruction module in preventing occlusions, and the superiority of our method over cascade approaches quantitatively and qualitatively. Our Project Page is available at https://excellent-wizard-cf6.notion.site/ieee-tcsvt-dynamic-camera-human-image-animation.**

***Index Terms*—Human image animation, 3D computer vision.**

I. Introduction

HUMAN image animation is one possible way to breathe life into the static people in a photo. Specifically, this task aims to generate a dynamic video from a source image by transferring a sequence of pose guidance to the human in the image. Recently, human image animation has been actively studied for its potential applications across various domains. For example, MagicAnimate [14] introduces a diffusion-based framework which employs an appearance encoder to maintain intricate details of the reference image throughout the generated video. Champ [13] proposes to use a 3D human parametric model for enhanced motion guidance. More recently, UniAnimate [1] achieved long-term human video generation. However, one should note that all these extensive research efforts focus only on dynamic human movements but not on camera movements. In other words, existing works largely assume a fixed camera viewpoint.

To bring not only the static character but also the entire scene in a photo to life, our intuition is that combining controllable cameras with human image animation can enhances the realism and vibrancy of a photo. A related direction is viewpoint synthesis, which generates images under changed viewpoints from one or more input images. In the case of single-image setting, prior methods mainly address moderate camera motion rather than full 360-degree view synthesis. For instance, Synsin [26] leverages latent 3D feature point clouds, and 3D Photography [25] uses Layered Depth Image (LDI) with explicit pixel connectivity as a representation for viewpoint synthesis. On the other hand, NerfDiff [20] introduces an image-conditioned NeRF [15] representation based on camera-aligned triplanes.

Since it is possible to control character movements or camera effects in an image with existing models, applying viewpoint synthesis to the output of human image animation models or vice versa seems to be the straightforward approach to our newly proposed task. However, combining these two tasks is not as simple. As shown in Fig. 1, cascaded approaches lead to artifacts and inconsistencies. This occurs because images are synthesized for each task separately. As a result, changes in camera parameters do not align with simultaneous changes in body pose, and adjustments in body pose do not integrate with concurrent camera movements. In other words, articulation and viewpoint change are handled independently, without a shared structural constraint to maintain consistent depth ordering and occlusion. Thus, to generate realistic and consistent videos, both changes should be considered simultaneously, within a common geometric representation, raising the need for joint training. In this paper, we propose a novel architecture which jointly models human image animation and viewpoint-conditioned synthesis under dynamic camera motion. Rather than applying the two tasks sequentially in image space, our key idea is to couple body articulation and camera transformation within a shared 3D representation before rendering. Throughout the framework, we separately handle the foreground and the background. First, the reference image goes through a human image animation network and an image inpainting network to gain animated frames and an inpainted background image. Then, depth values are estimated

This research was supported by the IITP(Institute of Information Communications Technology Planning Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT) (IITP-2026-RS-2023-00260091, 25%), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-16066849 and RS-2024-00414230 and RS-2024-00456152). *(Minseok Kang, Jihyun Kim, ChangHee Yang and Kyeongbo Kong contributed equally to this work.) (Corresponding author : Suk-Ju Kang)*

Minseok Kang, Jihyun Kim, ChangHee Yang and Suk-Ju Kang are with the Vision and Display Systems Laboratory for Electronic Engineering, Sogang University, Seoul 04017, Republic of Korea (E-mail: richkang715@gmail.com; jhkim5950@sogang.ac.kr; yangchanghee2251@gmail.com; sjkang@sogang.ac.kr).

Kyeongbo Kong is with the Computer Vision and Signal Processing Laboratory for Electronics Engineering, Pusan National University, Busan 46241, Republic of Korea (E-mail: kkb4723@gmail.com).

Fig. 1. **Importance of Joint Training.** Cascade approaches, where (a) the output of a human image animation network is used as input for a novel view synthesis model, or (b) vice versa, result in blurred and unrealistic visual content, as well as inconsistent backgrounds. In contrast, (c) our joint training method effectively addresses these issues.

for the animated frames and the background image, which are used to lift the foreground and the background into 3D space as point clouds. After combining the foreground and the background point clouds, by fixing or altering the camera poses, networks in the framework such as feature extractor, image decoder, and depth scaler are trained with both novel view synthesis and human image animation datasets simultaneously. Lastly, the combined point clouds are rendered from target camera poses and decoded by the image decoder which is trained to inpaint small gaps in the rendered frames.

Simply employing existing depth estimation models to predict depths for the background image results in occlusions in the rendered images, which we refer to as *depth ambiguity*. This issue arises from differences in depth scales, where the depth of the background image is not aligned with the actual background depth in the animated frames. To solve this problem, we incorporate a new module called Background Depth Reconstruction Module (BDRM), which is composed of a depth inpainter and a depth scaler.

Inspired by [31], we built the depth inpainter based on gated convolution networks, and employ it to fill in the masked region in the depth map of the reference image using the inpainted RGB background image. Moreover, for enhanced performance, we introduce a depth scaler which is trained with alignment and constraint losses. These components are important for preserving depth ordering and occlusion consistency when articulated human motion and viewpoint transformation occur together in the shared 3D space.

For the joint training of human image animation and viewpoint-conditioned synthesis, animated videos for each view is necessary which is impractical and difficult to collect. As such, we utilize a T-shaped training method for the joint training. T-shaped training method includes horizontal training and vertical training. During horizontal training, we use a viewpoint synthesis dataset. We start by selecting the first frame from each video sample in the TikTok dataset [33] and generate pseudo ground truth novel view frames using 3D photography [25]. We then randomly select target viewpoints to predict the corresponding novel view images. In the vertical training phase, we use the TikTok dataset [33] for human image animation. Here, we fix the camera pose and select random time steps. Then, our framework is trained using ground truth frames at the sampled time steps.

To summarize, our main contributions include:

- We introduce a novel task that significantly advances the field of human image animation by incorporating realistic and dynamic camera movement effects.
- We introduce Background Depth Reconstruction Module (BDRM), composed of a depth inpainter and a depth scaler, which prevents depth ambiguity such as occlusions of the human by the background in the rendered image.
- We formulate the task as a coupled 3D transformation problem, where body motion and camera transformation are handled within a shared geometric representation.
- We propose a joint training method which simultaneously learns to conduct human image animation and viewpoint-conditioned synthesis.

II. Related Work

A. Human Image Animation

There are two branches of work for human image animation task. The first line of work commonly deforms the reference image based on the motion guide sequence using a feature warping function, as in [2]–[5]. Subsequently, Generative Adversarial Networks (GANs) [6] are employed to fill in the missing regions considering the previously warped subjects. Specifically, Yang et al. [2] propose a Semantic Consistent Generative Adversarial Network (SCGAN), which is robust against noisy or abnormal poses. Zhang et al. [5] utilize an auxiliary task (source-to-source) to boost the performance. Yu et al. [3] leverage geometric kernel offsets with adaptive

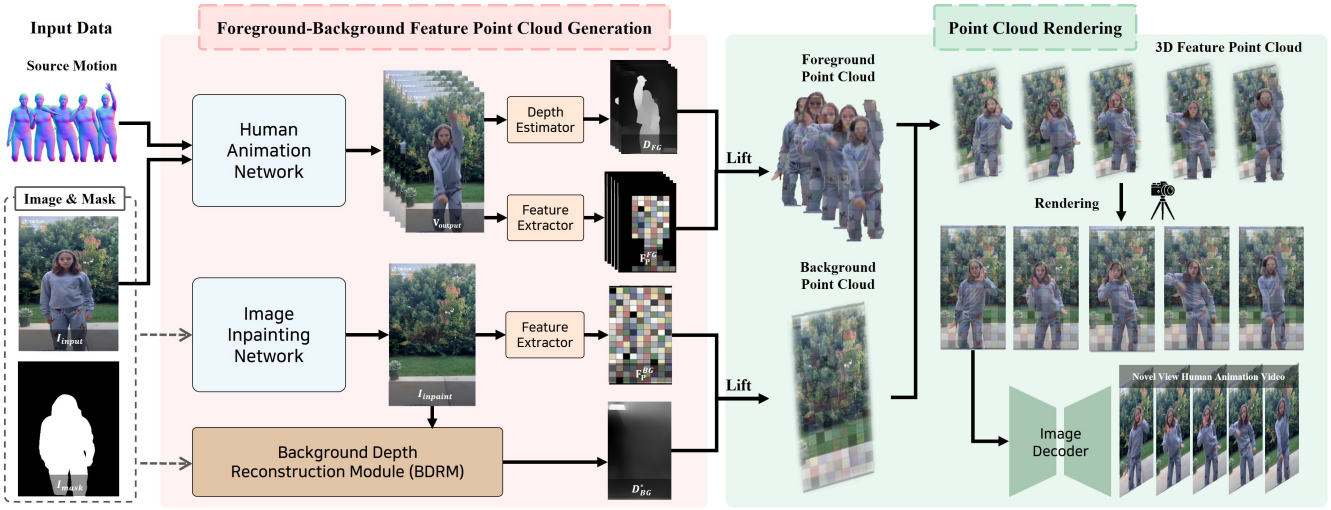

Fig. 2. **Overall framework.** Our framework is mainly composed of foreground-background feature point cloud generation and point cloud rendering process. In the generation processes, depth maps and feature maps of the foregrounds and the background are estimated. Particularly, we built BDRM for background depth estimation to resolve depth ambiguity. In the rendering process, we lift the foregrounds and background into 3D space as feature point clouds using the feature and depth information. Then, the fixed background and the changing foreground feature point clouds are rendered from target camera poses and the rendered feature maps are passed through an image decoder to obtain a novel view human animation video.

weight modulation to align features and perform style transfer simultaneously. The second line of work employs diffusion models (DMs) for video generation, achieving significant improvements in both quality and diversity, as in [7]–[14]. Particularly, some studies, such as [10], [12], and [13], leverage a large pre-trained DM combined with temporal modules and conditional guidance. In more detail, DreamPose [10] fine-tunes existing pre-trained image diffusion model and simplifies the task to finding the subspace of image with the conditioning image. DisCo [12] introduces human attribute pre-training method which disentangles control of the foreground and background region features. Champ [13] leverages a 3D human parametric model withing diffusion-based framework for enhanced body shape alignment and motion guidance. Furthermore, DM-based approaches exhibit superiority over GAN-based methods in terms of training stability and mode collapses. However, while these state-of-the-art models, including AnimateAnyone [9] and MagicAnimate [14], have dramatically improved the quality of human animation itself, they fundamentally assume a static camera.

B. Viewpoint Synthesis with Moderate Camera Motion

Synthesizing novel viewpoints with moderate camera translations has been explored through intermediate 3D-aware representations. Layered representations such as multi-plane images and layered depth images enable moderate viewpoint changes by modeling depth-aware image decomposition. For example, 3D Photography [25] synthesizes disoccluded regions via depth-aware inpainting. Point-based approaches such as SynSin [26] lift image features into latent point clouds to render new viewpoints. More recently, image-conditioned radiance field methods [15], [17]–[20], including NerfDiff, [20] infer volumetric representations from sparse inputs. Despite these advances, such methods generally assume static scenes without articulated motion. Moreover, they do not address the coupled transformation of non-rigid human motion and viewpoint change.

C. 3D Representations for Dynamic Scene Modeling

Recent dynamic scene modeling methods further extend 3D-aware representations to time-varying settings. For example, RoDynRF [46] jointly estimates dynamic radiance fields and camera parameters for robust reconstruction under challenging camera motion, while MoSca [47] models scene deformation using a 4D motion scaffold combined with dynamic Gaussian fusion. However, these methods are designed for reconstructing observed dynamic scenes from videos rather than synthesizing controllable human articulation and independently specified viewpoint changes from a single reference image.

III. Methodology

As illustrated in Fig. 2, our framework is designed to generate a video of animated humans with camera movement effects. Firstly, a reference image and a corresponding foreground mask are given as input to an image inpainting network to generate an inpainted background image. Consequently, the same reference image and a source video containing the target pose sequence is given to a human animation network to generate animated frames. Using DPT [34] as a depth estimator, we obtain depth maps for each animated frame. Then, with our BDRM, we estimate the depth of the background image. Subsequently, we generate feature point clouds using depth maps and RGB features, where the background point clouds are fixed and the foreground point clouds change as the time step progresses. Lastly, at target views, feature point clouds are rendered then decoded with an image decoder.

A. Foreground-Background Feature Point Cloud Generation

Human Image Animation Network. The human image animation network generates an animated video or a sequence

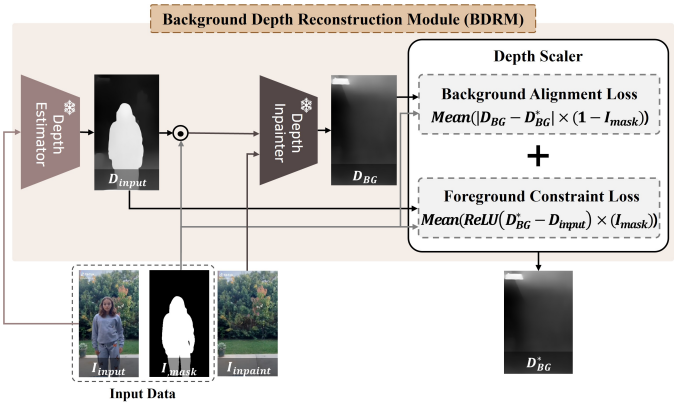

Fig. 3. **A detailed structure of BDRM.** The input image is first processed through a depth estimator to obtain an initial input depth map. After applying a mask to the input depth map, along with the input inpainted background image, it is passed to the depth inpainter for intermediate background depth estimation. Finally, the intermediate depth map is refined by the depth scaler, resolving the depth ambiguity.

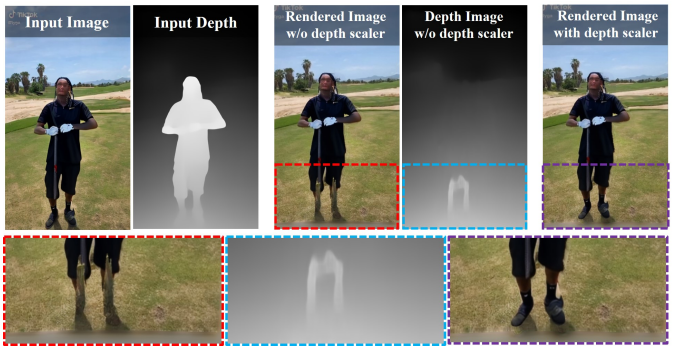

Fig. 4. **Qualitative comparison between w/o and w/. depth scaler.** In the depth map of the inpainted background image, estimated by our module without a depth scaler, inaccuracies near the legs result in occlusions in the rendered image. However, after implementing a depth scaler designed to resolve depth ambiguity, the previously occluded legs are accurately displayed.

of frames $V_{\text{Animation}} \in \mathbb{R}^{T\times H\times W}$, given a reference image $I_{input} \in \mathbb{R}^{H\times W}$ and a sequence of target human poses $V_{Cond} \in \mathbb{R}^{T\times H\times W}$. T denotes the number of frames, whereas H and W each represents height and width of the frames. V_{Cond} are pose guidance frames. Any human image animation model can be employed in our framework in a plug-and-play manner. We obtain $V_{Cond} \in \mathbb{R}^{T\times H\times W}$ from existing datasets, such as the TikTok dataset [33].

Image Inpainting Network. The image inpainting network generates visual contents for the masked region in a given image. In our framework, its objective is to produce a background image $I_{inpaint} \in \mathbb{R}^{H\times W}$ by removing the human figure, which is masked with a foreground mask $I_{mask} \in \mathbb{R}^{H\times W}$, from I_{input}. This allows us to handle the background and the foreground separately. For the inpainting network, we use Latent Diffusion Models [35] specifically trained to remove any object within the masked region.

Feature Extracting Network. Inspired by previous works [25], [32], we incorporate a feature extractor into our framework to extract the foreground feature map F_{FG} and the background feature map F_{BG}. These feature maps are combined with depth maps to be transferred into 3D space as feature point clouds $P = (X_i, f_i)$, which have been shown to produce high-quality results when rendered. X_i and f_i represent i-th 3D coordinate and the corresponding feature vector, respectively. To isolate the feature point clouds of the foreground P_{FG}, we apply I_{mask} to F_{FG}. As for the F_{BG}, we utilize the entire feature map without applying any mask.

Background Depth Reconstruction Module. In our framework, the BDRM is essential for reconstructing a consistent background that remains fixed throughout the output video while allowing the foreground to move. To achieve this, we developed a novel module as shown in Fig. 3. Note that for animated frames $V_{\text{Animation}}$, we utilize a pre-existing monocular depth estimator, DPT [34], to estimate depth for each frame but the following module specifically focuses on estimating the depth of the inpainted background $I_{inpaint}$. The temporal consistency of the background is ensured through the framework's architectural design. Our framework processes the static inpainted background image $I_{inpaint}$ only once to generate a single, fixed background depth map. This fixed depth map is then reused across all video frames, maintaining high background consistency throughout the video sequence.

Existing monocular depth estimators cannot determine the depth values of the inpainted background image in relation to the foreground. On the contrary, they can estimate the depth values of the background while considering the presence of the foreground, when using the input image. For this reason, we designed a depth inpainter which estimates an intermediate D_{BG}, using the depth of the background area in the I_{input} as the standard, and filling in the masked region using $I_{inpaint}$ as reference. Inspired by [31], the depth inpainter is designed based on gated convolution networks combined with self-attention mechanism on each convolution layer. However, the depth ambiguity in the mask region, which is well shown in Fig. 4, necessitates an additional scaling method. As such, we sophisticate the background depth estimation process by introducing a novel depth scaler, which refines the intermediate background depth D_{BG} to generate the final background depth D^*_{BG}. Our depth scaler is an encoder-decoder architecture resembling U-Net, which takes three key inputs: original depth D_{input}, intermediate background depth D_{BG}, and foreground mask I_{mask} to output the final background depth D^*_{BG}. To successfully refine D^*_{BG}, our depth scaler is trained with two novel loss terms - alignment loss and constraint loss:

$$\mathcal{L}_{\text{Scaler}} = \mathcal{L}_{\text{Alignment}} + \mathcal{L}_{\text{Constraint}}, \tag{1}$$

$$\mathcal{L}_{\text{Alignment}} = \text{Mean}(\text{Abs}(D_{BG} - D^*_{BG}) \cdot (1 - I_{mask})), \tag{2}$$

$$\mathcal{L}_{\text{Constraint}} = \text{Mean}(\text{ReLU}(D^*_{BG} - D_{input}) \cdot (I_{mask})), \tag{3}$$

where the alignment loss improves the depth consistency between intermediate background depth D_{BG} and final background depth D^*_{BG} outside the foreground masked areas, while the constraint loss ensures that the D^*_{BG} does not exceed D_{input} in the foreground masked regions.

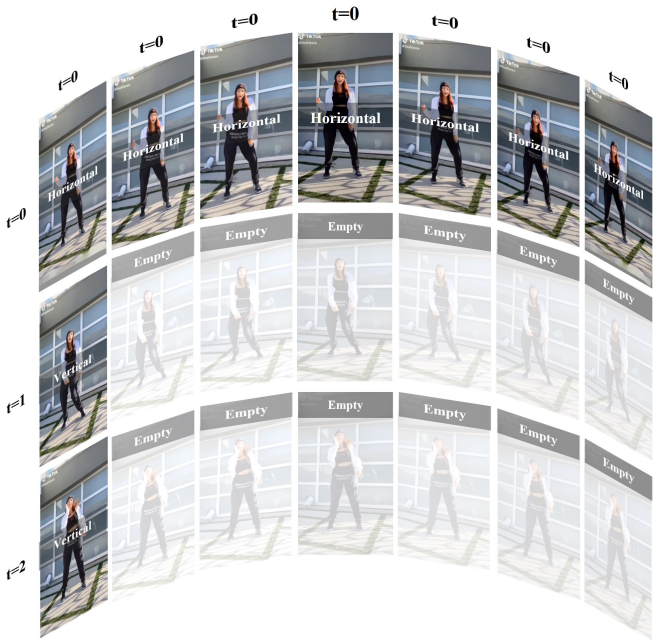

Fig. 5. **Joint training methodology.** The vertical axis represents the time axis showing a person's motion, while the horizontal axis represents images of the same time frame but viewed from different angles due to changes in the camera position. Without joint training, all data assigned to empty areas are necessary, which is impractical and costly to collect.

B. Point Cloud Rendering

As illustrated in Fig. 2, the F_{FG} and D_{FG}, and the F_{BG} and D^*_{BG} are combined to create P at each t in 3D space, which is subsequently rendered using a differentiable rendering equation R from the target camera pose c. Then, the rendered feature map $F_{Rendered}$ is passed to an image decoder network G to become the final result I_{Output}:

$$F_{Rendered} = R(P, c), \tag{4}$$

$$I_{Output} = G(F_{Rendered}). \tag{5}$$

Lastly, the sequence of I_{Output} is stacked into a video $V_{Output} \in \mathbb{R}^{T\times H\times W}$.

C. Training Strategy

We train our framework in three-stage manner. In the first stage, we train our depth inpainter in BRDM with masked images and pseudo GT depth images. In the second stage, we train the feature extractor and the image decoder using a human image animation dataset and a novel viewpoint-conditioned synthesis dataset. In the final stage, we fine-tune the whole framework including our depth scaler in BRDM while freezing the depth inpainter as well as the human animation network and the image inpainting network, with the same datasets used in the previous stage. As mentioned before, our framework uses existing models for both the human animation and image inpainting networks, supporting various models in a plug-and-play manner.

Joint Training Methodology. Our framework jointly learns to conduct human image animation and viewpoint-conditioned synthesis. We suggest T-shaped training for the joint training method. As for our task, animated videos for each different view would be necessary, which is impractical and difficult to collect. In particular, a fully supervised setting would require rendering or collecting multi-view animated videos across all timestamps, which is computationally expensive and difficult to scale in practice. Specifically, as demonstrated in Fig. 5, our framework learns viewpoint-conditioned synthesis horizontally and human image animation vertically. For the horizontal training, we fix the time step to 0 and use pseudo ground truth viewpoint-shifted images. We randomly select camera poses and train the model to render realistic viewpoint-conditioned images. On the other hand, for the vertical training, we fix the camera pose and use ground truth animated video frames. We randomly sample time steps, and train the model to synthesize images close to the ground truth frames. By alternating these two training directions, the model learns viewpoint variation and articulated motion in the same geometric space rather than learning them as two independent transformations. The overall joint training loss is as follows:

$$\mathcal{L}_{\text{Joint}} = \mathcal{L}_{\text{l}_1} + \mathcal{L}_{\text{VGG}}, \tag{6}$$

where l_1 and VGG loss [45] each accounts for per-pixel and perceptual dissimilarity between synthesized images and pseudo-ground truth images during horizontal training, and synthesized images and ground truth images during vertical training.

IV. Experiments

A. Datasets

We train the feature extractor, the image decoder and the depth inpainter with the TikTok dataset [33] which consists of approximately 350 video clips of a dancing human, with length of 10-15 seconds. The dataset is split into training and test sets with a ratio of 8:2. More specifically, for the vertical training (human image animation), we randomly sample frames from each clip from the original TikTok dataset whose camera pose is static. On the contrary, for the horizontal training (viewpoint-conditioned synthesis), we use the first frame of the training video clips to generate 100 pseudo GT viewpoint-shifted images for each video sample using 3D Photography [25], with a predefined camera trajectory (zoom in, zoom out, shift left and shift right).

The camera trajectory is predefined and limited to moderate viewpoint changes. Specifically, x- and y-axis translations are sampled within [-0.04, 0.04] in normalized camera space, z-axis translation is sampled within [-0.1333, 0.1333] to produce zoom effects, and no camera rotation is applied (i.e., the rotation matrix is fixed to identity) during both training and testing. Therefore, our setting focuses on viewpoint-conditioned synthesis under moderate viewpoint changes, rather than general large-baseline novel view synthesis.

To further evaluate robustness and general applicability under more diverse and challenging conditions, we additionally construct a new evaluation dataset, which we call Complex Scene Human Animation (CSHA). CSHA is built by combining selected samples from the evaluation dataset introduced in

Fig. 6. **Qualitative comparison.** We highlight artifacts, background inconsistency, and blurs with red boxes, on the TikTok dataset results generated by (a) Champ → NVS, (b) NVS → Champ, and (c) ours. In contrast to (a) and (b), our method (c) preserves both the details and the frame consistency. Detailed experimental results can be found at https://excellent-wizard-cf6.notion.site/ieee-tcsvt-dynamic-camera-human-image-animation.

RealisDance-DiT [48], which was designed to cover diverse real-world challenges for controllable character animation, with additional in-the-wild Internet videos. To further increase appearance and motion diversity, we also include generated reference images and additional motion-driven video samples. As a result, CSHA contains more diverse human actions, scene layouts, and complex backgrounds than TikTok, and is used to evaluate the robustness of our method in more challenging in-the-wild settings.

For the depth inpainter, we use MS COCO dataset [36] and generate corresponding pseudo GT depth maps with DPT [34].

B. Experimental Setup

For the human image animation network, we use three different models: Champ [13], MagicAnimate [14], and UniAnimate [1]. We employ Latent Diffusion Models [35] for the image inpainting network, and DPT [34] for the depth estimator. We trained our framework for 3 days, using a single A100 GPU. For inference, a single 3090 GPU was used. The full training procedure follows the three-stage strategy described in Section III. We first train the depth inpainter using masked images and pseudo-GT depth maps. We then train the feature extractor and image decoder jointly using both human image animation and viewpoint-conditioned synthesis data. Finally, we fine-tune the overall framework, including the depth scaler, while freezing the depth inpainter as well as the human animation and image inpainting networks.

For viewpoint-conditioned synthesis, target camera poses are uniformly sampled from a discrete set of 240 frames along a predefined circular trajectory. This trajectory is used only as an external control signal for pseudo-GT generation and target pose specification, rather than as a learned trajectory intrinsic to the model. In other words, camera trajectory generation and control are handled by external utility components, while our framework acts as a controllable renderer that follows the provided camera poses. This clarifies that the model is not restricted to a single learned motion pattern, but instead operates under explicitly specified camera controls within the supported motion ranges.

To balance the two training objectives, each mini-batch is constructed with a 1:1 ratio between vertical samples for human image animation and horizontal samples for viewpoint-conditioned synthesis. The overall training objective consists of a masked L1 loss and a perceptual loss based on VGG features. We optimize the model using an initial learning rate of 1e-4 or 3e-4 depending on the training stage, and decay the learning rate by a factor of 0.5 every 50,000 steps.

To the best of our knowledge, there is no prior work that serves as a baseline for our new task of combining human image animation generation and viewpoint-conditioned synthesis. For example, dynamic NeRF methods [42]–[44] require multiple view images, which is not suitable for our approach. The task we propose is to combine single image-based human animation with dynamic camera control under moderate viewpoint changes. State-of-the-art human image animation models such as MagicAnimate lack built-in camera control functionalities, making a direct head-to-head comparison with our method difficult for this specific task.

Therefore, we construct new baselines: human image animation → viewpoint-conditioned synthesis, and viewpoint-conditioned synthesis → human image animation. Since human image animation and viewpoint-conditioned synthesis are research topics that are actively studied, it is reasonable to jointly combine these two tasks to benchmark our method. It is important to clarify that the 'Dynamic Camera Effects' in our work do not refer to a module that generates camera trajectories. Rather, it signifies a framework that enables rendering from desired target camera poses, allowing for dynamic views that are impossible to achieve with existing animation models alone.

Human Image Animation → Viewpoint-Conditioned Synthesis. This baseline generates human motions from a static image using a human image animation model, and then applies viewpoint-conditioned rendering each frame. In particular, we use Champ [13], MagicAnimate [14], and UniAnimate [1] for human image animation, and then 3D Photography [25] to produce new viewpoints for each frame. We choose 3D Photography as the primary view synthesis component because our setting focuses on single-image viewpoint-conditioned synthesis under moderate camera motion, where handling disoccluded regions and depth-aware rendering is particularly important.

Viewpoint-Conditioned Synthesis → Human Image Animation. This baseline is the reverse of the aforementioned approach. We first generate viewpoint-shifted images from existing frames using 3D Photography [25], and then apply human image animation [1], [13], [14] to animate the viewpoint-shifted images. In other words, human image animation is conducted on each frame after its viewpoint has been altered.

C. Quantitative Comparisons

We adopt PSNR, SSIM, and LPIPS [40] to evaluate the single-frame rendering quality. PSNR (Peak Signal-to-Noise Ratio) measures the fidelity of the rendered image by comparing its similarity to the original image. SSIM (Structural Similarity Index) assesses the structural similarity between images, focusing on luminance, contrast, and the structure. LPIPS (Learned Perceptual Image Patch Similarity) evaluates perceptual similarity by considering human vision. Higher values indicate better quality as for both PSNR and SSIM, whereas lower LPIPS values reflect better perceptual quality. Since these three metrics are computed on individual frames, we additionally adopt FVD (Fréchet Video Distance) to evaluate temporal coherence at the video level. FVD measures the distance between the distributions of generated and reference videos in a spatiotemporal feature space, and is commonly used to evaluate temporal realism and consistency in video generation, thereby establishing it as a widely used metric across numerous studies [49]–[52]. The lower values indicate better temporal realism and consistency.

According to Table I, regardless of which human image animation network is used, our method shows competitive performance over the cascaded baselines across PSNR, SSIM, LPIPS, and FVD. Specifically, our method outperforms all cascaded baselines in PSNR, SSIM, and LPIPS across every

TABLE I
QUANTITATIVE COMPARISON. THREE BASELINES: HUMAN IMAGE ANIMATION → VIEWPOINT-CONDITIONED SYNTHESIS; VIEWPOINT-CONDITIONED SYNTHESIS → HUMAN IMAGE ANIMATION; OURS ARE EVALUATED ON THE TIKTOK DATASET. BEST SCORES ARE HIGHLIGHTED IN BOLD.

Method	PSNR(↑)	SSIM(↑)	LPIPS (↓)	FVD (↓)
Viewpoint-Conditioned Synthesis [25] → Human Image Animation [1], [13], [14]				
[25] → Champ [13]	17.079	0.699	0.372	**731.627**
[25] → UniAnimate [1]	18.304	0.736	0.336	652.925
[25] → MagicAnimate [14]	16.773	0.709	0.383	745.651
Human Image Animation [1], [13], [14] → Viewpoint-Conditioned Synthesis [25]				
Champ [13] → [25]	17.097	0.701	0.366	752.287
UniAnimate [1] → [25]	18.220	0.730	0.332	637.705
MagicAnimate [14] → [25]	16.661	0.708	0.394	853.101
Ours				
Champ [13] + Ours	**17.274**	**0.711**	**0.339**	835.137
UniAnimate [1] + Ours	**18.323**	**0.740**	**0.317**	**624.706**
MagicAnimate [14] + Ours	**17.719**	**0.728**	**0.337**	**714.238**

evaluated animation network, demonstrating enhanced per-frame fidelity and structural accuracy. This excellence in individual frame quality is further reflected in video-level performance, where our approach achieves the best FVD results for the UniAnimate and MagicAnimate backbones. While the FVD for the Champ backbone is slightly higher than the cascaded baselines, the consistent dominance in all other quantitative indicators confirms that our method still provides enhanced perceptual quality and structural consistency overall. This improvement in temporal coherence is consistent with our design choice of jointly modeling body motion and viewpoint transformation in a shared 3D representation, where depth ordering and occlusion are handled more consistently across time.

To further strengthen the comparison, we evaluate our method against recent 3D-aware models, RoDyNeRF [46] and MosCa [47]. Since these models were originally designed for video-based novel view synthesis reconstruction, we adapted them as reference comparisons by integrating them into a cascaded setup (e.g., Champ [13] + RoDyNeRF) to fit our single-image-based task. As shown in Table II, our method outperforms Champ + RoDyNeRF and Champ + MosCa on the TikTok dataset across all reported metrics, including FVD. This result demonstrates that the effectiveness of our framework extends beyond comparisons against a single view synthesis component, remaining valid even when compared with stronger recent NeRF- and 4DGS-based baselines. We also evaluate our method on our CSHA dataset, which contains diverse in-the-wild scenes, human actions, and complex backgrounds. As shown in Table II, our method maintains strong performance on this more challenging dataset, suggesting that the proposed framework generalizes beyond the original TikTok benchmark and remains robust under more diverse scene conditions.

D. Qualitative Comparisons

For qualitative results, we adopt Champ [13] for human image animation network, and 3D Photography [25] for view synthesis. We compare the results of the two baselines (human image animation → viewpoint-conditioned synthesis, viewpoint-conditioned synthesis → human image animation) with those of our approach. As shown in Fig. 6, both baselines

TABLE II
QUANTITATIVE COMPARISON WITH RECENT NERF AND 4DGS BASELINES. BEST SCORES ARE HIGHLIGHTED IN BOLD.

Dataset	Method	PSNR(↑)	SSIM(↑)	LPIPS(↓)	FVD(↓)
TikTok	Champ [13] + RoDyNerf [46]	16.882	0.693	0.420	1151.676
	Champ [13] + MosCa [47]	16.373	0.685	0.405	1304.043
	Champ [13] + Ours	**17.274**	**0.711**	**0.339**	**835.137**
CSHA	Champ [13] + RoDyNerf [46]	16.045	0.616	0.425	1537.022
	Champ [13] + MosCa [47]	18.610	0.658	0.399	1414.571
	Champ [13] + Ours	**18.819**	**0.741**	**0.305**	**853.943**

lack background consistency which leads to severe flickering artifacts. These artifacts are more noticeable in a video than in static frames. Moreover, around the edges of the foreground person such as between the legs, details are not properly rendered, which are well maintained in our method's results.

Additional qualitative results, shown in Fig. 7, demonstrate that our method effectively handles human image animation across a wide range of complex scenarios, even with diverse camera poses. Our approach consistently maintains quality and stability even in challenging conditions, which highlights its robustness and versatility for various applications.

To further validate the robustness of our method, we additionally compare our approach with stronger recent 3D-aware baselines, namely Champ + RoDyNeRF and Champ + MosCa. As shown in Fig. 8 and Fig. 9, these cascaded baselines produce noticeable distortions and inconsistent rendering around the human body and the background, especially under simultaneous human motion and viewpoint change. In contrast, our method preserves sharper human structure, and fewer geometric artifacts. We also provide qualitative results on our CSHA dataset. The results show that our method remains visually stable even in these more challenging settings. These observations further support the robustness and broader applicability of the proposed framework.

E. User Study

We conduct a user study by surveying 50 experts in the field of computer vision. We evaluated multiple videos, including those where the person is close to the target viewpoint (occupying a large area in the image) and those where the person is far from the camera (occupying a small area in the image). During the survey, we focused on three different criteria: frame consistency, image quality to assess if the images are generated clean, and smoothness of video to check if moving parts of the human body are not removed or if the person's scale remains consistent. As shown in Fig. 10, the frame consistency is 69.2%, image quality is 66.2%, and smooth video is 66.4%, which demonstrate that our approach is significantly more effective when viewed as a video compared to frame-by-frame analysis.

F. Ablation Study

Effect of each BDRM component. We further analyze the contribution of each component in the Background Depth Reconstruction Module (BDRM). The BDRM is designed in a hierarchical manner. Because the depth scaler takes the output of the depth inpainter as its input, the two components are not fully independent, and the depth scaler cannot be evaluated

Fig. 7. **Additional qualitative results.** The figure above presents the qualitative results of our method, highlighting its ability to maintain high frame consistency across a variety of videos. Unlike other methods that often struggle with background consistency or introduce artifacts, our approach effectively preserves frame-to-frame continuity, resulting in a more coherent and visually stable output throughout the video sequences.

without the preceding inpainter. Therefore, instead of isolating the depth inpainter as a completely standalone module, we perform a step-by-step ablation with three configurations: (1) without BDRM, where naive background depth is directly used, (2) with the depth inpainter only, and (3) with both the depth inpainter and the depth scaler. This provides the finest-grained component analysis that is structurally possible under the hierarchical dependency of BDRM. The results are shown in Table III. The results show that each stage of BDRM contributes to improved performance across all evaluated backbones and across Champ, UniAnimate, and MagicAnimate, confirming the effectiveness of the hierarchical BDRM design.

We also qualitatively compare the results of our framework before and after incorporating BDRM. As demonstrated in Fig. 11, due to the unaligned depths of the inpainted background and the actual background region in the input image, severe occlusions occur, where the background incorrectly covers the foreground. Furthermore, in the qualitative comparison shown in Fig. 4, when the depth scaler is not applied, inaccurate depth values in the inpainted region—particularly where depth

TABLE III
ABLATION STUDY ON THE COMPONENTS OF BDRM. WE COMPARE THREE CONFIGURATIONS: WITHOUT BDRM, WITH THE DEPTH INPAINTER ONLY, AND WITH BOTH THE DEPTH INPAINTER AND THE DEPTH SCALER. BEST SCORES ARE HIGHLIGHTED IN BOLD.

Method	PSNR (↑)	SSIM (↑)	LPIPS (↓)
Champ [13] + Ours			
Without BDRM	16.215	0.697	0.354
+ Depth Inpainter	17.203	0.710	0.341
+ Depth Inpainter + Depth Scaler	**17.274**	**0.711**	**0.339**
UniAnimate [1] + Ours			
Without BDRM	17.191	0.724	0.332
+ Depth Inpainter	18.094	0.739	0.320
+ Depth Inpainter + Depth Scaler	**18.323**	**0.740**	**0.317**
MagicAnimate [14] + Ours			
Without BDRM	16.569	0.714	0.350
+ Depth Inpainter	17.662	0.727	0.338
+ Depth Inpainter + Depth Scaler	**17.719**	**0.728**	**0.337**

Fig. 8. **Qualitative comparison with recent NeRF and 4DGS baselines on TikTok evaluation dataset.** Our results show that cascaded baselines (Champ + RoDyNeRF and Champ + MosCa) produce noticeable distortions and inconsistent rendering around the human body and the background, especially under simultaneous human motion and viewpoint change. In contrast, our method preserves sharper human structure and fewer geometric artifacts.

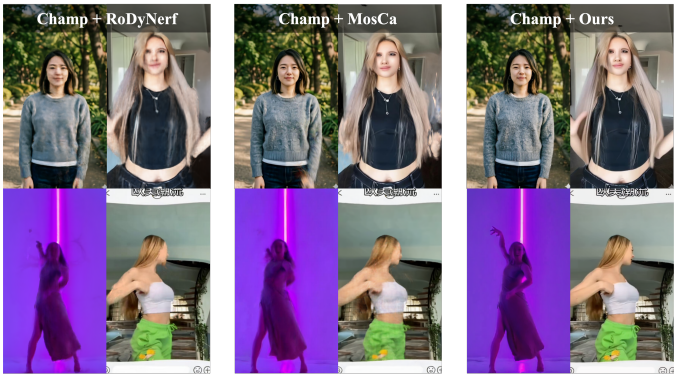

Fig. 9. **Qualitative comparison with recent NeRF and 4DGS baselines on our CSHA dataset.** Our results show that our method remains visually stable even in more challenging in-the-wild settings featuring complex scene layouts and diverse human actions.

values increase rapidly near the legs—result in noticeable occlusions. In contrast, after incorporating the depth scaler, these occlusions are effectively eliminated.

Effect of Foreground-Background Separation. We conduct an ablation study to validate the necessity of foreground-background separation. To this end, we test a variant that

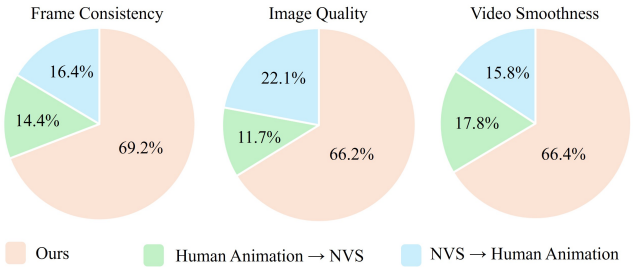

Fig. 10. **User study on output quality.** We conduct evaluations based on three criteria: frame consistency, image quality, and video smoothness. We then compare our method with cascade approaches. The user study confirms that our results significantly outperform the others in these aspects.

Fig. 11. **Effect of BDRM.** Without BDRM, due to the depth ambiguity, occlusions occur as highlighted in red boxes. On the contrary, with BDRM, depth ambiguity is resolved.

Fig. 12. **Ablation study on foreground-background separation.** Unified rendering without separation leads to severe geometric distortions and artifacts during viewpoint changes due to differing depth structures. In contrast, our method models them as separate point clouds to preserve depth ordering and ensure stable rendering under simultaneous motion and camera transformation.

renders the entire scene without explicitly separating the human subject from the background. As shown in Fig. 12, this variant produces severe artifacts when the viewpoint changes, especially around the human silhouette and newly disoccluded regions. Because the human body and the background exhibit substantially different depth structures, unified rendering often leads to holes and geometric distortions where occluded regions become newly visible. In contrast, our framework models the foreground and background as separate point clouds, which helps preserve depth ordering and maintain a consistent global coordinate system under simultaneous articulated motion and camera transformation. These results confirm that foreground-background separation is an important design choice for stable rendering in our setting.

V. Application

In this section, we show that our framework can be extended to other various tasks.

One can think of generating a camera dynamic human animation from input text prompt. T2M-GPT [41] is an outstanding model that generates human motion from textual descriptions. We use “A person is greeting with their hand raised.” and “A person is waving their hand.” as text prompt, and apply our framework to generate camera dynamic human animation. We adopt Champ [13] for human image animation network, using the input reference image and the output human motion of T2M-GPT. The results are shown in Fig. 14.

Also, our approach demonstrates robust adaptability when applied to unseen domains, highlighting its potential for broad applications across diverse scenarios. Unseen domain results

Fig. 13. **Unseen domain results.** Our method works well in new and unseen environments, keeping animations smooth and realistic. It adapts to different conditions without extra training, showing strong performance in diverse applications.

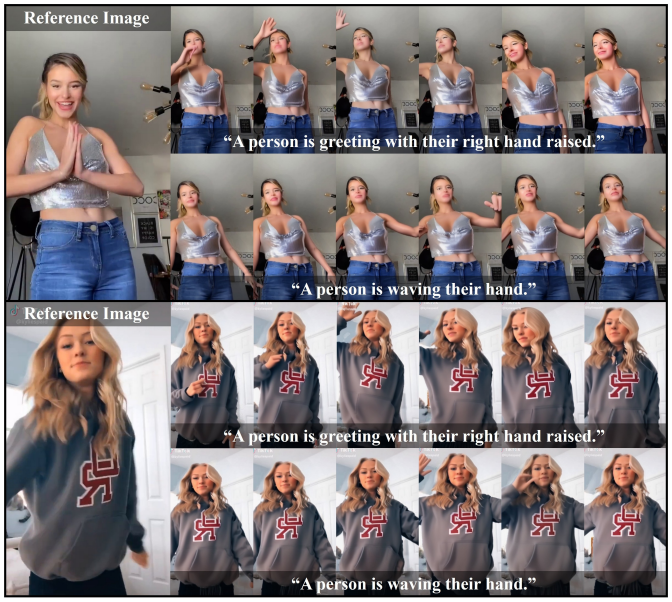

Fig. 14. **Applications on Text-to-Motion.** Our model accepts guidance motion videos. Thus, it is possible to obtain source videos from text-to-motion models, allowing for text control. Additionally, since users are able to generate camera trajectories themselves, our model offers high level of control over various attributes, enhancing dynamism of the output video.

in Fig. 13 shows that our model consistently delivers robust performance. This adaptability underlines the strength of our architecture for real-world applications where training data may not fully represent the target domains.

VI. Conclusion

In this work, we propose a novel task to bring the entire scene in a photo to life, by adding camera movements to static human image animation. To address the artifact problem that arises from a simple combination of conventional human image animation and viewpoint-conditioned synthesis methods, we develop a framework that allows both tasks to be learned jointly. We also introduce a BDRM to align the background depth, and convert foregrounds and background into 3D feature point clouds so that they can be rendered naturally according to the desired camera movement. Experiments show that our approach performs qualitatively and quantitatively better than simple cascade approaches. In addition, we open up new possibilities for adding camera movement to human image animation tasks, and is expected to be utilized in various applications in the future.

References

[1] X. Wang, S. Zhang, C. Gao, J. Wang, X. Zhou, Y. Zhang, L. Yan, and N. Sang, *"UniAnimate: Taming Unified Video Diffusion Models for Consistent Human Image Animation,"* arXiv preprint arXiv:2406.01188, 2024.

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

[3] W.-Y. Yu, L.-M. Po, R. C. C. Cheung, Y. Zhao, Y. Xue, and K. Li, *"Bidirectionally deformable motion modulation for video-based human pose transfer,"* in *Proc. IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7502–7512.

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

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

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *"Generative adversarial nets,"* in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[7] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, J. Laaksonen, M. Shah, and F. S. Khan, *"Person image synthesis via denoising diffusion model,"* in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5968–5976.

[8] D. Chang, Y. Shi, Q. Gao, H. Xu, J. Fu, G. Song, Q. Yan, Y. Zhu, X. Yang, and M. Soleymani, *"MagicPose: Realistic Human Poses and Facial Expressions Retargeting with Identity-aware Diffusion,"* in *Proc. Forty-first International Conference on Machine Learning*, 2023.

[9] L. Hu, *"Animate anyone: Consistent and controllable image-to-video synthesis for character animation,"* in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8153–8163.

[10] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman, *"Dreampose: Fashion image-to-video synthesis via stable diffusion,"* in *Proc. 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22623–22633.

[11] Y. Ma, Y. He, X. Cun, X. Wang, S. Chen, X. Li, and Q. Chen, *"Follow your pose: Pose-guided text-to-video generation using pose-free videos,"* in *Proc. AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4117–4125.

[12] T. Wang, L. Li, K. Lin, C.-C. Lin, Z. Yang, H. Zhang, Z. Liu, and L. Wang, *"Disco: Disentangled control for referring human dance generation in real world,"* arXiv e-prints, arXiv:2307, 2023.

[13] S. Zhu, J. L. Chen, Z. Dai, Y. Xu, X. Cao, Y. Yao, H. Zhu, and S. Zhu, *"Champ: Controllable and consistent human image animation with 3D parametric guidance,"* arXiv preprint arXiv:2403.14781, 2024.

[14] Z. Xu, J. Zhang, J. Liew, H. Yan, J.-W. Liu, C. Zhang, J. Feng, and M. Zheng Shou, *"Magicanimate: Temporally consistent human image animation using diffusion model,"* in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1481–1490.

[15] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, *"NeRF: Representing scenes as neural radiance fields for view synthesis,"* Commun. ACM, vol. 65, no. 1, pp. 99–106, 2021.

[16] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, *"3D Gaussian Splatting for Real-Time Radiance Field Rendering,"* ACM Trans. Graph., vol. 42, no. 4, pp. 139–1, 2023.

[17] T. Müller, A. Evans, C. Schied, and A. Keller, *"Instant neural graphics primitives with a multiresolution hash encoding,"* ACM Trans. Graph. (TOG), vol. 41, no. 4, pp. 1–15, 2022.

[18] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, *"Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields,"* in *Proc. IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5855–5864.

[19] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, *"Zip-NeRF: Anti-aliased grid-based neural radiance fields,"* in *Proc. IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19697–19705.

[20] J. Gu, A. Trevithick, K.-E. Lin, J. M. Susskind, C. Theobalt, L. Liu, and R. Ramamoorthi, *"NeRFdiff: Single-image view synthesis with NeRF-guided distillation from 3D-aware diffusion,"* in *Proc. International Conference on Machine Learning*, 2023, pp. 11808–11826.

[21] J. Luiten, G. Kopanas, B. Leibe, and D. Ramanan, *"Dynamic 3D Gaussians: Tracking by persistent dynamic view synthesis,"* arXiv preprint arXiv:2308.09713, 2023.

[22] Z. Yu, A. Chen, B. Huang, T. Sattler, and A. Geiger, *"Mip-Splatting: Alias-free 3D Gaussian Splatting,"* in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19447–19456.

[23] H. Dhamo, K. Tateno, I. Laina, N. Navab, and F. Tombari, *"Peeking behind objects: Layered depth prediction from a single image,"* Pattern Recognit. Lett., vol. 125, pp. 333–340, 2019.

[24] S. Tulsiani, R. Tucker, and N. Snavely, *"Layer-structured 3D scene inference via view synthesis,"* in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 302–317.

[25] M.-L. Shih, S.-Y. Su, J. Kopf, and J.-B. Huang, *"3D photography using context-aware layered depth inpainting,"* in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8028–8038.

[26] O. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson, *"SynSin: End-to-end view synthesis from a single image,"* in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7467–7477.

[27] C. Rockwell, D. F. Fouhey, and J. Johnson, *"Pixelsynth: Generating a 3D-consistent experience from a single image,"* in *Proc. IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14104–14113.

[28] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen, *"Unstructured Lumigraph Rendering,"* in *Proc. 28th Annual Conference on Computer Graphics and Interactive Techniques*, 2001, pp. 425–432.

[29] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, *"Stereo Magnification: Learning View Synthesis Using Multiplane Images,"* arXiv preprint arXiv:1805.09817, 2018.

[30] P. P. Srinivasan, R. Tucker, J. T. Barron, R. Ramamoorthi, R. Ng, and N. Snavely, *"Pushing the boundaries of view extrapolation with multiplane images,"* in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 175–184.

[31] Y.-K. Huang, T.-H. Wu, Y.-C. Liu, and W. H. Hsu, *"Indoor depth completion with boundary consistency and self-attention,"* in *Proc. IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[32] Q. Wang, Z. Li, D. Salesin, N. Snavely, B. Curless, and J. Kontkanen, *"3D Moments from Near-Duplicate Photos,"* in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3906–3915.

[33] Y. Jafarian and H. S. Park, *"Learning high fidelity depths of dressed humans by watching social media dance videos,"* in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12753–12762.

[34] R. Ranftl, A. Bochkovskiy, and V. Koltun, *"Vision Transformers for Dense Prediction,"* in *Proc. IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12179–12188.

[35] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, *"High-Resolution Image Synthesis with Latent Diffusion Models,"* arXiv preprint arXiv:2112.10752, 2021.

[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *"Microsoft COCO: Common Objects in Context,"* in *Proc. ECCV 2014*, pp. 740–755, 2014.

[37] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, *"Resolution-Robust Large Mask Inpainting with Fourier Convolutions,"* arXiv preprint arXiv:2109.07161, 2021.

[38] J. Jain, Y. Zhou, N. Yu, and H. Shi, *"Keys to Better Image Inpainting: Structure and Texture Go Hand in Hand,"* in *Proc. WACV*, 2023.

[39] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, *"Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer,"* IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 3, 2022.

[40] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, *"The Unreasonable Effectiveness of Deep Features as a Perceptual Metric,"* in *Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595, DOI: 10.1109/CVPR.2018.00068.

[41] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen, *"T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations,"* in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[42] S. Fridovich-Keil, G. Meanti, F. R. Warburg, B. Recht, and A. Kanazawa, *"K-planes: Explicit radiance fields in space, time, and appearance,"* in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12479–12488.

[43] A. Cao and J. Johnson, *"Hexplane: A fast representation for dynamic scenes,"* in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 130–141.

[44] F. Wang, S. Tan, X. Li, Z. Tian, Y. Song, and H. Liu, *"Mixed Neural Voxels for Fast Multi-View Video Synthesis,"* in *Proc. IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19706–19716.

[45] J. Johnson, A. Alahi, and L. Fei-Fei, *"Perceptual Losses for Real-Time Style Transfer and Super-Resolution,"* in *Proc. European Conference on Computer Vision (ECCV)*, 2016, pp. 694–711.

[46] Liu, Yu-Lun and Gao, Chen and Meuleman, Andreas and Tseng, Hung-Yu and Saraf, Ayush and Kim, Changil and Chuang, Yung-Yu and Kopf, Johannes and Huang, Jia-Bin, *"Robust dynamic radiance fields,"* in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13–23.

[47] Lei, Jiahui and Weng, Yijia and Harley, Adam W and Guibas, Leonidas and Daniilidis, Kostas, *"Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds,"* in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 6165–6177.

[48] Zhou, Jingkai and Wu, Yifan and Li, Shikai and Wei, Min and Fan, Chao and Chen, Weihua and Jiang, Wei and Wang, Fan, *"Realisdance-dit: Simple yet strong baseline towards controllable character animation in the wild,"* arXiv preprint arXiv:2504.14977, 2025.

[49] H. Li, Y. Li, Y. Yang, J. Cao, Z. Zhu, X. Cheng, and L. Chen, "Dispose: Disentangling pose guidance for controllable human image animation," arXiv preprint arXiv:2412.09349, 2024.

[50] Q. Wang et al., "VividPose: Vividly 3D-driven Stable Pose Diffusion of High Facial Fidelity," in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), 2025, pp. 1–6.

[51] A. Taghipour et al., "LatentMove: Towards Complex Human Movement Video Generation," arXiv preprint arXiv:2505.22046, 2025.

[52] S. Xu et al., "Hypermotion: DiT-based Pose-guided Human Image Animation of Complex Motions," arXiv preprint arXiv:2505.22977, 2025.

Minseok Kang received the B.S. degree in Computer Science and Engineering from Sogang University, Seoul, South Korea, in 2024, and the M.S. degree in electonic engineering from Sogang University, Seoul, South Korea, in 2026. His current research interests include computer vision applications using deep learning.

Jihyun Kim received the B.S. degree in business management from Sogang University, Seoul, South Korea, in 2021, and the M.S. degree in artifical intelligence from Sogang University, Seoul, South Korea, in 2024. Her current research interests include computer vision and deep learning.

ChangHee Yang received the B.S. degree in electronic engineering from Dankook University, Yongin-si, Gyeonggi-do, South Korea, in 2022, and the M.S. degree in electronic engineering from Sogang University, Seoul, South Korea, in 2024. His current research interests include human-related computer vision research using deep learning.

Kyeongbo Kong received the B.S. degree in Electronics Engineering from Sogang University, Seoul, Republic of Korea, in 2015, and the M.S. and Ph.D. degrees in Electrical Engineering from the Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea, in 2017 and 2020, respectively. From 2020 to 2021, he worked as a postdoctoral fellow with the Department of Electrical Engineering, POSTECH, Pohang, Republic of Korea. From 2021 to 2023, he was an assistant professor in the Media School at Pukyong National University, Busan. He is currently an associate professor of Electrical and Electronics Engineering at Pusan National University. His research interests include image processing, computer vision, machine learning, and deep learning.

Suk-Ju Kang (Member, IEEE) received the B.S. degree in electronic engineering from Sogang University, Seoul, South Korea, in 2006, and the Ph.D. degree in electrical and computer engineering from the Pohang University of Science and Technology, Pohang, South Korea, in 2011. He is currently a Professor of Electronic Engineering with Sogang University, Seoul. His research interests include image and video processing, multimedia signal processing, and deep learning systems. Dr. Kang was a recipient of the IEIE/IEEE Joint Award for Young IT Engineer of the Year in 2019 and the Merck Young Scientist Award in 2022.