

PFAvatar: Pose-Fusion 3D Personalized Avatar Reconstruction from Real-World Outfit-of-the-Day Photos

Dianbing Xi^{1,*}, Guoyuan An^{2,*}, Jingsen Zhu^{3,*}, Zhijian Liu¹, Yuan Liu⁴, Ruiyuan Zhang⁵, Jiayuan Lu¹, Yuchi Huo^{1,†}, Rui Wang¹

¹State Key Laboratory of CAD&CG, Zhejiang University

²Independent Contributor

³Cornell University

⁴Hong Kong University of Science and Technology

⁵Huawei Technologies Ltd

Abstract

We propose PFAvatar (Pose-Fusion Avatar), a new method that reconstructs high-quality 3D avatars from “Outfit of the Day” (OOTD) photos, which exhibit diverse poses, occlusions, and complex backgrounds. Our method consists of two stages: (1) fine-tuning a pose-aware diffusion model from few-shot OOTD examples and (2) distilling a 3D avatar represented by a neural radiance field (NeRF). In the first stage, unlike previous methods that segment images into assets (e.g., garments, accessories) for 3D assembly, which is prone to inconsistency, we avoid decomposition and directly model the full-body appearance. By integrating a pre-trained ControlNet for pose estimation and a novel Condition Prior Preservation Loss (CPPL), our method enables end-to-end learning of fine details while mitigating language drift in few-shot training. Our method completes personalization in just 5 minutes, achieving a $48\times$ speed-up compared to previous approaches. In the second stage, we introduce a NeRF-based avatar representation optimized by canonical SMPL-X space sampling and Multi-Resolution 3D-SDS. Compared to mesh-based representations that suffer from resolution-dependent discretization and erroneous occluded geometry, our continuous radiance field can preserve high-frequency textures (e.g., hair) and handle occlusions correctly through transmittance. Experiments demonstrate that PFAvatar outperforms state-of-the-art methods in terms of reconstruction fidelity, detail preservation, and robustness to occlusions/truncations, advancing practical 3D avatar generation from real-world OOTD albums. In addition, the reconstructed 3D avatar supports downstream applications such as virtual try-on, animation, and human video reenactment, further demonstrating the versatility and practical value of our approach.

Introduction

This paper focuses on a novel and essential task: transforming everyday photo albums into textured, personalized 3D human models (Xiu et al. 2024). These photo collections, commonly referred to as “Outfit of the Day” (OOTD) photos, exhibit several defining characteristics: 1) consistent identity, outfit, hairstyle, and accessories across images, 2) diverse poses and scales, 3) frequent occlusions

and significant truncations, and 4) varied viewpoints against complex backgrounds. These characteristics pose significant challenges to existing 3D avatar reconstruction methods (Alexander et al. 2010; Guo et al. 2017; Xiong et al. 2024; Shen et al. 2023; Işık et al. 2023), which typically require full visibility of the subject and precise camera calibration.

To address these challenges, PuzzleAvatar (Xiu et al. 2024) proposes an innovative framework that eliminates the need for camera calibration. It first decomposes OOTD photos into multiple semantic assets (e.g., garments, accessories, faces, and hair), each of which is associated with a unique token in a Stable Diffusion (Rombach et al. 2021) model—a component referred to as PuzzleBooth. After fine-tuning the diffusion model to personalize it for the given assets, the learned tokens are treated as “puzzle pieces” and used to assemble a 3D personalized avatar through a mesh-based representation using DM Tet (Gao et al. 2020).

However, this method faces several key challenges. First, although decomposing photos into multiple assets helps mitigate issues such as uncommon body poses, it heavily relies on the accuracy of segmentation, which may introduce visual inconsistencies. Second, since PuzzleBooth assembles the human body from individual pieces, it does not support pose-controllable image generation. Moreover, learning multiple individual components significantly increases the overall training time, making the process inefficient. As a result, applying Score Distillation Sampling (SDS) (Wang et al. 2022; Poole et al. 2022) for 3D reconstruction often leads to the Janus problem due to inconsistent pose alignment. Finally, the topology of DM Tet is restricted by the initial mesh structure, making it difficult to represent complex topological variations. Given the highly diverse appearances in OOTD images, this limitation often results in suboptimal 3D geometry with poor visual fidelity (e.g., hair strands or clothing textures).

This paper presents a novel approach, **PFAvatar**, to enhance the visual quality of 3D avatars. First, unlike methods that decompose OOTD images into multiple assets, we propose an end-to-end pose-aware diffusion model. We utilize a pre-trained ControlNet (Zhang, Rao, and Agrawala 2023) to predict the pose of each image, providing pose-prior information for training the personalized diffusion model.

*These authors contributed equally.

†These authors are the corresponding authors.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

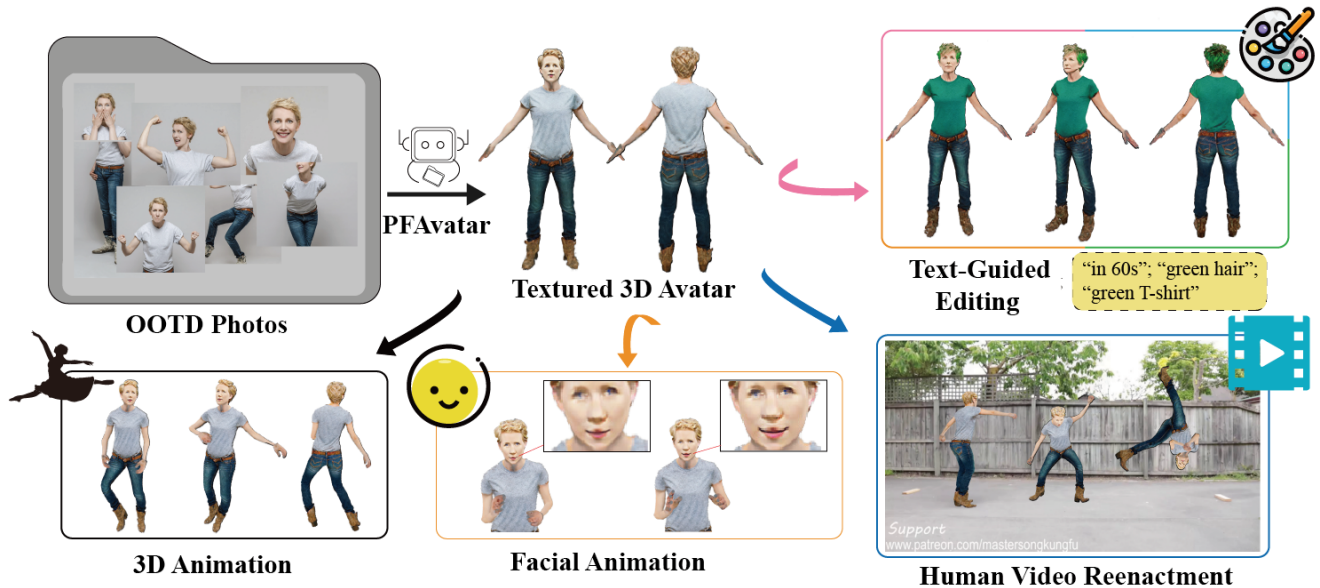


Figure 1: Using the “Outfit of the Day” (OOTD) photos from a personal collection (shown in the upper left), our PFAvatar reconstructs a personalized and fully textured 3D NeRF avatar (depicted in the middle). These OOTD photos can vary widely in terms of body poses, scales, camera angles, framing, frequent partial occlusions, or significant truncation. PFAvatar is designed to handle such variability robustly, enabling a range of downstream tasks. These include virtual try-on through text-guided editing, 3D animation, facial animation, and human video reenactment, all while meticulously preserving the subject’s identity and unique characteristics.

Additionally, we introduce a novel Condition Prior Preservation Loss (CPPL) to mitigate language and control drift issues that often arise during few-shot data finetuning (Ruiz et al. 2023). By eliminating asset decomposition, our method enables end-to-end learning of detailed structures, significantly improving both consistency and accuracy in the final output.

Second, we distill a Neural Radiance Field (NeRF) representation instead of a mesh from our trained diffusion model as the 3D avatar. Our motivation is that NeRF offers greater flexibility and robustness for this task than mesh-based representations. Specifically, NeRF’s volume density naturally handles occlusions through transmittance, ensuring that rays intersecting occluders contribute less to the final pixel and avoiding the generation of false surfaces. Additionally, Compared to mesh-based approach that struggles with high-frequency details due to resolution-dependent discretization, NeRF’s continuous volume rendering directly integrates high-frequency positional embeddings (e.g., hash-grids (Müller et al. 2022)) into the radiance field, which preserves fine-grained details, such as hair strands and intricate patterns, without loss of fidelity.

We conducted extensive explorations to identify the most effective method for distilling high-quality NeRF representations from the fine-tuned Pose-Aware Diffusion Model. Our final approach involves sampling from both the canonical SMPL-X space and the observation space. Sampling from the canonical space generates additional pose-conditioned images, which ensures 3D-consistent NeRF optimization. For the observation space, we employ a pro-

gressive sampling method (Huang et al. 2023a) to achieve higher-quality appearance details.

Additionally, we observed the instability issue of Score Distillation Sampling (SDS) (Poole et al. 2022; Lin et al. 2023), as highlighted in (Cao et al. 2024). This instability degrades our fine local structures due to a lack of human priors. To address this, we introduce a Local Geometry Loss during NeRF training. This loss leverages predefined meshes of body parts, such as hands and faces, to preserve intricate details and enhance the overall fidelity of the 3D avatar. Although NeRF representation is relatively new and lacks the extensive manipulation tools available for traditional mesh-based methods, we demonstrate that our NeRF-based 3D avatar is capable of animation and video reenactment. Moreover, it shows good imaging quality and detail preservation, showcasing its potential as a superior alternative for 3D avatar generation. Extensive experiments on real-world datasets, as shown in Fig. 4 and Fig. 6, demonstrate that PFAvatar outperforms state-of-the-art methods in avatar reconstruction and editing from OOTD photos. We also conducted a user study (see Appendix), which found that users prefer our model over state-of-the-art 3D avatar techniques, further validating the effectiveness and appeal of our approach. In summary, our main contributions include:

- We propose a Pose-Aware Diffusion Model trained from OOTD photos. By leveraging a pre-trained ControlNet for pose estimation and a Condition Prior Preservation Loss (CPPL), our method eliminates reliance on decomposition and enables end-to-end learning of detailed structures with minimal language/control drift un-

der few-shot settings.

- We distill a personalized NeRF avatar from this diffusion model using canonical SMPL-X sampling for 3D consistency and Multi-Resolution 3D-SDS for high-fidelity details. To address SDS instability, we introduce a Local Geometry Loss on anatomical part meshes, preserving fine structures such as hands and faces.
- Extensive experiments show that our NeRF-based avatars outperform mesh-based baselines in reconstruction quality, detail preservation, and robustness to occlusions and truncations.

Related Work

Text and Image-guided 3D Avatar Generation

Numerous studies have explored reconstructing clothed humans from multi-view images (Lin et al. 2024; Saito et al. 2019; Peng et al. 2021) or monocular video (Weng et al. 2022; Li et al. 2020). Recent work leverages text-to-image models and SDS to generate avatars from language descriptions, achieving finer details (Wang et al. 2023a; Liao et al. 2023; Kolotouros et al. 2023; Huang et al. 2023a; Cao et al. 2023). In addition, methods such as (Li et al. 2024; Qiu et al. 2025; Pan et al. 2024) typically depend on accurate pose estimation and require fully visible bodies, making them incapable of reconstructing avatars from images with significant truncations. In contrast, ours overcomes these limitations, enabling robust 3D avatar reconstruction from unconstrained, everyday photos without strict visibility constraints.

Finetuning of Personalized Diffusion Models

In recent years, with the increasing interest in the text-to-image domain, pioneering researchers have begun exploring methods for personalizing text-to-image models using photos of specific subjects. Work on model customization introduces new concepts through fine-tuning (either partial or whole) of pre-trained networks (Avrahami et al. 2023; Jain et al. 2022; Kumari et al. 2023; Liu et al. 2024; Ruiz et al. 2023). Other research re-purposes diffusion models for new tasks (Fu et al. 2024; Ke et al. 2024; Kocsis, Sitzmann, and Nießner 2024). One representative work is DreamBooth (Ruiz et al. 2023), which uses a rare token to represent a specific subject or style, while preventing overfitting through a prior preservation loss. Another approach, textual inversion (Gal et al. 2022), generates a new embedding for the input concept and optimizes this embedding vector with a few photos to enable subject-driven image generation.

Despite these methods achieving laudable results with common objects, the abundance of prior information in the human body poses challenges. This hinders the incorporation of such prior control when fine-tuning human images. Consequently, consistency may diminish when integrating with controllers like ControlNet (Zhang, Rao, and Agrawala 2023).

Pose-Free Reconstruction in the Wild

In our context, pose includes both camera position and body articulation. Accurate camera pose is critical for 3D recon-

struction (Mildenhall et al. 2021), but estimating it from unconstrained OOTD photos is highly challenging.

Prior works address this by jointly optimizing object and camera parameters (Xia et al. 2022; Wang et al. 2021; Lin et al. 2021), leveraging geometric priors (Meuleman et al. 2023; Fu et al. 2023; Bian et al. 2023), or using learning-based methods (Zhang et al. 2024; Wang et al. 2023c,b). Body pose estimation is even harder due to higher dimensionality, and most static-scene methods (Sun et al. 2022; Martin-Brualla et al. 2021) are not suitable for articulated subjects like humans.

PuzzleAvatar (Xiu et al. 2024) and AvatarBooth (Zeng et al. 2023) generate animatable avatars from a few images. However, AvatarBooth relies on dual diffusion models and SDS (Poole et al. 2022), while PuzzleAvatar, based on Break-A-Scene (Avrahami et al. 2023), must learn multiple concepts and requires 4 hours of training.

In contrast, our PFAvatar completes personalization in just 5 minutes. Unlike NFS (Katzir et al. 2023) used by PuzzleAvatar, our method incorporates human-specific priors, reducing Janus artifacts and improving reconstruction fidelity.

Method

Our PFAvatar reconstructs a 3D avatar that faithfully captures both geometry and appearance from input OOTD Photos $\{\mathcal{I}_i\}$. As shown in Fig. 2, it is composed of two primary stages. The first stage, **ControlBooth**, pre-processes the input images and uses them to fine-tune a Pose-Aware Diffusion Model \mathcal{M}_b . The second stage, **BoothAvatar**, distills a NeRF-based 3D avatar model from \mathcal{M}_b .

ControlBooth: Injecting Avatar Feature to Pose-Aware Diffusion Model

In this stage, we aim to train a personalized diffusion model \mathcal{M}_b to achieve novel view synthesis results based on the OOTD photos.

Pose-Aware Diffusion Model: OOTD images often exhibit diverse poses and frequent occlusions, making it challenging to train diffusion models directly on such data. Previous approaches, such as PuzzleAvatar (Xiu et al. 2024), address this by segmenting OOTD images into multiple assets (e.g., garments, accessories, faces, and hair), each associated with a unique token in a Stable Diffusion model (Rombach et al. 2021). However, this method can introduce segmentation inconsistencies, such as mismatched segment boundaries or incorrect labeling of parts.

To circumvent these issues, we propose a Pose-Aware Diffusion Model conditioned on input images $\{\mathcal{I}_i\}$, poses $\{\mathcal{P}_i\}$, and text prompts $\{\mathcal{T}_i\}$. Our data preprocessing pipeline begins by using Ground-SAM (Ren et al. 2024) to separate the foreground (avatar) from the background, yielding the input images $\{\mathcal{I}_i\}$. Importantly, this process only isolates the human region from the background, avoiding the segmentation inconsistency problems inherent in fine-grained part-based segmentation. Next, we employ a pre-trained ControlNet (Zhang, Rao, and Agrawala 2023) to predict the poses $\{\mathcal{P}_i\}$ for all images. Finally, we utilize a

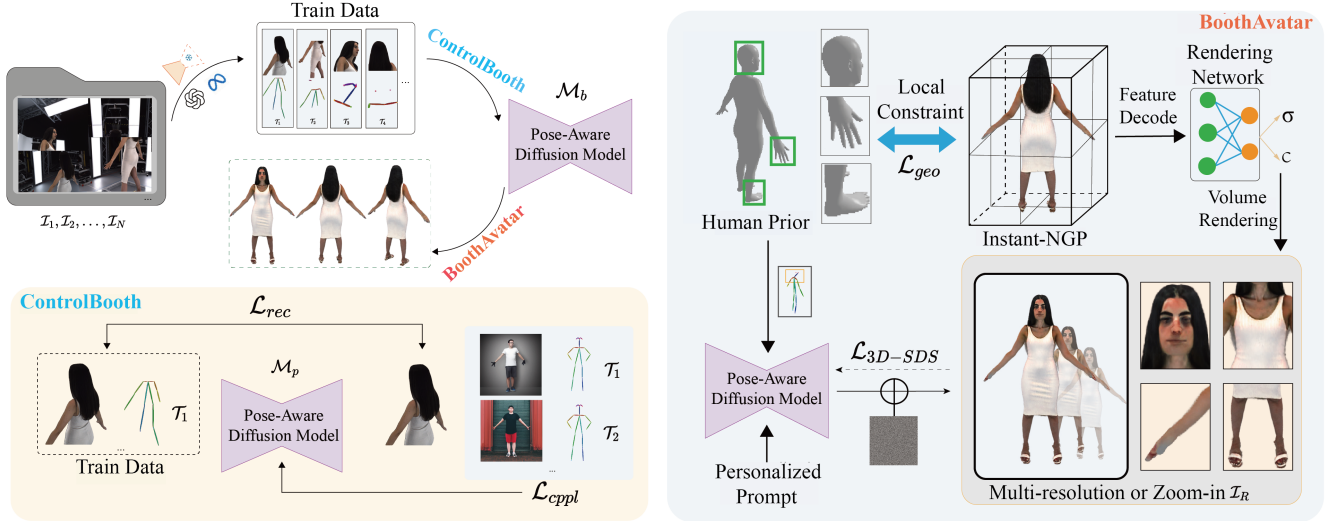


Figure 2: Overview of our PFAvatar pipeline. (1) The top left illustrates the overall flow of our framework, which consists of two main stages: ControlBooth and BoothAvatar. (2) The bottom left displays the training details of the ControlBooth stage. In this stage, our input data is composed of three parts: images, pose-conditioning, and captions. These are used to fine-tune a Pose-Aware Diffusion Model \mathcal{M}_b , where the Text-Encoder and the UNET are trained using the reconstruction diffusion loss, \mathcal{L}_{rec} (eq. (1)), and the condition-based prior preservation loss, \mathcal{L}_{cppl} (eq. (2)). (3) The right section shows the details of the BoothAvatar stage. In this stage, the avatar is represented as an A-posed canonical avatar. The model \mathcal{M}_b obtained from the previous stage is used to guide this reconstruction process. Using multi-resolution \mathcal{L}_{3D-SDS} (eq. (3)), we optimize a NeRF represented by Instant-NGP, with an additional loss \mathcal{L}_{geo} (eq. (4)) to stabilize local structures during SDS optimization.

vision-language model (GPT-4V (OpenAI 2024)) to generate detailed textual descriptions \mathcal{T}_i for each image. These descriptions are obtained through a specially designed query prompt (see Appendix), capturing fine-grained attributes such as body orientation, hairstyle, clothing, and other relevant characteristics.

Finetuning Pose-Aware Diffusion Model: Our diffusion model is finetuned on the inputs described above using two loss terms: a Reconstruction Diffusion Loss \mathcal{L}_{rec} to improve subject fidelity and a novel Condition Preservation Prior Loss (CPPL) \mathcal{L}_{cppl} to prevent the degradation of controllability. The overall training objective is defined as $\mathcal{L}_{total}^{CB} = \mathcal{L}_{rec} + \lambda_{cppl} \mathcal{L}_{cppl}$, where $\lambda_{cppl} = 1$.

The \mathcal{L}_{rec} encourages fidelity for each data tuple of prompt \mathcal{T}_i , pose \mathcal{P}_i , and image \mathcal{I}_i by ensuring that the model maintains high fidelity to the input description. For simplicity, we omit ϵ , t , and w_t in all subsequent loss equations. This loss is defined as follows:

$$\mathcal{L}_{rec} = \mathbb{E} \left[\left\| \mathcal{D}_\theta(\alpha_t \mathcal{I}_i + \sigma_t \epsilon, \mathbf{c}_{t_i}, \mathbf{c}_{p_i}) - \mathcal{I}_i \right\|_2^2 \right], \quad (1)$$

where the i -th image-space condition \mathcal{P}_i encodes a feature space conditioning vector \mathbf{c}_{p_i} , and \mathbf{c}_{t_i} represents its corresponding text conditioning vector.

When finetuning on a small set of OOTD images, there is a risk of reducing the variability in the output poses and view elements of the avatar (e.g., snapping to the few-shot views), as shown in Fig.3 (middle row). To address these issues, we propose a novel **Condition Prior Preservation Loss (CPPL)** to prevent the degradation of controllability,

which promotes diversity, counters language drift, and helps maintain control capabilities.

Specifically, we generate prior preservation (pr) data $\mathcal{I}_{pr} = \mathcal{D}_\theta(\epsilon, \mathbf{c}_{prt}, \mathbf{c}_{prp})$ using the ancestral sampler on the frozen pre-trained T2I diffusion model with random initial noise $\epsilon \sim \mathcal{N}(0, 1)$ and prior preservation text (prt) conditioning vectors $\mathbf{c}_{prt} := \Gamma(f_t(\mathcal{T}_{pr}))$ and prior preservation pose (prp) condition vectors $\mathbf{c}_{prp} := \mathcal{F}(\mathcal{P}_{pr})$. Here, f_t is used to convert the prompt \mathcal{T}_{pr} into the corresponding text embedding, while Γ represents the text encoder that transforms it into the corresponding text conditioning vectors. \mathcal{F} is a neural block that converts the output 2D map \mathcal{P}_{pr} into \mathbf{c}_{prp} . The form of \mathcal{L}_{cppl} is given by:

$$\mathbb{E} \left[\lambda w'_t \left\| \mathcal{D}_\theta(\alpha_t \mathcal{I}_{pr_i} + \sigma_t \epsilon, \mathbf{c}_{prt_i}, \mathbf{c}_{prp_i}) - \mathcal{I}_{pr_i} \right\|_2^2 \right], \quad (2)$$

where \mathbf{c}_{prp_i} , \mathbf{c}_{prt_i} , \mathcal{I}_{pr_i} are simplified as \mathbf{c}_{p_i} , \mathbf{c}_{t_i} , \mathcal{I}_i . \mathcal{L}_{cppl} is a condition prior-preservation term that supervises the model with its own generated images.

BoothAvatar: Reconstructing Avatar from the Finetuned Model

In this stage, we reconstruct a NeRF-based (Mildenhall et al. 2021) canonical A-posed subject-consistent avatar based on the Pose-aware diffusion model \mathcal{M}_b obtained from the previous stage. To fully leverage the controllability of \mathcal{M}_b , we employ the following techniques to reconstruct a high-quality 3D avatar.

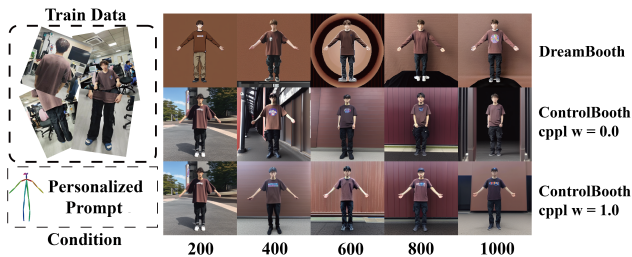


Figure 3: Balancing diversity and control with Condition Prior-Preservation Loss (CPPL). Using the fine-tuning strategy of Naive DreamBooth (Row 1) to generate images with new poses may introduce color discrepancies, significantly reducing thematic consistency. Training with only eq. (1) may lead to overfitting on the context of the input image and the subject’s appearance (e.g., pose). CPPL (Row 3) acts as a regularizer, mitigating overfitting while encouraging diversity and maintaining control.

NeRF-Based 3D Representation: Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) have shown more flexible 3D representation capabilities compared to mesh-based representations (e.g., PuzzleAvatar’s DMTet (Gao et al. 2020)), which is more effective for reconstructing complex clothed human avatars. We use Instant-NGP (Müller et al. 2022) as our canonical avatar representation, which has a high training speed while maintaining high-quality representation. We train the NeRF-based avatar model using score distillation sampling (SDS) (Poole et al. 2022) from controllable novel views generated by our personalized model \mathcal{M}_b .

Specifically, we sample camera poses from the observation space and sample avatar poses from the canonical SMPL-X space. An additional advantage of sampling from the canonical SMPL-X space is its capability to generate skeleton images from the current viewpoint, thereby facilitating convergence and ensuring thematic consistency.

3D-Consistent Score Distillation Sampling: Pose-aware 3D-consistent SDS sampling techniques allow for more controllable processes with high thematic consistency. Previous works that relied solely on text-guided SDS often faced Janus artifact issues (Xiu et al. 2024; Zeng et al. 2023). To address this, we incorporate 3D-aware conditioning images to refine SDS (Huang et al. 2023a) to facilitate 3D-consistent NeRF optimization. Specifically, an extra conditioning image c is used to compute the 3D-SDS gradient $\nabla_{\theta} \mathcal{L}_{3D-SDS}(\phi, \mathbf{x})$ as follows:

$$\mathbb{E} \left[w(t) (\epsilon_{\phi}(\mathbf{x}_t; y, t, c) - \epsilon) \frac{\partial \mathbf{z}_t}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \theta} \right], \quad (3)$$

where the conditioning image c may consist of one or a combination of elements such as skeletons or depth maps, $w(t)$ is a weighting function based on the timestep t , and y represents the corresponding text prompt. In our implementation, we choose skeletons as the conditioning image type due to their minimal structural priors to facilitate complex avatar generation.

Local Geometric Constraint: The instability of SDS optimization can compromise fine local structures derived from human priors. To address this, we introduce a local geometry loss during NeRF training based on predefined meshes of body parts, such as hands and faces. This loss aligns NeRF densities τ in local regions with the predefined meshes through a margin ranking loss \mathcal{L}_{geo} :

$$\mathcal{L}_{\text{geo}} = \begin{cases} (\max(0, \tau_{\max} - \tau(\mathbf{p})))^2 & \text{if } \mathbf{p} \text{ on mesh,} \\ (\max(0, \tau(\mathbf{p}) - \tau_{\min}))^2 & \text{if } \mathbf{p} \text{ not on mesh,} \end{cases} \quad (4)$$

where \mathbf{p} represents 3D points sampled on or near predefined meshes, $\tau(\mathbf{p})$ denotes the densities of 3D points \mathbf{p} predicted by NeRF, and τ_{\min} , τ_{\max} are constant hyperparameters.

In summary, at the BoothAvatar stage, the total loss to optimize the NeRF-based avatar representation is:

$$\mathcal{L}_{\text{total}}^{\text{BA}} = \mathcal{L}_{3D-SDS} + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}}, \quad (5)$$

where $\lambda_{\text{geo}} = 1.0$.

Multi-Resolution and Zoom-in Sampling: Directly rendering high-resolution images from NeRFs is computationally expensive. A common approach is to render a low-resolution image and then up-sample it to a higher resolution for SDS training (Chen et al. 2023b; Lin et al. 2023). However, directly increasing the upsampled resolution can lead to training collapse or inconsistent appearances. To address this, we leverage a multi-resolution and zoom-in optimization strategy, which progressively increases the up-sampling resolution for more stable SDS training while refining the avatar appearance.

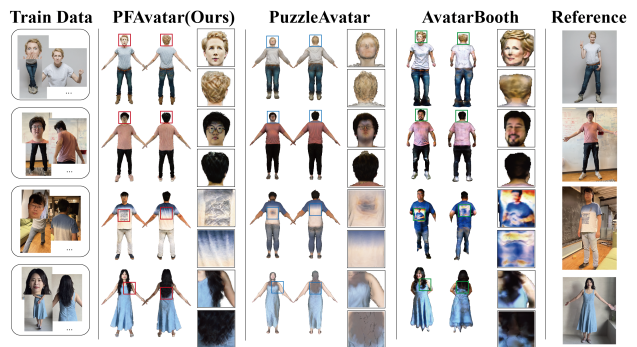


Figure 4: Qualitative Comparison I: Custom Dataset. Visual results on three distinct subjects compare PuzzleAvatar, AvatarBooth, and our method (PFAvatar). Our approach consistently preserves finer details and maintains structural coherence across multiple views, demonstrating clear superiority over the baselines. This highlights PFAvatar’s robustness in achieving high-quality, consistent reconstructions under challenging conditions.

Experiment

Implementation Details

Dataset. We evaluate our system’s performance using two distinct datasets: the Custom dataset and the PuzzleIOI

dataset (Xiu et al. 2024), which serves as a standardized benchmark for personalized human avatar reconstruction. In addition, we introduce a Custom dataset to assess the model’s performance on real-world, in-the-wild scenarios. This dataset comprises casual OOTD photos captured using mobile devices under unconstrained conditions. The images feature diverse poses, complex backgrounds, and varied camera viewpoints, posing significant challenges for 3D reconstruction. This setting enables a robust evaluation of the model’s generalization ability in realistic environments.

Baselines. To evaluate pose-aware diffusion model identity preservation, we compare our method with state-of-the-art (SOTA) approaches, including PuzzleAvatar (Xiu et al. 2024), InstantID (Wang et al. 2024), FreeCustom (Ding et al. 2024), and DisenBooth (Chen et al. 2023a). For reconstruction quality, we benchmark our method against PuzzleAvatar (Xiu et al. 2024) and AvatarBooth (Avrahami et al. 2023), both of which are designed to handle unconstrained inputs.

Metrics. One important aspect to evaluate is subject fidelity: the preservation of subject details in generated images. Fellow (Ruiz et al. 2023),(Raj et al. 2023). To evaluate subject identity preservation across the two stages of our pipeline, we adopt three metrics: DINO (Caron et al. 2021), CLIP-I (Radford et al. 2021), and CLIP-T (Radford et al. 2021).

In addition, for the PuzzleIOI benchmark, we further evaluate the quality of appearance reconstruction. To this end, we render multi-view color images of the reconstructed avatars and report three standard image quality metrics: PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index), and LPIPS (Learned Perceptual Image Patch Similarity). We denote the best results in **bold** and the second-best results with underline.



Figure 5: Qualitative result of personalized diffusion models. Compared to other methods, our *Pose-Aware Diffusion Model* achieves superior subject consistency while enabling precise control over character poses. The pose illustration in the lower-left corner represents the input control pose used for conditional generation. ✓ indicates support for control pose, while ✗ denotes lack of support.

Experiment Results

Analysis of ControlBooth: Subject Identity Preservation As illustrated in Table 1 and Fig. 5, our proposed Pose-Aware Diffusion Model achieves significantly improved

Method	CLIP-I↑		DINO↑		CLIP-T↑	
	body	head	body	head	body	head
PFAvatar(Ours)	0.9016	0.9432	0.7282	0.9352	0.3036	0.2996
PuzzleAvatar	0.8147	0.7705	0.6257	0.6096	0.2340	0.1849
InstantID	0.7687	0.8164	0.5977	0.8302	0.2164	0.2711
FreeCustom	<u>0.8573</u>	<u>0.9337</u>	<u>0.7022</u>	<u>0.9222</u>	0.2583	<u>0.2811</u>
DisenBooth	0.8445	0.7897	0.6930	0.8299	<u>0.2783</u>	0.2581

Table 1: The subject identity preservation performance of different baselines. Ours achieves the best performance.

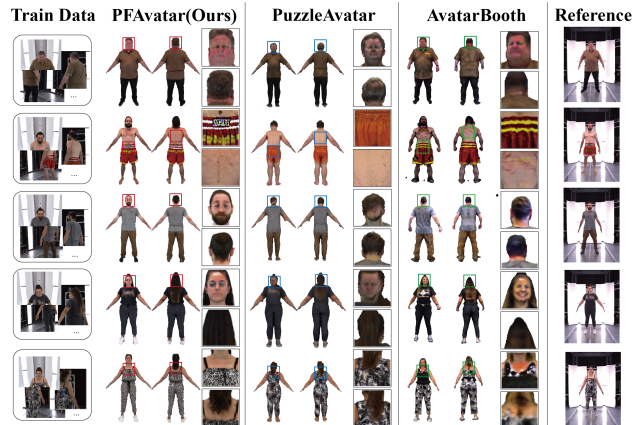


Figure 6: Qualitative Comparison-II: PuzzleIOI Dataset. Visual results on three distinct subjects compare PuzzleAvatar, AvatarBooth, and PFAvatar(Ours). Our approach consistently preserves finer details and maintains structural coherence across multiple views, demonstrating clear superiority over the baselines. This highlights PFAvatar’s robustness in achieving high-quality, consistent reconstructions under challenging conditions.

subject identity preservation. By leveraging a single diffusion model capable of pose-controllable generation, our method consistently maintains identity fidelity—particularly in the facial and head regions. In contrast, PuzzleAvatar (Xiu et al. 2024) lacks pose control capability at this stage, leading to suboptimal identity consistency during reconstruction. These results highlight the effectiveness of our strategy, which integrates both pose information and textual descriptions when training on OOTD images, thereby enabling high-fidelity and identity-consistent synthesis.

Analysis of BoothAvatar: Subject Identity Preservation and Appearance Reconstruction As shown in Table 2, Fig. 4, and Fig. 6, our method achieves superior rendering quality compared to PuzzleAvatar and AvatarBooth, particularly in texture detail preservation, geometric accuracy, and overall visual realism. These improvements demonstrate the robustness and versatility of our approach in generating high-quality 3D avatars from unconstrained, real-world inputs. Moreover, as illustrated in Fig. 6 and Table 3, our method consistently outperforms PuzzleAvatar (Xiu et al. 2024), TeCH (Huang et al. 2023b), and AvatarBooth (Zeng et al. 2023) on the PuzzleIOI benchmark across all evalua-

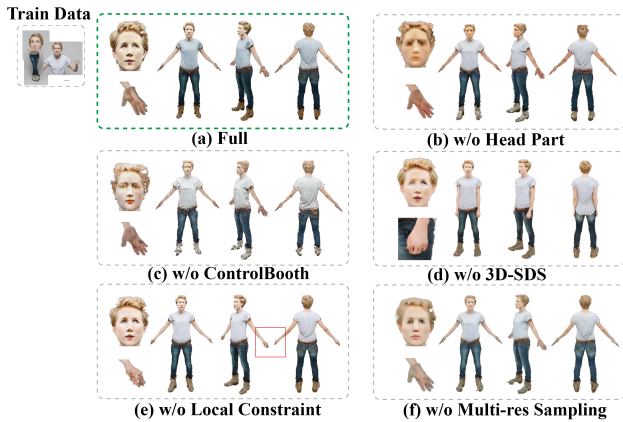


Figure 7: Ablation studies. Qualitative results showing the impact of removing key components. (a) Full model yields the best results; (b–f) each ablation leads to visible degradation in specific aspects.

tion metrics. Specifically, our method produces more realistic textures and achieves more accurate geometric reconstructions—even under challenging conditions such as complex poses—while preserving fine-grained attributes such as facial features, hair strands, and clothing patterns.

Ablation Study We assess the contributions of key components in our framework through ablation experiments. As shown in Fig. 7 and summarized in Table 4, each design choice plays a crucial role in overall performance.

ControlBooth: Excluding Head Region. To assess the effect of head data, we remove all head samples from training. As shown in Fig. 7 (b), this causes noticeable degradation in reconstruction quality, especially in facial regions.

ControlBooth: Using Vanilla DreamBooth. Replacing our customized ControlBooth with DreamBooth (Ruiz et al. 2023) reduces consistency and causes frequent color shifts (e.g., washed-out tones), as shown in Fig. 7 (c), validating the benefit of our targeted training scheme.

BoothAvatar: Using Vanilla SDS. Substituting our 3D-SDS with the SDS (Poole et al. 2022) compromises the generation of A-pose avatars (Fig. 7 (d)), limiting compatibility with downstream applications. Our 3D-SDS design enables more stable and controllable avatar generation.

BoothAvatar: Removing Local Geometric Constraint. Without the local geometric constraint, SDS optimization becomes unstable in fine-grained regions. As illustrated in Fig. 7 (e), this leads to simplified and ambiguous geometry—especially in hands—highlighting the necessity of enforcing localized shape consistency.

BoothAvatar: Disabling Multi-Resolution Sampling. Eliminating multi-resolution sampling and training at a fixed resolution results in slower convergence and degraded visual fidelity (Fig. 7 (f)). In contrast, our hierarchical sampling strategy enhances both training efficiency and rendering quality.

Method	CLIP-I \uparrow		DINO \uparrow		CLIP-T \uparrow	
	body	head	body	head	body	head
PFAvatar(Ours)	0.9125	0.9042	0.8072	0.8617	0.3546	0.3124
PuzzleAvatar	0.8722	0.8652	0.7220	0.8017	0.2836	0.2823
Avatarbooth	0.8533	0.8837	0.6778	0.7869	0.2907	0.2581

Table 2: Qualitative comparison on subject identity preservation. Final rendered avatars from PFAvatar, PuzzleAvatar, and AvatarBooth are compared.

Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
PFAvatar(Ours)	27.576	0.952	0.041
PuzzleAvatar	24.687	0.930	0.062
TECH	23.635	0.919	0.065
AvatarBooth	16.431	0.758	0.153

Table 3: Quantitative comparison of PuzzleIOI dataset. Compared to existing baselines, PuzzleAvatar, TECH, and Avatarbooth. PFAvatar achieves the best performance across all evaluation metrics.

Method	CLIP-I \uparrow		DINO \uparrow		CLIP-T \uparrow	
	full-body	head	full-body	head	full-body	head
Full (a)	0.9125	0.9042	0.8072	0.8617	0.3546	0.3124
w/o (b)	0.8702	0.8503	0.7154	0.8032	0.2912	0.2538
w/o (c)	0.8352	0.8234	0.7091	0.8023	0.2314	0.2313
w/o (d)	0.8021	0.8127	0.7281	0.7819	0.2281	0.2134
w/o (e)	0.8929	0.8831	0.8011	0.8590	0.3257	0.3081
w/o (f)	0.8654	0.8612	0.7486	0.8574	0.2812	0.2434

Table 4: Quantitative comparison of ablation results. Subject fidelity (DINO, CLIP-I) and prompt fidelity (CLIP-T) scores under different ablation settings.

Conclusion

In this paper, we introduce PFAvatar, a novel method for reconstructing high-quality 3D avatars from real-world “Outfit of the Day”(OOTD) photos. Our approach consists of two key stages: (1) fine-tuning a Pose-Aware Diffusion Model using few-shot OOTD examples, and (2) distilling a 3D NeRF-based avatar. By avoiding decomposition and directly modeling full-body appearance, integrating ControlNet for pose estimation, and introducing a Condition Prior Preservation Loss (CPPL), our method achieves consistent and detailed results while mitigating language drift. Additionally, our NeRF-based representation, optimized through canonical SMPL-X space sampling and Multi-Resolution 3D-SDS, effectively handles occlusions and preserves high-frequency textures. Experiments demonstrate that PFAvatar outperforms state-of-the-art methods in reconstruction fidelity, detail preservation, and robustness to occlusions/truncations, significantly advancing the practical generation of 3D avatars from real-world OOTD albums. In addition, we demonstrate the downstream applicability of the reconstructed 3D avatars, including virtual try-on, animation, and human video reenactment.

Acknowledgments

The work was partially supported by National Key R&D Program of China (No. 2023YFF0905102), Key R&D Program of Zhejiang Province (No. 2023C01039)

References

- Alexander, O.; Rogers, M.; Lambeth, W.; Chiang, J.-Y.; Ma, W.-C.; Wang, C.-C.; and Debevec, P. 2010. The Digital Emily Project: Achieving a Photorealistic Digital Actor. *IEEE Computer Graphics and Applications*, 30(4): 20–31.
- Avrahami, O.; Aberman, K.; Fried, O.; Cohen-Or, D.; and Lischinski, D. 2023. Break-A-Scene: Extracting Multiple Concepts from a Single Image. In *SIGGRAPH Asia 2023 Conference Papers*, SA '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703157.
- Bian, W.; Wang, Z.; Li, K.; Bian, J.; and Prisacariu, V. A. 2023. NoPe-NeRF: Optimising Neural Radiance Field with No Pose Prior.
- Cao, Y.; Cao, Y.-P.; Han, K.; Shan, Y.; and Wong, K.-Y. K. 2023. DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models. arXiv:2304.00916.
- Cao, Y.; Cao, Y.-P.; Han, K.; Shan, Y.; and Wong, K.-Y. K. 2024. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 958–968.
- Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. arXiv:2104.14294.
- Chen, H.; Zhang, Y.; Wu, S.; Wang, X.; Duan, X.; Zhou, Y.; and Zhu, W. 2023a. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In *The Eleventh International Conference on Learning Representations*.
- Chen, R.; Chen, Y.; Jiao, N.; and Jia, K. 2023b. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. arXiv:2303.13873.
- Ding, G.; Zhao, C.; Wang, W.; Yang, Z.; Liu, Z.; Chen, H.; and Shen, C. 2024. FreeCustom: Tuning-Free Customized Image Generation for Multi-Concept Composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Fu, X.; Yin, W.; Hu, M.; Wang, K.; Ma, Y.; Tan, P.; Shen, S.; Lin, D.; and Long, X. 2024. GeoWizard: Unleashing the Diffusion Priors for 3D Geometry Estimation from a Single Image. arXiv:2403.12013.
- Fu, Y.; Liu, S.; Kulkarni, A.; Kautz, J.; Efros, A. A.; and Wang, X. 2023. COLMAP-Free 3D Gaussian Splatting.
- Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion.
- Gao, J.; Chen, W.; Xiang, T.; Tsang, C. F.; Jacobson, A.; McGuire, M.; and Fidler, S. 2020. Learning Deformable Tetrahedral Meshes for 3D Reconstruction. arXiv:2011.01437.
- Guo, K.; Xu, F.; Yu, T.; Liu, X.; Dai, Q.; and Liu, Y. 2017. Real-Time Geometry, Albedo, and Motion Reconstruction Using a Single RGB-D Camera. *ACM Trans. Graph.*, 36(4).
- Huang, Y.; Wang, J.; Zeng, A.; Cao, H.; Qi, X.; Shi, Y.; Zha, Z.-J.; and Zhang, L. 2023a. DreamWaltz: Make a Scene with Complex 3D Animatable Avatars. arXiv:2305.12529.
- Huang, Y.; Yi, H.; Xiu, Y.; Liao, T.; Tang, J.; Cai, D.; and Thies, J. 2023b. TeCH: Text-guided Reconstruction of Life-like Clothed Humans. arXiv:2308.08545.
- Işık, M.; Rünz, M.; Georgopoulos, M.; Khakhulin, T.; Starck, J.; Agapito, L.; and Nießner, M. 2023. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *ACM Transactions on Graphics*, 42(4): 1–12.
- Jain, A.; Mildenhall, B.; Barron, J. T.; Abbeel, P.; and Poole, B. 2022. Zero-Shot Text-Guided Object Generation with Dream Fields. arXiv:2112.01455.
- Katzir, O.; Patashnik, O.; Cohen-Or, D.; and Lischinski, D. 2023. Noise-Free Score Distillation. arXiv:2310.17590.
- Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2024. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. arXiv:2312.02145.
- Kocsis, P.; Sitzmann, V.; and Nießner, M. 2024. Intrinsic Image Diffusion for Indoor Single-view Material Estimation. arXiv:2312.12274.
- Kolotouros, N.; Alldieck, T.; Zanfir, A.; Bazavan, E. G.; Fieraru, M.; and Sminchisescu, C. 2023. DreamHuman: Animatable 3D Avatars from Text. arXiv:2306.09329.
- Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-Concept Customization of Text-to-Image Diffusion. arXiv:2212.04488.
- Li, P.; Zheng, W.; Liu, Y.; Yu, T.; Li, Y.; Qi, X.; Li, M.; Chi, X.; Xia, S.; Xue, W.; et al. 2024. PSHuman: Photorealistic Single-view Human Reconstruction using Cross-Scale Diffusion. arXiv preprint arXiv:2409.10141.
- Li, R.; Xiu, Y.; Saito, S.; Huang, Z.; Olszewski, K.; and Li, H. 2020. Monocular Real-Time Volumetric Performance Capture. arXiv:2007.13988.
- Liao, T.; Yi, H.; Xiu, Y.; Tang, J.; Huang, Y.; Thies, J.; and Black, M. J. 2023. TADA! Text to Animatable Digital Avatars. arXiv:2308.10899.
- Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3D: High-Resolution Text-to-3D Content Creation. arXiv:2211.10440.
- Lin, C.-H.; Ma, W.-C.; Torralba, A.; and Lucey, S. 2021. BARF: Bundle-Adjusting Neural Radiance Fields. In *IEEE International Conference on Computer Vision (ICCV)*.
- Lin, L.; Peng, S.; Gan, Q.; and Zhu, J. 2024. FastHuman: Reconstructing High-Quality Clothed Human in Minutes. In *International Conference on 3D Vision, 3DV*.
- Liu, W.; Qiu, Z.; Feng, Y.; Xiu, Y.; Xue, Y.; Yu, L.; Feng, H.; Liu, Z.; Heo, J.; Peng, S.; Wen, Y.; Black, M. J.; Weller, A.; and Schölkopf, B. 2024. Parameter-Efficient Orthogonal Finetuning via Butterfly Factorization. arXiv:2311.06243.

- Martin-Brualla, R.; Radwan, N.; Sajjadi, M. S. M.; Barron, J. T.; Dosovitskiy, A.; and Duckworth, D. 2021. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. arXiv:2008.02268.
- Meuleman, A.; Liu, Y.-L.; Gao, C.; Huang, J.-B.; Kim, C.; Kim, M. H.; and Kopf, J. 2023. Progressively Optimized Local Radiance Fields for Robust View Synthesis. In *CVPR*.
- Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.
- Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.*, 41(4): 102:1–102:15.
- OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.
- Pan, P.; Su, Z.; Lin, C.; Fan, Z.; Zhang, Y.; Li, Z.; Shen, T.; Mu, Y.; and Liu, Y. 2024. HumanSplat: Generalizable Single-Image Human Gaussian Splatting with Structure Priors. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. arXiv:2012.15838.
- Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D using 2D Diffusion. arXiv:2209.14988.
- Qiu, L.; Gu, X.; Li, P.; Zuo, Q.; Shen, W.; Zhang, J.; Qiu, K.; Yuan, W.; Chen, G.; Dong, Z.; and Bo, L. 2025. LHM: Large Animatable Human Reconstruction Model from a Single Image in Seconds. In *arXiv preprint arXiv:2503.10625*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.
- Raj, A.; Kaza, S.; Poole, B.; Niemeyer, M.; Ruiz, N.; Mildenhall, B.; Zada, S.; Aberman, K.; Rubinstein, M.; Barron, J.; Li, Y.; and Jampani, V. 2023. DreamBooth3D: Subject-Driven Text-to-3D Generation. arXiv:2303.13508.
- Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; Zeng, Z.; Zhang, H.; Li, F.; Yang, J.; Li, H.; Jiang, Q.; and Zhang, L. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.
- Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. arXiv:2208.12242.
- Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; and Li, H. 2019. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. arXiv:1905.05172.
- Shen, K.; Guo, C.; Kaufmann, M.; Zarate, J. J.; Valentin, J.; Song, J.; and Hilliges, O. 2023. X-Avatar: Expressive Human Avatars. arXiv:2303.04805.
- Sun, J.; Chen, X.; Wang, Q.; Li, Z.; Averbuch-Elor, H.; Zhou, X.; and Snavely, N. 2022. Neural 3D Reconstruction in the Wild. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, SIGGRAPH '22. ACM.
- Wang, H.; Du, X.; Li, J.; Yeh, R. A.; and Shakhnarovich, G. 2022. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. arXiv:2212.00774.
- Wang, J.; Liu, Y.; Dou, Z.; Yu, Z.; Liang, Y.; Li, X.; Wang, W.; Xie, R.; and Song, L. 2023a. Disentangled Clothed Avatar Generation from Text Descriptions. arXiv:2312.05295.
- Wang, P.; Tan, H.; Bi, S.; Xu, Y.; Luan, F.; Sunkavalli, K.; Wang, W.; Xu, Z.; and Zhang, K. 2023b. PF-LRM: Pose-Free Large Reconstruction Model for Joint Pose and Shape Prediction. arXiv:2311.12024.
- Wang, Q.; Bai, X.; Wang, H.; Qin, Z.; and Chen, A. 2024. InstantID: Zero-shot Identity-Preserving Generation in Seconds. *arXiv preprint arXiv:2401.07519*.
- Wang, S.; Leroy, V.; Cabon, Y.; Chidlovskii, B.; and Revaud, J. 2023c. DUS3R: Geometric 3D Vision Made Easy. arXiv:2312.14132.
- Wang, Z.; Wu, S.; Xie, W.; Chen, M.; and Prisacariu, V. A. 2021. NeRF—: Neural Radiance Fields Without Known Camera Parameters. *arXiv preprint arXiv:2102.07064*.
- Weng, C.-Y.; Curless, B.; Srinivasan, P. P.; Barron, J. T.; and Kemelmacher-Shlizerman, I. 2022. HumanNeRF: Free-Viewpoint Rendering of Moving People From Monocular Video. In *CVPR*, 16210–16220.
- Xia, Y.; Tang, H.; Timofte, R.; and Gool, L. V. 2022. SiNeRF: Sinusoidal Neural Radiance Fields for Joint Pose Estimation and Scene Reconstruction. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press.
- Xiong, Z.; Li, C.; Liu, K.; Liao, H.; Hu, J.; Zhu, J.; Ning, S.; Qiu, L.; Wang, C.; Wang, S.; et al. 2024. MVHumanNet: A Large-scale Dataset of Multi-view Daily Dressing Human Captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19801–19811.
- Xiu, Y.; Ye, Y.; Liu, Z.; Tzionas, D.; and Black, M. J. 2024. PuzzleAvatar: Assembling 3D Avatars from Personal Albums. arXiv:2405.14869.
- Zeng, Y.; Lu, Y.; Ji, X.; Yao, Y.; Zhu, H.; and Cao, X. 2023. AvatarBooth: High-Quality and Customizable 3D Human Avatar Generation. arXiv:2306.09864.
- Zhang, J. Y.; Lin, A.; Kumar, M.; Yang, T.-H.; Ramanan, D.; and Tulsiani, S. 2024. Cameras as Rays: Pose Estimation via Ray Diffusion. arXiv:2402.14817.
- Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543.