DD-NeRF: Double-Diffusion Neural Radiance Field as a Generalizable Implicit Body Representation

Guangming Yao¹, Hongzhi Wu², Yi Yuan ¹*, Kun Zhou²

¹ NetEase Fuxi AI Lab
² State Key Lab of CAD&CG, Zhejiang University

yaoguangming@corp.netease.com, hwu@acm.org yuanyi@corp.netease.com, kunzhou@acm.org

Abstract

We present DD-NeRF, a novel generalizable implicit field for representing human body geometry and appearance from arbitrary input views. The core contribution is a double diffusion mechanism, which leverages the sparse convolutional neural network to build two volumes that represent a human body at different levels: a coarse body volume takes advantage of unclothed deformable mesh to provide the large-scale geometric guidance, and a detail feature volume learns the intricate geometry from local image features. We also employ a transformer network to aggregate image features and raw pixels across views, for computing the final high-fidelity radiance field. Experiments on various datasets show that the proposed approach outperforms previous works in both geometry reconstruction and novel view synthesis quality.

Introduction

Simultaneous reconstruction of body geometry and appearance from a sparse set of views is important in a wide variety of applications, including special effects, game production and virtual reality. However, this inverse problem is challenging, due to its illposedness with highly limited input and the large degree-of-freedom in the output.

Substantial research efforts have been made in the past, particularly with the help of powerful deep learning techniques. When the high-quality scanned 3D body models are available to supervise the network training, neural surface reconstruction can produce remarkable results, using signed distance filed (SDF) as the representation (Saito et al. 2019; 2020; Hong et al. 2021). However, such data are often tedious to acquire and not publicly accessible.

On the other hand, promising results on novel view synthesis are demonstrated with the neural radiance field, or NeRF (Mildenhall et al. 2020; Wang et al. 2021b; Martin-Brualla et al. 2021; Pumarola et al. 2021; Zhang et al. 2020). Unlike previous work that requires 3D supervision, NeRF only needs weak 2D supervision for training. Recently, NeuralBody (Peng et al. 2021) learns a dynamic neural representation from video frames captured by a sparse set of cameras. It requires hours of per-subject optimization, considerably limiting its practical applications. The closest work to ours is the generalizable radiance field (Yu et al. 2021; Chen et al. 2021). It avoids the time-consuming per-subject optimization and can be used to directly synthesize realistic images at novel viewpoints via fast network inference. However, these methods employ a voxel-based shape representation, which produces unsatisfactory results that suffer from bump artifacts.

To overcome the aforementioned weaknesses in existing work, we propose DD-NeRF, a novel generalizable neural representation for human body geometry and appearance (Fig. 2). It requires only weak 2D supervision, and supports fast inference and an arbitrary number of input views. At its core is a new mechanism called double diffusion: multi-view human body features on both the coarse and fine level are diffused into the space, which are subsequently aggregated to produce a high-quality neural reflectance field.

We make the key observation that many image features extracted by previous work are irrelevant or redundant for representing human bodies. Therefore, instead of storing dense neural features, we maintain in our feature volume only a sparse set of features that have high relevance with the body geometry and appearance. Specifically, we first sample 3D points on deformable human model (Loper et al. 2015) as anchors, and project them to each input view. Next, we sample pixel-aligned image features from different views and calculate their means and variances as the multi-view local features. Then, multi-layer perceptron (MLP) is adopted to decode the trilinearly interpolated neural features in the coarse- and fine-scale volumes to obtain signed distances as the implicit geometry. For appearance reasoning, we utilize the transformer to fuse multi-view image features and raw pixels. To ensure that the entire representation is learnable with weak 2D supervision, we employ the differential SDF rendering formulation of NeuS (Wang et al. 2021a).

In summary, this work has the following contributions:

• We present a novel generalizable neural representation of body geometry and appearance. By taking advantage of the coarse body prior and multi-view information, we design a geometry-aware method that can learn to recover high-quality 3D body. Meanwhile, our method supports further optimization in inference time for quality improvement.

• We propose a double diffusion mechanism, to effectively integrate multi-view features and coarse body prior for
Multi-view Images Reconstructed surface Synthesized Novel views

Figure 1: Our method can produce both detailed body surface and high-fidelity novel views, given a sparse set of input posed images.

Implicit field regression.

• Our method outperforms state-of-the-art techniques in both qualitative and quantitative comparisons.

Related Works

Single-View Body Shape Reconstruction

To regularize this highly under-constrained problem, many previous works rely on parametric body templates (Kanazawa et al. 2018; Kolotouros, Pavlakos, and Daniilidis 2019; Alldieck et al. 2019, 2018). Recently, researches focus on regressing a “freeform” 3D body shape. Impressive results are demonstrated with volumetric representations in modeling geometric details like clothing. For instance, BodyNet (Varol et al. 2018) learns to directly produce a voxel representation of a body in an end-to-end fashion. DeepHuman (Zheng et al. 2019) introduces the discretized volumetric representation for enhanced resolution. These explicit representations require a large memory footprint to express a high-precision model. On the other hand, implicit representations are more memory efficient. Pixel-Aligned Implicit Function (PIFU) regresses a function to determine the occupancy for any given 3D location (Saito et al. 2019). This approach infers both 3D surface and texture from a single image, and can handle highly detailed human bodies. Furthermore, PIFUHD (Saito et al. 2020) proposes a multi-level architecture. The coarse level focuses on holistic reasoning. NormalGAN (Wang et al. 2020) predicts the back-view RGBD image from the front view using adversarial learning.

All above methods are based on supervised learning. In comparison, our method only needs the posed multi-view images as weak 2D supervision.

Multi-View Body Shape Reconstruction

Traditional approaches in this category can produce a detailed human model with multi-view-stereo constraints. They focus on multi-view correspondence matching within the visual hull (Esteban and Schmitt 2004; Furukawa and Ponce 2006; Vlasic et al. 2008). Some recent methods apply deep learning to compute a high-quality result in a fully data-driven fashion. For instance, (Choy et al. 2016) deploys 3D convolution to learn 3D voxels. In addition, multi-view image features are combined with differentiable unprojection operations (Ji et al. 2017; Kar, Häne, and Malik 2017). Moreover, volumetric occupancy field is proposed to learn dynamic clothed body performances from sparse viewpoints (Gilbert et al. 2018; Huang et al. 2018). StereoPIFU (Hong et al. 2021) integrates the geometric constraints of stereo vision with implicit function representation, to alleviate the depth ambiguity problem. DoubleField (Shao et al. 2021) combines the surface field and radiance field for body representation with the supervision of ground-truth meshes. In comparison, our method only requires weak 2D supervision to output high-quality 3D geometry.

Neural Representations

Recent researches on neural implicit functions have emerged as an effective representation to learn 3D scenes from 2D images with differential rendering. NeRF (Mildenhall et al. 2020) learns implicit fields of density and color with volumetric rendering, producing impressive results on novel view synthesis. Following NeRF, NeuralBody (Peng et al. 2021) captures body performance from video frames and extracts the reconstructed surface by performing marching cube (Lorensen and Cline 1987) on a learned volumetric density field. Furthermore, UNISURF (Oechsle, Peng, and...)
Given multi-view input images and a corresponding SMPL model, the double diffusion builds two volumes to represent the coarse body prior and local features that depict the detailed geometry. Next, implicit field regression predicts the SDF and radiance. All modules are learned jointly, by minimizing the differences between the input and rendered images.

Figure 2: An overview of DD-NeRF. Given multi-view input images and a corresponding SMPL model, the double diffusion builds two volumes to represent the coarse body prior and local features that depict the detailed geometry. Next, implicit field regression predicts the SDF and radiance. All modules are learned jointly, by minimizing the differences between the input and rendered images.

Geiger [2021] improves the reconstruction quality, by shrinking the sampling region of volume rendering during optimization. DVR (Niemeyer et al. 2020), IDR (Yariv et al. 2020) and NLR (Kellnhofer et al. 2021) directly learn the surface radiance with implicit gradients. NeuS (Wang et al. 2021a) represents the scene with an SDF and thus can naturally extract the 3D surface with level set. We also employ SDF to represent the shape, and utilize the volumetric SDF rendering to train our network in an end-to-end way.

The above existing methods require a tedious per-subject optimization. The potential subject-independent priors are not exploited. To tackle this issue, PixelNeRF (Yu et al. 2021) and MVSNeRF (Chen et al. 2021) propose the conditional NeRF, which is trained across multiple subjects and can perform novel view synthesis via fast network inference. Our method also learns priors across different subjects. Therefore, it can run in a feed-forward way, given a sparse set of unseen images. We deploy the double diffusion mechanism to learn the coarse and detail features for geometric understanding. Additionally, we utilize the transformer to integrate multi-view information. As a result, our method outperforms concurrent works on generalizable radiance field reconstruction in terms of reconstruction quality.

**Methodology**

Given a sparse set of multi-view posed body images, our method computes an implicit radiance field that represents the geometry and appearance of the subject. We denote the input images as $I = \{I_1, I_2, ..., I_{N_v}\}$, where $N_v$ is the number of pre-calibrated cameras. The calibration results are stored as $\Phi = \{\Phi_1, \Phi_2, ..., \Phi_{N_v}\}$. In general, our network can be viewed as a radiance field function as follows:

$$s, c = F(x, d; I, \Phi),$$

where $x$ represents a 3D location, $d$ is a viewing direction, $s$ denotes a signed distance value at $x$ and $c$ is an RGB color as its appearance. The output radiance field can be used to synthesize a novel view via differentiable ray marching, or extract 3D surfaces with marching cube.

An overview of our model is illustrated in Fig. 2. There are three components, the image encoder, the double diffusion, and the implicit field regression. The image encoder is a stacked hourglass neural network (Newell, Yang, and Deng 2016) for image feature extraction. Next, the double diffusion builds two volumes, a coarse body volume and a detail feature volume. While the former represents the coarse body prior, the latter samples local image features that depict the detailed geometry for high-fidelity body representation. In addition, the implicit field regression consists of two sub-modules for regressing the SDF and radiance, respectively. The SDF regression generates the signed distance from trilinearly interpolated neural features inside both volumes. The radiance regression takes as input surface normal, multi-view raw pixels, image features and the view direction, and finally produces an RGB color as output. With the regressed signed distance and color, we can synthesize an image from any viewpoint by differential SDF rendering (Wang et al. 2021a). All the above modules are learned jointly, by minimizing the differences between the input and rendered images.

**Double Diffusion**

The idea of double diffuse is to exploit both coarse human body prior and detailed multi-view image features. We build a coarse body volume and a detail feature volume in parallel. We calculate the 3D bounding box to cover all possible SMPL models, and discretize the bounding box to build volumes. We use a resolution of $256^3$ for the coarse body volume, and $512^3$ for the detail feature one for capturing high-frequency geometric details.

The coarse body volume is similar to (Peng et al. 2021). While they utilize the volume to represent the dynamic body, we employ it to represent the static coarse human geometry. Given the multi-view images and the corresponding SMPL
parameters estimated with (Zheng et al. 2021), we obtain the vertices $V = \{v_1, v_2, ..., v_{6890}, v_i \in \mathbb{R}^3\}$ from the template model. Indices of the vertices are fed into the embedding network to compute the structured latent code $Z \in \mathbb{R}^{6890 \times 16}$, which represents the coarse body geometry and serves as a prior for implicit field reconstruction. Inspired by (Peng et al. 2021), we employ SparseConvNet (Graham, Engelcke, and Van Der Maaten 2018) to diffuse the structured latent codes into the volume, which will be further tri-linearly interpolated.

The detail feature volume is designed to integrate multi-view image features to facilitate detailed geometric understanding. Here the first step is to sample the on-body and far-body anchors. Similar to (Yu et al. 2018), the on-body anchors are relatively dense vertices computed with subdivision on the SMPL model. The far-body anchors are randomly sampled on the exterior surfaces of spheres centered at each vertex of the SMPL model, we empirically set the radius as 0.05m to account for objects that are relatively far away from the SMPL surface (e.g., loose clothing, long hair). In experiments, we uniformly sample 12446 far-body anchors, and utilize Trimesh (Dawson-Haggerty et al. 2019) to obtain 27554 on-body anchors from 6890 SMPL vertices with subdivision. Next, we project the anchors to different views and sample pixel-aligned multi-view features. The procedure can be expressed as:

$$\hat{v}_j^i = K_j \cdot [R_j^T \cdot R \cdot (a_i - T) + T_j],$$

$$f_j^i = E_j(\hat{a}_j^i),$$

where $a_i$ is 3D position the $i$-th anchor, $R$ and $T$ is the estimated rotation and translation of SMPL model. $K_j$ and $R_j/T_j$ are the intrinsic/extrinsic matrix of $j$-th camera. $E_j$ denotes the encoded image feature of $j$-th view, $f_j^i \in \mathbb{R}^c$ denotes the sampled $c$ channel features of $i$-th anchor from $j$-th view. Given the $N_e$ input views, we sampled $f_j = \{f_1, f_2, ..., f_{N_e}\}$ features for each anchor. We concatenate their mean and variance as the feature vector $\hat{f}_i \in \mathbb{R}^{2c}$, as common in multi-view geometry learning (Yao et al. 2018; Cheng et al. 2020). Then, our MLP encodes the feature vector across input views, explaining the appearance variations caused by both scene geometry and view-dependent shading effects. Finally, we apply the SparseConvNet again to efficiently diffuse the sparse anchor-wise feature to the detail feature volume.

Our double-diffusion mechanism incorporates the coarse body prior with elaborated geometry information. Thanks to the detailed local features, the learned implicit field can preserve the local details depicted in the input images.

**Implicit Field Regression**

This step consists of two sub-modules, one for SDF regression $F_S$ and the other radiance regression $F_R$. To obtain a high-quality output surface, we use signed distance instead of volumetric density to represent the body geometry. For a 3D point $x$ in the 3D bounding box $\Omega^3$, we trilinearly interpolate neural features in the coarse body volume and the detail feature one to obtain its coarse feature $f_c$ and detail feature $f_d$. We adopt MLP to regress signed distance $s$ conditioned on $x$, $f_c$ and $f_d$ as below:

$$s, f_s = F_S(x, f_c, f_d).$$

Here $f_s$ is the geometry-related feature, which will be fed to radiance regression. Similar with NeRF (Mildenhall et al. 2020), $x$ is mapped to a higher dimensional space with positional encoding. After computing $s$ for discrete $x$ around
the bounding volume, the body surface $\mathcal{S}$ can be expressed by the zero level-set of SDF, i.e., $\mathcal{S} = \{ x \in \Omega^3 | s = 0 \}$.

Once the geometry is obtained, the next step is to regress the corresponding appearance. Given the previous geometry-related feature $f_s$, we take the view direction, the geometry and the image feature into consideration. The radiance regression can be expressed by:

$$c = F_R(f_s, g, x, d, \{f_{xi}\}, \{p_{xi}\}|i = 1, 2, ..., N_c),$$  \hspace{1cm} (4)

where $d$ is the view-direction, and $g = \nabla F_S(x; f_c, f_d)$ represents the normal of surface $\mathcal{S}$ at position $x$. We transform $x$ to $i$-th view, and then sample pixel-aligned image feature $f_{xi}$, and the raw RGB value $p_{xi}$. Notably, $p_{xi}$ and $d$ are also mapped to a higher dimensional space with positional encoding for the learning of high-frequency variations.

Specifically, inspired by recent progress in NLP (Vaswani et al. 2017; Raffel et al. 2019), we adopt a transformer to effectively fuse features across views for appearance reasoning. The transformer consists of an encoder and a decoder. The encoder $E$ aims to encode the multi-view feature with stacked multi-head attention layers to yield the fused feature $\hat{f}_x$, computed as:

$$\hat{f}_x = E(\{f_{xi}\}, \{p_{xi}\}|i = 1, 2, ..., N_c).$$  \hspace{1cm} (5)

The MLP-based decoder $D$ regresses the final radiance:

$$c = D(f_s, g, x, d, \hat{f}_x).$$  \hspace{1cm} (6)

### Differential SDF Rendering

To optimize our representation, we adopt the SDF-based differential render (Wang et al. 2021a) to render images, which are compared with the input ones as supervision.
For a pixel in the target image, we denote the ray emitted from this pixel as \( \{x_i = o + td_i\mid t_{\min} \leq t \leq t_{\max}\} \), where \( o \) is the center of camera, \( d \) is a unit vector representing the direction of the ray, \( t_{\min}/t_{\max} \) is the minimal/maximal ray length, respectively. We accumulate the colors along the ray to predict the radiance at a pixel color by:

\[
\hat{C} = \sum_{i=1}^{n} T_i \alpha_i c_i. \tag{7}
\]

Here \( T_i \) is the discrete accumulated transmittance defined by \( T_i = \prod_{j=1}^{k-1}(1 - \alpha_j) \), and \( \alpha_i \) is discrete opacity value defined as:

\[
\alpha_i = \max \left( \Phi_s(f(s_i) - \Phi_s(s_{i+1})), 0 \right), \tag{8}
\]

\[
\Phi_s = (1 + e^{-k s})^{-1},
\]

where the inverse standard deviation \( k \) is a learnable scalar that increases with the number of training iterations.

### Loss Function

To learn the implicit body representation, we penalize the difference between the rendered pixel color and its counterpart in the input image with weak 2D supervision (i.e., the image and mask). We randomly select the input views, and then randomly sample a batch of pixels and the corresponding rays in the sampled views to train our network and the inverse standard deviation \( k \). The point sampling size is \( n \) and the batch size is \( m \). Similar to (Wang et al. 2021a; Yariv et al. 2020), our loss function \( L \) consists of a rendering loss \( L_r \), an Eikonal loss \( L_e \) and an optional mask loss \( L_{mask} \):

\[
L = \lambda_r L_r + \lambda_e L_e + \lambda_{mask} L_{mask}, \tag{9}
\]

where \( \lambda \) denotes the corresponding weights.

The core loss is the rendering loss \( L_r \), which aims to directly measure the difference between the rendered color and the corresponding ground-truth. The rendering loss \( L_r \) is the mean pixel-wise L1 loss defined as:

\[
L_r = \frac{1}{m} \sum_{i=1}^{m} |C_i - \hat{C}_i|. \tag{10}
\]

Here \( \hat{C}_i \) is the predicted pixel color and \( C_i \) is the ground-truth label.

To Eikonal loss servers as the Implicit Geometric Regularization (IGR) (2020), enforcing our SDF regression to be a approximately signed distance function. The Eikonal loss is calculated by:

\[
L_e = \frac{1}{nm} \sum_{i,j} \left( |\nabla(s_{ij})| - 1 \right)^2. \tag{11}
\]

The mask loss \( L_{mask} \) is defined as:

\[
L_{mask} = BCE(M_k, \hat{O}_k), \tag{12}
\]

where \( \hat{O}_k = \sum_{i=1}^{n} T_{ki} \alpha_k i \) is the sum of weights along the view ray, \( M_k \in \{0, 1\} \) is its optional mask value and BCE is the binary cross entropy loss.

### Experiments

#### Datasets

We perform experiments on the synthesized dataset Twidom\(^1\) captured datasets THuman2.0 (Zheng et al. 2021) and ZJU-Mocap (Peng et al. 2021). We use 1200 models from Twidom for training, 300 models for evaluation. For THuman, 400 models are used for training, and 100 for evaluation. All data in ZJU-Mocap are used for evaluation.

To render a scanned body model, we place it in the center of a unit sphere, and orient the camera toward the sphere center with a distance of 2.4m. We move the camera around the sphere, sample the yaw angle from \(-30^\circ\) to \(60^\circ\) with a step size of \(10^\circ\), and sample the roll angle from \(0^\circ\) to \(360^\circ\) with a step size of \(24^\circ\). For each body model, we render 135 images with a resolution of \(1024^2\).

#### Implement Details

The SMPL parameters are estimated with EasyMo-cap (Dong et al. 2021) from the input images. Note that the sampling strategy can heavily influence the final results as in other works on NeRF. Therefore, we adopt a similar hierarchical sampling method from (Wang et al. 2021a). We use the Adam optimizer (Kingma and Ba 2014), and set the learning rate to \(1 \times 10^{-3}\). The loss weights are set to 10, 1, 1 for \( \lambda_r, \lambda_e \) and \( \lambda_{mask} \), respectively.

#### Qualitative Comparison

We compare our approach with recent baseline methods to validate efficiency. PixelNeRF (Yu et al. 2021), PIFu (Saito et al. 2019) and our method are all generalizable methods, which are trained on many subjects and tested with randomly sampled unseen person. NeuralBody (Peng et al. 2021) is optimized for a specific person. We train it with 500 epochs for each tested subject. It is worth mentioning that, we use the public-released weights of PIFu (Saito et al. 2019) for comparison, which is trained with the supervision of scanned 3D meshes and takes a single image as input. For the remaining methods, we use 4 views (i.e., the left, front, right and back views) as the input.

We conduct comparison experiments on both the geometry (Fig. [3]) and the appearance (Fig. [4]). Please also refer to the accompanying for video animated results with varying views. The results show that PIFu (Saito et al. 2019) suffers from over-smoothing and reconstructs artifacted geometry and texture. Thanks to the double diffusion, our method can effectively learn both the coarse body prior and the detail appearance. PixelNeRF (Yu et al. 2021) and NeuralBody (Peng et al. 2021) generate plausible novel views. But bump artifacts appear in their generated shape results. Since we employ SDF to represent the geometry, our results do not exhibit such artifacts.

#### Quantitative Comparison

We adopt a variety of metrics to quantitatively evaluate our results at the image- and geometry-level. Specifically, we...
Table 1: Quantitative comparison of both human body geometry and appearance reconstruction. “Ft” indicates that fine-tuning is applied. Up/down arrows correspond to higher/lower values for better performance. Bold and underlined numbers correspond to the best and the second-best values for each metric.

| Model          | Supervision | PNSR↑ | SSIM↑ | LPIPS↓ | PNSR↑ | SSIM↑ | LPIPS↓ | PNSR↑ | SSIM↑ | LPIPS↓ |
|----------------|-------------|-------|-------|--------|-------|-------|--------|-------|-------|--------|
|                |             | THuman2.0 | Twindom | ZJU-Mocap | THuman2.0 | Twindom | ZJU-Mocap |
| PixelNeRF      | images      | 16.57  | 0.67  | 0.36   | 14.23 | 0.51  | 0.44   | 18.50 | 0.70  | 0.40   |
| PixelNeRF(Ft)  | images      | 19.50  | 0.82  | 0.32   | 17.75 | 0.62  | 0.26   | 20.70 | 0.81  | 0.31   |
| PIFu           | 3D meshes   | 18.39  | 0.58  | 0.33   | 16.43 | 0.70  | 0.30   | 19.12 | 0.75  | 0.43   |
| NeuralBody     | images      | 21.90  | 0.85  | 0.20   | 23.52 | 0.76  | 0.26   | 25.65 | 0.92  | 0.27   |
| Ours           | images      | 21.10  | 0.78  | 0.21   | 22.15 | 0.78  | 0.26   | 23.95 | 0.80  | 0.23   |
| Ours(Ft)       | images      | 23.08  | 0.85  | 0.16   | 27.67 | 0.84  | 0.21   | 25.60 | 0.88  | 0.22   |

Table 2: Quantitative comparison of surface reconstruction.

| Model          | Chamfer↓ | P2S↓ | Chamfer↓ | P2S↓ |
|----------------|----------|------|----------|------|
|                | THuman2.0 | Twindom | THuman2.0 | Twindom |
| PixelNeRF      | 1.597    | 1.146 | 1.528    | 1.126|
| PixelNeRF(Ft)  | 0.930    | 0.913 | 0.925    | 0.741|
| PIFu           | 1.510    | 1.524 | 1.708    | 1.630|
| NeuralBody     | 0.915    | 0.931 | 0.815    | 0.725|
| Ours           | 0.775    | 0.690 | 0.737    | 0.701|
| Ours(Ft)       | 0.709    | 0.634 | 0.690    | 0.631|

Table 3: Quantitative ablation study on the transformer.

| Model     | Chamfer↓ | P2S↓ | PSNR↑ | SSIM↑ |
|-----------|----------|------|-------|-------|
| w/o transformer | 0.861 | 0.81 | 19.89 | 0.66  |
| Ours      | 0.775    | 0.69 | 21.10 | 0.78  |

Table 4: Quantitative ablation study with different numbers of views.

| Views | Chamfer↓ | P2S↓ | PSNR↑ | SSIM↑ |
|-------|----------|------|-------|-------|
| 1     | 0.814    | 0.796| 19.76 | 0.77  |
| 3     | 0.787    | 0.701| 21.09 | 0.80  |
| 6     | 0.708    | 0.620| 22.50 | 0.88  |
| 9     | 0.680    | 0.617| 23.52 | 0.86  |

employ peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) and learned perceptual image patch similarity (LPIPS) (Zhang et al. 2018) to evaluate the similarity between the ground-truth and the synthesized image. The metrics of Chamfer distance (Chamfer) and average point-to-surface Euclidean distance (P2S) are applied to assess geometric quality.

For image-level comparison, we randomly select 20 target views as the ground truth. In the geometry-level comparison, marching cube is performed for surface extraction. To further inspect the capability of generalizable methods, we perform fine-tuning on PixelNeRF and our method with 100 epochs.

The quantitative comparison of image- and geometry-level are reported in Tab. 1 and 2, respectively. Our method outperforms others in most metrics, owing to the double diffusion that provides both the coarse and detailed geometric guidance. Notably, with the fine-turning in inference, our method achieves the highest scores in novel view synthesis. Even for the fast inference without fine-turning, our method achieves the highest scores at Chamfer and P2S, indicating the effectiveness of our SDF-based body representation. Moreover, all metrics are improved after fine-tuning, indicating that our method can learn a better representation when we perform specific-subject optimization based on the result of generalizable fast inference.

Ablation Studies

We conduct ablation studies to validate our method on fast inference. We first analyze the effect of transformer and design the “w/o transformer” variant: we remove the transformer from the full model, and multi-view features are stacked as input into the radiance decoder. The results are reported in Tab. 3. Most metrics are negatively changed with the absence of transformer. This indicates that the transformer can effectively integrate multi-view information and produce more satisfactory results. We then inspect the impact of the number of views. As reported in Tab. 4, the increase of the number of input views improves the performance of novel view synthesis.

Conclusion and Future Work

We propose a novel framework for generalizable neural representation for both human body geometry and appearance, from a sparse set of multi-view images. Our method deploys the double diffusion mechanism to harness a coarse body prior as well as aggregate detail features. It enable the efficient regression of both the signed distance and color. Moreover, only weak 2D supervision is needed with the help of SDF-based differential rendering. Our method considerably outperforms concurrent works in terms of reconstruction quality. In the future, we plan to consider the lighting condition and compute relightable radiance field for a better appearance understanding.
References
Alldieck, T.; Magnor, M.; Bhatnagar, B. L.; Theobalt, C.; and Pons-Moll, G. 2019. Learning to reconstruct people in clothing from a single RGB camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1175–1186.

Alldieck, T.; Magnor, M.; Xu, W.; Theobalt, C.; and Pons-Moll, G. 2018. Detailed human avatars from monocular video. In 2018 International Conference on 3D Vision (3DV), 98–109. IEEE.

Chen, A.; Xu, Z.; Zhao, F.; Zhang, X.; Xiang, F.; Yu, J.; and Su, H. 2021. MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo. arXiv preprint arXiv:2103.15595.

Cheng, S.; Xu, Z.; Zhu, S.; Li, Z.; Li, L. E.; Ramamoorthi, R.; and Su, H. 2020. Deep stereo using adaptive thin volume representation with uncertainty awareness. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2524–2534.

Choy, C. B.; Xu, D.; Gwak, J.; Chen, K.; and Savarese, S. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In European conference on computer vision, 628–644. Springer.

Dawson-Haggerty et al. 2019. trimesh.

Dong, J.; Fang, Q.; Jiang, W.; Yang, Y.; Bao, H.; and Zhou, X. 2021. Fast and Robust Multi-Person 3D Pose Estimation and Tracking from Multiple Views. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9054–9063.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end Recovery of Human Shape and Pose. In Computer Vision and Pattern Recognition (CVPR).

Kar, A.; Häne, C.; and Malik, J. 2017. Learning a multi-view stereo machine. arXiv preprint arXiv:1708.05375.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-nerf: Neural radiance fields for dynamic human shape reconstruction and novel view synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4287–4297.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Kolotouros, N.; Pavlakos, G.; and Daniilidis, K. 2019. Convolutional Mesh Regression for Single-Image Human Shape Reconstruction. In CVPR.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. ACM transactions on graphics (TOG), 34(6): 1–16.

Lorenzen, W. E.; and Cline, H. E. 1987. Marching cubes: A high resolution 3D surface construction algorithm. ACM siggraph computer graphics, 21(4): 163–169.

Martin-Brualla, R.; Radwan, N.; Sajjadi, M. S.; Barron, J. T.; Dosovitskiy, A.; and Duckworth, D. 2021. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7210–7219.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In European conference on computer vision, 405–421. Springer.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In European conference on computer vision, 483–499. Springer.

Niemeyer, M.; Mescheder, L.; Oechsle, M.; and Geiger, A. 2020. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3504–3515.

Oechsle, M.; Peng, S.; and Geiger, A. 2021. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. arXiv preprint arXiv:2104.10078.

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. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9054–9063.

Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In Proceedings of the IEEE International Conference on Computer Vision, 2307–2315.
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. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2304–2314.

Saito, S.; Simon, T.; Saragih, J.; and Joo, H. 2020. Pi-fuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 84–93.

Shao, R.; Zhang, H.; Zhang, H.; Cao, Y.; Yu, T.; and Liu, Y. 2021. DoubleField: Bridging the Neural Surface and Radiance Fields for High-fidelity Human Rendering. arXiv:2106.03798.

Varol, G.; Ceylan, D.; Russell, B.; Yang, J.; Yumer, E.; Laptev, I.; and Schmid, C. 2018. Bodynet: Volumetric inference of 3d human body shapes. In Proceedings of the European Conference on Computer Vision (ECCV), 20–36.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In Advances in neural information processing systems, 5998–6008.

Vlasic, D.; Baran, I.; Matusik, W.; and Popović, J. 2008. Articulated mesh animation from multi-view silhouettes. In ACM SIGGRAPH 2008 papers, 1–9.

Wang, L.; Zhao, X.; Yu, T.; Wang, S.; and Liu, Y. 2020. NormalGAN: Learning Detailed 3D Human from a Single RGB-D Image. In European Conference on Computer Vision, 430–446. Springer.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021a. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. arXiv preprint arXiv:2106.10689.

Wang, Z.; Wu, S.; Xie, W.; Chen, M.; and Prisacariu, V. A. 2021b. NeRF++: Neural Radiance Fields Without Known Camera Parameters. arXiv preprint arXiv:2102.07064.

Yao, Y.; Luo, Z.; Li, S.; Fang, T.; and Quan, L. 2018. Mvs-net: Depth inference for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision (ECCV), 767–783.

Yariv, L.; Kasten, Y.; Moran, D.; Galun, M.; Atzmon, M.; Basri, R.; and Lipman, Y. 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. arXiv preprint arXiv:2003.09852.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelfnerf: Neural radiance fields from one or few images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4578–4587.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In CVPR.

Zheng, Y.; Shao, R.; Zhang, Y.; Zheng, Z.; Yu, T.; and Liu, Y. 2021. DeepMultiCap: Performance Capture of Multiple Characters Using Sparse Multiview Cameras. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 7739–7749.