IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 February 2024, accepted 8 March 2024, date of publication 21 March 2024, date of current version 11 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3380059

RESEARCH ARTICLE

Single-View 3D Garment Reconstruction Using Neural Volumetric Rendering

YIZHENG CHEN[1], RENGAN XIE[2], SEN YANG[1], LINCHEN DAI[1], HONGCHUN SUN[3], YUCHI HUO[1,2], AND RONG LI[1]

[1]Zhejiang Laboratory, Hangzhou 311121, China
[2]Zhejiang University, Hangzhou 310027, China
[3]China Mobile (Hangzhou) Information Technology Company Ltd., China

Corresponding authors: Yuchi Huo (eehyc0@zju.edu.cn) and Rong Li (lirong@zhejianglab.edu.cn)

This work was supported in part by NSFC under Grant 62441205, in part by the Key Research and Development Program of Zhejiang Province under Grant 2023C01039, and in part by the Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

ABSTRACT Reconstructing 3D garment models usually requires laborious data-fetching processes, such as expensive lidar, multiple-view images, or SMPL models of the garments. In this paper, we propose a neat framework that takes single-image inputs for generating pseudo-sparse views of 3D garments and synthesizing multi-view images into a high-quality 3D neural model. Specifically, our framework combines a pretrained pseudo sparse view generator and a volumetric signed distance function (SDF) representation-based network for 3D garment modeling, which uses neural networks to represent both the density and radiance fields. We further introduce a stride fusion strategy to minimize the pixel-level loss in key viewpoints and semantic loss in random viewpoints, which produces view-consistent geometry and sharp texture details. Finally, a multi-view rendering module utilizes the learned SDF representation to generate multi-view garment images and extract accurate mesh and texture from them. We evaluate our proposed framework on the Deep Fashion 3D dataset and achieve state-of-the-art performance in terms of both quantitative and qualitative evaluations.

INDEX TERMS Computer graphics, garment reconstruction, single view, 3D reconstruction.

I. INTRODUCTION

Recently, image-based 3D reconstruction has made significant progress benefiting from neural rendering [1]. It also enabled the realization of implicit single-view garment reconstruction, which involves generating accurate texture and geometry of garments from a single image. These advances have resulted in improved performance and increased efficiency of garment reconstruction methods, leading to potential applications in various fields, such as fashion, e-commerce, and virtual try-on systems.

Several approaches have been proposed for 3D reconstruction or garment reconstruction in the past. In the context of 3D reconstruction, methods such as depth estimation using stereo images, structure-from-motion, and multi-view stereo [2], [3] have been used to generate 3D models from multiple images or viewpoints. However, these methods require multiple images and/or complex equipment, which may not always be feasible. In contrast, implicit reconstruction methods [1], [4] use deep learning techniques to learn a mapping between 2D images and 3D garment shapes, requiring single or sparse views as input. This makes them more practical for real-world applications. In the context of garment reconstruction, several methods have been proposed that leverage geometric and physical models of garments, such as deformable models, and physically-based models [5]. However, these methods can be computationally expensive and require specialized equipment for capturing the physical properties of the garment, such as deformation and texture.

In this paper, we introduce a novel framework for the implicit single-view reconstruction of garments that addresses these challenges. Our approach is designed to generate accurate 3D models of garments from a single image or viewpoint, taking into account the fine-grained details and intricate textures of garments, and accounting for the complex and non-rigid nature of garments. We evaluate our approach on several benchmark datasets and demonstrate its

The associate editor coordinating the review of this manuscript and approving it for publication was Songwen Pei.

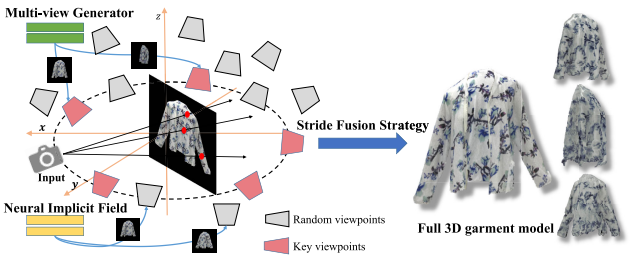

FIGURE 1. Given only a single garment image as input, our framework effectively fuses the weakly-aligned pseudo images predicted by the multi-view generator from sparse key viewpoints, and conducts semantic matching in random views to fill the gaps between key viewpoints, resulting in a high-quality full 3D model.

superiority over existing state-of-the-art methods. The main contributions include:

- We introduce a two-stage single-view 3D reconstruction framework. It utilizes an image-based generator to create a set of weakly-aligned key viewpoints, and then fuse these views into a view-consistent 3D SDF. The image-based generator can preserve reasonable texture details from any viewpoint, and the fusion stage ensures geometric and semantic consistency across different views.

- We propose a stride fusion strategy to synthesize these view-inconsistent pseudo views into 3D models. Specifically, we only optimize pixel-level MSE losses in some keyframes that have little overlapping between each other, and then utilize a self-supervised semantic loss to connect these striding keyframes to achieve view consistency.

- We leverage multi-resolution feature grids with MLP decoder to improve both the rendering clarity and geometric accuracy of reconstructed garments, resulting in high-quality output.

II. RELATED WORK

A. SINGLE-VIEW 3D RECONSTRUCTION

Reconstructing 3D objects from a single view is a challenging problem, as it is ill-conditioned and requires reconstructing the 3D structure of the scene from just one viewpoint.

One approach is to rely on collections of 3D primitives to approximate the target shape explicitly. These works obtain object embeddings from input RGB images and map them to the 3D space. Various 3D object representation methods are employed, such as mesh [6], [7], [8], point clouds [9], [10], and voxel [11], [12]. And the methods of embedding and mapping are influenced by 3D object representation methods. Some others leverage cues like texture [13] and defocus [14] to understand 3D shapes from a single image. The effectiveness of these approaches relies on the technique of estimating depth cues from images. In addition, [9] directly regresses the point clouds from the image using learned priors to complete the information of invisible parts.

Recently, there has been a remarkable development of NeRF-based approaches [15], [16] for 3D reconstruction, following the success of neural radiance fields (NeRF) [1]. Some researchers focus on improving the accuracy of sparse view reconstruction [17], [18], [19]. Furthermore, works like PixelNeRF [4] and PVSeRF [20] aim to reconstruct 3D scenes from a single image by incorporating prior knowledge of the object's structure. These methods train their models on ShapeNet [21], a database containing objects of simple shapes with available 3D annotation.

B. NOVEL VIEW SYNTHESIS

Novel view synthesis is a task of generating novel views of a scene from a new viewpoint. The generation of high-quality images from an unseen perspective is a challenging task, particularly when the object's position and orientation in the scene are not known.

One of the popular approaches to novel view synthesis involves using Generative Adversarial Network (GAN) [22]. In prior works, researchers explored the use of GAN models to discover latent semantic directions that could manipulate object rotation without reliance on underlying 3D models [21], [23], [24], [25]. Several recent works [26], [27] have extended this GAN-based approach to NeRF models and trained them using adversarial losses, resulting in significant performance improvements.

Another promising approach to novel view synthesis involves using diffusion models, specifically diffusion denoising probabilistic models. Diffusion models are a class of generative models that make use of a Markovian noising process to iteratively reverse the noise. In recent years, several researchers [28], [29], [30], [31] have explored the use of diffusion models in conjunction with radiance fields and have demonstrated excellent results in tasks such as conditional synthesis, completion, and other related tasks [32], [33].

In summary, while both GAN-based and diffusion models have individually shown remarkable progress in novel view synthesis, recent research has also explored their combination with NeRF models for even better results. Our work utilizes a GAN-based model to generate sparse pseudo views, but it is also possible to extend to a large diffusion model. Theoretically, the generation model is an orthogonal study of our fusion framework.

C. 3D GARMENT RECONSTRUCTION

Reconstructing textured 3D garments is a complex task faced with various challenges. Existing methods mainly rely on predefined 3D templates [5], [34], which provide strong priors for constraining the solution space of shape estimation. However, these methods are limited in their ability to model arbitrary clothing styles. Another challenge is to obtain a high-frequency displacement that accurately represents the geometric details of the garment. DeepWrinkles [35] aims to capture the fine details of the garment's geometry to achieve a high-quality 3D reconstruction. However, this method relies on pose estimation and cannot reconstruct individual in-shop clothing items.

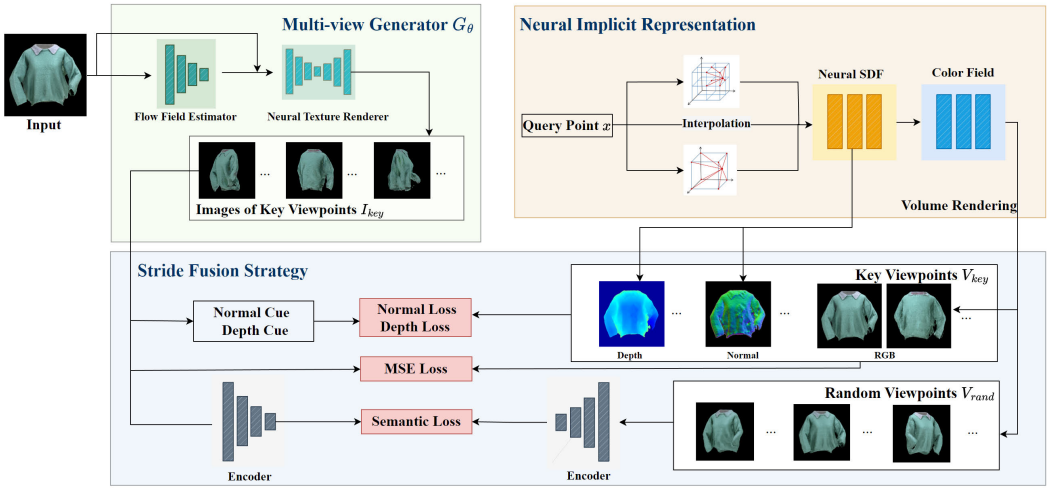

FIGURE 2. Overview of our single-image 3D garment reconstruction framework. The method takes as input a single image to generate pseudo views from 15 key viewpoints using a GAN-based generator, then fusees these sparse-pseudo images from random viewpoints with semantic loss and key viewpoints with pixel-wise loss.

In recent years, template-free garment reconstruction methods have also been developed. For example, [36] proposed a method that can obtain the garment's unsigned distance field from a single image without using templates. However, these methods do not recover textured or vertex-colored meshes, which are crucial for maintaining high-quality appearance details. Xcloth [37] attempted to reconstruct textured meshes. However, similarly, this method relies on SMPL depth prior from pose estimation. PIFu [38] and PIFu-HD [39] can be adapt for garment reconstruction, but they struggle to yield realistic outcomes.

Recently, studies have employed signed distance function (SDF) with volume rendering [16] to represent objects. Approaches such as 3PSDF [40] modified SDF to represent non-watertight geometries. However, it's important to note that these modifications are not compatible with established volumne rendering frameworks. Alternatively, the use of Unsigned Distance Functions (UDF) with volume rendering [41] to represent geometry is possible. Nevertheless, this often leads to excessive geometric holes or substantial errors.

Compared to existing methods, our work enables the accurate 3D reconstruction of clothing from single images, including both geometric and textural details, regardless of whether they are standalone or on a human model.

III. PROPOSED FRAMEWORK

We aim to reconstruct 3D garments from single-image inputs. There are three challenges: how to represent the 3D garment with high-fidelity geometric and texture details, how to predict unseen views of the input garment, and how to synthesize multiple views as a view-consistent 3D model. In this section, we introduce a neat framework based on the SDF-based neural implicit representation, multi-view generator, and the stride fusion strategy to achieve this challenging task.

Our approach for 3D reconstruction employs a neural SDF representation as the backbone, which generates an SDF value s and a color value c as outputs for each garment, based on a 3D query point x and viewing direction v. Furthermore, to capture more detailed information in the reconstruction process, we enhance the neural SDF representation by multi-resolution feature grids $\{\Phi_\theta^l\}_{l=1}^L$ to capture detailed information (Sec. III-A).

A notable aspect of our approach is the use of the stride fusion strategy to address blurry training results due to pixel-level misalignment and ensure high semantic similarity between the views. We adopt an asymmetric pixel-level RGB loss $\mathcal{L}_{rgb}$ optimizes the difference between the predicted images and a set of reference images in different weights according to the confidence of reference images. And a self-supervised semantic loss $\mathcal{L}_{sem}$ proposed to enforce consistency between different views in unseen viewpoints V_{rand} via a pre-trained Vision Transformer network. By combining these two losses, 3D models can achieve high semantic similarity to the target view and generate more realistic outcomes (Sec. III-B).

Additionally, we leverage the monocular depth and normal priors to improve neural implicit surface methods, resulting in better quality of reconstruction (Sec. III-C).

A. VOLUMETRIC SDF WITH DECODER

We considered both Signed Distance Field (SDF) and Unsigned Distance Field (UDF) for representing 3D geometry. While UDF has the advantage of being non-watertight, we found it tends to produce excessive geometric holes or significant errors in training, as demonstrated in Figure 13. Therefore, our framework employs an implicit neural SDF field to represent each cloth, which takes a 3d query point x and a viewing direction v as inputs, then predicts an SDF value $s(x)$ and a view-dependent color value $c(x)$.

The SDF value indicates the distance of the queried point from the surface boundary, and the sign indicates whether the point is inside or outside of a watertight surface. Following VolSDF [16], we convert the SDF value into the 3D density value σ as following:

$$\sigma(x) = \begin{cases} \frac{1}{\beta}\left(1 - \frac{1}{2}exp\left(\frac{s(x)}{\beta}\right)\right) & \text{if } s(x) < 0, \\ \frac{1}{2\beta}exp\left(-\frac{s(x)}{\beta}\right)\right) & \text{if } s(x) \geq 0, \end{cases} \quad (1)$$

where β is a learned parameter that controls the tightness of the density around the surface boundary.

In order to enhance the reconstruction details, our backbone employs multi-resolution feature grids $\left\{\Phi_\theta^l\right\}_{l=1}^L$ of resolution R_l to encode the feature $\phi(x)$ of each queried point x. The resolutions are sampled in geometric space [42] to combine features at different frequencies:

$$R_l := \left\lfloor R_{\min} b^l \right\rfloor, \quad b := \exp\left(\frac{\ln R_{\max} - \ln R_{\min}}{L-1}\right), \quad (2)$$

where $R_{\min}, R_{\max}$ are the coarsest and finest grid resolutions, respectively. Specifically, we extract the interpolated features at each level and concatenate them together:

$$\phi(x) = \left\{\text{interp}\left(\mathbf{x}, \Phi_\theta^l\right)\right\}_l, \quad (3)$$

where *interp* is the trilinear interpolation and $\gamma(x)$ corresponds to frequency encodings introduced in [43]. Then the extracted features are fed to an MLP decoder f_s together with encoded position x to predict the SDF value as follows:

$$s(x) = f_s\left(\gamma(\mathbf{x}), \phi(x)\right). \quad (4)$$

Additionally, we employ an MLP decoder f_c to produce the color of each queried point as well:

$$c(x) = f_c\left(x, v, \phi(x)\right), \quad (5)$$

where v is the viewing direction.

For each pixel, we query points on a ray that originates at the camera position $\mathbf{o}$ and points at the camera direction $\mathbf{r}(t) = \mathbf{o} + t\mathbf{v}$. Equipped with the SDF value of each point that could be converted to 3D density value σ using Equation (1), we render color images with volume rendering and calculate the RGB color as follows:

$$\mathbf{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t))dt,$$
$$\text{where} \quad T(t) = \exp\left(-\int_{t_n}^{t}\sigma(\mathbf{r}(s))ds\right). \quad (6)$$

B. STRIDE FUSION STRATEGY

Our goal is to reconstruct a 3D model of the cloth from a single image I_s. However, it is infeasible to train a neural implicit representation from a single image. To achieve this, we train a generative model G_θ introduced by [44] on a synthetic garment dataset containing the multiview images of more than thousands of garments. Once the generative model is trained, it is able to take a single-view image I_s of a garment and a set of N viewpoints $V_{key} = \{v_1, v_2 \ldots v_i, i \in N\}$ as input and generate pseudo views $I_{key} = \{m_1, m_2 \ldots m_i, i \in N\}$ of corresponding viewpoint, which leverages prior knowledge to compensate for missing information. It is assumed that the pseudo view I_{key} correspond to various angles V_{key} of the garment and display similarities with the input image at the feature level.

Intuitively, we can train an SDF field directly from the pseudo views utilizing volume rendering. However, generating a dense set of pseudo views I_{key} can lead to significant pixel-level misalignment, resulting in blurry training results. To address this issue, we propose a stride fusion strategy that aims to mitigate pixel-level misalignment while ensuring high semantic similarity between the 3D reconstruction and the input image, resulting in a 3D model with a high degree of fidelity. Specifically, there are two main points to this strategy:

a: ASYMMETRIC PIXEL-LEVEL LOSS.

We adopt a selective optimization approach that focuses on optimizing pixel-wise mean squared error (MSE) losses only in a limited set of key viewpoints that exhibit minimal overlap with each other. In practice, we sample 15 views V_{key} around the garment in a horizontal direction. And 15 pseudo images I_{key} are predicted by pretrained generator G_θ. Then we adopt an asymmetric pixel-level RGB loss to minimize the per-pixel difference between the predicted view and the reference images that contain input single I_s and pseudo views I_{key}. The pixel-level RGB loss is defined as:

$$\mathcal{L}_{rgb} = w_i \sum_r^R ||\hat{C}(r) - C(r)||_2^2, i \in N+1 \quad (7)$$

where r denotes a ray from the camera at each frame i and R denotes a set of sampled rays. $\hat{C}(r)$ is the predicted pixel color at ray r, $C(r)$ is the pixel color in reference images, and $N+1$ is the total number of viewpoints in the train images containing I_{key} indexed from 1 to N and I_s indexed $N+1$. w_i is the weight of loss, which reflects the credibility of different views. In our work, we apply $w_i = |v_i - v_s|$, where we assume the input single view v_s to be the origin view and $w_{N+1} = 1$ for the input single view. This approach helps to minimize the impact of pixel-level misalignment and generate sharper training results with reduced blurring.

b: SEMANTIC CONSISTENCY LOSS.

Noticing that the content and style of the two views are similar, though pixel-level misalignment exists between different views, we propose to employ a self-supervised semantic loss to connect the striding key viewpoints to further enhance view consistency across the generated images. This involves incorporating a pre-trained Vision Transformer (ViT) network, which has been proven to be an expressive semantic prior even between images with misalignment [45], [46]. Inspired by [47], we random sample M unseen viewpoints $V_{rand} = \left\{p_1, p_2 \ldots p_j, j \in M\right\}$ around the garment.

Then, different from I_{key} which predict by generator G_θ, we predict the images $I_{rand} = \{r_1, r_2 \dots r_j, j \in M\}$ from the SDF field utilizing volumetric rendering. Furthermore, we adopt a pre-trained ViT model E_{vit} to extract feature embedding from images and enforce semantic consistency by minimizing the difference of feature embedding between different views:

$$\mathcal{L}_{sem} = ||E_{vit}(I_s) - E_{vit}(r_j)||_2^2 + \sum_{i=1}^{N} w_i ||E_{vit}(m_i) - E_{vit}(r_j)||_2^2, \quad (8)$$

where $E_{vit}(I_s)$,$E_{vit}(m_i)$ and $E_{vit}(r_j)$ are the semantic features of the input image, pseudo images, and the rendered image from a random viewpoint, respectively. In addition, $r_j \in I_{rand}$, and w_i is the weight of loss same as Equation (7). This term compares the semantic features of the reference images and the rendered images from any random viewpoint to ensure the consistency of the underlying scene structure. In practice, we adopt CLIP-ViT [48], a self-supervised vision transformer trained on ImageNet [49] dataset.

C. DEPTH AND NORMAL CUES

Combining volume rendering and implicit surfaces has shown impressive 3D reconstruction outcomes [15], [16], [50]. However, this technique faces challenges in intricate scenes, especially in textureless and sparsely observed regions. Related work [50] has already demonstrated the effectiveness of geometric priors in enhancing reconstruction quality. Therefore, following MonoSDF [50], we use readily off-the-shelf monocular geometric priors thereby improving neural implicit surface methods.

a: MONOCULAR DEPTH CUES

One common monocular geometric cue is a monocular depth map, which can be easily obtained via an off-the-shelf monocular depth predictor. In our case, we employ the pre-trained Omnidata model [51] to predict a depth map $\bar{D}$ for each input RGB image. Since the absolute scale is difficult to estimate in general scenes, $\bar{D}$ must be considered as a relative cue.

b: MONOCULAR NORMAL CUES

We also leverage surface normals as an additional geometric cue. Following a similar approach to the depth cues, we utilize the pre-trained Omnidata model to generate a normal map $\bar{N}$ for every RGB image. Unlike depth cues that provide semi-local relative information, normal cues are local and capture geometric detail.

D. LOSS

a: EIKONAL LOSS

Following common practice [16], we also add an Eikonal term [52] on the sampled points $\mathcal{X}$ to regularize SDF values in 3D space:

$$\mathcal{L}_{eik} = \sum_{\mathbf{x} \in \mathcal{X}} \left(\|\nabla f_\theta(\mathbf{x})\|_2 - 1 \right)^2. \quad (9)$$

b: DEPTH CONSISTENCY LOSS

We also enforce consistency between our rendered expected depth $\hat{D}$ and the monocular depth $\bar{D}$:

$$\mathcal{L}_{dep} = \sum_{\mathbf{r} \in \mathcal{R}} \|(p\hat{D}(\mathbf{r}) + q) - \bar{D}(\mathbf{r})\|^2, \quad (10)$$

where p and q are the scale and shift used to align $\hat{D}$ and $\bar{D}$ since $\bar{D}$ is defined only up to scale. Note that these factors have to be estimated individually per batch as the depth maps predicted for different batches can differ in scale and shift. p and q are solved with a least-squares criterion [53].

c: NORMAL CONSISTENCY LOSS

Similarly, we impose consistency on the volume-rendered normal $\hat{N}$ and the predicted monocular normals $\bar{N}$ transformed to the same coordinate system with angular and L1 losses [51]:

$$\mathcal{L}_{\text{norm}} = \sum_{\mathbf{r} \in \mathcal{R}} \|\hat{N}(\mathbf{r}) - \bar{N}(\mathbf{r})\|_1 + \left\| 1 - \hat{N}(\mathbf{r})^\top \bar{N}(\mathbf{r}) \right\|_1. \quad (11)$$

The combined loss function is given by:

$$\mathcal{L} = \mathcal{L}_{rgb} + \mathcal{L}_{sem} + \lambda_1 \mathcal{L}_{eik} + \lambda_2 \mathcal{L}_{dep} + \lambda_3 \mathcal{L}_{norm}, \quad (12)$$

where λs are hyperparameters that control the relative importance of each loss term.

IV. EXPERIMENTS RESULTS

A. IMPLEMENTATION DETAILS

a: TRAINING THE GENERATOR.

During the training of the generator, we use Adam optimizer with a learning rate of 1e-4 and a batch size of 32. We train 100K iterations taking approximately 3 days. Given the limited number of training examples, we also apply data augmentation techniques to improve generalization performance. Specifically, we applied color jittering transformations to change the hue, saturation, and brightness of the input image, expanding the range of possible color distributions that the network could learn from. This approach helped to stabilize training and improve the generator's ability to capture the richness of real-world color distributions.

b: TRAINING THE IMPLICIT SIGNED DISTANCE FIELD.

During the SDF training process, we set $R_{\min=16}$ and $R_{\max=2048}$, and used 16 multi-resolution feature grids $\{\Phi_\theta^l\}_{l=1}^{L}$ to capture detailed information about the 3D garment models. In our experiments, it takes 50k iterations for the SDF field, we use a batch size of 2048 sampled rays. In addition, we set $\lambda_1 = \lambda_2 = 0.1$, and $\lambda_3 = 0.05$.

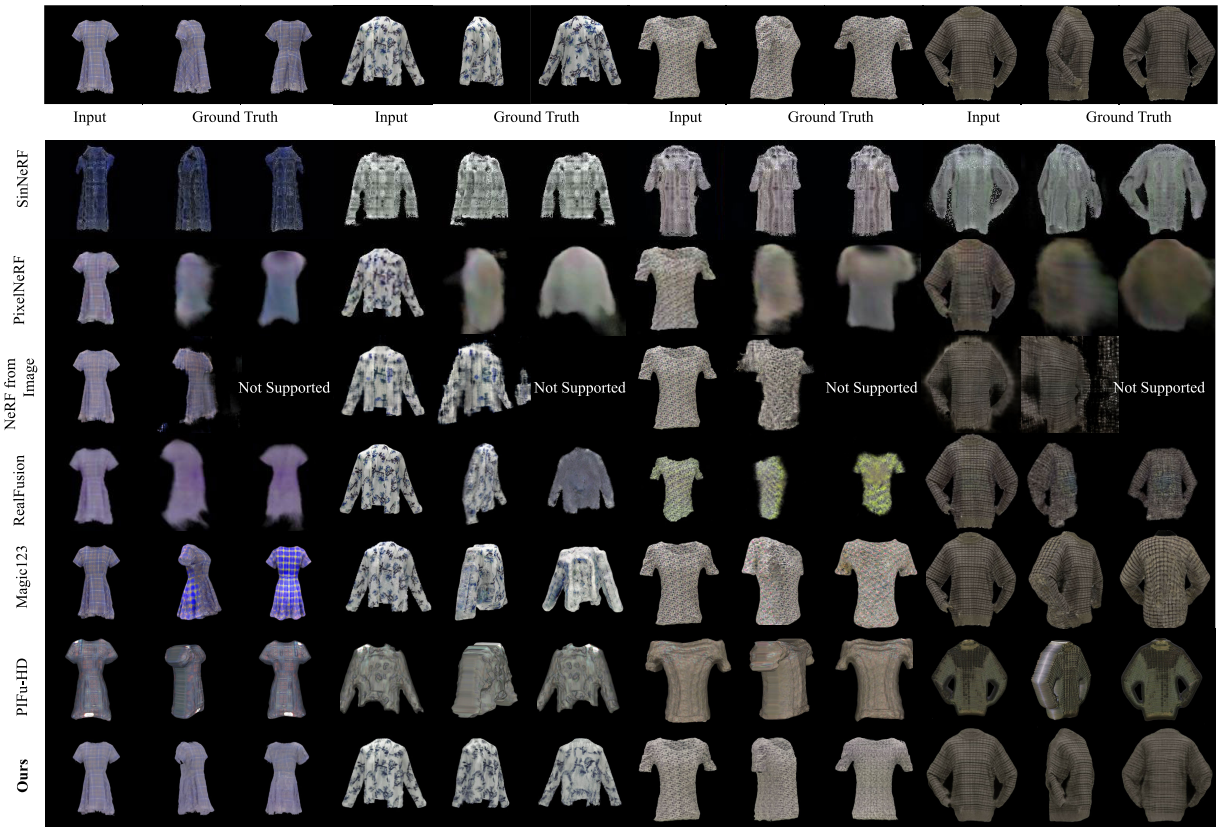

FIGURE 3. Qualitative comparison of our method with the existing SOTA methods. The novel-view images are rendered at 60 and 180 degrees.

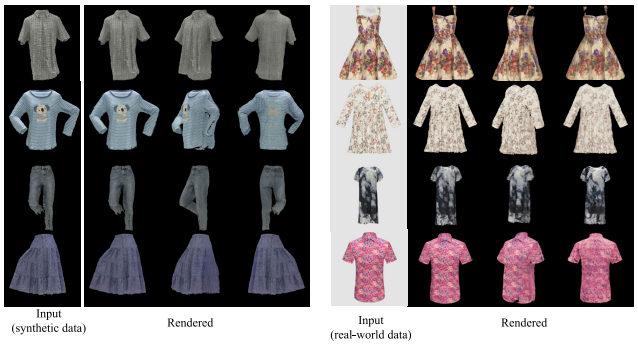

FIGURE 4. Qualitative results of various types of clothing, including both synthetic data (left) and real images from the internet (right).

Our input data is at a resolution of 512*512, and we compute the pixel loss at the same resolution. However, computing the semantic loss requires rendering the entire image, which is computationally intensive. To address this, we render images at a lower resolution of 256*256 to extract semantic features, and only need 25% of the rays as full-resolution training images, which helps improve training efficiency. Furthermore, we found that the semantic loss converges faster than the pixel loss. As a result, in our stride fusion strategy, we only compute the semantic loss every 10 iterations, exploiting its rapid convergence to improve the efficiency of our training process.

B. EVALUATION METRICS

We used standard evaluation metrics to quantitatively evaluate the performance of our proposed framework: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [54], and Learned Perceptual Image Patch Similarity (LPIPS) [55]. PSNR measures the quality of the reconstructed 3D garments in terms of signal-to-noise ratio, while SSIM measures the structural similarity between the reconstructed 3D garments and ground truth. LPIPS is a perceptual distance metric that is learned from a deep neural network and provides a more accurate evaluation of the visual quality of the reconstructed 3D garments. Additionally, we also use Chamfer distance and volume IoU to measure the quality of shape reconstruction.

C. DATASETS

For training and evaluating our proposed framework, we used the Deep Fashion 3D dataset [35]. This dataset consists of 2078 models reconstructed from real garments, covering 10 different categories and 563 garment instances. One of the unique features of this dataset is that each garment is randomly posed to enhance the variety of real clothing deformations, which makes it more challenging and realistic. In practice, when training generator G_θ, we sample 1729 garment models for training and 349 models for testing, where the training and test sets have disjoint instances. We get

TABLE 1. Quantitative comparison of our method with the existing SOTA methods using PSNR, SSIM, and LPIPS on DeepFashion 3D dataset.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF from image [59]	16.26	0.614	0.456
PixelNeRF [4]	20.87	0.714	0.396
SinNeRF [58]	20.13	0.688	0.417
RealFusion [28]	19.62	0.629	0.426
PIFu-HD [39]	18.17	0.638	0.439
Magic123 [60]	21.68	0.716	0.390
Our method	**23.21**	**0.769**	**0.373**

a collection of images by rendering 15 images for each garment model in the Deep Fashion 3D with viewpoints sampled uniformly in the yaw axis. MGN [56] contains 5 garment categories and 154 textured garments models. 134 garments models are randomly selected as the training set and the remaining 20 models form the test set. The SIZER dataset [57] comprises clothing size variations and encompasses 100 different subjects wearing casual clothing items in various sizes, including approximately 2000 scans, 1800 scans are randomly selected as the training set and the remaining 200 models form the test set.

D. QUANTITATIVE EVALUATION

In the quantitative evaluation, we conduct experiments on DeepFashion 3D dataset. We compare our proposed method with 6 state-of-the-art single-image to neural representation algorithms: PixelNeRF [4], SinNeRF [58], NeRF from image [59] RealFusion [28], PIFu-HD [39] and Magic123 [60]. In SinNeRF, we used ground truth depth values for training. Furthermore, introducing back-facing training can cause the results to collapse since it's highly dependent on the depth of the single-view input. Therefore, we only train and evaluate scenes between -60 and 60 degrees. To ensure a fair comparison, we fine-tuned the diffusion prior used in RealFusion on the DeepFashion 3D dataset. And as the pretrained models of PixelNeRF, NeRF from image and PIFu-HD are not trained for the reconstruction of garments, we train and infer results using the same dataset split as the one used in training our generator G_θ. We evaluate the performance of these methods using three 2D image-based metrics (PSNR, SSIM, LPIPS). Our proposed method outperforms other approaches, showcasing our advantages in terms of accuracy and robustness. Please see Table 1 for a detailed illustration of the experimental results.

E. QUALITATIVE EVALUATION

In this section, we provide a qualitative evaluation of our proposed framework by presenting some visual results. The reconstructed 3D garments are shown from different viewpoints and compared with the results of the state-of-the-art methods. The results demonstrate that our framework generates high-quality 3D garment models with realistic details, accurate geometry, and textures.

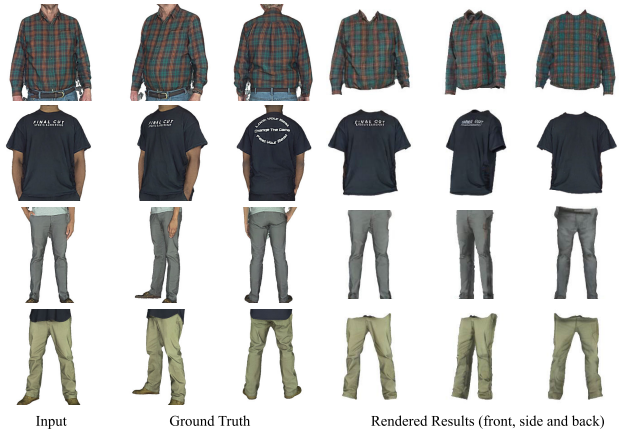

FIGURE 5. Qualitative results of our method on different kinds of garments in MGN dataset.

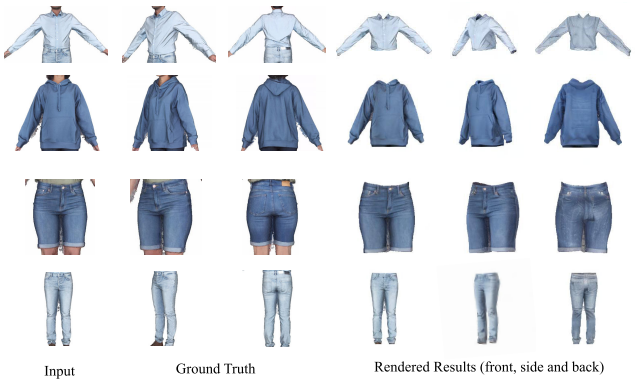

FIGURE 6. Qualitative results of our method on different kinds of garments in SIZER dataset.

Figure 3 shows some examples of reconstructed 3D garments using our proposed framework compared to SOTA methods. The first line shows the input and ground-truth of novel view, the second to seventh columns shows the results of the SOTA method, and the eighth column shows the results of our framework. We can observe that each of these algorithms has its own limitations. Diffusion-based methods like NeRF from image [59], RealFusion [28] and Magic123 [60] suffered from color differences between the input images and the reconstructed 3D models. This issue could be attributed to the image inversion used in these algorithms The output results of PixelNeRF [4] and PIFu-HD [39] exhibit blurriness in views. SinNeRF [58] can produce acceptable results for novel views close to the input view, but it exhibit significant distortion and deformation when the camera viewpoint is far away from the input view, making it unable to provide predictions from backside. Compared to these existing approaches, our proposed framework can reconstruct high-quality garments from both a shape and texture perspective, being able to produce more detailed and accurate 3D models.

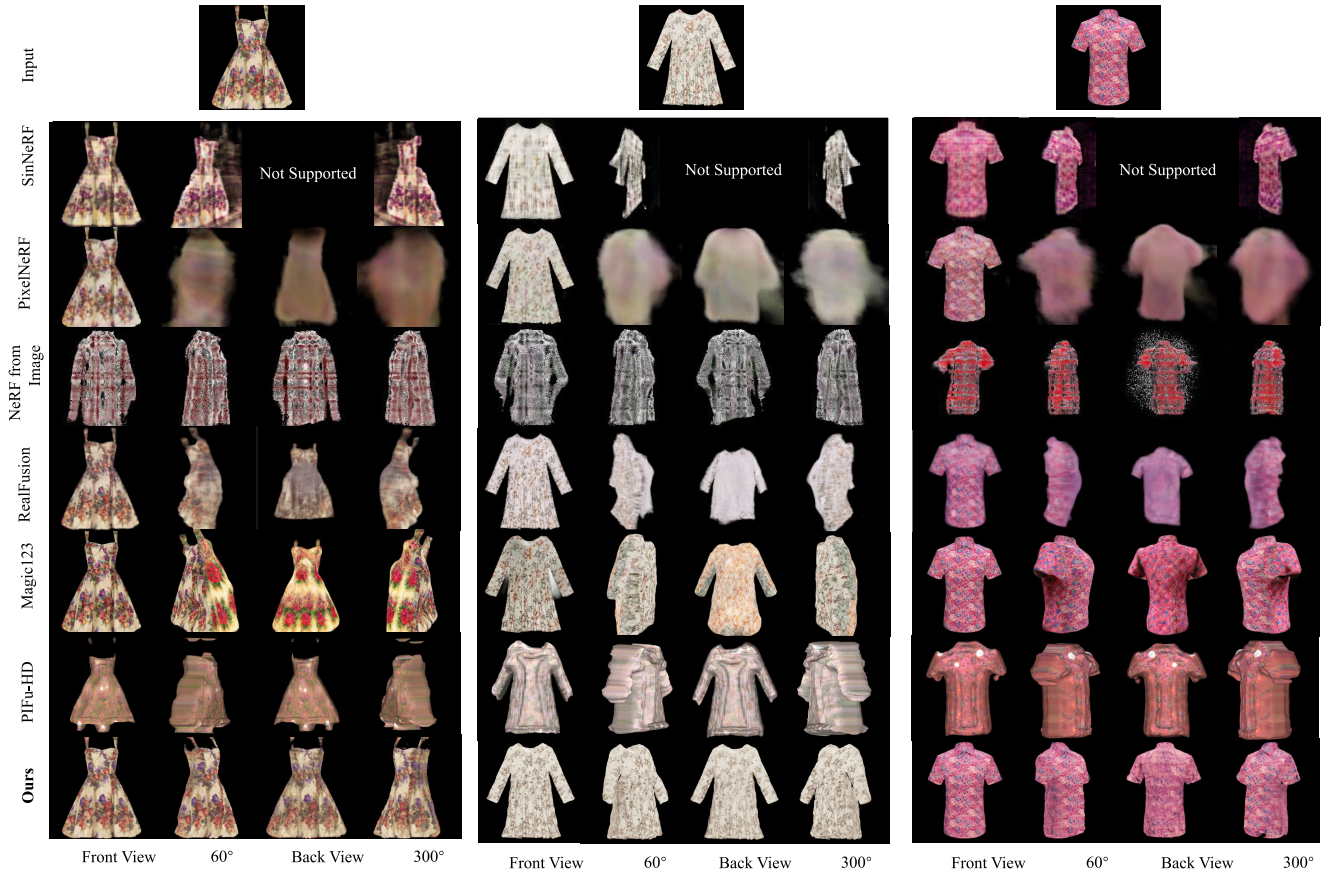

FIGURE 7. Qualitative comparison of our method with the existing SOTA methods on in-the-wild (real-world) garment images.

Figure 4 shows the reconstructed 3D garments from different viewpoints using our proposed framework. The inputs encompass diverse types of clothing, including both synthetic data (left) and real images from the internet (right). We can see that the reconstructed garments are visually consistent from different viewpoints, indicating that our framework is able to generate high-quality 3D garment models that are suitable for various applications.

We alse provide an additional qualitative evaluation of our proposed framework. In Figure 7, we present a comparison between our method and the state-of-the-art methods (PixelNeRF [4], SinNeRF [58], NeRF from image [59] and RealFusion [28]) on real-world data. It is worth noting that the SinNeRF method used in the main text employs ground truth depth as a prior, which is not feasible in real-world scenarios. Therefore, for the in-the-wild images presented in this section, we use Omnidata [51] to estimate the depth of inputs as training prior. The results clearly demonstrate that our method outperforms the SOTA methods and exhibits good generalization ability.

In addition, to further demonstrate the versatility and effectiveness of our framework, we present more reconstruction results in Figure 7. These results showcase our framework's performance on various garment types, including dresses, shirts, and pants. Our method can achieve high-quality reconstruction results across different clothing styles and shapes, highlighting its broad applicability in real-world scenarios.

Apart from the DeepFashion 3D dataset, to assess the generalization capability of our framework, we conducted experiments on the MGN [56] and SIZER [57] datasets, achieving promising results (Figure 5 and Figures 6).

Overall, the qualitative evaluation results demonstrate that our proposed framework is capable of generating high-quality 3D garment models with realistic details and accurate geometry and textures, outperforming the state-of-the-art methods in terms of visual quality.

F. ABLATION STUDY

We perform ablation analyses to assess the impact of different components of our model. Specifically, we examining the two components of stride fusion strategy (SFS): the asymmetric pixel-level loss and the semantic consistency loss (Table 2, Figure 1 and Figure 10). It is observed that the asymmetric pixel-level loss significantly reduces blurriness arising from pixel-level misalignment and enhances the clarity of generated objects and. And the semantic consistency loss mitigates three-dimensional shape deformation caused by inconsistencies among pseudo images. Additionally, we conduct an ablation study to evaluate the contributions of

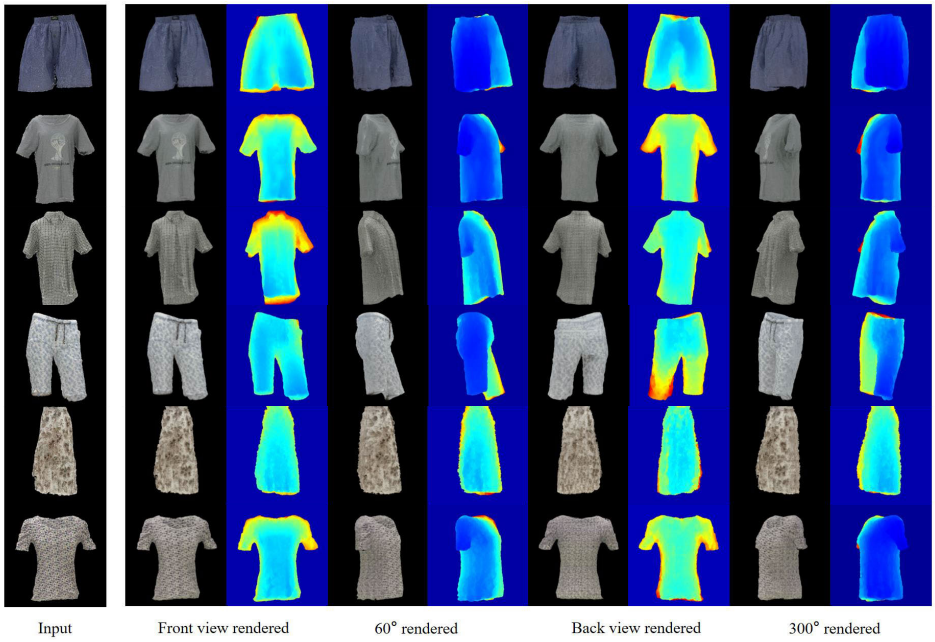

FIGURE 8. Results of our method on different kinds of garments.

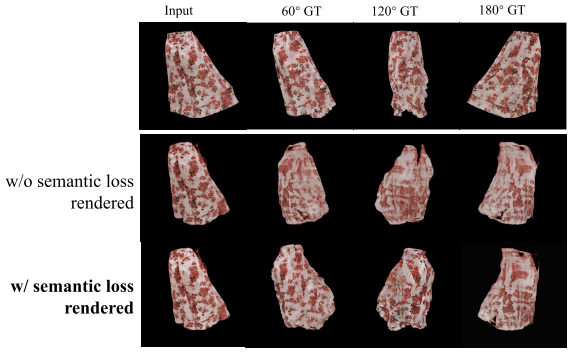

FIGURE 9. Visual comparison of our model with or without the stride fusion strategy.

TABLE 2. Ablation study on stride fusion strategy using PSNR, SSIM, and LPIPS on DeepFashion 3D dataset.

	PSNR↑	SSIM↑	LPIPS↓
w/o Stride Fusion Strategy	20.19	0.732	0.417
w/o Semantic Consistency Loss	20.82	0.745	0.393
w/o Asymmetric Pixel-level Loss	21.26	0.728	0.402
Full Model (Ours)	**23.21**	**0.769**	**0.373**

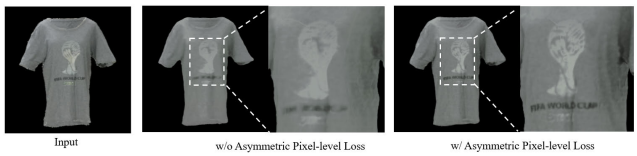

FIGURE 10. Visual comparison of with and without the asymmetric pixel-level loss.

feature grid (Feat. Grid), depth prior and normal prior in our framework (Table 3 and Figure 11).

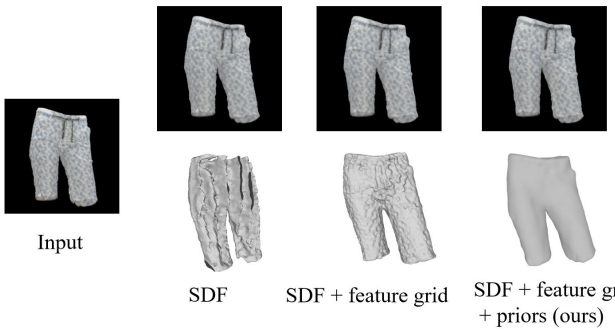

FIGURE 11. Ablation study of different components in our framework.

TABLE 3. Ablation study on various designs of our model using PSNR, SSIM, LPIPS, CD (chamfer distance) and volume IoU on DeepFashion 3D dataset.

	PSNR↑	SSIM↑	LPIPS↓	CD ↓	IoU ↑
w/o Feat. Grid	22.02	0.752	0.379	3.256	0.072
w/o depth cue	22.67	0.722	0.380	0.953	0.142
w/o normal cue	22.13	0.747	0.378	1.273	0.162
Full Model (Ours)	**23.21**	**0.769**	**0.373**	**0.162**	**0.27**

Additionally, we investigated the impact of the number of pseudo images on our outcomes. In the final experiment, we opted for 15 images, as we observed that both excessively high and low number of pseudo images had detrimental effects on results, as depicted in Figure 12 and Table 4. Insufficient images led to gaps within the representation, while an excessive number resulted in minor gaps and texture blurring.

We conducted distinct evaluations employing NEUDF [41] and NeAT [61] for geometry representation. Specifically, NEUDF employed Unsigned Distance Functions (UDF) alongside volume rendering, whereas NeAT utilized a Signed

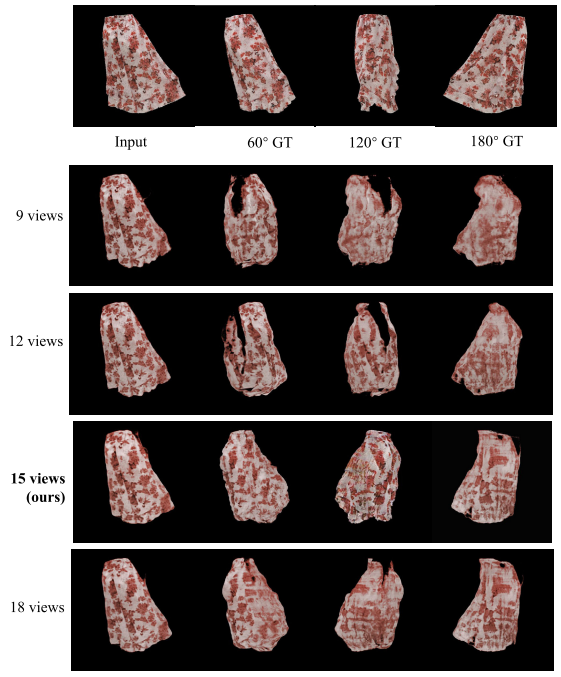

FIGURE 12. Ablation study of different number of pseudo images.

TABLE 4. Ablation study on different number of pseudo images of our model using PSNR, SSIM, and LPIPS on DeepFashion 3D dataset.

	PSNR↑	SSIM↑	LPIPS↓
9 views	20.45	0.737	0.421
12 views	21.26	0.748	0.392
15 views (ours)	**23.21**	**0.769**	**0.373**
18 views	22.75	0.759	0.383

TABLE 5. Ablation study on various designs of our model using CD (chamfer distance) and volume IoU on DeepFashion 3D dataset.

Method	CD ↓	IoU ↑
NEUDF	7.256	0.019
NEAT	6.953	0.022
SDF	3.473	0.062
Full Model (Ours)	**0.162**	**0.27**

Distance Function (SDF) alongside a validity probability function to depict geometry. Both approaches have the capability to generate non-watertight geometries. However, as demonstrated in Figure 13 and Table 5, their ability to represent arbitrary geometries resulted in unintended issues in our context, such as holes, significant deviations from ground truth geometry, and highly fragmented structures. This phenomenon could be attributed to the misalignment present among pseudo images generated by the generator, which has led to an overly fragmented geometry representation. Consequently, we continue using the vanilla SDF-based approach for clothing representation (the last column).

V. LIMITATIONS AND FUTURE WORK

Despite the promising outcomes of our single-view 3D garment reconstruction framework, certain limitations and avenues for future research need consideration. The model's reliance on front-view predictions for side and back views may yield visually plausible but unreal results. To address this issue, future research may increase the diversity of the training dataset. This enhanced dataset would better capture the intricate variations in garment appearance, ensuring a more accurate and realistic reconstruction across all perspectives. Besides, while the stride fusion strategy successfully tackles pixel-level misalignment, it may encounter challenges in regions with occlusions. Enhancing the model's capability to handle occlusions and partial visibility is crucial for future improvements in robustness and versatility.

Furthermore, the computational complexity of the proposed approach poses a potential obstacle for applications or deployment in resource-constrained environments. Therefore, it is imperative to explore optimizations and efficiency-enhancing techniques to make the framework more practical. Addressing these identified limitations will contribute significantly to refining our proposed method, ultimately expanding its applicability in the domain of 3D garment reconstruction.

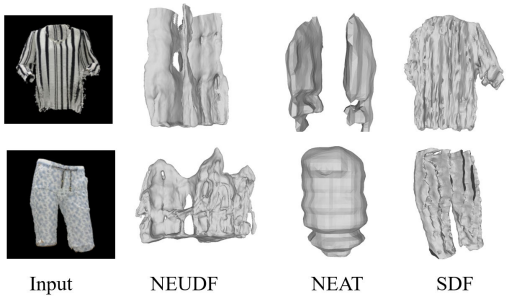

FIGURE 13. Ablation study of different representation of geometry.

VI. CONCLUSION

In this paper, we proposed a neat framework for 3D garment reconstruction. Our framework consists of a pseudo sparse view generator, a volumetric SDF network, and a multi-view rendering module. We further introduce a stride fusion strategy to produce view-consistent geometry and clear texture details. Extensive experiments and ablation studies verified our framework's superiority and the impact of each component.

In summary, our proposed framework provides an effective method for 3D garment modeling and multi-view image synthesis, which can be applied to various domains such as virtual try-on, e-commerce, and fashion design.

ACKNOWLEDGMENT

This work was supported in part by NSFC (No. 62441205), Key R&D Program of Zhejiang Province (No. 2023C01039), and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University. *(Yizheng Chen and Rengan Xie contributed equally to this work.)*

REFERENCES

[1] 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, Jan. 2022.

[2] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon, "KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera," in *Proc. 24th Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2011, pp. 559–568.

[3] K. Zhang, M. Liu, J. Zhang, and Z. Dong, "PA-MVSNet: Sparse-to-dense multi-view stereo with pyramid attention," *IEEE Access*, vol. 9, pp. 27908–27915, 2021.

[4] A. Yu, V. Ye, M. Tancik, and A. Kanazawa, "PixelNeRF: Neural radiance fields from one or few images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4576–4585.

[5] C. Patel, Z. Liao, and G. Pons-Moll, "TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7363–7373.

[6] M. Worchel, R. Diaz, W. Hu, O. Schreer, I. Feldmann, and P. Eisert, "Multi-view mesh reconstruction with neural deferred shading," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6177–6187.

[7] Q. Xu, W. Wang, D. Ceylan, R. Mech, and U. Neumann, "DISN: Deep implicit surface network for high-quality single-view 3D reconstruction," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[8] B. Ahmad, P. A. Floor, I. Farup, and Ø. Hovde, "3D reconstruction of gastrointestinal regions using single-view methods," *IEEE Access*, vol. 11, pp. 61103–61117, 2023.

[9] H. Fan, H. Su, and L. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2463–2471.

[10] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3D reconstruction in function space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4455–4465.

[11] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta, "Learning a predictable and generative vector representation for objects," in *Proc. Eur. Conf. Comput. Vis.* Springer, Sep. 2016, pp. 484–499.

[12] J. Wu, Y. Wang, T. Xue, X. Sun, B. Freeman, and J. Tenenbaum, "MarrNet: 3D shape reconstruction via 2.5 D sketches," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[13] Z. Li and N. Snavely, "MegaDepth: Learning single-view depth prediction from internet photos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2041–2050.

[14] P. Favaro and S. Soatto, "A geometric approach to shape from defocus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 406–417, Mar. 2005.

[15] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," 2021, *arXiv:2106.10689*.

[16] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, "Volume rendering of neural implicit surfaces," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, Dec. 2021, pp. 4805–4815.

[17] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su, "MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14104–14113.

[18] J. Chibane, A. Bansal, V. Lazova, and G. Pons-Moll, "Stereo radiance fields (SRF): Learning view synthesis for sparse views of novel scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7907–7916.

[19] Q. Wang, Z. Wang, K. Genova, P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser, "IBRNet: Learning multi-view image-based rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4688–4697.

[20] X. Yu, J. Tang, Y. Qin, C. Li, X. Han, L. Bao, and S. Cui, "PVSeRF: Joint pixel-, voxel- and surface-aligned radiance field for single-image novel view synthesis," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 1572–1583.

[21] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, and H. Su, "Shapenet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[23] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1532–1540.

[24] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "GANSpace: Discovering interpretable GAN controls," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9841–9850.

[25] J. Zhang, Y. Dong, M. Kuang, B. Liu, B. Ouyang, J. Zhu, H. Wang, and Y. Meng, "The art of defense: Letting networks fool the attacker," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 3267–3276, 2023.

[26] E. R. Chan, M. Monteiro, P. Kellnhofer, J. Wu, and G. Wetzstein, "Pi-GAN: Periodic implicit generative adversarial networks for 3D-aware image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5795–5805.

[27] M. Niemeyer and A. Geiger, "GIRAFFE: Representing scenes as compositional generative neural feature fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11448–11459.

[28] L. Melas-Kyriazi, C. Rupprecht, I. Laina, and A. Vedaldi, "Realfusion: 360° reconstruction of any object from a single image," 2023, *arXiv:2302.10663*.

[29] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D using 2D Diffusion," 2022, *arXiv:2209.14988*.

[30] G. Li, H. Zheng, C. Wang, C. Li, C. Zheng, and D. Tao, "3DDesigner: Towards photorealistic 3D object generation and editing with text-guided diffusion models," 2022, *arXiv:2211.14108*.

[31] M. A. Bautista, P. Guo, S. Abnar, W. Talbott, A. Toshev, Z. Chen, L. Dinh, S. Zhai, H. Goh, and D. Ulbricht, "GAUDI: A neural architect for immersive 3D scene generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 25102–25116.

[32] X. Zeng, A. Vahdat, F. Williams, Z. Gojcic, O. Litany, S. Fidler, and K. Kreis, "LION: Latent point diffusion models for 3D shape generation," 2022, *arXiv:2210.06978*.

[33] L. Zhou, Y. Du, and J. Wu, "3D shape generation and completion through point-voxel diffusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5806–5815.

[34] A. Mir, T. Alldieck, and G. Pons-Moll, "Learning to transfer texture from clothing images to 3D humans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7021–7032.

[35] H. Zhu, "Deep Fashion3D: A dataset and benchmark for 3D garment reconstruction from single images," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Springer, 2020.

[36] F. Zhao, W. Wang, S. Liao, and L. Shao, "Learning anchored unsigned distance functions with gradient direction alignment for single-view garment reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12654–12663.

[37] A. Srivastava, C. Pokhariya, S. S. Jinka, and A. Sharma, "XCloth: Extracting template-free textured 3D clothes from a monocular image," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 2504–2512.

[38] S. Saito, Z. Huang, R. Natsume, S. Morishima, H. Li, and A. Kanazawa, "PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2304–2314.

[39] S. Saito, T. Simon, J. Saragih, and H. Joo, "PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 81–90.

[40] W. Chen, C. Lin, W. Li, and B. Yang, "3PSDF: Three-pole signed distance function for learning surfaces with arbitrary topologies," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18501–18510.

[41] Y.-T. Liu, L. Wang, J. Yang, W. Chen, X. Meng, B. Yang, and L. Gao, "NeUDF: Leaning neural unsigned distance fields with volume rendering," 2023, *arXiv:2304.10080*.

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

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[44] Y. Ren, X. Yu, J. Chen, T. H. Li, and G. Li, "Deep image spatial transformation for person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7687–7696.
[45] N. Tumanyan, O. Bar-Tal, S. Bagon, and T. Dekel, "Splicing ViT features for semantic appearance transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10738–10747.
[46] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel, "Deep ViT features as dense visual descriptors," 2021, *arXiv:2112.05814*.
[47] A. Jain, M. Tancik, and P. Abbeel, "Putting NeRF on a diet: Semantically consistent few-shot view synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5865–5874.
[48] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[50] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger, "MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction," 2022, *arXiv:2206.00665*.
[51] A. Eftekhar, A. Sax, J. Malik, and A. Zamir, "Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3D scans," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10766–10776.
[52] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," 2020, *arXiv:2002.10099*.
[53] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–12.
[54] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[55] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[56] B. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll, "Multi-garment net: Learning to dress 3D people from images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5419–5429.
[57] G. Tiwari, B. L. Bhatnagar, T. Tung, and G. Pons-Moll, "Sizer: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, Aug. 2020.
[58] D. Xu, Y. Jiang, P. Wang, Z. Fan, H. Shi, and Z. Wang, "SinNeRF: Training neural radiance fields on complex scenes from a single image," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel., Springer, 2022, pp. 736–753.
[59] D. Pavllo, D. J. Tan, M.-J. Rakotosaona, and F. Tombari, "Shape, pose, and appearance from a single image via bootstrapped radiance field inversion," 2022, *arXiv:2211.11674*.
[60] G. Qian, J. Mai, A. Hamdi, J. Ren, A. Siarohin, B. Li, H.-Y. Lee, I. Skorokhodov, P. Wonka, and S. Tulyakov, "Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors," 2023, *arXiv:2306.17843*.
[61] X. Meng, W. Chen, and B. Yang, "NeAT: Learning neural implicit surfaces with arbitrary topologies from multi-view images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 248–258.

YIZHENG CHEN received the M.S. degree in operations research from Columbia University. She is currently an Engineer with Zhejiang Laboratory. Her research interests include computer graphics, computer vision, and machine learning.

RENGAN XIE is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Zhejiang University. His research interests include data visualization, computer graphics, and machine learning.

SEN YANG is currently an Engineer with Zhejiang Laboratory. His research interests include machine learning, computer vision, and computer graphics.

LINCHEN DAI received the Ph.D. degree from the State Key Laboratory of CAD and CG, Zhejiang University. He is currently a Researcher with Zhejiang Laboratory. His research interests include computer graphics and computer vision.

HONGCHUN SUN is currently the Senior Product Manager of the Home Solutions Department, China Mobile (Hangzhou) Information Technology Company Ltd. She focuses on the planning, design, and implementation of products in the home domain, with a particular emphasis on leveraging AI algorithms for practical applications.

YUCHI HUO received the Ph.D. degree from the State Key Laboratory of CAD and CG, Zhejiang University. He is currently an Assistant Professor with the State Key Laboratory of CAD and CG, Zhejiang University. His research interests include computer graphics, computer vision, machine learning, and computational optics.

RONG LI received the Ph.D. degree from Zhejiang University, China, in 2015. He is currently a Senior Researcher with Zhejiang Laboratory. His current research interests include computer vision, computer graphics, and deep learning, specifically, for theory and practice of virtual content generation.