electronics

Article

Deform2NeRF: Non-Rigid Deformation and 2D–3D Feature Fusion with Cross-Attention for Dynamic Human Reconstruction

Xiaolong Xie [1,†], Xusheng Guo [2,†], Wei Li [3,*], Jie Liu [3] and Jianfeng Xu [3]

[1] School of Mathematics and Computer Science, Nanchang University, Nanchang 330031, China; 416100210092@email.ncu.edu.cn
[2] School of Information Engineering, Nanchang University, Nanchang 330031, China; 6118120023@email.ncu.edu.cn
[3] School of Software, Nanchang University, Nanchang 330031, China; ndliujie@ncu.edu.cn (J.L.); jianfeng_x@ncu.edu.cn (J.X.)
* Correspondence: weili.cs@ncu.edu.cn
† These authors contributed equally to this work.

Abstract: Reconstructing dynamic human body models from multi-view videos poses a substantial challenge in the field of 3D computer vision. Currently, the Animatable NeRF method addresses this challenge by mapping observed points from the viewing space to a canonical space. However, this mapping introduces positional shifts in predicted points, resulting in artifacts, particularly in intricate areas. In this paper, we propose an innovative approach called Deform2NeRF that incorporates non-rigid deformation correction and image feature fusion modules into the Animatable NeRF framework to enhance the reconstruction of animatable human models. Firstly, we introduce a non-rigid deformation field network to address the issue of point position shift effectively. This network adeptly corrects positional discrepancies caused by non-rigid deformations. Secondly, we introduce a 2D–3D feature fusion learning module with cross-attention and integrate it with the NeRF network to mitigate artifacts in specific detailed regions. Our experimental results demonstrate that our method significantly improves the PSNR index by approximately 5% compared to representative methods in the field. This remarkable advancement underscores the profound importance of our approach in the domains of new view synthesis and digital human reconstruction.

Keywords: computer vision; NeRF; deep learning; digital human; point cloud; image synthesis

Citation: Xie, X.; Guo, X.; Li, W.; Liu, J.; Xu, J. Deform2NeRF: Non-Rigid Deformation and 2D–3D Feature Fusion with Cross-Attention for Dynamic Human Reconstruction. *Electronics* **2023**, *12*, 4382. https://doi.org/10.3390/electronics12214382

Academic Editors: Dah-Jye Lee and Dong Zhang

Received: 19 September 2023
Revised: 12 October 2023
Accepted: 16 October 2023
Published: 24 October 2023

1. Introduction

Recently, obtaining a three-dimensional model of the human body through dynamic video or multi-view images has become a popular research topic in the computer vision field. This technology has significant applications in the meta-universe, virtual reality, animation, and more. However, modeling the dynamic human body solely based on multi-view images [1–4] is challenging because human motion typically involves non-rigid transformations. Additionally, due to occlusion and limited views, accurately estimating human poses from multi-view images is an extremely difficult problem [5].

With the increasing popularity of NeRF research, one of the mainstream approaches is to render new perspectives and postures using NeRF [6–8]. Studies such as NT [9], NHR [10], and Animatable NeRF [1] utilize multi-view images to obtain rendered new-view images and create a rough human body model. Animatable NeRF also introduces the SMPL [11] model as an a priori model for the first time, which plays a role in constraining the movement of points and has achieved good results. However, the SMPL model has limited effectiveness in modeling large non-rigid deformations of the human body. Previous studies [1–3,12] also have certain artifacts and issues with view inconsistency. Moreover, these studies only utilized multi-view images for light sampling and did not effectively utilize 2D–3D feature fusions. Therefore, we aim to leverage 2D–3D feature fusion more effectively to eliminate artifacts and reduce view inconsistency.

We propose a method called Deform2NeRF, which builds upon Animatable NeRF by introducing two new modules: the non-rigid deformation field and the 2D–3D feature fusion field. The non-rigid offset correction field corrects the offset problem in non-rigid transformation by addressing large-scale motion. The 2D–3D feature fusion fusion field extracts features from each image and more effectively utilizes the features of the image, thereby alleviating problems with artifacts and view inconsistency. We tested and evaluated our designed network on ZJU-Mocap [2] and H36M [13] datasets, and the results demonstrate that our design is reasonable and has achieved remarkable results. The results are shown in Figure 1.

In summary, our contributions are as follows:

- We propose a new method called Deform2NeRF to correct the offset in non-rigid deformation and extract more information.
- We propose a 2D–3D feature fusion field to reduce artifacts and inconsistent views.
- Our Deform2NeRF network can achieve good results without additional training for the synthesis of new postures, indicating that our model is robust.

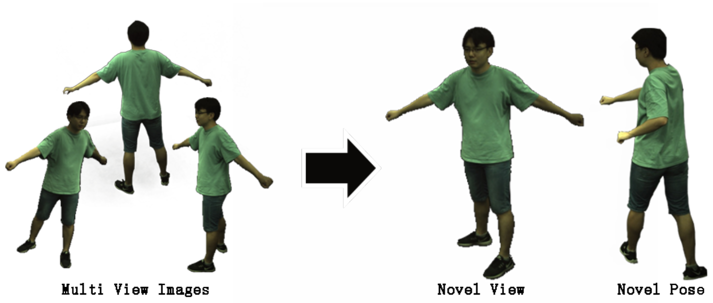

Figure 1. Given a multi-view dynamic human body image, we enable the synthesis of new perspectives and new poses, thereby implicitly reconstructing a three-dimensional representation of the human body.

2. Related Work

2.1. Neural Radiance Field

NeRF [6–8] proposes to render new perspectives by inputting the 5D coordinate points of an object and outputting the density and RGB values of the point cloud, thereby implicitly representing static objects or scenes. Thanks to the introduction of the volume rendering formula [14], NeRF has achieved excellent results and has become a hot research topic in computer vision. NeRF++ [15] extends NeRF to a wide range of boundaryless scenarios. Plenoxels, Instant-NGP, and other works [6,8,16] have optimized the sampling and training strategies of NeRF, greatly reducing the training time. Consistent-NeRF [17] employs depth-derived geometry information and a depth-invariant loss to concentrate on pixels that exhibit 3D correspondence and maintain consistent depth relationships. This significantly improves the performance of the model under sparse view conditions. Head NeRF [3] applies NeRF to the head and achieves real-time high-fidelity reconstruction of the human head. D-NeRF [18] designed a deep learning model to implicitly encode a scene and synthesize novel views at an arbitrary time. Other methods [1,9,10,19–21] extend NeRF from static scenes to moving human bodies, providing new ideas for the generation of digital human images and expanding the application range of NeRF.

2.2. Three-Dimensional Reconstruction of Human Bodies Based on NeRF

The 3D reconstruction of dynamic human bodies based on NeRF has recently gained significant attention in the 3D computer vision community. Several approaches [21–23] have been proposed to enhance the capability of NeRF in capturing the non-rigid deformation of human bodies. For instance, NHR [10] extracts 3D point cloud features through PointNet++ [24] to achieve human body rendering. Neural Human Performer [19] utilizes an attention mechanism to solve the non-rigid transformation of the human body. Neural Body [2] describes the motion state of the human body by introducing potential coding. Animatable NeRF [1] introduces the SMPL model as an a priori model to constrain the human bone structure to some extent. However, since human body movements are complex non-rigid deformations, these methods may generate artifacts in the details. The SLRF [23] method addresses this issue by introducing structured local radiance fields to better describe the details of the folds of clothing on human bodies. SelfNeRF [20] incorporates KNN and hash coding to constrain the movement of human point clouds and improve computational efficiency. Nevertheless, effectively computing the non-rigid deformation of the human body, and eliminating artifacts while preserving details, remains a challenging problem to be solved.

3. Methods

3.1. Neural Blend Weight Fields

Animatable NeRF [1] proposes the use of neural blend weight fields based on a three-dimensional human skeleton and skeleton-driven deformation framework [25] to solve the problem of under-constraint in human deformation.

Specifically, it defines the human skeleton as K parts [11] and generates the K transform matrix $\{G_k\} \in SE(3)$. Using the linear hybrid peeling algorithm, the point $\mathbf{x}^{\mathbf{can}}$ in the canonical space [22] can be transformed into the observation space $\mathbf{x}^{\mathbf{obs}}$. The specific formula is defined as:

$$\mathbf{x}^{\mathbf{obs}} = \left(\sum_{k=1}^{K} w(\mathbf{x}^{\mathbf{can}})_k G_k\right)\mathbf{x}^{\mathbf{can}}, \tag{1}$$

Similarly, we can also convert points in the observation space to points in the standard space. Our method is outlined in Figure 2. $w^o(\mathbf{x})$ is a hybrid weight function defined in the observation space.

$$\mathbf{x}^{\mathbf{can}} = \left(\sum_{k=1}^{K} w^o(\mathbf{x}^{\mathbf{obs}})_k G_k\right)^{-1}\mathbf{x}^{\mathbf{obs}} \tag{2}$$

However, training the mixed weight field with the NeRF network does not achieve good results. Therefore, Animatable NeRF [1] samples for any three-dimensional point by first assigning the initial mixed weights according to the body model, and then uses the residual vector learned by the network to correct the model. The residual network is represented by an MLP (Multi-Layer Perception Machine) network:

$$F_{\Delta\mathbf{w}(\mathbf{x},\psi_i)} \rightarrow \Delta\mathbf{w}_i \tag{3}$$

where ψ is the potential code obtained by the inter-frame embedding layer, and the residual vector $\Delta\mathbf{w}_i \in R_K$. Thus, we can define the neural mixed weight field $\mathbf{w}_i$ of the i-th image, where $\mathbf{w}^{\mathbf{s}}$ represents the initial neural blender weights and S_i denotes the human body model:

$$\mathbf{w}_i(\mathbf{x}) = \text{norm}(F_{\Delta\mathbf{w}}(\mathbf{x}, \boldsymbol{\psi}_i) + \mathbf{w}^{\mathbf{s}}(\mathbf{x}, S_i)), \tag{4}$$

In this way, we obtain the blended weight corresponding to this picture, allowing us to realize the mutual mapping of sampling points in canonical space and observation space [26]. That is, we map the sampling point $\mathbf{x}$ in the observation space to the canonical space to obtain the point $\mathbf{x}'$, and input it into the NeRF networks F_c and F_σ to obtain the color $c_i(\mathbf{x})$ and opacity $\sigma_i(\mathbf{x})$ of the point x for view direction d as follows:

$$\begin{gathered} \sigma_i(\mathbf{x}), z_i(\mathbf{x}) \leftarrow F_\sigma(\gamma_{\mathbf{x}}(\mathbf{x}')) \\ c_i(\mathbf{x}) \leftarrow F_c(z_i(\mathbf{x}), \gamma_d(d), \ell_i) \end{gathered} \tag{5}$$

where $z_i(\mathbf{x})$ denotes the shape feature of the human body, ℓ_i denotes the latent code, and γ denotes the position encoding.

Then, we use the volume rendering formula to render [14]. In this manner, we derive the color information $\hat{C}(\mathbf{r})$ for each pixel in the generated image from this view direction:

$$\begin{aligned} \hat{C}(\mathbf{r}) &= \textstyle\sum_{i=1}^{N} T_i(1-\exp(-\sigma_i\delta_i))\mathbf{c}_i, \\ \text{where } T_i &= \exp\left(-\textstyle\sum_{j=1}^{i-1}\sigma_j\delta_j\right) \end{aligned} \tag{6}$$

where δ_i denotes the distance between successive points and N denotes the number of sampled point clouds in the ray sampling. $\mathbf{c}$ and σ are the RGB value and density of the point cloud, respectively.

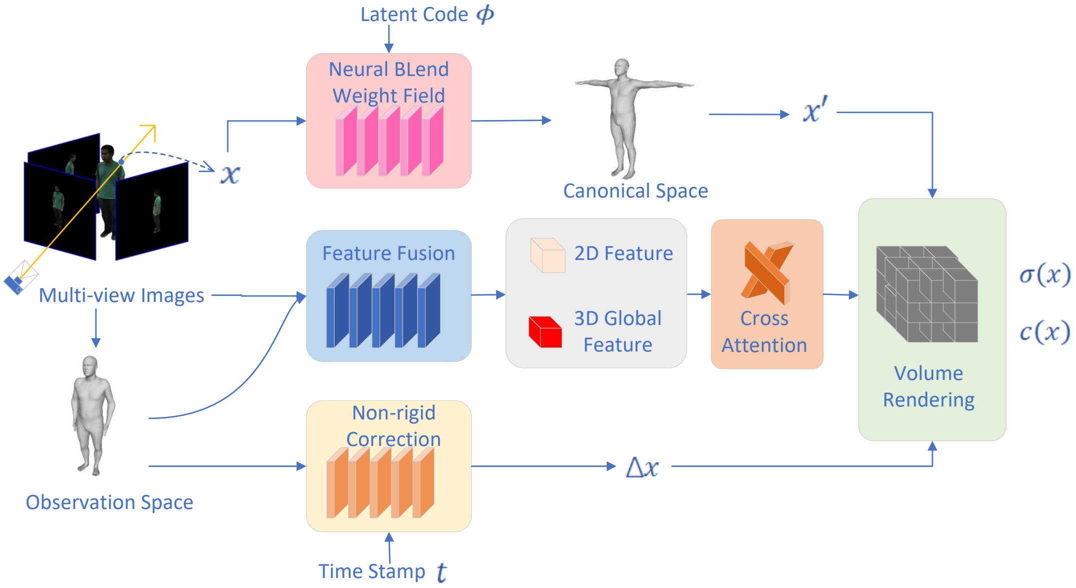

Figure 2. Our network employs dynamic multi-view images of the human body as input to generate an a priori model using the SMPL framework. The mixed weight field predicts the positions of points in the canonical space. To address non-rigid deformations, we utilize a non-rigid deformation field to correct the offset resulting from such deformations. Additionally, our 2D–3D feature fusion network combines and integrates abundant feature information to enhance the effectiveness of human body reconstruction.

3.2. The 2D–3D Feature Fusion Network with Cross-Attention

We observe that previous approaches in human NeRF only utilize image features in light sampling as ground truth, neglecting the potential of using image feature information to guide network training. We consider this as a missed opportunity to leverage valuable features. To address this limitation, we propose to extract feature information from the multi-view image I using a well-established ResNet [27] network:

$$f_{2D} = ResNet(I) \tag{7}$$

By extracting these features, we can effectively discriminate the actions depicted in each picture. It is important to note that each frame of the picture corresponds to a three-dimensional point cloud model, denoted as S. However, these models may exhibit roughness or contain certain interference points, thereby affecting the learning process of the NeRF network to a certain extent. To overcome this challenge, we propose the extraction of global features from the point cloud model through a dedicated point cloud feature network [24,28,29]. By capturing the global characteristics of the point cloud, we aim to represent the human body using these aggregated features. This approach enables the NeRF network to better handle the impact of local irregularities or interference points, thereby enhancing the overall reconstruction performance:

$$f_{3D} = PointNet{+}{+}(S) \tag{8}$$

The fusion of 2D and 3D features is achieved by merging their respective feature representations. Notably, N represents the number of 3D point clouds, g_θ denotes the feature extraction network, and $\otimes$ represents the cross-attention and feature fusion method:

$$f = g_\theta([f_{3D} \otimes repeat(f_{2D}, N)]) \tag{9}$$

Subsequently, the fused feature f is incorporated into the NeRF network, allowing for enhanced constraints on human motion. This fusion feature plays a crucial role in capturing both the spatial information from the 3D point clouds and the visual characteristics from the 2D images. By leveraging this combined feature representation, our approach can effectively constrain the modeling of human motion within the NeRF framework, leading to more accurate and realistic reconstructions.

3.3. Non-Rigid Deformation Field with NeRF

It is well-known that tracking non-rigid deformations in human motion presents a significant challenge. While methods such as the neural mixed weight field proposed in Animatable NeRF [1] have been developed, serious artifacts may still occur for non-rigid deformations with large amplitudes. Building upon the insights gained from SLRF [23], we propose the use of a non-rigid deformation field to correct for any offset induced by non-rigid deformation. Specifically, we introduce an MLP network, denoted as Δx, defined as follows:

$$\Delta \mathbf{x}_i = F_{\Delta \mathbf{x}}(P_i, \ell_i, t_i) \tag{10}$$

where P represents the SMPL [11] model parameters, t corresponds to the timestamp of the image, and ℓ is the corresponding potential coding.

To further improve the performance of NeRF, we aim to incorporate 2D–3D feature fusions into the rendering process. Specifically, we propose to extract 2D–3D feature fusions using a 2D–3D feature fusion extraction network; the 2D–3D feature fusions are then integrated into the NeRF network to address the issues of artifacts and inconsistent views. Currently, NeRF only utilizes multi-view images for light sampling, without exploiting their inherent features. Therefore, we propose to incorporate these 2D–3D feature fusions to further enhance the rendering process.

After extracting the 2D–3D feature fusions, we integrate them into the NeRF network to use the two-dimensional features of the image and fuse them with the three-dimensional point cloud features, as follows:

$$\begin{gathered} \sigma_i(\mathbf{x}), z_i(\mathbf{x}) = F_\sigma(\gamma_{\mathbf{x}}(\mathbf{x}' + \Delta \mathbf{x})) \\ c_i(\mathbf{x}) = F_c(z_i(\mathbf{x}), \gamma_d(d), \ell_i, f) \end{gathered} \tag{11}$$

The above networks are optimized jointly through the NeRF network, and the blend weight field is continuously updated:

$$\mathbf{w}^{\text{new}}(\mathbf{x}, \psi^{\text{new}}) = \text{norm}(F_{\Delta \mathbf{w}}(\mathbf{x}, \psi^{\text{new}}) + \mathbf{w}^s(\mathbf{x}, S^{\text{new}})) \tag{12}$$

By jointly optimizing the above networks through the NeRF network, we are able to continuously track the deformation of human motion and update the parameters of the network. This enables us to effectively address the artifacts caused by non-rigid deformation and the blurring of details.

Following the acquisition of more precise color and opacity data for each point, we employ the well-established NeRF strategy to synthesize a composite view from this perspective using the volume rendering equation (Equation (6)). Subsequently, the synthesized image undergoes a comparative analysis with the ground truth to compute the loss function:

$$L_{\text{rgb}} = \sum_{r \in \mathcal{R}} \|\tilde{\mathbf{C}}_i(\mathbf{r}) - \mathbf{C}_i(\mathbf{r})\|_2, \tag{13}$$

4. Results

4.1. Datasets

To evaluate the effectiveness of our proposed network, we conducted experiments on the widely used H36M [13] and ZJU-Mocap datasets [2]. Following the standard evaluation strategy for human NeRF, to facilitate an equitable comparison, we have embraced a congruent evaluation strategy and utilized identical indicators as the predecessor baseline method. Additionally, the dataset employed holds status as a widely accepted benchmark for assessing the reconstruction of human dynamics. We used the ZJU-Mocap dataset (313, 315, 377, 386), which uses 21 cameras to capture multi-view videos. We selected four views as training data and the remaining views for evaluation. Similarly, for the H36M dataset, which includes complex human actions, we used S1, S5, S6, S7, S8, S9, and S11 as our data sets, where we used three camera views for training and the remaining views for testing. Two common metrics, PSNR and SSIM, were used to assess the performance of our model in new view synthesis and to compare it with previous works.

4.2. Evaluation

To evaluate the effectiveness of our approach in the field of 3D vision, we conducted evaluation and ablation experiments using two widely used datasets, H36M and ZJU-Mocap. Following the evaluation strategies used in previous studies [1] on human NeRF, we used ZJU-Mocap, which includes 21 cameras and records multi-view videos. We selected four viewpoints for training and used the remaining viewpoints for testing. Similarly, for the H36M dataset, we used multi-view videos recorded using four cameras, including a series of complex human actions. We used S1, S5, S6, S7, S8, S9, and S11 as datasets, where we selected images from three viewpoints for training [17] and used the remaining images for testing.

We compared our approach with previous methods such as Neural Body and Animatable NeRF, and conducted quantitative analysis using two common metrics, PSNR and SSIM. The results of our experiments are presented in Table 1 and demonstrate the effectiveness of our approach. As shown in Figure 3, the images rendered by our model have a more complete human body.

In the ZJU-Mocap dataset, we selected four characters (313, 315, 377, 386) to evaluate our approach for new viewpoint and new pose synthesis. Table 1 presents the quantitative analysis of the new viewpoint assessment, while Table 2 displays the results of the new pose assessment. Additionally, Table 3 shows the outcomes of the new view assessment. Similarly, we evaluated our approach on the H36M dataset and presented the results of image comparison, which show that our approach effectively eliminates artifacts.

Table 1. Quantitative comparison of novel view synthesis on the ZJU-Mocap dataset.

	PSNR↑		**SSIM↑**	
	AN	**Ours**	**AN**	**Ours**
313	26.77	**27.30**	0.943	**0.944**
315	20.00	**20.62**	0.867	0.851
377	23.69	**25.71**	0.919	**0.921**
386	26.62	**27.19**	0.886	**0.898**
Average	24.27	**25.21**	0.904	**0.904**

Note: Bold formatting is used to emphasize the data in the table. The up arrow (↑) indicates that a larger indicator is better.

Table 2. Quantitative comparison of novel pose synthesis on the H36M dataset.

PSNR↑				**SSIM↑**			
NT	**NHR**	**AN**	**Ours**	**NT**	**NHR**	**AN**	**Ours**
20.42	20.93	22.41	**23.01**	0.842	0.858	0.876	**0.881**

Note: Bold formatting is used to emphasize the data in the table. The up arrow (↑) indicates that a larger indicator is better.

Table 3. Results of novel view synthesis on the H36M dataset in terms of PSNR and SSIM (higher is better).

	PSNR↑				**SSIM↑**			
	NT	**NHR**	**AN**	**Ours**	**NT**	**NHR**	**AN**	**Ours**
S1	20.98	21.08	22.76	**23.90**	0.860	0.872	0.894	**0.902**
S5	19.87	20.64	23.32	**24.16**	0.855	0.872	0.891	**0.902**
S6	20.18	20.40	22.77	**23.89**	0.816	0.830	0.867	**0.891**
S7	20.47	20.29	21.95	**23.07**	0.856	0.868	0.889	**0.905**
S8	16.77	19.13	22.88	**23.27**	0.837	0.871	0.899	**0.902**
S9	22.96	23.04	24.62	**25.49**	0.873	0.879	0.904	**0.918**
S11	22.96	23.04	24.66	**25.64**	0.859	0.871	0.903	**0.911**
Average	20.42	20.93	23.28	**24.20**	0.851	0.866	0.892	**0.904**

Note: Bold formatting is used to emphasize the data in the table. The up arrow (↑) indicates that a larger indicator is better.

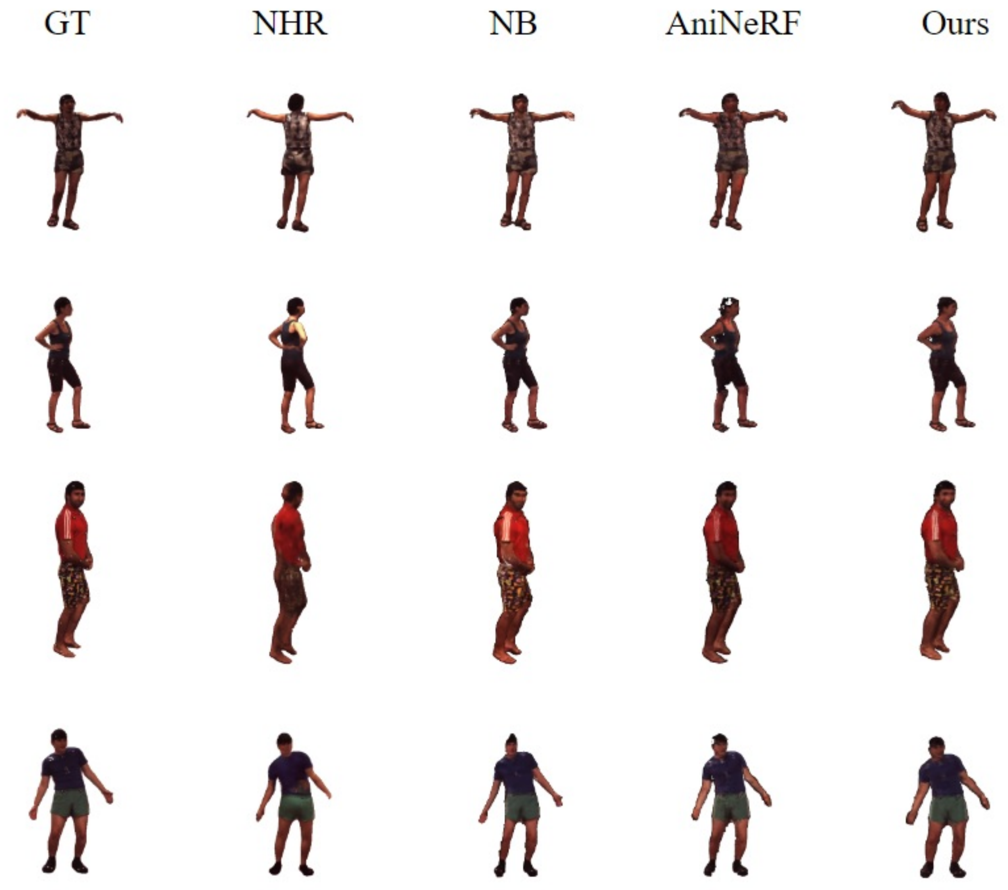

Figure 3. Qualitative results of novel view synthesis on the H36M dataset.

4.3. Ablation Studies

To validate the effectiveness of our network, we conducted ablation experiments on the H36M dataset, specifically on the S11 tester. In these experiments, we analyzed the impact of the offset correction network and the feature extraction network. Additionally, we examined the influence of the time step parameter t on the results of synthesizing new perspectives and new postures of the human body. Furthermore, we explored the effects of different time steps and training rounds on the outcomes. The results of these ablation experiments are summarized in Tables 4–6.

Impact of the non-rigid deformation field network. In Table 4, we present a comparison of the performance of the non-rigid deformation correction network. The results clearly demonstrate that our proposed offset network is capable of accurately capturing and describing the complex non-rigid deformations of the human body. Furthermore, it effectively alleviates artifacts that may arise during the rendering process. This analysis confirms the effectiveness of our offset correction network in improving the quality of synthesized human body representations.

Table 4. Comparison with and without non-rigid offset correction network on subject "S11".

	PSNR↑	SSIM↑
with non-rigid deformation field	**25.64**	**0.911**
without non-rigid deformation field	24.65	0.903

Note: Bold formatting is used to emphasize the data in the table. The up arrow (↑) indicates that a larger indicator is better.

Impact of the 2D–3D feature fusion fusion network. To effectively utilize the features present in the images, we devised a 2D–3D feature fusion extraction network. The experimental results substantiate the benefits of incorporating 2D–3D feature fusion information, as it significantly improves the rendering quality. Table 5 provides a quantitative comparison, further supporting the superior performance achieved through the integration of 2D–3D feature fusions into our approach.

Table 5. Comparison with and without feature fusion network on subject "S11".

	PSNR↑	SSIM↑
with feature fusion	**25.64**	**0.911**
without feature fusion	24.68	0.903

Note: Bold formatting is used to emphasize the data in the table. The up arrow (↑) indicates that a larger indicator is better.

Impact of the time stamp. To enhance the temporal coherence and stability of the variables, particularly the residual position mapping of the sampled points, we incorporated the time step of the input images into our network. This addition enables a more accurate depiction of the image sequence. Table 6 visually illustrates the impact of different time steps on the rendering results, further highlighting the significance of considering temporal information in our approach.

Table 6. Comparison with and without time stamp on subject "S11".

	PSNR↑	SSIM↑
with time stamp	**25.64**	**0.911**
without time stamp	25.34	0.908

Note: Bold formatting is used to emphasize the data in the table. The up arrow (↑) indicates that a larger indicator is better.

5. Discussion

Previous methods in the field of human NeRF have demonstrated certain limitations in effectively tracking and estimating non-rigid deformations caused by human motion. To address this challenge, we introduced the non-rigid correction module into our framework. However, it is important to note that this problem requires further exploration and improvement, potentially through the incorporation of additional feature point information or a more accurate human reference SMPL model. While our proposed method alleviates the artifact problem to a certain extent, future research may benefit from the integration of techniques such as KNN [20] to further constrain human deformation and enhance the overall performance of human NeRF.

Furthermore, we observed that previous human NeRF approaches did not fully exploit the potential of 2D–3D feature fusion information and point cloud feature information in the 3D model. In our work, we designed a feature fusion network to extract and integrate two-dimensional and three-dimensional features, aiming to achieve better results. The under-constrained nature of human NeRF and the sparsity of input images highlight the importance of extracting as much information as possible from 2D images. Maximizing the utilization of feature information from both image and 3D point cloud sources may prove to be a crucial factor in improving the quality of reconstruction results.

In conclusion, while our proposed method represents a step forward in addressing the challenges of human NeRF, further advancements are needed to fully capture and model non-rigid deformations. Additionally, exploring more effective methods to leverage feature information and extracting comprehensive information from limited input images are areas that can contribute to improving the overall effectiveness of human NeRF in future research.

6. Conclusions

In conclusion, our proposed method, Deform2NeRF, represents a significant advancement in the field of computer vision, particularly in the context of dynamic human body modeling using multi-view images. We have addressed several key challenges associated with this technology and made notable contributions:

(1) Offset correction for non-rigid deformation: We introduced a novel module, the non-rigid deformation field, which effectively corrects the offset problem inherent in non-rigid transformations. This module addresses large-scale motion and significantly improves the accuracy of dynamic human body modeling.

(2) The use of 2D–3D feature fusion with a cross-attention network: Our 2D–3D feature fusion field method is a pioneering approach that extracts and utilizes features from each image more effectively. Instead of merely using multi-view images for light sampling, this module fuses 2D and 3D features to mitigate problems related to artifacts and view inconsistency.

(3) Improved model generalization: Our experiments clearly demonstrate the effectiveness of our approach in eliminating artifacts and enhancing the generalization ability of the model. Notably, our method obviates the need for separate training to synthesize new poses, showcasing its robustness and versatility.

(4) Outstanding results on benchmark datasets: We conducted rigorous evaluations on the ZJU-Mocap and H36M datasets, and our method consistently outperformed previous state-of-the-art approaches. This indicates the practicality and real-world applicability of our Deform2NeRF model.

However, it is important to acknowledge the following limitations of our method:

(1) Challenges with non-rigid clothing: Our method may encounter difficulties in accurately reconstructing non-rigid clothing or fabrics, as it primarily focuses on modeling the human body itself.

(2) Sensitivity to lighting conditions: Like many computer vision techniques, our model may be sensitive to variations in lighting conditions, potentially affecting the quality of the reconstructed models.
(3) Statistical body model: Our study employs the SMPL model as the foundational framework for representing the human body—a prevalent approach in the realm of human body reconstruction. It is essential to acknowledge that alternative methodologies, including, but not limited to, STAR [30] and SMPL-X [31], have simpler and more efficient means of human body model representation, potentially yielding superior results.

Despite these limitations, Deform2NeRF represents a promising step forward in dynamic human body modeling. We believe that ongoing research and development can address these challenges and further enhance the capabilities of our approach, making it even more valuable for applications in the meta-universe, virtual reality, animation, and related fields.

Author Contributions: Conceptualization, X.X. and X.G.; methodology, X.G. and X.X.; software, X.G. and X.X.; validation, X.G. and X.X.; formal analysis, X.X., X.G. and W.L.; investigation, X.X. and X.G.; resources, J.L. and J.X.; data curation, X.G.; writing—original draft preparation, X.G. and X.X.; writing—review and editing, W.L. and X.G.; visualization, X.X.; supervision, J.L. and J.X.; project administration, W.L.; funding acquisition, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Foundations of China (Grant No. 62361041) awarded to Wei Li, the National Natural Science Foundation of Jiangxi Province (NSFC) under Grant 20212BAB212012 awarded to Wei Li, the National Natural Science Foundations of China (Grant No. 62266032) awarded to Yanyang Xiao, the National Natural Science Foundation of China under Grant 62102174 awarded to Jianfeng Xu, and the Jiangxi Training Program for Academic and Technical Leaders in Major Disciplines-Leading Talents Project (Grant No. 20225BCI22016) awarded to Jianfeng Xu.

Data Availability Statement: In this study, the ZJU-Mocap dataset was downloaded from ZJU-Mocap dataset (https://chingswy.github.io/Dataset-Demo/, accessed on 25 November 2022) and the H36M dataset was downloaded from H36M dataset (http://vision.imar.ro/human3.6m/description.php, accessed on 25 November 2022).

Acknowledgments: I would like to express my heartfelt gratitude to my supervisor for their invaluable guidance and support throughout this research. Additionally, I extend my appreciation to all those who contributed to this study in any way. The authors would like to acknowledge the anonymous reviewers for their valuable comments.

Conflicts of Interest: I declare that there are no conflicts of interest regarding the research presented in this paper.

References

1. Peng, S.; Dong, J.; Wang, Q.; Zhang, S.; Shuai, Q.; Zhou, X.; Bao, H. Animatable neural radiance fields for modeling dynamic human bodies. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 14314–14323.
2. Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; Zhou, X. 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, Online, 19–25 June 2021; pp. 9054–9063.
3. Hong, Y.; Peng, B.; Xiao, H.; Liu, L.; Zhang, J. Headnerf: A real-time nerf-based parametric head model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20374–20384.
4. Zhao, F.; Yang, W.; Zhang, J.; Lin, P.; Zhang, Y.; Yu, J.; Xu, L. Humannerf: Efficiently generated human radiance field from sparse inputs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7743–7753.
5. Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; Li, H. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2304–2314.

6. Barron, J.T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; Srinivasan, P.P. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 5855–5864.
7. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **2021**, *65*, 99–106. [CrossRef]
8. Fridovich-Keil, S.; Yu, A.; Tancik, M.; Chen, Q.; Recht, B.; Kanazawa, A. Plenoxels: Radiance fields without neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5501–5510.
9. Xiang, F.; Xu, Z.; Hasan, M.; Hold-Geoffroy, Y.; Sunkavalli, K.; Su, H. Neutex: Neural texture mapping for volumetric neural rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 19–25 June 2021; pp. 7119–7128.
10. Wu, M.; Wang, Y.; Hu, Q.; Yu, J. Multi-view neural human rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1682–1691.
11. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A skinned multi-person linear model. *ACM Trans. Graph. (TOG)* **2015**, *34*, 1–16. [CrossRef]
12. Pavlakos, G.; Zhu, L.; Zhou, X.; Daniilidis, K. Learning to estimate 3D human pose and shape from a single color image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 459–468.
13. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [CrossRef] [PubMed]
14. Kajiya, J.T.; Von Herzen, B.P. Ray tracing volume densities. *ACM SIGGRAPH Comput. Graph.* **1984**, *18*, 165–174. [CrossRef]
15. Zhang, K.; Riegler, G.; Snavely, N.; Koltun, V. Nerf++: Analyzing and improving neural radiance fields. *arXiv* **2020**, arXiv:2010.07492.
16. Müller, T.; Evans, A.; Schied, C.; Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph. (ToG)* **2022**, *41*, 1–15. [CrossRef]
17. Hu, S.; Zhou, K.; Li, K.; Yu, L.; Hong, L.; Hu, T.; Li, Z.; Lee, G.H.; Liu, Z. ConsistentNeRF: Enhancing Neural Radiance Fields with 3D Consistency for Sparse View Synthesis. *arXiv* **2023**, arXiv:2305.11031.
18. Pumarola, A.; Corona, E.; Pons-Moll, G.; Moreno-Noguer, F. D-nerf: Neural radiance fields for dynamic scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 19–25 June 2021; pp. 10318–10327.
19. Kwon, Y.; Kim, D.; Ceylan, D.; Fuchs, H. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24741–24752.
20. Peng, B.; Hu, J.; Zhou, J.; Zhang, J. SelfNeRF: Fast Training NeRF for Human from Monocular Self-rotating Video. *arXiv* **2022**, arXiv:2210.01651.
21. Dong, J.; Shuai, Q.; Zhang, Y.; Liu, X.; Zhou, X.; Bao, H. Motion capture from internet videos. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part II 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 210–227.
22. Lewis, J.P.; Cordner, M.; Fong, N. Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, USA, 23–28 July 2000; pp. 165–172.
23. Zheng, Z.; Huang, H.; Yu, T.; Zhang, H.; Guo, Y.; Liu, Y. Structured local radiance fields for human avatar modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 15893–15903.
24. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
25. Park, K.; Sinha, U.; Barron, J.T.; Bouaziz, S.; Goldman, D.B.; Seitz, S.M.; Martin-Brualla, R. Nerfies: Deformable neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 11–17 October 2021; pp. 5865–5874.
26. Gong, K.; Liang, X.; Li, Y.; Chen, Y.; Yang, M.; Lin, L. Instance-level human parsing via part grouping network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 770–785.
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
28. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. Pointcnn: Convolution on x-transformed points. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Volume 31.
29. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.

30. Osman, A.A.; Bolkart, T.; Black, M.J. Star: Sparse trained articulated human body regressor. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part VI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 598–613.
31. Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A.A.; Tzionas, D.; Black, M.J. Expressive body capture: 3d hands, face, and body from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10975–10985.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.