

AirNeRF: 3D Reconstruction of Human with Drone and NeRF for Future Communication Systems

Alexey Kotcov, Maria Dronova, Vladislav Cheremnykh, Sausar Karaf and Dzmitry Tsetserukou

Abstract—In the rapidly evolving landscape of digital content creation, the demand for fast, convenient, and autonomous methods of crafting detailed 3D reconstructions of humans has grown significantly. Addressing this pressing need, our AirNeRF system presents an innovative pathway to the creation of a realistic 3D human avatar. Our approach leverages Neural Radiance Fields (NeRF) with an automated drone-based video capturing method. The acquired data provides a swift and precise way to create high-quality human body reconstructions following several stages of our system. The rigged mesh derived from our system proves to be an excellent foundation for free-view synthesis of dynamic humans, particularly well-suited for the immersive experiences within gaming and virtual reality.

I. INTRODUCTION

Creating detailed 3D reconstructions of humans is a challenging topic in the research community, film, and game industries, where high-fidelity human recreation typically involves pre-recorded templates and multiple camera systems. However, such requirements extend beyond the application scenarios of general uses, such as personalised avatars for telepresence, AR/VR, Metaverse, virtual fitting, and more. Thus, the direct reconstruction of high-fidelity digital avatars from either video footage or image sets is of great practical importance. Classical methods, e.g. photogrammetry [1], have been used for the purpose of 3D object reconstruction, generating a dense point cloud of a scene from different perspectives. Nonetheless, it requires expensive equipment, a long setup time, and has limitations when it comes to measuring non-uniform surfaces and producing detailed reconstructions. Reflections and transparent objects can be challenging due to texture changes. Recently, the quality of 3D reconstruction has improved due to the advances in NeRF-based approaches. This paper, among other aspects, explores the potential applications of this technology in 3D human reconstruction for VR and telepresence purposes.

In this study we present AirNeRF, a novel approach designed to create realistic personalized digital avatars for VR applications. It leverages advancements in data collection through drone-based imaging, novel NeRF-based view synthesis, 3D point cloud-driven mesh creation, segmentation, and automatic rigging.

II. RELATED WORKS

Given that our research leverages NeRF as a backbone, we review existing literature related to NeRF, with a particular

The authors are with the Intelligent Space Robotics Laboratory, CDE, Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30, bld. 1, 121205, Moscow, Russia. {alexey.kotcov, maria.dronova, vladislav.cheremnykh, sausar.karaf, d.tsetserukou}@skoltech.ru

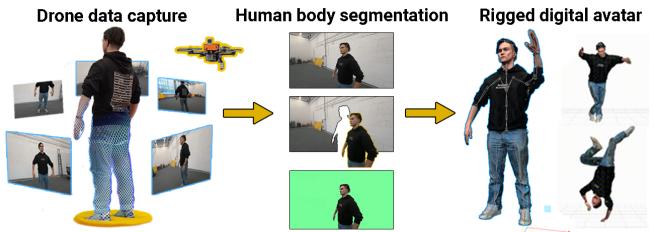


Fig. 1: Layout of AirNeRF system for 3D human reconstruction.

focus on realistic 3D human avatar creation. This focus is imperative for the rendering of human subjects within a scene, especially in the context of developing visual and immersive experiences [2], [3]. We also briefly review works that aim to reconstruct and perform novel view synthesis of provided scenes.

Neural Radiance Fields have gained widespread popularity as a method for scene modeling and rendering from novel perspectives since their initial publication [4]. This popularity stems from the high-quality rendering capabilities of NeRF [4]. The inherent advantage of representing a scene as a radiance field lies in its built-in capability to render the scene seamlessly from various supported viewpoints through volume rendering.

Attempts have been made to customize NeRF for dynamic scenes [5]–[7] and to manipulate and assemble scenes using different NeRF models [8], [9], broadening their potential applications. While these approaches have produced intriguing and promising outcomes, they often necessitate the separate training of editable instances [8] or manual curation of training data [3]. However, this work focuses on a more feasible ‘in-the-wild’ setting. In the context of our specific task, considerable efforts have been directed towards NeRF models rigged modern methods, including neural networks such as RigNet [10] or SMPL [11] with the use of camera prediction algorithm COLMAP [12], [13].

Concurrently, HumanNeRF [14] aims for the free-viewpoint rendering of humans from a single video. While sharing similar goals with our approach, there are a few main distinctions. Firstly, HumanNeRF [14] depends on manual mask annotation for human-background separation, while our method utilizes modern segmentation techniques such as Semantic Guided Human Mating Segmentation [15]. Secondly, HumanNeRF [14] represents motion using a combination of skeletal and non-rigid transformations, potentially resulting in ambiguous or unknown transformations under



Fig. 2: Comparison of 3D reconstruction methods: a) Photogrammetry using scanner. b) Photogrammetry from the set of images. c) Nerfacto model. d) Instant-NGP model. e) Gaussian Splatting render, on a static reduced-size human figure.

novel poses. In contrast, our method addresses this ambiguity by utilizing auto rigging methods, which allows to project joints on a 3D model of the human body. A similar study, Vid2Actor [16], constructs animatable humans from a single video by jointly learning a voxelized canonical volume and skinning weights. Despite similar objectives, our method achieves precise human geometry reconstruction using less than 300 images, compared to Vid2Actor’s requirement for thousands of frames. This demonstrates the data-efficiency of our approach.

Human body reconstruction. This approach has seen widespread adoption of markerless techniques for achieving free-viewpoint videos or reconstructing geometric representations. Some recent endeavors have embraced a lightweight and single-view setup [17], [18]. However, these methods necessitate pre-scanned templates or naked human models, presenting challenges in achieving photo-realistic view synthesis. On the other end of the spectrum, high-end approaches [2], [19]–[21] demonstrate the capability to deliver high-quality surface motion and appearance reconstruction. Nonetheless, they demand dense camera setups and a controlled imaging environment, posing accessibility challenges.

The SelfRecon method was proposed in the following work [22]. It merges implicit and explicit representations to achieve coherent space-time reconstructions of clothed human bodies from monocular self-rotating video input. As a result, it gives a high-fidelity deformable 3D human body reconstruction, though it has some limitations. Its optimisation process is time-consuming, which limits its practical use. Moreover, relying on a normal prediction network may result in inaccurate predictions for real-world applications and inconsistency across different frames.

Monocular RGB-D-based methods [23]–[27] adopt traditional modeling and rendering pipelines for synthesizing novel human views. Despite their contributions, these methods grapple with the inherent self-occlusion constraints, limiting their ability to capture motions in occluded regions. Lightweight multi-view solutions [28]–[30], resembling our approach the most, strike a balance between hardware demands and high-fidelity reconstruction. However, these solutions still rely on 3 to 8 RGB-D streams as input.

TABLE I: Comparison of Methods on Static 3D Model

	Time, min	Vertices	Triangles	File size, Mb
Photogrammetry (scanner)	30	648 861	1 297 650	23.5
Photogrammetry (images)	16	73 319	143 026	11
NeRF (Nerfacto)	6	230 749	455 547	17.7
NeRF (Instant-NGP)	5	207 362	402 727	15.8
Gaussian splatting	5	557 002	0	138

III. QUALITY COMPARISON OF THE 3D RECONSTRUCTION APPROACHES

The first stage of our study involved selecting the suitable method for 3D reconstruction. The primary criteria we chose dictate that a method must be both fast and efficient, enabling seamless integration into the AirNeRF pipeline for human body avatar creation. We performed the first comparison step on a static model of a reduced-size human figure, thereby eliminating the dynamic factors introduced when working with a person. In the second step, a comparison was conducted using a human model.

A. Static model

We performed a comparative analysis that included conventional photogrammetry approaches and novel volume rendering techniques, such as NeRF-based and Gaussian Splatting methodologies. The traditional method, in turn, comprised two approaches: the first included the use of an Artec 4 scanner and Artec Studio software, and the second entailed reconstructing the object from a set of images by the Agisoft Metashape software. The results of the comparison are illustrated in Table I and Fig. 2.

When comparing the latest methods with traditional ones, we have selected the following examples: the Nerfacto model, which comprises elements of diverse NeRF-based methodologies for efficient volume rendering, the Instant-NGP model, distinguished by its enhanced sampling technique compared to vanilla NeRF, and Gaussian Splatting, a novel approach for representing radiance fields by explicitly storing a collection of 3D volumetric Gaussians.

All of the above methods require a dataset consisting of images of the model and the corresponding spatial positions of the camera. In this study, we acquired these datasets by utilizing a video capture of the model and the COLMAP application [12], [13]. To ensure a fair comparison of the methods, the images used for training Gaussian Splatting-based and Gaussian Splatting models, as well as photogrammetry with images, were identical.

Resulting Table I presents key metrics crucial for our application, including the time required to generate a mesh from having a physical model, vertices and triangles of the output mesh, and file size of the PLY format in MB. Upon analyzing these results, we have concluded that neither of the photogrammetry methods fulfills our requirements in terms of small processing time and the number of triangles in the model. Overload in triangles number may result in a lack of engine support, while an insufficient amount may lead to non-representative avatar model. It is noteworthy that Gaussian Splatting demonstrated excellent performance according to render visualization results, as can be seen in Fig. 2(e). However, it should be taken into account that due to its methodological specifics, Gaussian Splatting has its own model representation that differs from conventional triangles or polygons. This representation needs further post-processing to achieve a standard 3D model format. Such a pipeline requires high-power GPU efficiency and sufficient preprocessing time. The aforementioned factors compel us to discard this method in favor of developing an effective and swift system for reconstruction.

B. Human model

During this experiment, photogrammetry with a scanner encountered challenges when reconstructing large objects such as the human body. This particular case necessitates the use of larger and more expensive scanners. Additionally, as we concluded from the last experiment, the obtained results are user-dependant due to the manual nature of the scanning process. Furthermore, achieving a comprehensive 3D reconstruction of the human body through this method demanded extensive post-processing manipulation within the Artec Studio software. The reconstruction of the human body from images using photogrammetry provides poor results as well. Based on this, as well as on the outcomes of the static model experiment, in this section we compare only the two most promising NeRF-based approaches yielding optimal results.

Table II presents a comparison of meshes obtained for a human model through the implementation of Nerfacto and Instant-NGP NeRF models. Given that the reconstruction quality metrics (PSNR, SSIM, and LPIPS) are nearly identical, we should favor the NeRF model that produces results with fewer triangles and vertices, as our main goal is to achieve suitable 3D model format for AR/VR engines support while ensuring high-quality 3D representation of the person. Based on the experimental results, we can conclude that the Nerfacto model performs better on application for our specific task of human body 3D reconstruction.

IV. SYSTEM PIPELINE

In the following section, we provide a comprehensive description of the proposed AirNeRF system configuration for obtaining a human avatar. To create a digital avatar using our method, it is required to have a GPU capable of NeRF training, a drone, and an open space of at least 2 by 2 meters. Through this pipeline, our work aims to address the need for personalized avatars, potentially enhancing interaction within VR environments.

A. Drone configuration

Fig. 3 provides an overview of the proposed system, which utilizes a single unmanned aerial vehicle (UAV) equipped with a DJI Action 2 camera to capture visual data for the construction of a realistic 3D human avatar. To ensure accurate 3D avatar reconstruction and autonomous scanning via the drone, a precise localization technique is necessary. While onboard sensors could potentially be used, this study employs the Vicon V5 tracking system for improved accuracy. The motion capture system consists of 14 infrared cameras detecting retroreflective markers placed on the UAV. By using the information about the markers' position, the tracking system can triangulate the UAV's position with high precision. The Vicon tracking system enables the UAV's location to be determined at a high update rate, allowing for safe and collision-free camera capture while maintaining a one-meter distance from the person. In our system, the refresh rate of the tracking system is set to 100 Hz.

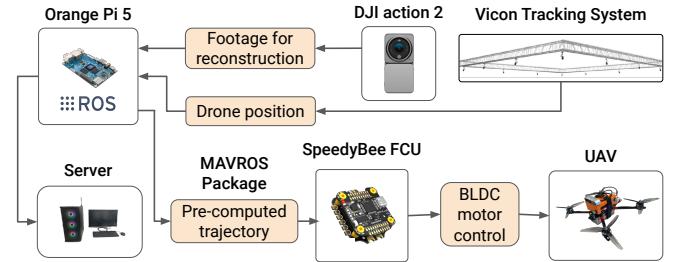


Fig. 3: Hardware architecture of the AirNeRF system.

The UAV setup includes an OrangePi 5 single-board computer, working in tandem with a SpeedyBee F405 flight controller and electronic speed controller stack. The single-board computer utilizes the Robot Operating System (ROS) framework for communications with the flight controller via the MAVROS package. Upon system initialization, a helical trajectory is calculated using the following equations:

$$x(\theta) = r \cos(\theta) \quad (1)$$

$$y(\theta) = r \sin(\theta) \quad (2)$$

$$z = \frac{h_{max} - h_{min}}{\theta_{max}} \theta - h_{min} \quad (3)$$

where r represents the radius of the helical shape, h_{min} and h_{max} are the minimum and maximum flight heights, respectively, and, θ_{max} is the maximum value of the parameter

TABLE II: Comparison of Methods on 3D Human Model

	Time, min	Vertices	Triangles	PSNR	SSIM	LPIPS	File size, Mb
NeRF (Nerfacto)	7	209 288	413 858	19.761	0.921	0.117	15.3
NeRF(Instant-NGP)	6	641 241	1 234 369	19.238	0.915	0.129	46.4

θ . As the value of θ increases the points $(x(\theta), y(\theta), z(\theta))$ maintain a helical trajectory. The drone then flies through all points of the trajectory while maintaining the camera's orientation towards the center of the helix. Once the scanning process is complete, the drone returns to its starting position and lands.

B. NeRF model configuration

As an integral component of our system, the choice of NeRF model configuration plays a crucial role in achieving optimal results, as we discussed in Section III. Among the spectrum of different NeRF models, the Nerfacto emerged as our preferred selection due to its fast processing time and high quality of resulting reconstruction. Nerfacto is under active development and incorporates the advantages of various approaches of model implementation using different parts from recent researches that are proven to be effective. This model's distinctive ability to handle complex spatial patterns aligns seamlessly with our objective of crafting detailed 3D reconstructions of humans.

During the training phase, our model underwent 30,000 iterations, a process driven by a dataset comprising 300 RGB images collected by the drone. This approach ensures the robust learning of intricate features necessary for the accurate representation of human subjects.

Executing on the Ubuntu 22.04 system, powered by the RTX 4090 graphics card, the Nerfacto configuration provided the ability for efficient and precise digital content creation. This strategic selection formed the backbone of our system and enabled to achieve the rapid and autonomous generation of high-quality 3D reconstructions with the training process completed in 6 minutes on average.

V. SYSTEM EVALUATION

We now provide details on the method used in our experiments for reconstructing digital human avatars. Our overall architecture consists of several stages, which we describe in the following subsections.

A. Dataset collection

Our data collection approach aims to be autonomous and robust, as this stage directly impacts the quality of the final reconstruction. We use a drone rather than manual image capture to reduce human error and variability during acquisition. The drone allows us to precisely program reproducible scanning trajectories around the subject. For this work, we implement a helical trajectory maintaining a constant distance of one meter to the scanned human while capturing a full 360-degree view from bottom to top, as depicted in Fig. 4. This helical path provides stability throughout the footage and maximizes the information captured about the human body from all angles.

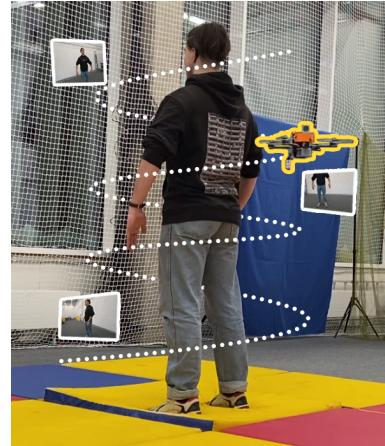


Fig. 4: Data capturing process by the drone following the helical trajectory.

Avoiding suboptimal viewpoints and occlusion is crucial for generating complete and accurate reconstructions. By using an automated drone-based capture method, we obtain consistent high-quality data that fully observe the subject. This improves reconstruction fidelity compared to casual handheld footage. Our experiments demonstrate the benefits of this controlled and robust data collection approach before segmentation, NeRF training, and rigging stages which rely on comprehensive input.

B. Human body segmentation

After obtaining the multi-view dataset, preprocessing and segmentation of the human subject are necessary preparatory steps before NeRF reconstruction. Segmenting out the human body from each frame helps focus the subsequent reconstruction on the relevant subject while ignoring unnecessary computation on background regions, therefore speeding up the overall NeRF training process. The segmentation pipeline used in our system is shown in Fig. 5.

We implement human matting using a state-of-the-art SemanticGuidedHumanMatting module with a pretrained ResNet50 backbone [31]. This network processes each input frame to extract pixel-wise human segments based on semantic guidance. Depending on the inputs, the module outputs either a folder of processed images or a segmented video sequence. By removing background clutter, the segmentation preprocesses the data into a domain-specific set focused on reconstructing the human. The resulting segments can then be passed directly to the NeRF training stage, avoiding excess computation and accelerating avatar reconstruction. In summary, targeted preprocessing transforms the raw captured footage into a concise representation adapted for full-body 3D human reconstruction from multiple views.

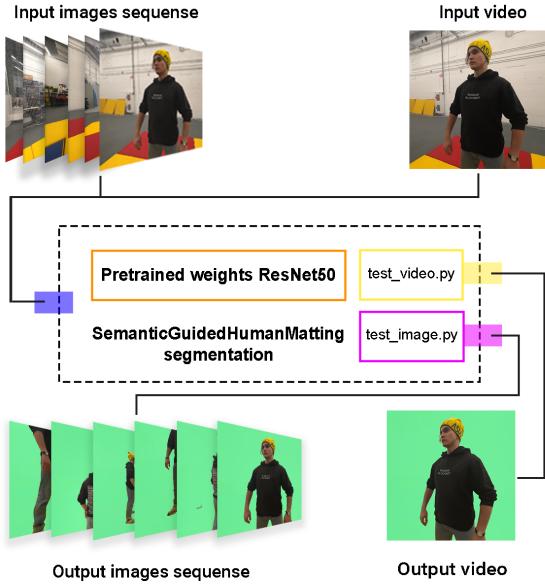


Fig. 5: Segmentation pipeline for extracting human outline from the collected data.

C. NeRF training

At this stage, we have fully prepared data for training the NeRF model, either as an image sequence or as a video file. Prior to starting the training process, it is necessary to compute the positional data for each image within the global coordinate system. To accomplish this task, we employ the COLMAP Structure from Motion (SfM) tool. In our experiment, we input a video in MOV format, necessitating the configuration of COLMAP to use sequential matcher and SIFT features. These parameters are more appropriate for detecting poses from video data. Following the completion of the positional calculations, the training phase is started. The visual representation in Fig. 6, with the depth channel displayed on the right side, illustrates the efficacy of our segmentation preprocessing. Notably, only one object is reconstructed within the scene, according to the intended outcome of the process.

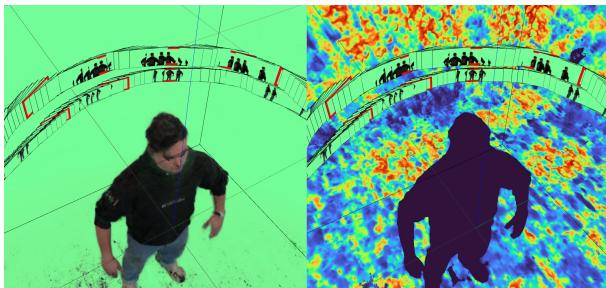


Fig. 6: NeRF training process.

D. Mesh rigging

During the dataset collection phase, our volunteers stand in a specific A-pose or T-pose (Fig. 7(a)) that is required for



a) Human avatar models



b) Rigged digital avatar after AccuRIG processing

Fig. 7: System results: digital avatars for rigging.

the auto-rigging tool AccuRIG. The tool is used to estimate the correct joint positions of the human body with precision. After processing our static human mesh through the auto-rigging module, the output comprises a rigged human body including 22 joints. This amount of joints, particularly the additional two for each hand and each leg, enables the model to make natural rotations and movements. As a result, we have a skeleton that meets the requirements for applying any modern humanoid animation file to the final digital avatar, as can be seen in Fig. 7(b). The Table III presents surface characteristics of the resulting rigged avatars, as well as the file size of the FBX format.

TABLE III: Reconstructed 3D Models

	Vertices	Triangles	Size, Mb	Supported Engines
Human 1	150756	50252	2.81	Unity, Unreal
Human 2	108594	36198	2.71	Engine, Lens Studio,
Human 3	154254	51418	3.00	Meta Spark Studio

As indicated by the metrics, amount of triangles ranges from approximately 36,000 to 51,000 which meets modern 3D engines requirements for triangle count. The resulting file size for all avatars falls within a single range and is relatively lightweight. The following results allow the utilization of the acquired avatars in various popular game engines.

VI. CONCLUSION AND FUTURE WORK

We have developed AirNeRF, a novel system for generating the digital human avatars by leveraging NeRF and drone-based motion capturing. The proposed pipeline features

NeRF rendering and automatic rigging to digitize a real human without 3D modeling.

The results of the research experiment, as demonstrated by the rigged avatars and their corresponding metrics, showcase the system's capability to generate fully prepared human avatars for 3D engines from lightweight drone footage. Specifically, the resulting avatars comprise 22 joints, with a triangle count ranging from approximately 36,000 to 51,000, meeting modern 3D engine requirements. The file sizes of the avatars, in FBX format, fall within a lightweight range, around 3 MB, making them easily integrated into various popular game engines.

VR, AR, and Metaverse fields can be enhanced in terms of user experience by lowering the threshold for the human avatar creation stage preserving realistic appearance and correct proportions. AirNeRF system can potentially considerably improve the immersion of the human into personalized Metaverse and 3D teleconferencing, making such technologies affordable for everyone. Additionally, Industry 4.0 and BIM can potentially benefit from 3D reconstruction of the large scale objects, e.g. engines and buildings. In further research, our development group will focus on the possibility of enhancing the system through the implementation of a Gaussian Splatting mesh reconstruction pipeline.

ACKNOWLEDGEMENTS

Research reported in this publication was financially supported by the Russian Science Foundation grant No. 24-41-02039.

REFERENCES

- [1] E. M. Mikhail, J. S. Bethel, and J. C. McGlone, *Introduction to modern photogrammetry*. John Wiley & Sons, 2001.
- [2] L. Liu, M. Habermann, V. Rudnev, K. Sarkar, J. Gu, and C. Theobalt, “Neural actor: Neural free-view synthesis of human actors with pose control,” *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021.
- [3] S. Peng, C. Geng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, X. Zhou, and H. Bao, “Implicit neural representations with structured latent codes for human body modeling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [4] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. Mcallister, J. Kerr, and A. Kanazawa, “Nerfstudio: A modular framework for neural radiance field development,” in *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings, SIGGRAPH '23*, ACM, July 2023.
- [5] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, “Nerfies: Deformable neural radiance fields,” *ICCV*, 2021.
- [6] K. Park, U. Sinha, P. Hedman, J. T. Barron, S. Bouaziz, D. B. Goldman, R. Martin-Brualla, and S. M. Seitz, “Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields,” *ACM Trans. Graph.*, vol. 40, dec 2021.
- [7] A. Nenashev, M. Kurenkov, A. Potapov, I. Zhura, M. Katerishich, and D. Tsetserukou, “Loconerf: A nerf-based approach for local structure from motion for precise localization,” in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 641–646, 2023.
- [8] M. Guo, A. Fathi, J. Wu, and T. Funkhouser, “Object-centric neural scene rendering,” 2020.
- [9] M. Dronova, V. Cheremnykh, A. Kotcov, A. Fedoseev, and D. Tsetserukou, “Flynerf: Nerf-based aerial mapping for high-quality 3d scene reconstruction,” 2024.
- [10] Z. Xu, Y. Zhou, E. Kalogerakis, C. Landreth, and K. Singh, “Rignet: Neural rigging for articulated characters,” *ACM Trans. on Graphics*, vol. 39, 2020.
- [11] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “Smpl: a skinned multi-person linear model,” *ACM Trans. Graph.*, vol. 34, oct 2015.
- [12] J. L. Schönberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [13] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, “Pixel-wise view selection for unstructured multi-view stereo,” in *European Conference on Computer Vision (ECCV)*, 2016.
- [14] F. Zhao, W. Yang, J. Zhang, P. Lin, Y. Zhang, J. Yu, and L. Xu, “Humannerf: Efficiently generated human radiance field from sparse inputs,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7743–7753, June 2022.
- [15] X. Chen, Y. Zhu, Y. Li, B. Fu, L. Sun, Y. Shan, and S. Liu, “Robust human matting via semantic guidance,” in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2022.
- [16] C.-Y. Weng, B. Curless, and I. Kemelmacher-Shlizerman, “Vid2actor: Free-viewpoint animatable person synthesis from video in the wild,” 2020.
- [17] M. Habermann, W. Xu, M. Zollhöfer, G. Pons-Moll, and C. Theobalt, “Livecap: Real-time human performance capture from monocular video,” *ACM Trans. Graph.*, vol. 38, mar 2019.
- [18] L. Xu, W. Xu, V. Golyanik, M. Habermann, L. Fang, and C. Theobalt, “Eventcap: Monocular 3d capture of high-speed human motions using an event camera,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 4967–4977, IEEE Computer Society, jun 2020.
- [19] M. Habermann, L. Liu, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt, “Real-time deep dynamic characters,” *ACM Transactions on Graphics*, vol. 40, aug 2021.
- [20] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, and C. Theobalt, “Markerless motion capture of multiple characters using multiview image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2720–2735, 2013.
- [21] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt, “Fast articulated motion tracking using a sums of gaussians body model.” in *2011 International Conference on Computer Vision*, pp. 951–958, 2011.
- [22] B. Jiang, Y. Hong, H. Bao, and J. Zhang, “Selfrecon: Self reconstruction your digital avatar from monocular video,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [23] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu, “Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7287–7296, 2018.
- [24] K. Guo, F. Xu, T. Yu, X. Liu, Q. Dai, and Y. Liu, “Real-time geometry, albedo, and motion reconstruction using a single rgbd camera,” *ACM Trans. Graph.*, vol. 36, jul 2017.
- [25] Y. Jiang, S. Jiang, G. Sun, Z. Su, K. Guo, M. Wu, J. Yu, and L. Xu, “Neuralhofusion: Neural volumetric rendering under human-object interactions,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6155–6165, 2022.
- [26] L. Xu, W. Cheng, K. Guo, L. Han, Y. Liu, and L. Fang, “Flyfusion: Realtime dynamic scene reconstruction using a flying depth camera,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 1, pp. 68–82, 2021.
- [27] M. Matrosov, O. Volkova, and D. Tsetserukou, “Lightair: a novel system for tangible communication with quadcopters using foot gestures and projected image,” pp. 1–2, 07 2016.
- [28] M. Dou, P. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rheemann, V. Tankovich, and S. Izadi, “Motion2fusion: real-time volumetric performance capture,” *ACM Trans. Graph.*, vol. 36, nov 2017.
- [29] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escalano, C. Rhemann, D. Kim, J. Taylor, P. Kohli, V. Tankovich, and S. Izadi, “Fusion4d: real-time performance capture of challenging scenes,” *ACM Trans. Graph.*, vol. 35, jul 2016.
- [30] L. Xu, Z. Su, L. Han, T. Yu, Y. Liu, and L. Fang, “Unstructuredfusion: Realtime 4d geometry and texture reconstruction using commercial rgbd cameras,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2508–2522, 2020.
- [31] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.