EG UK Computer Graphics & Visual Computing (2024)
A. Slingsby and D. Hunter (Editors)

View-Consistent Virtual Try-on of Glasses using a Hybrid NeRF-Mesh Rendering Approach

A. Rak[1], T. Wirth[1], T. Lindemeier [3], V. Knauthe[1] and A. Kuijper[1,2],

[1]TU Darmstadt, Germany [2] Fraunhofer IGD, Germany [3] Carl Zeiss AG, Germany

Abstract

In recent times, an increasing fraction of global purchases is conducted via the world wide web. For individual accessories, such as glasses, a purchase commonly involves trying on multiple products to fit individual aesthetic preferences. The experience of the try-on process differs greatly between online and offline shopping. While there are real-time methods that facilitate virtual try-on of glasses, they usually project them onto a 2D image. This leads to inconsistent positioning of the glasses model between different views, negatively influencing the shopping experience. We propose a strategy, that enables the virtual try-on of glasses using a Neural Radiance Field as head avatar and a meshed glasses model, leading to consistent positioning of the spectacle frame through multiple views while maintaining real world like visual quality. We contribute an approach for placing and aligning the glasses in relation to the human head in the given NeRF context. Furthermore, we propose a framework for real-time hybrid rendering of meshes and Neural Radiance Fields in the same scene. The proposed method requires training times around one minute and produces a freely explorable 3D model that achieves interactive framerates on end-consumer hardware.

1. Introduction

Neural Radiance Fields (NeRFs) [MST*20] are an ongoing trend in the computer graphics community, due to their unprecedented photorealism in image-based novel-view-synthesis. NeRFs model scenes as continuous volumetric functions in the form of multi-layer perceptrons (MLPs) that describe the radiance of light at each point in space. By training on input images captured from multiple viewpoints, NeRFs learn to capture intricate lighting effects and generate renderings with detailed appearance information. This facilitates highly realistic novel-view-synthesis of scenes containing fine structures such as fur and hair.

Consequently, derivative works have dealt with the representation of human beings with NeRFs [PSB*21; PSH*21; ZBT23], facing additional challenges such as dynamic movements during capturing. These works have shown that human NeRF representations yield rendered novel views that are hardly discernible from real images, which poses the question whether virtual try-on applications can benefit from this technology. Although some works have explored the composition of virtual volumetric humans and clothing [SOTC22], as well as glasses [LSS*23], achieving both realistic appearance and interactions, the training times of up to eight days on a single GPU prevent their usage in an online virtual try-on setting. This work attempts to bridge this gap.

Manufacturers typically store existing models of glasses frames in mesh formats. To combine these pre-existing models with the expressiveness of human face avatars, we propose the usage of a specialized renderer, combining NeRFs and meshes in one scene. This design decision maintains visual quality of the human face avatar in comparison to a fully meshed-based renderer. It also mitigates conversion inaccuracies in comparison to a setup, where the spectacle frames are converted to a NeRF format and rendered using NeRF composition techniques. Furthermore, the runtime performance of our hybrid renderer supersedes the performance of a renderer relying on NeRF models exclusively, due to early ray termination on intersections with mesh surfaces.

While recent work has dealt with the insertion of meshes into NeRF scenes [YWTZ24; QGX*23], to the best of our knowledge, no published work exists that leverages hybrid rendering for virtual try-on applications, which necessitates both quick training- and rendering times. Therefore, we propose our own framework, available on GitHub [RWL*24], that coherently displays both NeRFs and surfaces in one scene for the purpose of virtual try-on of glasses. Our renderer achieves real-time performance on consumer-grade hardware. To achieve this, the renderer is based on the highly runtime-efficient, albeit complex Instant-NGP NeRF variant [MESK22], utilizing custom CUDA kernels for hybrid rendering. To allow for rapid prototyping and seamless integration with Python libraries for landmark detection in human faces [LTN*19], the renderer exposes Python bindings for loading and transforming meshes and NeRFs.

In summary, our main contributions are:

- a framework for real-time hybrid NeRF-mesh rendering,
- an automatic approach for placing and aligning virtual objects on human heads in a NeRF context,
- an integration of both components into a cohesive pipeline enabling photorealistic and view-consistent virtual try-on for glasses.

DOI: 10.2312/cgvc.20241223

2. Related Work

Neural Radiance Fields (NeRFs) [MST*20] estimate a continuous radiance representation of a 3D scene using a multi-layer perceptron. Using standard volume rendering techniques, this representation can be leveraged to render novel views with high levels of realism from multi-view image inputs.

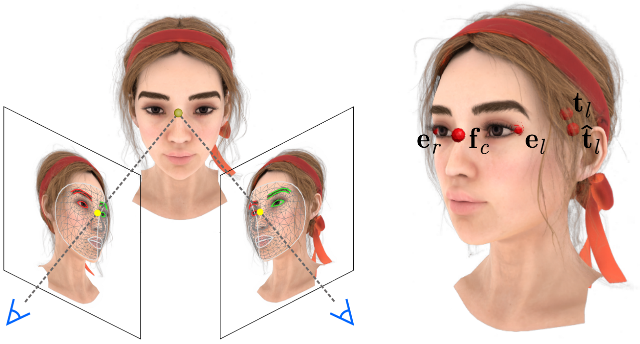
Human NeRF Avatars. NeRFs are generally restricted to static scenes. Therefore, the rendering of a human face is often difficult due to micro movements during the capturing process. Park et al. mitigate these effects by introducing a deformation field [PSB*21] and lifting NeRFs into a higher dimensional representation space [PSH*21]. Subsequent publication [WCS*22; HPX*22; ZAB*22; GTZN21; ZBT23] allow the generation of morphable 3D avatars, i.e., face models that can depict movement and different emotions, based on NeRF architectures. These architectures usually require long training times. For our proposed framework, we opt for the accelerated NeRF variant Instant-NGP [MESK22], enabling interactive framerates and subminute training times on end consumer hardware. However, this design choice has the drawback, that small alterations in the captured human's pose reduce the result's quality.

Inserting Objects into NeRFs. A multitude of publications propose strategies to compose multiple NeRF models in one scene [TCWZ22; LGO*23; JKK*23]. Other recent work deals with the insertion of meshes into NeRF scenes [YWTZ24; QGX*23]. They disentangle the baked lighting setup to enable dynamic relighting of the NeRF components and allow realistic occlusion and shadow generation. This complicates the training process, increasing training times considerably, which is undesirable for practical application. Thus, our proposed framework ignores dynamic lighting effects. The results, however show, that even though the lighting is baked into the model, the rendered novel views have a high level of realism, provided they are captured under conditions without strong directional light influence.

Virtual Try-On of Glasses. Virtual Try-On for glasses has been extensively examined in related publications [YKF*11; ADD16; Zha18; FJS18; MA21; KSSU19; TZT*14; ZGL*17]. These approaches use techniques of augmented reality, i.e., they use an image of the human face and overlay a projection of the glass frame, to achieve a high level of photorealism. Therefore, these strategies are restricted to view directions, from which an input image is available and can not guarantee consistency between views. Li et al. [LSS*23] have recently introduced a morphable and relightable eyeglass model composed together with volumetric head avatars. While their approach produces photorealistic renderings, training times of eight hours and a capturing setup involving a light-stage prevent practical usage for try-on applications. In contrast, our approach allows for a casual capturing environment and produces renderings in less than two minutes at the expense of visual fidelity.

3. Method

Our work deals with the virtual try-on of meshed glasses for NeRF-based avatars of human subjects. We outline our method for aligning the NeRF and the glasses into the same coordinate system, ensuring a proper fit on the virtual face in section 3.1. To facilitate a cohesive visualization, both representations have to be rendered in the same scene. This necessitates a specialised rendering approach, which is detailed in section 3.2.



(a) Reconstructing 3D landmarks from multiple views.

(b) Nose, eye and temple landmarks.

Figure 1: Landmarks on a synthetic head. Triangulation is used to reconstruct 3D landmarks (a), leveraging multiple rendered views of a trained human NeRF model and corresponding 2D landmarks identified with MEDIAPIPE FACE MESH. The 3D landmarks obtained through this process are shown in (b).

3.1. Glasses Placement

Fitting the glasses involves aligning predefined reference points on the glasses' frame with dynamically identified landmarks on the face. Three reference and three landmark points are required for this operation to optimize translation and rotation. The predefined reference points on the spectacle frame's temple tips $\mathbf{g}_r, \mathbf{g}_l$ and nose bridge $\mathbf{g}_c$ are manually annotated for each glasses model.
The corresponding 3D face landmarks are dynamically identified on a trained human NeRF avatar. The NeRF model enables rendering novel views of the face with known camera poses and a static appearance, i.e., the face is not moving between different views. This facilitates the triangulation of three-dimensional landmarks. We leverage MEDIAPIPE FACE MESH [LTN*19] to detect two-dimensional face landmarks on multiple rendered views. With the corresponding known camera parameters, we project rays through the 2D landmarks. The 3D point with the smallest Euclidean distance to all rays associated with a landmark is identified as the reconstructed 3D landmark. This process is illustrated in Fig. 1a.
As FACE MESH does not include ear landmarks, we approximate their positions leveraging landmarks on the nose $\mathbf{f}_c$, on the outer corners of the eyes $\mathbf{e}_l$, $\mathbf{e}_r$, and on the temples $\mathbf{t}_l$, $\mathbf{t}_r$ (see Fig. 1b). In accordance to Yuan et al. [YKF*11], we assume the ears – the region where the glasses frame rests on the ears in particular – to be at the same height as the eyes in a typical facial structure. Thus, the ear landmark $\mathbf{f}_l$ is approximated as follows ($\mathbf{f}_r$ analogous):

$$\mathbf{d}_{\text{eye}} = \frac{\mathbf{e}_l - \mathbf{e}_r}{||\mathbf{e}_l - \mathbf{e}_r||} \qquad \mathbf{d}_{\text{up}} = \mathbf{d}_{\text{eye}} \times \mathbf{d}_{\text{fwd}} \tag{1}$$

$$\hat{\mathbf{t}}_l = \mathbf{t}_l + \left(\frac{(\mathbf{e}_l - \mathbf{t}_l) \cdot \mathbf{d}_{\text{up}}}{\mathbf{d}_{\text{up}} \cdot \mathbf{d}_{\text{up}}} \right) \mathbf{d}_{\text{up}} = \mathbf{t}_l + ((\mathbf{e}_l - \mathbf{t}_l) \cdot \mathbf{d}_{\text{up}})\, \mathbf{d}_{\text{up}} \tag{2}$$

$$\mathbf{f}_l = \hat{\mathbf{t}}_l - \alpha \, ||\mathbf{e}_l - \mathbf{e}_r|| \, \mathbf{d}_{\text{fwd}} \tag{3}$$

where $\mathbf{d}_{\text{eye}}$, $\mathbf{d}_{\text{up}}$ and $\mathbf{d}_{\text{fwd}}$ define the head's orientation, $\hat{\mathbf{t}}_l$ is the left temple landmark projected onto the eye plane along the $\mathbf{d}_{\text{up}}$ vector and α determines the ratio between the distance of the eyes and the distance between the ear and temple, for which a value of 0.5 produced satisfactory results. We assume the face's $\mathbf{d}_{\text{fwd}}$ vector to point in the opposite direction of a camera forward vector corresponding to camera extrinsics that belong to a virtual picture depicting a frontally oriented face. The camera transform with this property is estimated by comparing the point cloud of the detected landmarks – MEDIAPIPE FACE MESH provides pseudo depth values for each landmark – with the point cloud of reference landmarks derived from an image of a forward-facing generic face. Procrustes analysis provides a transformation matrix that best aligns these two point clouds, resulting in a front-facing camera view.

With the glasses reference points $\mathbf{g}_c, \mathbf{g}_r, \mathbf{g}_l$ and the landmarks $\mathbf{f}_c, \mathbf{f}_l, \mathbf{f}_r$ identified, the spectacle frame can now be positioned on the face. First, the frame is scaled with factor $\frac{|\mathbf{t}_l - \mathbf{t}_r|}{|\mathbf{g}_l - \mathbf{g}_r|}$ to ensure a proper fit. We then employ the KABSCH algorithm [Kab76], which finds the optimal rotation matrix that minimizes the least-squares difference between two sets of paired points. The translation $\mathbf{t}$ and rotation $\mathbf{R}$ required to position the glasses are computed as follows:

$$\mathbf{t} = -\mathbf{g_c} + \mathbf{f}_c \tag{4}$$

$$\mathbf{R} = \text{KABSCH}\left(\{\mathbf{g}_r - \mathbf{g}_c, \mathbf{g}_l - \mathbf{g}_c\}, \{\mathbf{f}_r - \mathbf{f}_c, \mathbf{f}_l - \mathbf{f}_c\}\right) \tag{5}$$

3.2. Hybrid Rendering

To coherently render surfaces and NeRFs in the same scene, we employ a two-stage renderer. The first stage involves rendering the meshes. We adopt a simplified Disney BRDF model in OPTIX to ray trace the meshes [BS12]. To mitigate jagged edges in the final composed scene, the first pass is anti-aliased by multisampling every pixel with random jitter [Ken13]. For a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, the first rendering stage yields the color $S_c(\mathbf{r})$ and opacity $S_\alpha(\mathbf{r})$ of the rendered pixel, as well as the depth $S_t(\mathbf{r})$ of the surface that was hit by $\mathbf{r}$. If the ray does not hit a surface, $S_\alpha(\mathbf{r}) = 0$. For fully opaque surfaces, $S_\alpha(\mathbf{r}) = 1$.

The second rendering stage produces the final rendered image, compositing volumetric NeRF data with the rendered surfaces. For a ray $\mathbf{r}$ between a near and far depth t_n and t_f, the rendered color $C(\mathbf{r}, t_n, t_f)$ of a NeRF is defined as follows [MST*20]:

$$C(\mathbf{r}, t_n, t_f) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\,\mathbf{c}(\mathbf{r}(t), \mathbf{d})\,dt \tag{6}$$

where $T(t)$ denotes the accumulated transmittance from t_n to t along $\mathbf{r}$, σ is the volume density and $\mathbf{c}$ the RGB color at a given point (and viewing direction) in space. For a more detailed explanation we refer to [MST*20]. To incorporate surface data into an integrated NeRF-mesh rendering approach, we introduce the formulation of the rendered color $X(\mathbf{r}, t_n, t_f)$ as:

$$X(\mathbf{r}, t_n, t_f) = \begin{cases} C(\mathbf{r}, t_n, t_s) + T(t_s)S_c(\mathbf{r}) & \text{if } S_\alpha(\mathbf{r}) = 1 \\ C(\mathbf{r}, t_n, t_s) + T(t_s)S_c(\mathbf{r}) \\ \quad + (T(t_s) - S_\alpha(\mathbf{r}))\,C(\mathbf{r}, t_s, t_f) & \text{if } S_\alpha(\mathbf{r}) > 0 \\ C(\mathbf{r}, t_n, t_f) & \text{otherwise} \end{cases} \tag{7}$$

where t_s is short for $S_t(\mathbf{r})$. In Eq. 7, the mesh color $S_c(\mathbf{r})$ from the first rendering pass is composited with NeRF color and density information. The formulation considers both Mesh-NeRF and NeRF-Mesh occlusions.

In NeRF rendering, the continuous integral is numerically estimated using ray marching. If the ray marching process reaches a depth $S_t(\mathbf{r})$ at which the first rendering pass encountered a ray hit on an opaque surface, the ray marching process is terminated early, as the volumetric data behind the surface is occluded and can thus be discarded. This is also accounted for in our hybrid formulation, which is the case when $S_\alpha(\mathbf{r}) = 1$.

	Frame Times [ms]					
	Head Avatar			Head Avatar + Glasses		
	480p	720p	1080p	480p	720p	1080p
Synth	12.7	19.7	35.6	12.5	19.6	34.9
Real	14.4	28.4	54.1	14.1	27.6	51.6

Table 1: Frame times for various rendering resolutions using an RTX 3090. All resolutions have a 16:9 aspect ratio. Frame times are captured both with and without rendering the glasses frame.

4. Results

We implement our hybrid renderer on top of Instant-NGP [MESK22], as it provides state-of-the-art training and rendering times. The CUDA rendering code is adapted to facilitate composition of mesh and NeRF color information, as well as early ray termination upon surface hits. The routines for loading NeRFs and meshes from files, as well as the routines for transforming them in 3D space are exposed with Python bindings. Thereby, the performance critical rendering code, particularly ray tracing and -marching, is implemented in a capsuled and runtime-optimized C++/CUDA library, while less demanding tasks, particularly the identification of face landmarks and optimization of the glasses position is implemented in Python. This allows us to conveniently leverage existing Python libraries for these tasks.

We evaluate our approach for virtual try-on of glasses on two human subjects – one synthetic and one real – with two glasses frames each. Each NeRF model is trained for 1 minute on an RTX 3090 GPU. After training, the placement process outlined in Sec. 3.1 is conducted, taking another 5-10 seconds. Finally, the hybrid scene can be explored in our renderer with interactive frame rates.

Table 1 lists the frame times for three different render resolutions and the two introduced NeRF avatars. The frame time experiments were conducted on an RTX 3090. Interactive frame rates of 20 - 30 FPS are achieved for Full HD resolution (1080p). Frame times differ between the two avatars, since in the real scene, the shoulders of the subject are included and thus cause a larger volume to be rendered. Notably, adding the glasses to the rendering process reduces rendering times by 2-4%, even though an additional ray-tracing rendering pass is computed for the glasses frame. This is caused by early ray termination during NeRF rendering that occurs when the ray marching depth reaches an opaque mesh surface.

The hybrid scene can be viewed in an interactive fashion, allowing the user to freely orbit around the head to view the glasses from any angle. The obtained results are presented in Fig. 2. In all four

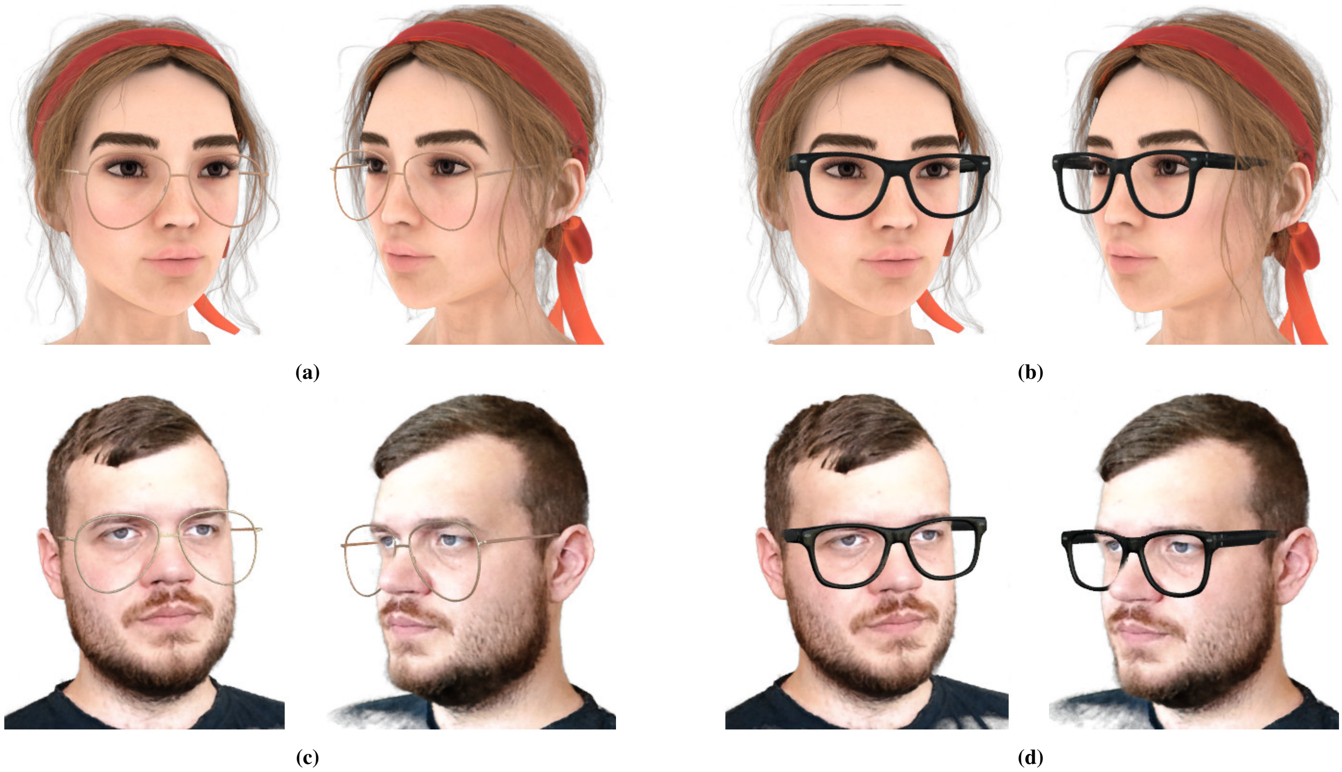


Figure 2: Rendered views of hybrid scenes with a synthetic (a, b) and real (c, d) human NeRF and two glasses frames. Video results are provided as supplementary material. The glasses correctly rest on the subject's nose and ears and are properly occluded by hair and facial geometry. They are positioned consistently across all views, as they share the same scene space with the NeRF head avatars.

scenarios, the glasses are properly aligned on the faces, resting on the nose and ears and not clipping into the NeRF geometry. The glasses' edges do not appear jagged and the semi-transparent pixels resulting from anti-aliasing blend seamlessly with the colors of the NeRF. Notably, this semi-transparent occlusion works both ways, which is evident in Fig. 2a and 2b, where hair strands convincingly occlude the frames' arms.

Although the results may suggest that the real human subject was videographed in a controlled environment with a white background, this is not the case. The subject was captured in a casually lit room and instructed to sit on a chair, in order to minimize movements during the capturing process. The neutral white background is achieved by limiting the bounding box for volume marching, effectively pruning any scene backgrounds.

5. Conclusion

In this work, we propose virtual try-on for glasses incorporating a hybrid rendering approach, that combines mesh and NeRF rendering. We leverage projections of 2D facial landmarks into 3D space in combination with an approximation of the position of the human ear to position meshed models of glasses in relation to a NeRF human head avatar represented as a NeRF. The proposed combined renderer integrates the two representations in one coherent scene. Images of the human subject can be captured in an environment with unconstrained lighting using a handheld camera or smartphone. The required preprocessing time for NeRF training and placement of the glasses mesh, after which the rendering process is available, constitutes less than 2 minutes. The presented experiments show consistent placement of the glasses frame and plausible occlusions, as well as improved rendering times due to early ray termination.

5.1. Limitations & Future Work

Our approach does currently not incorporate spectacle lenses. As the severity of ametropia has an effect on the shape and thickness of the utilized spectacle lenses, the lenses refractive properties vary between customers. Consequently, future work should consider lenses with variable refraction in the ray marching process [LSS*23]. Furthermore, the proposed methodology does not employ real world measurements of the human head avatar, leading to estimated relative sizes between the glasses and the human head. With the proliferation of depth sensors in modern phones, accurate measurements may be obtained during the capturing process without additional visual cues [MAH*22]. Future work may exploit these sensors to gauge head dimensions and thereby place glasses with accurate real world measurements into the scene. Finally, our approach leverages Instant-NGP and is therefore not robust against movements during the capturing process. To mitigate this, strategies that are more robust against dynamic scenes [PSB*21; PSH*21] could be employed in future work.

References

[ADD16] Azevedo, Pedro, Dos Santos, Thiago Oliveira, and De Aguiar, Edilson. "An augmented reality virtual glasses try-on system". *2016 XVIII Symposium on Virtual and Augmented Reality (SVR)*. IEEE. 2016, 1–9. doi: 10.1109/SVR.2016.12 2.

[BS12] Burley, Brent and Studios, Walt Disney Animation. "Physically-based shading at disney". *Acm Siggraph*. Vol. 2012. vol. 2012. 2012, 1–7 3.

[FJS18] Feng, Zhuming, Jiang, Fei, and Shen, Ruimin. "Virtual glasses try-on based on large pose estimation". *Procedia computer science* 131 (2018), 226–233. doi: 10.1016/j.procs.2018.04.207 2.

[GTZN21] Gafni, Guy, Thies, Justus, Zollhofer, Michael, and Niessner, Matthias. "Dynamic neural radiance fields for monocular 4d facial avatar reconstruction". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, 8649–8658. doi: 10.1109/CVPR46437.2021.00854 2.

[HPX*22] Hong, Yang, Peng, Bo, Xiao, Haiyao, et al. "Headnerf: A real-time nerf-based parametric head model". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 20374–20384 2.

[JKK*23] Jambon, Clément, Kerbl, Bernhard, Kopanas, Georgios, et al. "NeRFshop: Interactive Editing of Neural Radiance Fields". *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 6.1 (2023). doi: 10.1145/3585499 2.

[Kab76] Kabsch, W. "A solution for the best rotation to relate two sets of vectors". *Acta Crystallographica Section A* 32.5 (Sept. 1976), 922–923. doi: 10.1107/S0567739476001873 3.

[Ken13] Kensler, Andrew. *Correlated multi-jittered sampling*. Technical Memo #13-01. Pixar. 2013 3.

[KSSU19] Kobayashi, Takumi, Sugiura, Yuta, Saito, Hideo, and Uema, Yuji. "Automatic eyeglasses replacement for a 3d virtual try-on system". *Proceedings of the 10th Augmented Human International Conference 2019*. 2019, 1–4. doi: 10.1145/3311823.3311854 2.

[LGO*23] Lazova, Verica, Guzov, Vladimir, Olszewski, Kyle, et al. "Control-nerf: Editable feature volumes for scene rendering and manipulation". *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, 4340–4350. doi: 10.1109/WACV56688.2023.00432 2.

[LSS*23] Li, Junxuan, Saito, Shunsuke, Simon, Tomas, et al. "Megane: Morphable eyeglass and avatar network". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 12769–12779. doi: 10.1109/CVPR52729.2023.01228 1, 2, 4.

[LTN*19] Lugaresi, Camillo, Tang, Jiuqiang, Nash, Hadon, et al. *MediaPipe: A Framework for Building Perception Pipelines*. 2019. arXiv: 1906.08172 [cs.DC] 1, 2.

[MA21] Milanova, Mariofanna and Aldaeif, Fatima. "Markerless 3D virtual glasses try-on system". *New Approaches for Multidimensional Signal Processing: Proceedings of International Workshop, NAMSP 2020*. Springer. 2021, 99–111. doi: 10.1007/978-981-33-4676-5_7 2.

[MAH*22] Mikalai, Zamotsin, Andrey, Dyagilev, Hawas, Hafez S, et al. "Human body measurement with the iPhone 12 Pro LiDAR scanner". *AIP Conference Proceedings*. Vol. 2430. 1. AIP Publishing. 2022. doi: 10.1063/5.0078310 4.

[MESK22] Müller, Thomas, Evans, Alex, Schied, Christoph, and Keller, Alexander. "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding". *ACM Trans. Graph.* 41.4 (July 2022), 102:1–102:15. doi: 10.1145/3528223.3530127 1–3.

[MST*20] Mildenhall, Ben, Srinivasan, Pratul P., Tancik, Matthew, et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". *ECCV*. 2020. doi: 10.1145/3503250 1–3.

[PSB*21] Park, Keunhong, Sinha, Utkarsh, Barron, Jonathan T, et al. "Nerfies: Deformable neural radiance fields". *ICCV*. 2021, 5865–5874. doi: 10.1109/ICCV48922.2021.00581 1, 2, 4.

[PSH*21] Park, Keunhong, Sinha, Utkarsh, Hedman, Peter, et al. "HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields". *ACM Trans. Graph.* 40.6 (Dec. 2021). doi: 10.1145/3478513.3480487 1, 2, 4.

[QGX*23] Qiao, Yi-Ling, Gao, Alexander, Xu, Yiran, et al. "Dynamic Mesh-Aware Radiance Fields". *ICCV* (2023). doi: 10.1109/ICCV51070.2023.00042 1, 2.

[RWL*24] Rak, Arne, Wirth, Tristan, Lindemeier, Thomas, et al. *NeRF-Glasses*. 2024. url: https://github.com/arnerak/nerf-glasses 1.

[SOTC22] Santesteban, Igor, Otaduy, Miguel, Thuerey, Nils, and Casas, Dan. "Ulnef: Untangled layered neural fields for mix-and-match virtual try-on". *Advances in Neural Information Processing Systems* 35 (2022), 12110–12125 1.

[TCWZ22] Tang, Jiaxiang, Chen, Xiaokang, Wang, Jingbo, and Zeng, Gang. "Compressible-composable nerf via rank-residual decomposition". *Advances in Neural Information Processing Systems* 35 (2022), 14798–14809 2.

[TZT*14] Tang, Difei, Zhang, Juyong, Tang, Ketan, et al. "Making 3D eyeglasses try-on practical". *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE. 2014, 1–6 2.

[WCS*22] Weng, Chung-Yi, Curless, Brian, Srinivasan, Pratul P, et al. "Humannerf: Free-viewpoint rendering of moving people from monocular video". *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*. 2022, 16210–16220. doi: 10.1109/CVPR52688.2022.01573 2.

[YKF*11] Yuan, Miaolong, Khan, Ishtiaq Rasool, Farbiz, Farzam, et al. "A mixed reality system for virtual glasses try-on". *Proceedings of the 10th international conference on virtual reality continuum and its applications in industry*. 2011, 363–366. doi: 10.1145/2087756.2087816 2.

[YWTZ24] Ye, Keyang, Wu, Hongzhi, Tong, Xin, and Zhou, Kun. "A Real-time Method for Inserting Virtual Objects into Neural Radiance Fields". *IEEE Transactions on Visualization and Computer Graphics* (2024). doi: 10.1109/TVCG.2024.3422814 1, 2.

[ZAB*22] Zheng, Yufeng, Abrevaya, Victoria Fernández, Bühler, Marcel C, et al. "Im avatar: Implicit morphable head avatars from videos". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 13545–13555 2.

[ZBT23] Zielonka, Wojciech, Bolkart, Timo, and Thies, Justus. "Instant volumetric head avatars". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 4574–4584. doi: 10.1109/CVPR52729.2023.00444 1, 2.

[ZGL*17] Zhang, Qian, Guo, Yu, Laffont, Pierre-Yves, et al. "A virtual try-on system for prescription eyeglasses". *IEEE computer graphics and applications* 37.4 (2017), 84–93. doi: 10.1109/MCG.2017.3271458 2.

[Zha18] Zhang, Boping. "Augmented reality virtual glasses try-on technology based on iOS platform". *EURASIP Journal on Image and Video Processing* 2018.1 (2018), 1–19. doi: 10.1186/s13640-018-0373-8 2.