Holoported Characters: Real-time Free-viewpoint Rendering of Humans from Sparse RGB Cameras

Ashwath Shetty[1,2] Marc Habermann[1,3] Guoxing Sun[1] Diogo Luvizon[1,3]
Vladislav Golyanik[1] Christian Theobalt[1,3]
[1] Max Planck Institute for Informatics, Saarland Informatics Campus [2] Saarland University
[3] Saarbrücken Research Center for Visual Computing, Interaction and AI

Abstract

We present the first approach to render highly realistic free-viewpoint videos of a human actor in general apparel, from sparse multi-view recording to display, in real-time at an unprecedented 4K resolution. At inference, our method only requires four camera views of the moving actor and the respective 3D skeletal pose. It handles actors in wide clothing, and reproduces even fine-scale dynamic detail, e.g. clothing wrinkles, face expressions, and hand gestures. At training time, our learning-based approach expects dense multi-view video and a rigged static surface scan of the actor. Our method comprises three main stages. Stage 1 is a skeleton-driven neural approach for high-quality capture of the detailed dynamic mesh geometry. Stage 2 is a novel solution to create a view-dependent texture using four test-time camera views as input. Finally, stage 3 comprises a new image-based refinement network rendering the final 4K image given the output from the previous stages. Our approach establishes a new benchmark for real-time rendering resolution and quality using sparse input camera views, unlocking possibilities for immersive telepresence. Code and data is available on our project page.

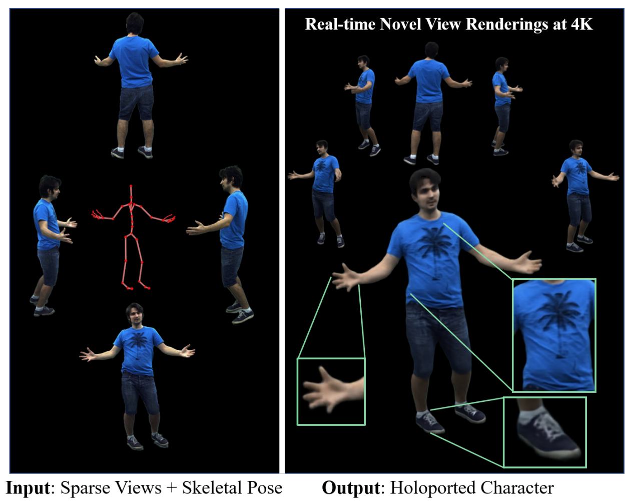

Figure 1. We propose *Holoported Characters*, a novel approach for real-time free-view point rendering of humans at 4K resolution. During inference, our method only requires four sparse images observing the human and the respective 3D skeletal pose. Then, our three-stage pipeline generates novel views of the performance in real-time and at an unprecedented resolution of 4K. We highlight that our approach can account for detailed effects such as clothing wrinkles, facial expressions, and hand gestures.

1. Introduction

Human free-viewpoint rendering is a long-standing and highly challenging problem in Vision and Graphics. The goal is to render any virtual viewpoint of the character given a discrete set of input camera views. Earlier approaches resorted to variants of explicit multi-view photometric reconstruction [4, 33, 34], reconstruction based on light fields [22], point primitives [40], or template-based scene representations [1, 5, 43] to compute an estimate of dynamic shape and appearance from the multi-view input. Despite great progress, these solutions are often limited along multiple dimensions, for example: They often require a very high number of camera views for good quality; computation times are often far from real-time; ghosting artifacts in rendered appearance are prevalent since even the best methods fail to capture error-free scene geometry.

In recent years, a new class of approaches to human free-viewpoint rendering that combines explicit dynamic scene representation with neural-network-based representation and image formation has led to a boost in result quality. These methods utilize neural implicit scene representations to encode the moving human [6, 7, 15, 20, 23, 25–29, 37]. However, even the most advanced learning-based approaches face clear limitations: Real-time processing from capture to rendering on the basis of neural implicit representations is hard to achieve [7]. Rendering resolution is often limited [17, 26, 28]. Capturing and displaying fine-scale dynamic effects, such as face expressions, hand gestures, clothing dynamics, or the waving of hair, is often impossible [6, 7, 20] or requires complicated tracking and registration pipelines [39, 42]. Finally, even for some of the best approaches, there are clear differences between ground-truth reference views and renderings, since methods aim for plausibility (where details can differ) rather than truthful detail reproduction [6, 7, 15, 23, 27, 37] and loose types of apparel are often times out of reach [28].

We, therefore, present Holoported Characters, a new method for human free-viewpoint rendering that is the first combining the following properties: It is end-to-end real-time at test time, potentially enabling live capture and free-viewpoint rendering of a human in general wide clothing at unprecedented 4K image resolution. It achieves state-of-the-art and truthful, not merely plausible, free-viewpoint rendering quality of even fine details (see Fig. 1). It only requires four camera views at test time. It truthfully reproduces even face expressions, hand and finger gestures, and dynamic details of loose clothing. This combination paves the way for high-quality rendering in real-time merged reality and telepresence.

The training phase of our algorithm requires as input a static 3D scan of the person rigged with a skeleton, as well as dense multi-view video of the moving person. Our algorithm operates in three stages, each of which introduces important contributions: In stage 1, our first contribution is an improved real-time approach for neural network-based skeleton-driven deformation of the template mesh. It extends the approach of Habermann et al. [6] by training the deformation method on the multi-view video using both color-based supervision as well as supervision from neural SDF reconstructions, which greatly enhances 3D shape quality. In stage 2, given stage-1 human mesh reconstructions, we propose a real-time projective texturing pipeline that maps the images onto the texture space of the mesh, and a new neural-network-based method computes a dynamic and view-dependent surface texture and feature map from this projected texture. It learns to compute a complete coherent surface texture with minimal distortion despite potential inaccuracies in stage-1 geometry. In stage 3, our new image-based refinement network takes the stage-1 geometry rendered with the stage-2 texture and features as input and transforms it into the final high-resolution rendering.

We validate our design through thorough ablations, and demonstrate state-of-the-art quality in our experiments.

2. Related Work

Free-view Replay. Methods for novel view synthesis of general non-rigid scenes [16, 21, 32, 35] can be applied to videos with humans but struggle with large articulations due to the absence of human-specific priors. They enable scene replay but do not allow changes in the human pose without model retraining. Other human-specific replay methods can be trained using monocular [41] or multi-view videos [9, 49]. The visual quality of monocular methods can suffer due to the lack of explicit 3D information [41]. Multi-view methods [9, 49] can also render complex appearance elements like hair and clothing, even in real-time [49]. However, as they can solely replay the multi-view sequence they were trained on, they are not well suited for interactive telepresence applications, which is the focus of this work.

Animatable Avatars. Recent approaches generate novel 2D views of humans in novel poses but do not allow freely changing the 3D viewpoint [2, 11, 18, 19, 30]. Specialized approaches for the 3D free-viewpoint rendering of humans from monocular or multi-view RGB videos achieve high-fidelity results [10, 20, 41, 44, 47]; some of them generalize to poses unseen during training [6, 7, 15, 23, 27, 29, 37]. Their major limitation is that details such as clothing wrinkles are blurred as the pose information is ambiguous: A single pose can induce a variety of wrinkles, which often happens during training and which causes averaging of fine appearance details. While HDHumans [7] and other concurrent works [14, 24] can generate high-frequency wrinkles (due to the generative formulation), those are often hallucinated and are not consistent with the ground-truth observations. Moreover, many methods also strongly rely on a neural rendering component, which makes them slow and not suitable for immersive applications. DDC [6] is one of the few methods which run in real-time and enable animatable control over the character: It generates an explicit mesh and texture and supports loose clothing. However, it still suffers from blurred details as most other methods.

Image-driven Dynamic Scene Rendering. Approaches that use images for novel view rendering of dynamic scenes could also be applied to our task [17, 36, 45]. The recent work ENeRF [17] is an interactive and real-time approach for free-viewpoint rendering with a neural representation driven by a sparse set of multi-view images. However, due to the lack of human priors, the method suffers from multi-view-consistency artifacts and does not generalize well to new poses and views. In contrast, our method achieves view consistent and high-quality results in real-time as we explicitly account for the humanoid structure, i.e. its articulation and non-rigid deformations.

Sparse Image-driven Avatars. In contrast to the previous paragraph, some image-driven works [3, 13, 28] explicitly model human priors. Neural Image-based Avatars [13] is a generalizable approach, which can drive arbitrary human performers from sparse images and 3D poses. However, it fails to generate high-quality wrinkles and expres-

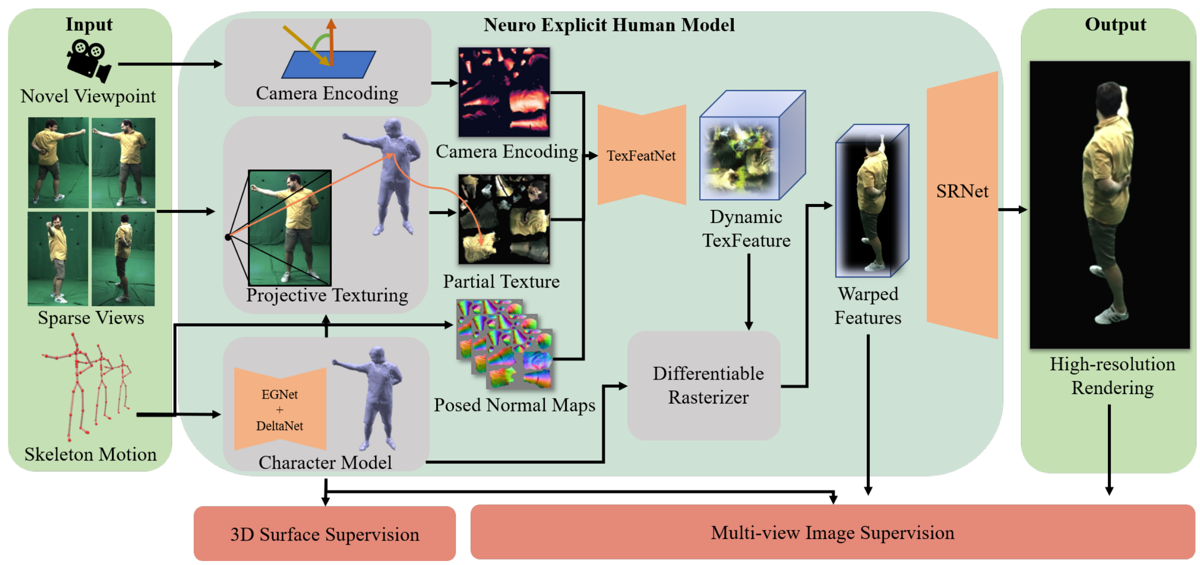

Figure 2. **Method Overview.** *Holoported Characters* takes sparse camera views, the respective 3D skeletal pose, and the camera parameters of the novel view as input and generates high-resolution rendering in real-time. Our character model takes the motion as input and predicts a pose-dependent deformation of the template mesh. Then, our projective texturing pipeline maps the sparse views onto this mesh's texture space. This texture, camera encoding, and posed normal maps are then fed into our TexFeatNet, producing a view-dependent dynamic texture feature. Finally, our SRNet takes those low-resolution features in image space and generates the high-resolution rendering.

sions for new identities. UV Volumes [3] is a recent work that leverages a neural texture stack and generates a UV volume for real-time free-viewpoint rendering. However, they require a dense capture setting at test time, whereas our approach requires solely four cameras. Drivable Volumetric Avatars (DVA) [28] is most closely related to ours in that it achieves high-quality free-viewpoint renderings of humans in real-time from 3D skeletal pose and sparse RGB images. In contrast to our method, they represent the virtual character as volumetric primitives loosely attached to a skinned mesh. We found that their formulation is limited to tight types of apparel, and their model does not scale well for a large variety of poses. Instead, our approach features a deformable character model capable of dealing with loose clothing. Moreover, as we optimize explicit structures like meshes and textures, we demonstrate high-quality appearance recovery as well as generalizability to arbitrary poses.

3. Method

Given multi-view images of an actor, our objective is to train a person-specific model capable of generating photoreal renderings during inference. The model takes *sparse* multi-view images, the corresponding 3D skeletal pose, and the target virtual camera as input, and produces a rendering for the specified view (see Fig. 2). First, we introduce our character model (Sec. 3.1), which generates a posed character mesh from the 3D skeletal pose (stage 1). Then, we explain our texture projection module (Sec. 3.2), which projects the input images onto the mesh and generates a partial texture map effectively encoding the high-frequency visual information of the sparse views (stage 2). Following this, our TexFeatNet (Sec. 3.3) generates temporally stable, view-conditioned, and complete texture and features for the target view, which are rendered into image space and fed finally to our SRNet module (Sec. 3.4), which generates the final 4K renderings (stage 3).

3.1. Deformable Character Model

First, we introduce our deformable mesh-based model of the human, which is image-agnostic, i.e. only depends on the skeletal motion. It will later serve us as a proxy for our projective texturing (Sec. 3.2) as we aim at learning the appearance in 2D texture/image space rather than in 3D, as this is efficient and compatible with 2D convolutions. This model takes the 3D motion M_t as input, and poses a template mesh where t denotes the time. We build on top of the learned character model of Habermann et al. [6],

$$C_i(M_t, f_{\text{eg}}(M_t), f_{\text{delta}}(M_t)) = \boldsymbol{v}_i, \tag{1}$$

as it is differentiable, real-time, and has the capacity to effectively model loose clothing. The model incorporates a structure-aware graph neural network $f_{\text{eg}}(M_t)$ to address coarse deformations. This network predicts the rotation and translation of the nodes of an Embedded Graph [31] of the template. Simultaneously, the network $f_{\text{delta}}(M_t)$ addresses fine deformations, predicting per-vertex displacements of the template.

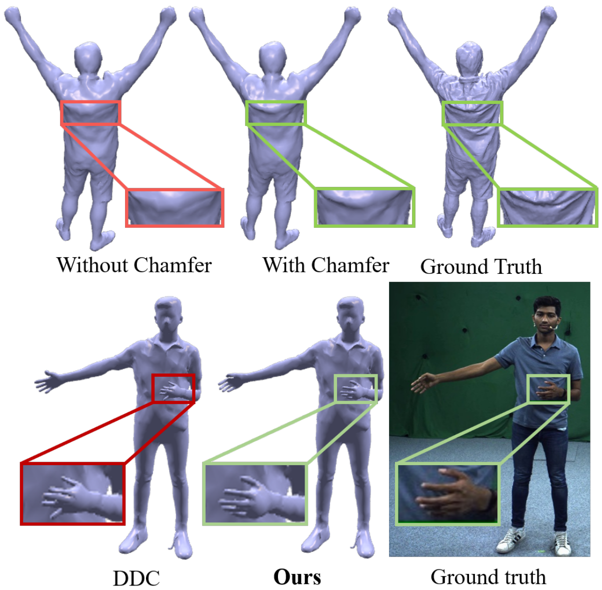

Figure 3. **Recovery of Geometric Details.** Our proposed point-cloud supervision and hand modeling helps us recover more details such as wrinkles and hand gestures compared to the baseline.

The resultant deformations are applied to the template in canonical space, followed by posing using Dual Quaternion Skinning [12], yielding the final positioned vertex $\boldsymbol{v}_i \in \mathbb{R}^3$ for each vertex i of the template mesh. The aggregate of these posed vertices $\boldsymbol{V} \in \mathbb{R}^{N\times 3}$ is derived by stacking each $\boldsymbol{v}_i$, where N represents the number of vertices in the template. Their entire model is trained in multiple stages, using only multi-view data. Please refer to the supplementary material and the original work [6] for more technical details.

Improvements to the Character Model. In practice, relying solely on multi-view images for model training yields imprecise reconstruction, which hinders our projective texturing pipeline. To improve the character model's surface quality, we first reconstruct a high-quality surface $\mathcal{S}$ per frame using recent state-of-the-art surface reconstruction methods [38]. We then leverage $\mathcal{S}$ to provide stronger 3D supervision to the displacement network $f_{\text{delta}}(M_t)$. In practice, we do this with an additional Chamfer loss:

$$\mathcal{L}_{\text{cham}}(\boldsymbol{V},\boldsymbol{S}) = \sum_{\boldsymbol{v}_\text{s}\in\boldsymbol{S}} \min_{\boldsymbol{v}\in\boldsymbol{V}} ||\boldsymbol{v}_\text{s}-\boldsymbol{v}||_2^2 + \sum_{\boldsymbol{v}\in\boldsymbol{V}} \min_{\boldsymbol{v}_\text{s}\in\boldsymbol{S}} ||\boldsymbol{v}_\text{s}-\boldsymbol{v}||_2^2 \quad (2)$$

where $\boldsymbol{S} \in \mathbb{R}^{S\times 3}$ is obtained by using Poisson disk sampling [46] on the surface $\mathcal{S}$. We observe that this term significantly improves the surface quality of the model. Fig. 3 shows the difference in geometry. Notice that the wrinkle on the back is significantly better captured.

Additionally, we observe that the coarse deformation, which $f_{\text{eg}}(M_t)$ predicts, causes artifacts for the hands, as they are a part of the body with high articulation very close together, which cannot be modeled by coarse deformation. To address this issue, we set specific parameters at hand vertices to zero, preventing the model to deviate from the initial skinning deformation. This significantly improves the performance, as illustrated in Fig. 3.

3.2. Efficient Projective Texturing

Now that we have a reasonable 3D proxy from our previous stage, we discuss how to encode the sparse image information, i.e. the four camera views, into a common texture space. We choose the texture space as it presents an efficient way of encoding appearance, as it is spatially aligned per frame, and rendering via a generated texture encodes a 3D bias into the network enforcing multi-view consistency. This motivates us to generate partial textures from the sparse input images to effectively encode appearance into a 2D representation while still being 3D-aware.

For each view i, we generate the partial texture $\boldsymbol{T}_{\text{part},i} \in \mathbb{R}^{W\times H\times 3}$ via inverse texture mapping, where W and H denote the texture map dimensions. In more detail, each texel (u, v) is transformed into 3D space using the UV mapping of our deformable mesh $\mathbf{V}$ and projected into the input view i where (x, y) denotes the projected image-space coordinates. We then store the color of the image at (x, y) in the texture space at (u, v). However, this operation does not account for the fact that some texels might not be visible from a particular view.

Thus, we compute per-view visibility masks $\boldsymbol{T}_{\text{vis},i} \in \mathbb{R}^{W\times H}$, which encode if a texel is visible in view i. For a more detailed derivation of the texel visibility, we refer to the supplement. Moreover, when the 3D surface normal of a texel (u, v) is almost perpendicular to the camera ray $\boldsymbol{d} \in \mathbb{R}^3$ of the pixel (x, y), texture projection suffers from distortions. Thus, despite the visibility, we add another condition encoded as the boolean texture map:

$$\boldsymbol{T}_{\text{angle},i}[u, v] = arcos(\boldsymbol{T}_{\text{norm}}[u, v] \cdot \boldsymbol{d}) < \delta, \quad (3)$$

which ensures that dot product between the texel normal $T_{\text{norm}} \in \mathbb{R}^{W\times H\times 3}$ and the camera viewing direction is smaller than a threshold angle δ. Our final computation for whether a texel is valid or not is defined as:

$$\boldsymbol{T}_{\text{valid},i} = \boldsymbol{T}_{\text{vis},i} \wedge \boldsymbol{T}_{\text{angle},i}, \quad (4)$$

where "$\wedge$" is the logical *and* operator.

Finally, we fuse the textures from all views via

$$\boldsymbol{T}_{\text{part}} = (\sum_{i\in C_{\text{in}}} \boldsymbol{T}_{\text{part},i} \circ \boldsymbol{T}_{\text{valid},i}) \circ (\frac{1}{\boldsymbol{T}_{\text{count}}}), \quad (5)$$

where "$\circ$" is the Hadamard product, C_{in} is the number of input camera views, $\boldsymbol{T}_{\text{part}} \in \mathbb{R}^{W\times H\times 3}$ is the final partial texture, and $\boldsymbol{T}_{\text{count}} \in \mathbb{R}^{W\times H}$ stores the number of valid views per texel. In practice, all operations can be efficiently

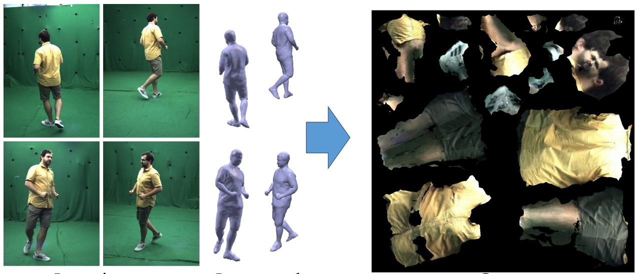

Figure 4. **Projective Texturing.** Given a sparse set of cameras and the posed character mesh, our method recovers a partial texture map using projective texturing, where pixels in screen space are mapped to texels of the texture map.

implemented using tensor operations, thus, leading to real-time computations. Fig. 4 shows a partial texture obtained by our method.

3.3. Tex-2-Tex Translation Network

While the generated partial textures effectively encode image information, refinement is essential for achieving photorealistic rendering since imperfect reconstruction and calibration often lead to artifacts. Additionally, some details may not be observed from the sparse input views. Hence, we propose a texture-to-texture network (TexFeatNet):

$$f_{\text{tex}}(\boldsymbol{T}_{\text{part}}, \boldsymbol{T}_{\text{motion}}, \boldsymbol{T}_{\text{cam}}) = \boldsymbol{T}_{\text{dyn}} = |\boldsymbol{T}_{\text{rgb}}|\boldsymbol{T}_{\text{feat}}| \quad (6)$$

$$\boldsymbol{T}_{\text{cam}} = \frac{\boldsymbol{T}_{\text{pos}} - \boldsymbol{o_c}}{||\boldsymbol{T}_{\text{pos}} - \boldsymbol{o_c}||} \quad (7)$$

that takes the sparse image information, geometry as well as viewing direction, and generates a view-dependent and dynamic texture and feature map. $\boldsymbol{T}_{\text{motion}} \in \mathbb{R}^{W \times H \times T \times 3}$ encodes information about the skeletal motion by baking the 3D surface normals of the posed mesh into a texture and stacking them over a time window of size T. This allows the network to recover pose-dependent appearance features even if such information is not present in the partial texture, i.e. the sparse input views. $\boldsymbol{o_c} \in \mathbb{R}^3$ is the camera origin, $\boldsymbol{T}_{\text{pos}} \in \mathbb{R}^{W \times H \times 3}$ stores the 3D coordinates of each texel and $\boldsymbol{T}_{\text{cam}} \in \mathbb{R}^{W \times H \times 3}$ encodes the viewing direction of the target view, enabling the network to learn view-dependent effects. $\boldsymbol{T}_{\text{part}}$ are the previously derived partial textures. Without them, we notice that results are blurrier due to the one-to-many mapping [20] between skeletal pose and surface appearance.

$\boldsymbol{T}_{\text{dyn}} \in \mathbb{R}^{W \times H \times 75}$ is the output of the network, and contains color ($\boldsymbol{T}_{\text{rgb}}$) in the first three channels and texel features $\boldsymbol{T}_{\text{feat}}$, which are both the input to our super-resolution module introduced next.

3.4. Differentiable Rendering and Super-Resolution

Given the deformable mesh $\mathbf{V}$ (Sec. 3.1) and the first three channels, i.e. $\boldsymbol{T}_{\text{rgb}}$, of the dynamic texture map $\boldsymbol{T}_{\text{dyn}}$, we can already render an image $\mathbf{I}_{\text{mesh},c}$ of the character with dynamic appearance using a standard rasterizer R as

$$\mathbf{I}_{\text{mesh},c} = R_c(\mathbf{V}, \boldsymbol{T}_{\text{rgb}}). \quad (8)$$

Here, c denotes the target camera view. However, this suffers from typical mesh rendering artifacts such as staircasing artifacts on the borders and also is limited by the resolution of the texture map.

To overcome this, we propose a super-resolution module (SRNet), which takes as input $\boldsymbol{I}_{\text{feat},c}$ computed as

$$\mathbf{I}_{\text{feat},c} = R_c(\mathbf{V}, \boldsymbol{T}_{\text{dyn}}), \quad (9)$$

i.e. by rendering the features and RGB texture onto screen space with a resolution of $W' \times H'$, and outputs a super-resolved image

$$f_{\text{sr}}(\boldsymbol{I}_{\text{feat},c}) = \boldsymbol{I}_{\text{sr},c}, \quad (10)$$

which has a resolution of $4W' \times 4H'$. As the SRNet purely operates in image space, we implement it as a shallow 2D CNN with small receptive fields. Thus, it is powerful enough to super-resolve the image with detail recovery while maintaining multi-view consistency.

Supervision. TexFeatNet f_{tex} (Sec. 3.3) and the SRNet f_{sr} (Sec. 3.4) are jointly trained. The TexFeatNet network is supervised using the rendering loss

$$\mathcal{L}_{\text{ren}} = \sum_{c,(x,y)\in\mathbb{R}^2} \|\boldsymbol{F}_c[x,y] \circ (\mathbf{I}_{\text{mesh},c}[x,y] - \boldsymbol{I}^l_c[x,y])\|_1, \quad (11)$$

whereas SRNet is supervised using

$$\mathcal{L}_{\text{sr}} = \sum_{c,(x,y)\in\mathbb{R}^2} \|\mathbf{I}_{\text{sr},c}[x,y] - (\boldsymbol{F}_c[x,y] \circ \boldsymbol{I}^h_c[x,y])\|_1. \quad (12)$$

$\boldsymbol{F}_c$ are the foreground matting masks and $\boldsymbol{I}^l_c$ and $\boldsymbol{I}^h_c$ are the low- and high-resolution ground-truth images, respectively. In $\mathcal{L}_{\text{sr}}$, we omit applying the matting mask to the super-resolved image, prompting the model to learn complete images with improved border details.

4. Results

Runtime. Utilizing dual NVIDIA A100 GPUs, our inference pipeline achieves real-time rendering at the full 4112×3008 resolution. GPU 1 runs the character model and projective texturing at 22 FPS, while GPU 2 handles TexFeatNet and SRNet at approximately 25 FPS. See the supplementary material for more details.

Figure 5. **Qualitative Results for Novel Poses and Views.** Our method generates high-quality renderings showing realistic wrinkle patterns and high-frequency details such as hands gestures and facial expressions. Note that our method is robust to challenging poses like squats and complicated clothing types such as loose skirts and highly textured garments, e.g. the pullover.

Dataset. We evaluate our method on two sequences from the *DynaCap* dataset [6], including one subject wearing tight clothes and one subject in loose clothes. Since in DynaCap the actors are with fists clenched, we also recorded a new dataset with three novel sequences where the actors were instructed to use both hands in a natural manner; Therefore, we can evaluate the capability of our method to represent fingers. We captured our new sequences in a multi-view camera setup with 120 synced 4K resolution cameras. Each sequence is performed by an actor wearing different clothes and is split into around 20K frames for training and 7K frames for testing. The camera streams are further divided into training and testing views. To ensure a fair comparison to other methods and to avoid a bias in the results due motion tracking errors (which is not the focus of this work), we run markerless motion capture on 20 cameras.

4.1. Qualitative Results

We evaluate our method qualitatively considering *novel view synthesis* and *novel pose synthesis*. Our method generalizes to novel poses at test time as demonstrated in Fig. 5, where we show results for unseen poses rendered under new viewpoints. Note the quality of hands, clothing wrinkles, and expressions that can be generated by our method in real time, even for unseen poses. We would also like to point out that our approach is able to model wrinkles faithfully, even for very challenging loose clothes, as shown for subject S2. This is due to the ability of our character model to generalize in this setting. Finally, note the quality of hands and facial expressions, for subjects S4 and S5, which is pivotal for immersive telepresence. Additionally, we provide more visualizations and applications such as texture editing as well as telepresence in the supplemental material.

4.2. Comparison

Comparison to Animatible Methods. In Fig. 6, we qualitatively compare our method to approaches that solely take

Figure 6. **Comparison with Animatible Approaches.** Here, we show results for novel poses. Note that DDC [6] fails to produce high-frequency details and cloth wrinkles. HDHumans [7] can hallucinate high-frequency details, however, they do not match the ground truth. In contrast, our method produces high-quality renderings, which are consistent with the ground truth.

Figure 7. **Comparison with Image-driven Approaches.** ENeRF [17] produces artifacts under far views and DVA [28] suffers from blurred renderings. In contrast, our method generalizes well to far views and produces sharp results.

the skeletal pose as input during test time. DDC [6] is the state of the art in terms of runtime as it demonstrates real-time performance, while HDHumans [7] can be considered the state of the art in terms of photorealism though not achieving real-time performance. Note that DDC often synthesizes blurred textures resulting in a lack of detail as the pose to appearance mapping is not a bijection and they do not explicitly account for that. While HDHumans generates sharper wrinkles, they do not match the ground truth caused by their adversarial training (see insets in Fig. 6). Moreover, both methods cannot account for facial expressions and hand gestures. In contrast, our method faithfully recovers the ground truth details, and is able to generate facial expressions and hand gestures.

Comparison to Image-driven Methods. Next, we compare to methods, that take sparse multi-view images as input and render free-viewpoint videos in real time. ENeRF [17] is a method designed for general scenes, while DVA [28] is specifically designed for humans, thus, most closely related to our approach. Originally, ENeRF uses the *nearest* two views *at test time* from a dense setup to produce the target view, which effectively violates our sparse view requirement during inference. Instead, we retrain ENeRF using all the training views/frames and always perform inference using four *fixed cameras*, which is the same setting used for all methods. As shown in Fig. 7, ENeRF, because of their constrained color formulation, fails to maintain multi-view consistency and produces blurry results for novel views, especially for those that are far away from the input views. We further compare to DVA [28]. Since DVA's character model is solely based on skinning and regularizes large deformations, it performs poorly in representing high-frequency details and loose types of apparel. We also found that their volume-primitive-based formulation is hard to train on a large variety of poses leading to a degradation in quality when training on large-scale data (such as the DynaCap dataset), which is essential to achieve desired pose generalization at test time. In consequence, we found that DVA produces blurred results and noticeable artifacts (see insets in Fig. 7). In contrast, our method results in high-quality renderings with sharp details, wrinkles, and it can also handle loose clothing. We refer to the supplemental for a more detailed analysis of these methods.

Evaluation Protocol. We quantitatively compare our method with the state-of-the-art approaches on two subjects of the DynaCap dataset. We trained all methods on every frame of the training sequence using the training camera views. For evaluating the novel view synthesis accuracy, we compute metrics on holdout views (cameras: 7, 18, 27, 40) on every 10th frame of the training sequence and report the average. For novel poses, we compute metrics on the same views but on the testing sequence for every 10th frame. Again, we report the average across frames. We report the Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) [48], and Frechet Inception Distance (FID) [8] in the following. For a fair comparison, we trained all approaches on 1K resolution.

Quantitative Comparisons. In Tab. 1, we report the quantitative comparisons for both tasks, i.e. novel view and pose synthesis. Our method outperforms all previous real-time methods by a significant margin, especially in LPIPS and FID metrics. This demonstrates its strong superiority in producing realistic renderings with fine details. Comparatively, our method performs on par with HDHumans, while running orders of magnitude faster. Most importantly, our method recovers the real details present in the ground truth, while HDHumans produces details that are perceptually plausible, but which not necessarily align with the

		Subject S1 (tight clothing)			Subject S2 (loose clothing)		
Method	**RT**	**PSNR**↑	**LPIPS**↓ (×1000)	**FID**↓	**PSNR**↑	**LPIPS**↓ (×1000)	**FID**↓
DDC	✔	32.96 (28.05)	20.07 (30.43)	27.73 (38.37)	27.92 (25.92)	36.33 (41.07)	47.23 (56.43)
ENeRF	✔	30.75 (30.54)	28.03 (29.41)	32.81 (36.39)	29.83 (**29.61**)	35.07 (36.08)	36.51 (41.08)
DVA	✔	31.65 (**30.60**)	26.27 (29.41)	35.68 (43.11)	27.32 (24.06)	40.95 (45.73)	144.21 (142.07)
Ours 1K	✔	**33.93** (30.19)	**14.38** (**24.95**)	**8.27** (**12.80**)	**33.18** (28.85)	**23.04** (**30.50**)	**11.10** (**18.01**)
HDHumans	✘	31.00 (27.69)	14.61 (24.00)	**4.93** (**9.25**)	28.98 (26.32)	26.28 (33.33)	6.83 (**11.87**)
Ours	✔	**33.93** (**30.24**)	**12.98** (**23.74**)	6.28 (11.62)	**31.45** (**28.03**)	**21.86** (**28.49**)	**5.95** (13.26)

Table 1. **Quantitative Comparison.** We present *novel view* and *novel pose* (brackets) synthesis results on subjects *S1* and *S2* from the DynaCap dataset [6]. Our method outperforms real-time approaches and particularly excels in loose clothing (*S2*). Bold indicates **best**.

Method	**PSNR**↑	**LPIPS**↓ (×1000)	**FID**↓	**Res.**
w/o Texture	26.44	44.34	66.13	1K
w/o Features	29.89	25.52	14.02	1K
w/o Chamfer	30.04	27.47	13.81	1K
w/o SR	29.76	27.82	15.17	1K
w/o 4K Train	30.19	24.95	12.80	1K
Ours	**30.24**	**23.74**	**11.62**	1K
w/o 4K Train	**28.81**	33.35	18.69	4K
Ours	28.75	**32.4**	**17.42**	4K

Table 2. **Quantitative Ablations.** We evaluate the main components of our method considering *novel pose synthesis* on subject *S1*. Every component in our pipeline contributes to our final results. Notably, the texture input, the super resolution (SR) module, and the Chamfer loss are the most relevant components.

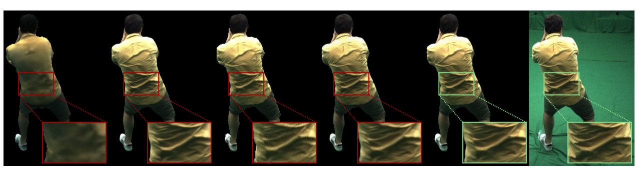

Figure 8. **Qualitative Ablations.** Note that all our design choices, i.e. partial texture features, super resolution, and 4K supervision, contribute to the final result quality.

ground truth. This is further verified by the difference in the PSNR metric between our methods.

4.3. Ablation

Partial Texture Input. Removing the partial textures as an input to our TexFeatNet turns our method into an animatable representation, which suffers from the ambiguous mapping from skeletal pose to appearance. Adding the partial texture instead effectively provides additional cues and, thus, results in better PSNR, LPIPS, and FID scores (Tab. 2). We additionally ablate the usefulness of the additional features, which are the output of our TexFeatNet along with the texture itself or, conversely, the input to the SRNet. The additional texture features boost all the metrics and also the general sharpness of the result (see "w/o Features" and "Ours" in Fig. 8). The additional features help to preserve some of the high-frequency information that might be lost in a pure color-based texture with finite spatial resolution, as well as the forward texture mapping process.

Geometric Details. We also evaluate the effect of the Chamfer loss for the 3D surface supervision in Tab. 2. It significantly improves the rendering quality, as it leads to surfaces that are better aligned with the input images.

4K Resolution. The super-resolution (SR) module significantly contributes to improving the visual quality and the metrics. Importantly, the SRNet removes observable noise, which is due to the discrete mapping of texels, and the typical mesh boundary artifacts (see the inset "w/o SR" and "Ours" in Fig. 8). Considering the rendering resolution, we observe that training our approach on 4K images translates into better metrics for both evaluation cases (1K and 4K image resolution), especially when considering LPIPS and FID, which are more sensitive to high-frequency details in comparison to PSNR.

5. Conclusions

We introduced Holoported Characters, a novel method for real-time free-viewpoint rendering of humans from sparse RGB cameras and 3D skeletal poses. Our approach achieves unprecedented 4K resolution and runtime performance by seamlessly integrating neural and explicit components. We believe our work is an important step towards telepresence enabling immersive communication across the globe. However, our method is not without limitations. For example, we cannot handle topological changes such as opening a jacket. For the future, we plan to explore multi-layered human representations potentially being able to model topological changes, which our method currently cannot handle.

Acknowledgments. This project was supported by the Saarbrücken Research Center for Visual Computing, Interaction and AI. Christian Theobalt was supported by ERC Consolidator Grant 4DReply (770784).

References

[1] Dan Casas, Marco Volino, John Collomosse, and Adrian Hilton. 4D Video Textures for Interactive Character Appearance. *Computer Graphics Forum (Proceedings of EUROGRAPHICS)*, 33(2):371–380, 2014. 1

[2] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. *IEEE International Conference on Computer Vision (ICCV)*, 1:0–0, 2019. 2

[3] Yue Chen, Xuan Wang, Xingyu Chen, Qi Zhang, Xiaoyu Li, Yu Guo, Jue Wang, and Fei Wang. Uv volumes for real-time rendering of editable free-view human performance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16621–16631, 2023. 2, 3

[4] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)*, 35(4): 114, 2016. 1

[5] Martin Eisemann, Bert De Decker, Marcus Magnor, Philippe Bekaert, Edilson de Aguiar, Naveed Ahmed, Christian Theobalt, and Anita Sellent. Floating textures. *Computer Graphics Forum (Proc. of Eurographics EG)*, 27(2):409–418, 2008. Received the Best Student Paper Award at Eurographics 2008. 1

[6] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics (TOG)*, 40(4):1–16, 2021. 1, 2, 3, 4, 6, 7, 8

[7] Marc Habermann, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhoefer, and Christian Theobalt. Hdhumans: A hybrid approach for high-fidelity digital humans. *Proc. ACM Comput. Graph. Interact. Tech.*, 6(3), 2023. 1, 2, 6, 7

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. 7

[9] Mustafa Iş ık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. HumanRF: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics*, 42(4): 1–12, 2023. 2

[10] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[11] Moritz Kappel, Vladislav Golyanik, Mohamed Elgharib, Jann-Ole Henningson, Hans-Peter Seidel, Susana Castillo, Christian Theobalt, and Marcus Magnor. High-fidelity neural human motion transfer from monocular video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1541–1550, 2021. 2

[12] Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Skinning with dual quaternions. In *Proceedings of the 2007 Symposium on Interactive 3D Graphics and Games*, page 39–46, New York, NY, USA, 2007. Association for Computing Machinery. 4

[13] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural image-based avatars: Generalizable radiance fields for human avatar modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2

[14] Youngjoong Kwon, Lingjie Liu, Henry Fuchs, Marc Habermann, and Christian Theobalt. Deliffas: Deformable light fields for fast avatar synthesis. *Advances in Neural Information Processing Systems*, 2023. 2

[15] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[16] Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, and Zhaoyang Lv. Neural 3d video synthesis from multi-view video. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[17] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields with learned depth-guided sampling. In *SIGGRAPH Asia Conference Proceedings*, 2022. 2, 7

[18] Lingjie Liu, Weipeng Xu, Michael Zollhöfer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)*, 38(5), 2019. 2

[19] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *Transactions on Visualization and Computer Graphics (TVCG)*, PP:1–1, 2020. 2

[20] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.*, 40(6), 2021. 1, 2, 5

[21] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[22] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 1

[23] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *International Conference on Computer Vision*, 2021. 1, 2

[24] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. 2023. 2

[25] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Animatable neural radiance fields for human body modeling. *ICCV*, 2021. 1
[26] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. *CVPR*, 1(1): 9054–9063, 2021. 2
[27] Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. Anr: Articulated neural rendering for virtual avatars. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
[28] Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, et al. Drivable volumetric avatars using texel-aligned features. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2, 3, 7
[29] Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, Alexander Vakhitov, and Victor Lempitsky. Textured neural avatars. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2
[30] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 2
[31] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing*. Eurographics Association, 2007. 3
[32] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2021. 2
[33] T. Tung, S. Nobuhara, and T. Matsuyama. Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *International Conference on Computer Vision (ICCV)*, pages 1709–1716. IEEE, 2009. 1
[34] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM Transactions on Graphics (TOG)*, 28(5):174, 2009. 1
[35] Chaoyang Wang, Lachlan Ewen MacDonald, László A. Jeni, and Simon Lucey. Flow supervision for deformable nerf. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
[36] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 2
[37] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *European Conference on Computer Vision*, 2022. 1, 2
[38] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 3295–3306, 2023. 4
[39] Ziyan Wang, Giljoo Nam, Tuur Stuyck, Stephen Lombardi, Michael Zollhoefer, Jessica Hodgins, and Christoph Lassner. Hvh: Learning a hybrid neural volumetric representation for dynamic hair performance capture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
[40] Michael Waschbüsch, Stephan Würmlin, Daniel Cotting, Filip Sadlo, and Markus Gross. Scalable 3d video of dynamic scenes. *The Visual Computer*, 21:629–638, 2005. 1
[41] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 2
[42] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, Yaser Sheikh, Jessica Hodgins, and Chenglei Wu. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *ACM Trans. Graph.*, 41(6), 2022. 2
[43] Feng Xu, Yebin Liu, Carsten Stoll, James Tompkin, Gaurav Bharaj, Qionghai Dai, Hans-Peter Seidel, Jan Kautz, and Christian Theobalt. Video-based characters: Creating new human performances from a multi-view video database. In *ACM SIGGRAPH 2011 Papers*, pages 32:1–32:10, New York, NY, USA, 2011. ACM. 1
[44] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *Neural Information Processing Systems*, 2021. 2
[45] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 2
[46] Cem Yuksel. Sample elimination for generating poisson disk sample sets. *Computer Graphics Forum*, 34, 2015. 4
[47] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Transactions on Graphics (TOG)*, 40(4):1–18, 2021. 2
[48] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, Los Alamitos, CA, USA, 2018. IEEE Computer Society. 7
[49] Fuqiang Zhao, Yuheng Jiang, Kaixin Yao, Jiakai Zhang, Liao Wang, Haizhao Dai, Yuhui Zhong, Yingliang Zhang, Minye Wu, Lan Xu, and Jingyi Yu. Human performance modeling and rendering via neural animated mesh. *ACM Trans. Graph.*, 41(6), 2022. 2