Deep Relightable Textures

Volumetric Performance Capture with Neural Rendering

ABHIMITRA MEKA*, Google, MPI Informatics, Saarland Informatics Campus
ROHIT PANDEY*, CHRISTIAN HÄNE, SERGIO ORTS-ESCOLANO, PETER BARNUM, PHILIP DAVIDSON, DANIEL ERICKSON, YINDA ZHANG, JONATHAN TAYLOR, SOFIEN BOUAZIZ, CHLOE LEGENDRE, WAN-CHUN MA, RYAN OVERBECK, THABO BEELER, PAUL DEBEVEC, and SHAHRAM IZADI, Google
CHRISTIAN THEOBALT, MPI Informatics, Saarland Informatics Campus
CHRISTOPH RHEMANN and SEAN FANELLO, Google

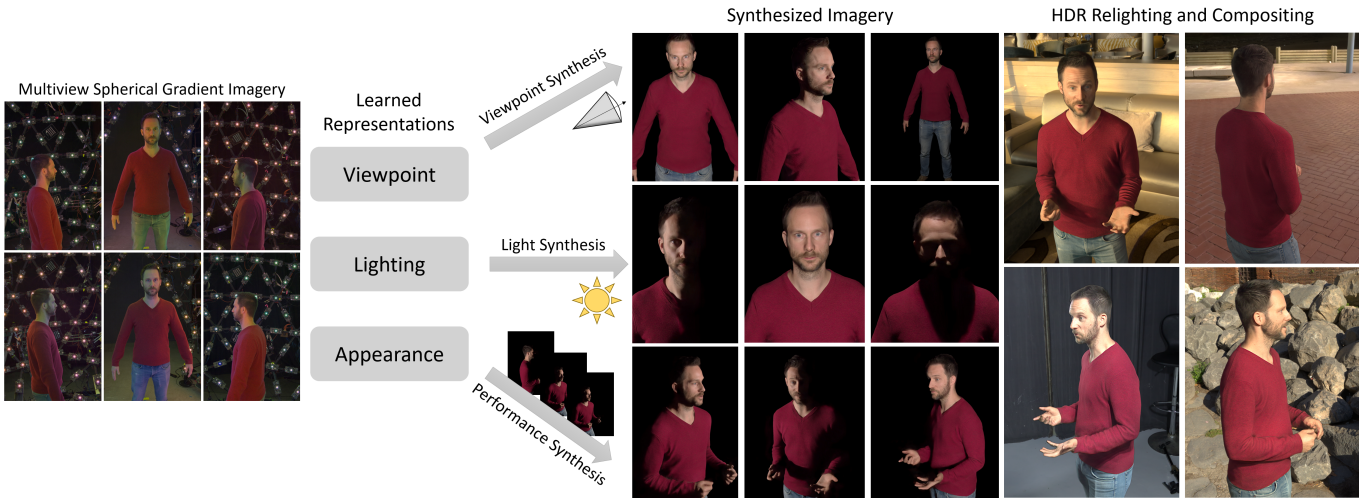

Fig. 1. Deep Relightable Textures – Our method is able to photo-realistically synthesize and composite *dynamic* performers under any *lighting condition* from a desired *camera viewpoint*.

The increasing demand for 3D content in augmented and virtual reality has motivated the development of volumetric performance capture systemsnsuch as the Light Stage. Recent advances are pushing free viewpoint relightable videos of dynamic human performances closer to photorealistic quality. However, despite significant efforts, these sophisticated systems are limited by reconstruction and rendering algorithms which do not fully model complex 3D structures and higher order light transport effects such as global illumination and sub-surface scattering. In this paper, we propose a system that combines traditional geometric pipelines with a neural rendering scheme to generate photorealistic renderings of dynamic performances under desired viewpoint and lighting. Our system leverages deep neural networks that model the classical rendering process to learn implicit features that represent the view-dependent appearance of the subject independent of the geometry layout, allowing for generalization to unseen subject poses and even novel subject identity. Detailed experiments and comparisons demonstrate the efficacy and versatility of our method to generate high-quality results, significantly outperforming the existing state-of-the-art solutions.

CCS Concepts: • **Computing methodologies** → **Computer vision**; **Machine learning**; **Rendering**.

Additional Key Words and Phrases: neural rendering, volumetric capture, reflectance estimation, novel view synthesis, relighting

ACM Reference Format:
Abhimitra Meka, Rohit Pandey, Christian Häne, Sergio Orts-Escolano, Peter Barnum, Philip Davidson, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, Chloe Legendre, Wan-Chun Ma, Ryan Overbeck, Thabo Beeler, Paul Debevec, Shahram Izadi, Christian Theobalt, Christoph Rhemann, and Sean Fanello. 2020. Deep Relightable Textures: Volumetric Performance Capture with Neural Rendering. *ACM Trans. Graph.* 39, 6, Article 259 (December 2020), 21 pages. https://doi.org/10.1145/3414685.3417814

*Authors contributed equally to this work.

Authors' addresses: Abhimitra Meka, Google, MPI Informatics, Saarland Informatics Campus; Rohit Pandey; Christian Häne; Sergio Orts-Escolano; Peter Barnum; Philip Davidson; Daniel Erickson; Yinda Zhang; Jonathan Taylor; Sofien Bouaziz; Chloe Legendre; Wan-Chun Ma; Ryan Overbeck; Thabo Beeler; Paul Debevec; Shahram Izadi, Google; Christian Theobalt, MPI Informatics, Saarland Informatics Campus; Christoph Rhemann; Sean Fanello, Google.

0730-0301/2020/12-ART259
https://doi.org/10.1145/3414685.3417814

1 INTRODUCTION

Capturing and rendering photorealistic human performances with controllable viewpoint and lighting is one of the most active research areas in the fields of computer graphics and vision. Such a technology has applications in augmented and virtual reality [Kelly et al. 2019; Lombardi et al. 2018; Orts-Escolano et al. 2016], movie production [Seymour 2020], and game development.

The problem can be split into two parts: 3D acquisition and rendering. Acquisition systems often focus on generating an accurate 3D volumetric model of a dynamic performer which can be rendered from any arbitrary viewpoint [Collet et al. 2015; Tanco and Hilton 2000]. In order to composite the model into novel environments with believable appearance, the model must be relit by the local lighting of the environment. Most previous systems build a parameterized mesh of restricted resolution and detail with a fixed lighting condition, and as a result, it can be difficult to change the lighting on the model realistically. Some other earlier works showed plausible template-based reconstruction of relightable 3D video [Li et al. 2013; Wu et al. 2013], but parametric reflectance models and the use of mesh templates still limited the re-rendering fidelity.

On the other hand, image-based relighting systems use additional data capture to solve the rendering problem. Rather than acquiring the full 3D shape of the subject, these methods capture 2D images under different illumination conditions to record a complete reflectance field of the performer [Debevec et al. 2000; Meka et al. 2019; Peers et al. 2006]. While this technology enables photorealistic renderings of humans, the lack of full body 3D geometry and the complexity of physically based rendering makes these approaches suitable only for very specialized applications, or requiring considerable post-processing and manual touch ups.

Capturing and rendering a large number of performers in a photorealistic manner with no manual intervention is still a very active research area. Recently, [Guo et al. 2019] built a fully distributed pipeline for volumetric performance capture that combines volumetric capture with relighting based on time-multiplexed color gradient illuminations. These alternating lighting conditions have been shown to enable the estimation of the material properties such as albedo, roughness, and surface normals [Fyffe et al. 2009].

Despite impressive results and the evident quality improvement over previous capture solutions, such a system still lacks photorealism. Although the *3D capture* pipeline uses custom high quality depth sensors [Guo et al. 2019; Kowdle et al. 2018], it cannot precisely reconstruct thin complex structures such as hair and clothing. The *reflectance estimation* proposed in [Guo et al. 2019] relies on a simplistic cosine lobe reflectance model [Fyffe et al. 2009] and therefore cannot accurately model higher-frequency view-dependent specular effects or the complex luster of cloth and skin.

While classical graphics pipelines have come a long way in modeling light-surface interactions for complex materials such as human skin and hair to render high-quality human images [Seymour 2020], they still lack the capacity to cross over the 'uncanny valley' and approach photorealism that is indistinguishable from the groundtruth. We refer the reader to [Pharr et al. 2016] for a comprehensive review on the open research topic.

Neural rendering [Eslami et al. 2018; Kim et al. 2018; Liu et al. 2019; Lombardi et al. 2019; Martin-Brualla et al. 2018; Sitzmann et al. 2018, 2019; Thies et al. 2019], has shown very promising results that overcome the shortcomings of traditional computer graphics by applying deep learning to learn disentangled representations of appearance and viewpoint. However, these systems usually assume a fixed lighting condition [Lombardi et al. 2019; Martin-Brualla et al. 2018; Thies et al. 2019], or they cannot generalize to unseen objects and subjects [Lombardi et al. 2019; Thies et al. 2019]. These issues limit their applicability in volumetric performance capture scenarios.

In this work, we present the first capture and rendering framework that learns representations of *appearance*, *viewpoint* and *lighting* of moving humans in arbitrary clothing (see Table 1). The proposed method relies on a high end Light Stage setup, (for detailed description of the hardware setup, please see the appendix), and combines traditional reflectance and geometry capture pipelines with a neural rendering approach to produce nearly photorealistic renderings of performers from any viewpoint and under any desired illumination condition.

Our method builds neural textures on the fly by extracting features from multi-view imagery. These features are pooled into a common texture-space parameterization (UV parameterization) obtained from a coarse geometry estimate. The pooled features encode both local and global geometric properties and 4D reflectance. The features are then reprojected to image space based on the desired viewpoint, and finally evaluated by a neural renderer along a desired lighting direction to correct the imperfections due to the coarse geometry and synthesize the final relit image, without any manual intervention.

In summary, our main contributions are:

- A volumetric capture framework that leverages neural rendering to synthesize photorealistic humans from arbitrary viewpoints under desired illumination conditions.
- An approach to build neural textures from multi-view images to render the full reflectance field for unseen dynamic performances of humans, including occlusion shadows and an alpha compositing mask. This overcomes the issues of previous works using neural textures [Thies et al. 2019] that need to be re-trained for every new UV parameterization.
- High quality results on free-viewpoint videos with dynamic performers, extensive evaluations and comparisons to show the efficacy of the method and substantial improvements over existing state-of-the-art systems.

Our framework presents a significant step towards bridging the gap between image-based rendering methods and volumetric videos, enabling exciting possibilities in mixed reality productions.

2 RELATED WORK

Synthesizing realistic relightable humans it is often tackled using image-based relighting techniques [Debevec et al. 2000; Einarsson et al. 2006; Wenger et al. 2005b], or Parametric models with priors [Barron and Malik 2015; Blanz and Vetter 1999; Garrido et al. 2013, 2016; Gotardo et al. 2018; Ichim et al. 2015; Meka et al. 2017; Pons-Moll et al. 2015; Theobalt et al. 2007; Thies et al. 2016; Wen et al.

Table 1. Machine learning methods achieve a high degree of photorealism, whereas traditional capture pipelines [Guo et al. 2019] are better at generalization, rendering capabilities, and they can capture moving performers. Our algorithm brings together the capabilities of both state-of-the-art approaches.

	Free Viewpoint	Relightable	Moving Performers	Higher-order Appearance Model	Generalization
Guo et al. [2019]	Yes	Yes	Yes	No	Yes
Thies et al. [2019]	Yes	No	No	Yes	No
Martin-Brualla et al. [2018]	Yes	No	Yes	Yes	Yes
Wenger et al. [2005a]	No	Yes	Yes (with high FPS)	Yes	Yes
Meka et al. [2019]	No	Yes	Yes	Yes	Yes
Xu et al. [2019]	Yes (half hemisphere)	No	Not Demonstrated	Yes	Yes
Hedman et al. [2018]	Yes	No	Not Demonstrated	Yes	Yes
Proposed	Yes	Yes	Yes	Yes	Yes

2003]. More recently, related work relies on sophisticated multi-view 3D performance capture systems [Beeler et al. 2011; Cagniart et al. 2010; Collet et al. 2015; de Aguiar et al. 2008; Dou et al. 2017; Guo et al. 2019; Tanco and Hilton 2000; Vlasic et al. 2008], and neural rendering techniques [Dosovitskiy et al. 2015; Eslami et al. 2018; Kulkarni et al. 2015; Zhu et al. 2014] to increase photorealism of the final renderings.

Multi-view 3D Performance Capture. These systems explicitly estimate the deforming geometry of the performer using multi-view setups [Beeler et al. 2011; Cagniart et al. 2010; Collet et al. 2015; de Aguiar et al. 2008; Dou et al. 2017; Guo et al. 2019; Tanco and Hilton 2000; Vlasic et al. 2008]. These sophisticated pipelines may focus on face performance capture [Beeler et al. 2010, 2011; Ma et al. 2007] or full body reconstructions [Cagniart et al. 2010; Collet et al. 2015; de Aguiar et al. 2008; Guo et al. 2019; Tanco and Hilton 2000; Vlasic et al. 2008].

For instance the works of Beeler et al. [2010, 2011] have shown impressive *facial performance results*, but they do not explicitly estimate reflectance information required for photorealistic rendering.

Full body performance capture, also known as free viewpoint videos or volumetric videos, are very popular since the works of Carranza et al. [2003]; Tanco and Hilton [2000]. More recently, the Microsoft volumetric capture system [Collet et al. 2015] has been used in many commercial productions for mixed reality including through licensees such as Metastage.

Recent advances in high speed depth sensing [Fanello et al. 2016, 2017; Tankovich et al. 2018] have enabled *real-time* performance capture [Dou et al. 2017, 2016], showing compelling applications such as virtual telepresence [Orts-Escolano et al. 2016]. These methods [Collet et al. 2015; Dou et al. 2017] usually rely on sparse correspondences [Innmann et al. 2016; Wang et al. 2016; Zaharescu et al. 2009] to guide a non-rigid alignment method between the reconstructed meshes, providing a temporally consistent reconstruction over time.

Despite all these efforts, these systems lack photorealism due to missing high frequency details [Orts-Escolano et al. 2016] and baked-in diffuse texture [Collet et al. 2015], which does not allow for accurate and convincing re-lighting of these models in arbitrary scenes.

Guo et al. [2019] have made an attempt to overcome most of these issues. Their sophisticated pipeline combines the image based rendering technique proposed in [Fyffe et al. 2009], with custom high resolution depth sensors to estimate geometry and material properties such as albedo, roughness and normals. This achieves unprecedented quality and photorealism for free viewpoint videos.

Although this recent work is a substantial improvement over previous approaches, the method of Guo et al. [2019] does not achieve true photorealism due to the assumptions made in their reflectance maps which rely on a simple cosine lobe model.

Nevertheless, these multi-view capture systems offer a foundation for machine learning approaches and in this paper we show how to leverage them to acquire ground truth data and train a deep learning based solution for volumetric performance capture.

Neural Rendering. An orthogonal perspective is given by the recent advances in the machine learning community in the area of *neural rendering* [Dosovitskiy et al. 2015; Eslami et al. 2018; Kulkarni et al. 2015; Zhu et al. 2014], see Tewari et al. [2020] for a comprehensive review.

These approaches aim at using deep learning to control specific parameters of a scene such as: lighting, geometry, camera view, pose/layout, etc.

For instance, in the context of view synthesis, the work of Hedman et al. [2018] learns to blend multiple RGB images to capture view-dependent effects on static scenes. Similarly, Xu et al. [2019] proposes an architecture to perform view synthesis of an object using photometric images. When combined with Xu et al. [2018], the method can also enable relighting capabilities, however its applicability is limited to a *half hemisphere* around the captured object with a *fixed* distance from the camera.

In a similar direction, the work of Philip et al. [2019] shows compelling relighting results on static scenes, correctly synthesizing shadowing effects. The very recent work of Mildenhall et al. [2020] achieves highly realistic view synthesis, however, similar to Xu et al. [2019], the model does not enable relightability.

Other related works are focusing on synthesis of humans [Balakrishnan et al. 2018; Chan et al. 2019; Kim et al. 2019, 2018; Liu et al. 2019; Ma et al. 2017, 2018; Neverova et al. 2018; Si et al. 2018; Thies et al. 2019; Zhao et al. 2017], with particular emphasis on performance capture [Lombardi et al. 2019; Martin-Brualla et al. 2018; Pandey et al. 2019; Shysheya et al. 2019] and relighting [Meka et al. 2019; Sun et al. 2019; Zhou et al. 2019].

The *LookinGood* system by Martin-Brualla et al. [2018] introduces the concept of neural rerendering for performance capture of human

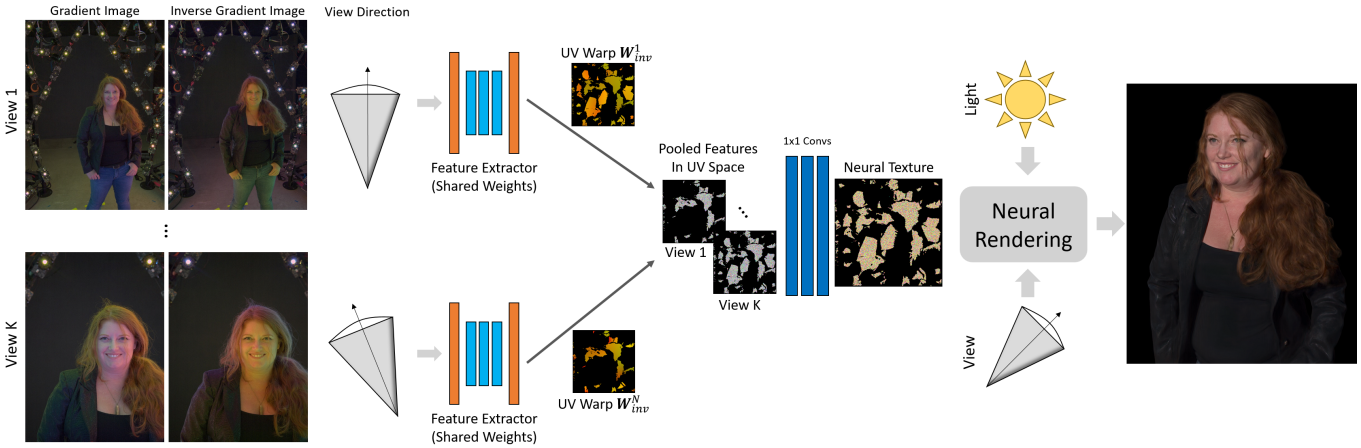

Fig. 2. **Proposed Pipeline.** We propose a Neural Rendering pipeline for rendering of moving humans in any desired viewpoint and lighting. Features are extracted from the raw images and then pooled in UV space with a learned blending function. 1×1 convolutions in texture space allow for generalization to different parameterizations. A final Neural Rendering synthesizes the image in camera space. See text for details.

actors. The authors leverage a real-time volumetric capture system [Dou et al. 2017] augmented with "witness cameras" that provide groundtruth unseen views that are used by a deep convolutional neural network to re-render the final image. The method, however does not control the lighting condition, therefore does not allow relighting the subject.

Follow up work by Lombardi et al. [2019], proposes to use *Neural Volumes*, which are able to render a performer from any arbitrary view without relying on a pre-captured 3D mesh such as in Martin-Brualla et al. [2018]. The main drawback is the lack of a disentangled appearance representation: indeed the volume is trained per-subject with fixed illumination and it does not generalize to different performers or arbitrary lighting conditions.

Orthogonal trends such as the work of Pandey et al. [2019], try to reduce the infrastructure requirements (e.g. multiple cameras at test time) and they leverage a semi-parametric model to synthesize humans in arbitrary poses and viewpoints from a single RGBD image. Similarly to Martin-Brualla et al. [2018] and Lombardi et al. [2018], the method has a fixed lighting condition.

The work of Meka et al. [2019] uses two spherical gradient illumination images similar to [Guo et al. 2019] and generates compelling renderings under any desired illumination. However the approach does not allow for free viewpoint rendering and cannot model self occlusions due to the lack of geometry. Similar to Meka et al. [2019], the methods proposed by Sun et al. [2019] and Zhou et al. [2019] aim at controlling the lighting condition of portrait images, without changing the viewpoint.

Proposed Approach. Our method achieves nearly photorealistic renderings of *dynamic performers* from *arbitrary viewpoints*, with any *desired illumination* condition. Moreover, our approach is *scalable* and does not require manual intervention.

In contrast to related works [Chen et al. 2019; Thies et al. 2019; Zhang et al. 2020], we propose a framework that combines geometric pipelines and neural rendering that enables *simultaneous* disentanglement of appearance, viewpoint and lighting. In Table 1 we summarize the main capabilities of state-of-the-art methods.

3 NEURAL RENDERING FOR PERFORMANCE CAPTURE

Our goal is to generate photorealistic renderings of humans in motion under arbitrary viewpoint and lighting conditions. Moreover we wish to generalize to dynamic performers, enabling renderings at scale with no manual post-processing used in movie products.

Based on these goals, our method leverages deep learning to address the following substantial drawbacks of the traditional geometric pipelines (see appendix for a more detailed discussion on the drawbacks.):

(1) **Inadequate Geometric Model.** Meshes or 3D voxels of any reasonable density are not expressive enough to capture fine grained details such as hair.
(2) **3D Acquisition Errors.** Even if a mesh could accurately model the geometry, the reconstruction can be inaccurate due to erroneous calibration or approximations in the many stages of a reconstruction pipeline.
(3) **Approximate Reflectance Model.** Typical BRDFs (e.g. the Phong [Li et al. 2013] or cosine lobe models used in previous work [Guo et al. 2019]) are not expressive enough to take into account the complex image formation process that would lead to a photo-realistic rendering of a human.
(4) **Approximate Rendering.** Even when the BRDF model suffices, many assumptions/approximations ignore high order light transport effects such as sub-surface scattering and global illumination, leading to unrealistic renderings.

In order to overcome these difficulties, we employ a neural architecture (Fig. 2) that extracts features from each of the multi-view images and pools them into texture space (UV space) based on a pre-acquired coarse geometry estimate. The pooled features are further transformed using 1x1 convolutions to extract implicit reflectance and local geometry information, which are then reprojected into the image space of a novel desired viewpoint. The reprojected features, in combination with classical graphics buffers such as light visiblity maps and reflection maps, are provided as input to a neural renderer (Fig. 3) to generate the final output image of the subject lit

under a desired lighting direction. By sampling the lighting direction over the unit sphere, the neural renderer can generate a set of images that form the full reflectance basis for the frame. This basis can be used to relight the image under arbitrary lighting environments. The neural rendering replaces the use of an explicit BRDF and allows for modeling higher-order light transport effects directly from the training data. Further, it removes the strict dependency on accurate geometry as the rendering network can compensate for inaccuracies (e.g. filling in missing hair cf. Fig. 12) as also shown by Martin-Brualla et al. [2018].

In contrast to previous work such as Thies et al. [2019], which learns a fixed neural texture, we first extract features from images and then pool them in texture space using the pre-computed warp fields that remap the images to UV space. As a consequence, the neural textures can be *regressed* from input images, rather than optimizing them through back-propagation such as [Chen et al. 2019; Shysheya et al. 2019; Thies et al. 2019], which limits generalization.

Crucially, the extracted features have a certain spatial extent thanks to the receptive fields of the feature extraction network. This implies that in texture space we can resort to simple 1×1 convolutions which do not depend on the UV arrangement. The use of 1×1 convolutions is also justified by geometric capture systems (e.g. Guo et al. [2019]), which obtain reflectance maps with simple *per-pixel* operations in RGB space. We argue that learned 1×1 operators on feature vectors are superior to hand-crafted per-pixel operations in RGB space. These observations are essential to disentangling appearance: indeed at test-time a new neural texture can be built from a set of multi-view images and an approximate parameterized geometry (i.e. pre-computed warps from image space to UV parameterization). As a byproduct, we can generalize to unseen performances and we do not need to re-train the network even if the UV parameterization changes. On the contrary, related works relying on neural textures [Chen et al. 2019; Shysheya et al. 2019; Thies et al. 2019] need to re-train every time the UV parameterization changes (e.g. dynamic sequences), even for a fixed subject.

The proposed pipeline allows for simultaneous synthesis of appearance, viewpoint and lighting of dynamic performances: to the best of our knowledge this is the first neural rendering system with this capability. In the following we detail each of these steps.

3.1 Inputs and Feature Extraction

Our neural rendering pipeline assumes the availability of an approximate geometry of the subject for every frame of the performance. We use a hardware setup similar to that of Guo et al. [2019] to obtain such geometry, as explained in Section 4.1. This geometry estimate is used to generate a UV map of the surface along with warp fields that map multi-view images into the texture space and vice verse. Note that it is generally very challenging to achieve a temporally coherent UV parameterization for a non-rigidly deforming geometry as is the case in dynamic performances. Our method assumes no such temporal correspondences for even consecutive frames and is in fact designed to be completely robust to arbitrary texture space changes and hence provides generalization of appearance synthesis across subject pose and identity.

To capture the input 2D images, our method leverages a Light Stage, a studio device containing a capture volume inside a spherical dome fitted with calibrated RGB lights and multi-view cameras. Previous works that employ the Light Stage [Fyffe et al. 2009; Guo et al. 2019; Meka et al. 2019] have shown that spherical gradient illumination conditions can be used to extract information regarding surface normals, albedo and roughness. Deep learning methods have been successfully applied to these inputs to obtain convincing relighting results in image space [Meka et al. 2019].

Following this trend, our system takes as input two images captured under spherical gradient illumination conditions from N camera viewpoints, where each image has 2000×1500 pixels. These complementary lighting conditions are aligned using 2D optical flow such as in Meka et al. [2019]. Additionally, we concatenate to each pixel a view direction vector, i.e. the ray going from the optical center to the center of the pixel in world space, resulting in a 3D unit vector that can be encoded in 2 channels. The view direction provides the network with some guidance regarding the view-dependent effects on a given image.

A U-Net architecture [Ronneberger et al. 2015] is used to extract features from each viewpoint. The architecture takes as input 8 channels: 6 for the two gradient images, 2 for the view direction. The specific network has 5 encoder/decoder layers with 16, 32, 64, 128, 256 filters, extracted with 3×3 convolutions followed by blur pool [Zhang 2019] in the encoder and blur unpool in the decoder. A final output layer infers a tensor of 16 channels with 2000×1500 resolution.

Crucially, this U-Net extracts features with receptive fields with a reasonable spatial extent (478×478). Additionally, the final output has the same resolution of the input images (2000×1500), to preserve all the high frequency details. The feature extraction is carried out for each view and a total of N feature tensors with 16 channels are generated. See Fig. 2 (left) for an overview of the feature extractor.

3.2 Learn to Regress the Texture Space

At this stage we have one tensor $\mathbf{F}$ with 16 channels and 2000×1500 for each camera view. Assuming that a 3D geometry with parameterization is available, we can compute warp fields that map each pixel from image space to the UV texture space.

The warp fields are pre-computed using the 3D geometry to map between texture UV coordinates and camera image coordinates with explicit occlusion handling via ray casting. We generate a 2000×1500 warp field $\mathbf{W}^k(x, y) = (u, v)$ as a 2-channel map from each pixel of camera k to UV coordinates of parameterization – in essence, the rasterization of raw UV coordinates on our geometry for camera k. For the inverse mapping $\mathbf{W}^k_{inv}(u, v) = (x, y)$ we construct a 1000×1000 warp field matching our UV texture dimension, where the 2-channel value at each UV texel is the visibility-tested projection from the parameterized geometry into the image coordinates of camera k. These warp fields can be used in an end-to-end framework in a fully differentiable manner as shown in Jaderberg et al. [2015].

The warped feature tensors $\mathbf{F}^1_w, \ldots, \mathbf{F}^N_w$ are pooled together into a single tensor which removes the dependency on the order of the input images. To do so we perform a weighted sum of the features, where the weights are computed using the dot product between the

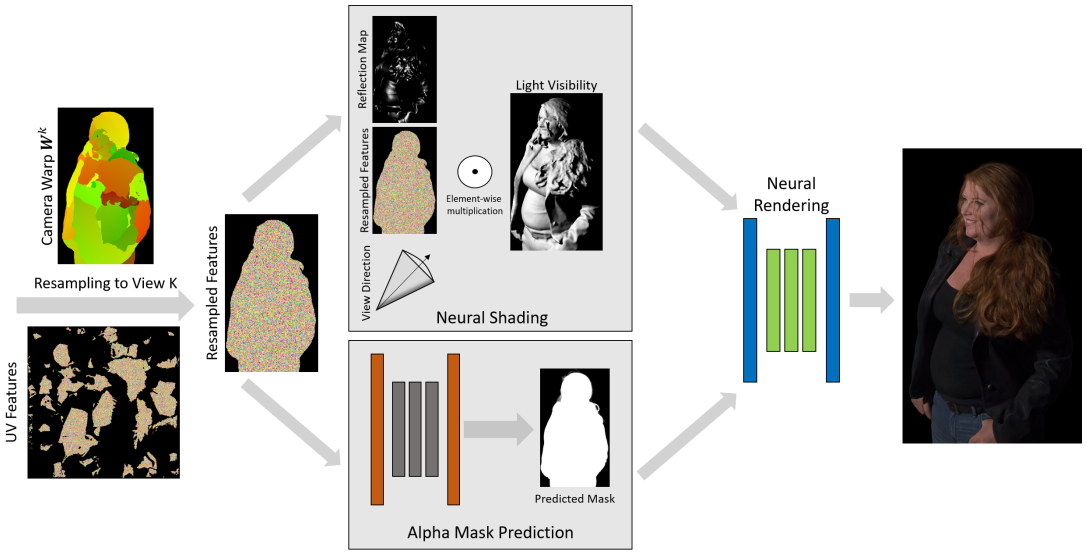

Fig. 3. **Neural Rendering** – The learned texture space is resampled given the desired rendering camera. A Neural Shading module adds lighting information to the resampled features; an alpha matting prediction module predicts an alpha mask. Finally, a U-Net performs the actual rendering.

camera viewing direction and the surface normals. This is inspired by traditional volumetric capture pipelines [Collet et al. 2015; Guo et al. 2019] that utilize a similar weighted scheme to stitch together multiple views in the UV space. This generates a texture space tensor of $1000 \times 1000 \times 16$.

Thanks to this high dimensional feature vector we can rely on a few 1×1 convolutions followed by non-linearities in texture space, which allows for generalization for different parameterizations. In particular we perform three 1×1 convolutions followed by ReLU activations to obtain a final texture space tensor with 16 channels.

3.3 Neural Rendering

The final part of the pipeline is a Neural Renderer module and it is depicted in Fig. 3. This component takes as input a target camera view, which is used to generate the warp $\mathbf{W}^k$. This warp is employed to resample features from the texture space to the image space. The Neural Renderer module consists of two branches: a *Neural Shading* and an *Alpha Matting* network. The output of these two modules are then passed through a final U-Net that generates the actual rendered images.

Neural Shading. The resampled features do not contain information regarding the desired viewpoint or light condition, but mostly encode surface and material properties. To explicitly encourage the network to learn the shading function, we borrow components from computer graphics rendering such as the light visibility map and reflection map, and cast them in a neural network framework.

The light visibility map is computed per-pixel via the dot product between the surface normal $\mathbf{n}$ and the target lighting direction $\mathbf{l}$. We also handle occlusions explicitly via ray casting, which result in black pixels in the map (see Fig. 3, Neural Shading module).

The reflection map is inspired by traditional Phong shading, and defined as $(\mathbf{r} \cdot \mathbf{v})^{\alpha}$, where $\mathbf{v}$ is the view direction of the target camera and $\mathbf{r} = 2(\mathbf{l} \cdot \mathbf{n})\mathbf{n} - \mathbf{l}$. Such a reflection map has been shown by Meka et al. [2018] to guide the network towards specularities and view dependent effects. While Meka et al. [2018] attempts to estimate such a 'mirror-like' reflection map using a neural network to augment the reflectance estimation pipeline, we feed this map as an input to the neural renderer to aid the specularity synthesis.

The resampled neural features, the reflection map and the view direction, encoded per-pixel in 2 channels, are concatenated into a single tensor $\mathbf{S}$ of dimensions 2000×1500 with 19 channels (16 for the features, 1 for reflection map, 2 for view direction) and multiplied element-wise with the light visibility map, simulating a *neural diffuse rendering*.

Alpha Matting. The second branch of the neural renderer consists of a small U-Net with skip connections that takes as input the resampled features of size 2000×1500 and comprehends 6 fully convolutional layers for encoder and decoder with 3×3 filters, with outputs 8, 16, 32, 64, 128, 256. The output of this network is an alpha mask: we will show the importance of the alpha mask for the application of compositing in virtual environments.

Finally, the output of the Neural Shading and Alpha Matting network are concatenated and passed to the final U-Net to perform the final rendering. This rendering architecture takes as input a tensor of 20 channels (19 for the neural shader, 1 for the alpha mask) of size 2000×1500 and it passes it through 5 levels for the encoder and 5 for the decoder. We use 3×3 convolutions with outputs 64, 128, 256, 512, 1024. Additionally, we employ skip connections between the encoder and decoder, except for the last layer, which generates the final RGB image.

At test time, given multi-view images of a performer, we need to build the neural texture only once and use the neural renderer module to synthesize any novel illumination condition from any desired viewpoint.

Fig. 4. A subject illuminated under both color gradient and inverse color gradient illumination. Intuitively, a red observation corresponds to either red albedo or a surface normal pointing towards the red lights. As the sweater is largely red under both illumination conditions, the network receives a strong cue that the albedo of the sweater is red. In contrast, the subject's right forehead appears red under only one illumination condition, providing a strong constraint on surface orientation.

Fig. 5. Pseudo-groundtruth alpha masks obtained with a variant of Sengupta et al. [2020]. These masks are employed at training time to predict an alpha mask that is then used for compositing. Despite the small imperfections, we found them very effective for this task.

4 TRAINING DETAILS

In this section we describe the acquisition process and how to train the proposed framework in an end-to-end fashion.

4.1 Groundtruth Acquisition

The described pipeline relies on: multiview imagery of performers acquired with two spherical gradient illumination conditions, the knowledge of a 3D parameterized geometry, and groundtruth images under a specific illumination condition and alpha masks. In order to acquire this data, we rely on a Light Stage [Debevec 2012], a custom spherical dome with 331 fully programmable LEDs.

Multi-view Imagery. Similarly to Guo et al. [2019], we use 58 high resolution RGB cameras to record video at 60 hertz with 12.4 megapixel resolution. As previously described, we interleave two different visible lighting conditions based on spherical gradient illumination [Fyffe et al. 2009]. A spherical gradient image is obtained by programming the LEDs to emit a color that changes with respect to its position in the Light Stage. In particular, given the lighting direction vector θ of a LED relative to the center of the stage, the light emitted by that LED for the first gradient image is programmed to have the RGB color $((1+\theta_x)/2, (1+\theta_y)/2, (1+\theta_z)/2)$, and the second gradient image is programmed to have the RGB color $((1-\theta_x)/2, (1-\theta_y)/2, (1-\theta_z)/2)$. An example of spherical gradient images can be seen in Fig. 4.

Geometry Acquisition. To acquire the base geometry needed to pre-compute the warp fields, we additionally use 32 high resolution infrared (IR) cameras. These are coupled with 16 custom structured light projectors such that they can be used for active stereo depth estimation. A multi-view stereo algorithm followed by a Poisson reconstruction step and a parameterization phase are used to retrieve the final geometry. For details see Guo et al. [2019] and Collet et al. [2015].

Given the base geometry, we can also compute the *light visibility maps* and the *reflection maps* described above, that are passed to our neural renderer.

Groundtruth Images. Target images are acquired by collecting full reflectance fields. In particular, we capture a sequence of so called One-Light-At-a-Time (OLAT) images. In each OLAT image, only one of the 331 LEDs is turned on and this has a known light direction pointing from the center of the LightStage to the LED position. A single sequence consists of 331 OLATs for 58 high resolution RGB cameras and 32 active IR sensors. Due to the large amount of data, we are limited to a framerate of 60 hz. In other words we need ~ 6 seconds per sequence acquisition, during which time the subject may move a little, causing misalignments in the training data. To overcome these issues, we follow [Meka et al. 2019] and introduce additional "tracking frames" with all the LEDs turned on. These fully lit images are acquired after every 10th OLAT and they are used to perform an optical flow alignment in image space for each view with respect to a selected keyframe [Anderson et al. 2016]. The optical flow is then interpolated to align OLAT images between two tracking frames. Finally, we capture the spherical gradient illumination conditions that are used as input to the system.

Note that since all the 2D imagery is aligned to a given reference frame, we do not need to compute the geometry for all the OLATs but we simply rely on the parameterized mesh computed for a selected keyframe.

The alpha masks used during training are obtained with a variant of Sengupta et al. [2020] trained on Light Stage data [Debevec et al. 2002], see Figure 5. Despite not being perfect, we found these masks very effective for this task.

4.2 Loss Functions

We train the feature extractor (Section 3.1), neural texture (Section 3.2), and neural rendering (Section 3.3) components of our pipeline end-to-end, using a combination of multiple losses. In particular, our loss function is defined by four components.

Photometric Loss in Feature Space: $L_{\text{VGG}}(I, \hat{I})$. Similarly to Martin-Brualla et al. [2018], we use the squared ℓ_2 distance between features extracted from the target image I and the predicted image $\hat{I}$ using a VGG network pre-trained on the ImageNet classification task [Zhang et al. 2018]. As demonstrated by multiple previous works [Martin-Brualla et al. 2018; Meka et al. 2019; Pandey et al. 2019] this loss leads to sharper results compared to a traditional ℓ_1 distance in image space.

Alpha Loss: $\ell_1(M, \hat{M})$. In order to infer the alpha mask, we simply compute an ℓ_1 norm between the groundtruth mask M and the inferred mask $\hat{M}$.

Reflection Saliency Loss: $L_{\text{VGG}}(S, \hat{S})$. To encourage the network to learn specular highlights and view dependent effects, we also propose an additional reflection loss. We define $S = R \odot I$, where R is the reflection map defined in Section 3.3 and $\odot$ indicates element-wise multiplication. Similarly we define $\hat{S} = R \odot \hat{I}$ for the predicted image $\hat{I}$. The reflection loss is then computed as ℓ_2 distance of S and $\hat{S}$ in feature space using the VGG network. In the ablation study we demonstrate how this loss is vital to recovering view dependent effects. In Figure 6 we show visual examples of these terms.

Texture Loss: $L_{\text{VGG}}(I, N)$. Similarly to Thies et al. [2019], we add a loss between the target image I and the first 3 channels of the resampled neural texture N. This forces the network to represent part of its texture space as an actual RGB image, which is again inspired by computer graphics pipelines.

Our total loss is finally defined as:

$$L_{\text{total}} = w_1 L_{\text{VGG}}(I, \hat{I}) + w_2 \ell_1(M, \hat{M}) + w_3 L_{\text{VGG}}(S, \hat{S}) + w_4 L_{\text{VGG}}(I, N), \tag{1}$$

where w_i are used to control the contribution of the individual loss functions to the total loss. For our experiments we use $w_1 = 1.0$, $w_2 = 0.25$, $w_3 = 0.5$, and $w_4 = 1.0$.

4.3 Implementation Details

We implemented our training pipeline in TensorFlow where we distribute the training across 8 NVIDIA Tesla V100 GPUs. At each iteration we randomly pick a target OLAT per GPU. The feature extractor module in Sec. 3.1 is run on each camera independently. Since our system consists of 58 RGB cameras, this requires a substantial amount of computation and memory. Indeed during training, the system will need to extract and pool features multiple times, for potentially millions of iterations. In order to speed up this phase we consider only a neighborhood of 3 cameras around the target OLAT, which are computed to cover all the target pixels in UV texture space. Note that these features are average-pooled in UV space, making this operation independent of the order and the number of cameras.

We use the ADAM optimizer [Kingma and Ba 2014] with a learning rate of 10^{-4}, employing an exponential decay of the learning rate of 0.1 every $100k$ iterations. We optimize our network for $500k$ iterations before the training converges, which usually takes 2-3 days.

At test-time, we pre-compute the learned texture by extracting features from all the cameras. This allows us to only run the neural render to infer the desired viewpoint and lighting conditions. The evaluation of all the OLAT's for a given frame is performed in a fully parallelized fashion on the cloud using Intel 2.6 Ghz processors with 32 cores.

5 EVALUATION

In this section we provide an exhaustive evaluation of the proposed approach. We specifically focus our experiments as to elucidate our approach's ability to disentangle lighting, viewpoint and appearance.

Note that in many of the experiments we aim to synthesize an image from a desired viewpoint under a single directional light source. This scenario actually emphasizes complex light transport effects such as specular highlights and subsurface scattering, which are crucial to achieve true photorealism. Additionally, since lighting is additive, a simple linear combination of these directional light sources can be formed to generate arbitrary relighting provided, for instance, by an HDRI map [Debevec et al. 2000]. The results obtained using HDRI relighting are shown and discussed in Section 6.

Dataset Acquisition. For this work we acquired data from 70 participants with different skin color and clothing. Moreover, we asked participants to perform specific poses, in order to present the network with sufficient variability. Following ML fairness practices, we ensured that our dataset was as diverse as possible. For each subject we acquired 331 OLATs, 2 gradient images from 58 RGB images and 32 IR images. Each user provides $\sim 30,000$ images with 2000×1500 pixels that are used as training examples. A few representative examples of the acquired groundtruth data can be seen in Figure 7. We hold out a few subjects from the dataset to show the generalization capabilities of the method. All the results are shown under conditions that were not presented to the network during training.

Notably, each subject is reconstructed independently and therefore they do not share the same mesh topology and parameterization. Nevertheless, we prove that our approach is able to generalize to unseen performers and changing parameterizations.

Finally, we automatically detect and discard OLAT images that contain strong lens flares such as the ones shown in Figure 8. In order to find those cases, we check if the projected 3D position of a light is in the field of view of the camera or if the angle between the camera viewpoint and the light is greater than 130°. We found this simple strategy to be very effective and it is able remove most of the images that could corrupt the groundtruth data, at the same time the learning scheme is able to generalize well to those missing training examples.

Metrics. In order to analyze the quantitative performance of the system we report the following measurements: PSNR, MultiScale-SSIM, Photometric error, i.e. ℓ_1-loss, and Perceptual loss [Zhang

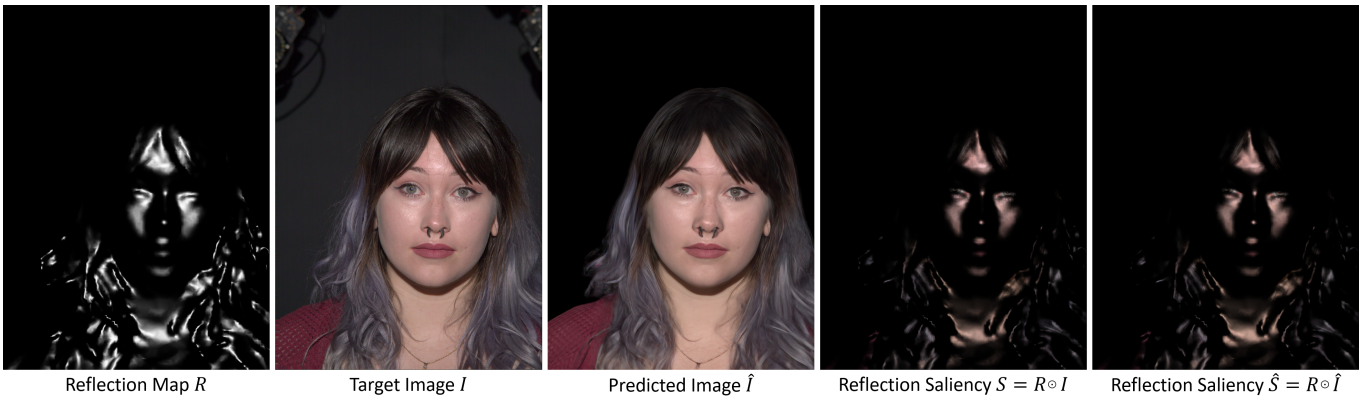

Fig. 6. Visualization of the reflection saliency used in the loss function. The contribution of this component is crucial to recover specular highlights and view dependent effects.

Fig. 7. Our dataset includes 70 subjects performing a wide range of poses. Our capture setup allows us to capture images simultaneously from multiple viewpoints under 331 One-Light-At-a-Time lighting conditions.

et al. 2018]. The perceptual loss is calculated using an ℓ_1 distance in feature space. Features are extracted using a pre-trained VGG architecture: in particular only the two first convolutional layers are used.

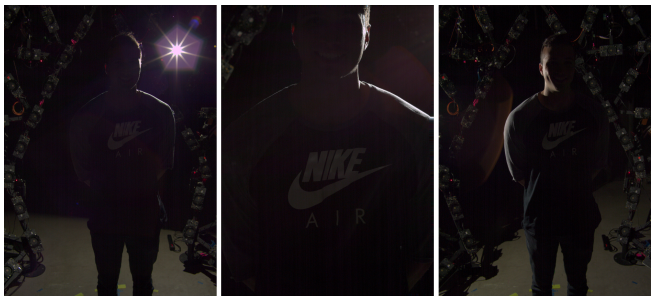

Fig. 8. Examples of automatically discarded OLAT images that could corrupt the training data due to lens flares.

5.1 Qualitative Results

We here show multiple test results generated with our system and analyze the various capabilities.

Rendering Under Novel Illumination. We first show that the system is able to synthesize any desired directional lighting condition. Results are shown in Figure 9, top row. Please note the recovered high frequency details such as specularities on the forehead and view dependent effects. In Section 6 we show how to linearly combine multiple directional light sources to perform HDRI relighting.

Rendering Novel Viewpoints. In Figure 9, middle row, we demonstrate the capabilities of our system by rendering a subject from completely novel viewpoints. As can be seen, the images are compelling throughout the multiple rendered viewpoints. In the supplementary video we show multiple fly through examples where the method handles changing direction as well as scaling.

Rendering Diverse Appearances. In Figure 9, bottom row we show additional results of generated lighting and viewpoints on a diverse set of multiple performers. Note how we are able to render complex hair shapes (third and sixth examples), general apparel (first example), and self occlusions (fifth column).

Fig. 9. Examples of synthesized results. Top: a fixed subject and viewpoint with changing lighting direction. Middle: a fixed subject and illumination with changing viewpoint. Bottom: rendered subjects with different viewpoints and illumination.

Simultaneous Light, View and Appearance Rendering. In Figure 10, we demonstrate our systems ability to relight *moving* subjects under new lighting conditions and any desired viewpoint. Additionally, thanks to the use of 1×1 convolutions in UV space, the system handles different UV parameterizations changing appearance over time.

In contrast to other work [Thies et al. 2019] that would need to train a neural texture for every new UV parameterization, our framework directly infers a neural texture at test time. We observe that such generated temporal sequences show a remarkable amount of temporal stability and a high degree of photorealism. More examples are shown in the supplementary video.

Alpha Mask Prediction. As by product, our method can predict an alpha mask which is key for convincing compositing of the performers in any desired environment. In Figure 11 we show the predicted mask for selected viewpoints and subjects. The inferred matte captures a fair amount of details, however it is still far from the groundtruth quality. This is somehow expected since the mask prediction network needs to handle any desired viewpoint, differently from the groundtruth mattes that can be only acquired from fixed camera positions. Please see more results in the supplementary video, where temporal stability can also be appreciated. For practical applications such as relighting and compositing (Section 6) we found the alpha prediction to be fairly effective.

5.2 Comparisons with State-of-the-Art

We here compare the method with the state-of-the-art. In particular we selected traditional computer graphics approaches ([Guo et al. 2019; Wenger et al. 2005b]) as well as recent neural rendering-based methods [Martin-Brualla et al. 2018; Meka et al. 2019; Thies et al. 2019]. Note that many of these methods can only perform one single task at a time such as view synthesis or light interpolation and may not be applicable to dynamic sequences. As discussed in Table 1, our method uniquely supports all of the capabilities of these methods simultaneously.

Rendering Novel Viewpoints. In Figure 12 we compare our method to other state-of-the-art methods for novel view synthesis. Notice how our results are comparable and often superior to the state-of-the-art method of Thies et al. [2019], which requires to be trained per object, or every time the UV parameterization changes. Our method instead is able to support dynamic sequences with changing UV parameterization and relighting (Figure 10), all capabilities that Thies et al. [2019] cannot achieve.

Additionally, despite being trained on images containing the masked foreground, Thies et al. [2019] does not predict an explicit

Fig. 10. Simultaneous Illumination, Viewpoint and Appearance Synthesis of Dynamic Performers. Thanks to the use of 1×1 convolutions in UV space, our method handles different UV parameterizations depicted in the top right corner of each frame.

background or alpha mask, hence the method cannot be used directly for compositing applications and often leads to inconsistent background over time as shown in the supplementary video.

Compared to Martin-Brualla et al. [2018], our method better generalizes to free-viewpoint camera trajectories, keeps high frequency

Fig. 11. Alpha prediction results. The predicted mattes are key for convincing compositing effects from any desired viewpoint.

details sharper and view dependent effects follow the target viewpoint more naturally (see supplementary video). This is somehow expected as Martin-Brualla et al. [2018] method is a particular case of the framework we propose: indeed in Martin-Brualla et al. [2018], authors use a neural re-renderer on images rendered by a computer graphics pipeline, whereas our approach generates neural features directly from input images.

In comparison with state-of-the-art volumetric capture pipelines [Guo et al. 2019], our approach leads to more photorealistic results since it does not make any approximation on the BRDF model and learns the rendering function directly from the data. Additionally, like other neural rendering methods, our approach is more robust to geometry imperfections.

Rendering Under Novel Illumination. In Figure 13 we compare our method to other state-of-the-art algorithms for relighting. In this case, our results outperform the competitors, while maintaining the additional capabilities outlined in Table 1.

Indeed, methods like Guo et al. [2019] and Wenger et al. [2005b] rely on standard computer graphics rendering techniques with approximated BRDF models. These systems cannot capture complex light transport effects, limiting their realism. For instance Guo et al. [2019] relies on a cosine lobe model [Fyffe et al. 2009], where the shininess factor does not capture specularities and view dependent effects. Whereas HDRI lighting results shown in the original paper are impressive, these maps are usually well approximated by low-frequency lighting conditions where specularities play a marginal role. On the contrary, when we use a single directional light instead (e.g. sunlight), like in our comparisons, these high order, light transport effects become more evident and crucial to achieve photorealism.

On the other hand, machine learning based methods can learn these complex light transport interactions directly from the data.

Fig. 12. Comparisons with state-of-the-art methods on the view synthesis task. Note how our framework is on par and often better than other neural rendering methods [Martin-Brualla et al. 2018; Thies et al. 2019] and computer graphics pipelines [Guo et al. 2019].

The method of Meka et al. [2019] was retrained on our dataset and, as shown in Figure 13, it is able to generalize well to full body captures, despite being designed for facial performances. However, since it is an image based method, it cannot handle self-occlusions accurately as demonstrated in Figure 13, first column. Moreover the approach is limited to *fixed* viewpoints and it cannot synthesize novel views.

Thanks to the underlying geometry, the proposed approach is designed to handle full body performance capture, relightability from multiple viewpoints and complex light transport effects that are learned directly from the data.

Fig. 13. Above we demonstrate how our results are comparable or superior to state-of-the-art neural rendering methods such as Meka et al. [2019] that can control the light from a *fixed* view during a performance capture. The underlying geometry allows for better occlusion handling compared to Meka et al. [2019]. In comparisons with computer graphics pipelines [Guo et al. 2019; Wenger et al. 2005a], our approach achieves more photorealistic results.

5.3 Quantitative Results

To complete our analysis, we also perform quantitative evaluations on test subjects, reporting the metrics described in Section 5.

Table 2 shows a quantitative analysis for the proposed method and compares it with state-of-the-art approaches. We conducted the evaluation on two different tasks: novel view synthesis and relighting. Our method is able to simultaneously perform both tasks. The results are given in Table 2. Note that the only method that is able to perform simultaneous view synthesis and relighting is the approach of Guo et al. [2019]: the proposed framework is able to outperform it substantially on both tasks on all absolute image error metrics.

The top part of Table 2 shows how our method performs comparable to other related works for novel view synthesis, achieving similar results in terms of photometric, PSNR, MS-SSIM and also perceptual dissimilarity. While other neural rendering methods such as Martin-Brualla et al. [2018] perform slightly better than our approach, note however that the method performs poorly for arbitrary camera trajectories, exhibiting strong temporal artifacts that are not captured by these metrics (see supplementary video).

Regarding the relighting task, the bottom part of Table 2 shows how our method performs very similarly and marginally better than the relighting method of Meka et al. [2019] and outperforms the others significantly. Please note that our method can also perform novel-view synthesis, unlike the competing method of Meka et al. [2019].

In practice, it is important to note that these metrics only capture a holistic view of the rendered images and they do not really provide useful insights on important details such as view dependent effects, specular highlights, shadows, complex geometric shapes such as hair. These details are very important to achieve photorealism and we refer the reader to the qualitative results and the supplementary video for more exhaustive comparisons.

5.4 Ablation Study

In this section we study the effects of various design decisions in our pipeline.

5.4.1 Use of Light Visibility Maps. We evaluate the advantage of using a light visibility map. This map provides very strong cues on occlusion shadows and intensity of the light. Without such a map, the network would spend a lot of capacity trying to add convincing shadows. As can be seen in Figure 14, the realism of the novel renderings suffers without this map. Notably, these shadows come from the approximate acquired geometry, which generates rough visibility maps and does not model high frequency details, nevertheless, the network is able to compensate for many of these issues shown in the Figure.

5.4.2 Use of Reflection Maps. The reflection map guides the network towards specularities and view dependent effects. The importance of this input is shown in Figure 15. For this particular experiment we focused on very challenging illumination conditions, to highlight the view dependent effects. Note how the renderings generated with the network without reflection map have a more diffuse look and they are overall lower quality. This is expected since the network has no clue where specularities should occur and will try to hallucinate (memorize) them. On the other hand, by using the reflection map as input, the network's task reduces to learning the material properties (e.g. clothing vs skin).

Table 2. Quantitative evaluations on test images. We compare our method on two tasks: view synthesis and novel lighting. Note that some competitors cannot handle both the tasks simultaneously and we mark them as *Not Applicable (N/A)*. Perceptual error [Zhang et al. 2018] corresponds to the squared Euclidean distance between feature representations extracted using a pre-trained VGG model (the highest resolution feature map of the encoder is used). In **bold** we show the best overall metrics, in **red** we show the best method among the approaches that support simultaneous view synthesis and relighting.

		Proposed	Guo et al. [2019]	Thies et al. [2019]	Martin-Brualla et al. [2018]	Wenger et al. [2005a]	[Meka et al. 2019]
Novel View	Photometric Error ↓	1.5672	3.2339	2.0499	**1.5211**	N/A	N/A
	PSNR ↑	34.0750	28.6620	31.9682	**34.3952**	N/A	N/A
	MS-SSIM ↑	0.9705	0.9425	0.9630	**0.9729**	N/A	N/A
	Perceptual ↓	0.0583	0.0834	0.0713	**0.0561**	N/A	N/A
Novel Illumination	Photometric Error ↓	**2.4560**	4.054	N/A	N/A	2.6830	2.5336
	PSNR ↑	**31.5059**	27.5980	N/A	N/A	30.6604	31.4041
	MS-SSIM ↑	**0.9495**	0.9130	N/A	N/A	0.9424	0.9449
	Perceptual ↓	**0.0634**	0.0870	N/A	N/A	0.0660	0.0649

Fig. 14. Light visibility map ablation. See how self occlusions and shadows are better captured thanks to the proposed approach.

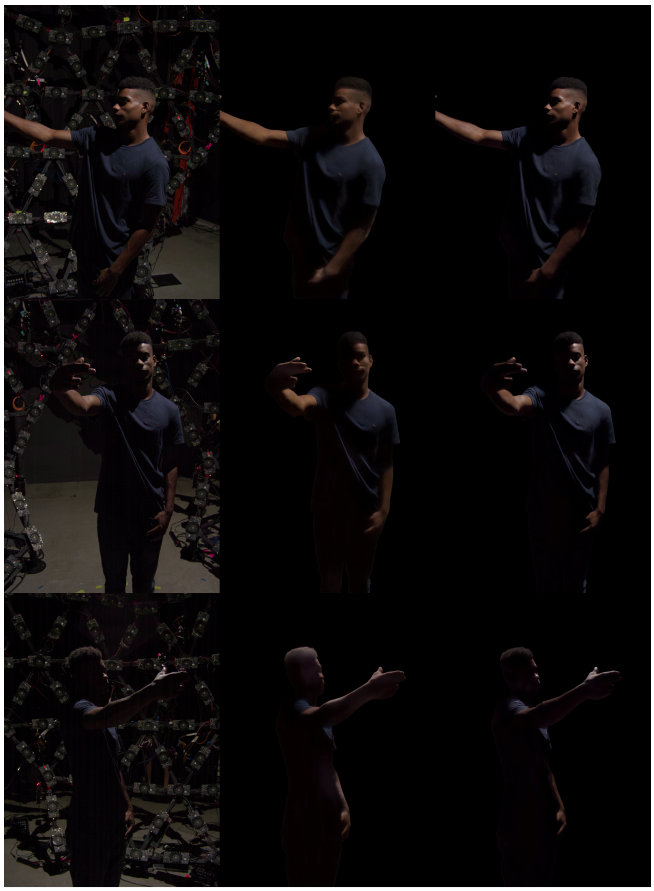

Fig. 15. Reflection map ablation. The reflection map is the key to recover view dependent effects in challenging illumination conditions. Note how without this map, the subject looks more diffuse and lacks photorealism.

5.4.3 Neural Renderer Size. In this experiment we evaluate the impact of the neural rendering size. In particular we consider the architecture presented in Section 3.3 in the main paper and remove half of the layers. In Figure 16 we show a comparison between the two architectures. The small neural rendering module is able to capture the majority of the details including view dependent effects on the skin. Not surprisingly, the bigger architecture does a better job in terms of sharpness, however this demonstrates that smaller architectures are still valuable when efficiency becomes a critical constraint.

Fig. 16. Neural renderer size ablation. We compare the full size neural rendering module described in Section 3.3 in the main paper, with a model that has half the number of layers and therefore lightweight. Notice how even the small model is able to capture the majority of the details including specularities on skin. The large model is however able to retrieve sharper results as expected.

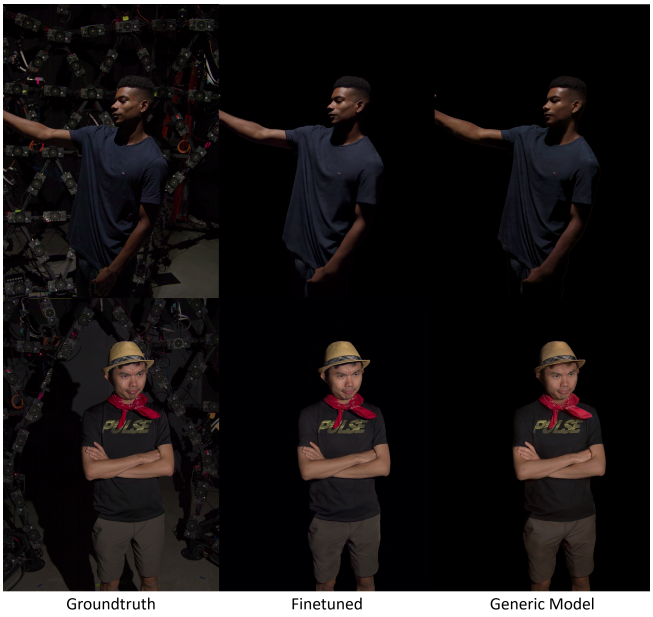

Fig. 17. Generalization experiment comparing an unseen subject (right) with the same fine-tuned model on the subject (middle). View and light condition are both unseen to the network.

5.4.4 Seen vs Unseen Subjects. Here we compare our model trained on all the subjects with the same model but fine-tuned on an unseen performer. The fine-tuning can be efficiently done in a few hours of training: empirically we found the neural renderer is the component that benefits the most from this. Indeed, whereas the feature extractors may capture more general properties of the human bodies, the neural renderer module is more dependent on the specific geometry as also shown by Martin-Brualla et al. [2018].

The goal of this experiment is to demonstrate generalization capabilities with respect to appearance. Figure 17 shows the results. Overall the model generalizes well on the unseen performer, thanks to the feature extractors that can build neural textures on the fly and the 1×1 convolutions that do not depend on the specific parameterization.

Some loss in sharpness is however noticeable and for brightly-hued scenes a slight color shift may occur, which we argue is due to the small training dataset, which does not contain enough variability. In practice, this can be mitigated by recording a short OLAT sequence for each new subject before the actual performance, and fine-tune the neural renderer. This seems a reasonable approach since introduces very little overhead to the system.

6 FREE-VIEWPOINT RELIGHTING AND COMPOSITING

The ambitious goal of a volumetric capture system consists of rendering captured humans photorealistically in virtual scenes with convincing illumination. To do so, three main components are necessary: free viewpoint rendering, relighting and compositing. Relatively few earlier methods [Einarsson et al. 2006; Guo et al. 2019; Li et al. 2013; Theobalt et al. 2007] have demonstrated this for *dynamic performances*, with the latest method by Guo et al. [2019] showing the highest visual quality. Although the improvements over previous work are consistent, their final results are still far from photorealistic (discussed in the Appendix). In this section we show that our framework is the first neural rendering approach able to achieve free-viewpoint relighting and compositing of *dynamic* performances, surpassing the state-of-the-art in terms of photorealism.

6.1 Static HDRI Relighting

Our system is able to infer OLAT images from any desired viewpoint, and, due the additive property of light, an image can be rendered under a completely arbitrary lighting condition, albeit with a fair amount of computation.

To achieve this, following the seminal work of Debevec et al. [2000], we render a dense set of (331) OLAT images covering the shell of the Light Stage. For an arbitrary lighting condition, encoded as a full HDR map, one can look up a weight in the map for each light in the stage. The OLAT images can then be linearly combined, using the looked up weights as coefficients, to obtain a realistic image of the subject under the desired lighting condition.

In order to generate compelling compositing results, we leverage the alpha mask prediction networks to blend the relit foreground into a new background. Thus we realistically augment virtual scenes with arbitrary lighting conditions.

Figure 18 shows our synthesized images and compares them with the groundtruth compositing and with the method of Guo et al. [2019]. Note how our framework generates visually pleasant images with an increase in photorealism. Although we rely on the same geometry of Guo et al. [2019], our neural renderer is able to mitigate imperfections, particularly in hair and on the neck. View dependent effects and specularities are better captured, especially in environments with strong directional light sources. Finally, our alpha mask prediction provides a more convincing blending with the virtual environments.

Fig. 18. Relighting and Compositing results from multiple viewpoints. For a set of fixed viewpoints and a static subject, the acquired OLATs and alpha masks are used to generate the groundtruth compositing. Our method shows an increase of photorealism compared to Guo et al. [2019]: note how imperfections in the geometry are mitigated, view dependent effects are better captured and the predicted alpha masks allow for more convincing compositing results.

6.2 Dynamic HDRI Relighting

When we consider sequences with *moving* performers, OLAT images cannot be captured since they require the subject static for the whole acquisition. Only a few methods can provide simultaneous free-viewpoint synthesis and relighting of dynamic performances [Einarsson et al. 2006; Guo et al. 2019; Li et al. 2013; Theobalt et al. 2007], and the work of Guo et al. [2019] is the most advanced one achieving the highest visual fidelity. Here we demonstrate that our work achieves more convincing results also in this setting.

To perform HDRI relighting of a dynamic sequence, OLAT images must be available for each frame. In particular, for a 10 seconds clip at 60Hz (600 frames per sequence), we need to run our method to generate 200, 000 OLATs. We recall that once we build the neural textures, only the neural rendering needs to run to generate novel viewpoints and lighting conditions. This can be efficiently done with a cloud implementation, that takes roughly 2 hours to generate all the outputs. In fact, a non-optimized CPU implementation runs roughly at 1 fps per rendering, which is already sufficient to generate renderings at scale in the cloud with no manual intervention.

In Figure 19, we compare our approach with Guo et al. [2019]. Note how our approach generates more realistic images in every environment. In particular, our renderings look more visually pleasant when a strong directional light is in the scene (first column). We are also able to capture complex geometric structure such as hair (second and third column). View dependent effects are better modeled by the neural renderer (see first and fourth example). Finally, because of the predicted alpha mask, our renderings have a

Fig. 19. Dynamic HDRI Relighting and Compositing from arbitrary viewpoints. Our method provides more visually pleasant results compared to the previous state-of-the-art approach of Guo et al. [2019]. Note how our approach mitigates geometric errors in complex structure such as hair, view dependent effects are better captured by the neural renderer and our alpha mask prediction enables a more natural blend of the subject in the virtual scene.

more natural silhouette and they blend more naturally into the environment. More examples can be appreciated in the supplementary video.

7 LIMITATIONS

Although our neural rendering approach achieves consistently higher quality compared to standard graphics rendering pipelines, it still has its limitations. Similar to previous performance capture works [Collet et al. 2015; Guo et al. 2019], it requires an elaborate multi-view setup with custom hardware, due to which this high-quality performance capture can only be performed in a studio and does not generalize to in-the-wild settings. Note that while our framework uses 58 RGB cameras, this is still substantially lower than neural rendering based volumetric capture methods [Mildenhall et al. 2020; Thies et al. 2019] that require hundreds or even thousands of views.

Our method also generates very strong view and light-direction dependent specular effects. But these specularities can sometimes have a 'flat' appearance due to the low dynamic range of our input and output images, as apparent in the eye and hair regions in Figure 20. This can be improved using HDR cameras with higher bit-depth. The limitation of convolutional neural networks in generating very high spatial frequencies, or the implicit spatial smoothness in their output, can also be a contributing factor. Recent work has shown that using periodic activation functions [Sitzmann et al. 2020] in the network architecture or applying fourier feature mapping [Tancik et al. 2020] to the input and output images may help in resolving this issue.

While the method offers a good degree of robustness to imperfections in geometry estimate, it still suffers when poor 3D reconstruction leads to large pieces of missing surfaces. These are likely irrecoverable as the projected neural texture must be in the receptive field of the feature extractor and neural renderer for infilling to occur. Similarly, non-opaque surfaces such as glasses or jewelry that exhibit transparency, scattering and refraction effects will introduce large geometric errors causing the neural texture to be projected to incorrect portions of the image. Examples of these limitations can be observed in Figure 20.

While our result look photorealistic, the network introduces some blur and loss in resolution when compared to the actual groundtruth images. In brightly-hued apparel we also sometimes notice a slight color shift. The predicted alpha mask is also not always of the same quality as the groundtruth matte. These limitations are more evident in unseen subjects, although this can be mitigated either by acquiring a larger training set with more variability or simply by capturing an extra static OLAT sequence per performer, which introduces a 6 second recording overhead that is, in our experience, relatively minor.

Finally, leveraging the additive property of light to perform relighting under an arbitrary lighting condition produces a bottleneck as our system must first render 331 OLATs per frame. Although this can be done efficiently with a parallel processing system in the

Fig. 20. Examples of failure cases. Despite the achieved photorealism compared to previous work, very high frequency details such as earrings and individual hair strands are not correctly recovered.

cloud, more compact lighting representations can be used to enable real-time HDRI relighting.

8 DISCUSSION

We presented a novel system that makes significant progress towards achieving photorealism for relightable humans in motion synthesized from arbitrary viewpoints. The key intuition is the employment of a neural renderer that leverages an approximate mesh, but crucially can compensate for errors in the geometry and reflectance properties.

This is achieved by sampling a learned neural texture and using an image-space neural renderer to synthesize the final output image. Contrary to other work that optimizes a neural texture to reproduce a set of ground truth images [Thies et al. 2019], our system can directly infer neural textures from multi-view color gradient images for a new subject. As result, the proposed framework provides a practical way to control lighting and viewpoint while changing the appearance of the subject.

Interesting directions of future work could be to remove the dependency on a mesh with an explicit UV parameterization. Also, removing the computationally intensive requirement of rendering a full set of OLATs in order to generate a relit image for a given HDRI map and instead directly rendering the desired image could also be practically impactful. Finally, the proposed system gives an effective way to generate ground truth at scale of moving performers.

REFERENCES

Robert Anderson, David Gallup, Jonathan T Barron, Janne Kontkanen, Noah Snavely, Carlos Hernández, Sameer Agarwal, and Steven M Seitz. 2016. Jump: virtual reality video. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 1–13.

Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frédo Durand, and John V. Guttag. 2018. Synthesizing Images of Humans in Unseen Poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Jonathan T. Barron and Jitendra Malik. 2015. Shape, Illumination, and Reflectance from Shading. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 8 (2015).

Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. 2010. High-quality Single-shot Capture of Facial Geometry. In *ACM SIGGRAPH 2010*.

Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross. 2011. High-quality Passive Facial Performance Capture Using Anchor Frames. In *ACM SIGGRAPH 2011*.

Volker Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *Proc. of the Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '99)*.

Cedric Cagniart, Edmond Boyer, and Slobodan Ilic. 2010. Probabilistic Deformable Surface Tracking From Multiple Videos. In *Proc. ECCV (Lecture Notes in Computer Science)*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.), Vol. 6314. Springer, Heraklion, Greece, 326–339. https://doi.org/10.1007/978-3-642-15561-1_24

Joel Carranza, Christian Theobalt, Marcus Magnor, and Hans-Peter Seidel. 2003. Free-Viewpoint Video of Human Actors. *ACM Trans. Graph.* 22 (07 2003), 569–577. https://doi.org/10.1145/882262.882309

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019. Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision*. 5933–5942.

Zhang Chen, Anpei Chen, Guli Zhang, Chengyuan Wang, Yu Ji, Kiriakos N. Kutulakos, and Jingyi Yu. 2019. A Neural Rendering Framework for Free-Viewpoint Relighting. *CoRR* (2019).

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-quality Streamable Free-viewpoint Video. *ACM TOG* (2015).

Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. 2008. Performance Capture from Sparse Multi-View Video. *ACM Trans. Graph.* 27, 3 (Aug. 2008), 1–10. https://doi.org/10.1145/1360612.1360697

Paul Debevec. 2012. The Light Stages and Their Applications to Photoreal Digital Actors. In *SIGGRAPH Asia*. Singapore.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the Reflectance Field of a Human Face. In *Proceedings of SIGGRAPH 2000 (SIGGRAPH '00)*.

Paul Debevec, Andreas Wenger, Chris Tchou, Andrew Gardner, Jamie Waese, and Tim Hawkins. 2002. A Lighting Reproduction Approach to Live-Action Compositing. In *SIGGRAPH*.

Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. 2015. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1538–1546.

Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. 2017. Motion2Fusion: Real-time Volumetric Performance Capture. *SIGGRAPH Asia* (2017).

Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi. 2016. Fusion4D: Real-time Performance Capture of Challenging Scenes. *SIGGRAPH* (2016).

Per Einarsson, Charles-Felix Chabert, Andrew Jones, Wan-Chun Ma, Bruce Lamond, Tim Hawkins, Mark Bolas, Sebastian Sylwan, and Paul Debevec. 2006. Relighting Human Locomotion with Flowed Reflectance Fields. In *Proceedings of the 17th Eurographics Conference on Rendering Techniques (EGSR)*.

S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. 2018. Neural scene representation and rendering. *Science* 360, 6394 (2018).

S. R. Fanello, C. Rhemann, V. Tankovich, A. Kowdle, S. Orts Escolano, D. Kim, and S. Izadi. 2016. HyperDepth: Learning Depth from Structured Light Without Matching. In *CVPR*.

Sean Ryan Fanello, Julien Valentin, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, Carlo Ciliberto, Philip Davidson, and Shahram Izadi. 2017. Low Compute and Fully Parallel Computer Vision with HashMatch. In *ICCV*.

Graham Fyffe, Cyrus A. Wilson, and Paul Debevec. 2009. Cosine Lobe Based Relighting from Gradient Illumination Photographs. In *SIGGRAPH '09: Posters (SIGGRAPH '09)*.

Pablo Garrido, Levi Valgaert, Chenglei Wu, and Christian Theobalt. 2013. Reconstructing Detailed Dynamic Face Geometry from Monocular Video. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 32, 6, Article 158 (Nov. 2013), 10 pages.

Pablo Garrido, Michael Zollhoefer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Perez, and Christian Theobalt. 2016. Reconstruction of Personalized 3D Face Rigs from Monocular Video. (2016).

Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2018. Practical Dynamic Facial Appearance Modeling and Acquisition. In *SIGGRAPH Asia*.

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram

Izadi. 2019. The Relightables: Volumetric Performance Capture of Humans with Realistic Relighting. In *ACM TOG*.
Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. 2018. Deep Blending for Free-Viewpoint Image-Based Rendering. *SIGGRAPH Asia* (2018).
Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3D Avatar Creation from Hand-held Video Input. *ACM Trans. Graph.* 34, 4, Article 45 (July 2015), 14 pages.
Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. 2016. VolumeDeform: Real-time Volumetric Non-rigid Reconstruction.
Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. 2015. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems (NIPS)*.
Michael Kazhdan and Hugues Hoppe. 2013. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)* 32, 3 (2013), 1–13.
Sean Kelly, Samantha Cordingley, Patrick Nolan, Christoph Rhemann, Sean Fanello, Danhang Tang, Jude Osborn, Jay Busch, Philip Davidson, Paul Debevec, Peter Denny, Graham Fyffe, Kaiwen Guo, Geoff Harvey, Shahram Izadi, Peter Lincoln, Wan-Chun Alex Ma, Jonathan Taylor, Xueming Yu, Matt Whalen, Jason Dourgarian, Genevieve Blanchett, Narelle French, Kirstin Sillitoe, Tea Uglow, Brenton Spiteri, Emma Pearson, Wade Kernot, and Jonathan Richards. 2019. AR-ia: Volumetric Opera for Mobile Augmented Reality. In *SIGGRAPH Asia 2019 XR*.
Hyeongwoo Kim, Mohamed Elgharib, Hans-Peter Zollöfer, Michael Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. 2019. Neural Style-Preserving Visual Dubbing. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 178:1–13.
Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollöfer, and Christian Theobalt. 2018. Deep Video Portraits. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 163.
Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* (2014).
Adarsh Kowdle, Christoph Rhemann, Sean Fanello, Andrea Tagliasacchi, Jon Taylor, Philip Davidson, Mingsong Dou, Kaiwen Guo, Cem Keskin, Sameh Khamis, David Kim, Danhang Tang, Vladimir Tankovich, Julien Valentin, and Shahram Izadi. 2018. The Need 4 Speed in Real-Time Dense Visual Tracking. *SIGGRAPH Asia* (2018).
Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. 2015. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*. 2539–2547.
Guannan Li, Chenglei Wu, Carsten Stoll, Yebin Liu, Kiran Varanasi, Qionghai Dai, and Christian Theobalt. 2013. Capturing Relightable Human Performances under General Uncontrolled Illumination. *Computer Graphics Forum (Proc. EUROGRAPHICS 2013)* (2013).
Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. 2019. Neural Rendering and Reenactment of Human Actor Videos. *ACM Transactions on Graphics* (2019).
Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep Appearance Models for Face Rendering. *ACM Trans. Graph.* 37, 4, Article 68 (July 2018).
Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *SIGGRAPH* (2019).
Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose guided person image generation. In *NIPS*.
Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Disentangled Person Image Generation. *CVPR* (2018).
Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. 2007. Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination. In *Proceedings of the Eurographics Conference on Rendering Techniques (EGSR'07)*.
Ricardo Martin-Brualla, Rohit Pandey, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Julien Valentin, Sameh Khamis, Philip Davidson, Anastasia Tkach, Peter Lincoln, Adarsh Kowdle, Christoph Rhemann, Dan B Goldman, Cem Keskin, Steve Seitz, Shahram Izadi, and Sean Fanello. 2018. LookinGood: Enhancing Performance Capture with Real-time NeuralRe-Rendering. In *SIGGRAPH Asia*.
Abhimitra Meka, Gereon Fox, Michael Zollhöfer, Christian Richardt, and Christian Theobalt. 2017. Live User-Guided Intrinsic Video For Static Scene. *IEEE Transactions on Visualization and Computer Graphics* 23, 11 (2017).
Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhoefer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul Debevec, Christian Theobalt, Julien Valentin, and Christoph Rhemann. 2019. Deep Reflectance Fields - High-Quality Facial Reflectance Field Inference From Color Gradient Illumination. *ACM Transactions on Graphics (Proceedings SIGGRAPH)*.
Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. 2018. LIME: Live Intrinsic Material Estimation. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. 11.
Microsoft. 2014. UVAtlas - isochart texture atlasing. http://github.com/Microsoft/UVAtlas
Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. arXiv:cs.CV/2003.08934
Natalia Neverova, Riza Alp Güler, and Iasonas Kokkinos. 2018. Dense Pose Transfer. In *European Conference on Computer Vision (ECCV)*.
Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. 2016. Holoportation: Virtual 3D Teleportation in Real-time. In *UIST*.
Rohit Pandey, Anastasia Tkach, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Ricardo Martin-Brualla, Andrea Tagliasacchi, George Papandreou, Philip Davidson, Cem Keskin, Shahram Izadi, and Sean Fanello. 2019. Volumetric Capture of Humans with a Single RGBD Camera via Semi-Parametric Learning. In *CVPR*.
Pieter Peers, Tim Hawkins, and Paul E. Debevec. 2006. *A Reflective Light Stage*. Technical Report.
Matt Pharr, Wenzel Jakob, and Greg Humphreys. 2016. *Physically Based Rendering: From Theory to Implementation* (3rd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei Efros, and George Drettakis. 2019. Multi-view Relighting Using a Geometry-Aware Network. *SIGGRAPH* (2019).
Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. 2015. Dyna: A Model of Dynamic Human Shape in Motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH)* 34, 4 (Aug. 2015), 120:1–120:14.
Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI* (2015).
Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. 2020. Background Matting: The World is Your Green Screen. In *Computer Vision and Pattern Recognition (CVPR)*.
Mike Seymour. 2020. Face it Will: Gemini Man. *https://www.fxguide.com/fxfeatured/face-it-will-gemini-man/* (2020).
Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, R S Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, I. M. Pasechnik, Dmitry Ulyanov, Alexander Vakhitov, and Victor S. Lempitsky. 2019. Textured Neural Avatars. *CVPR* (2019).
Chenyang Si, Wei Wang, Liang Wang, and Tieniu Tan. 2018. Multistage Adversarial Losses for Pose-Based Human Image Synthesis. In *CVPR*.
Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. 2020. Implicit Neural Representations with Periodic Activation Functions. In *arXiv*.
Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. 2018. DeepVoxels: Learning Persistent 3D Feature Embeddings. *CoRR* abs/1812.01024 (2018).
Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. 2019. Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations. arXiv:cs.CV/1906.01618
J. Starck and A. Hilton. 2007. Surface Capture for Performance-Based Animation. *IEEE Computer Graphics and Applications* (2007).
Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. 2019. Single image portrait relighting. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 79.
Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. 2020. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. *arXiv preprint arXiv:2006.10739* (2020).
L. M. Tanco and A. Hilton. 2000. Realistic synthesis of novel human movements from a database of motion capture examples. In *Proceedings Workshop on Human Motion*.
Vladimir Tankovich, Michael Schoenberg, Sean Ryan Fanello, Adarsh Kowdle, Christoph Rhemann, Max Dzitsiuk, Mirko Schmidt, Julien Valentin, and Shahram Izadi. 2018. SOS: Stereo Matching in O(1) with Slanted Support Windows. *IROS* (2018).
Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, Rohit Pandey, Sean Fanello, Gordon Wetzstein, Jun-Yan Zhu, Christian Theobalt, Maneesh Agrawala, Eli Shechtman, Dan B Goldman, and Michael Zollhoefer. 2020. State of the Art on Neural Rendering. In *Eurographics*.
Christian Theobalt, Naveed Ahmed, Hendrik P. A. Lensch, Marcus A. Magnor, and Hans-Peter Seidel. 2007. Seeing People in Different Light-Joint Shape, Motion, and Reflectance Capture. *IEEE TVCG* 13, 4 (2007), 663–674.
Justus Thies, Michael Zollhoefer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. 2016. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *Proc. CVPR*.

Justus Thies, Michael Zollhöfer, and Matthias Niessner. 2019. Deferred Neural Rendering: Image Synthesis Using Neural Textures. *SIGGRAPH and ACM TOG* (2019).
Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. 2008. Articulated Mesh Animation from Multi-View Silhouettes. *ACM Trans. Graph.* 27, 3 (Aug. 2008), 1–9. https://doi.org/10.1145/1360612.1360696
Shenlong Wang, Sean Ryan Fanello, Christoph Rhemann, Shahram Izadi, and Pushmeet Kohli. 2016. The Global Patch Collider. *CVPR* (2016).
Zhen Wen, Zicheng Liu, and T. S. Huang. 2003. Face relighting with radiance environment maps. In *CVPR*.
Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. 2005a. Performance Relighting and Reflectance Transformation with Time-Multiplexed Illumination. In *SIGGRAPH*.
Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. 2005b. Performance Relighting and Reflectance Transformation with Time-multiplexed Illumination. In *ACM SIGGRAPH 2005 Papers (SIGGRAPH '05)*.
Chenglei Wu, Carsten Stoll, Levi Valgaerts, and Christian Theobalt. 2013. On-set Performance Capture of Multiple Actors with a Stereo Camera. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 32, 6, Article 161 (Nov. 2013), 11 pages.
Zexiang Xu, Sai Bi, Kalyan Sunkavalli, Sunil Hadap, Hao Su, and Ravi Ramamoorthi. 2019. Deep View Synthesis from Sparse Photometric Images. *SIGGRAPH* (2019).
Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. 2018. Deep image-based relighting from optimal sparse samples. *ACM Trans. on Graphics* (2018).
A. Zaharescu, E. Boyer, K. Varanasi, and R. Horaud. 2009. Surface feature detection and description with applications to mesh matching. In *CVPR*.
Richard Zhang. 2019. Making Convolutional Networks Shift-Invariant Again. In *ICML*.
Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. *IEEE conference on computer vision and pattern recognition (CVPR)* (2018).
Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, Jonathan T. Barron, Ravi Ramamoorthi, and William T. Freeman. 2020. Neural Light Transport for Relighting and View Synthesis. *CoRR* (2020).
Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, and Jiashi Feng. 2017. Multi-View Image Generation from a Single-View. *CoRR* (2017).
Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David Jacobs. 2019. Deep Single Image Portrait Relighting. In *International Conference on Computer Vision (ICCV)*.
Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2014. Multi-view perceptron: a deep model for learning face identity and view representations. In *NIPS*.

PRELIMINARIES ON VOLUMETRIC PERFORMANCE CAPTURE

In this Appendix, we give the reader more context regarding state-of-the-art volumetric performance capture pipelines and their main drawbacks. The goal is to provide additional motivations for our work and in more general, the use of neural rendering for performance capture.

Volumetric capture, also known as free viewpoint videos, started with early works of Starck and Hilton [2007]; Tanco and Hilton [2000] and it has recently gained a lot of attention in the context of AR/VR [Collet et al. 2015]. Previous works [Li et al. 2013; Theobalt et al. 2007] have also shown how to capture dynamic subjects and relight them for more convincing renderings. Recent advances in Guo et al. [2019] pushed the technology even further, achieving convincing results.

In the following, we analyze the recent work of Guo et al. [2019], being currently the most advanced volumetric capture pipeline with realistic relighting, and describe its drawbacks in detail. Note that the elements discussed here are general and valid for many related works, since often these systems [Collet et al. 2015; Dou et al. 2017, 2016; Guo et al. 2019] share many similarities and building blocks.

A performance capture system usually consists of three main components: geometry acquisition, reflectance estimation, and rendering, which are depicted in Figure 21.

In Guo et al. [2019], *3D geometry* is captured using 16 high resolution custom active depth sensors, which provide *mm* precision in the considered range. The well established screened Poisson reconstruction [Kazhdan and Hoppe 2013] is used to obtain a 3D mesh. Despite the advances in 3D reconstruction, any mesh of a reasonable size and resolution would struggle at capturing very complex structures such as hair, which is one of the main limitations highlighted in Guo et al. [2019]. Once a 3D mesh is available, these systems [Collet et al. 2015; Guo et al. 2019] retrieve a parameterization often using the UVAtlas algorithm [Microsoft 2014]. The parameterized mesh allows for a common 2D texture space for all the multiview images. In order to build a texture map, a typical approach [Carranza et al. 2003; Collet et al. 2015] is to compute a weighted average of all views based on the surface normals and view direction of each camera. However, due to imperfect geometry and sub-pixel camera miscalibration, the rendered images are less detailed compared to the original raw images.

Reflectance Estimation in [Guo et al. 2019] relies on two lighting conditions proposed in [Fyffe et al. 2009]. In particular, the system captures two interleaved gradient illumination conditions. Consecutive frames are aligned through 2D optical flow, allowing for per-pixel correspondences in UV space between the two complementary illumination conditions. A cosine lobe model is then used to retrieve material properties such as albedo, gloss map and surface normals via simple per-pixel operations in RGB space. This component has multiple limitations: first, it relies on perfectly aligned gradient illumination conditions in texture space; second it assumes a hand-crafted cosine lobe BRDF, which, as detailed in the original paper [Guo et al. 2019], cannot accurately model specular highlights and view dependent effects due to its approximations; finally, per-pixel operations in RGB space are very sensitive to image noise.

Photorealistic Rendering of humans using traditional computer graphics pipelines remains a very challenging and an active topic of research. Recent movie productions [Seymour 2020] model geometry and texture up to pore-level and use multiple post-production effects to create realistic looking imagery of humans. But even given perfect geometry and material properties, photorealistic rendering of many people at scale (i.e. with no manual post-processing by skilled 3D artists) remains an unsolved problem.

Examples of these drawbacks are depicted in Figure 22. In b), we show a model with pre-baked (fixed) lighting (Figure 22, matching the groundtruth photograph (Figure 22, a). This shows a high degree of photorealism, and when a desired lighting condition is known, capturing in this modality always guarantees the best possible results, note however that this model is not relightable. Additionally, despite this subject having a simple geometry, fine details like earrings are not correctly captured and view dependent effects (e.g. specular highlights) are also baked-in and do not follow the actual camera viewpoint.

More commonly, volumetric capture systems usually acquire performers with fully lit illumination, which approximates their albedo (Figure 22, c). This modality still looks compelling in terms of quality and photorealism, but it does not match the desired illumination.

The system of Guo et al. [2019], allows for the estimation of reflectance maps such as albedo, photometric normals and gloss map. This model offers relightability and can be rendered in new environments. In Figure 22, d,e, we show examples of diffuse renderings and

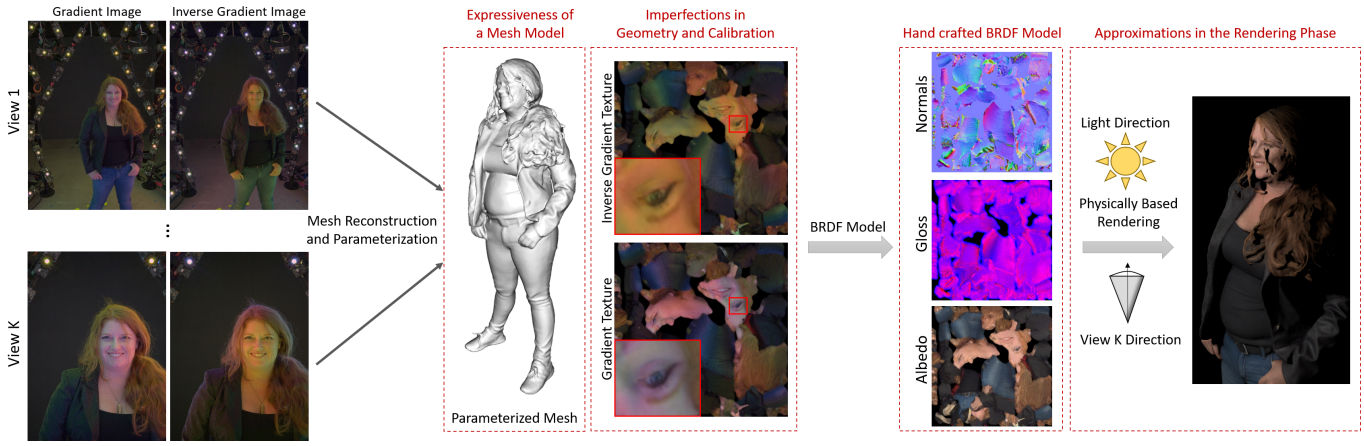

Fig. 21. Preliminaries on performance capture. A state-of-the-art volumetric capture and rendering pipeline (e.g. [Guo et al. 2019]) estimates a parameterized mesh using custom high resolution depth sensors. RGB images are acquired with interleaved spherical gradient illuminations that are used to estimate the reflectance maps. However, dense but finite resolution of the mesh model does not capture fine details such as hair; imperfect geometry and calibration may cause oversmooth texture maps. Hand-crafted BRDF models cannot accurately model high frequency details and view dependent effects. Finally, Physically Based Rendering is still a challenging topic especially when performed at scale with no manual post-processing. See text for details.

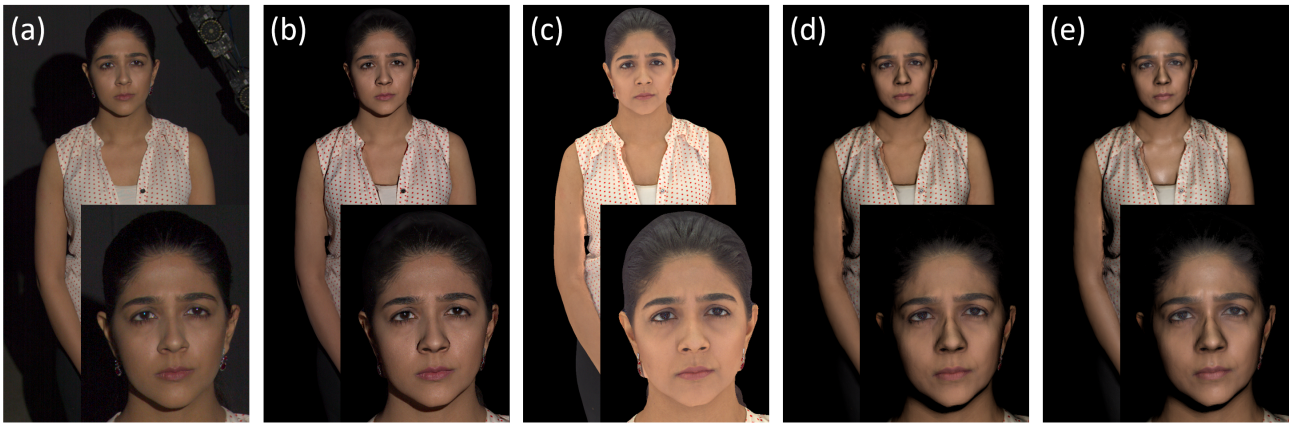

Fig. 22. Comparison between a captured photograph (a) and various rendering techniques. A 3D model captured with baked-in lighting (matching the photograph conditions) achieves the highest level of photorealism (b), this model is however not relightable and does not capture view dependent effects. The rendered albedo computed using Guo et al. [2019] also exhibits high degree of photorealism (c), but it does not match the photo illumination. When we relight the same model under the desired lighting condition, then the results start to move away from photorealism (d,e). See text for details.

the full model used by Guo et al. [2019]. Despite the evident improvements over previous work the final renderings start to look uncanny and not realistic due to the approximations discussed above.

These preliminaries motivate our approach and, more in general, the use of neural rendering for performance capture. Whereas improving traditional geometric reconstruction and rendering pipelines will still cover a key role for the next few years, we believe that neural rendering approaches, such as ours, will become more popular to overcome the limitations we discussed. As demonstrated in this paper, our goal is to enhance traditional performance capture systems and push the limits of photorealism.