Deep Reflectance Fields

High-Quality Facial Reflectance Field Inference from Color Gradient Illumination

ABHIMITRA MEKA, MPI Informatics, Saarland Informatics Campus, Google
CHRISTIAN HÄNE and ROHIT PANDEY, Google
MICHAEL ZOLLHÖFER, Stanford University
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, and PAUL DEBEVEC, Google
CHRISTIAN THEOBALT, MPI Informatics, Saarland Informatics Campus
JULIEN VALENTIN and CHRISTOPH RHEMANN, Google

Fig. 1. Deep Reflectance Fields – given only *two* observations (color gradient images) of an actor, our method is able to relight the subject under any lighting condition. Our approach generalizes to unseen subjects, viewpoints, illumination conditions and can handle dynamic performances.

We present a novel technique to relight images of human faces by learning a model of facial reflectance from a database of 4D reflectance field data of several subjects in a variety of expressions and viewpoints. Using our learned model, a face can be relit in arbitrary illumination environments using only two original images recorded under spherical color gradient illumination. The output of our deep network indicates that the color gradient images contain the information needed to estimate the full 4D reflectance field, including specular reflections and high frequency details. While capturing spherical color gradient illumination still requires a special lighting setup, reduction to just two illumination conditions allows the technique to be applied to dynamic facial performance capture. We show side-by-side comparisons which demonstrate that the proposed system outperforms the state-of-the-art techniques in both realism and speed.

Authors' addresses: Abhimitra Meka, MPI Informatics, Saarland Informatics Campus, Google, ameka@mpi-inf.mpg.de; Christian Häne; Rohit Pandey, Google; Michael Zollhöfer, Stanford University; 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, Google; Christian Theobalt, MPI Informatics, Saarland Informatics Campus; Julien Valentin; Christoph Rhemann, crhemann@google.com, Google.

0730-0301/2019/7-ART77 $15.00
https://doi.org/10.1145/3306346.3323027

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

Additional Key Words and Phrases: Reflectance estimation, machine learning

ACM Reference Format:
Abhimitra Meka, Christian Häne, Rohit Pandey, Michael Zollhöfer, 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 Trans. Graph.* 38, 4, Article 77 (July 2019), 12 pages. https://doi.org/10.1145/3306346.3323027

1 INTRODUCTION

Modifying the lighting in a facial portrait image is a much sought after capability that would benefit many visual effects including portrait photography, and virtual or augmented reality applications. This relighting is particularly challenging since the facial appearance is the result of a complex interaction of light with the many materials that make up the skin, eyes, hair, teeth, and clothing, each of which have complex geometry and varying amounts of specular reflection and subsurface scattering. Further, ignoring or approximating these properties is especially perilous as humans are highly capable of detecting the subtle cues of realism in facial renderings. While today's computer graphics techniques can produce photo-realistic digital human models which can be rendered in any lighting and from any viewpoint, creating such models is still extremely laborious

and expensive. Indeed, progress towards automated avatar creation still falls far short of photo-realism.

In order to reach the highest level of photo-realism, image-based relighting systems capture actors at high resolution under a large number of lighting conditions. For instance, high quality pore-level 4D reflectance fields of humans can be acquired with the Light Stage proposed by Debevec et al. [2000] – a spherical dome equipped with a large number of controllable light sources and cameras. The 4D reflectance field from one camera view can be sampled by capturing hundreds of one-light-at-a-time (OLAT) images, each of them capturing the subject illuminated by a single light on the Light Stage. By projecting the environment map of a new illumination condition onto this captured illumination basis, photo-realistically re-lit images of a subject can be created as a weighted combination of the OLAT images. The relighting results exhibit the full range of local and global effects, including diffuse lighting, specular reflections, inter-reflection, subsurface scattering, and self-shadowing. Unfortunately, capturing several hundreds OLAT images, a number typically required for high quality reflectance field capture, requires several seconds, e.g., ≈ 8 seconds using the Light Stage 2 [Debevec 2012]. Capturing a time-varying reflectance field of dynamic scenes in this way is challenging, and relies on a hardware setup variant equipped with high speed cameras, as well as an error-prone optical flow alignment step [Einarsson et al. 2006].

To allow the capture of dynamic scenes, the key is to be able to rely on a *small* set of input images that can be captured at real-time frame rates – while the actor is performing freely. In this setting, strong priors can help to better constrain reconstruction, but they introduce significant trade-offs. For instance, [Saito et al. 2017] and Yamaguchi et al. [2018] only handle skin, and can not correctly relight facial hair, eyes, teeth, accessories, or upper body clothing, since their underlying assumptions do not hold in these regions. An alternative to manually crafted priors is the use of learnable pipelines such as the one proposed by Xu et al. [2018]. Their deep neural network seeks to relight a scene under novel illumination based on a set of five optimal images captured under predefined directional lighting. The approach provides compelling results on synthetic data, but fails to handle complex object shapes and high frequency details such as shadows, and can only handle low image resolutions (128×128 pixels).

We introduce a new approach for the acquisition of high-quality time-varying 4D reflectance fields of a human actor at 30 fps in a Light Stage, without having to resort to time-multiplexing, motion compensation techniques, or priors. Our approach uses a deep neural network to learn a mapping from only *two* images, captured under spherical gradient illumination, to the full 4D reflectance field. As such, it can reconstruct *any* OLAT image from a given lighting direction. The predicted dynamic reflectance fields come very close in quality to models captured with a dense set of OLAT images. Our method enables quasi-photorealistic relighting of the *complete* human head as it handles skin subsurface scattering, wrinkle details, skin specularity, facial hair, and teeth, as well as the complex appearance of the human eyes in a unified manner, and in a way that generalizes across different identities. While a Light Stage only generates a *discrete* illumination basis due to the finite number of mounted light sources, we recover a *continuous* illumination basis, since the network can be evaluated for any illumination direction. Our core technical contributions can be summarized as:

- A capture system that enables 4D reflectance field estimation of moving subjects.
- A machine learning-based formulation that maps spherical gradient images to the OLAT image corresponding to a particular lighting direction.
- A task-specific perceptual loss trained to pick up specularities and high frequency details.
- A sliding window based pooling loss that robustly handles the small misalignments between the spherical gradient images and the groundtruth OLAT images.

Our experiments show that our method is effective in real applications such as relighting in arbitrary lighting environments and compares favorably with off-line capture systems and other state-of-the-art approaches.

2 RELATED WORK

Modeling photorealistic humans is an active research topic in the computer vision, graphics, and machine learning communities. Here we categorize related works that are representative of different trends in the literature as *parametric model fitting*, *image-based*, and *learning-based* solutions.

Parametric Model Fitting. These approaches assume strong priors, typically performing an explicit reconstruction while employing hand-designed reflectance and/or lighting models. General shape, illumination, and reflectance can be recovered based on a set of hand-crafted priors and optimization [Barron and Malik 2015; Meka et al. 2017]. Parametric models of geometry, surface reflectance, or illumination have been employed for reconstruction and relighting in the context of human bodies [Theobalt et al. 2007], faces [Blanz and Vetter [n. d.]; Garrido et al. 2013, 2016; Gotardo et al. 2018; Hawkins et al. 2004; Ichim et al. 2015; Thies et al. 2016], eyes [Bérard et al. 2016], eyelids [Bermano et al. 2015], and hair [Hu et al. 2015; Zhang et al. 2017]. Faces can be relit under a diffuse appearance assumption based on radiance environment maps and ratio-images [Wen et al. 2003]. Other approaches jointly estimate parametric BRDF models and wavelet-based incident illumination to relight 3D videos of humans [Li et al. 2013]. Relighting of the human head can be formulated as a mass transport problem [Shu et al. 2017] based on position and normal estimates recovered by a parametric face model. Cosine lobe relighting can be performed analytically based on a pair of spherical gradient illumination images [Fyffe et al. 2009], but secondary effects such as shadows are of low quality due to the use of approximations in modeling the face geometry. Some recent deep learning-based approaches [Saito et al. 2017; Yamaguchi et al. 2018] estimate the *parameters* of a predefined reflectance model from single images. The approach of Gotardo et al. [2018] for dynamic appearance estimation extracts SVBRDF (diffuse and specular) and geometry (also fine scale) from images captured under uniform lighting, but their approach is restricted to the skin region. Recently, multiple works have also focused on the challenging problem of extracting the SVBRDF from a *single* image using a flash [Li et al. 2018a,b; Nam et al. 2018]. Since, all model-based approaches use hand-crafted priors, they are typically limited to specific parts of

the human body and only handle these in isolation. Many of these approaches only work under low-frequency illumination conditions and do not handle the specularity of skin and sub-surface scattering effects. In contrast, our model-free approach enables relighting of the *complete* human head.

Image-Based Relighting. To reach the highest level of realism, image-based relighting techniques capture actors at high resolution under a large number of lighting conditions. High quality pore-resolution 4D reflectance fields of humans can be acquired with a Light Stage [Debevec et al. 2000]. Einarsson et al. [2006] illuminates the scene with a smaller set of approximately 30 lighting basis functions with larger spatial support to enable real-time capture, but this comes at the expense of lighting resolution. Other techniques use high framerate video and time-multiplex the sampling of the lighting basis over a window of several frames [Wenger et al. 2005], but this requires expensive and error prone motion estimation. An alternative approach is to use a reference subject's 4D reflectance field to modify the lighting on a target subject's performance using an aligned ratio image [Peers et al. 2007]. However, this requires having a 4D reflectance field available of a similar-looking subject and can transfer high-frequency details from the reference subject to the target. And for dynamic performances, this solution is approximate as it interpolates from a sparsely sampled collection of static poses. The style transfer technique of [Shih et al. 2014] matches local image statistics from a reference portrait to a target portrait and thereby is also able to perform some degree of relighting of the target portrait. However, the technique can require manual touch-up and can be challenged by harsh lighting scenarios. Unfortunately, the acquisition of 4D reflectance fields is a slow process and thus the subject would have to move in a stop-motion manner. This makes capturing high quality reflectance fields of dynamic facial performances very difficult, requiring expensive high speed cameras running at thousands of frames per second and potentially uncomfortable light levels [Wenger et al. 2005]. To the best of our knowledge, we introduce the first approach for deriving *time-varying* 4D reflectance fields of a human actor at 30 fps in a Light Stage.

Learning-Based Techniques. Deep learning based techniques have recently been applied to the problem of relighting arbitrary objects [Meka et al. 2018; Ren et al. 2015; Xu et al. 2018] and human bodies [Kanamori and Endo 2018]. The method of [Nalbach et al. 2017] showed that appearance synthesis can be cast as a learning based screen-space shading problem based on per-pixel scene attributes such as position, normal and reflectance. Based on a set of OLAT images, the approach of [Xu et al. 2018] is trained to relight a scene under novel illumination based on an optimal set of five jointly selected OLAT images. While results are compelling, it fails to handle complex object shapes, high-frequency specularities, and shadows caused by grazing angle illumination and non-convex geometry. The data-driven rendering of Lombardi et al. [2018] learns a joint representation of facial geometry and appearance from a multi-view capture setup, but this technique does not address the problem of relighting. The approach by Kanamori and Endo [2018] enables occlusion-aware inverse rendering for the human body, but results are restricted to Lambertian surfaces and low-frequency illumination. In contrast, we propose a novel machine learning-based

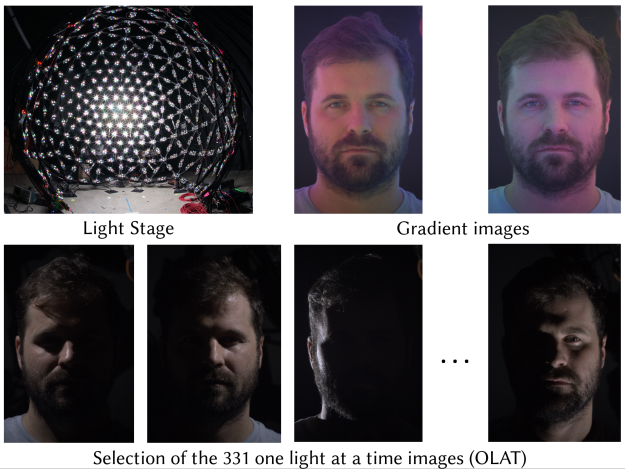

Fig. 2. **Capture setup** – We use programmable light sources mounted on a geodesic dome to light the subject under RGB color gradient images and OLAT images for training, inference, and validation data.

formulation that maps spherical gradient images to a full dataset of one-light-at-a-time (OLAT) images. This enables *model-free* relighting of dynamic scenes captured in a Light Stage.

In contrast to all other approaches, we leverage the insight of Fyffe et al. [2009], where spherical gradient images are used to derive full-color diffuse/specular albedo and surface normals for dielectric materials. Fyffe et al. [2009] assumed a simple cosine lobe model, which uses only *local* information and low frequency statistics to fit the BRDF. We argue that, if we provide a more expressive underlying model, spherical gradient images contain all the information necessary to generate the full reflectance field. In our method, this model is a neural network that infers the complex mapping from spherical gradient images to every possible directional lighting condition. Our model can take advantage of *non-local* information and contextual cues. For the first time, this enables estimating full reflectance fields of dynamic subjects without any explicit prior or BRDF model.

3 DEEP REFLECTANCE FIELDS

As light follows the superposition principle, one can photo realistically apply any desired lighting configuration to a given actor by combining a finite set of lighting conditions. In more detail, by capturing a set of images where only one light is turned on at a time (OLAT), one can linearly combine the RGB channels of these images in order to simulate a desired environment map; see Figure 3. In practice, sufficiently high sampling resolution in both captured images and light sources is key to ensure that details in both the surface (e.g., skin pores) and directional effects (e.g., specularities) are captured. The main disadvantage of this approach is the extended duration of time during which the subject has to remain still while the OLAT images are captured. As there are 331 lights in our system, the acquisition of the corresponding 331 OLAT images would take several seconds, making the capture and relighting of dynamic performances a real challenge. One of the contributions of this work is overcoming this limitation by directly regressing an

arbitrary OLAT image using *only two* observations of the subject captured under spherical gradient illumination.

3.1 Spherical Color Gradient Images

Spherical color gradient illumination images for reflectance estimation were originally proposed in [Fyffe et al. 2009]. Given the lighting direction vector θ of a LED relative to the center of the Light 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 $((1-\theta_x)/2, (1-\theta_y)/2, (1-\theta_z)/2)$. Figure 2 shows the two gradient images captured from a single camera viewpoint. Although simple to form, these images can be leveraged to recover important reflectance information about the surface being captured [Fyffe and Debevec 2015; Fyffe et al. 2009]. In particular, the patterns, when summed, produce a full-on white light condition which reveals the subject's *total reflectance* (diffuse plus specular), and the difference of the images encodes the average reflectance direction into the RGB color channels (a strong cue for surface normals). Further, the magnitude of the difference image relative to the sum image is a function of not only the BRDF but also the local self-shadowing (cues to shadow estimation). In this sense, the photographs under the two illumination patterns provide both geometric and albedo information to the inference algorithm. In contrast to previous work that interpreted gradient images using simple local parametric reflectance models, we employ deep learning to leverage the spatial context of the gradient images to infer far more realistic reflectance estimates.

3.2 Hardware and Data Capture

To acquire the necessary spherical gradient observations with corresponding ground truth OLAT images for training, we leveraged a LED sphere Light Stage capture setup [Debevec 2012]. Our Light Stage is a 3.5m diameter spherical dome on which 331 custom LED light sources with red, green, blue, and white controllable LEDs are evenly distributed as in Figure 2. Each of these LEDs is fully controllable by a driver, allowing it to emit light of any desired intensity and color. In order to capture actors at high resolution and under different viewpoints, we leverage nine Sony IMX253 cameras, capable of capturing 12.4 MP images at 60 Hz. All of the lights and cameras are synchronized via a hardware trigger.

Data Capture and Post-Processing. As we cast relighting as a supervised regression problem, we require corresponding inputs and outputs to train the neural network; see Figure 4. The input consists of two color spherical gradient images and a desired lighting direction, while the output is an OLAT image corresponding to that lighting direction. In order to relight an image at test time, we predict a collection of OLAT images (the full 4D reflectance field) using only the two spherical gradient images as input. Note that the OLAT images are only captured for *training* purposes and thus, at *inference* time, we only capture gradient images for the dynamic sequences we wish to relight. Precise pixel-to-pixel correspondence between OLATs and gradient images at training time is crucial to infer sharp OLAT images at inference time. Unfortunately, it is challenging for actors to remain completely still for the extended amount of time required to capture all 331 OLAT images. To overcome this challenge, when capturing training data, we interleave "tracking frames" into the capture sequence:

(1) Capture the 331 OLAT images, however:
 (1.1) After every 11 OLAT captures, capture a "tracking frame"
(2) Capture the two gradient images

A tracking frame is an image captured where *all* the lights on the Light Stage are turned on to generate homogeneous illumination. Once the capture session is over, we consider the last tracking image as a reference, and compute a dense optical flow-field across tracking frames using the method by Anderson et al. [2016]. The homogeneous illumination in the tracking frames is what makes the computation of dense optical flow possible. The optical flow field computed over tracking frames, is then linearly interpolated through time to provide correspondences across the OLATs. Although this procedure generally provides flow-fields of sufficiently good quality, the motion compensated frames can still present mis-alignments that could hinder the performance of our regressor. We address this issue by proposing a training loss that effectively compensates for small mis-alignments in image space; see Section 3.3.

3.3 Predicting Photo-Realistic 4D Reflectance Fields

In this section we describe our main algorithmic contribution: a deep neural network capable of predicting photo-realistic 4D reflectance fields for previously unseen faces. In more detail, given two gradient images and a lighting direction as input, we want to predict how any subject would look under white light coming from a specified spotlight direction. Our OLAT prediction can be seen as solving an image-to-image translation task [Chen and Koltun 2017; Isola et al. 2016; Zhu et al. 2017], where the goal is to start from input images from a certain domain and "translate" them into another domain. Our scenario is similar in the sense that we are transforming gradient illumination images to another image with the same content, but different illumination.

As such, the architecture that we employed is inspired by U-NET [Ronneberger et al. 2015] which has recently shown impressive results on image-to-image translation tasks involving photo-realistic images of humans [Martin-Brualla et al. 2018]. We employ a fully convolutional variant [Long et al. 2015] allowing efficient training of our network on *patches* and processing of high resolution images at inference time. At inference time, the input of our network is two spherical gradient images of resolution $W = 2560$ and $H = 3072$. Similar to [Eslami et al. 2018], we concatenate the lighting direction to *each* pixel of the input tensor. This results in an input tensor of size $W \times H \times 9$. The output of the network is an RGB image of size $W \times H \times 3$.

The U-NET encoder takes the input tensor and runs $M = 8$ convolutional layers using 3×3 convolutions. The output of the convolutions is immediately passed through a ReLU activation function, followed by a batch-normalization layer, and a max-pool layer. In the decoder stage we use bilinear upsampling followed by a convolutional layer. We use skip connections between the encoder and the decoder by concatenating the output from the encoder convolutional layer to the features at the corresponding decoder layer. The network is illustrated in Figure 4.

Fig. 3. **Image-Based Relighting** – An environment map (left) can be approximated with the 331 lighting directions of the light stage (middle). With a denser sampling of 1301 light directions, as enabled by our method, we obtain a better lighting environment approximation (right).

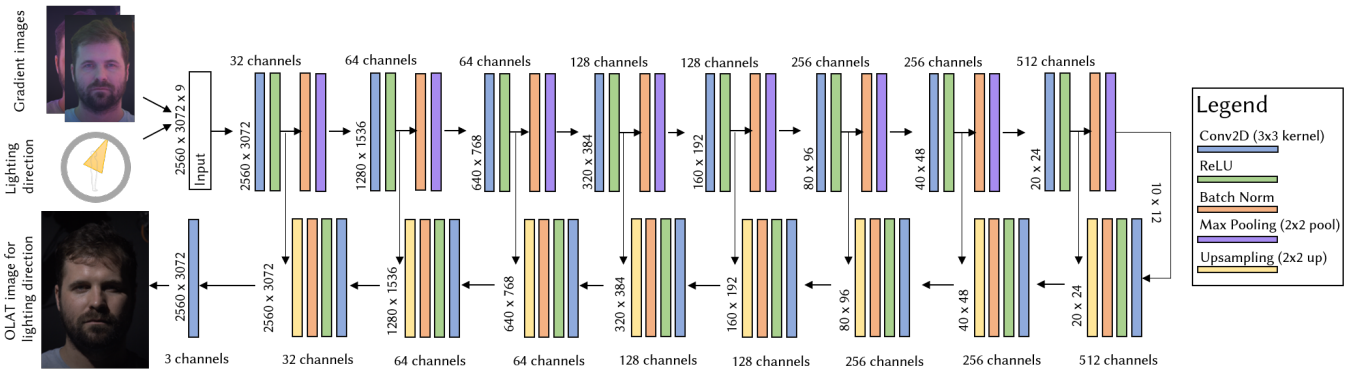

Fig. 4. **Pipeline** – Our network receives as input a pair of gradient images and a lighting direction. Via a U-Net architecture, it regresses the OLAT image that is corresponding to that particular lighting configuration. Network legend: Conv2D, ReLU, Batch Norm, MaxPooling and Upsampling.

3.4 Training

At training time we employ random crops with resolution $512 \times 512 \times 9$ as input to the network. After the $M = 8$ convolutional layers this produces a tensor of size $2 \times 2 \times 512$. Crops are crucial to train fast enough on high resolution images and to achieve the highest level of quality. They effectively limit the amount of context the network is able to see and hence prevent over-fitting [Kuo et al. 2018]. Using crops during training also enables the formulation of a novel patch-based *local* alignment strategy.

Training Setup. In order to hasten training, we distribute the training across 12 NVIDIA Tesla V100 GPUs. At each training epoch, we randomly pick a training frame, a patch within that frame and one OLAT per GPU. We use the ADAM optimizer [Kingma and Ba 2014] with a learning rate of 10^{-4}, and use exponential decay of the learning rate, with a rate of 0.1 every 10^6 iterations. We optimize our network for 1 million iterations before the training converges.

Training Losses. Choosing the appropriate loss for a new task is non-trivial and requires systematic trial and error. For example, a simple photometric loss does not lead to photo-realistic output as also shown by previous works [Martin-Brualla et al. 2018]. We therefore employ a loss function to specifically address this problem. Let I_{pred} be the prediction of our network and I_{gt} the ground truth OLAT image, we define our loss as:

$$\mathcal{L} = \|\text{Perc}(I_{\text{pred}}) - \text{Perc}(I_{\text{gt}})\|_2^2 \,. \tag{1}$$

The loss is the squared ℓ_2-norm of the difference in feature space between the predicted image and the ground truth image. Here, we indicate feature space by $\text{Perc}(\cdot)$. A common choice in the literature is to use a VGG network [Simonyan and Zisserman 2014] pre-trained on ImageNet to compute the perceptual loss [Zhang et al. 2018]. While this loss is well suited for generic natural images, our task at hand is specific and such an ImageNet trained model would lead to sub-optimal results, especially when regressing specularities and other high frequency details; see Figure 5. Therefore, we propose to enhance the loss using a VGG architecture that has been trained on a more relevant task: we consider as input a random image patch sampled from a groundruth OLAT image I_{gt} and the goal is to correctly determine which light direction generated the given patch – we recall that in total we have 331 light directions. We cast the problem as a regression task and hence train the network to minimize the ℓ_1-loss between the predicted direction and the ground truth direction. Training was stable with ℓ_2 or ℓ_1 losses. We used an ℓ_1 loss as it tends to produce sharper results for image-to-image translation tasks.

As specularities heavily depend on the direction of incoming light, a perceptual loss using our new task specific VGG network is particularly sensitive to these high-frequency effects, but is inferior to a perceptual loss using a network trained on ImageNet when it comes to reconstructing lower frequencies; see Figure 5. As these two losses capture complementary aspects of the desired result, we combine them $\mathcal{L} = \mathcal{L}_{\text{pretrained}} + \lambda \mathcal{L}_{\text{specific}}$, where the two components $\mathcal{L}_{\text{pretrained}}$ and $\mathcal{L}_{\text{specific}}$ are obtained by using the pre-trained VGG and the task specific VGG loss respectively. In more detail, we use five convolutional layers from each VGG and rescale activations by their corresponding feature length to ensure that they all contribute in the same manner to the final loss. We use $\lambda = 0.5$ in our training.The effect of this loss is shown in Figure 5.

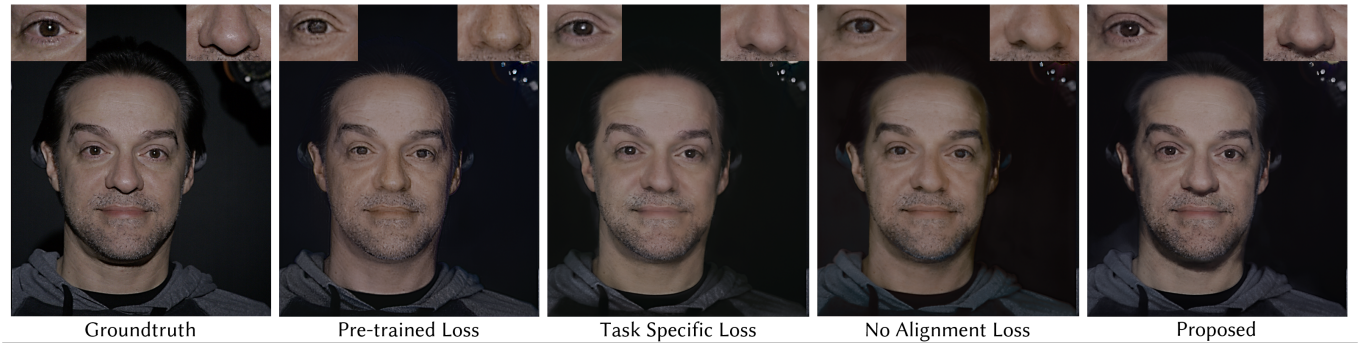

Fig. 5. **Training losses** – Effect of different training losses on the final results. (a) ground truth, results generated with: (b) VGG pre-trained on ImageNet, (c) task dependent specific loss, (d) without alignment loss, (e) proposed loss.

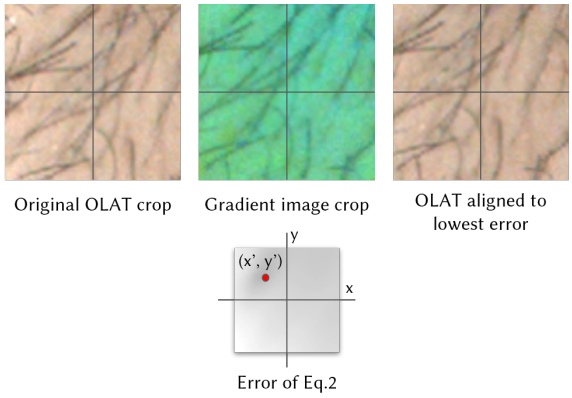

Fig. 6. **Alignment Loss**. Our slide-pooling loss accounts for misalignment between the ground truth OLAT crop (top left) and the gradient crop (top center). In the (x, y) coordinate frame (bottom) the energy of Equation 2 has a minimum at (x', y') slightly up and to the left, marked with a red dot. The ground truth OLAT image aligned to the minimum energy (top right) appears well aligned to the gradient image.

Sliding window pooling loss. Slight misalignments of gradient and ground truth OLAT images leads to complications with losses that assume pixel-perfect alignment; see Sec. 3.2. Indeed, naively computing the pixel difference loss will result in blurred results. To solve this problem, we propose a novel alignment strategy:

$$x', y' = \underset{x,y}{\operatorname{argmin}} \sum_u \sum_v \|I_{\text{gt}}(u-x, v-y) - I_{\text{pred}}(u, v)\|_1, \quad (2)$$

where $I(u, v)$ is the intensity value for a certain pixel location (u, v), the offsets x, y are sampled in a $[-20, 20]\times[-20, 20]$ window, and $\hat{x}, \hat{y}$ are the optimal offsets that correspond to the best aligning image, denoted $\hat{I}_{\text{gt}}$. The image $\hat{I}_{\text{gt}}$ is then used in Equation 1 instead of I_{gt}, effectively producing a slide-pooling loss that takes into account translational mis-alignments; see Figure 6.

3.5 Inference

As described in Section 3.2, we only capture two gradient images per frame, allowing us to capture relightable data at 30Hz. Once the data is captured, the user only needs to define the lighting environment that should be used for relighting the captured sequences. A dense set of light directions from which to sample the environment map also has to be defined. We run each of these directions together with the two gradient images through the network to estimate the corresponding OLAT images. Once all the OLAT images have been obtained, they can be combined according to the environment map to form the relit images. It is interesting to note that the number of lights composing that environment map can be *much greater* than the 331 used during training, leading to more detailed relit images. For input images of $2560x3072$ resolution, the inference time of our network for a single OLAT image, averaged over 100 runs, is $270.14ms$ on a workstation using an Nvidia TitanXp GPU and $1360.65ms$ on a workstation with only 2 Intel Xeon Gold 6154 CPUs. Although the inference time seems quite high, we use parallel GPU clusters to speed up the OLAT inference.

4 EXPERIMENTS

In this section, we perform an in-depth analysis of the proposed approach. To this end, we captured a dataset with 18 subjects. For each subject, we recorded sets of 331 ground truth OLAT images, 2 gradient illuminations, and 33 fully lit tracking frames. As mentioned in Section 3.2, tracking and OLAT images are only used at training time. For each person, we recorded their imagery from 9 different viewpoints. We additionally recorded each subject giving a dynamic facial performance for 5 seconds while interleaving the two color gradient lighting conditions. We split the captured data into a training set consisting of frames from (10) training subjects and a test set consisting of frames from (8) test subjects. We only used 5 viewpoints for training, leaving 4 unseen viewpoints for testing.

4.1 Qualitative Comparisons

In this section we show qualitative results on different test sequences and under different conditions. It is important to note that none of the subjects used for these comparisons are part of the training set.

OLAT Inference. In Figure 7, we show some examples of OLAT images inferred by our neural network. Our method reproduces

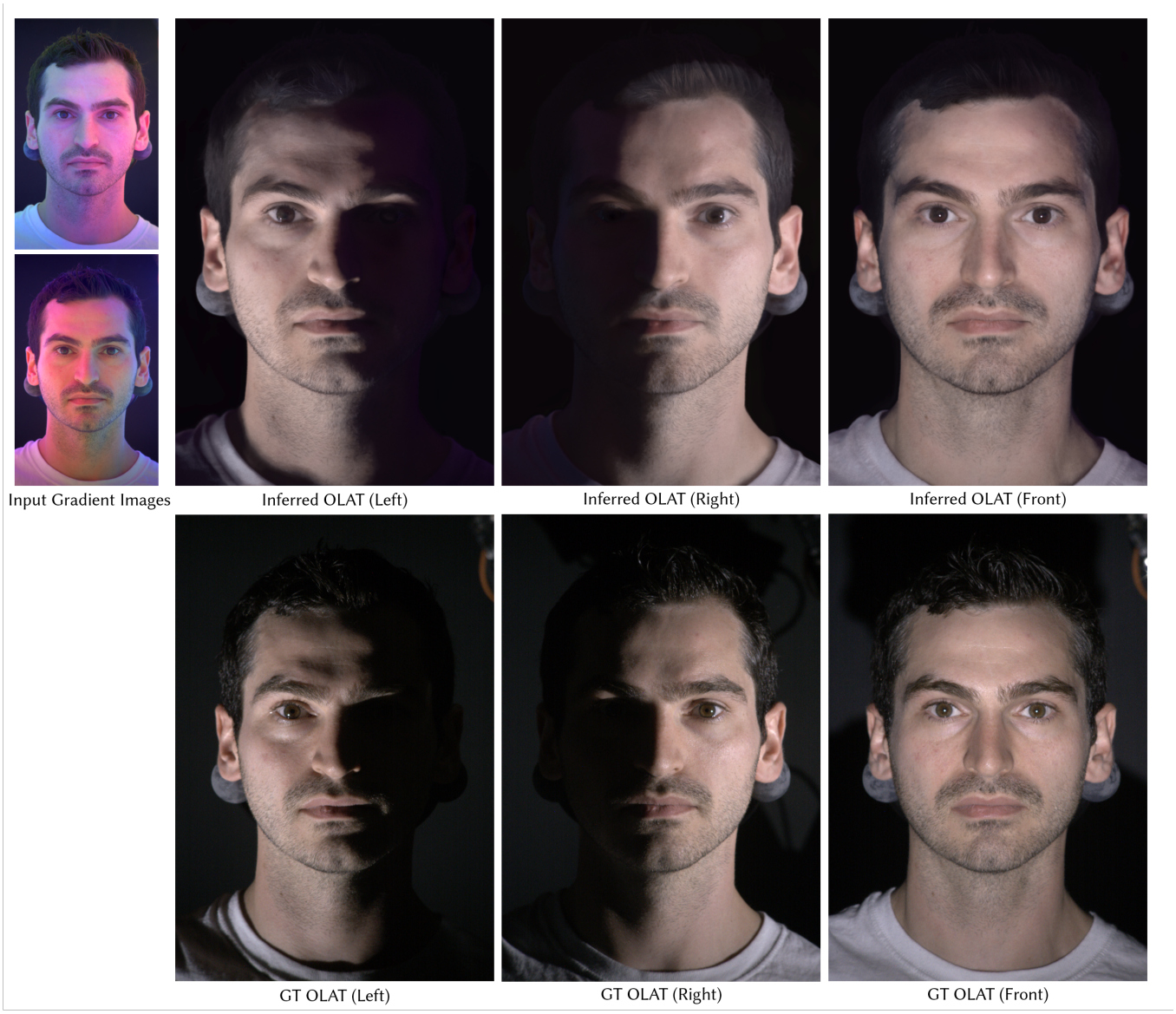

Fig. 7. **Qualitative results** – examples of gradient input images, inferred OLAT images and ground truth. Notice how high frequency details such as specularities, shadows and skin texture are correctly extracted from the gradient images.

both coarse- and high-frequency details and achieves realistic reconstructions which closely approximate the ground truth imagery. Shadows, reflections and details present in the original OLAT image data are faithfully reconstructed by our approach.

Light Direction Interpolation. In Figure 8, we show how our system is able to infer lighting directions that are not part of the dataset, demonstrating our method's ability to generalize. Our ground truth comprises OLAT images from 331 lighting directions. As such, a direct application of this discretized reflectance field to relight a sequence is limited in lighting resolution; for example, specular highlights that would be caused by lighting directions that are not part of the 331 sampled directions are not seen. In contrast, using our network we can *infer* an OLAT image for *any* lighting direction, essentially recovering a continuous reflectance field, as opposed to a discretized version obtained by the Light Stage.

Viewpoint Generalization. In Figure 9, we demonstrate generalization with respect to viewpoints. As discussed, our rig includes 9 cameras of which only 5 are used for training, leaving 4 unseen viewpoints for testing. As it can be observed, the proposed method does not introduce any specific artifacts with respect to the viewpoint. This demonstrates that the gradient images contain enough information so that the network can infer some notion of geometry.

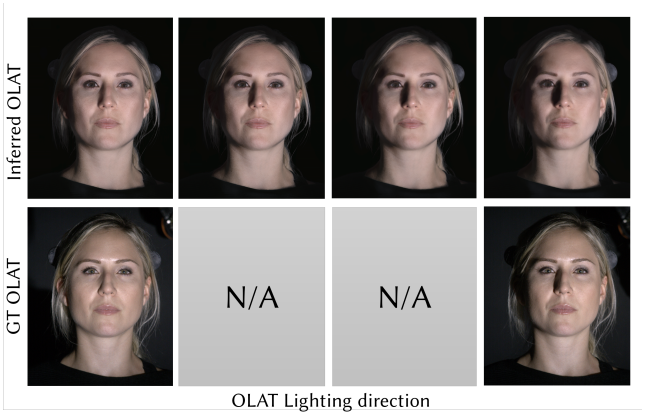

Fig. 8. **Generalization w.r.t. light direction** – Top row: inferred OLAT images; bottom row: ground truth OLAT images. The center two columns correspond to two lighting directions interpolated between the lighting directions in the far left and far right columns. Despite, not having ground truth images for these directions, due to the sparsity of lights on the light stage, our method can infer OLAT images (top middle images).

Fig. 9. **Generalization w.r.t. viewpoint** – The method is able to generalize across views, showing that the input incorporates some form of geometrical information that the network can exploit.

Comparison with the State of the Art. In Figure 10, we compare our results to the state-of-the-art approaches of Fyffe et al. [2009] and Shu et al. [2017]. The method of Fyffe et al. [2009] also takes as input two gradient illumination images, but relies on the cosine lobe reflectance model to generate images under arbitrary lighting conditions. Note this method has the drawback that it requires an additional color correction calibration to account for differences in camera color primaries vs light color primaries, whereas our proposed method simply learns to relight in whichever color space the input data is given (note that in our experiments, due to the missing color calibration step, there is a purple color cast in the

Fig. 10. **Comparison of OLAT images.** We compare OLAT images generated with different methods. See text for details.

results). Shu et al. [2017] uses a light transport approach: the method transfers lighting from a source portrait image to a target portrait image. We utilize it to transfer the lighting from on OLAT image of the source subject to the fully lit image of a target subject thereby generating a single OLAT image of the source subject. Conducting the transfer for each source OLAT image allows us to generate all the OLAT images for the target subject from a single fully lit image of the target subject. The final relighting with the environment map is based on the generated OLAT images. Notice how the results produced by the other baselines lack details where we are able to infer even extreme oblique spotlights.

Dynamic Capture. In Figure 11, we show how our method is able to handle *dynamic* subjects performing arbitrary motions and expressions. Note that no ground truth is available for these sequences as OLAT ground truth acquisition is feasible only for *static* scenes. Importantly, our network is able to generalize to facial expressions

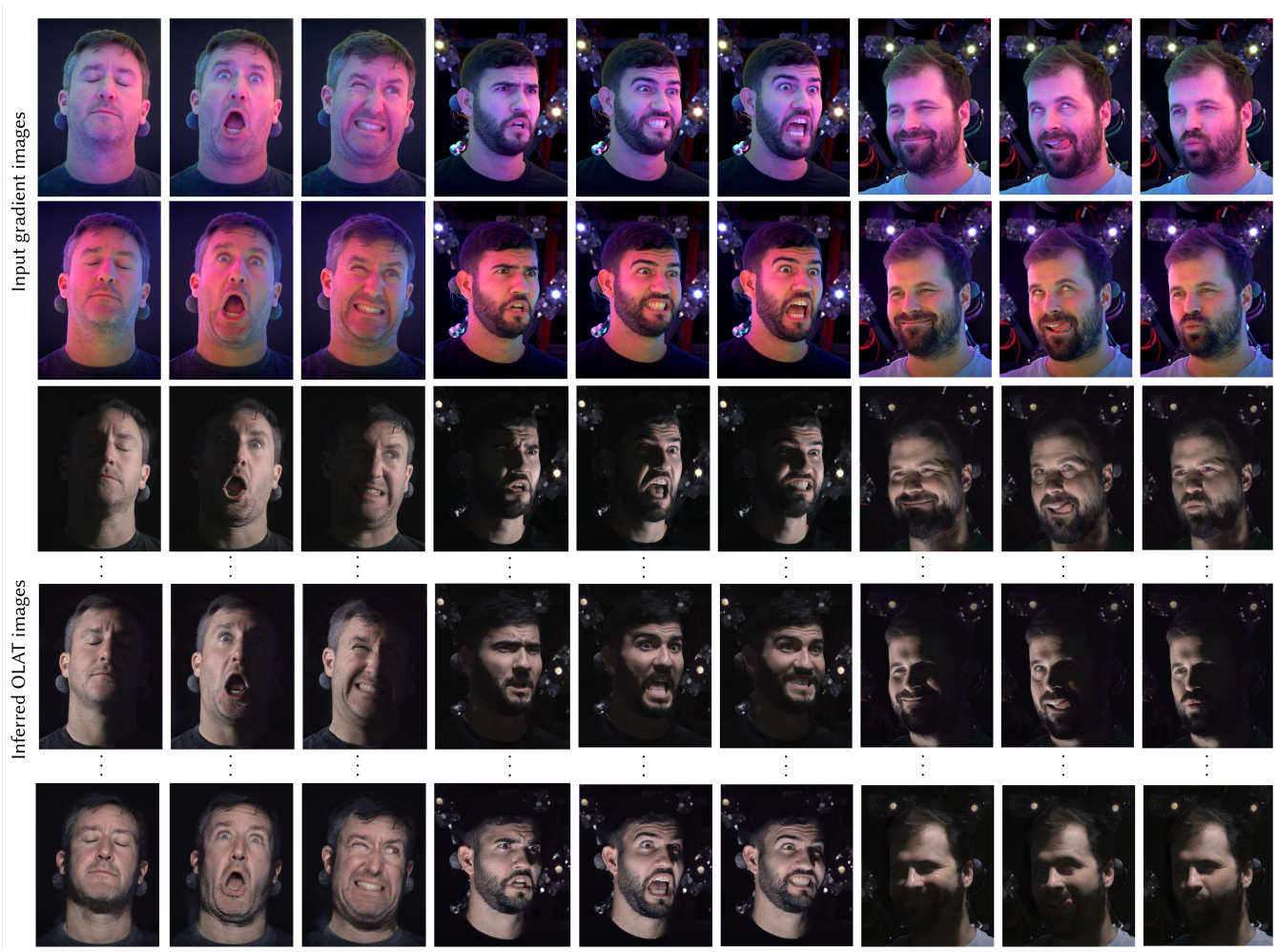

Fig. 11. **Dynamic Capture** – Top rows: input gradients for a moving subject. Bottom rows: inferred OLAT images. See more examples in our video.

which are not present in the training data, which is captured with a neutral expression. Compared to Fyffe et al. [2009], our technique produces more natural skin reflectance, a better reproduction of specular highlights, and significantly better shadows; see our **supplementary video** for more examples.

4.2 Ablation Study

In Figure 5, we evaluate the effect of each component of the proposed loss function. The proposed loss outperforms a VGG network pre-trained on ImageNet that is not able to pick up specularities, shadows and high frequency detail. Furthermore, the alignment strategy we propose in Equation 2 leads to sharper results.

In Figure 12, we explored different input modalities by training a network that takes as input a subset of the OLAT images (with wide and narrow baseline between the input lighting directions) and infers the remaining ones. Note how these networks failed to recover high-frequency shadows and texture, proving that the proposed gradient images are a better choice for the relighting task.

In Table 1, we report *quantitative* evaluations. In particular we compute metrics such as photometric error and MS-SSIM by training multiple architectures where we selectively use one or more losses.

4.3 User Study

In order to objectively gauge the quality of our predicted OLATs, we executed two user studies, one with static images, and the other with videos. In the first user study, we randomly sampled 10 ground truth OLAT images and 10 images predicted by our network for 140 users to assess. We show the users each OLAT image and with no additional information, ask them if they believe that the image is real, i.e., captured using an actual camera, or synthetic. Among the 2800 responses, participants were able to correctly identify the real or synthetic images 79% of the time, indicating that there is room for improvement of the quality of the OLAT images generated by our method. Among the wrong assessments, 50.8% of the real images were wrongly determined to be synthetic and 49.2% of the synthetic images were wrongly classified as real.

Table 1. Quantitative evaluations on test sequences of subjects. Photometric error is measured via the ℓ_1-norm. We fixed the architecture and we compared the proposed loss function with the other baselines. We obtained significantly lower MSE with the ground truth while the SSIM score is similar to the other networks. Do note that these statistical measured often do not quantify well the subjective photorealism of the images.

	Ours	$\mathcal{L}_{\text{pretrained}}$	$\mathcal{L}_{\text{specific}}$	No Alignment Loss	3-OLAT Input
Photometric Error	**808.64**	917.82	914.81	956.98	1320.51
MS-SSIM	0.222	0.217	0.216	**0.290**	0.216

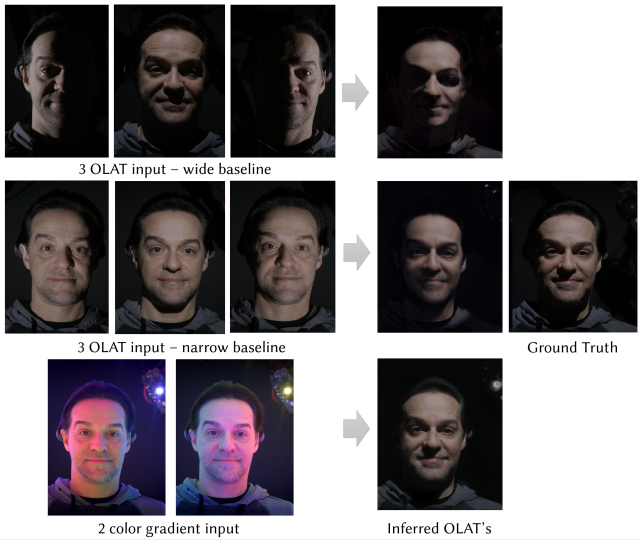

Fig. 12. **Ablation study** – Comparisons with different input modalities. Taking three OLAT as input (top) does not perform as well as the proposed gradient images (bottom).

In the second user study with 58 participants, we showed users 6 randomly selected video relighting results, to gauge whether the video appeared real or fake. Of the net 348 responses, more than 66% were marked real. This shows that even though the inferred OLATs might appear synthetic, the relighting results under high-frequency environment maps were mostly considered realistic. We also asked the participants what *cues* they used to decide if an image was real or synthetic. Common responses included issues with eye highlights, teeth texture and general blurriness.

4.4 Environment Map Based Relighting

We use our estimated reflectance field to relight an image under a new lighting environment. Given an environment map, we use the OLAT images to produce an image under the desired illumination, see Figure 3. For each OLAT image we assign an RGB scaling factor based on the intensity of the environment map as in [Debevec et al. 2000]. The final relit image is generated as a linear combination of the weighted OLAT images. In Fig. 13, we compare our results to Fyffe et al. [2009] which uses exactly the same input but requires a prior in the form of a parametric BRDF representation. In Fig. 14, we compare to the relighting capability of the state-of-the-art works of [Shu et al. 2017] and [Yamaguchi et al. 2018]. Our results show state-of-the-art quality for a method that can perform relighting of dynamic sequences *without* resorting to parametric priors.

5 CONCLUSIONS

We have proposed a novel approach for capturing high-quality time-varying 4D reflectance field at 30 FPS without requiring high-speed cameras, motion compensation, or parametric priors. This enables our method to generate relightable dynamic sequences of human actors. Our approach provides a simple and effective model and process which can be applied not only to producing high-quality time-varying 4D reflectance fields of faces, but potentially to *any* static or dynamic object or scene.

Limitations. While our results are generally realistic and are a significant improvement over existing techniques, our analysis shows that the synthetic OLAT images can in some cases be detected due to small artifacts, and occasional over-smoothing. We note that specular highlights in our estimated OLAT images are often attenuated, and at times missing on particularly noticeable regions of the face, such as eyes and teeth. This may be a result from the low-frequency nature of the color gradient illumination. This type of illumination captures enough information to infer a simplistic reflectance model. Additional (static) capture sessions could be used to better capture the reflectance of skin. In other words, an interesting question for future research is the relationship between lighting pattern configuration and reflectance information that can be captured.

As common in time-multiplexed capture, the slight misalignment between the two input images causes temporal artifacts. We expect that the continuous improvements in camera hardware will help mitigate these issues. Nonetheless, note that in environment relighting results these artifacts are averaged out by the integration process, resulting in very plausible relighting.

Future Work. It remains of interest to improve the output quality while making the algorithm more efficient. An interesting way to improve performance is to reuse features extracted early on in the network for all the OLAT directions one would predict, for example by using a late fusion technique for the lighting direction instead of the current early fusion. To further improve output quality, we can imagine exploring novel task-specific perceptual losses and neural generative techniques (GANs) to further aid in recovering high frequency details. An additional interesting avenue of research is to explore input representations beyond spherical gradient images which could lead to higher quality outputs. Such representations could be handcrafted, or even learned as part of the neural network training process. The success of our method in recovering detailed reflectance fields suggests the tantalizing possibility of high quality multi-view geometry and BRDF capture along the lines of [Ghosh et al. 2011; Ma et al. 2007] in dynamic settings.

Fig. 13. **Relighting with HDRI lighting environments** – Row 1: Ground truth OLAT base relighting, Row 2: cosine lobe relighting [Fyffe et al. 2009], Row 3: our relighting results. Notice how our method outperforms all state-of-the-art methods and comes very close to the ground truth.

ACKNOWLEDGMENTS

The authors would like to thank all participants of the lightstage recordings. We also thank the authors of Yamaguchi et al. [2018] and Shu et al. [2017] for providing the results of their methods on our data. C. Theobalt was supported by the ERC Consolidator Grants *4DRepLy* (770784). M. Zollhöfer was supported by the Max Planck Center for Visual Computing and Communications.

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. *SIGGRAPH Asia* (2016).

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

Pascal Bérard, Derek Bradley, Markus Gross, and Thabo Beeler. 2016. Lightweight Eye Capture Using a Parametric Model. *ACM Trans. Graph. (Proc. SIGGRAPH)* 35, 4, Article 117 (July 2016), 12 pages.

Amit Bermano, Thabo Beeler, Yeara Kozlov, Derek Bradley, Bernd Bickel, and Markus Gross. 2015. Detailed Spatio-temporal Reconstruction of Eyelids. *ACM Trans. Graph.* (2015).

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

Qifeng Chen and Vladlen Koltun. 2017. Photographic Image Synthesis with Cascaded Refinement Networks. *ICCV* (2017).

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)*.

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 '06)*.

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).

Graham Fyffe and Paul Debevec. 2015. Single-Shot Reflectance Measurement from Polarized Color Gradient Illumination. In *ICCP*.

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).

Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview Face Capture Using Polarized Spherical Gradient Illumination. *ACM Trans. Graph.* (2011).

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

Tim Hawkins, Andreas Wenger, Chris Tchou, Andrew Gardner, Fredrik Göransson, and Paul E Debevec. 2004. Animatable Facial Reflectance Fields. *Rendering Techniques* (2004).

Fig. 14. **Relighting Comparisons** – Comparisons with other state of art approaches. Our method outperforms the latest work in the field.

Liwen Hu, Chongyang Ma, Linjie Luo, and Hao Li. 2015. Single-View Hair Modeling Using A Hairstyle Database. *ACM Trans. on Graphics (SIGGRAPH)* (2015).

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.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2016. Image-to-Image Translation with Conditional Adversarial Networks. *arxiv* (2016).

Yoshihiro Kanamori and Yuki Endo. 2018. Relighting Humans: Occlusion-aware Inverse Rendering for Full-body Human Images. In *SIGGRAPH Asia*. ACM.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* (2014).

Weicheng Kuo, Christian Häne, Esther Yuh, Pratik Mukherjee, and Jitendra Malik. 2018. PatchFCN for Intracranial Hemorrhage Detection. *arXiv preprint arXiv:1806.03265* (2018).

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).

Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. 2018a. Materials for Masses: SVBRDF Acquisition with a Single Mobile Phone Image. In *ECCV (Lecture Notes in Computer Science)*. Springer.

Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. 2018b. Learning to Reconstruct Shape and Spatially-varying Reflectance from a Single Image. In *SSIGGRAPH Asia*.

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).

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

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, 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.

Oliver Nalbach, Elena Arabadzhiyska, Dushyant Mehta, Hans-Peter Seidel, and Tobias Ritschel. 2017. Deep Shading: Convolutional Neural Networks for Screen-Space Shading. 36, 4 (2017).

Giljoo Nam, Joo Ho Lee, Diego Gutierrez, and Min H. Kim. 2018. Practical SVBRDF Acquisition of 3D Objects with Unstructured Flash Photography. In *SIGGRAPH Asia*.

Pieter Peers, Naoki Tamura, Wojciech Matusik, and Paul Debevec. 2007. Post-production Facial Performance Relighting Using Reflectance Transfer. In *ACM SIGGRAPH 2007 Papers (SIGGRAPH '07)*. ACM, Article 52.

Peiran Ren, Yue Dong, Stephen Lin, Xin Tong, and Baining Guo. 2015. Image Based Relighting Using Neural Networks. *ACM Trans. Graph.* 34, 4 (July 2015).

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI* (2015).

Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. 2017. Photorealistic Facial Texture Inference Using Deep Neural Networks. In *CVPR*. IEEE Computer Society, 2326–2335.

YiChang Shih, Sylvain Paris, Connelly Barnes, William T. Freeman, and Frédo Durand. 2014. Style Transfer for Headshot Portraits. *ACM Trans. Graph.* (2014).

Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. 2017. Portrait Lighting Transfer Using a Mass Transport Approach. *ACM Trans. Graph.* (2017).

K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

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*.

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. 2005. Performance Relighting and Reflectance Transformation with Time-multiplexed Illumination. In *ACM SIGGRAPH 2005 Papers (SIGGRAPH '05)*.

Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. 2018. Deep image-based relighting from optimal sparse samples. *ACM Trans. on Graphics* (2018).

Shuco Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-fidelity Facial Reflectance and Geometry Inference from an Unconstrained Image. *ACM Trans. Graph.* 37, 4, Article 162 (July 2018).

Meng Zhang, Menglei Chai, Hongzhi Wu, Hao Yang, and Kun Zhou. 2017. A Data-driven Approach to Four-view Image-based Hair Modeling. *ACM ToG* 36 (2017).

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. *CVPR* (2018).

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *ICCV*.