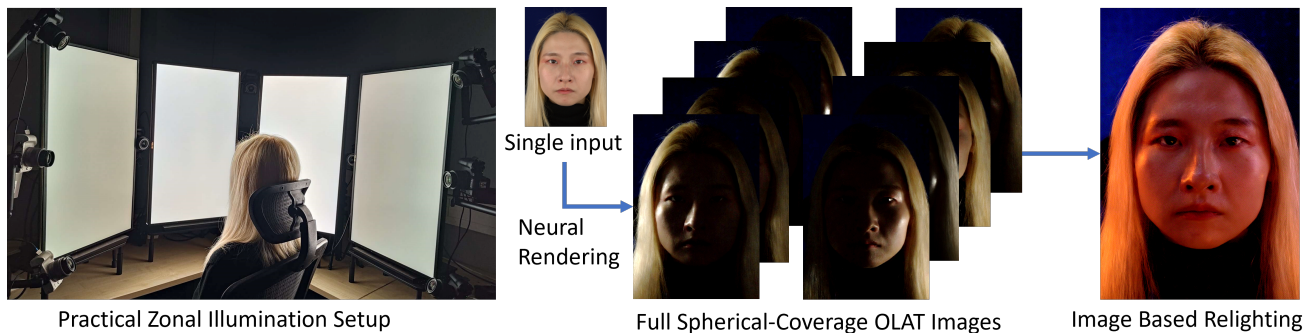# High Quality Neural Relighting using Practical Zonal Illumination

A. Lin[1,2] 🔵   Y. Lin [2] 🔵   X. Li[1,2] 🔵   A. Ghosh[1,2] 🔵

[1]Imperial College London, UK
[2]Lumirithmic Ltd.

Figure 1: *Given an image of a face lit with uniform frontal illumination, our method generates smooth OLAT images extending beyond the zonal illumination field of the capture setup. Our approach achieves light-stage quality relighting with a practical consumer-available hardware setup.*

Abstract

We present a method for high-quality image-based relighting using a practical limited zonal illumination field. Our setup can be implemented with commodity components with no dedicated hardware. We employ a set of desktop monitors to illuminate a subject from a near-hemispherical zone and record One-Light-At-A-Time (OLAT) images from multiple viewpoints. We further extrapolate sampling of incident illumination directions beyond the frontal coverage of the monitors by repeating OLAT captures with the subject rotation in relation to the capture setup. Finally, we train our proposed skip-assisted autoencoder and latent diffusion based generative method to learn a high-quality continuous representation of the reflectance function without requiring explicit alignment of the data captured from various viewpoints. This method enables smooth lighting animation for high-frequency reflectance functions and effectively manages to extend incident lighting beyond the practical capture setup's illumination zone. Compared to state-of-the-art methods, our approach achieves superior image-based relighting results, capturing finer skin pore details and extending to passive performance video relighting.

CCS Concepts
• Computing methodologies → Reflectance modeling; Image-based rendering; Computational photography;

1. Introduction

Photo-realistic portrait relighting has long been an active research topic and has many direct applications in visual effects. This problem is, however, difficult to solve since facial relighting depends on light interaction with various materials and complex geometry. Standard model-based approaches that estimate reflectance (BRDFs) and shape (geometry) are generally limited by the accuracy and complexity of the underlying model. Image-based relighting approaches overcome these limitations by densely capturing the reflectance field through One-Light-At-A-Time (OLAT) images.

While image-based approaches can produce outstanding photo-realistic results, OLAT captures typically require dedicated hardware such as a light stage [DHT*00]. More recently, various sophisticated light stages have been built with higher lighting resolution and faster cameras to capture better data, including supporting dynamic performance capture. However, the cost and complexity of building and operating a light stage is very high, necessitating specialized engineering and professional operation, thereby making it inaccessible to non-expert users.

There are two main advantages of a high-end light stage for

image-based relighting. First, a light-stage setup efficiently produces basis OLAT illumination from a full sphere of lighting directions. Second, high-end light stages nowadays can synchronize OLAT illumination and high-speed cameras to capture data at video rates [MHP*19]. This minimizes movement of subjects between measurements to a point which any minor motion can be largely fixed using optical-flow alignment.

While light stages offers great advantages for reflectance acquisition, the complexity in building such device makes it impractical to be implemented outside of a dedicated laboratory. On the other hand, smartphone-based approaches for relighting struggle to match the quality that a light stage can produce. Recent state-of-the-art relighting results with smartphone images still require data obtained from the light stage as training data [PEL*21].

In this work we demonstrate light-stage quality relighting with a much simpler consumer-grade hardware setup using a combination of an appropriate capture process and a well-designed neural architecture. We employ a practical setup proposed by Lattas et al. [LLK*22], which can be easily implemented by plugging a few consumer-grade cameras and 4 LCD monitors to a computer. Specifically, we propose a generative method that generates a full spherical, continuous reflectance field for a desired subject, using a limited zonal illumination field setup built using off-the-shelf hardware, as illustrated in Fig. 2a. Our method works with data captured under slower capturing speed (e.g. 3 fps), thus eliminating the need for high-speed cameras. Synchronization at lower frame rate is also much easier as the cameras' shutter can be triggered between each OLATs and there are sufficient time between each OLATs for the cameras to record an image.

Due to the limited frontal illumination coverage provided by our setup, we require the subject to be rotated a few times in relation to the capture setup, repeating OLAT captures per rotation. This capture process enables extensive incidental illumination coverage with respect to the subject. Subsequently, we train our novel neural architecture to both handle misalignments between photographs and extrapolate for full spherical reflectance field sampling.

Although our method is simple to set up and operate, its technical complexity resides in the network design and training process. Our latent diffusion network is specifically designed and trained to handle both paired (input and OLAT pair) and unpaired data (OLAT image only). Additionally, our decoder incorporates a specially designed perceptual loss that effectively addresses misalignment in the paired data caused by natural body movement. Aside from reducing the hardware requirements from a light stage to a setup built with consumer components, our method requires only one reference image of a subject (uniformly lit) at inference time to facilitate high quality image-based relighting. The requirement of only one image thus enables video performance relighting with pronounced facial movement that can be passively recorded using just a consumer video camera. Our main contributions can be summarized as:

- A practical capture method that enables high-quality reflectance field estimation based on off-the-shelf consumer hardware.
- A novel loss function that addresses the alignment issues caused by movements between a pair of a given OLAT image and a reference image.

- A skip-assisted decoder that utilizes features from the reference image to reconstruct a target OLAT image with finer details and sharper highlights.
- A conditionally generative formulation of reflectance field that efficiently utilizes paired and unpaired data to account for lack of alignment of captured data due to natural subject motion.

2. Related Work

After the light stage was first utilized for reflectance capture [DHT*00], the light stage has become a standard device in high quality capture and relighting of human subjects. Data captured from a light stage enables photo-realistic applications such as transferring reflectance from an unseen subject [PTMD07] and even relighting dynamic scenes [CEJ*06]. Even for SVBRDF acquisitions a light stage is essential for providing the necessary illumination patterns for high quality captures. For example, [MHP*07] utilized polarized spherical gradients to efficiently capture specular and diffuse normal maps. This work is later extended by [GFT*11] for multiview consistent facial captures. These works all require a dedicated light stage in order to provide the needed illumination patterns.

Deep Learning Enabled by Light Stage

More recently, with the advances of deep learning, the ability of a light stage to capture large amounts of high quality data makes it almost a prerequisite for reflectance capture and modelling. Meka et al. [MHP*19] utilized a light stage to capture gradient patterns and OLAT images of subjects under multiple expressions to enable dynamic video relighting. Sun et al. [SXZ*20] used the same setup to acquire data for learning a continuous high-frequency OLAT image synthesis. The acquired OLAT images can also be used to synthesize virtually unlimited amount of data with different environmental illumination, providing a deep learning architecture with sufficient data to learn arbitrary relighting from single portrait images [SBT*19; PEL*21; NLML20]. In addition, light stages have also been demonstrated to provide sufficient training data for whole body captures [GLD*19; MPH*20; ZFT*21] and creating photo-realistic relightable avatars [LBZ*20; BLS*21; YZF*23; SSS*23].

Approaches without Controlled Illumination

While light stages have driven a lot of state-of-the-art relighting methods, such a specialized device is difficult to build and operate, rendering it inaccessible to most users. A number of works have attempted reflectance acquisition and relighting without access to a light stage. Wang et al. [WHZZ23] proposed using the sun as a primary light source and rotating the subject to measure 'OLAT-like' photographs. Han et al. [HLX23] used a sequence of images illuminated with single co-located smart phone flashlight for geometry and reflectance estimation. Azinovic´ et al. [AMH*23] places polarizers on a smartphone's flash light and camera to capture both parallel and cross polarized images for facial appearance acquisition. Chen et al. [CL22] decomposed geometry and reflectance into BRDF and occlusion fields using a neural field as a 4D representation to achieve dynamic human relighting without explicitly capturing OLAT images. These works, however, rely on explicit shape

and BRDF models and are limited in the quality of relighting results. Sengupta et al. [SCKS21] trained a deep network for direct neural relighting based on images illuminated by a video sequence lit on screen. Their method, however, does not extrapolate the limited zonal illumination from a single screen and can only relight the subject under different screen patterns.

Devices with Limited Zonal Illumination Field

A more practical alternative to a light stage is to build an illumination device with limited zonal illumination field. This approach finds a middle ground between building a complex hardware and relying on uncontrolled lighting for reflectance capture. In fact, capture setups providing only frontal illumination with multiple cameras are increasingly common in geometry research where a certain amount of relighting is desired [XZC*23; SBL*23; XCW*23]. Limited zonal illumination devices have also been shown to be sufficient in model-based SVBRDF captures [LLK*22] and has also been used for lighting reproduction [OTII07]. These works, however, do not overcome the challenge of the limited zonal illumination and would either rely on model-based relighting or can only relight a subject within the limited zonal illumination field.

3. Method

While much neural relighting work has been done using a light stage, most work has not investigated full-coverage relighting with a limited zonal illumination field device and consumer-grade cameras. These types of setups, however, are more practical than a light stage and quite commonly used [SBL*23; XZC*23; LLK*22]. While a limited zonal illumination field device as shown in Figure 2a can only provide limited One-Light-At-A-Time (OLAT) lighting basis, by placing multiple cameras across the zonal illumination field device and having the subject face each camera separately, it is possible to capture a set of photographs that contains the images of the subject being lit from all lighting directions as illustrated in Figure 2b.

The data acquired with this approach differs fundamentally from data acquired using a light stage. While OLAT data obtained from a light stage can be sufficiently aligned to a set of reference images, facilitating either direct image relighting or deep learning approaches that predict an OLAT image based on a set of reference images [MHP*19], a limited zonal illumination setup presents challenges. Capturing different lighting directions by facing different cameras inherently creates unaligned images for each lighting direction. Moreover, the slower capture process leads to sporadic subject movements between photographs, making it impossible to perfectly align these photographs to any reference image, even with optical flow alignment. Consequently, we end up with a dataset comprising unaligned OLAT images and their corresponding reference 'Full On' images. In the remainder of this section, we will discuss how to capture such a dataset and train a reflectance model using this data.
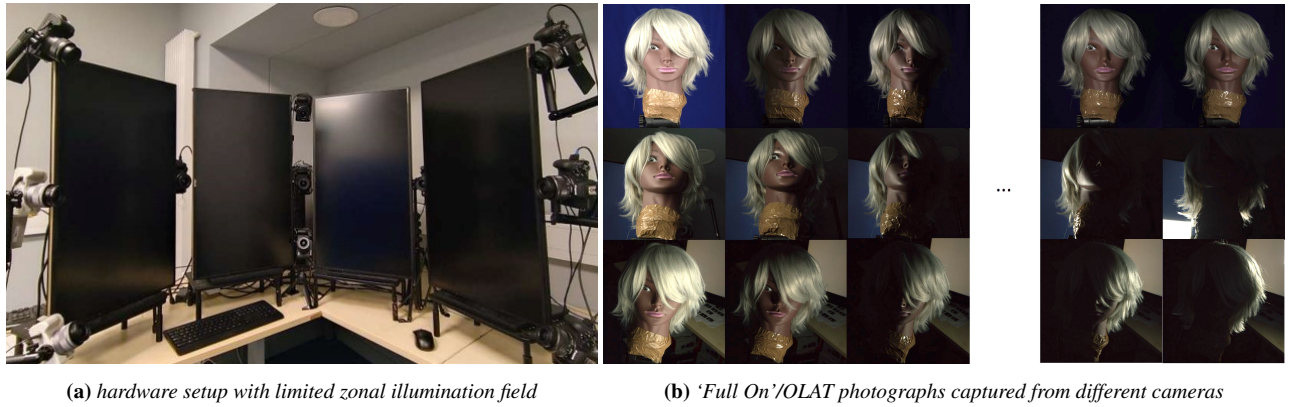
3.1. Hardware Setup and Data Acquisition

Our setup is illustrated in Figure 2. The setup consists of four vertically placed screens covering the frontal zone of the subject, inspired by Lattas et al. [2022]. Eleven cameras are positioned around the screens: three cameras in the leftmost column, three in the middle column, three in the rightmost column, and one camera each in the middle left and middle right columns, respectively. Each monitor is divided into eight equally sized blocks, composing 32 different basis illumination blocks across all four monitors. We perform One Light at a Time (OLAT) captures while each basis block illuminates in sequence. Between each OLAT captures in the sequence, there is also a 'Full On' image taken with the screens fully illuminating, to make each OLAT image become an OLAT/'Full On' pair. The above process is repeated 11 times, with the subject facing each of the 11 cameras in turn. The central five cameras simultaneously capture images while the subject is facing any among the five cameras, and the three outermost cameras on either side are also simultaneously capturing when the subject is facing in that direction. After capturing the subject, a mirrorball is used to capture the light directions for each OLAT pattern with respect to each camera.

Due to the natural movement of the human body, achieving pixel-perfect alignment between neighbouring 'Full On' images and OLAT (One-Light-At-A-Time) images is rarely possible. To address this, an optical flow step is typically applied. However, natural body movements can lead to significant pixel misalignment, which optical flow methods do not precisely correct, potentially degrading the final result if an image-to-image translation network is trained directly on these images. Instead, we compute the optical-flow field with the method proposed by Anderson et al. [2016] to detect significant movements, empirically set to a threshold of more than 8 pixels. If significant movement is detected, we discard the 'Full On' image. Otherwise, the intermediate OLAT image is considered sufficiently aligned with the 'Full On' image, though not exactly on a pixel level. Our experiments show that in a typical capture session, 60% of the OLAT images have tolerable movement. However, this percentage largely depends on the subject.

3.2. Latent Conditional-Generative Approach to Reflectance Modelling

With our capture setup we can obtain two classes of acquired data: 1. OLAT/'Full On' image pairs, and 2. single OLAT images when the 'Full On' image is discarded. Typical image-to-image translation approaches do not efficiently use such captured data as they require perfectly paired data. Therefore, we opt for a latent conditional generative approach to model reflectance. Our conditional generative model learns to generate random OLAT images by training on all captured OLAT images, then additionally learns to generate OLAT images that are aligned to a reference input by using the OLAT/'Full On' image pairs, with the 'Full On' image used as the condition. A high-level illustration of our training method is shown in Figure 3. For all captured images of resolution $2304 \times 3328 \times 3$, a Vector-Quantized Variational Autoencoder (VQ-VAE) [VV*17] is trained to encode all images into a latent image of resolution $288 \times 416 \times 32$. We assume that the paired 'Full On' and OLAT images that have subject movement within 8 pixels between the paired image have corresponding latent images that are sufficiently

(a) *hardware setup with limited zonal illumination field*   **(b)** *'Full On'/OLAT photographs captured from different cameras*

Figure 2: *While a limited zonal illumination cannot directly capture OLAT photographs spanning the entire lighting coverage, by placing multiple cameras in different positions and facing each cameras separately while capturing, it is possible to capture a set of data that represents an extended zonal illumination field, though the images from different cameras are not corresponded (aligned) to each other.*
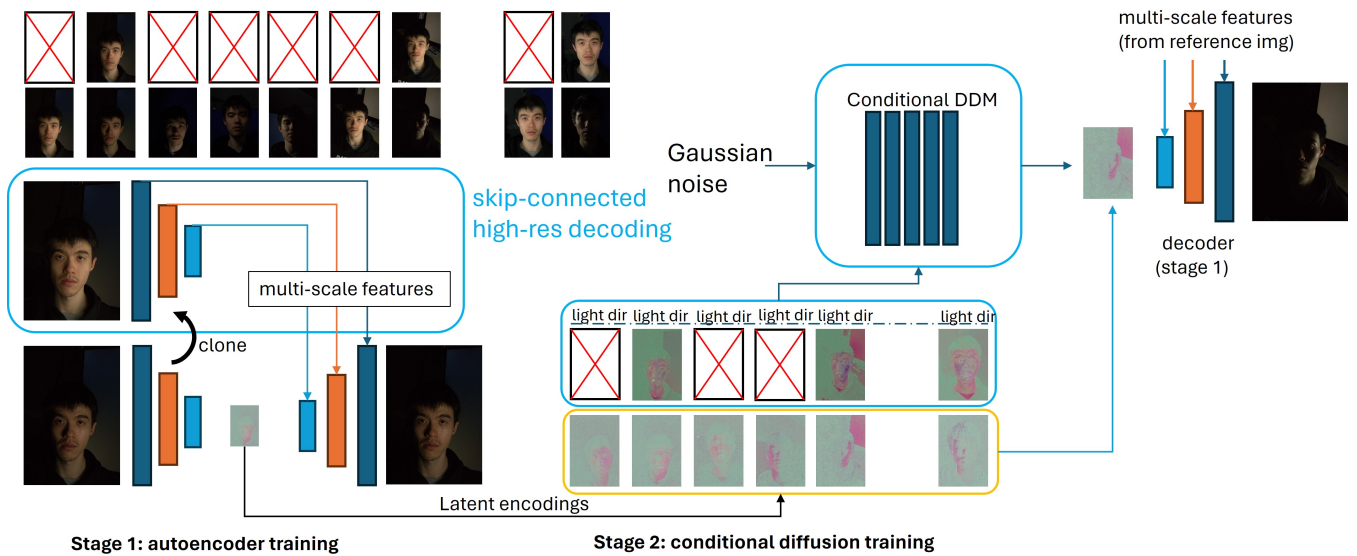
Figure 3: *Data Training Pipeline. Our method takes a set of OLAT photos in which only part of the OLAT photos have a reference photo. We train a skip-assisted high-resolution autoencoder with a vector-quantized latent space [VV*17]. When a reference image is present, multi-scale features from the reference image are fed to the decoder in a UNet-like fashion [RFB15] to assist the decoding of OLAT images. The details of training the autoencoder is illustrated in algorithm 1. After training the autoencoder, latent features of the OLAT and reference photographs, as well as the corresponding lighting directions are used to train a conditional diffusion denoising process, resulting in a generator that generates OLAT images given a reference well-lit photograph and a target lighting direction.*

aligned, since the latent space is scaled down to 1/8 from its original resolution. This eliminates the need for explicit alignment at the original image resolution.

Thereafter, the latent images are used to train a deterministic conditional Denoising Diffusion Model, following [HBC23]. The light direction is used as the condition given to the model, along with the reference latent image as an optional condition. This method allows the model to generalize lighting effects from OLAT images without any input image, adapting to scenarios where the 'Full On' image is discarded. The final model is trained to infer a

plausible OLAT latent image given a light direction, with or without a conditional 'FULL ON' latent image. The inferred OLAT latent image is then passed through the pre-trained decoder to obtain the final result.

3.3. Multiscale Skip-Assisted High Res Decoding

While encoding images into a smaller latent space ensures latent alignment, high-frequency details are inevitably lost during the encoding process. Our main contribution is the introduction of multiscale guided decoding, which produces high-resolution decoded

OLAT images guided by a loosely aligned 'Full On' image. Adapting the U-Net skip connections proposed in [RFB15], we encode both the reference 'Full On' image and the OLAT image. Then, we pass the multi-scale features from each layer throughout encoding the 'Full On' image to the corresponding layers in the decoder, which decodes to the OLAT image. Instead of simply concatenating the skip features with the decoded features, we apply an attention mechanism utilizing the additive attention gates from [OSF*18]. Contrast to directly feeding reference images through additional convolution layers such as [ZFC*23], our extracted features contain more high frequency information that is important in reconstructing sharp details. This additional information during decoding facilitates the reconstruction of high-quality OLAT images with enhanced photorealism.

Misalignment-aware Perceptual Loss

Decoding an image with multi-scale features from a reference image without pixel exact alignment to the target image reintroduces the alignment issue. While contextual loss functions have been proposed to deal with non-aligned data [MTZ18], these address global contextual features and requires heavy computation. In comparison, our task call for a loss function that match local features with slight misalignment. To address this, we introduce our misalignment-aware perceptual loss.

A typical perceptual loss employs a pre-trained image classifier, such as VGG [SZ14], to extract features from both images, upon which the perceptual loss is calculated. We utilize the fact that the VGG network reduces the feature size by a factor of 2 at each incremental depth. For two images where the misalignment is constrained within 2^n pixels, we can assume that the features downscaled by a factor of 2^n are sufficiently aligned. We further hypothesize that the features downscaled by a factor of $2^m, m < n$, experience a misalignment of 2^{n-m} pixels. This understanding forms the basis for our proposed perceptual loss calculation.

Given two images I_{gt} and I_{pred} with a misaligned kernel size 2^n, our perceptual loss is expressed as:

$$P^n = \sum_{k \in \Omega} (F_k(I_{gt}) - F_k(I_{pred}))^2, F_k(I) = maxpool(f(I)_k, max(0, 2^{n-m_k})) \tag{1}$$

where $f(I)_k$ denotes the feature grid extracted from image I from the kth convolution layer, and 2^{m_k} is the scale by which the feature grid is downsampled with respect to the original image resolution. Figure 4 illustrates our proposed misalignment-aware perceptual loss.

Training the Skip-Assisted Encoder Decoder

We train the encoder and the skip-assisted decoder together, following the training schema outlined below:

- For OLAT/'Full On' paired images, both images are encoded, and the OLAT image is decoded with skip assistance as described in Section 3.3. The features used for skip assistance is considered to be misaligned from the OLAT image.
- For a single OLAT input, the input is encoded and decoded without skip assistance.
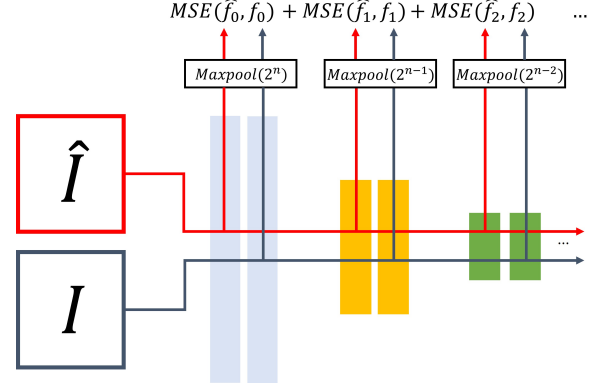
Figure 4: *Illustration of our misalignment-aware perceptual loss. We show that a simple modification of calculating max-pooling before taking the mean square error is sufficient to make a standard perceptual loss robust to constrained misalignment in images*

- For a random probability of 50% the OLAT image is decoded with skip assistance using the features extracted from the OLAT image itself. In this case the skip assisting features are perfectly aligned with the OLAT image.

We propose two loss functions, each associated with the aligned and misaligned data, respectively:

$$L_{aligned}(\hat{I}, I) = (\hat{I} - I)^2 + w_1 P^0_{pretrained}(\hat{I}, I) + w_2 P^0_{custom}(\hat{I}, I) \tag{2}$$

$$L_{misaligned}(\hat{I}, I) = P^n_{pretrained}(\hat{I}, I) + w P^n_{custom}(\hat{I}, I). \tag{3}$$

We utilize both pre-trained and custom perceptual loss as proposed by Meka et al. [MHP*19]. For $L_{unpaired}$, we choose $n = 3$ as per our threshold for movement detection. In terms of data augmentation, we randomly use only the OLAT image in the paired data and train with it as unpaired data. For unpaired OLAT images, we randomly use the OLAT image itself as the 'Full On' image, effectively making it a paired training scenario. The training process is illustrated in the following pseudocode:

Implementation Details

Our skip-assisted encoder decoder follows the implementation of Van et al. [VV*17] with 3 down-sampling convolutional layers. 4 residual layers are used in the residual stack, with all hidden convolutional layers having 128 channels. For vector quantization we have 1024 independent embedding vectors, each having a dimension of 32. Our skip-assested encoder decoder is trained over 200 epochs with an Adam optimizer [KB14] using an intial learning rate of 10^{-5}. For our latent diffusion model we modify the model proposed by Nichol et al. [ND21] to take a latent image and light direction vector as conditional input, the remaining structure stays the same as the original implementation. The diffusion model is also trained over 200 epochs with an initial learning rate of 10^{-5} using the Adam optimizer.

ALGORITHM 1: Loss function inside the training loop

$target\_im \leftarrow$ sample_training_data;
$embedding \leftarrow encoder(target\_im)$;
if $random\_sample([True, False])$ **then**
    # decode without assisted features
    $pred \leftarrow decoder(embedding, features = None)$;
    $recon\_loss = L_{aligned}(pred, target\_im)$;
else
    # decode using assisted features
    if $aligned\_reference\_available$ **then**
        $ref\_im \leftarrow$ get_reference_image$(target\_im)$;
        $L = L_{misaligned}$;
    else
        $ref\_im \leftarrow target\_im$;
        $L = L_{aligned}$;
    end
    $feat \leftarrow encoder.get\_features(ref\_im)$;
    $pred \leftarrow decoder(embedding, features = feat)$;
    $recon\_loss \leftarrow L(pred, target\_im)$;
end
vqvae uses embedding loss for vector quantization
$embedding\_loss \leftarrow$
$encoder.get\_quantized\_loss(embedding)$
$total\_loss \leftarrow recon\_loss + embedding\_loss$;

---

4. Results and Analysis

In this section, we show and analyze the results generated with our proposed method. We also compare our results with the method proposed in [MHP*19]. We captured 10 subjects, encompassing various races and both genders. Additionally, we captured two gradient illumination images before our OLAT/'Full On' paired image captures to produce the comparison results using the method described in Meka et al[MHP*19]. The subsequent subsections present both qualitative and quantitative analyzes of our proposed approach.

4.1. Expanding Zone of Illumination from Single Image

Figure 5 presents comparison results between actual photographs (top row) and generated OLAT images (bottom row). The generation process exclusively uses a single input photograph. The comparative results demonstrate that our method can generalize effectively in terms of pose, expression, and lighting directions for the desired subject. The inferred OLAT images, derived from this process, act as the foundation for high-quality, image-based relighting.

4.2. Relighting with Extended Zone of Illumination

Figure 6 shows the high quality relighting results that are achievable with our method. To illustrate the necessity of expanding the zone of illumination, we compare the relighting results of our method with those using only limited zonal illumination in Figure 7. For this comparison, a static mannequin is utilized, and we capture additional OLAT images of this mannequin in a light stage

equipped with 168 lights. Our results demonstrate that in scenarios where side or back lighting is predominant, our method successfully generates very plausible relighting results, while limited zonal illumination fails to replicate similar lighting effects. Note that the color difference between the light stage result (figure 7 a) and the monitor illumination-based results (figure 7 b, c) is due to the difference in the spectral distributions of the two illumination setups [LYL*16]. Figure 8 further underscores the necessity of extending the zone of illumination in limited illumination devices. Limited zonal illumination does not capture the effects of back lighting (figure 8 a), and using the lighting components of a Voroni grid creates side lighting effects instead of true back lighting (figure 8 b). Additional relighting results are shown in Figure 12 and the accompanying video.

4.3. Passive Video Performance Capture

Our method also empowers high-quality dynamic video performance capture and relighting in the absence of a lightstage. By training the network with our proposed OLAT/'Full On' images of the same subject under multiple static expressions, we obtain a trained newtork that infers OLAT images for each frame of a video sequence with the subject lit under a 'Full On' lighting. Figure 9 shows the high quality relighting results for such a performance video obtained with a practical setup (see accompanying video for more results).

4.4. Comparisons

While a number of subject relighting works have been proposed over the years [PEL*21; SBT*19], these work uses data of multiple subjects for their training process. In constrast, our method focuses on relighting a single subject without acquiring data from other subjects, thus we focus on comparing we compare our method with Deep Reflectance Fields [MHP*19], which represents the state-of-the-art method for per-subject reflectance capture. Following the approach of Meka et al., we utilize the 'Full On' images to estimate the optical flow between images and attempt to align the OLAT images with the color-multiplexed gradient images. In Figure 10, it can be observed that the method by Meka et al. produces results that are much blurrier and lack fine-scale texture and highlights. This issue arises from the imperfect alignment between the OLAT images and the multiplexed images after applying optical flow. Although [MHP*19] proposed a slide-pooling loss to offset the image patch for optimal alignment, this strategy proves insufficient when movements involve rotation and scaling, scenarios more common in the natural body wobbles. It is also important to note that Meka et al.'s method requires two input images (color gradients), whereas our method requires only one ('Full On') for generating the OLATs, making it suitable for passive performance capture relighting (see Section 4.3). Table 1 provides a quantitative comparison between our method and [MHP*19]. Our method surpasses [MHP*19] in perceptual similarity [ZIE*18] when tested with both pre-trained AlexNet and VGG networks. While our method scores lower in structural similarity, we acknowledge that this metric does not fully capture the quality of subjective photo-realism.

Figure 5: *Top row: Captured OLAT images. Bottom row: OLAT images generated with our method using the input image. All generated images do not have aligned OLAT captures present in the training data. The input image features a different pose and a slightly different expression and hairstyle; therefore, the actual generated OLAT results differ from the reference images. However, the network successfully learns to generate highly realistic lighting effects and synthesize novel OLAT images with highly believable shadows.*
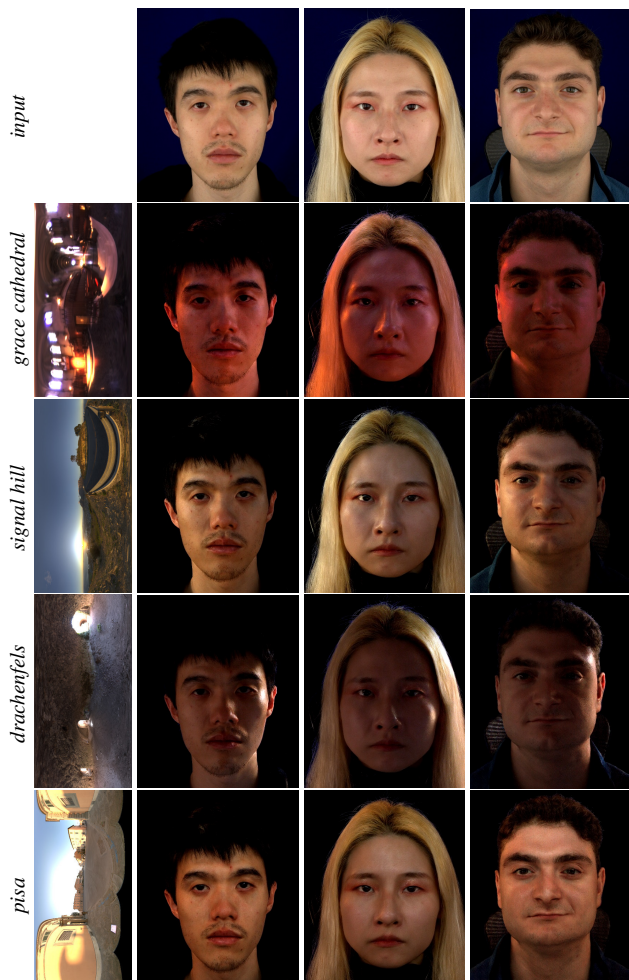
**Figure 6:** *Results of relighting under various HDR environments.*

(a) *environment map* **(b)** *light-stage (reference)* **(c)** *limited zonal relighting* **(d)** *full-coverage relighting (ours)*

Figure 7: *Comparing the relighting results between the partial illumination field (c) and our extended zonal illumination (d), our method (d) produces back lighting effects the matches the light-stage reference (a) for environments where back lighting is dominant. This effect is otherwise not reproducible with limited zonal illumination devices (b).*

4.5. Ablation Study

In Figure 11, we examine the different effects of our proposed novel components in network design. Our skip-assisted decoding block demonstrates superior performance compared to standard autoencoders that do not utilize multi-scale features from a reference image. The skip-assisted decoding strategy results in images with sharper textures and highlights. When comparing our loss function to the slide-pooling loss [MHP*19] for handling misalignment, our loss function proves to be more robust, producing results with

(a) *limited zonal relighting*　(b) *approx. with front OLATs*　(c) *full-coverage relighting (ours)*
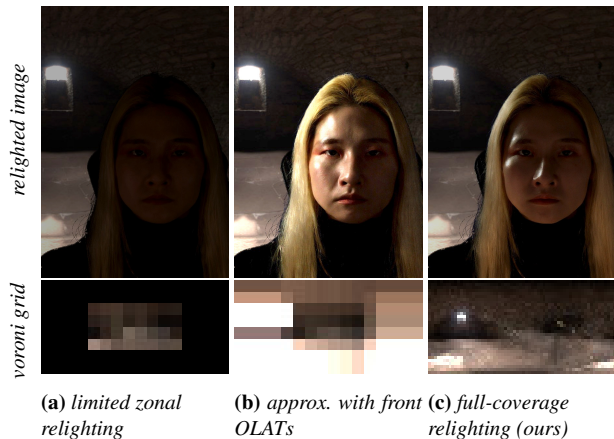
Figure 8: *Our method (c) recreates back-lighting effects that cannot be reproduced with only sampling frontal illumination (a) or even integrating the back-lights into the side-most OLATs (b). While (b) does produce some side-lighting effects, the variation of back-lighting gets integrated to side OLATs and thus cannot reflected in the relighted image. The background environment map is manually matted for better visualization. This effect is more prominent when relighting a video under rotating environment maps (see supplemental video).*

Table 1: *Quantitative comparison of our reconstructed result with the method proposed by Deep Reflectance Field [MHP*19]. We provide results for structural similarity (SSIM) and perceptual similarity (LPIPS) [ZIE*18] using both AlexNet and VGG as the backbone ($LPIPS_a$ and $LPIPS_v$). Note that perceptual similarity is much more correlated to the perceptual quality of photo-realism.*

Subjects	Subject 1			Subject 2		
	SSIM	*LPIPS$_a$*	*LPIPS$_v$*	*SSIM*	*LPIPS$_a$*	*LPIPS$_v$*
Meka et al.	**0.751**	0.186	0.301	**0.713**	0.219	0.331
Ours	0.718	**0.127**	**0.282**	0.672	**0.137**	**0.303**

sharper specular highlights. Failure to adequately handle misalignment leads to overall blurrier images with less detail, as illustrated in Figure 11(c).

5. Conclusion

In this work, we present a novel method for the acquisition of high-quality reflectance fields employing a practical limited zonal illumination field setup. Our approach learns to extrapolate lighting directions beyond the predetermined limits of the hardware setup. Experiments show that our method generates more realistic relighting results in contrast to simplistic image-based relighting from OLAT images with limited zonal coverage. Furthermore, comparative studies demonstrate that our method outperforms existing state-of-the-art approaches.

We further demonstrate the performance and practicality through its application in predicting the reflectance field of a dynamic performance video with varying facial expressions. Our method facilitates reflectance field acquisition research without the necessity for specialized and dedicated light stage hardware.

Limitations. Although our proposed method enables photo-realistic relighting results, we do observe a asymmetric performance of our OLAT extrapolation. Specifically, we find that the extrapolation is more effective when applied along the longitudinal direction compared to the latitudal direction. This discrepancy is caused by inherent hardware limitations, where the camera placements facilitate complete measurements of OLAT images along the longitudinal line, while only slightly increasing the latitudinal coverage. Additionally, we acknowledge that our method is not designed for capturing complex objects with sharp reflective and refractive properties.

We also acknowledge that skip-connections add grid artifacts in the darker areas of the generated image. We believe this is a worthy trade-off since we attain sharper details on the face, and in practice, the darker area (which is usually the background) often gets replaced with alpha matting.

Acknowledgement

References

[AGB*16] ANDERSON, ROBERT, GALLUP, DAVID, BARRON, JONATHAN T, et al. "Jump: virtual reality video". *ACM Transactions on Graphics (TOG)* 35.6 (2016), 1–13 3.

[AMH*23] AZINOVIĆ, DEJAN, MAURY, OLIVIER, HERY, CHRISTOPHE, et al. "High-res facial appearance capture from polarized smartphone images". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 16836–16846 2.

[BLS*21] BI, SAI, LOMBARDI, STEPHEN, SAITO, SHUNSUKE, et al. "Deep relightable appearance models for animatable faces". *ACM Transactions on Graphics (TOG)* 40.4 (2021), 1–15 2.

[CEJ*06] CHABERT, CHARLES-FÉLIX, EINARSSON, PER, JONES, ANDREW, et al. "Relighting human locomotion with flowed reflectance fields". *ACM SIGGRAPH 2006 Sketches*. 2006, 76–es 2.

[CL22] CHEN, ZHAOXI and LIU, ZIWEI. "Relighting4d: Neural relightable human from videos". *European Conference on Computer Vision*. Springer. 2022, 606–623 2.

[DHT*00] DEBEVEC, PAUL, HAWKINS, TIM, TCHOU, CHRIS, et al. "Acquiring the reflectance field of a human face". *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 2000, 145–156 1, 2.

[GFT*11] GHOSH, ABHIJEET, FYFFE, GRAHAM, TUNWATTANAPONG, BOROM, et al. "Multiview face capture using polarized spherical gradient illumination". *ACM Transactions on Graphics (TOG)* 30.6 (2011), 1–10 2.

[GLD*19] GUO, KAIWEN, LINCOLN, PETER, DAVIDSON, PHILIP, et al. "The relightables: Volumetric performance capture of humans with realistic relighting". *ACM Transactions on Graphics (ToG)* 38.6 (2019), 1–19 2.
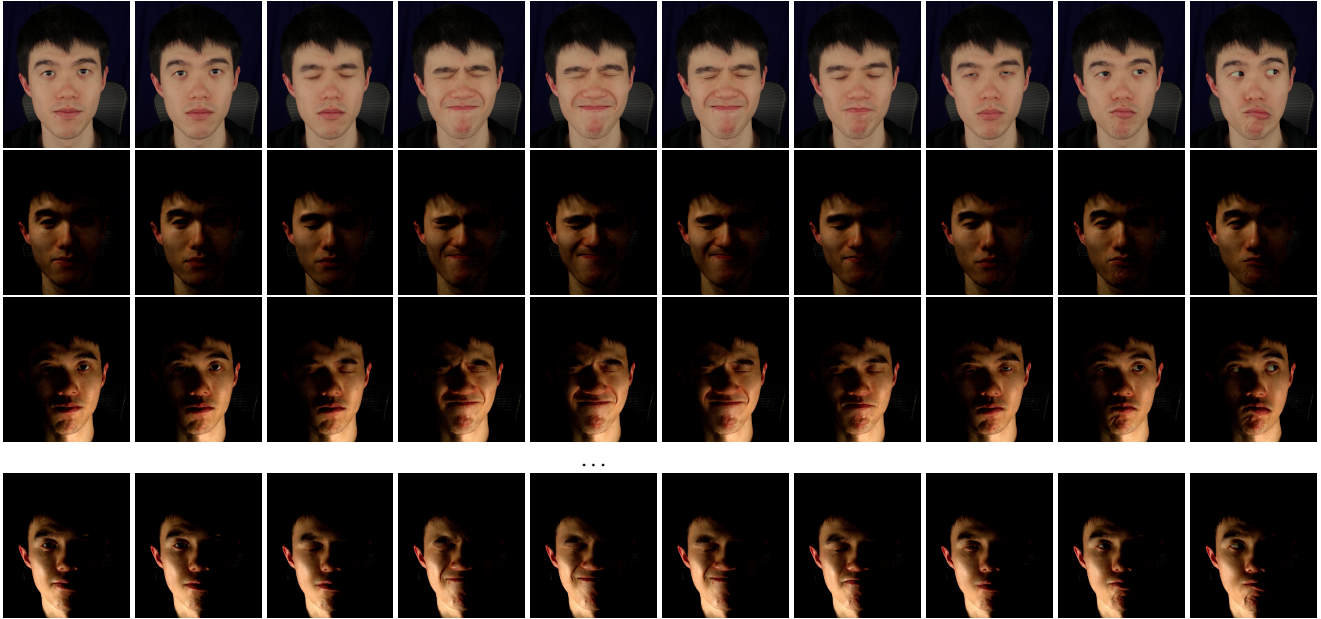
Figure 9: *Relighting of video sequence acquired under passive uniform lighting (top row) to generate corresponding OLATs.*
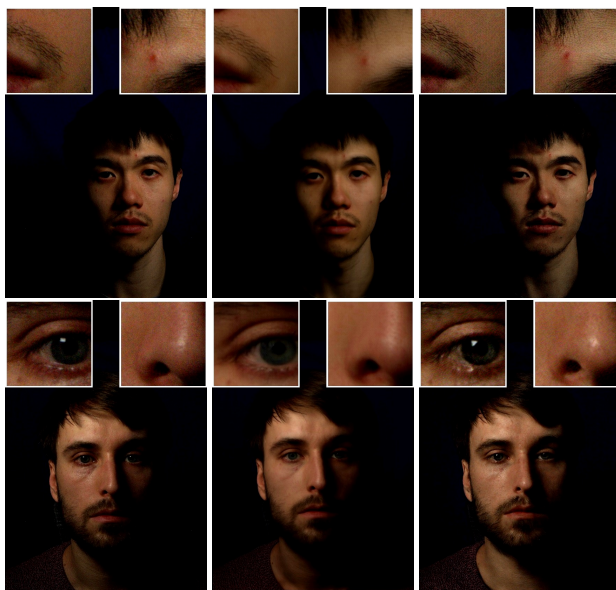
(a) *ground truth*   **(b)** *[MHP*19]*   **(c)** *ours*

Figure 10: *Comparing our methods with Deep Reflectance Fields [MHP*19] using our captured photographs. Our method produces sharper results both in texture and specular highlights.*

[HBC23] HEITZ, ERIC, BELCOUR, LAURENT, and CHAMBON, THOMAS. "Iterative α -(de)Blending: a Minimalist Deterministic Diffusion Model". *ACM SIGGRAPH 2023 Conference Proceedings*. SIGGRAPH '23. Los Angeles, CA, USA: Association for Computing Machinery, 2023. ISBN: 9798400701597. DOI: 10.1145/3588432.3591540. URL: https://doi.org/10.1145/3588432.3591540 4.

[HLX23] HAN, YUXUAN, LYU, JUNFENG, and XU, FENG. "High-Quality Facial Geometry and Appearance Capture at Home". *arXiv preprint arXiv:2312.03442* (2023) 2.

[KB14] KINGMA, DIEDERIK P and BA, JIMMY. "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980* (2014) 5.

[LBZ*20] LI, RUILONG, BLADIN, KARL, ZHAO, YAJIE, et al. "Learning formation of physically-based face attributes". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, 3410–3419 2.

[LLK*22] LATTAS, ALEXANDROS, LIN, YIMING, KANNAN, JAYANTH, et al. "Practical and scalable desktop-based high-quality facial capture". *European Conference on Computer Vision*. Springer. 2022, 522–537 2, 3.

[LYL*16] LEGENDRE, CHLOE, YU, XUEMING, LIU, DAI, et al. "Practical multispectral lighting reproduction". *ACM Transactions on Graphics (TOG)* 35.4 (2016), 1–11 6.

[MHP*07] MA, WAN-CHUN, HAWKINS, TIM, PEERS, PIETER, et al. "Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination." *Rendering Techniques* 2007.9 (2007), 10 2.

[MHP*19] MEKA, ABHIMITRA, HAENE, CHRISTIAN, PANDEY, ROHIT, et al. "Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination". *ACM Transactions on Graphics (TOG)* 38.4 (2019), 1–12 2, 3, 5–10.

[MPH*20] MEKA, ABHIMITRA, PANDEY, ROHIT, HAENE, CHRISTIAN, et al. "Deep relightable textures: volumetric performance capture with neural rendering". *ACM Transactions on Graphics (TOG)* 39.6 (2020), 1–21 2.
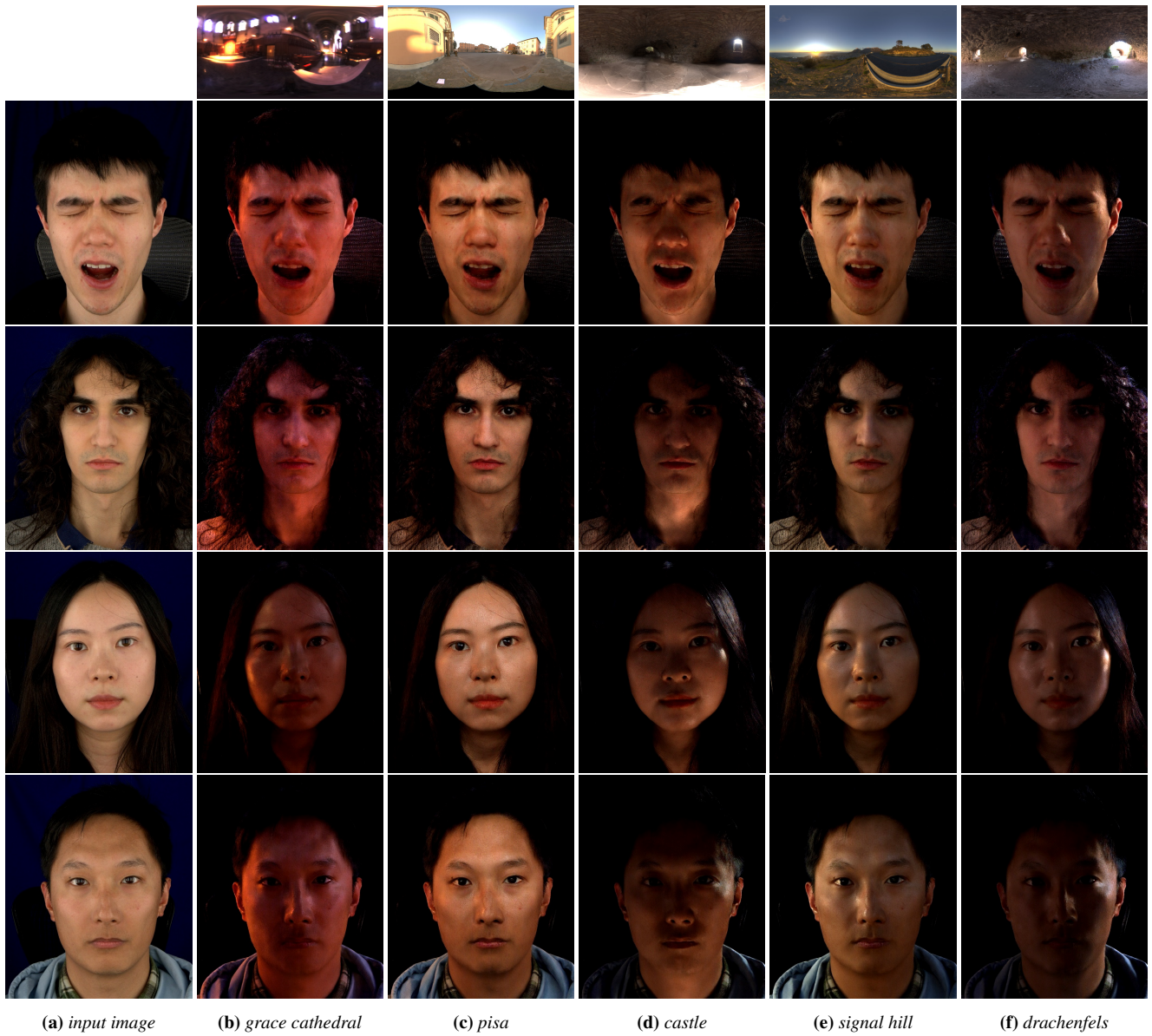
(a) *ground truth*   (b) *no skip-assisted decoding*   (c) *slide-pooling loss [MHP*19]*   (d) *ours*
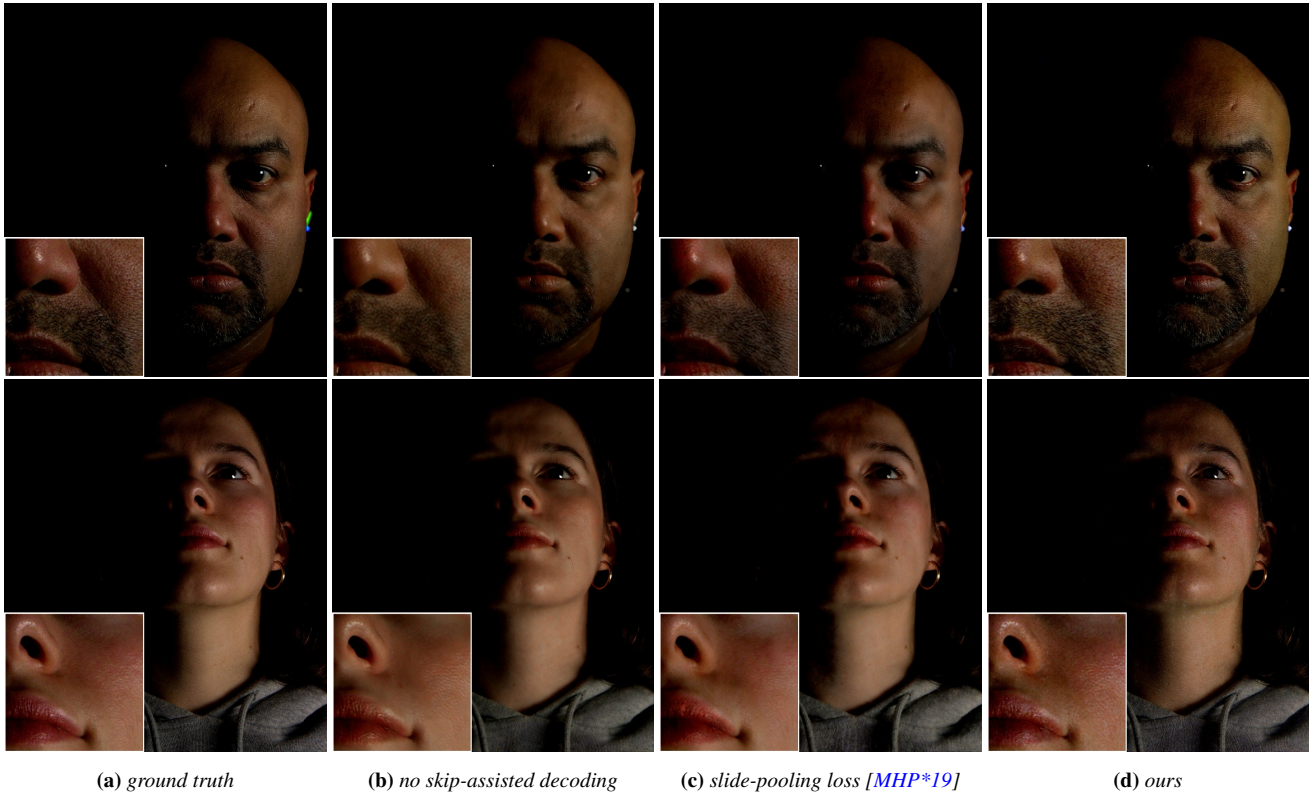
Figure 11: *Ablation study - second column (b): decoded without the proposed skip-assistance, third column (c): decoded with the skip-assistance and the slide-pooling loss proposed by [MHP*19]. Our skip-assisted decoder trained with our proposed loss functions produces sharper-detailed results (d) compared to (b) and (c).*

[MTZ18] MECHREZ, ROEY, TALMI, ITAMAR, and ZELNIK-MANOR, LIHI. "The contextual loss for image transformation with non-aligned data". *Proceedings of the European conference on computer vision (ECCV)*. 2018, 768–783 5.

[ND21] NICHOL, ALEXANDER QUINN and DHARIWAL, PRAFULLA. "Improved denoising diffusion probabilistic models". *International Conference on Machine Learning*. PMLR. 2021, 8162–8171 5.

[NLML20] NESTMEYER, THOMAS, LALONDE, JEAN-FRANÇOIS, MATTHEWS, IAIN, and LEHRMANN, ANDREAS. "Learning physics-guided face relighting under directional light". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, 5124–5133 2.

[OSF*18] OKTAY, OZAN, SCHLEMPER, JO, FOLGOC, LOIC LE, et al. "Attention u-net: Learning where to look for the pancreas". *arXiv preprint arXiv:1804.03999* (2018) 5.

[OTII07] OKABE, MAKOTO, TAKAYAMA, KENSHI, IJIRI, TAKASHI, and IGARASHI, TAKEO. "Light shower: a poor man's light stage built with an off-the-shelf umbrella and projector". *ACM SIGGRAPH 2007 sketches*. 2007, 62–es 3.

[PEL*21] PANDEY, ROHIT, ESCOLANO, SERGIO ORTS, LEGENDRE, CHLOE, et al. "Total relighting: learning to relight portraits for background replacement". *ACM Transactions on Graphics (TOG)* 40.4 (2021), 1–21 2, 6.

[PTMD07] PEERS, PIETER, TAMURA, NAOKI, MATUSIK, WOJCIECH, and DEBEVEC, PAUL. "Post-production facial performance relighting using reflectance transfer". *ACM Transactions on Graphics (TOG)* 26.3 (2007), 52–es 2.

[RFB15] RONNEBERGER, OLAF, FISCHER, PHILIPP, and BROX, THOMAS. "U-net: Convolutional networks for biomedical image segmentation". *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, 234–241 4, 5.

[SBL*23] SARKAR, KRIPASINDHU, BÜHLER, MARCEL C, LI, GENGYAN, et al. "LitNeRF: Intrinsic Radiance Decomposition for High-Quality View Synthesis and Relighting of Faces". *SIGGRAPH Asia 2023 Conference Papers*. 2023, 1–11 3.

[SBT*19] SUN, TIANCHENG, BARRON, JONATHAN T, TSAI, YUN-TA, et al. "Single image portrait relighting". *ACM Transactions on Graphics (TOG)* 38.4 (2019), 1–12 2, 6.

[SCKS21] SENGUPTA, SOUMYADIP, CURLESS, BRIAN, KEMELMACHER-SHLIZERMAN, IRA, and SEITZ, STEVEN M. "A light stage on every desk". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, 2420–2429 3.

[SSS*23] SAITO, SHUNSUKE, SCHWARTZ, GABRIEL, SIMON, TOMAS, et al. "Relightable gaussian codec avatars". *arXiv preprint arXiv:2312.03704* (2023) 2.

[SXZ*20] SUN, TIANCHENG, XU, ZEXIANG, ZHANG, XIUMING, et al. "Light stage super-resolution: continuous high-frequency relighting". *ACM Transactions on Graphics (TOG)* 39.6 (2020), 1–12 2.

[SZ14] SIMONYAN, KAREN and ZISSERMAN, ANDREW. "Very deep convolutional networks for large-scale image recognition". *arXiv preprint arXiv:1409.1556* (2014) 5.

[VV*17] VAN DEN OORD, AARON, VINYALS, ORIOL, et al. "Neural discrete representation learning". *Advances in neural information processing systems* 30 (2017) 3–5.

[WHZZ23] WANG, YIFAN, HOLYNSKI, ALEKSANDER, ZHANG, XIUMING, and ZHANG, XUANER. "Sunstage: Portrait reconstruction and relighting using the sun as a light stage". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 20792–20802 2.

[XCW*23] XU, YINGYAN, CHANDRAN, PRASHANTH, WEISS, SEBASTIAN, et al. "Artist-Friendly Relightable and Animatable Neural Heads". *arXiv preprint arXiv:2312.03420* (2023) 3.

[XZC*23] XU, YINGYAN, ZOSS, GASPARD, CHANDRAN, PRASHANTH, et al. "Renerf: Relightable neural radiance fields with nearfield lighting". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, 22581–22591 3.

[YZF*23] YANG, HAOTIAN, ZHENG, MINGWU, FENG, WANQUAN, et al. "Towards practical capture of high-fidelity relightable avatars". *SIGGRAPH Asia 2023 Conference Papers*. 2023, 1–11 2.

[ZFC*23] ZHU, ZIXIN, FENG, XUELU, CHEN, DONGDONG, et al. "Designing a better asymmetric vqgan for stablediffusion". *arXiv preprint arXiv:2306.04632* (2023) 5.

[ZFT*21] ZHANG, XIUMING, FANELLO, SEAN, TSAI, YUN-TA, et al. "Neural light transport for relighting and view synthesis". *ACM Transactions on Graphics (TOG)* 40.1 (2021), 1–17 2.

[ZIE*18] ZHANG, RICHARD, ISOLA, PHILLIP, EFROS, ALEXEI A, et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". *CVPR*. 2018 6, 8.

A. Lin, Y. Lin X. Li & A. Ghosh / High Quality Neural Relighting using Practical Zonal Illumination

(a) *input image*   **(b)** *grace cathedral*   **(c)** *pisa*   **(d)** *castle*   **(e)** *signal hill*   **(f)** *drachenfels*

Figure 12: *Additional results of relighting under various HDR environments.*