Neural Reflectance Capture in the View-Illumination Domain

Kaizhang Kang, Minyi Gu, Cihui Xie, Xuanda Yang, Hongzhi Wu and Kun Zhou *Fellow, IEEE*

Abstract—We propose a novel framework to efficiently capture the unknown reflectance on a non-planar 3D object, by learning to probe the 4D view-lighting domain with a high-performance illumination multiplexing setup. The core of our framework is a deep neural network, specifically tailored to exploit the multi-view coherence for efficiency. It takes as input the photometric measurements of a surface point under learned lighting patterns at different views, automatically aggregates the information and reconstructs the anisotropic reflectance. We also evaluate the impact of different sampling parameters over our network. The effectiveness of our framework is demonstrated on high-quality reconstructions of a variety of physical objects, with an acquisition efficiency outperforming state-of-the-art techniques.

Index Terms—multi-view illumination multiplexing, neural acquisition

1 Introduction

Digital acquisition of real-world appearance is one central problem in computer graphics and vision. A digitized object, represented as a 3D mesh and a 6D Spatially-Varying Bidirectional Reflectance Distribution Function (SVBRDF), can be rendered to faithfully reproduce the original look in the virtual world, with any view and lighting conditions.

Recently, the demand for **efficient, high-quality** digitization surges in various important applications. For cultural heritage, museums around the world are eager to digitally preserve the intricate details of millions of precious artifacts. In e-commerce, a large number of different products must be scanned with high precision for online display, usually under a tight time budget. Even for research in computer graphics and vision, the lack of a large-scale database of high-quality captured 3D objects substantially hinders the development of novel learning-based algorithms.

While geometry digitization has received considerable progress in the past decades [1], [2], the acquisition of general reflectance remains challenging. One fundamental difficulty in image-based reflectance capture is the dimensionality mismatch: the image sensors are designed to efficiently probe the 2D spatial domain only; directly measuring a 6D SVBRDF with a 2D sensor would take a prohibitively long time for capturing photographs under all possible combinations of view and lighting directions, in order to preserve high-frequency features like sharp highlights [3], [4].

Significant research efforts have been made, to improve the sampling efficiency in the 4D view-illumination domain. With a single point light source, various hand-crafted [5] or data-driven priors [6] have been proposed, to properly regularize the reflectance solution from a sparser number of samples. When multiple lights are available, illumination multiplexing [7] can be employed to physically encode more light-dependent information in each measurement, and then computationally recover challenging appearance such as anisotropic reflectance. This results in a higher signal-to-noise ratio with a shorter capture time [8]. However, while existing work on illumination multiplexing can take input images from each view at a time for reconstructing non-planar reflectance, state-of-the-art techniques [9], [10] do not exploit the coherence among photometric measurements across different views, leading to a suboptimal acquisition efficiency.

Fig. 1: By learning to efficiently acquire the appearance information in the view-illumination domain with a near-field lightstage, we faithfully reconstruct complex, spatially-varying non-planar reflectance of physical objects. Here we show the captured results of a variety of real-world objects under novel lighting and view conditions. Please refer to the accompanying video for animated results. Background texture courtesy of Design Connected EOOD.

In this paper, we present a novel framework to efficiently

• *H. Wu is the corresponding author. All authors are with State Key Lab of CAD & CG, Zhejiang University, Hangzhou, China, 310058, except that X. Yang is with University of California San Diego. K. Zhou is also affiliated with ZJU-FaceUnity Joint Lab of Intelligent Graphics. E-mail: hwu@acm.org*

Manuscript received Oct 15, 2020; revised Sep 24, 2021.

acquire the reflectance on a non-planar 3D object, by learning to probe and aggregate the information in the 4D view-illumination domain, with a high-performance lighting multiplexing hardware prototype. The core of the framework is a deep neural network: for each visible surface point, it takes photometric measurements under learned lighting patterns at different views as input, and reconstructs the normal and the reflectance. The network is carefully designed to tackle the challenges, as well as take the unique opportunities in the joint domain of view and illumination: we propose (1) an efficient network structure to exploit the rotational equivariance of input/output and support a variable number of visible input views, (2) a primary view selection mechanism to reduce the variation of input data for network efficiency, and (3) a view-dependent smoothing factor to address the non-differentiability of the input. Our work belongs to the recent line of research on neural reflectance capture [8], [10], in which the entire pipeline of physical sampling and computational reconstruction of reflectance is automatically and jointly optimized by mapping to a mixed-domain neural network.

The effectiveness of our framework is demonstrated with the efficient acquisition of physical objects with considerable variations in reflectance. One typical configuration in our framework requires 12 views and 6 lighting patterns per view. Adding the 24 lighting patterns for geometry reconstruction, the total number of photographs is **96**, corresponding to a camera acquisition time of 30 seconds. In comparison, **768** images are used in [10] and **1,100** in [9], which are two techniques most similar to ours. We also validate our results with the photographs, as well as compare against related work. Finally, we evaluate the impact of different sampling parameters over appearance reconstruction, and provide guidelines that might be useful for future work.

2 Related Work

In this section, we mainly review previous work on spatially-varying reflectance capture with *controlled lighting*, which is closely related to our paper. For a broader view of the field, we direct interested readers to excellent recent surveys [11], [12], [13], [14]. Depending on the type of light source used during acquisition, related work can be categorized into the following two classes.

2.1 Point Light Source

The most general acquisition approach is to exhaustively sample the 6D domain of reflectance, producing high-quality direct measurements [3], [4]: a mechanical system places a point light and a camera at densely sampled lighting and view directions, and then the camera takes one photograph of the sample in each case. The capture time is prohibitively long. One solution is to employ a large number of cameras and lights and switch one pair on at a time, to avoid mechanical movements [15]. The other way is to introduce various priors to improve the acquisition efficiency, while ensuring that the reconstruction stays well-determined with the reduction in the number of samples.

2.1.1 Hand-Crafted Priors

Reflectance can be reconstructed from a small number of photographs, by constraining it to be a linear combination of basis materials, using a point light in the visible spectrum [5], [16] or even an IR emitter [17]. Zickler et al. [18] trade the spatial resolution for the angular accuracy via a scattered-data interpolation, essentially sharing the reflectance over its 6D domain. Dong et al. [19] propose a two-stage acquisition method, assuming that the reflectance lies on a low-dimensional manifold.

Wang et al. [20] exploit the spatial similarity of reflectance and the spatial variations of local frames on a typical sample, to synthesize complete microfacet distributions of BRDFs. A projector-camera pair is introduced in [21] for joint acquisition of shape and reflectance, with a strong prior imposed on the latter due to the limited sampling in the angular domain. Aittala et al. [22] use 2 input photographs to reconstruct stationary materials. The number is further reduced to 1 in [23], where the correspondences among pixels of presumably similar appearance are established.

2.1.2 Data-Driven Priors

Priors obtained from a large amount of data via machine learning techniques often lead to more robust and efficient reconstruction algorithms, compared with hand-crafted, heuristic ones. Matusik et al. [24] perform principal component analysis on measured isotropic reflectance and derive a sampling scheme with 800 samples. Nielsen et al. [25] reduce the number to about 20 with an algorithm that optimizes both lighting and view sampling directions. A subsequent work further cuts it to 2, exploiting the variations of view directions on a near-field camera [26].

The development of deep learning over the past years enables even more powerful data-driven priors. Li et al. [27] estimate an SVBRDF from a single image of a planar sample under unknown environment illumination, with a self-augmentation training process. Deschaintre et al. [28] propose a method that takes a single input photograph lit by a flash, and outputs an SVBRDF with a network trained over a large dataset of procedural materials. They further introduce a pooling-based network to aggregate appearance information from 1 to 10 input images [29]. The isotropic reflectance and a depth map can be directly regressed from a single image under unknown environment illumination and flash lighting [30]. By learning a latent embedding, it is possible to efficiently optimize an SVBRDF with respect to an arbitrary number of input images [31], [32]. Recently, excellent isotropic reflectance reconstruction is demonstrated, by aggregating photographs captured with a camera and a collocated point light from 6 sampled directions [6], [33]. Gao et al. [34] propose a neural appearance representation for free-viewpoint relighting, which is computed from more than 5,000 flash-lit images.

2.2 Illumination Multiplexing

This class of techniques program the intensities of multiple lights simultaneously, record the responses of a sample under different lighting configurations, and recover the reflectance from the measurements. With multiple lights on, more appearance information is physically packed into each

measurement, making it possible to robustly and efficiently reconstruct challenging cases such as anisotropic reflectance. Furthermore, unlike many point-light-based approaches, no spatial coherence assumption is typically needed, which results in higher quality reconstructions [9], [35].

Lightstage systems take photographs under carefully designed lighting patterns, and recover the reflectance from an inverse lookup table [7], [9], or via an alternating optimization [36]. With 2 images captured under color gradient illumination, Meka et al. [37] infer the isotropic reflectance of human faces with a deep network. In [38], [39], [40], a linear light source is moved over a planar material sample, and the SVBRDF is estimated from the corresponding appearance variations. Aittala et al. [41] use a camera and an LCD panel as the light source, to capture isotropic reflectances based on a frequency domain analysis.

Recently, Kang et al. [8] map the illumination patterns and the single-view reconstruction algorithm to an autoencoder, which enables the automatic optimization of both factors. The idea is further extended to the joint acquisition of reflectance and shape [10]. While state-of-the-art work [9], [10] takes multi-view images as input, they are essentially single-view techniques: the inter-view coherence among photometric measurements is not exploited, and the reflectance is independently estimated at each view. It is not clear how to perform efficient sampling in the joint 4D domain of view and illumination, especially with a low per-view bandwidth that is not sufficient to produce satisfactory results using existing work.

3 Hardware Prototype

We build a high-performance, box-shaped lightstage to conduct the acquisition experiments (Fig. 2). Its size is 80cm $\times$ 80cm $\times$ 77cm. The sample object, with a maximum size of 20cm $\times$ 20cm $\times$ 20cm, is placed on a digital turntable near the center of the device, and rotated to different angles for multi-view imaging. A single FLIR BFS-U3-123S6C-C vision camera captures high-dynamic-range (HDR) photographs at a resolution of 4,096$\times$3,000. We illuminate the sample with 24,576 LEDs on the six faces of the device with polycarbonate diffusers attached. The total LED power is about 2,000W, and the pitch of adjacent LEDs is 1cm. We achieve a binary pattern projection speed of 48,000 frames per second, with home-designed circuits for high-speed control and synchronization.

We carefully calibrate the intrinsic/extrinsic parameters of the camera, as well as the positions, orientations, angular intensity of LEDs. Color calibration is performed with an X-Rite ColorChecker Passport. We resolve the scale ambiguity of diffuse/specular albedo with a planar diffuse patch of a uniform albedo [38]. The rotation angle of the turntable is computed from printed markers on its surface [42].

4 Preliminaries

4.1 Assumptions

We assume a known, coarse geometry computed with a state-of-the-art shape reconstruction technique. We also assume that the appearance of interest can be well described as a 6D anisotropic SVBRDF. The reflectance at each spatial location is independently reconstructed, with no assumption on the spatial coherence. In addition, no polarization filter is used in acquisition. Similar to the majority of related work, we do not explicitly handle interreflections and self-occlusions, and leave their formal treatment to future work.

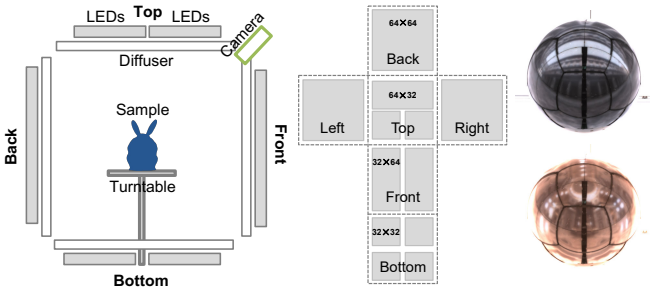


Fig. 2: The lighting layout. From the left to right, a side view of our prototype, the vertical-cross parameterization of all lights with 4,096 LEDs on each face, and the calibration sphere under physically projected *uffizi*/*stpeters* environment maps.

4.2 Equations

The following derivations are based on a gray-scale channel for brevity, which can be easily extended to the spectral domain. First, the outgoing radiance L from a surface point $\mathbf{p}$ towards a fixed camera can be modeled as:

$$L(I,\mathbf{p}) = \sum_l I(l) \int \frac{1}{||\mathbf{x_l}-\mathbf{x_p}||^2} \Psi(\mathbf{x_l}, -\omega_\mathbf{i}) V(\mathbf{x_l}, \mathbf{x_p}) f(\omega_\mathbf{i}'; \omega_\mathbf{o}', \mathbf{p}) (\omega_\mathbf{i} \cdot \mathbf{n_p})^+ (-\omega_\mathbf{i} \cdot \mathbf{n_l})^+ d\mathbf{x_l}. \quad (1)$$

Here $\mathbf{x_p}/\mathbf{n_p}$ is the position/normal of $\mathbf{p}$, while $\mathbf{x_l}/\mathbf{n_l}$ is the position/normal of a point on a locally planar light source with an index of l. $\omega_\mathbf{i}/\omega_\mathbf{o}$ denotes the lighting/view direction in the world space, and $\omega_\mathbf{i}'/\omega_\mathbf{o}'$ is expressed in the local frame of $\mathbf{p}$. Note that $\omega_\mathbf{i} = \frac{\mathbf{x_l}-\mathbf{x_p}}{||\mathbf{x_l}-\mathbf{x_p}||}$. $I(l)$ is the intensity for the light with an index of l, in the range of [0, 1]. The array $\{I(l)\}$ corresponds to a lighting pattern. $\Psi(\mathbf{x_l}, \cdot)$ represents the angular distribution of the light intensity. V is a binary visibility function between $\mathbf{x_l}$ and $\mathbf{x_p}$. The operator $(\cdot)^+$ computes the dot product between two vectors, and clamps any negative result to zero. $f(\cdot; \omega_\mathbf{o}', \mathbf{p})$ is a 2D BRDF slice, which is a function of the lighting direction.

We employ the anisotropic GGX model [43] to represent f:

$$f(\omega_\mathbf{i}'; \omega_\mathbf{o}', \mathbf{p}) = \frac{\rho_d}{\pi} + \rho_s \frac{D(\omega_\mathbf{h}'; \alpha_x, \alpha_y) F(\omega_\mathbf{i}', \omega_\mathbf{h}') G(\omega_\mathbf{i}', \omega_\mathbf{o}'; \alpha_\mathbf{x}, \alpha_\mathbf{y})}{4(\omega_\mathbf{i}' \cdot \mathbf{n})(\omega_\mathbf{o}' \cdot \mathbf{n})}, \quad (2)$$

where ρ_d/ρ_s is the diffuse/specular albedo, α_x/α_y is the roughness, and ω_h' is the half vector. In addition, D is the microfacet distribution function, F is the Fresnel term, and G is the geometry term for shadowing/masking effects. An index of refraction of 1.5 is used.

4.3 Definitions

As L is linear with respect to I (Eq. 1), it can be expressed as the dot product between I and a **lumitexel** m:

$$L(I, \mathbf{p}) = \sum_l I(l) m(l; \mathbf{p}), \tag{3}$$

where m is a function of the light index l, defined on the surface point $\mathbf{p}$ of the sample object:

$$m(l; \mathbf{p}) = L(\{I(l) = 1, \forall_{j \neq l} I(j) = 0\}, \mathbf{p}). \tag{4}$$

Each entry of m records a "virtual" measurement of L, with only one light turned on and set to its maximum intensity, and all other lights off.

We denote the number of views/lighting patterns as #v/#l, respectively. During acquisition, the sample object is rotated to #v equally spaced angles for multi-view imaging. Consequently, a point p on the object surface is also rotated #v times. We define a **multi-view lumitexel** q at $\mathbf{p}$ as a collection of lumitexels at #v different views:

$$q(l, v; \mathbf{p}) = V(\mathbf{p}^v, \mathbf{x_c}) m(l; \mathbf{p}^v), \tag{5}$$

where v is the view index in the range of [0, #v-1]; $\mathbf{p}^v$ is the position of $\mathbf{p}$ at the view v, after a rotation of $\frac{v}{\#v} \times 2\pi$ along the rotation axis of our digital turntable; $\mathbf{x_c}$ is the optical center of the camera, and V tests if the $\mathbf{p}^v$ is visible to the camera. Please refer to Fig. 13 for examples of multi-view lumitexels.

We observe that the acquisition is equivalent to projecting the multi-view lumitexel of $\mathbf{p}$ with the lighting patterns $\{I_j\}_{j=0,1,\ldots,\#l-1}$ in the physical domain, which yields a set of multi-view photometric **measurements** r, defined as:

$$r(v, j; \mathbf{p}) = \sum_l q(l, v; \mathbf{p}) I_j(l). \tag{6}$$

Finally, we define the **primary view** ϕ of a point $\mathbf{p}$ among all $\#v$ views, as the view index where the normal is closest to the view direction:

$$\phi(\mathbf{n}) = \arg\max_v (\mathbf{n}^v \cdot \omega_o^v)^+. \tag{7}$$

Here $\mathbf{n}^v/\omega_\mathbf{o}^\mathbf{v}$ are the normal/view direction at the view v, after a corresponding turntable rotation. Please refer to Sec. 6.1 for why we need to define ϕ. Note that our framework is not tied to the above definition. One can also plug in other definitions of the primary view.

5 Overview

To acquire the appearance of a physical object, we rotate it to $\#v$ equally spaced angles, and take photographs under $\#l$ learned lighting patterns at each view. First, a coarse shape is reconstructed with multi-view stereo. Next, for each surface point $\mathbf{p}$, our neural network predicts its shading normal from the corresponding multi-view photometric measurements. The result is subsequently used to optimize the coarse geometry and select a primary view. With the multi-view correspondences computed on the refined shape, we run our network again to produce as output a diffuse/specular lumitexel at the primary view. Finally, we fit parametric BRDFs to the lumitexels, and store in texture maps as the reflectance results. Please refer to Fig. 3 for an illustration of our pipeline.

6 Our Neural Network

6.1 Design Decisions

Below we describe our key design decisions, before introducing the details of the network. Similar to previous work [8], [10], we choose not to directly regress anisotropic BRDF parameters, as the mapping from a multi-view lumitexel to various parameters is complex and not amenable for machine learning. Instead, we opt to output lumitexels and leave the parameter estimation to the non-linear fitting.

Next, since the network takes a physical multi-view lumitexel as input, it would be straightforward to propose an autoencoder that reproduces the input. However, such a network is inefficient, as it does not exploit the **rotational equivariance** of multi-view lumitexels: if the input multi-view lumitexel is circularly permuted by k views, we expect a permutation on the output as well. A naïve network would have to learn this property from the training data with no quality guarantees. Instead, we decide to output one view of the multi-view lumitexel at a time, and permute the input so that the view angle of each input view relative to the output remains constant. This not only explicitly enforces the equivariance, but also simplifies the task for the network. An illustration is shown in Fig. 4.

Furthermore, we observe that it is not necessary to recover the complete input multi-view lumitexel. Existing work on reflectance capture (e.g., [8], [40]) has demonstrated that the observations at a non-extreme view are sufficient to recover an anisotropic BRDF. In addition, fitting a single-view lumitexel avoids the undesired blur, which might occur due to even a slight inconsistency in the multi-view fitting scenario. Therefore, we decide to reconstruct the lumitexel at a primary view only (Eq. 7). This is roughly equivalent to group all possible samples into #v slices; and then our network only needs to recover samples whose normal lies in a single slice, which is a substantial reduction compared with a hemisphere as in related work [9], [10], resulting in a superior efficiency. Please see Fig. 5 for a visualization. Note that the variation of the input lumitexels decreases, with the increase of #v.

Finally, unlike existing work on neural reflectance capture [8], [10], our input multi-view lumitexel is inherently non-differentiable, due to the discontinuous visibility change. Also, the lumitexels at grazing views often contain strong Fresnel peaks, making it numerically challenging for proper handling in a neural network. Moreover, the footprint of a pixel on the object surface grows much larger at such views, which may lead to correspondence inaccuracies. To alleviate the above issues, we propose to multiply the multi-view measurements with a view-dependent smoothing factor, as detailed in the next subsection.

6.2 Input/Output

The input to our network is the view- and light- dependent measurements r (Eq. 5) of a physical multi-view lumitexel. For the aforementioned differentiability/grazing angle suppression reasons, we transform the measurements by multiplying with a view-dependent smoothing factor, before feeding them to the network:

$$r'(v, j; \mathbf{p}) = r(v, j; \mathbf{p}) \frac{H[(\mathbf{n}^v \cdot \omega_\mathbf{o})^+; \#v]}{\max_v H[(\mathbf{n}^v \cdot \omega_\mathbf{o})^+; \#v]}. \tag{8}$$

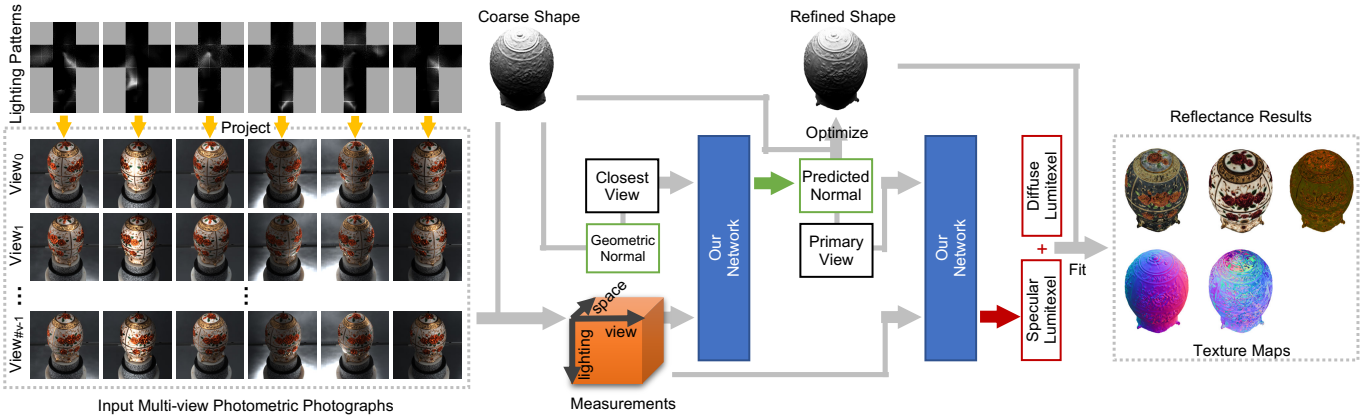


Fig. 3: Our processing pipeline. First, we take photographs of a sample object under learned lighting patterns at equally spaced views. For each visible surface point, we feed its multi-view photometric measurements to our network to obtain a more accurate normal, expressed in a view determined by the geometric normal. The predicted normal helps optimize the initial coarse 3D geometry, and selects a primary view. We then send the measurements to our network again to produce the diffuse and specular lumitexel at the primary view. Finally, we fit parametric BRDFs to the lumitexels, and store the reflectance results as texture maps.

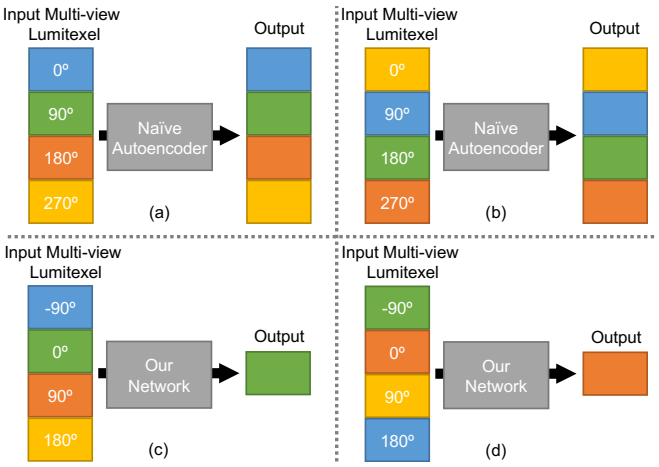


Fig. 4: The structural benefit of our network. The rotational equivariance of input/output must be learned via training data in a naïve autoencoder (a & b). Absolute rotation angles are used here. Our network outputs the lumitexel at one view at a time, assuming only fixed, relative rotation angles with the input views (c & d). This structure explicitly enforces the rotational equivariance. No re-training is needed, if all input views are rotated by a multiple of 90° in this case.

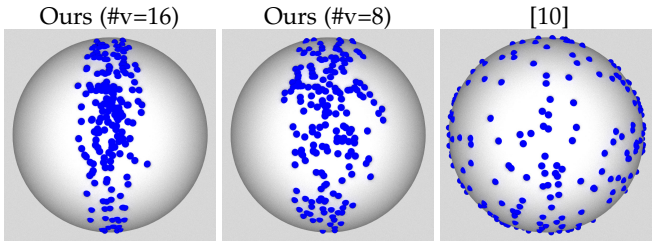


Fig. 5: Comparison of the normal distributions of input data of different networks. Each blue dot represents the normal of a randomly sampled training lumitexel. The number of samples are the same in all three cases. Due to the primary view selection mechanism, our networks are tuned to reconstruct lumitexels with smaller variations in normal, resulting in a higher efficiency. In comparison, existing work [10] handles normals that lies on an entire visible hemisphere.

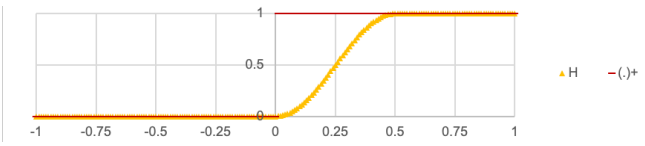


Fig. 6: A plot of the ramp function H, in comparison with the original non-differentiable $(\cdot)^+$. β is indicated on the horizontal axis. We use #v=12 in this plot.

Here H is a ramp function that smoothes out the abrupt visibility change, defined as follows:

$$H(\beta;\#v) = \begin{cases} 0, & \text{for } \beta \le 0, \\ \frac{1}{2}[1-\cos(\frac{\beta}{\beta_0}\pi)], & \text{for } 0 < \beta < \beta_0, \\ 1, & \text{for } \beta_0 \le \beta, \end{cases} \quad (9)$$

where $\beta_0 = \cos(\frac{\pi}{2} - \frac{2\pi}{\#v})$. A plot of $H(;12)$ is shown in Fig. 6. Note that at a grazing view angle, $(\mathbf{n}^v \cdot \omega_\mathbf{o})^+$ produces a small value close to 0. The normalization term $\frac{1}{\max_v H}$ is needed, to ensure that at least the measurements at one visible view are not attenuated. Otherwise, the network would have no clue to resolve the scale ambiguity between ρ_d/ρ_s and the smoothing factor in certain cases. Here we actually use the coarse geometric normal $\mathbf{n}_{\text{geo}}$ in computing the smoothing factor, as $\mathbf{n}$ is not known at the time.

The output of the network is a refined normal at the view selected by $\mathbf{n}_{\text{geo}}$, and the diffuse/specular lumitexel at the primary view determined by the predicted normal. Similar to [10], we employ a uniformly resampled cube map parameterization of $6 \times 8^2/6 \times 64^2$ for output diffuse/specular lumitexels, based on their different frequency natures. The idea of resampling is to decouple the output parameterization from the physical lighting layout, which often has discontinuities due to practical factors. Note that

the predicted normal is used not only for the primary view selection, but also for shape refinement in geometry reconstruction (Sec. 7.1).

6.3 Architecture

The first part of the network is a single fully connected (fc) layer that corresponds to the lighting patterns used at each view. The layer has $\frac{\#l}{2}\times h$ linear weights, where $\#l$ is the number of lighting patterns and h is the dimension of a single-view lumitexel ($h = 24,576$ in our experiments). For physical realization, we split every h linear weights into two lighting patterns: one pattern is obtained with all positive weights unchanged and negative ones clamped to zero, and the other is computed with all negative weights negated and positive ones set to zero, similar to [9], [10]. There is also a normalization (norm) layer for each set of h weights so that the l^2-norm is enforced to be 1, in order to bound the values and prevent degradation in the presence of noise (Sec. 6.5). Note that the same set of lighting patterns are used in all views.

The lighting pattern layer is followed by two branches, for predicting the normal (7 fc layers and 2 norm layers) and the diffuse/specular lumitexel (20 fc layers). We use mostly fc layers to avoid making assumptions about the relationships among components of the measurements. Each fc layer, except for the one that outputs the result, is followed by a leaky ReLU activation layer. For the normal branch, the first norm layer is to take away the effect of albedos over the measurements, which is irrelevant with the normal; the second norm layer explicitly produces a unit vector as output.

Note that the outputs of the normal branch and the lumitexel branch are connected, as illustrated in Fig. 3. Once we obtain the predicted normal, we use it to select a primary view and ask the lumitexel branches to output results at that view. As a result, the input multi-view lumitexel measurements are circularly permuted accordingly, to maintain a fixed rotation angle between each input view and the primary view.

6.4 Loss Function

The loss function $\mathcal{L}$ consists of 3 data terms, which penalize the deviation from the network predictions to their ground-truths.

$$\mathcal{L} = \lambda_n \mathcal{L}_n(\mathbf{n}^{\phi(\mathbf{n}_{\text{geo}})}) + \lambda_d \mathcal{L}_d(m_d^{\phi(\mathbf{n})}) + \lambda_s \mathcal{L}_s(m_s^{\phi(\mathbf{n})}), \quad (10)$$

where

$$\mathcal{L}_n(\mathbf{n}) = ||\mathbf{n}^{\phi(\mathbf{n}_{\text{geo}})} - \tilde{\mathbf{n}}^{\phi(\mathbf{n}_{\text{geo}})}||_2,$$
$$\mathcal{L}_s(m_s) = \Sigma_l[\log(1 + m_s^{\phi(\mathbf{n})}(l)) - \log(1 + \tilde{m}_s^{\phi(\mathbf{n})}(l))]^2, \quad (11)$$
$$\mathcal{L}_d(m_d) = \Sigma_l[m_d^{\phi(\mathbf{n})}(l) - \tilde{m}_d^{\phi(\mathbf{n})}(l)]^2. \quad (12)$$

Here $\mathcal{L}_s$ is the specular lumitexel loss, $\mathcal{L}_d$ is the diffuse lumitexel loss, $\mathcal{L}_n$ is the normal loss, $\phi(\mathbf{n}_{\text{geo}})$ is the view selected using the geometric normal $\mathbf{n}_{\text{geo}}$, and $\phi(\mathbf{n})$ is the primary view chosen by the predicted normal $\mathbf{n}$. m_d/m_s is the predicted diffuse/specular lumitexel at the view $\phi(\mathbf{n})$. The corresponding ground-truths are denoted with a tilde. When computing $\mathcal{L}_s$, a $\log$ transform is applied to compress the high dynamic range. We use $\lambda_n = 1$, $\lambda_d = 1$ and $\lambda_s = 0.01$ in all experiments.

6.5 Training

Our network is implemented with PyTorch, using the Adam optimizer [44] with mini-batches of 50 and a momentum of 0.9. Xavier initialization is applied to all weights in the network. We train 2 million iterations with a learning rate of 1×10^{-4}.

6.5.1 Data

To generate sufficient training data, we synthesize lumitexels with randomly sampled geometric and reflectance properties. For geometric properties, we randomly pick a location p from a valid volume in the lightstage (Sec. 3); we uniformly sample $\mathbf{n}$ that is visible with respect to the camera from at least one view, and $\mathbf{t}$ as a random unit vector orthogonal to $\mathbf{n}$. For reflectance properties, we use the anisotropic GGX model and randomly sample ρ_d/ρ_s uniformly in the range of $[0, 1]^3$, and α_x/α_y uniformly on the log scale in the range of $[0.006, 0.5]$.

To model the coarse geometric normal $\mathbf{n}_{\text{geo}}$, we perturb the original $\mathbf{n}$ with a randomly sampled orthogonal vector, whose length is drawn from a Gaussian distribution (μ=0, σ=0.5). The result is normalized and stored as $\mathbf{n}_{\text{geo}}$. Note that due to the possible difference between $\mathbf{n}_{\text{geo}}$ and $\mathbf{n}$, there are views that are visible with $\mathbf{n}_{\text{geo}}$ but not $\mathbf{n}$. For such views, we sample a new set of parameters and generate a novel lumitexel accordingly, to simulate a correspondence error in practice (i.e., other surface points are observed at these views).

The calibration data of the acquisition device (Sec. 3) are used in the evaluation of Eq. 5 for multi-view lumitexel synthesis. We split the synthetic data with a ratio of 8:2 into the training/validation set.

6.5.2 Noise

To add the robustness of the network to physical acquisition noise, we multiply each photometric measurement with a Gaussian noise (μ=1, σ=5%). In addition, we perturb the primary view angle by a Gaussian noise (μ=0, σ=0.3 $\times$ $\frac{2\pi}{\#v}$), so that non-primary view predictions are also trained with a non-zero probability, to cope with the possible view prediction inaccuracy. Furthermore, to increase the tolerance for multi-view-related errors (e.g., minor misalignments/the changing footprint of a pixel w.r.t. the view angle), we perturb the reflectance parameters before computing the lumitexels at non-primary views. Specifically, we multiply a Gaussian noise (μ=1, σ=10%) to each channel of ρ_d/ρ_s; another Gaussian noise (μ=1, σ=15%) is multiplied to α_x/α_x. Finally, dropout regularization with a rate of 10% is applied to all fc layers, except the first one corresponding to lighting patterns, and the ones right preceding the outputs.

We set the standard deviations of the above noises larger than the statistics observed in a pilot study, to account for phenomena like shadowing/interreflection that are not explicitly modeled in synthetic samples. The effectiveness of this strategy is also demonstrated in [45] with a number of concave objects.

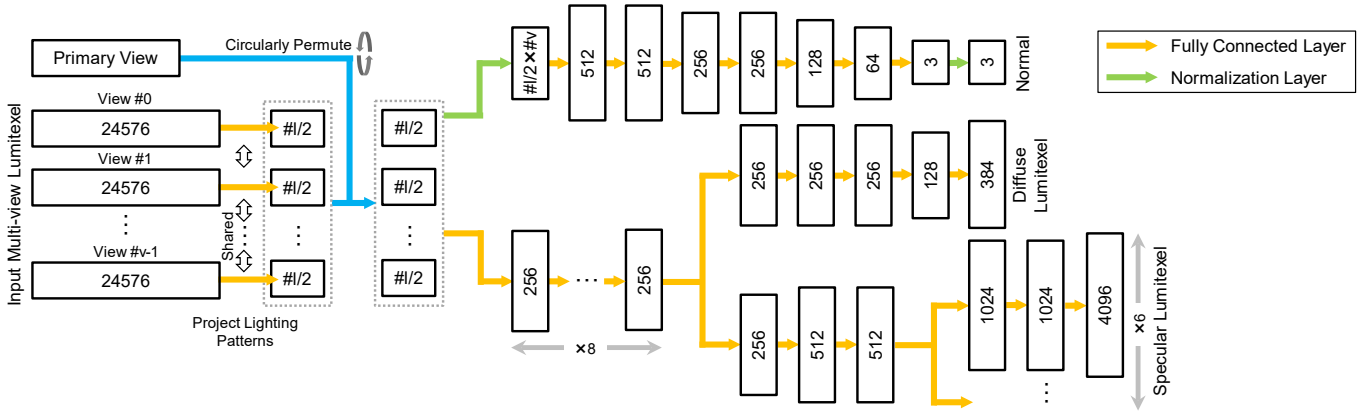


Fig. 7: The architecture of our neural network. The physical multi-view lumitexel of a point p is transformed into a small number of measurements by a set of lighting patterns shared across all views, implemented as a single fully connected layer. The per-view measurements are circularly permuted according to the primary view. Then they are aggregated to compute a normal, a gray-scale diffuse and specular lumitexel at the primary view as output.

7 Implementation Details

7.1 Geometry Reconstruction

We use input photographs of the object from multiple views under a full-on lighting pattern. The idea is to physically minimize lighting-dependent effects. These photographs are fed to a state-of-the-art multi-view stereo technique [2] to generate a point cloud as output. We then apply screened Poisson surface reconstruction [46] to compute a coarse 3D mesh from the point cloud. Next, remeshing is performed to roughly match the vertex density with the resolution of the input images. Finally, with the coarse 3D mesh for establishing correspondences among input images, we directly predict more refined normals with our network, and optimize the position of each vertex accordingly using [47]. Note that other methods for shape reconstruction can also be employed in our framework.

7.2 Reflectance Computation

Once the refined geometry is constructed, we compute spatially-varying GGX BRDF parameters and store them in texture maps as the reflectance result. First, a uv-parameterization is built on the mesh. Then for each texel, we determine its corresponding surface point $\mathbf{p}$, and compute its visibility with respect to the camera at each view via ray tracing. Next, the photometric measurements at all sampled views are sent to the network, to produce a normal, and a diffuse/specular lumitexel at the primary view selected by the predicted normal. Note that the measurements at invisible views are filled with zeros. Subsequently, we fit α_x, α_y, $\mathbf{n}$, $\mathbf{t}$ and a gray-scale ρ_s to the spectrally averaged specular lumitexel with L-BFGS-B [48]. Finally, we compute the RGB ρ_d and ρ_s using non-negative linear least squares with respect to the predicted RGB lumitexels, while fixing all other known parameters.

8 Results & Discussions

All experiments are performed on a workstation with an Intel Core i9-10940X CPU, 256GB memory, and a GeForce GTX 2080 Ti video card. During acquisition, we compute high-dynamic-range (HDR) images of the physical object, by merging 3 LDR ones with different exposures via bracketing. As currently there is only one camera in the setup, we rotate the turntable $\#v$ times with an angle of $\frac{2\pi}{\#v}$ each, to capture multiple views of the physical object. As mentioned in Sec. 7.1, the geometry acquisition is decoupled from reflectance capture. In our experiments, we take 24 photographs with all lights on for shape reconstruction. With a typical set of learned lighting patterns (#v=12, #l=6), the total capture time of all 24+72=96 HDR photographs is 30 seconds, excluding the turntable rotation time. The total size of photographs is 3GB. In the future, the acquisition time may be further reduced by deploying more cameras to capture multiple views simultaneously.

It takes 80 hours to train our neural network. The typical time to run the network on the multi-view photometric measurements of an object is 2 minutes. It takes 15 minutes for geometry reconstruction with multi-view stereo, and 2 hours for SVBRDF fitting with our unoptimized code. We use a resolution of 1024^2 for all texture maps. The results in this paper are rendered with path tracing via NVIDIA OptiX. We visualize in Fig. 8 the learned lighting patterns, trained with anisotropic GGX samples and different parameters in sampling the view/illumination domain. The photographs of a physical object under the patterns with #v=12 and #l=6 are also shown. The appearance variations of the sample object under our optimized lighting patterns will be exploited in subsequent reflectance reconstruction.

8.1 Modeling Results & Comparisons

In Fig. 13, we first show reconstruction results on synthetic lumitexels. Compared with baseline methods (using fixed SH patterns or [10]), our network faithfully recovers the normal and the diffuse/specular lumitexels from input multi-view lumitexel with considerable variations. Moreover, as the input bandwidth increases, the quality of the reconstruction results are improved. Note that all reconstruction results in the figure are the direct outputs from the networks, prior to fitting.

In Fig. 14, we test our framework on non-planar physical objects with varied reflectances. The texture maps that rep-

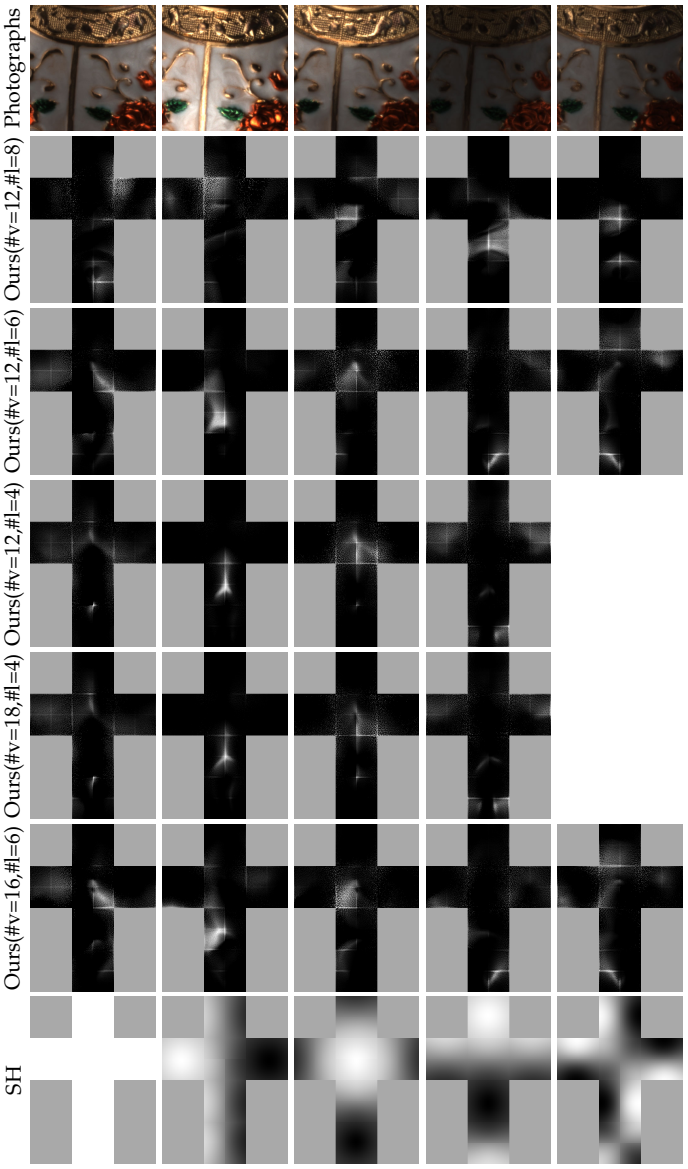


Fig. 8: Different lighting patterns. From the top row to bottom: the photographs of a physical sample lit with the corresponding lighting patterns in the third row, our learned patterns (#v=12, #l=8), (#v=12, #l=6), (#v=12, #l=4), (#v=18, #l=4) and (#v=16, #l=6), and spherical harmonics patterns (SH). Only a subset of all patterns are shown due to the limited space.

resent the reflectance results are shown. Please also refer to the accompanying video for animated results. In Fig. 15, we validate our reconstruction results by qualitatively comparing against the photographs of the physical samples under a novel lighting condition not used in the acquisition. In addition, quantitative errors in structural similarity index (SSIM) are listed.

We further compare the our results against one state-of-the-art technique [10] in Fig. 11. Our network is designed to efficiently probe the 4D view-illumination domain, while [10] samples the illumination only and does not exploit the multi-view coherence. As a result, our network produces superior-quality reflectance over [10], with the

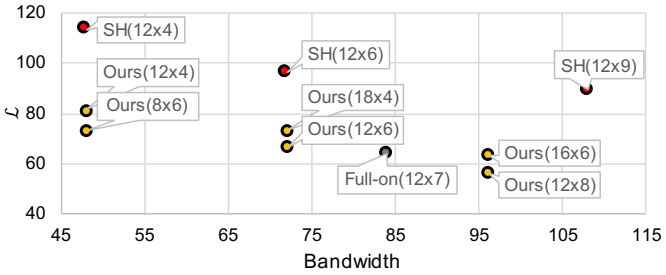


Fig. 9: Comparisons of prediction qualities of networks with different parameters. Parameters #v and #l are indicated in parentheses as (#v×#l). We also add one full-on pattern per-view to our learned patterns (12×6) for comparison. The network loss $\mathcal{L}$ is computed on the validation dataset, according to Eq. 10.

same input bandwidth (#v=12, #l=6). In Fig. 11-c, the direction of anisotropic reflection is not correctly estimated, and the specular color on the right part of the fabric is reconstructed as green, rather than gold. Note that here we use the same geometry computed with our approach, to exclude the impact of geometry over appearance reconstruction. For a more direct/end-to-end comparison, please refer to Fig. 12, where the geometry is obtained with our approach and [10], respectively. Our appearance reconstruction achieves a higher quality in comparison with [10] at the same input bandwidth (#v=12, #l=6).

8.2 Evaluations

In Fig. 9, we evaluate the impact of various parameters/inputs over the prediction quality of our network. The horizontal axis represents the input bandwidth for reflectance reconstruction (i.e., #v×#l), and the vertical axis indicates the network loss $\mathcal{L}$ defined in Eq. 10. Note that the loss is computed on the synthetic validation dataset, rather than the surface points of a specific physical object. Part of the corresponding lighting patterns can be visualized in Fig. 8.

The first thing to observe is that our prediction quality improves with the increase of the input bandwidth. There are 3 groups of networks with equal bandwidth in the figure (bandwidth = 48, 72 and 96). Our network learns to exploit the available information measured in the 4D view-illumination domain to reduce the prediction error. In addition, for each group of the same input bandwidth, allocating more sampling efforts to the illumination domain results in a slightly lowered $\mathcal{L}$. This suggests that we should use more lighting patterns and less view angles in sampling. However, in practice, the number of sampled view angles is constrained by the complexity of the geometry and cannot be arbitrarily reduced. Otherwise, certain surface points may receive no reflectance measurements at all.

Next, we evaluate two other techniques. Instead of using learned patterns, we fix the patterns to spherical harmonics (marked as SH in Fig. 9, with #v=12 and #l=4, 6 and 9). The network loss using SH is considerably higher, compared to our network of the same bandwidth. The main reason is that in these cases, the light patterns are fixed and cannot fully enjoy the power of the joint optimization of the entire

SSIM = 0.904 SSIM = 0.898

Fig. 10: Impact of geometry accuracy over appearance reconstruction. From the left to right: coarse geometry from COLMAP, our refined mesh, and corresponding appearance reconstructions on the coarse/refined shape.

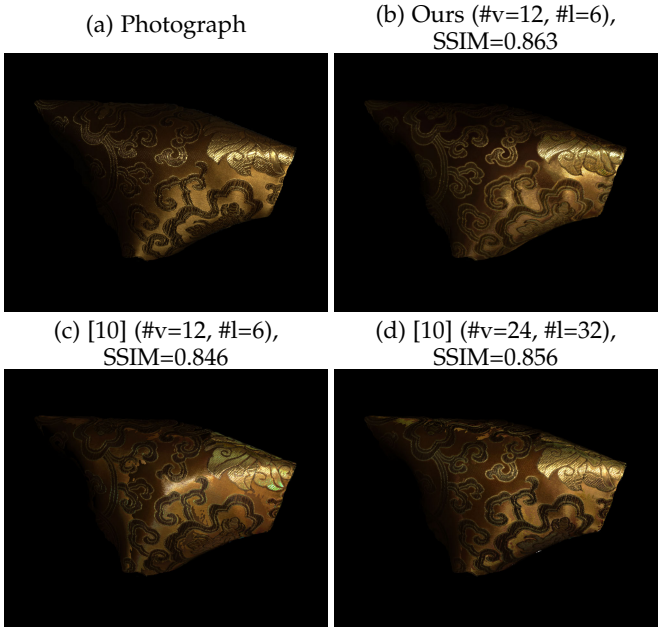


Fig. 11: Comparisons of results using our networks and the previous work of [10], with different sampling parameters: a photograph (a); our result with #v=12, #l=6 (b); the results of [10], with the same number of input images as ours (#v=12, #l=6) (c), and the much higher number of images used in the original paper (#v=24, #l=32) (d). While [10] produces reasonable results at a high input bandwidth in (d), artifacts such as inaccurate anisotropic reflection direction and overly green specular color appear in (c), when using the same low bandwidth as ours.

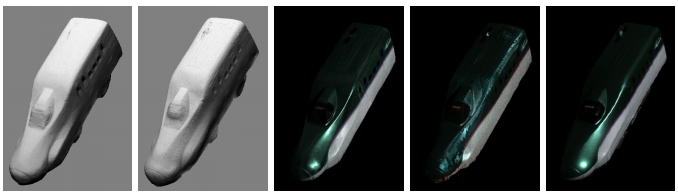


Fig. 12: Comparisons of results using our network (#v=12, #l=6) and the previous work of [10] (#v=12, #l=6), with the same input bandwidth. From the left to right: our geometry reconstruction, the shape computed with [10], the appearance reconstructed by our method/ [10], and the corresponding photograph.

acquisition pipeline. In addition, we augment our network (#v=12, #l=6) with a "free" fixed full-on pattern (marked as Full-on in Fig. 9), as it is used in the geometry reconstruction anyway. As expected, the prediction loss is slightly reduced with the extra input information.

In a pilot study, we also test learning a different set of lighting patterns for each different view, to gain more input information with more varied patterns. However, the initial results are not as good as our current network with the same set of patterns for each view. We believe that even though the amount of measured information is expected to be higher, this network with view-varying patterns cannot support rotational equivariance, leading to inferior performance. In comparison, the structure of our network is designed to explicitly enforce the rotational equivariance in the first place (Fig. 4), and can be trained more efficiently.

In Fig. 10, we evaluate the sensitivity of our approach with respect to the accuracy of the geometry. Appearance reconstruction results on the coarse shape from COLMAP and our refined mesh show that our approach is tolerant to minor geometric inaccuracies.

Finally, we test the ability of our network to make full use of multi-view input information in Fig. 16. As the number of visible input views decreases, the quality of reconstructed lumitexels generally lowers as well, due to the lack of information. On the other hand, it shows that our network successfully learns to aggregate the multi-view measured information whenever it is available.

9 Limitations & Future Work

Our work is subject to a number of limitations. First, similar to previous work [9], [10], [16], we do not explicitly handle inter-reflections or self-shadowing. In addition, as a data-driven approach, our network cannot produce results that substantially deviate from the training samples (see Fig. 17 for an example). Also, the reflectance cannot be recovered, if it is not observed from any input view.

We hope that this work will inspire future research along various directions. It will be interesting to consider polarization as an extra channel for multiplexing to further increase the acquisition efficiency. We are also interested in handling more challenging appearance (e.g., with strong scattering effects). It would be promising to further extend our work to less or even uncontrolled illumination conditions. Finally, it is of practical value to automatically determine the parameters in our networks (e.g., #v/#l) from a rapid initial scan.

Acknowledgments

The authors would like to thank Ruisheng Zhu, Yaxin Yu, Xiaohe Ma and Mingqi Yi for their help. This work is partially supported by NSF China (61772457, 62022072 & 61890954).

References

[1] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *CVPR*, 2003.
[2] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *ECCV*, 2016.

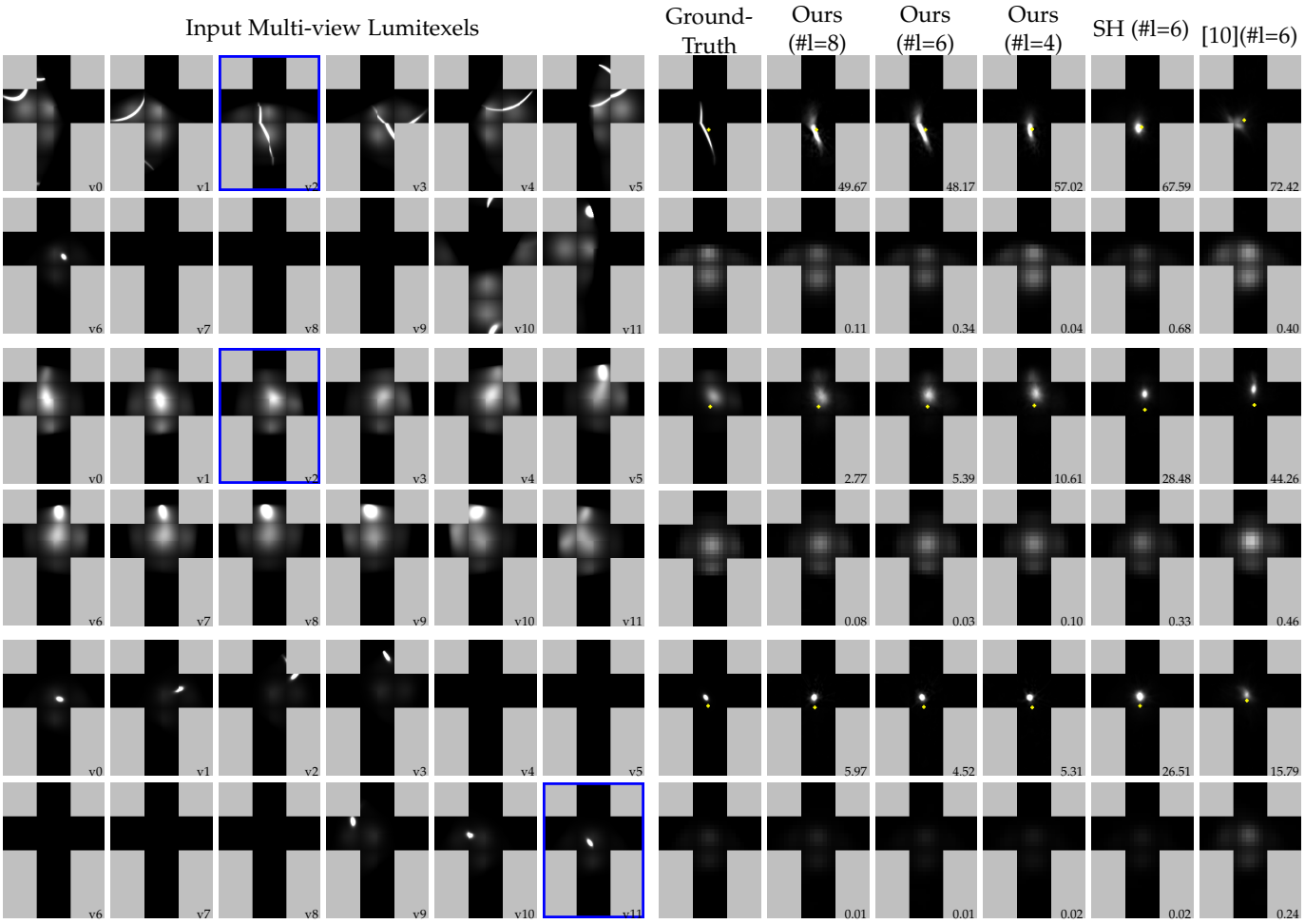


Fig. 13: Reconstruction results on synthetic lumitexels. For every two rows, the left six columns are the input multi-view lumitexels, with the view index marked at the bottom-right corner, and the primary view shown in a blue box; the right six columns are the ground-truths and the reconstruction results using different methods, with the specular lumitexel shown in the first row, the diffuse one in the second row, and the normal indicated with a yellow cross; the lumitexel reconstruction loss ($\mathcal{L}_s/\mathcal{L}_d$) is marked at the bottom-right corner, computed with Eq. 11/12. All results are the direct network outputs before fitting. The sampled view number #v is 12 in all cases.

[3] K. J. Dana, B. van Ginneken, S. K. Nayar, and J. J. Koenderink, "Reflectance and texture of real-world surfaces," *ACM Trans. Graph.*, vol. 18, no. 1, pp. 1–34, Jan. 1999.

[4] J. Lawrence, A. Ben-Artzi, C. DeCoro, W. Matusik, H. Pfister, R. Ramamoorthi, and S. Rusinkiewicz, "Inverse shade trees for non-parametric material representation and editing," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 735–745, Jul. 2006.

[5] H. P. A. Lensch, J. Kautz, M. Goesele, W. Heidrich, and H.-P. Seidel, "Image-based reconstruction of spatial appearance and geometric detail," *ACM Trans. Graph.*, vol. 22, no. 2, pp. 234–257, Apr. 2003.

[6] S. Bi, Z. Xu, K. Sunkavalli, D. Kriegman, and R. Ramamoorthi, "Deep 3D Capture: Geometry and reflectance from sparse multi-view images," in *CVPR*, 2020.

[7] A. Ghosh, T. Chen, P. Peers, C. A. Wilson, and P. Debevec, "Estimating specular roughness and anisotropy from second order spherical gradient illumination," *CGF*, vol. 28, no. 4, pp. 1161–1170, 2009.

[8] K. Kang, Z. Chen, J. Wang, K. Zhou, and H. Wu, "Efficient reflectance capture using an autoencoder," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 127:1–127:10, Jul. 2018.

[9] B. Tunwattanapong, G. Fyffe, P. Graham, J. Busch, X. Yu, A. Ghosh, and P. Debevec, "Acquiring reflectance and shape from continuous spherical harmonic illumination," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 109:1–109:12, Jul. 2013.

[10] K. Kang, C. Xie, C. He, M. Yi, M. Gu, Z. Chen, K. Zhou, and H. Wu, "Learning efficient illumination multiplexing for joint capture of reflectance and shape," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 165:1–165:12, Nov. 2019.

[11] T. Weyrich, J. Lawrence, H. P. A. Lensch, S. Rusinkiewicz, and T. Zickler, "Principles of appearance acquisition and representation," *Found. Trends. Comput. Graph. Vis.*, vol. 4, no. 2, pp. 75–191, 2009.

[12] D. Guarnera, G. C. Guarnera, A. Ghosh, C. Denk, and M. Glencross, "Brdf representation and acquisition," *CGF*, vol. 35, no. 2, pp. 625–650, 2016.

[13] M. Weinmann and R. Klein, "Advances in geometry and reflectance acquisition," in *SIGGRAPH Asia Courses*, 2015, pp. 1:1–1:71.

[14] Y. Dong, "Deep appearance modeling: A survey," *Visual Informatics*, vol. 3, no. 2, pp. 59 – 68, 2019.

[15] C. Schwartz, M. Weinmann, R. Ruiters, and R. Klein, "Integrated high-quality acquisition of geometry and appearance for cultural heritage." in *VAST*, vol. 2011, 2011, pp. 25–32.

[16] G. Nam, J. H. Lee, D. Gutierrez, and M. H. Kim, "Practical svbrdf acquisition of 3d objects with unstructured flash photography," in *SIGGRAPH Asia Technical Papers*, 2018, p. 267.

[17] H. Wu and K. Zhou, "AppFusion: Interactive appearance acquisition using a Kinect sensor," *CGF*, vol. 34, no. 6, pp. 289–298, 2015.

[18] T. Zickler, S. Enrique, R. Ramamoorthi, and P. Belhumeur, "Reflectance sharing: Image-based rendering from a sparse set of images," in *Proc. EGSR*, 2005, pp. 253–264.

[19] Y. Dong, J. Wang, X. Tong, J. Snyder, Y. Lan, M. Ben-Ezra, and

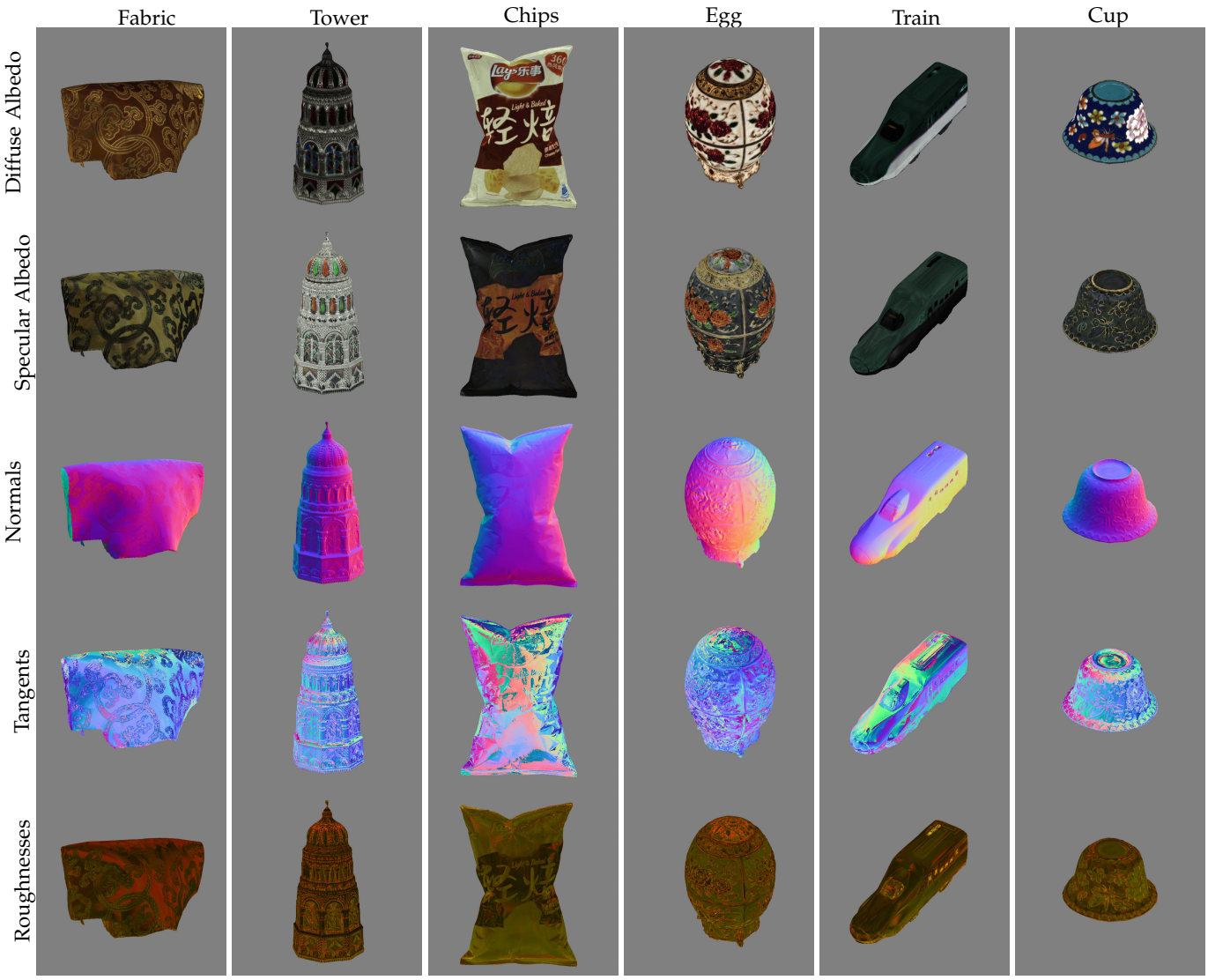


Fig. 14: Reflectance reconstruction results with our network (#v=12, #l=6). Each normal/tangent is added with $(1, 1, 1)$ and then divided by 2 to fit to the range of $[0, 1]^3$ for visualization. The roughness α_x/α_y is visualized in the red/green channel.

B. Guo, "Manifold bootstrapping for svbrdf capture," *ACM Trans. Graph.*, vol. 29, no. 4, pp. 98:1–98:10, Jul. 2010.

[20] J. Wang, S. Zhao, X. Tong, J. Snyder, and B. Guo, "Modeling anisotropic surface reflectance with example-based microfacet synthesis," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 41:1–41:9, Aug. 2008.

[21] M. Holroyd, J. Lawrence, and T. Zickler, "A coaxial optical scanner for synchronous acquisition of 3d geometry and surface reflectance," *ACM Trans. Graph.*, vol. 29, no. 4, pp. 99:1–99:12, Jul. 2010.

[22] M. Aittala, T. Weyrich, and J. Lehtinen, "Two-shot svbrdf capture for stationary materials," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 110:1–110:13, Jul. 2015.

[23] M. Aittala, T. Aila, and J. Lehtinen, "Reflectance modeling by neural texture synthesis," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 65:1–65:13, Jul. 2016.

[24] W. Matusik, H. Pfister, M. Brand, and L. McMillan, "Efficient isotropic brdf measurement," in *Proc. EGWR*, 2003, pp. 241–247.

[25] J. B. Nielsen, H. W. Jensen, and R. Ramamoorthi, "On optimal, minimal brdf sampling for reflectance acquisition," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 186:1–186:11, Oct. 2015.

[26] Z. Xu, J. B. Nielsen, J. Yu, H. W. Jensen, and R. Ramamoorthi, "Minimal brdf sampling for two-shot near-field reflectance acquisition," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 188:1–188:12, Nov. 2016.

[27] X. Li, Y. Dong, P. Peers, and X. Tong, "Modeling surface appearance from a single photograph using self-augmented convolutional neural networks," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 45:1–45:11, Jul. 2017.

[28] V. Deschaintre, M. Aittala, F. Durand, G. Drettakis, and A. Bousseau, "Single-image svbrdf capture with a rendering-aware deep network," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 128:1–128:15, Jul. 2018.

[29] ——, "Flexible svbrdf capture with a multi-image deep network," *CGF*, vol. 38, no. 4, pp. 1–13, 2019.

[30] Z. Li, Z. Xu, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker, "Learning to reconstruct shape and spatially-varying reflectance from a single image," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 269:1–269:11, Dec. 2018.

[31] D. Gao, X. Li, Y. Dong, P. Peers, K. Xu, and X. Tong, "Deep inverse rendering for high-resolution svbrdf estimation from an arbitrary number of images," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 134:1–134:15, Jul. 2019.

[32] Y. Guo, C. Smith, M. Hašan, K. Sunkavalli, and S. Zhao, "Materialgan: Reflectance capture using a generative svbrdf model," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 254:1–254:13, 2020.

[33] Z. Xu, S. Bi, K. Sunkavalli, S. Hadap, H. Su, and R. Ramamoorthi, "Deep view synthesis from sparse photometric images," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 76:1–76:13, Jul. 2019.

[34] D. Gao, G. Chen, Y. Dong, P. Peers, K. Xu, and X. Tong, "Deferred neural lighting: free-viewpoint relighting from unstructured photographs," *ACM*, vol. 39, no. 6, pp. 1–15, 2020.

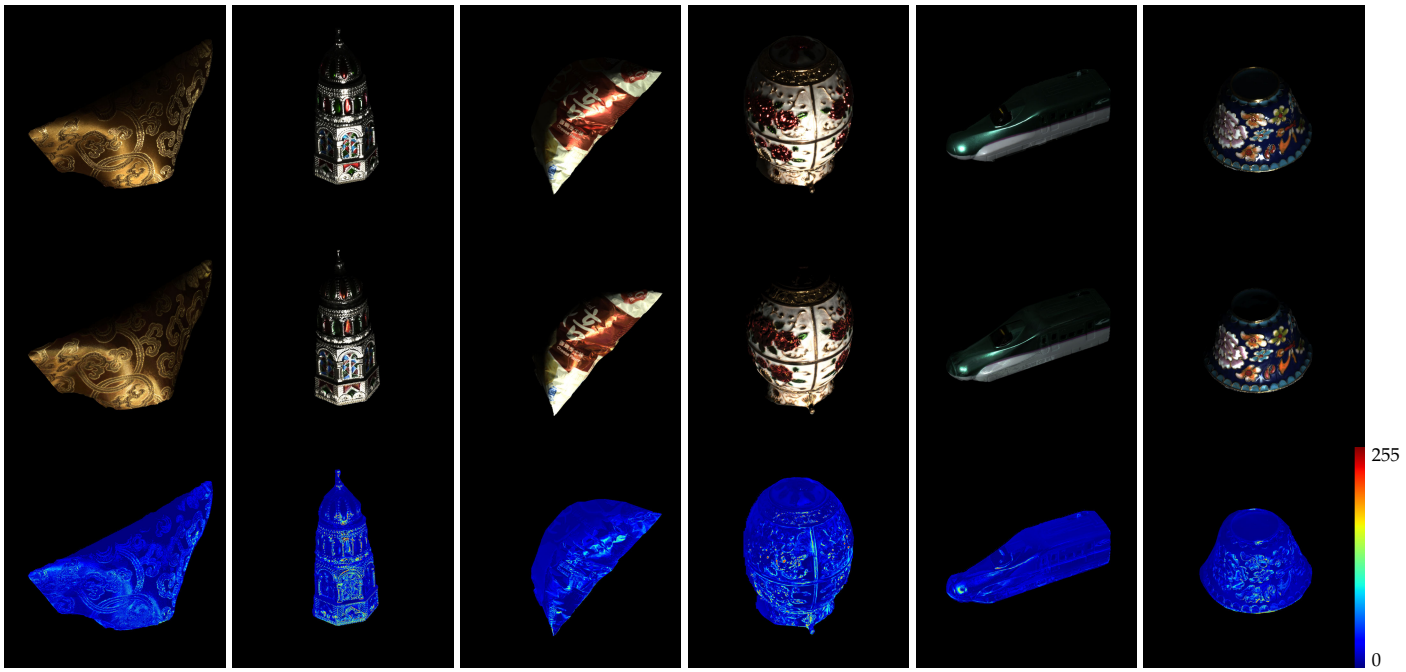


Fig. 15: Validation results. The top row shows photographs of physical objects, the second row are the rendered images of the reconstruction results with our network (#v=12, #l=6), and the last row shows color-coded differences. The quantitative errors of our results with respect to the photographs are reported at the bottom in SSIM.

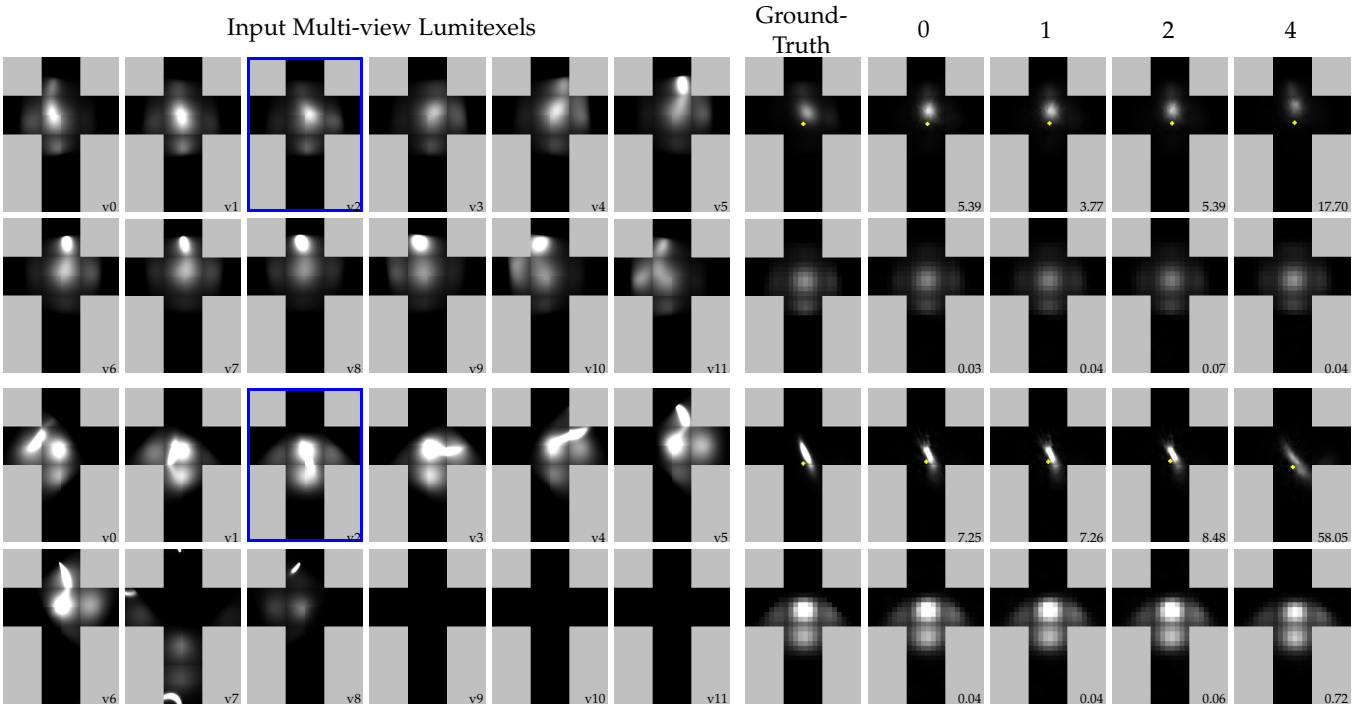


Fig. 16: Impact of the reduced number of visible input views on our network. For every two rows, the left six columns are the input multi-view lumitexels, with the view index marked at the bottom-right corner, and the primary view shown in a blue box; the seventh column are the ground-truths. Starting from the eighth column, the number marked on top indicates the number of visible input views that are intentionally zeroed out. We begin with the views distant from the primary one and move gradually towards it. The lumitexel reconstruction loss ($\mathcal{L}_s/\mathcal{L}_d$) is marked at the bottom-right corner, computed with Eq. 11/12.

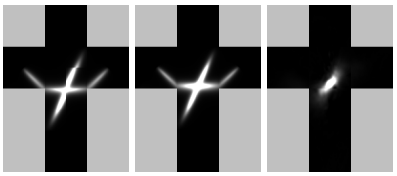


Fig. 17: A failure case. From the left to right, the input multi-view lumitexel at the primary view, the ground-truth output and our reconstruction.

[35] D. Den Brok, H. C. Steinhausen, M. B. Hullin, and R. Klein, "Multiplexed acquisition of bidirectional texture functions for materials," in *Measuring, Modeling, and Reproducing Material Appearance 2015*, vol. 9398. SPIE, 2015, p. 93980F.

[36] G. Nam, J. H. Lee, H. Wu, D. Gutierrez, and M. H. Kim, "Simultaneous acquisition of microscale reflectance and normals," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 185:1–185:11, Nov. 2016.

[37] A. Meka, C. Häne, R. Pandey, M. Zollhöfer, S. Fanello, G. Fyffe, A. Kowdle, X. Yu, J. Busch, J. Dourgarian, P. Denny, S. Bouaziz, P. Lincoln, M. Whalen, G. Harvey, J. Taylor, S. Izadi, A. Tagliasacchi, P. Debevec, C. Theobalt, J. Valentin, and C. Rhemann, "Deep reflectance fields: High-quality facial reflectance field inference from color gradient illumination," *ACM Trans. Graph.*, vol. 38, no. 4, Jul. 2019.

[38] A. Gardner, C. Tchou, T. Hawkins, and P. Debevec, "Linear light source reflectometry," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 749–758, 2003.

[39] P. Ren, J. Wang, J. Snyder, X. Tong, and B. Guo, "Pocket reflectometry," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–10, 2011.

[40] G. Chen, Y. Dong, P. Peers, J. Zhang, and X. Tong, "Reflectance scanning: Estimating shading frame and brdf with generalized linear light sources," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 117:1–117:11, Jul. 2014.

[41] M. Aittala, T. Weyrich, and J. Lehtinen, "Practical SVBRDF capture in the frequency domain," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 110:1–110:12, Jul. 2013.

[42] M. Fiala, "Artag, a fiducial marker system using digital techniques," in *CVPR*, June 2005.

[43] B. Walter, S. R. Marschner, H. Li, and K. E. Torrance, "Microfacet Models for Refraction through Rough Surfaces," in *Proc. EGWR*, 2007.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[45] X. Ma, K. Kang, R. Zhu, H. Wu, and K. Zhou, "Free-form scanning of non-planar appearance with neural trace photography," *ACM Trans. Graph.*, vol. 40, no. 4, Jul. 2021.

[46] M. Kazhdan and H. Hoppe, "Screened poisson surface reconstruction," *ACM Trans. Graph.*, vol. 32, no. 3, pp. 29:1–29:13, Jul. 2013.

[47] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi, "Efficiently combining positions and normals for precise 3D geometry," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 536–543, Jul. 2005.

[48] J. L. Morales and J. Nocedal, "Remark on "algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound constrained optimization"," *ACM Trans. Math. Softw.*, vol. 38, no. 1, pp. 7:1–7:4, Dec. 2011.

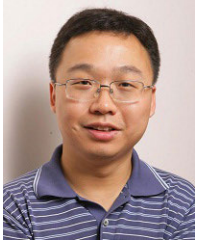
Kaizhang Kang is currently a Ph.D. student in the State Key Lab of CAD & CG, Zhejiang University. He received his B.Eng. degree from College of Computer Science & Technology, and Honors Degree from Chu Kochen Honors College, Zhejiang University, in 2018. His research interests include appearance acquisition/modeling and rendering.

Minyi Gu is a master student in the State Key Lab of CAD & CG, Zhejiang University. She received her B.Eng. from the same university in 2018. Her research interests include appearance acquisition and rendering.

Cihui Xie received his B.Eng./M.Eng.in Computer Science from Zhejiang University in 2017/2020. His research interests include physically based rendering and computational photography.

Xuanda Yang received his B.Eng. in Computer Science from Zhejiang University in 2020. He is now a Ph.D. student at University of California San Diego. His research interests include appearance acquisition, rendering and high-performance computer graphics.

Hongzhi Wu is an associate professor in the State Key Lab of CAD & CG, Zhejiang University. He received B.Sc. in computer science from Fudan University in 2006, and Ph.D. in computer science from Yale University in 2012. His current research interests include high-density illumination multiplexing devices and differentiable acquisition. He has served on the program committees of conferences including PG, EGSR and HPG.



Kun Zhou is a Cheung Kong Professor in the Computer Science Department of Zhejiang University, and the Director of the State Key Lab of CAD & CG. Prior to joining Zhejiang University in 2008, Dr. Zhou was a Leader Researcher of the Internet Graphics Group at Microsoft Research Asia. He received his B.S. degree and Ph.D. degree in computer science from Zhejiang University in 1997 and 2002, respectively. His research interests are in visual computing, parallel computing, human computer interaction, and virtual reality. He is a Fellow of IEEE.