Deep View Synthesis from Sparse Photometric Images

ZEXIANG XU, University of California, San Diego
SAI BI, University of California, San Diego
KALYAN SUNKAVALLI, Adobe Research
SUNIL HADAP*, Lab126, Amazon
HAO SU, University of California, San Diego
RAVI RAMAMOORTHI, University of California, San Diego

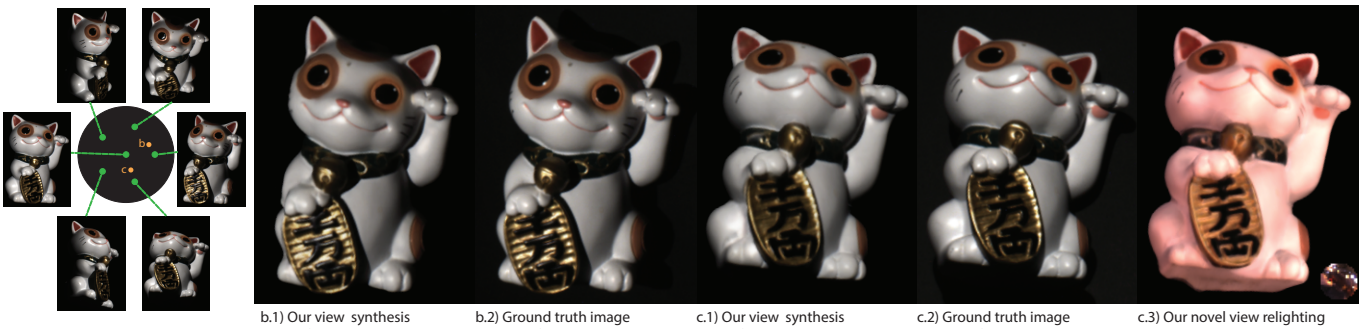

Fig. 1. We present a method to synthesize scene appearance from a novel view by interpolating only six wide-baseline images (a). We do this by using a structured setup to capture photometric images under directional lighting and interpolating them using a novel deep neural network. Our method can reproduce complex appearance effects like specularities, shadows, and occlusions (b.1,c.1) resulting in images that are close to ground truth captured images (b.2,c.2). These results can be combined with image-based relighting methods to visualize the scene under novel view and lighting (c.3).

The goal of light transport acquisition is to take images from a sparse set of lighting and viewing directions, and combine them to enable arbitrary relighting with changing view. While relighting from sparse images has received significant attention, there has been relatively less progress on view synthesis from a sparse set of "photometric" images—images captured under controlled conditions, lit by a single directional source; we use a spherical gantry to position the camera on a sphere surrounding the object. In this paper, we synthesize novel viewpoints across a wide range of viewing directions (covering a 60° cone) from a sparse set of just six viewing directions. While our approach relates to previous view synthesis and image-based rendering techniques, those methods are usually restricted to much smaller baselines, and are captured under environment illumination. At our baselines, input images have few correspondences and large occlusions; however we benefit from structured photometric images. Our method is based on a deep convolutional network trained to directly synthesize new views from the six input views. This network combines 3D convolutions on a plane sweep volume with a novel per-view per-depth plane attention map prediction network to effectively aggregate multi-view appearance. We train our network with a large-scale synthetic dataset of 1000 scenes with complex geometry and material properties. In practice, it is able to synthesize novel viewpoints for captured real data and reproduces complex appearance effects like occlusions, view-dependent specularities and hard shadows. Moreover, the method can also be combined with previous relighting techniques to enable changing both lighting and view, and applied to computer vision problems like multiview stereo from sparse image sets.

CCS Concepts: • **Computing methodologies → Image-based rendering**.

Additional Key Words and Phrases: appearance acquisition, novel view synthesis

ACM Reference Format:
Zexiang Xu, Sai Bi, Kalyan Sunkavalli, Sunil Hadap, Hao Su, and Ravi Ramamoorthi. 2019. Deep View Synthesis from Sparse Photometric Images. *ACM Trans. Graph.* 38, 4, Article 76 (July 2019), 13 pages. https://doi.org/10.1145/3306346.3323007

1 INTRODUCTION

A central problem in computer graphics and vision is to acquire images of a scene and reproduce its appearance under arbitrary lighting and viewpoint. This has traditionally been accomplished by densely sampling the scene's "reflectance field" [Debevec et al. 2000] and interpolating these images using a combination of image-based rendering and relighting methods. Recent work has demonstrated image-based relighting from sparse "photometric" images captured

*This work was done prior to joining Amazon.

Authors' addresses: Zexiang Xu, University of California, San Diego, zexiangxu@cs.ucsd.edu; Sai Bi, University of California, San Diego, bisai@cs.ucsd.edu; Kalyan Sunkavalli, Adobe Research, sunkaval@adobe.com; Sunil Hadap, Lab126, Amazon, sunilhadap@acm.org; Hao Su, University of California, San Diego, haosu@eng.ucsd.edu; Ravi Ramamoorthi, University of California, San Diego, ravir@cs.ucsd.edu.

under controlled directional lighting [Xu et al. 2018], thus significantly reducing the acquisition cost. However, novel view synthesis methods still require a dense sampling of a scene's "light field" [Gortler et al. 1996; Levoy and Hanrahan 1996]. Consequently, they capture hundreds of images, especially when the scene has complex surface reflectance [Wood et al. 2000]. While recent work has addressed novel view synthesis from sparse images [Flynn et al. 2016; Kalantari et al. 2016; Srinivasan et al. 2017], these methods are highly restricted in the range of viewpoints they can synthesize.

Our goal is to make appearance acquisition and rendering more practical by synthesizing *a wide range of novel viewpoints* from *a sparse set of images*. To do so, we image the scene, consisting of several objects, with six cameras placed on a vertex and the centers of the adjoining faces of a regular icosahedron (see Fig. 2). This results in a central camera, and five distributed symmetrically around it at an angle of about 37°. At this large baseline, the captured images have significant occlusions (see Fig. 1(b)). State-of-the-art multi-view stereo methods fail to reconstruct complete geometry from such sparse views. Yet, we show that our method can interpolate the entire convex hull of these six viewpoints — a cone of more than 60° — while accurately reproducing effects like complex occlusions and high-frequency, view-dependent specularities (see Fig. 1(c)).

This is made possible by a combination of our structured acquisition procedure and a novel learning-based interpolation scheme. Unlike previous view synthesis approaches that capture images under environment illumination, we acquire images under a single directional light. These "photometric" images capture appearance information like shading, shadows, and specularities, and have been used for scene reconstruction via Photometric Stereo methods [Woodham 1980] (after which they are named) and image-based relighting [Debevec et al. 2000]. We show that using such inputs leads to view synthesis results that capture detailed scene appearance; this also enables other applications like novel view relighting (see Fig. 1, 8).

We introduce a novel deep convolutional neural network that learns to interpolate the wide-baseline photometric images and render arbitrary output viewpoints between them. Similar to previous learning-based view synthesis methods [Flynn et al. 2016], our network projects the input images onto multiple depth planes of the output view to construct a view-dependent plane sweep volume. This volume is processed to predict the final output view (and depth) using 3D convolutional layers with downsampling and upsampling operations; this allows the network to reason about both geometry and appearance at multiple spatial scales. A key, novel component of our method is that we explicitly predict *per-plane per-input-view attention maps*. These attention maps are used to modulate the 3D plane sweep volume and allow the network to aggregate multi-view appearance while accounting for occlusions, viewpoint variations, etc. (similar to blending weights in IBR methods), and lead to sharper, more accurate results. We supervise the output image and depth map, and our network learns to predict the attention maps in an unsupervised fashion.

We train our CNN with a large-scale, synthetic dataset consisting of 1000 scenes with procedurally generated shapes and complex spatially-varying BRDFs. We render these scenes with our six-view camera configuration under random directional lighting using path tracing. The rendered images approximate real-world appearance and light transport well, allowing the network to generalize well to real captured scenes. This can be seen in Fig. 1, where our method generates photorealistic interpolated results for a real object with complex geometry and appearance effects, including challenging occlusions, high-frequency specularities and cast shadows.

We also demonstrate extensions of our approach that go beyond view synthesis. We can add multiple lights to our acquisition setup to capture a sparse set of multi-view, multi-light images. Our view synthesis network can be extended to synthesize appearance under *novel view and lighting* from this sparse data (see Fig. 1c3). We also show that our view synthesis network can be used to "densify" the captured views of a scene; these dense views can then be input to a multi-view stereo algorithm to produce reconstructions that are significantly better than what is possible with just the sparse inputs (see Fig. 9). By making these methods work with sparse image sets, our work takes a step towards eliminating the need for dense capture systems and making scene acquisition more practical.

2 RELATED WORKS

Light transport acquisition. Traditionally, light transport acquisition methods use complex systems to capture images of a scene under a dense set of lighting and view directions. These samples can then be used to explicitly reconstruct scene geometry and reflectance [Debevec et al. 2000; Furukawa et al. 2002; Holroyd et al. 2010; Schwartz et al. 2011; Zhou et al. 2013]. Recent methods have demonstrated geometry and reflectance reconstruction under more relaxed settings such as handheld capture [Nam et al. 2018] and unknown environment lighting [Xia et al. 2016]. For a more detailed discussion of the previous work on appearance acquisition, we refer readers to [Weinmann and Klein 2015; Weyrich et al. 2009]. While these methods can produce high-quality rendering results, they still require capturing hundreds of images.

This requirement can be relaxed for specific applications. For example, image-based relighting methods focus on directly combining images captured under varying lighting conditions to relight the scene under novel lighting. While earlier image-based relighting methods required brute-force sampling [Debevec et al. 2000; Malzbender et al. 2001], recent methods have leveraged light transport coherence to reduce the number of images required [Peers et al. 2009]. In particular, Xu et al. [2018] demonstrate image-based relighting from only five images. Sparse capture has also been shown to be sufficient for reflectance estimation for planar samples [Hui et al. 2017; Li et al. 2018a; Xu et al. 2016], scenes with known geometry [Zhou et al. 2016a] or single-view reflectance [Barron and Malik 2015; Li et al. 2018b]. Our work focuses on capturing multi-view scene appearance and can reproduce complex effects like view-dependent specularities and occlusions from only six images. Moreover, our method can be extended to capture scene appearance under both varying view and lighting from only 36 images. This is a significant change from traditional light transport acquisition methods that require hundreds to thousands of images.

Novel view synthesis. Novel view synthesis methods focus on interpolating the appearance of the scene between captured images [Chen and Williams 1993], i.e., the view interpolation aspect of the

light transport acquisition problem. This can be done by re-sampling rays from a densely captured light field [Gortler et al. 1996; Levoy and Hanrahan 1996]; such light fields can be also reconstructed from sparse samples by leveraging various forms of correspondence between multiple views [Dąbała et al. 2016; Vagharshakyan et al. 2018; Yao et al. 2016]. Alternatively, image-based rendering methods project captured images onto proxy geometry (that can be given or reconstructed from the captured images), blend and resample them from the novel viewpoint [Buehler et al. 2001; Chaurasia et al. 2013, 2011; Debevec et al. 1996; Sinha et al. 2009].

Penner and Zhang [2017] propose soft 3D, which reconstructs depths and visibilities for each input view to achieve better blending. Other works present different blending techniques to achieve ghosting free synthesis results for inaccurate 3D reconstruction [Bi et al. 2017; Eisemann et al. 2008; Zhou and Koltun 2014]. Hedman et al. [2018] present a method that learns to predict the blending weights. These blending techniques require scanned or MVS-reconstructed geometry as an input, for which densely sampled viewpoints are necessary. All these methods rely on capturing tens to hundreds of images with large overlap and reasonable baseline; without this, the geometric reconstruction and view synthesis would fail. Moreover, these methods are usually designed for free-viewpoint navigation and do not focus on reproducing detailed appearance like sharp specularities. In contrast, our method works with only six images that are captured with a fairly large baseline, and reproduces complex scene appearance accurately. It does so in an end-to-end fashion and learns to predict appearance, depths and attention/blending maps.

Surface light fields [Wood et al. 2000] are designed to capture complex scene appearance under high-frequency point lighting. These methods reconstruct the surface geometry and represent appearance on the surface using lumispheres. Because of the high-dimensionality of this representation, methods have focused on compressing this data using PCA and vector quantization [Wood et al. 2000] or deep networks [Chen et al. 2018]. Our work is able to capture similar appearance effects with vastly fewer images. Moreover, our network can be thought of as a scene-agnostic representation that can interpolate any scene's images to a new viewpoint.

Learning-based novel view synthesis. Recently, deep learning techniques have been applied to novel view synthesis to achieve unstructured multi-view interpolation [Flynn et al. 2016], narrow-baseline stereo extrapolation [Zhou et al. 2018], narrow-baseline interpolation [Kalantari et al. 2016] and single-view extrapolation [Srinivasan et al. 2017] in the context of light fields. All these methods estimate geometry; either a single depth [Kalantari et al. 2016; Srinivasan et al. 2017] or per-plane depth probabilities (or blending weights) [Flynn et al. 2016; Zhou et al. 2018] are predicted for either one input view [Srinivasan et al. 2017; Zhou et al. 2018] or each novel view [Flynn et al. 2016; Kalantari et al. 2016]. All these methods resolve visibilities in an implicit way, whereas our network learns multi-view correspondence and explicitly predicts per-input-view visibility-aware attention maps jointly with depth probability maps at a novel viewpoint. Most importantly, we demonstrate that we can synthesize a much wider range of new viewpoints compared to these works.

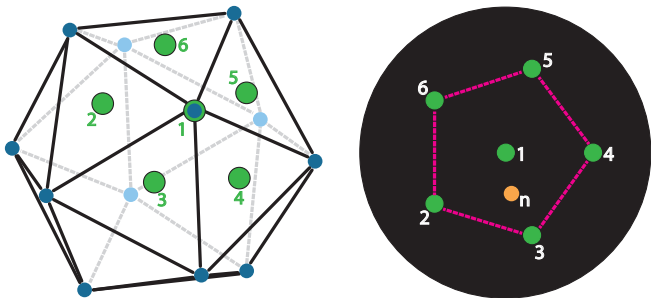

Fig. 2. Acquisition configuration. Left: a regular convex icosahedron with 12 spherically symmetric vertices. Right: a projective view of our configuration, where the black background circle represents a hemisphere towards the central view. We consider a setup with six known views (green circles, denoted by numbers 1-6), in which one is located at a vertex and five at the adjoining face centers. Our goal is to synthesize a novel view (orange circle, noted by n), in the convex hull of the six known views (red dash-lines).

Geometry-free learning-based methods can directly generate pixels for a novel view from one or multiple input views [Tatarchenko et al. 2015; Yang et al. 2015], though these methods are restricted to specific shape classes. Other methods leverage flow prediction to warp pixels from source views [Park et al. 2017; Zhou et al. 2016b]. Flow-based warping has been combined with per-input-view confidence maps to improve aggregation [Sun et al. 2018]. While we also use attention maps, ours are predicted from geometric correspondences, inferred jointly with depths, and incorporate information about occlusions, viewpoint differences, etc. We show that this leads to results that are more realistic than flow-based methods.

Learning-based appearance acquisition. Deep learning-based methods have been applied to appearance acquisition applications like reflectance capture [Li et al. 2017], reflectance map estimation [Rematas et al. 2016], and depth estimation [Eigen and Fergus 2015]. Using photometric images has been shown to be better for single-shot BRDF acquisition [Deschaintre et al. 2018; Li et al. 2017, 2018b], and image-based relighting from sparse samples [Xu et al. 2018]. All these methods focus on the single-view setting. In this work, we explore multi-view appearance acquisition using photometric measurements and achieve photorealistic novel view synthesis under a single directional light from six sparse views.

3 ACQUISITION CONFIGURATION

Similar to previous work on reflectance field acquisiton [Debevec et al. 2000; Weyrich et al. 2006], we choose an acquisition setup where cameras are placed on a sphere. We assume that the cameras are distant with respect to the scale of the scene and image the scene with a field-of-view that ensures sufficient pixel coverage. Our goal is to acquire the appearance of a generic scene, without making any assumptions about scene composition. Therefore we utilize a symmetric camera configuration, that would be optimal on average. Given these design decisions, we need to choose a configuration that symmetrically samples a sphere. This sampling has to balance two constraints: we would like a sparse sampling that minimizes

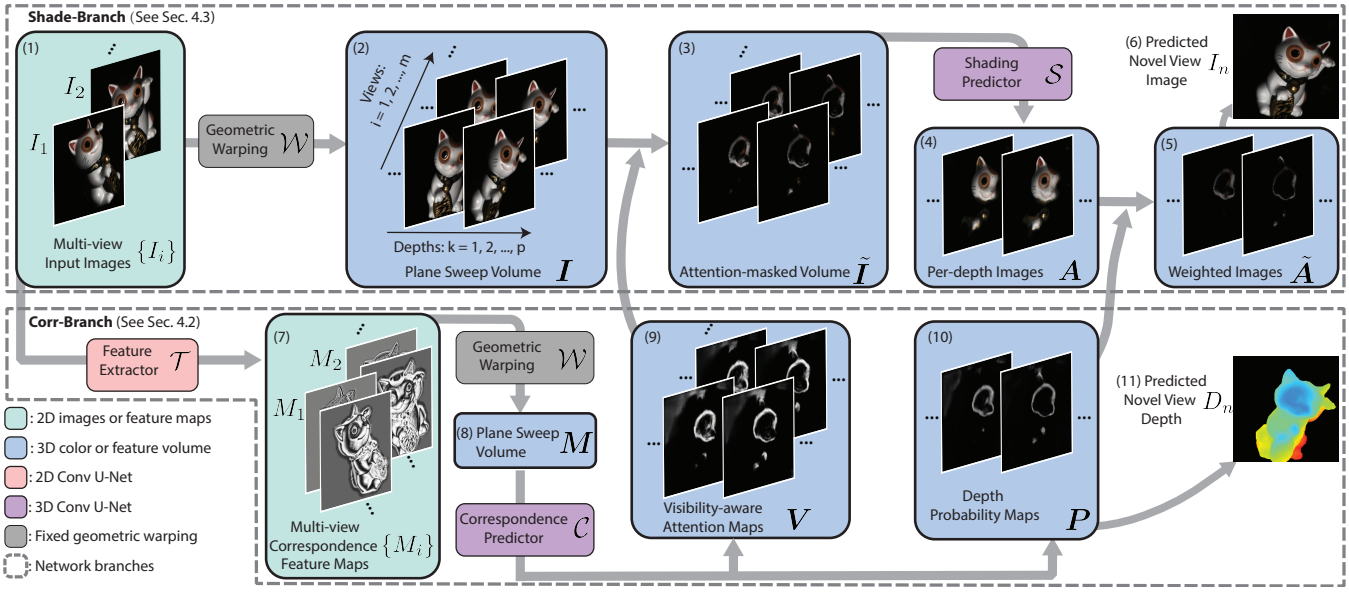

Fig. 3. Network Overview. Our view synthesis network consists of two branches that operate on plane sweep volumes (shown in blue) that are constructed using geometric warping (gray boxes). **Corr-Branch** (bottom) extracts image features (7) and estimates attention maps (9) and depth probability maps (10). The depth probabilites are used to estimate scene depth from the novel view (11). **Shade-Branch** (top) processes the input image volume using the attention maps and depth probabilities to synthesize the novel view image (6). Please refer to the supplementary material for details of $\mathcal{T}$, $\mathcal{C}$ and $\mathcal{S}$.

acquisition cost while ensuring that views have sufficient overlap to allow high-quality view interpolation.

Our configuration of choice, as shown in Fig. 2, is inspired by a standard spherically symmetric shape: the regular convex icosahedron. An icosahedron has 12 vertices with 20 equilateral triangular faces, symmetrically distributed around a sphere. Given an icosahedron, we investigate a setup with $m = 6$ cameras, in which one camera is positioned on a vertex and five "boundary" cameras are positioned at the centers of the faces that surround the central vertex. In this configuration, the angular distance from the central view to each boundary view is about 37°. Note that this is a very wide baseline, and our cameras observe very different parts of a scene with limited overlapping regions. Our goal is to synthesize an arbitrary view point in the convex hull of these six known views; this represents a cone with an angular baseline of more than 60°. Because of the symmetry of the icosahedron, our six-view setup can potentially be extended for full sphere acquisition. Cameras can be placed at every vertex and every face center, in total requiring merely 32 views. This is a very sparse camera setup with significantly fewer views compared to previous techniques [Chen et al. 2018; Wood et al. 2000] that capture hundreds of images.

We focus on a practical case where a scene is composed of one or a few objects that are placed on a flat platform. We capture photometric images of this scene lit by an arbitrary directional light from the frontal hemisphere (the black circular region in Fig. 2 right). This region is most likely to illuminate the scene and light coming from behind the scene will have little contribution. For our multi-view multi-light extension (Sec. 4.4), we use six fixed lights, with one collocated with the central camera, and five located on the surrounding neighboring five vertices. Fixing the light directions in this scenario allows the network to leverage this structured input and produce higher quality synthesis results.

4 ALGORITHM

Given images from our six pre-defined, calibrated views under a single directional light, our goal is to synthesize a new image from a specified novel view between the six inputs. Inspired by the recent success of deep learning, we propose to train a deep CNN to directly regress the final output image. Our method is a geometry-based IBR method that uses scene geometry to align and blend the multi-view data. However, instead of relying on a precomputed geometric proxy [Buehler et al. 2001; Hedman et al. 2018], we design our network to infer multi-view correspondence information for each novel view, and predict novel view depths as side products. Since we are using a synthetic dataset, the ground truth for both final images and view-dependent depths can be generated and used as supervision. Our network leverages this supervision to learn shading and correspondence simultaneously in a single end-to-end system. A key component in our network architecure, is that we modulate the input views with per-view, per-scene-depth attention maps that are learned without any supervision. We find that these attention maps indirectly learn a combination of visibility, viewpoint-based weighting, and other factors that when taken into account lead to high-quality view synthesis results. In this section, we discuss the details of our network design (Sec. 4.1, 4.2, 4.3, 4.4), data generation (Sec. 4.5) and training details (Sec. 4.6).

4.1 Inputs and basic architecture

The inputs to our network are a fixed number of m input images, I_1,I_2,..., I_m from m views under a single directional light. While our architecture can be potentially generalized to setups with other fixed number of views, we consider $m = 6$ in this paper as discussed in Sec. 3. Our network regresses an image I_n and the associated depth map D_n from a novel view point. This involves aligning and aggregating multi-view inputs, given input view camera parameters $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2$,..., $\mathbf{\Theta}_m$ and novel view camera parameters $\mathbf{\Theta}_n$.

Our network uses a plane sweep volume representation. Plane sweep volumes are constructed by warping colors or features from multiple input views to a novel view at multiple pre-defined depth planes. The warping function, denoted as $\mathcal{W}$, is a homography-based geometric function, and is implemented as a non-learnable (but differentiable) layer. Plane sweep volumes contain geometrically aligned structured data, which is suitable for deep learning-based IBR methods [Flynn et al. 2016]. To help the network better understand the multi-view data, we provide it with information about the input and output viewpoints. Specifically, we supply b_1, b_2,..., b_m, which describe the angular distances from a novel view to each input view, and the depth values d_1, d_2,..., d_p of the p planes in a plane sweep volume, as additional inputs to the network. As such, our network is a regression function $\mathbf{\Phi}$:

$$I_n, D_n = \mathbf{\Phi}(I_1, \mathbf{\Theta}_1, b_1 ..., I_m, \mathbf{\Theta}_m, b_m; d_1, ..., d_p; \mathbf{\Theta}_n) \quad (1)$$

$$= \mathbf{\Phi}(\{I_i\}, \{\mathbf{\Theta}_i\}, \{b_i\}; \{d_k\}; \mathbf{\Theta}_n), \quad (2)$$

where $\{\cdot\}$ represents a set containing either multi-view data (denoted with subscript i) or multiple depths (denoted with subscript k). We use this notation convention in the rest of the paper.

As shown in Fig. 3, our network uses two separate branches. The first, **Corr-Branch**, seeks to analyze multi-view correspondences and reconstruct scene geometry. The second branch, denoted as **Shade-Branch**, reconstructs the output image. Both the correspondence predictor and shading predictor networks use a 3D U-Net architecture, comprised of 3D convolutions, upsampling and downsampling operations, to process their corresponding plane sweep volumes [Huang et al. 2018]. This allows the network to reason about multi-view scene geometry and appearance while predicting the output view and depth.

The plane sweep volume representation contains incorrect or redundant features (e.g., features that are at incorrect depths or occluded or view-dependent). Therefore, we estimate "attention" maps that account for how much information should be used from each view at each depth, and incorporate factors such as per-view visibility and view weighting. Because these attention maps are likely to be highly correlated with the scene geometry, we predict them jointly with the depth probabilities in Corr-Branch. We pre-modulate the input image pixel volume with these attention maps before it is processed by the shading predictor. This ensures that each depth plane of the input pixel volume only has meaningful color information at the beginning of the shading prediction, leading to signicantly improved view synthesis results.

We provide ground truth rendered images as supervision, which allows the network to reason about scene appearance and photo consistency. We also supervise the output depth images, which allows the network to learn about scene geometry and correspondences. While the learnable parameters of the correspondence and shading branches are separate, their data flows are highly correlated because they share information coming from the image and depth supervision (Fig. 3).

4.2 Learning multi-view correspondences: Corr-Branch

Multi-view reconstruction and re-rendering methods fundamentally rely on finding correspondences across the input images. We achieve this using deep CNNs that process a plane sweep cost volume with deep-learned filters at multiple scales. Depth and visibility information are highly correlated: depth expresses the distance at which multi-view appearance is consistent, and visibility expresses if single-view appearance is consistent with all other views. Therefore, we propose a novel correspondence estimation network, Corr-Branch, to jointly infer depth and visibility-aware attention maps for a novel view.

Corr-Branch consists of a feature extractor $\mathcal{T}$ and a correspondence predictor C. Photometric images are captured under directional lighting and can have high frequency view-dependent specularities which complicate correspondence reasoning. Therefore, we apply a small U-Net style feature extractor, $\mathcal{T}$, to pre-filter each input image I_i. $\mathcal{T}$ learns to extract specular-invariant features like edges and orientations (examples shown in Fig. 3) that are meaningful for correspondence estimation. Specifically, $\mathcal{T}$ transforms each 3-channel RGB image I_i to an 8-channel feature map:

$$M_i = \mathcal{T}(I_i). \quad (3)$$

Given p discrete, pre-defined depth values $\{d_k | k = 1, .., p\}$, we construct a plane sweep volume $\boldsymbol{M}$ at a novel view, $\mathbf{\Theta}_n$, from the extracted per-input-view feature maps $\{M_i | i = 1, .., m\}$:

$$\boldsymbol{M} = \mathcal{W}(\{M_i\}, \{\mathbf{\Theta}_i\}; \{d_k\}; \mathbf{\Theta}_n). \quad (4)$$

$\mathcal{W}$ geometrically warps every M_i onto every depth plane at a distance d_k using input view, $\mathbf{\Theta}_i$, and novel view, $\mathbf{\Theta}_n$, to form the volume $\boldsymbol{M}$. We use $\boldsymbol{M}_{k,i}$ to denote the warped M_i at the k^{th} depth in $\boldsymbol{M}$. Correspondence inference requires the network to understand photometric consistency across multiple views. We achieve this by processing the volume $\boldsymbol{M}$ with 3D filters with gradually expanding receptive fields. To make it easier for the network to consider multi-view consistency (and inconsistency), we pre-process the feature volume by removing the average multi-view feature at each depth plane as:

$$\tilde{\boldsymbol{M}}_{k,i} = \left(\boldsymbol{M}_{k,i} - \frac{\sum_j \boldsymbol{M}_{k,j}}{m} \right)^2. \quad (5)$$

This operation is similar to variance volumes that have been used in other deep learning-based reconstructions methods [Yao et al. 2018]. However, while a variance volume expresses only global information across all views, $\tilde{\boldsymbol{M}}$ only subtracts the mean and retains per-view information in the form of per-view maps $\tilde{\boldsymbol{M}}_{k,i}$. This allows our network to infer view-dependent attention maps while also leveraging global information. In fact, if required, the network can compute the variance volume in subsequent convolutional layers by averaging $\tilde{\boldsymbol{M}}$ across views.

$\tilde{\boldsymbol{M}}$ is thus a 3D (depth × image height × image width) volume, where each "voxel" has $8m$ channels. We augment these features

with view differences b_i and per-plane depth values d_k to make the network utilize the novel view's location (vis-a-vis the input views) for correspondence reasoning:

$$\boldsymbol{N}_k = \tilde{\boldsymbol{M}}_k \oplus \{b_i\} \oplus d_k, \tag{6}$$

where $\oplus$ is a per-voxel/per-pixel concatenation operator. Since our views lie on a sphere in our acquisition configuration, we use the cosine of the angle between a pair of views as b_i. The volume, $\boldsymbol{N}$, thus has $9m + 1$ channels.

Our correspondence predictor C is a 3D U-Net style network. It processes volume $\boldsymbol{N}$ through a series of 3D convolutional layers, each followed by group normalization (GN) and ReLU layers. We use downsampling and upsampling along the depth and image dimensions to analyze multi-view correspondence at multiple spatial scales. The details of C are shown in Fig. 3. The correspondence predictor outputs an $(m + 1)$-channel volume, in which the first m channels represent a per-view attention volume $\boldsymbol{V}$ and the last one channel represents a depth probability volume $\boldsymbol{P}$:

$$\boldsymbol{V}, \boldsymbol{P} = C(\boldsymbol{N}). \tag{7}$$

Each channel in $\boldsymbol{V}$ corresponds to the attention information for the corresponding input view, and incorporates both visibility and viewpoint-based weighting information. $\boldsymbol{V}_{k,i}$ is an attention map for the i^{th} view at the k^{th} depth plane; it provides a pixel-wise attention mask that is used during shading prediction. $\boldsymbol{P}_k$, on the other hand, provides a pixel-wise depth probability for the k^{th} depth plane.

$\boldsymbol{P}$ is processed with a depth-wise softmax operation to produce actual probability maps for each depth plane:

$$\boldsymbol{P}^d = \text{soft-max}(\alpha^d \boldsymbol{P}), \tag{8}$$

where α^d is a learnable scalar parameter. The output view depth image is finally predicted as:

$$D_n = \sum_k d_k P_k^d. \tag{9}$$

We provide ground truth depth images for D_n as supervision. We expect the attention maps and depth estimates to be highly correlated. Therefore, we estimate them from the single correspondence predictor network and share information until the last layer. This ensures that the attention prediction utilizes the depth supervision to learn meaningful features. As noted before, both the attention, $\boldsymbol{V}$, and depth probability, $\boldsymbol{P}$, are provided to the shading prediction branch to help aggregate multi-view appearance.

4.3 Learning to predict shading: Shade-Branch

Inferring scene appearance from a novel view is a highly challenging task in our configuration: our wide angular baseline input views can see different parts of a scene with limited overlap. Morevover, complex shading effects like high frequency specular highlights vary significantly across views. We resolve these challenges using our image prediction branch, Shade-Branch. As shown in Fig. 3, Shade-Branch has a shading predictor $\mathcal{S}$ that reasons about scene appearance from multi-view images $\{I_i\}$, using the multi-view correspondence information (attention maps, $\boldsymbol{V}$, and depth probability, $\boldsymbol{P}$) predicted by Corr-Branch.

Similar to Eqn. (4), a plane sweep volume $\boldsymbol{I}$ is constructed from original images $\{I_i\}$ using homography-based warping $\mathcal{W}$. This volume contains the original color information warped onto multiple planes. This volume can have highly redundant and potentially inconsistent information from multiple views due to strong occlusions. A key feature of our network is that we pre-mask this color volume $\boldsymbol{I}$ using the visibility-aware attention maps inferred from the correspondence branch, to disentangle this redundancy and inconsistency. The masking is achieved by a voxel-wise multiplication, for every view at every depth:

$$\tilde{\boldsymbol{I}}_{k,i} = \boldsymbol{I}_{k,i} \boldsymbol{V}_{k,i}. \tag{10}$$

This directly connects Corr-Branch and Shade-Branch; Corr-Branch can thus leverage appearance information from Shade-Branch to estimate attention maps that allow the shading prediction to be as accurate as possible.

Our shading predictor $\mathcal{S}$ is a 3D-Unet style network similar to the correspondence predictor C, but with more channels at each layer for better appearance reasoning. It processes the masked volume $\tilde{\boldsymbol{I}}$, the original attention maps $\boldsymbol{V}$ and other information, and predicts a 3-channel appearance volume

$$\boldsymbol{A} = \mathcal{S}(\tilde{\boldsymbol{I}}, \boldsymbol{V}, \{b_i\}, \{d_k\}), \tag{11}$$

where $\tilde{\boldsymbol{I}}, \boldsymbol{V}, \{b_i\}, \{d_k\}$ are concatenated voxel-wise, similar to Eqn. (6). Supplying the original attention maps, $\boldsymbol{V}$ (in addition to the modulated appearance volume) gives the network more freedom to reconstruct scene appearance. Note that each plane $\boldsymbol{A}_k$ is a predicted image, containing the predicted appearance of the scene at the k^{th} depth plane. These per-plane predicted images are weighted by the depth probability maps from Corr-Branch to reconstruct the final image as:

$$\tilde{\boldsymbol{A}}_k = \boldsymbol{A}_k P_k^a, \tag{12}$$

$$I_n = \sum_k \tilde{\boldsymbol{A}}_k. \tag{13}$$

where P^a is the normalized depth probability volume using soft-max and a scalar parameter $\alpha^a = 4$ similar to Eqn. (8). Each plane $(\tilde{\boldsymbol{A}}_k)$ of $\tilde{\boldsymbol{A}}$, as shown in Fig. 3, is a clean weighted image, with depth-incorrect outlier pixels completely masked out.

We provide ground truth rendered novel view images as supervision for I_n, and train our network end to end. As noted before, by transferring information between Shade-Branch and Corr-Branch (in the form of the attention maps and depth estimates), our novel network design consolidates both appearance synthesis and correspondence estimation, allowing the two branches to leverage each other for better inference.

4.4 Extension to multi-light inputs.

Many photometric applications, like image-based relighting and photometric stereo, acquire photometric images under multiple light sources from a single viewpoint. We propose an extension of our approach to multi-view and *multi-light* datasets. As noted in Sec. 3, for this application we capture images under six fixed views and six fixed lights.

Multi-light data allows for better geometric reconstruction by giving us more information about scene appearance [Davis et al.

2005]. For example, specularities under one light source can disappear under another light source, and shadowed regions under one light can be illuminated by another. This leads to better estimation of normals, edges and other features. We design an extended feature extractor $\mathcal{T}_q$ to take advantage of this for better multi-view correspondence inference. The difference between the single-light $\mathcal{T}$ and the multi-light $\mathcal{T}_q$ feature extractors is merely the number of input channels. $\mathcal{T}_q$ uses a fixed number of structured inputs; the subscript q specifies the number of lights. For each view, we stack q images $I_i^1, I_i^2, ..., I_i^q$ under q different light sources together as a $3q$-channel feature map as an input for $\mathcal{T}_q$. While Corr-Branch predicts the attention maps and depth probability maps from multi-light data, the input for our Shade-Branch remains the same; it still predicts a novel view image under a single light source using the multi-view images under that light.

We use our network to synthesize novel views for each input light. In Section 6, we show that it can be combined with image-based relighting methods to re-render scene appearance under novel viewpoint and lighting — the full scene acquisition scenario.

4.5 Data generation

To the best of our knowledge, there is no existing large-scale dataset that contains multi-view photometric images. Previous novel view synthesis methods have trained with data from on-line videos [Zhou et al. 2018], car driving scenes [Geiger et al. 2012] or simple objects from specific classes [Chang et al. 2015], none of which apply to our high-quality appearance acquisition scenario. Therefore, we create a novel large-scale synthetic dataset. As in Xu et al. [2018], we procedurally generate shapes by combining primitives with randomly generated bump maps and randomly merging 1 to 9 primitives. We create 1000 training scenes and 50 testing scenes using this method. We texture map these shapes with SVBRDFs from the Adobe Stock 3D material dataset[1]. This dataset contains 1329 SVBRDFs, and we separate it into 1129 training and 200 test materials.

We render our dataset using path tracing with 1000 samples; we use Optix to achieve fast path tracing, similar to [Li et al. 2018b]. This physically based rendering method ensures that our images contain realistic light transport. We render 512×512-resolution images and tone-map them using gamma 2.2. We render our scenes using the camera configuration described in Sec. 3. For each scene, we randomly select a field-of-view angle from 5° to 60°, and correspondingly calculate a distance for cameras based on the angle and the scene's size to ensure good pixel coverage. Each training image set consists of 6 input views and one novel view under a single directional light. For each training scene, we create 30 such sets by randomly placing 30 novel views between the 6 input views and selecting 30 random directional lights for rendering. In total, we have 30000 such sets in our training set. For each test scene, we randomly select 36 novel views and 3 directional lights, and render each view under each light, which creates 108 image sets, for a total of 5400 image sets in our test set.

[1]https://stock.adobe.com/3d-assets

4.6 Training details

Loss functions. We use an L1 loss on both the ground truth images and depths from each novel view. Specifically, let $\mathcal{L}_a$ be the image L1 loss and $\mathcal{L}_d$ be the depth L1 loss. Our final loss $\mathcal{L}$ is given by

$$\mathcal{L} = \mathcal{L}_a + \beta \mathcal{L}_d. \tag{14}$$

We use $\beta = 0.1$ for all our experiments.

Parameters and strategies. Our six views are radially symmetric (see Fig. 2); we leverage this symmetry to provide structured inputs to our network and make training easier. Given a novel view, we first rotate the image and depth around the central viewing direction to make its up direction point towards the central input view. We then reorder the input views to achieve a canonical input view layout, where the first view is the central view, and the other views are ordered in counter-clockwise order starting with the view on the left (see Fig. 2 right).

We assume that the test scenes are captured by a calibrated spherical gantry and that the physical size of a scene and the distance from a camera to the center of the scene is approximately known. This allows the depth values $\{d_k | k = 1, 2..., p\}$ to be specified correspondingly. While our network is fully convolutional and can support an arbitrary number of depth planes, p, we use $p = 64$ for our training and all experiments. During training, we have perfectly symmetric cameras distributed at a known distance, $\hat{d}$, from the center of a scene. We also know the ground truth scene size Δ (the maximum difference between $\hat{d}$ and each pixel depth). To get $\{d_k\}$, we set $d_1 = \hat{d} - \gamma\Delta$, $d_p = \hat{d} + \gamma\Delta$, and uniformly divide this range for other d_k, which can be expressed by:

$$d_k = \hat{d} - \gamma\Delta + \gamma \frac{2\Delta(k-1)}{p-1}. \tag{15}$$

During training, we randomly pick γ between 0.8 to 2.0, so that our network sees different sweep volume scales; we thus only need to specify a roughly correct distance $\hat{d}$ and size Δ at test time for real scenes (using $\gamma = 1$).

We train our networks using three or four NVIDIA Titan Xp or 1080Ti GPUs, using a batch size of 4 for each GPU, for a total batch size of 12 or 16. We apply group normalizations in both $\mathcal{C}$ and $\mathcal{S}$, and observe better performance than batch normalization with our small per-GPU batch size, similar to [Wu and He 2018]. During training, we randomly crop 64×64 patches from novel view images and depths for data augmentation. Since our images can have large regions of black background, we only select crops that have at least 50% non-background pixels. Our network generally converges after training for 400 epochs (about 4 days with 4 GPUs).

5 EXPERIMENTS

We now present a comprehensive evaluation of our method on both synthetic and real data.

Ablation study on synthetic data. We first justify the design of our network architecture through ablations on the synthetic dataset. We compare our proposed single-light photometric novel view synthesis network, $\mathcal{TCS}$, i.e., with a feature extractor ($\mathcal{T}$) and correspondence and shading predictors ($\mathcal{C}$ and $\mathcal{S}$), against a number of variants. We also compare it to our multi-light network, $\mathcal{T}_6\mathcal{CS}$, with six different

Table 1. Ablation study. We evaluate different versions of our networks on our synthetic testing dataset, and compare the image L1 error, PSNR, SSIM, and Depth L1 error on the central 256 × 256 crops.

	Image L1	Image PSNR	Image SSIM	Depth L1
$\mathcal{T}(\mathcal{CS})_{\text{noV}}$	0.0451	30.74	0.9391	0.0567
$\mathcal{TC}_{\text{noD}}\mathcal{S}$	0.0345	32.05	0.9520	0.1448
$\mathcal{TCS}$	0.0318	32.61	0.9573	0.0437
$\mathcal{T}_6\mathcal{CS}$	0.0307	33.03	0.9602	0.0246

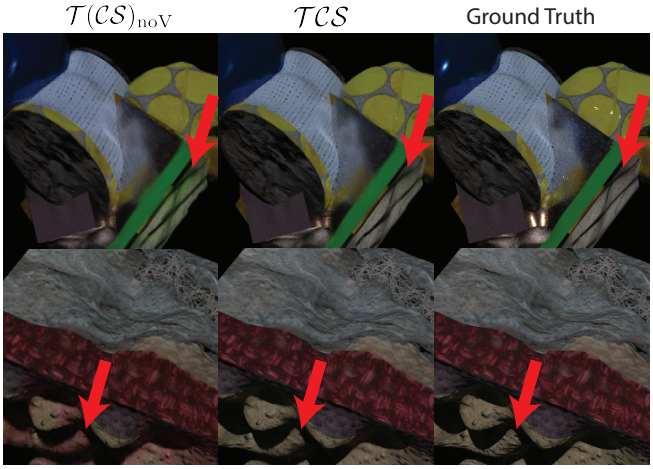

Fig. 4. Qualitative comparisons on synthetic test set between $\mathcal{TCS}$ and $\mathcal{T}(\mathcal{CS})_{\text{noV}}$ (i.e., with and without attention maps, respectively). $\mathcal{T}(\mathcal{CS})_{\text{noV}}$ suffers from color bleeding artifacts (red arrows), that are resolved by $\mathcal{TCS}$.

directional lights as inputs for Corr-Branch and a multi-light feature extractor. Specifically, we compare against the following networks: $\mathcal{TC}_{\text{noD}}\mathcal{S}$, that doesn't have depth supervision, and $\mathcal{T}(\mathcal{CS})_{\text{noV}}$, a network that does not use attention maps. We evaluate all these networks on our synthetic testing dataset and compare the image L1 loss, PSNR, SSIM, and depth L1 loss. To avoid biases in these metrics from the large black backgrounds in our rendered images, we crop the central 256 × 256 regions from all testing images for evaluation; these crops have 75% non-background pixels on average. We also calculate depth L1 only for the foreground as we do for training.

The numerical comparisons of these different networks are shown in Tab. 1. As demonstrated by $\mathcal{TCS}$ vs $\mathcal{T}(\mathcal{CS})_{\text{noV}}$, our visibility-aware attention maps significantly improve reconstruction performance; image L1 loss reduces by about 30% and is accompanied by a large improvement in PSNR and SSIM. Figure 4 shows qualitative comparisons between $\mathcal{TCS}$ and $\mathcal{T}(\mathcal{CS})_{\text{noV}}$. We observe many color-bleeding artifacts with $\mathcal{TCS}$ because it is unable to resolve the large occlusions in these scenes. We also observe that the attention maps can be inferred and help the synthesis in an unsupervised way; network $\mathcal{TC}_{\text{noD}}\mathcal{S}$ with attention maps, but without depth supervision, performs much better at view synthesis than $\mathcal{T}(\mathcal{CS})_{\text{noV}}$ without attention maps and with depth supervision. These comparisons demonstrate the key role our visibility-aware attention maps play in effectively eliminating the incorrect, inconsistent information in a plane sweep volume. Also, comparing $\mathcal{TC}_{\text{noD}}\mathcal{S}$ vs $\mathcal{TCS}$ shows that depth supervision improves reconstruction accuracy, though in a subtle way.

Finally, when multi-light data is provided, our multi-light network $\mathcal{T}_6\mathcal{CS}$ has the best performance. This confirms that the multi-light feature extractor $\mathcal{T}_6$ extracts better features leading to both better depth prediction and better image synthesis than our single light version. While the multi-light version requires more acquired images, it can be naturally combined with an image-based relighting method to enable view and lighting changes (see Fig. 8).

Real data capture. We capture our real scenes—composed of one or more real objects placed on a platform—using a spherical gantry. We capture each scene from the six input views under either a central directional light (for single-light results) or six directional lights (for multi-light results and additional applications). We also capture 50 novel views under these lights, as ground truth to validate our view synthesis quality. Our cameras are about 50cm away from the platform. We thus set $\hat{d}$ = 50cm for all our real scenes. While the sizes and scales of our real scenes vary, we find that Δ = 6cm works well for most cases. Our network is robust to these variations because of the randomized Δ during training. We mask out the background of the captured images before passing them to our network. Our network is fully convolutional and we directly apply our method on these images, although we are training on 64×64 crops. We use an Nvidia Titan Xp to process our real results and it takes about 2 seconds to generate a 400×400 image.

Comparisons against previous view synthesis methods. We now compare our method to previous state-of-the-art view synthesis methods. We considered comparing against learning-based IBR methods. However, these methods are usually designed for general IBR applications with densely sampled views [Flynn et al. 2016; Hedman et al. 2018]. We tried training an implementation of Flynn et al. [2016] on our dataset, but it failed to predict reasonable results in our wide-baseline case. Hedman et al. [2018] require estimating geometry using multi-view stereo, which does not work from our sparse inputs. Besides, methods designed for specific scenarios, like light fields [Kalantari et al. 2016; Srinivasan et al. 2017] or pair-view extrapolation [Zhou et al. 2018] are also not easy to apply in our case. We also tried training the released network of [Kalantari et al. 2016] on our dataset; it doesn't converge to any reasonable results because of a significantly more challenging task. Therefore, we compare against [Penner and Zhang 2017], a state-of-the-art non-learning based IBR method. Note that, Penner and Zhang [2017] have already demonstrated that their method performs better than [Flynn et al. 2016] and [Kalantari et al. 2016].

We also compare against a flow-based view synthesis method [Sun et al. 2018]. Directly applying their model, trained on KITTI [Geiger et al. 2012] or ShapeNet [Chang et al. 2015], doesn't work on our data. Therefore, we retrain their model using our dataset, albeit with a white background; this is the same scenario Sun et al.[2018] use in their paper.

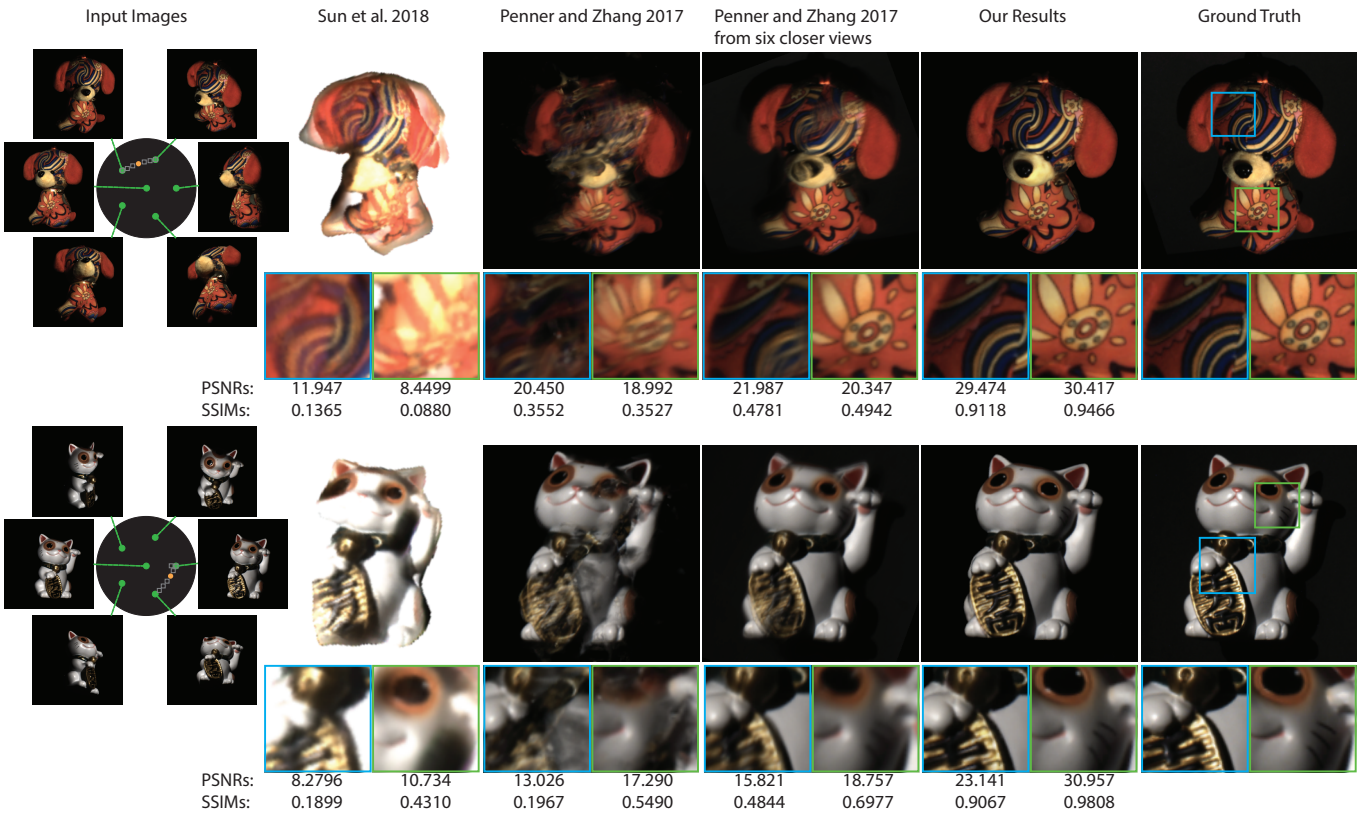

Fig. 5. Comparisons with flow-based view synthesis [Sun et al. 2018] (second column) and Soft3D [Penner and Zhang 2017] (third column) using wide-baseline inputs (first column, input viewing directions shown in green, and novel view direction in yellow). We also compare with [Penner and Zhang 2017] (fourth column) that uses a much closer set of views as input (marked in grey rectangles in the first column). Our results are significantly better than other methods and differences almost imperceptible from ground truth. We show cropped insets of all results with corresponding PSNRs and SSIMs (bottom).

In Fig. 5, we show qualitative and quantitative comparisons on two real captured examples. The first scene has a complicated surface texture and the second has complex specularities and hard shadows. In the second and third columns of Fig. 5, we compare with [Sun et al. 2018] and [Penner and Zhang 2017] using the same six-view images we use for our method. We can see that our method performs significantly better than the two methods qualitatively, as shown by the synthesized images (details in insets), and quantitatively, as shown by the PSNR and SSIM values. Both previous methods fail to handle this challenging wide-baseline configuration. Sun et al. [2018] produce blurred results with no appearance details and many mis-aligned ghosting artifacts. The results of [Penner and Zhang 2017] also contain serious ghosting artifacts. Our method, on the other hand, produces photorealistic view synthesis results with significantly higher PSNR and SSIM values. As shown in the insets, our method recovers both the complicated texture of the first example as well as the challenging hard shadows and view-dependent specularities in the second example.

We also select six closer views as input for [Penner and Zhang 2017]. As shown in the fourth column, these "easier" inputs improve their results. However, their results with the small-baseline inputs are still worse than our results from the six wide-baseline inputs, highlighting the accuracy and robustness of our method.

View synthesis on real photometric data. Figures 1 and 10 show our view synthesis results from our single-light network compared with captured ground truth. Our method produces photo-realistic novel view images for these real scenes, which accurately match the ground truth. As demonstrated in many examples, our method generates high-quality view interpolation results even at challenging viewing directions that are close to the boundary of the pentagonal cone, where very limited input information can be used. These results are consistent across a wide variety of scenes, in terms of both materials (pottery, cloth, mental, wood, plastic and candy) and geometry (single and multiple objects; small and big objects).

Comparison between single-light and multi-light networks. We show a challenging scene under directional lighting in Fig. 6, and compare our single-light network with our extended multi-light network. The scene contains thin structures (like the arms and thumbs) which are very distinct from our training geometry primitives. These highly non-convex structures exhibit challenging cast shadows that complicate correspondence inference. Nevertheless, our single-light network performs well for most viewing directions

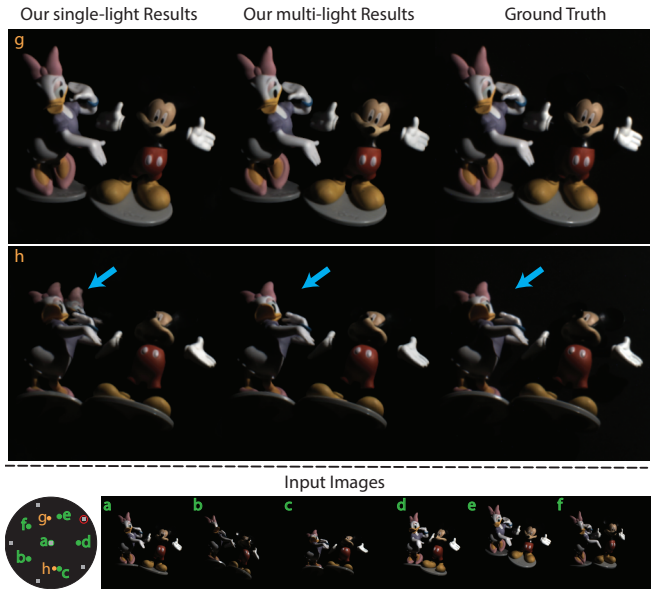

Fig. 6. Single-light vs. multi-light network comparison. For a complex scene captured under a challenging light direction (marked by red hollow circle), our single-light network may generate ghosting artifacts from some challenging viewing directions (second row, marked by blue arrows). Our multi-light network resolves these issues using images under multiple light sources (marked by gray). For most viewing directions, our single-light network produces high-quality results (first row). The six input images are shown on the bottom, with light directions marked in green with labels (a-f), and novel views marked in yellow with labels (g,h).

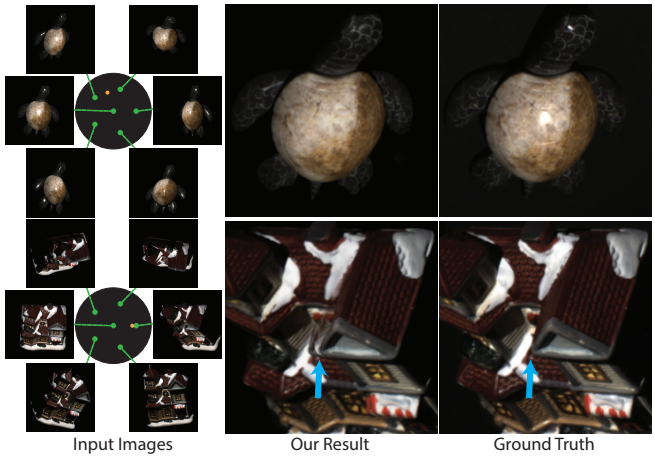

Fig. 7. Limitations. Our method fails to reconstruct sharp specularies that have long-range motion (top) and highly non-convex occlusions (bottom).

(e.g. Fig. 6.g). As shown in Fig. 6.h, our single-light network may generate obvious ghosting artifacts for some challenging directions, but our multi-light network can resolve these issues thanks to more reliable correspondence inference from multi-light images.

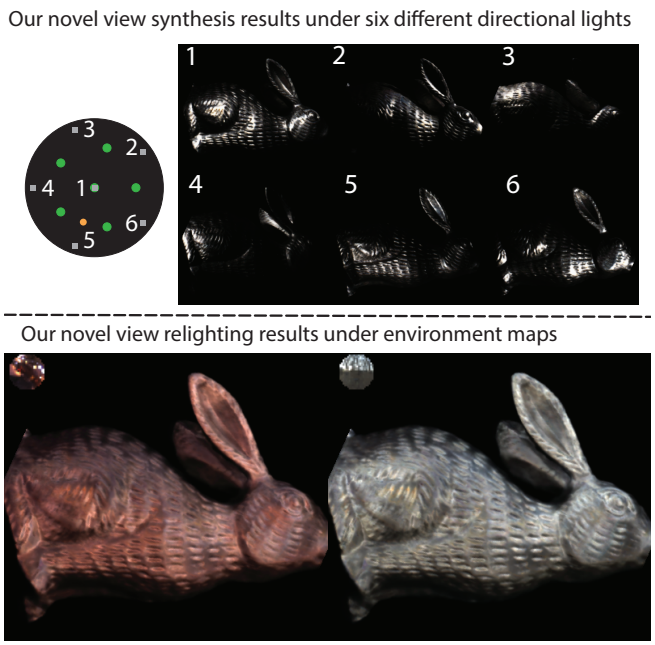

Fig. 8. Novel view relighting using our novel view synthesis results. We apply our multi-light network to synthesize six novel view images under six directional lights (marked in gray with labels 1-6), each synthesized from the same six views (marked in green). We use these high-quality synthesis results as inputs for a deep image-based relighting technique [Xu et al. 2018], and create relighting results (on the bottom) under environment maps (shown on the top left of each result).

Limitations. Our method only handles opaque scenes, which is a limitation of our training dataset. Our network is trained on 64×64 cropped images, which limits the spatial scale of appearance reasoning. Consequently, long-range effects like sharp specularities that move significantly are not reconstructed well (see Fig. 7). Our method might blur sharp specularities (see Fig. 10). Also, our network generates blurred results with ghosting for highly non-convex scenes with parts that are visible in only one or two views (see Fig. 7).

6 ADDITIONAL APPLICATIONS

Our view synthesis method can be combined with other scene acquisition and rendering techniques to enable a broad set of applications. We now demonstrate a few examples.

Novel view relighting. Our method can synthesize novel views from images captured under different directional lights. These, in turn, can be used with image-based relighting methods to enable rendering under novel view and lighting. One such relighting example is shown in Fig. 1c3. We apply our multi-light network with six different directional lights and synthesize novel view images for each light separately (see Fig. 8 top). We train the Relight-Net network from [Xu et al. 2018] for our six-light setting. We pass the synthesized novel view images for the six lights to this network to generate images under novel directional lights. Similar to [Xu

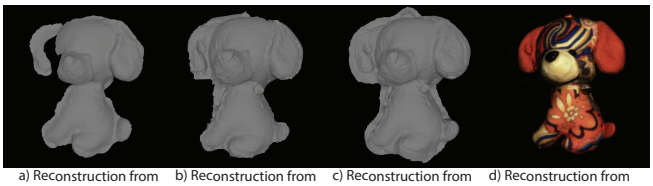

Fig. 9. Multi-view stereo reconstruction from our synthesized images. We synthesize 56 high-quality novel views from six images and use them as inputs for a multi-view stereo algorithm, COLMAP [Schönberger et al. 2016], to generate 3D reconstructions (c,d) that are comparable to a reconstruction from 56 captured images (b). COLMAP reconstructs incomplete geometry with holes from only the six sparse images (a).

et al. 2018], we achieve relighting under novel environment maps by linearly combining the relit images as shown in Fig. 8 bottom. Note that our network and the Relight-Net are trained separately, without end-to-end refinement or any other special processing.

While our results from a novel view under changing environment map often look realistic, some aspects still need improvment. For example, some blurriness, incorrect shadow motion and temporal inconsistency that become obvious when changing the view under a novel environment map. Jointly training our network with the Relight-Net using a larger task-specific training dataset can potentially resolve, or at least alleviate, these issues. That said, to our knowledge, this is the first attempt at synthesizing a full reflectance field, enabling changes to both lighting and viewpoint, from such sparse samples (36 images from 6 views under 6 lights).

Multi-view stereo. Multi-view stereo methods often require a dense set of input views, and can fail to reconstruct complete hole-free geometry for sparse views such as ours. We apply our method to "densify" the captured scene and synthesize 56 novel view images around a scene from six images. We pass these 56 synthesized images to a multi-view stereo system COLMAP [Schönberger et al. 2016] and achieve 3D reconstruction of the scene (see Fig. 9 c and d). As a baseline, we pass the 56 captured ground truth images to COLMAP for reconstruction and we observe qualitatively similar results as shown in Fig. 9. Note that COLMAP reconstructs incomplete geometry with missing parts from the original six sparse views. By making MVS methods work better with such sparse viewpoints, our method makes them more robust and general.

7 CONCLUSION AND FUTURE WORK

We have demonstrated a method to synthesize photometric scene appearance at a wide range of novel viewpoints from a sparse set of only six images captured at large baselines. This is in contrast to previous methods that rely on densely sampling the scene with hundreds of viewpoints. We achieve this by training a novel deep CNN that can simultaneously infer correspondences and shading from structured photometric images. Our network predicts visibility-aware attention maps that effectively address photometric and geometric inconsistencies and allow for the accurate aggregation of multi-view scene appearance. We present evaluations and comparisons to previous view synthesis methods, and show that we can generate significantly more accurate and photorealistic images across a wide range of scenes. Fundamentally, our work takes a step towards capturing and rendering scene appearance from sparse image sets. This is a classic problem in vision and graphics and we believe that our work can enable many other applications. For example, we demonstrate that our synthesized images can be used to achieve novel view relighting and multi-view stereo from sparse images. In the future, it would be interesting to explore extensions of our technique to other challenging scene acquisition tasks, like multi-view BRDF reconstruction, 360° scene reconstruction, and the acquisition of dynamic scene appearance from sparse images.

ACKNOWLEDGEMENTS

We thank Pratul Srinivasan for helpful suggestions and code for comparison with [Penner and Zhang 2017]. We thank Zhengqin Li for providing a renderer for the synthetic dataset used in the paper. Parts of this work were done while Zexiang Xu was an intern at Adobe Research. This work was supported in part by NSF grants 1617234, 1703957, ONR grant N000141712687, Adobe (including an Adobe Research Fellowship), a Powell-Bundle Fellowship, the Ronald L. Graham Chair and the UC San Diego Center for Visual Computing.

REFERENCES

Jonathan T Barron and Jitendra Malik. 2015. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)* 37, 8 (2015), 1670–1687.

Sai Bi, Nima Khademi Kalantari, and Ravi Ramamoorthi. 2017. Patch-based optimization for image-based texture mapping. *ACM Transactions on Graphics (TOG)* 36, 4 (2017).

Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. 2001. Unstructured lumigraph rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 425–432.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

Gaurav Chaurasia, Sylvain Duchene, Olga Sorkine-Hornung, and George Drettakis. 2013. Depth synthesis and local warps for plausible image-based navigation. *ACM Transactions on Graphics (TOG)* 32, 3 (2013), 30.

Gaurav Chaurasia, Olga Sorkine, and George Drettakis. 2011. Silhouette-Aware Warping for Image-Based Rendering. In *Computer Graphics Forum*, Vol. 30. Wiley Online Library, 1223–1232.

Anpei Chen, Minye Wu, Yingliang Zhang, Nianyi Li, Jie Lu, Shenghua Gao, and Jingyi Yu. 2018. Deep Surface Light Fields. *Proc. ACM Comput. Graph. Interact. Tech.* 1, 1, Article 14 (July 2018), 17 pages. https://doi.org/10.1145/3203192

Shenchang Eric Chen and Lance Williams. 1993. View Interpolation for Image Synthesis. In *Proceedings of SIGGRAPH*. 279–288.

Łukasz Dąbała, Matthias Ziegler, Piotr Didyk, Frederik Zilly, Joachim Keinert, Karol Myszkowski, H-P Seidel, Przemyslaw Rokita, and Tobias Ritschel. 2016. Efficient Multi-image Correspondences for On-line Light Field Video Processing. In *Computer Graphics Forum*, Vol. 35. Wiley Online Library, 401–410.

James Davis, Diego Nehab, Ravi Ramamoorthi, and Szymon Rusinkiewicz. 2005. Spacetime Stereo: A Unifying Framework for Depth from Triangulation. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)* 27, 2 (Feb. 2005), 296–302.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 145–156.

Paul E Debevec, Camillo J Taylor, and Jitendra Malik. 1996. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 11–20.

Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. 2018. Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 128.

David Eigen and Rob Fergus. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2650–2658.

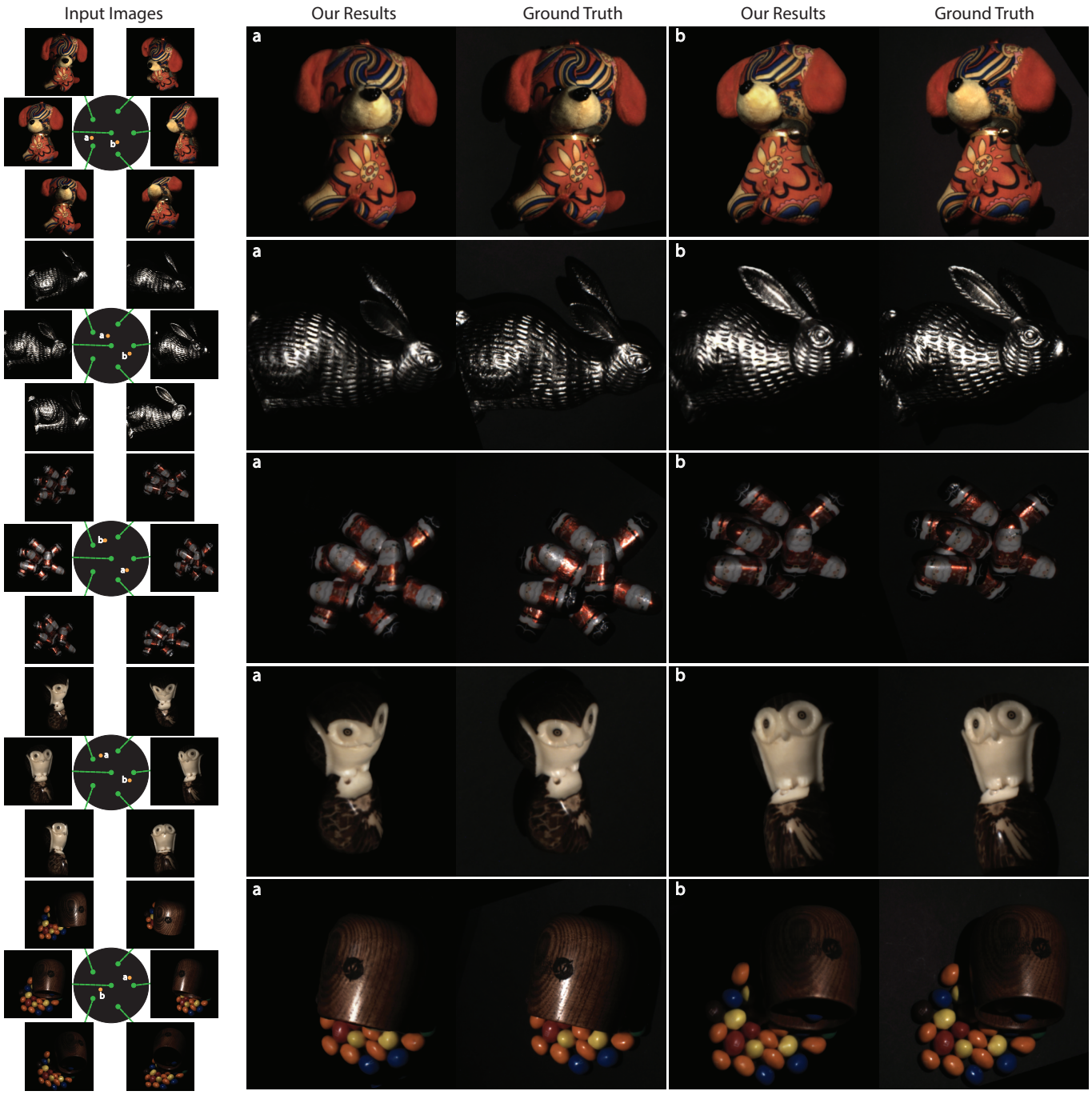

Fig. 10. Novel view synthesis results from our single-light network on real scenes. For each scene, we show two novel view synthesis results (second and forth columns) compared with captured ground truth images (third and fifth columns), whose viewing directions are marked in yellow with corresponding labels (a and b). We also show the inputs in the first column marked with corresponding viewing directions in green.

Martin Eisemann, Bert De Decker, Marcus Magnor, Philippe Bekaert, Edilson De Aguiar, Naveed Ahmed, Christian Theobalt, and Anita Sellent. 2008. Floating textures. In *Computer graphics forum*, Vol. 27. Wiley Online Library, 409–418.

John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. 2016. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5515–5524.

Ryo Furukawa, Hiroshi Kawasaki, Katsushi Ikeuchi, and Masao Sakauchi. 2002. Appearance Based Object Modeling using Texture Database: Acquisition Compression and Rendering.. In *Rendering Techniques*. 257–266.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3354–3361.

Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. 1996. The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 43–54.

Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. 2018. Deep blending for free-viewpoint image-based rendering. In *SIGGRAPH Asia 2018 Technical Papers*. ACM, 257.

Michael Holroyd, Jason Lawrence, and Todd Zickler. 2010. A Coaxial Optical Scanner for Synchronous Acquisition of 3D Geometry and Surface Reflectance. *ACM Trans. Graph.* 29, 4, Article 99 (July 2010), 99:1–99:12 pages.

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. 2018. DeepMVS: Learning Multi-View Stereopsis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhuo Hui, Kalyan Sunkavalli, Joon-Young Lee, Sunil Hadap, Jian Wang, and Aswin C Sankaranarayanan. 2017. Reflectance capture using univariate sampling of brdfs. In *The IEEE International Conference on Computer Vision (ICCV)*, Vol. 2.

Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. 2016. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 193.

Marc Levoy and Pat Hanrahan. 1996. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. ACM, 31–42.

Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2017. Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 45.

Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. 2018a. Materials for Masses: SVBRDF Acquisition with a Single Mobile Phone Image. In *ECCV*.

Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. 2018b. Learning to reconstruct shape and spatially-varying reflectance from a single image. In *SIGGRAPH Asia 2018 Technical Papers*. ACM, 269.

Tom Malzbender, Dan Gelb, and Hans Wolters. 2001. Polynomial Texture Maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*. 519–528.

Giljoo Nam, Joo Ho Lee, Diego Gutierrez, and Min H Kim. 2018. Practical SVBRDF acquisition of 3D objects with unstructured flash photography. In *SIGGRAPH Asia 2018 Technical Papers*. ACM, 267.

Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. 2017. Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 702–711.

Pieter Peers, Dhruv K Mahajan, Bruce Lamond, Abhijeet Ghosh, Wojciech Matusik, Ravi Ramamoorthi, and Paul Debevec. 2009. Compressive light transport sensing. *ACM Transactions on Graphics (TOG)* 28, 1 (2009), 3.

Eric Penner and Li Zhang. 2017. Soft 3d reconstruction for view synthesis. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 235.

Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Efstratios Gavves, and Tinne Tuytelaars. 2016. Deep reflectance maps. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4508–4516.

Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. 2016. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 501–518.

Christopher Schwartz, Michael Weinmann, Roland Ruiters, and Reinhard Klein. 2011. Integrated High-Quality Acquisition of Geometry and Appearance for Cultural Heritage.. In *VAST*, Vol. 2011. 25–32.

Sudipta Sinha, Drew Steedly, and Rick Szeliski. 2009. Piecewise planar stereo for image-based rendering. (2009).

Pratul P Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. 2017. Learning to synthesize a 4d rgbd light field from a single image. In *IEEE International Conference on Computer Vision (ICCV)*. 2262–2270.

Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim. 2018. Multi-view to Novel View: Synthesizing Novel Views with Self-Learned Confidence. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. 2015. Single-view to multi-view: Reconstructing unseen views with a convolutional network. *CoRR abs/1511.06702* 1, 2 (2015), 2.

Suren Vagharshakyan, Robert Bregovic, and Atanas Gotchev. 2018. Light field reconstruction using shearlet transform. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)* 40, 1 (2018), 133–147.

Michael Weinmann and Reinhard Klein. 2015. Advances in Geometry and Reflectance Acquisition (Course Notes). In *SIGGRAPH Asia 2015 Courses*. Article 1, 1:1–1:71 pages.

Tim Weyrich, Jason Lawrence, Hendrik P. A. Lensch, Szymon Rusinkiewicz, and Todd Zickler. 2009. Principles of Appearance Acquisition and Representation. *Found. Trends. Comput. Graph. Vis.* 4, 2 (Feb. 2009), 75–191.

Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, and Markus Gross. 2006. Analysis of Human Faces Using a Measurement-based Skin Reflectance Model. *ACM Trans. Graph.* 25, 3 (July 2006), 1013–1024.

Daniel N Wood, Daniel I Azuma, Ken Aldinger, Brian Curless, Tom Duchamp, David H Salesin, and Werner Stuetzle. 2000. Surface light fields for 3D photography. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 287–296.

Robert J Woodham. 1980. Photometric method for determining surface orientation from multiple images. *Optical engineering* 19, 1 (1980), 191139.

Yuxin Wu and Kaiming He. 2018. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 3–19.

Rui Xia, Yue Dong, Pieter Peers, and Xin Tong. 2016. Recovering shape and spatially-varying surface reflectance under unknown illumination. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 187.

Zexiang Xu, Jannik Boll Nielsen, Jiyang Yu, Henrik Wann Jensen, and Ravi Ramamoorthi. 2016. Minimal BRDF sampling for two-shot near-field reflectance acquisition. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 188.

Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. 2018. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 126.

Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. 2015. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Advances in Neural Information Processing Systems*. 1099–1107.

Li Yao, Yunjian Liu, and Weixin Xu. 2016. Real-time virtual view synthesis using light field. *EURASIP Journal on Image and Video Processing* 2016, 1 (2016), 25.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. 2018. MVSNet: Depth Inference for Unstructured Multi-view Stereo. *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).

Qian-Yi Zhou and Vladlen Koltun. 2014. Color map optimization for 3D reconstruction with consumer depth cameras. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 155.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. 2018. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 65.

Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. 2016b. View synthesis by appearance flow. In *European conference on computer vision (ECCV)*. Springer, 286–301.

Zhiming Zhou, Guojun Chen, Yue Dong, David Wipf, Yong Yu, John Snyder, and Xin Tong. 2016a. Sparse-as-possible SVBRDF acquisition. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 189.

Zhenglong Zhou, Zhe Wu, and Ping Tan. 2013. Multi-view photometric stereo with spatially varying isotropic materials. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1482–1489.