Specular-to-Diffuse Translation for Multi-View Reconstruction

Shihao Wu[1] [0000−0003−4778−8520], Hui Huang[2] [0000−0003−3212−0544] ⋆,
Tiziano Portenier[1] [0000−0003−1766−1705], Matan Sela[3] [0000−0002−0808−9041],
Daniel Cohen-Or[2,4] [0000−0001−6777−7445], Ron Kimmel[3] [0000−0002−3180−7961], and
Matthias Zwicker[5] [0000−0001−8630−5515]

[1]University of Bern [2]Shenzhen University [3]Technion - Israel Institute of Technology
[4]Tel-Aviv University [5]University of Maryland

Abstract. Most multi-view 3D reconstruction algorithms, especially when shape-from-shading cues are used, assume that object appearance is predominantly diffuse. To alleviate this restriction, we introduce S2Dnet, a generative adversarial network for transferring multiple views of objects with specular reflection into diffuse ones, so that multi-view reconstruction methods can be applied more effectively. Our network extends unsupervised image-to-image translation to multi-view "specular to diffuse" translation. To preserve object appearance across multiple views, we introduce a Multi-View Coherence loss (MVC) that evaluates the similarity and faithfulness of local patches after the view-transformation. In addition, we carefully design and generate a large synthetic training data set using physically-based rendering. During testing, our network takes only the raw glossy images as input, without extra information such as segmentation masks or lighting estimation. Results demonstrate that multi-view reconstruction can be significantly improved using the images filtered by our network.

Keywords: Generative adversarial network, multi-view reconstruction, multi-view coherence, specular-to-diffuse, image translation

1 Introduction

Three-dimensional reconstruction from multi-view images is a long standing problem in computer vision. State-of-the-art shape-from-shading techniques achieve impressive results [1, 2]. These techniques, however, make rather strong assumptions about the data, mainly that target objects are predominantly diffuse with almost no specular reflectance. Multi-view reconstruction of glossy surfaces is a challenging problem, which has been addressed by adding specialized hardware (e.g., coded pattern projection [3] and two-layer LCD [4]), imposing surface constraints [5,6], or making use of additional information like silhouettes and environment maps [7], or the Blinn-Phong model [8].

In this paper, we present a generative adversarial neural network (GAN) that translates multi-view images of objects with specular reflection to diffuse ones. The network aims

⋆ Corresponding author: Hui Huang (hhzhiyan@gmail.com)

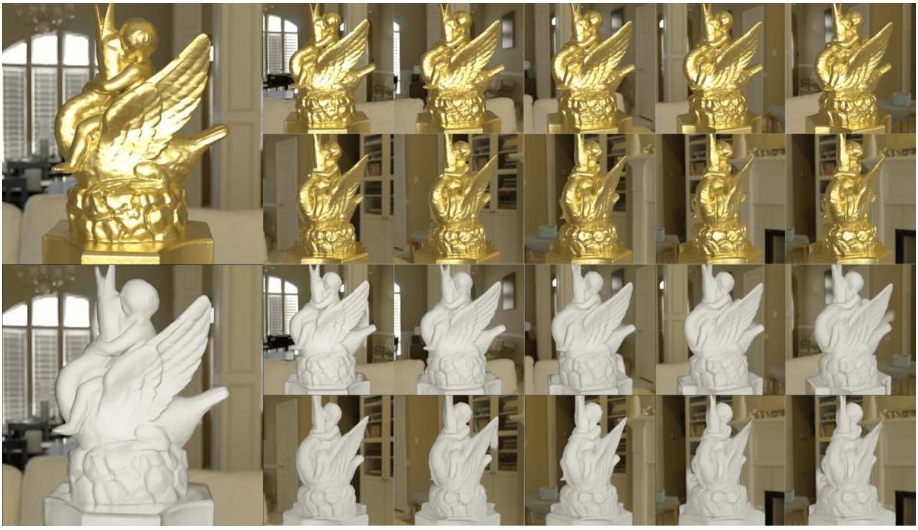

Fig. 1: Specular-to-diffuse translation of multi-view images. We show eleven views of a glossy object (top), and the specular-free images generated by our network (bottom).

to generate a specular-free surface, which then can be reconstructed by a standard multi-view reconstruction technique as shown in Figure 1. We name our translation network, S2Dnet, for Specular-to-Diffuse. Our approach is inspired by recent GAN-based image translation methods, like pix2pix [9] or cycleGAN [10], that can transform an image from one domain to another. Such techniques, however, are not designed for multi-view image translation. Directly applying these translation techniques to individual views is prone to reconstruction artifacts due to the lack of coherence among the transformed images. Hence, instead of using single views, our network considers a triplet of near-by views as input. These triplets allow learning the mutual information of neighboring views. More specifically, we introduce a global-local discriminator and a perceptual correspondence loss that evaluate the multi-view coherency of local corresponding image patches. Experiments show that our method outperforms baseline image translation methods.

Another obstacle of applying image translation techniques to specularity removal is the lack of good training data. It is rather impractical to take enough paired or even unpaired photos to successfully train a deep network. Inspired by the recent works of simulating training data by physically-based rendering [11–14] and domain adaptation [15–18], we present a fine-tuned process for generating training data, then adapting it to real world data. Instead of using Shapenet [19], we develop a new training dataset that includes models with richer geometric details, which allows us to apply our method to complex real-world data. Both quantitative and qualitative evaluations demonstrate that the performance of multi-view reconstruction can be significantly improved using the images filtered by our network. We show also the performance of adapting our network on real world training and testing data with some promising results.

2 Related work

Specular Object Reconstruction. Image based 3D reconstruction has been widely used for AR/VR applications, and the reconstruction speed and quality have been improved dramatically in recent years. However, most photometric stereo methods are based on the assumption that the object surface is diffuse, that is, the appearance of the object is view independent. Such assumptions, however, are not valid for glossy or specular objects in uncontrolled environments. It is well known that modeling the specularity is difficult as the specular effects are largely caused by the complicated global illumination that is usually unknown. For example, Godard et al. [7] first reconstruct a rough model by silhouette and then refine it using the specified environment map. Their method can reconstruct high quality specular surfaces from HDR images with extra information, such as silhouette and environment map.

In contrast, our method requires only the multi-view images as input. Researchers have proposed sophisticated equipment, such as a setup with two-layer LCDs to encode the directions of the emitted light field [4], taking advantages of the IR images recorded by RGB-D scanners [20, 21] or casting coded patterns onto mirror-like objects [3]. While such techniques can effectively handle challenging non-diffuse effects, they require additional hardware and user expertise. Another way to tackle this problem is by introducing additional assumptions, such as surface constraints [5, 6], the Blinn-Phong model [8], and shape-from-specularity [22]. These methods can also benefit from our network that outputs diffuse images, where strong specularities are removed from uncontrolled illumination. Please refer to [23] for a survey on specular object reconstruction.

GAN-based Image-to-Image Translation. We are inspired by the latest success of learning based image-to-image translation methods, such as ConditionalGAN [9], cycleGAN [10], [24] dualGAN, and discoGAN [17]. The remarkable capacity of Generative Adversarial Networks (GANs) [25] in modeling data distributions allows these methods to transform images from one domain to another with relatively small amounts of training data, while preserving the intrinsic structure of original images faithfully. With improved multi-scale training techniques, such as Progressive GAN [26] and pix2pixHD [27], image-to-image translation can be performed at mega pixel resolutions and achieve results of stunning visual quality.

Recently, modified image-to-image translation architectures have been successfully applied to ill-posed or underconstrained vision tasks, including face frontal view synthesis [28], facial geometry reconstruction [29–32], raindrop removal [33], or shadow removal [34]. These applications motivate us to develop a glossiness removal method based on GANs to facilitate multi-view 3D reconstruction of non-diffuse objects.

Learning-based Multi-View 3D Reconstruction. Learning surface reconstruction from multi-view images end-to-end has been an active research direction recently [35–38]. Wu et al. [39] and Gwak et al. [40] use GANs to learn the latent space of shapes and apply it to single image 3D reconstruction. 3D-R2N2 [36] designs a recurrent network for unified single and multi-view reconstruction. Image2Mesh [41] learns parameters

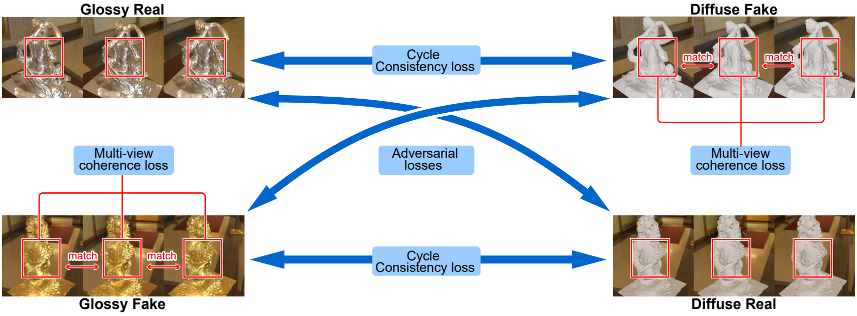

Fig. 2: Overview of S2Dnet. Two generators and two discriminators are trained simultaneously to learn cross-domain translations between the glossy and the diffuse domain. In each training iteration, the model randomly picks and forwards a real glossy and diffuse image sequence, computes the loss functions and updates the model parameters.

of free-form-deformation of a base model. Nonetheless, in general, the reconstruction quality of these methods cannot really surpass that of traditional approaches that exploit multiple-view geometry and heavily engineered photometric stereo pipelines. To take the local image feature coherence into account, we focus on removing the specular effect on the image level and resort to the power of multi-view reconstruction as a post-processing and also a production step.

On the other hand, there are works, closer to ours, that focus on applying deep learning on subparts of the stereo reconstruction pipeline, such as depth and pose estimation [42], feature point detection and description [43,44], semantic segmentation [45], and bundle adjustment [46,47]. These methods still impose the Lambertian assumption for objects or scenes, where our method can serve as a preprocessing step to deal with glossiness.

Learning-based Intrinsic Image Decomposition. Our method is also loosely related to some recent works on learning intrinsic image decomposition. These methods include training a CNN to reconstruct rendering parameters, e.g., material [48, 49], reflectance maps [50], illumination [51], or some combination of those components [13, 48, 52]. These methods are often trained on synthetic data and are usually applied to the re-rendering of single images. Our method shares certain similarity with these methods. However, our goal is not to recover intrinsic images with albedos. Disregarding albedo, we aim for output images with a consistent appearance across the entire training set that reflects the structure of the object.

3 Multi-view Specular-to-Diffuse GAN

In this section, we introduce S2Dnet, a conditional GAN that translates multi-view images of highly specular scenes into corresponding diffuse images. The input to our model is a multi-view sequence of a glossy scene without any additional input such as segmentation masks, camera parameters, or light probes. This enables our model

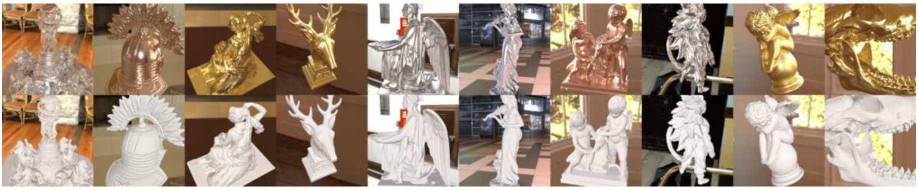

Fig. 3: Gallery of our synthetically rendered specular-to-diffuse training data.

to process real-world data, where such additional information is not readily available. The output of our model directly serves as input to state-of-the-art photometric stereo pipelines, resulting in improved 3D reconstruction without additional effort. Figure 2 shows a visualization of the proposed model. We discuss the training data, one of our major contributions, in Section 3.1. In Section 3.2 we introduce the concept of inter-view coherence that enables our model to process multiple views of a scene in a consistent manner, which is important in the context of multi-view reconstruction. Then, we outline in Section 3.3 the overall end-to-end training procedure. Implementation details are discussed in Section 3.4. Upon publication we will release both our data (synthetic and real) and the proposed model to foster further work.

3.1 Training Data

To train our model to translate multi-view glossy images to diffuse correspondents, we need appropriate data for both domains, i.e., glossy source domain images as inputs, and diffuse images as the target domain. Yi et al. [24] propose a MATERIAL dataset consisting of unlabeled data grouped in different material classes, such as plastic, fabric, metal, and leather, and they train GANs to perform material transfer. However, the MATERIAL dataset does not contain multi-view images and thus is not suited for our application. Moreover, the dataset is rather small and we expect our deep model to require a larger amount of training data. Hence, we propose a novel synthetic dataset consisting of multi-view images, which is both sufficiently large to train deep networks and complex to generalize to real-world objects. For this purpose, we collect and align 91 watertight and noise-free geometric models featuring rich geometric details from SketchFab (Figure 3). We exclude three models for testing and use the remaining 88 models for training. To obtain a dataset that generalizes well to real-world images, we use PBRT, a physically based renderer [53] to render these geometric models in various environments with a wide variety of glossy materials applied to form our source domain. Next, we render the target domain images by applying a Lambertian material to our geometric models.

Our experiments show that the choice of the rendering parameters has a strong impact on the translation performance. On one hand, making the two domains more similar by choosing similar materials for both domains improves the translation quality on synthetic data. Moreover, simple environments, such as a constant ground plane, also increase the quality on synthetic data. On the other hand, such simplifications cause the model to overfit and prevent generalization to real-world data. Hence, a main goal of our dataset is to provide enough complexity to allow generalization to real data. To achieve realistic illumination, we randomly sample one of 20 different HDR indoor environment maps and randomly rotate it for each scene. In addition, we orient a directional light source pointing from the camera approximately towards the center of the scene and position two additional light sources above the scene. The intensities, positions, and directions of these additional light sources are randomly jittered. This setup guarantees a rather even, but still random illumination. To render the source domain images, we applied the various metal materials defined in PBRT, including copper, silver, and gold. Material roughness and index of refraction are randomly sampled to cover a large variety of glossy materials. We randomly sample camera positions on the upper hemisphere around the scene pointing towards the center of the scene. To obtain multi-view data, we always sample 5 close-by, consecutive camera positions in clock-wise order while keeping the scene parameters fixed to mimic the common procedure of taking photos for stereo reconstruction. Since we collect 5 images of the same scene and the input to our network consists of 3 views, we obtain 3 training samples per scene. All rendered images are of 512×512 resolution, which is the limit for our GPU memory. However, it is likely that higher resolutions would further improve the reconstruction quality. Finally, we render the exact same images again with a white, Lambertian material, i.e., the mapping from the source to the target domain is bijective. The proposed procedure results in a training dataset of more than 647k images, i.e., more than 320k images per domain. For testing, we rendered 2k sequences of images, each consisting of 50 images. All qualitative results on synthetic data shown in this paper belong to this test set.

3.2 Inter-view Coherence

Multi-view reconstruction algorithms leverage corresponding features in different views to accurately estimate the 3D geometry. Therefore, we cannot expect good reconstruction quality if the glossy images in a multi-view sequence are translated independently using standard image translation methods, e.g., [9, 10]. This will introduce inconsistencies along the different views, and thus cause artifacts in the subsequent reconstruction. We therefore propose a novel model that enforces inter-view coherence by processing multiple views simultaneously. Our approach consists of a global and local consistency constraint: the global constraint is implemented using an appropriate network architecture, and the local consistency is enforced using a novel loss function.

Global Inter-view Coherence. A straightforward idea to incorporate multiple views is to stack them pixel-by-pixel before feeding them to the network. We found that this does not lead to strong enough constraints, since the network can still learn independent filter weights for the different views. This results in blurry translations, especially if corresponding pixels in different views are not aligned, which is typically the case. Instead, we concatenate the different views along the spatial axis before feeding them to the network. This solution, although simple, enforces the network to use the same filter weights for all views, and thus effectively avoids inconsistencies on a global scale.

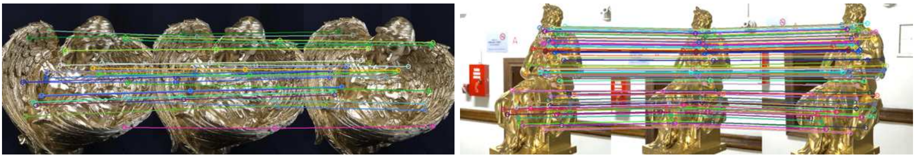

Fig. 4: Two examples of the SIFT correspondences pre-computed for our training.

Local Inter-view Coherence. Incorporating loss functions based on local image patches has been successfully applied to generative adversarial models, such as image completion [54] or texture synthesis [55]. However, comparing image patches at random locations is not meaningful in a multi-view setup for stereo reconstruction. Instead, we encourage the network to maintain feature point correspondences in the input sequence, i.e., inter-view correspondences should be invariant to the translation. Since the subsequent reconstruction pipeline relies on such correspondences, maintaining them during translation should improve reconstruction quality. To achieve this, we first extract SIFT feature correspondences for all training images. For each training sequence consisting of three views, we compute corresponding feature points between the different views in the source domain; see Figure 4 for two examples. During training, we encourage the network output at the SIFT feature locations to be similar along the views using a perceptual loss in VGG feature space [27,56–58]. The key idea is to measure both high- and low-level similarity of two images by considering their feature activations in a deep CNN like VGG. We adopt this idea to keep local image patches around corresponding SIFT features perceptually similar in the translated output. The perceptual loss in VGG feature space is defined as:

$$\mathcal{L}_{VGG}(x, \hat{x}) = \sum_{i=1}^{N} \frac{1}{M_i} \| F^{(i)}(x) - F^{(i)}(\hat{x}) \|_1, \tag{1}$$

where $F^{(i)}$ denotes the i-th layer in the VGG network consisting of M_i elements. Now consider a glossy input sequence consisting of three images X_1, X_2, X_3, and the corresponding diffuse sequence $\tilde{X_1}, \tilde{X_2}, \tilde{X_3}$ produced by our model. A SIFT correspondence for this sequence consists of three image coordinates p_1, p_2, p_3, one in each glossy image, and all three pixels at the corresponding coordinates represent the same feature. We then extract local image patches $\tilde{x_i}$ centered at p_i from $\tilde{X_i}$, and define the perceptual correspondence loss as:

$$\mathcal{L}_{corr}(\tilde{X_1}, \tilde{X_2}, \tilde{X_3}) = \mathcal{L}_{VGG}(\tilde{x_1}, \tilde{x_2}) + \mathcal{L}_{VGG}(\tilde{x_2}, \tilde{x_3}) + \mathcal{L}_{VGG}(\tilde{x_1}, \tilde{x_3}). \tag{2}$$

3.3 Training Procedure

Given two sets of data samples from two domains, a source domain A and a target domain B, the goal of image translation is to find a mapping T that transforms data points $X_i \in A$ to B such that $T(X_i) = \tilde{X_i} \in B$, while the intrinsic structure of X_i should be preserved under T. Training GANs has been proven to produce astonishing results on this task, both in supervised settings where the data of the two domains are paired [9], and in unsupervised cases using unpaired data [10]. In our experiments, we observed that both approaches (ConditionalGAN [9] and cycleGAN [10]) perform similarly well on our dataset. However, while paired training data might be readily available for synthetic data, paired real-world data is difficult to obtain. Therefore we come up with a design for unsupervised learning that can easily be fine-tuned on unpaired real-world data.

Cycle-consistency Loss. Similar to CycleGAN [10], we learn the mapping between domain A and B with two translators $G_B : A \to B$ and $G_A : B \to A$ that are trained simultaneously. The key idea is to train with cycle-consistency loss, i.e., to enforce that $G_A(G_B(X)) \approx X$ and $G_B(G_A(Y)) \approx Y$, where $X \in A$ and $Y \in B$. This cycle-consistency loss guarantees that data points preserve their intrinsic structure under the learned mapping. Formally, the cycle-consistency loss is defined as:

$$\mathcal{L}_{cyc}(X,Y) = \|G_A(G_B(X)) - X\|_1 + \|G_B(G_A(Y)) - Y\|_1. \tag{3}$$

Adversarial Loss. To enforce the translation networks to produce data that is indistinguishable from genuine images, we also include an adversarial loss to train our model. For both translators, in GAN context often called generators, we train two additional discriminator networks D_A and D_B that are trained to distinguish translated from genuine images. To train our model, we use the following adversarial term:

$$\mathcal{L}_{adv} = \mathcal{L}_{GAN}(G_A, D_A) + \mathcal{L}_{GAN}(G_B, D_B), \tag{4}$$

where $\mathcal{L}_{GAN}(G, D)$ is the LSGAN formulation [59].

Overall, we train our model using the following loss function:

$$\mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{corr}\mathcal{L}_{corr}, \tag{5}$$

where λ_{adv}, λ_{cyc}, and λ_{corr} are user-defined hyperparameters.

3.4 Implementation Details

Our model is based on cycleGAN and implemented in Pytorch. We experimented with different architectures for the translation networks, including U-Net [60], ResNet [61], and RNN-blocks [62]. Given enough training time, we found that all networks produce similar results. Due to its memory efficiency and fast convergence, we chose U-Net for our final model. As shown in Figure 5, we use the multi-scale discriminator introduced in [27] that downsamples by a rate of 2, which generally works better for high resolution images. Our discriminator also considers the local correspondence patches as additional input, which helps to produce coherent translations. Followed by the training guidances proposed in [26], we use pixel-wise normalization in the generators and add a 1-strided

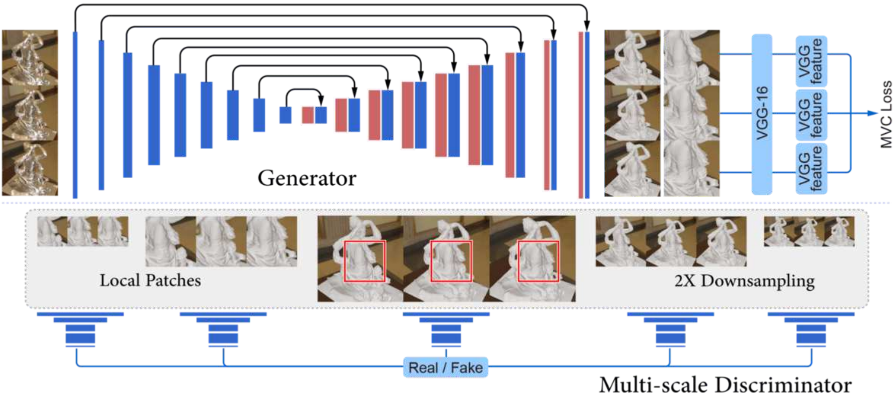

Fig. 5: Illustration of the generator and discriminator network. The generator uses the U-net architecture and both input and output are a multi-view sequence consisting of three views. A random SIFT correspondence is sampled during training to compute the correspondence loss. The multi-scale joint discriminator examines three scales of the image sequence and two scales of corresponding local patches. The width and height of each rectangular block indicate the channel size and the spatial dimension of the output feature map, respectively.

convolutional layer after each deconvolutional layer. For computing the correspondence loss, we use a patch size of 256×256 and sample a single SIFT correspondence per training iteration randomly. The discriminator follows the architecture as: C64-C128-C256-C512-C1. The generator's encoder architecture is: C64-C128-C256-C512-C512-C512-C512-C512. We use $\lambda_{adv} = 1, \lambda_{cyc} = 10, \lambda_{corr} = 5$ in all our experiments and train using the ADAM optimizer with a learning rate of 0.0002.

4 Evaluation

In this section, we present qualitative and quantitative evaluations of our proposed S2Dnet. For this purpose, we evaluate the performance of our model on both the translation task and the subsequent 3D reconstruction, and we compare to several baseline systems. In Section 4.1 we report results on our synthetic test set and we also perform an evaluation on real-world data in Section 4.2.

To evaluate the benefit of our proposed inter-view coherence, we perform a comparison to a single-view translation baseline by training a cycleGAN network [10] on glossy to diffuse translation. Since our synthetic dataset features a bijective mapping between glossy and diffuse images, we also train a pix2pix network [9] for a supervised baseline on synthetic data. In addition, we compare reconstruction quality to performing stereo reconstruction directly on the glossy multi-view sequence to demonstrate the benefit of translating the input as a preprocessing step. For 3D reconstruction, we apply a state-of-the-art multi-view surface reconstruction method [1] on input sequences consisting

	Glossy	pix2pix	cycleGAN	S2Dnet
Image MSE	118.39	56.20	69.15	57.78

Table 1: Quantitative evaluation of the image error on our synthetic testing data.

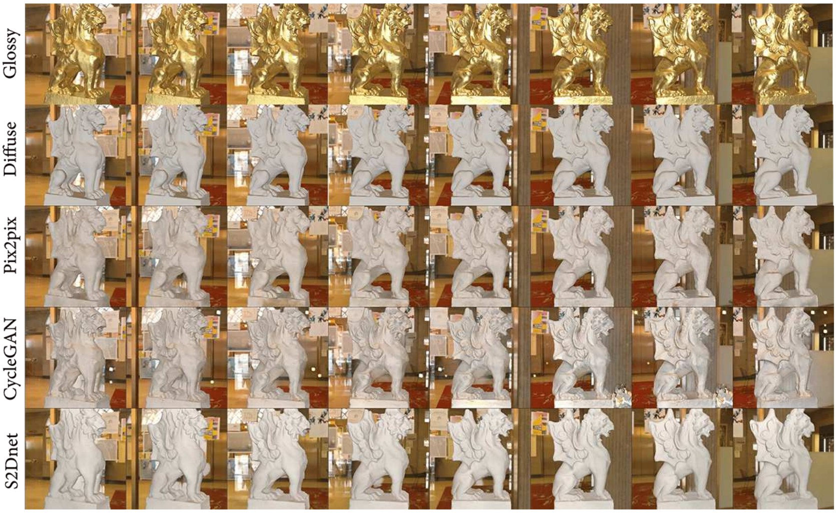

Fig. 6: Qualitative translation results on a synthetic input sequence consisting of 8 views. From top down: the glossy input sequence, the ground truth diffuse rendering, and the translation results for the baselines pix2pix and cycleGAN, and our S2Dnet. The output of pix2pix is generally blurry. The cycleGAN output, although sharp, lacks inter-view consistency. Our S2Dnet produces both crisp and coherent translations.

of 10 to 15 views. For our method, we translate each input view sequentially but we feed the two neighboring views as additional inputs to our multi-view network. For the two baseline translation methods, we translate each view independently. The 3D reconstruction pipeline then uses the entire translated multi-view sequence as input.

4.1 Synthetic Data

For a quantitative evaluation of the image translation performance, we compute MSE with respect to the ground truth diffuse renderings on our synthetic test set. Table 1 shows a comparison of our S2Dnet to pix2pix and cycleGAN. Unsurprisingly, the supervised pix2pix network performs best, closely followed by our S2Dnet, which outperforms the unsupervised baseline by a significant margin. In Figure 6 we show qualitative translation results. Note that the output of pix2pix is generally blurry. Since MSE penalizes outliers and prefers a smooth solution, pix2pix still achieves a low MSE error. While the output of cycleGAN is sharper, the translated sequence lacks inter-view consistency, whereas our S2Dnet produces both highly detailed and coherent translations.

Model	1	2	3	4	5	6	7	8	9	10	AVG
Glossy	0.67	0.88	1.35	0.76	1.15	1.13	1.15	0.78	0.54	0.66	0.90
cycleGAN	1.18	0.72	0.89	0.59	1.35	0.72	0.99	0.62	0.51	0.42	0.80
S2Dnet	0.52	0.67	0.72	0.43	0.87	0.54	0.92	0.65	0.55	0.56	0.64

Table 2: Quantitative evaluation of surface reconstruction performance on 10 different scenes. The error metric is the percentage of bounding box diagonal. Our S2Dnet performs best, and the translation baseline still performs significantly better than directly reconstructing from the glossy images. The numbering of the models follows the visualization in Figure 7, using the same left to right order.

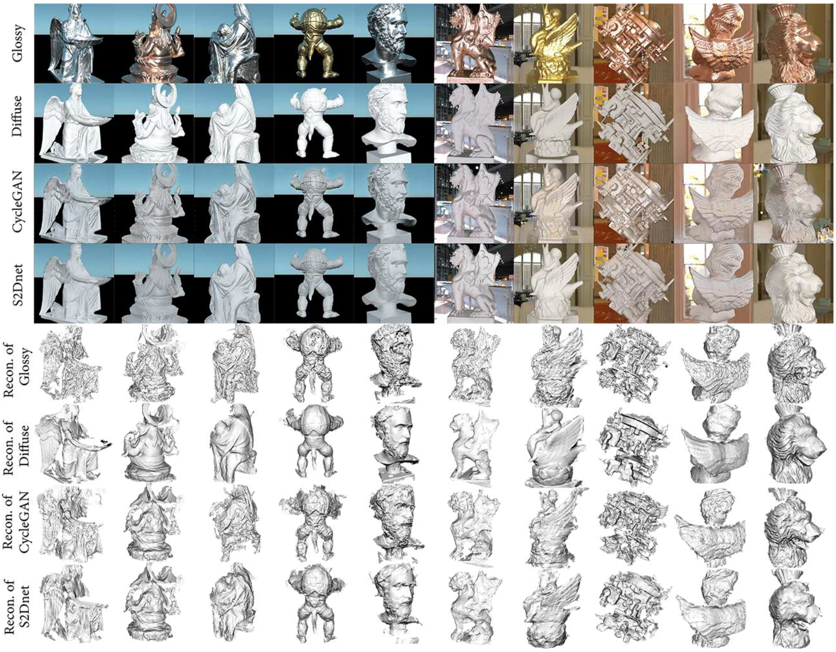

Fig. 7: Qualitative surface reconstruction results on 10 different scenes. From top to bottom: glossy input, ground truth diffuse renderings, cycleGAN translation outputs, our S2Dnet translation outputs, reconstructions from glossy images, reconstructions from ground truth diffuse images, reconstructions from cycleGAN output, and reconstructions from our S2Dnet output. All sequences are excluded from our training set, and the objects in column 3 and 4 have not even been seen during training.

Next, we evaluate the quality of the surface reconstruction by feeding the translated sequences to the reconstruction pipeline. We found that the blurry output of pix2pix is not suitable for stereo reconstruction, since already the first step, estimating camera parameters based on feature correspondences, fails on this data. We therefore exclude pix2pix from the surface reconstruction evaluation but include the trivial baseline of directly reconstructing from the glossy input sequence to demonstrate the benefit of

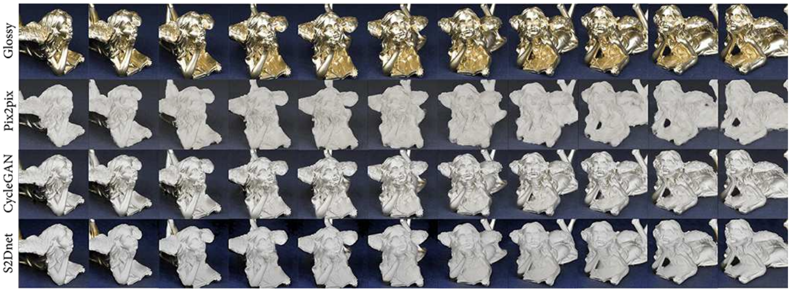

Fig. 8: Qualitative translation results on a real-world input sequence consisting of 11 views. The first row shows the glossy input sequence and the remaining rows show the translation results of pix2pix, cycleGAN, and our S2Dnet. All networks are trained on synthetic data only. Similar to the synthetic case, cycleGAN outperforms pix2pix, but it produces high-frequency artifacts that are not consistent along the views. Our S2Dnet is able to remove most of the specular effects and preserves all the geometric details in a consistent manner.

the translation step. In order to compute the geometric error of the surface reconstruction output, we register the reconstructed geometry to the ground truth mesh using a variant of ICP [63]. Next, we compute the Euclidean distance of each reconstructed surface point to its nearest neighbor in the ground truth mesh and report the per-model mean value. Table 2 shows the surface reconstruction error for our S2Dnet in comparison to the three baselines. The numbers show that our S2Dnet performs best, and that preprocessing the glossy input sequences clearly helps to obtain a more accurate reconstruction, even when using the cycleGAN baseline. In Figure 7 we show qualitative surface reconstruction results for 10 different scenes in various environments.

4.2 Real-world Data

Since we do not have real-world ground truth data, we compile a real-world test set and perform a qualitative comparison on it. For all methods, we compare generalization performance when training on our synthetic dataset. Moreover, we evaluate how the different models perform when fine-tuning on real-world data, or training on real-world data from scratch. For this purpose, we compile a dataset by shooting photos of real-world objects. We choose 5 diffuse real-world objects and take 5k pictures in total from different camera positions and under varying lighting conditions. Next, we use a glossy spray paint to cover our objects with a glossy coat and shoot another 5k pictures to represent the glossy domain. The resulting dataset consists of unpaired samples of glossy and diffuse objects under real-world conditions, see Figure 10 a) and b).

In Figure 8 we show qualitative translation results on real-world data. All networks are trained on synthetic data only here, and they all manage to generalize to some extent to real-world data, thanks to our high-quality synthetic dataset. Similar to the synthetic

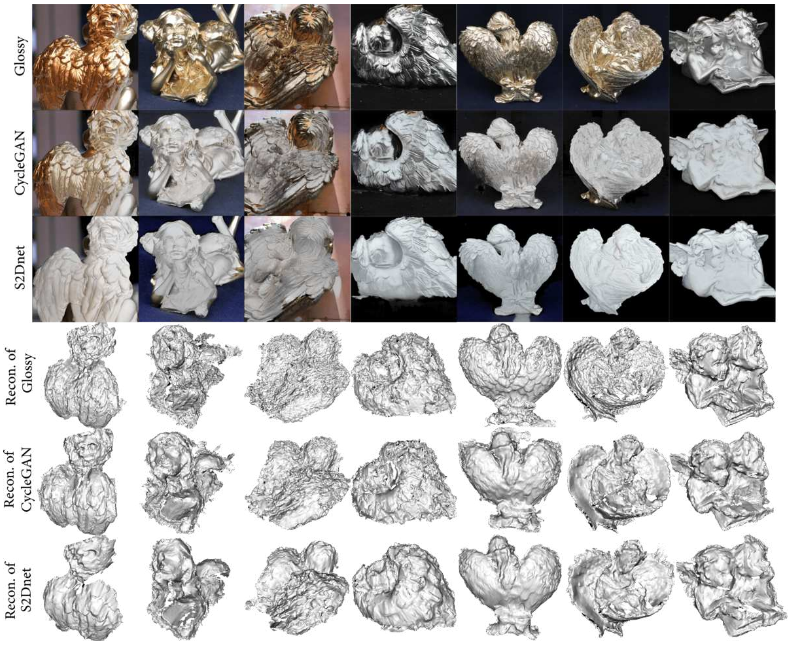

Fig. 9: Qualitative surface reconstruction results on 7 different real-world scenes. Top to bottom: glossy input, cycleGAN translation outputs, our S2Dnet translation outputs, reconstructions from glossy images, reconstructions from cycleGAN output, and reconstructions from our S2Dnet output. All networks are trained on synthetic data only.

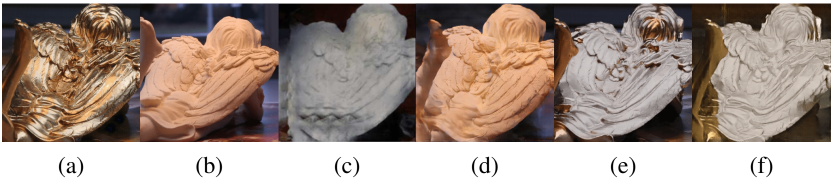

Fig. 10: a), b) A sample of our real-world dataset. c) translation result of cycleGAN when training from scratch on our real-world dataset. d) S2Dnet output, trained from scratch on our real-world dataset. e) S2Dnet output, trained on synthetic data only. f) S2Dnet output, trained on synthetic data, fine-tuned on real-world data.

results in Figure 6, pix2pix produces blurry results, while cycleGAN introduces inconsistent high-frequency artifacts. S2Dnet is able the remove most of the specular effects and preserves geometric details in a consistent manner. In Figure 9 we show qualitative surface reconstruction results for 7 different scenes. Artifacts occur mainly close to the object silhouettes in complex background environments. This could be mitigated by training with segmentation masks.

Finally, we evaluate performance when either fine-tuning or training from scratch on real-world data. We retrain or fine-tune S2Dnet and cycleGAN on our real-world dataset, but cannot retrain pix2pix for this purpose, since it relies on a supervision signal that is not present in our unpaired real-world dataset. Our experiments show that training or fine-tuning using such a small dataset leads to heavy overfitting. The translation performance for real-world objects that were not seen during training decreases significantly compared to the models trained on synthetic data only. In Figure 10 c) and d) we show image translation results of cycleGAN and S2Dnet when training from scratch on our real-world dataset. Since the scene in Figure 10 is part of the training set (although the input image itself is excluded from the training set), our S2Dnet produces decent translation results, which is not the case for scenes not seen during training. Fine-tuning our S2Dnet produces similar results (Figure 10 f)).

5 Limitations and Future Work

Although the proposed framework enables reconstructing glossy and specular objects more accurately compared to state-of-the-art 3D reconstruction algorithms, a few limitations do exist. First, since the network architecture contains an encoder and a decoder with skip connections, the glossy-to-Lambertian image translation is limited to images of a fixed resolution. This resolutions might be too low for certain types of applications. Next, due to the variability of the background in real images, the translation network might treat a portion of the background as part of the reconstructed object. Similarly, the network occasionally misclassifies the foreground as part of the background, especially in very light domains on specular objects. Finally, as the simulated training data was rendered by assuming a fixed albedo, the network cannot consistently translate glossy materials with spatially varying albedo into a Lambertian surface. We predict that given a larger and more diverse training set in terms of shapes, backgrounds, albedos and materials, the accuracy of the proposed method in recovering real object would be largely enhanced. Our current training dataset includes the most common types of specular material. The proposed translation network has potential to be extended to other more challenging materials, such as transparent objects, given proper training data.

Acknowledgement

We thank the anonymous reviewers for their constructive comments. This work was supported in parts by Swiss National Science Foundation (169151), NSFC (61522213, 61761146002, 61861130365), 973 Program (2015CB352501), Guangdong Science and Technology Program (2015A030312015), ISF-NSFC Joint Research Program (2472/17) and Shenzhen Innovation Program (KQJSCX20170727101233642).

References

1. Langguth, F., Sunkavalli, K., Hadap, S., Goesele, M.: Shading-aware multi-view stereo. In: Proceedings of the European Conference on Computer Vision (ECCV). (2016)
2. Maier, R., Kim, K., Cremers, D., Kautz, J., Niessner, M.: Intrinsic3d: High-quality 3d reconstruction by joint appearance and geometry optimization with spatially-varying lighting. 2017 IEEE International Conference on Computer Vision (ICCV) (2017) 3133–3141
3. Tarini, M., Lensch, H.P.A., Goesele, M., Seidel, H.P.: 3d acquisition of mirroring objects using striped patterns. Graph. Models **67**(4) (July 2005) 233–259
4. Tin, S.K., Ye, J., Nezamabadi, M., Chen, C.: 3d reconstruction of mirror-type objects using efficient ray coding. In: 2016 IEEE International Conference on Computational Photography (ICCP). (May 2016) 1–11
5. Ikeuchi, K.: Determining surface orientations of specular surfaces by using the photometric stereo method. IEEE Trans. Pattern Analysis & Machine Intelligence (6) (1981) 661–669
6. Savarese, S., Perona, P.: Local analysis for 3d reconstruction of specular surfaces. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. Volume 2. (2001) II–738–II–745 vol.2
7. Godard, C., Hedman, P., Li, W., Brostow, G.J.: Multi-view Reconstruction of Highly Specular Surfaces in Uncontrolled Environments. In: 3DV. (2015)
8. Khanian, M., Boroujerdi, A.S., Breuß, M.: Photometric stereo for strong specular highlights. Computational Visual Media (Feb 2018)
9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)
10. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
11. Zhang, Y., Song, S., Yumer, E., Savva, M., Lee, J.Y., Jin, H., Funkhouser, T.: Physically-based rendering for indoor scene understanding using convolutional neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)
12. Movshovitz-Attias, Y., Kanade, T., Sheikh, Y.: How useful is photo-realistic rendering for visual learning? In: European Conference on Computer Vision, Springer (2016) 202–217
13. Shi, J., Dong, Y., Su, H., Yu, S.X.: Learning non-lambertian object intrinsics across shapenet categories. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)
14. Meka, A., Maximov, M., Zollhoefer, M., Chatterjee, A., Richardt, C., Theobalt, C.: Live intrinsic material estimation. arXiv preprint arXiv:1801.01075 (2018)
15. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213 (2017)
16. Benaim, S., Wolf, L.: One-sided unsupervised domain mapping. In Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., eds.: Advances in Neural Information Processing Systems 30. Curran Associates, Inc. (2017) 752–762
17. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In Precup, D., Teh, Y.W., eds.: Proceedings of the 34th International Conference on Machine Learning. Volume 70 of Proceedings of Machine Learning Research., International Convention Centre, Sydney, Australia, PMLR (06–11 Aug 2017) 1857–1865
18. Kang, G., Zheng, L., Yan, Y., Yang, Y.: Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. arXiv preprint arXiv:1801.10068 (2018)

19. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago (2015)
20. Or-El, R., Hershkovitz, R., Wetzler, A., Rosman, G., Bruckstein, A.M., Kimmel, R.: Real-time depth refinement for specular objects. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. (2016) 4378–4386
21. Or-El, R., Rosman, G., Wetzler, A., Kimmel, R., Bruckstein, A.M.: Rgbd-fusion: Real-time high precision depth recovery. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. (2015) 5407–5416
22. Chen, T., Goesele, M., Seidel, H.P.: Mesostructure from specularity. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Volume 2. (2006) 1825–1832
23. Ihrke, I., Kutulakos, K.N., Lensch, H.P.A., Magnor, M., Heidrich, W.: Transparent and Specular Object Reconstruction. Computer Graphics Forum (2010)
24. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
25. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., eds.: Advances in Neural Information Processing Systems 27. Curran Associates, Inc. (2014) 2672–2680
26. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations. (2018)
27. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. arXiv preprint arXiv:1711.11585 (2017)
28. Huang, R., Zhang, S., Li, T., He, R.: Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
29. Richardson, E., Sela, M., Kimmel, R.: 3d face reconstruction by learning from synthetic data. In: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE (2016) 460–469
30. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
31. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition, IEEE (2017) 5553–5562
32. Sengupta, S., Kanazawa, A., Castillo, C.D., Jacobs, D.: Sfsnet: Learning shape, reflectance and illuminance of faces in the wild. arXiv preprint arXiv:1712.01261 (2017)
33. Qian, R., Tan, R.T., Yang, W., Su, J., Liu, J.: Attentive generative adversarial network for raindrop removal from a single image. arXiv preprint arXiv:1711.10098 (2017)
34. Wang, J., Li, X., Hui, L., Yang, J.: Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. arXiv preprint arXiv:1712.02478 (2017)
35. Tatarchenko, M., Dosovitskiy, A., Brox, T.: Multi-view 3d models from single images with a convolutional network. In: European Conference on Computer Vision (ECCV). (2016)
36. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: Proceedings of the European Conference on Computer Vision (ECCV). (2016)

37. Lin, C.H., Kong, C., Lucey, S.: Learning efficient point cloud generation for dense 3d object reconstruction. In: AAAI Conference on Artificial Intelligence (AAAI). (2018)
38. Tulsiani, S., Zhou, T., Efros, A.A., Malik, J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)
39. Wu, J., Zhang, C., Xue, T., Freeman, B., Tenenbaum, J.: Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: In Advances in Neural Information Processing Systems (NIPS). (2016) 82–90
40. Gwak, J., Choy, C.B., Chandraker, M., Garg, A., Savarese, S.: Weakly supervised 3d reconstruction with adversarial constraint. In: 3D Vision (3DV), 2017 Fifth International Conference on 3D Vision. (2017)
41. Pontes, J.K., Kong, C., Sridharan, S., Lucey, S., Eriksson, A., Fookes, C.: Image2mesh: A learning framework for single image 3d reconstruction. arXiv preprint arXiv:1711.10669 (2017)
42. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)
43. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: Lift: Learned invariant feature transform. In: European Conference on Computer Vision, Springer (2016) 467–483
44. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. arXiv preprint arXiv:1712.07629 (2017)
45. Ma, L., Stueckler, J., Kerl, C., Cremers, D.: Multi-view deep learning for consistent semantic mapping with rgb-d cameras. In: iros, Vancouver, Canada (Sep 2017)
46. Zhu, R., Wang, C., Lin, C.H., Wang, Z., Lucey, S.: Object-centric photometric bundle adjustment with deep shape prior. arXiv preprint arXiv:1711.01470 (2017)
47. Zhu, R., Wang, C., Lin, C.H., Wang, Z., Lucey, S.: Semantic photometric bundle adjustment on natural sequences. arXiv preprint arXiv:1712.00110 (2017)
48. Liu, G., Ceylan, D., Yumer, E., Yang, J., Lien, J.M.: Material editing using a physically based rendering network. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
49. Yu, Y., Smith, W.A.: Pvnn: A neural network library for photometric vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 526–535
50. Rematas, K., Ritschel, T., Fritz, M., Gavves, E., Tuytelaars, T.: Deep reflectance maps. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2016)
51. Georgoulis, S., Rematas, K., Ritschel, T., Fritz, M., Van Gool, L., Tuytelaars, T.: Delight-net: Decomposing reflectance maps into specular materials and natural illumination. arXiv preprint arXiv:1603.08240 (2016)
52. Georgoulis, S., Rematas, K., Ritschel, T., Fritz, M., Tuytelaars, T., Van Gool, L.: What is around the camera? In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. (2017) 5170–5178
53. Pharr, M., Humphreys, G.: Physically based rendering, second edition: From theory to implementation. (2010)
54. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Transactions on Graphics (TOG) **36**(4) (2017) 107
55. Xian, W., Sangkloy, P., Lu, J., Fang, C., Yu, F., Hays, J.: Texturegan: Controlling deep image synthesis with texture patches. arXiv preprint arXiv:1706.02823 (2017)
56. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on, IEEE (2016) 2414–2423
57. Wang, C., Xu, C., Wang, C., Tao, D.: Perceptual adversarial networks for image-to-image transformation. arXiv preprint arXiv:1706.09138 (2017)

58. Vansteenkiste, E., Kern, P.: Taming adversarial domain transfer with structural constraints for image enhancement. arXiv preprint arXiv:1712.00598 (2017)
59. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: ICCV, IEEE (2017) 2813–2821
60. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241
61. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition. (June 2016)
62. Chaitanya, C.R.A., Kaplanyan, A.S., Schied, C., Salvi, M., Lefohn, A., Nowrouzezahrai, D., Aila, T.: Interactive reconstruction of monte carlo image sequences using a recurrent denoising autoencoder. ACM Trans. Graph. **36**(4) (July 2017) 98:1–98:12
63. Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: 3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on, IEEE (2001) 145–152