Semantically Supervised Appearance Decomposition for Virtual Staging from a Single Panorama

TIANCHENG ZHI, Carnegie Mellon University, USA
BOWEI CHEN, Carnegie Mellon University, USA
IVAYLO BOYADZHIEV, Zillow Group, USA
SING BING KANG, Zillow Group, USA
MARTIAL HEBERT, Carnegie Mellon University, USA
SRINIVASA G. NARASIMHAN, Carnegie Mellon University, USA

We describe a novel approach to decompose a single panorama of an empty indoor environment into four appearance components: specular, direct sunlight, diffuse and diffuse ambient without direct sunlight. Our system is weakly supervised by automatically generated semantic maps (with floor, wall, ceiling, lamp, window and door labels) that have shown success on perspective views and are trained for panoramas using transfer learning without any further annotations. A GAN-based approach supervised by coarse information obtained from the semantic map extracts specular reflection and direct sunlight regions on the floor and walls. These lighting effects are removed via a similar GAN-based approach and a semantic-aware inpainting step. The appearance decomposition enables multiple applications including sun direction estimation, virtual furniture insertion, floor material replacement, and sun direction change, providing an effective tool for virtual home staging. We demonstrate the effectiveness of our approach on a large and recently released dataset of panoramas of empty homes.

CCS Concepts: • Computing methodologies → Computational photography.

Additional Key Words and Phrases: appearance decomposition, reflection removal, sunlight detection, object insertion

1 INTRODUCTION

With the current pandemic-related restrictions and many working from home, there is a significant uptick in home sales. Remote home shopping is becoming more popular and effective tools to facilitate virtual home tours are much needed. One such tool is virtual staging: how would furniture fit in a home and what does it look like if 3D model credits: cake©HQ3DMOD/Adobe Stock, lounge©JaskaranSingh/Adobe Stock, cherries©VIZIPARK/Adobe Stock, table©Claudio Naviglio/Adobe Stock, tea©Brandon Westlake/Adobe Stock, dining set©Oleg Kuch/Adobe Stock, dinnerware©HQ3DMOD/Adobe Stock, utensil©Dmitrii Iaplatov/Adobe Stock, pot©Brandon Westlake/Adobe Stock, side table©JaskaranSingh/Adobe Stock, frames©Kamil/Adobe Stock, sofa©HQ3DMOD/Adobe Stock, envelope©adobestock3d/Adobe Stock, pillow©TurboSquad/Adobe Stock, beanbag©FongAudiashonza/Adobe Stock, stand©Dmitrii Iaplatov/Adobe Stock, plant©Kids Creative Agency/Adobe Stock, basket©adobestock3d/Adobe Stock.
certain settings (e.g. sun direction, flooring) are changed? To provide effective visualization for staging, panoramas are increasingly being used to showcase homes. Panoramas provide surround information but methods designed for limited-FOV perspective photos cannot be directly applied.

However, inserting virtual furniture into a single panoramic image of a room in a photo-realistic manner is hard. Multiple complex shading effects should be taken into consideration. Such complex interactions are shown in Fig. 1; the insertion induces effects such as occlusion, sunlight cast on objects, partially shadowed specular reflections, and soft and hard shadows. This task is challenging due to the lack of ground truth training data. The ground truth annotations for appearance decomposition tasks either require significant human labor [Bell et al. 2014], or specialized devices in a controlled environment [Grosse et al. 2009; Shen and Zheng 2013], which is hard to be extended to a large scale. Previous approaches [Fan et al. 2017; Li et al. 2020a; Lin et al. 2019] rely on synthetic data for supervised training. However, there is the issue of domain shift and the costs of designing and rendering scenes [Li et al. 2021b].

We present a novel approach to insert virtual objects into a single panoramic image of an empty room in a near-photo-realistic way. Instead of solving a general inverse rendering problem from a single image, we identify and focus on two ubiquitous shading effects that are important for visual realism: interactions of inserted objects with (1) specular reflections and (2) direct sunlight (see Fig. 1). This is still a challenging problem that has no effective solutions because of the same reasons mentioned above. We observe that while ground truth real data for these effects is hard to obtain, by contrast, it is easier to obtain semantics automatically, given recent advances on layout estimation [Pintore et al. 2020; Sun et al. 2019; Zou et al. 2018] and semantic segmentation [Chen et al. 2017; Long et al. 2015; Wang et al. 2020]. An automatically generated semantic map with ceiling, wall, floor, window, door, and lamp classes is used to localize the effects coarsely. These coarse localizations are used to supervise a GAN based approach to decompose the input image into four appearance components: (1) diffuse, (2) specular (on floor), (3) direct sunlight (on floor and walls), and (4) diffuse ambient without sunlight. The appearance decompositions can then be used for several applications including insertion of furniture, changing flooring, albedo estimation and estimating and changing sun direction.

We evaluate our method on ZInD [Cruz et al. 2021], a large dataset of panoramas of empty real homes. The homes include a variety of floorings (tile, carpet, wood), diverse configurations of doors, windows, indoor light source types and positions, and outdoor illumination (cloudy, sunny). We analyze our approach by comparing against ground truth specular and sunlight locations and sun directions, and by conducting ablation studies. Most previous approaches for diffuse-specular separation and inverse rendering were not designed for the setting used in this work to enable direct apples-to-apples comparisons; they require perspective views or supervised training of large scale real or rendered data. But we nonetheless show performances of such methods as empirical observations and not necessarily to prove that our approach is better in their settings.

Our work also has several limitations: (1) we assume the two shading effects occur either on the planar floor or walls, (2) our methods detect for mid-to-high frequency shading effects but not subtle low-frequency effects or specular/sunlight interreflections, and (3) our methods can induce artifacts if the computed semantic map is erroneous. Despite these limitations, our results suggest that the approach can be an effective and useful tool for virtual staging. Extending the semantics to furniture can enable appearance decomposition of already-furnished homes.
To sum up, our key idea is using easier-to-collect and annotate discrete signals (layout/windows/lamps) to estimate harder-to-collect continuous signals (specular/sunlight). This is a general idea that can be useful for any appearance estimation task. Our contributions include: (1) A semantically and automatically supervised framework for locating specular and direct sunlight effects (Sec. 3). (2) A GAN-based appearance separation method for diffuse, specular, ambient, and direct sunlight component estimation (Sec. 4 and 5). (3) Demonstration of multiple virtual staging applications including furniture insertion and changing flooring and sun direction (Sec. 7). To our knowledge, we are the first to estimate direct sunlight and sun direction from indoor images. The overall pipeline is illustrated in Fig. 2. Our code is released at: https://github.com/tiancheng-zhi/pano_decomp.

2 RELATED WORK

Inverse Rendering. The goal of inverse rendering is to estimate various physical attributes of a scene (e.g., geometry, material properties and illumination) given one or more images. Intrinsic image decomposition estimates reflectance and shading layers [Baslamisli et al. 2021; Janner et al. 2017; Li and Snively 2018a,b; Liu et al. 2020; Tappen et al. 2003]. Other methods attempt to recover scene attributes with simplified assumptions. Methods [Barron and Malik 2015; Boss et al. 2020; Kim et al. 2017; Li and Snively 2018b] for a single object use priors like depth-normal consistency and shape continuity. Some methods [Sengupta et al. 2018; Shu et al. 2017] use priors of a particular category (e.g., no occlusion for faces). Some works assume near-planar surfaces [Aittala et al. 2015; Deschaingt et al. 2018, 2019; Gao et al. 2019; Hu et al. 2022; Li et al. 2018]. In addition, human assistance with calibration or annotation is studied for general scenes [Karsch et al. 2011; Yu et al. 1999].

Data-driven methods require large amounts of annotated data, usually synthetic images [Li et al. 2021a, 2020a, 2021b; Wang et al. 2021a]. The domain gap can be reduced by fine-tuning on real images using self-supervised training via differentiable rendering [Sengupta et al. 2019]. The differentiable rendering can be used to optimize a single object [Kaya et al. 2021; Zhang et al. 2021]. Recent works [Boss et al. 2021a,b] extend NeRF [Mildenhall et al. 2020] to appearance decomposition. The combination of Bayesian framework and deep networks is explored by [Chen et al. 2021] for reflectance and illumination estimation. In our work, we model complex illumination effects on real 360° panoramas of empty homes. Similar to [Karsch et al. 2011], we believe that discrete semantic elements (like layout, windows, lamps, etc.) are easier to collect and train good models for. By contrast, diffuse and specular annotations are continuous spatially varying signals that are harder to label.

Illumination Estimation. Many approaches represent indoor lighting using HDR maps (or its spherical harmonics). Some estimate lighting from a single LDR panorama [Eilersten et al. 2017; Gkitsas et al. 2020], a perspective image [Gardner et al. 2017; Somanath and Kurz 2020], a stereo pair [Srinivasan et al. 2020], or object appearance [Georgoulis et al. 2017; Park et al. 2020; Weber et al. 2018]. Recent approaches [Garon et al. 2019; Li et al. 2020a; Wang et al. 2021a] extend this representation to multiple positions, enabling spatially-varying estimation. Others [Gardner et al. 2019; Jiddi et al. 2020; Karsch et al. 2014] estimate parametric lighting by modeling the position, shape, and intensity of light sources. Zhang et al. [2016] combine both representations and estimate a HDR map together with parametric light sources. However, windows are treated as the source of diffuse skylight without considering directional sunlight. We handle the spatially-varying high-frequency sun illumination effects, which is usually a challenging case for most methods.

Some techniques estimate outdoor lighting from outdoor images. Early methods [Lalonde et al. 2012, 2010] use analytical models to describe the sun and sky. Liu et al. [Liu et al. 2014] estimates sun direction using 3D object models. Recently, deep learning methods [Hold-Geoffroy et al. 2019, 2017; Zhang et al. 2019] regress the sun/sky model parameters or outdoor HDR maps by training on large scale datasets. A recent work [Swedish et al. 2021] estimates high-frequency illumination from shadows. However, they use outdoor images as input, where the occlusion of the sunlight by interior walls is not as significant as that for indoor scenes.

Specular Reflection Removal. There are two main classes of specular reflection removal techniques. One removes specular highlights on objects. Non-learning based approaches usually exploit appearance or statistical priors to separate specular reflection, including chromaticity-based models [Akashi and Okatani 2016; Shafer 1985; Tan and Ikeuchi 2005; Yang et al. 2010], low-rank model [Guo et al. 2018], and dark channel prior [Kim et al. 2013]. Recently, data-driven methods [Fu et al. 2021; Shi et al. 2017; Wu et al. 2020] train deep networks in a supervised manner. Shi et al. [2017] train a CNN model using their proposed object-centric synthetic dataset. Fu et al. [2021] present a large real-world dataset for highlight removal, and introduce a multi-task network to detect and remove specular reflection. However, the reflection on floors is more complex than highlights, because it may reflect window textures and occupy a large region.

The second class removes reflections from a glass surface in front of the scene. Classical methods use image priors to solve this ill-posed problem, including gradient sparsity [Arvanitopoulos et al. 2017; Levin and Weiss 2007], smoothness priors [Li and Brown 2014; Wan et al. 2016], and ghosting cues [Shih et al. 2015]. Recently, deep learning has been used for this task [Dong et al. 2021; Fan et al. 2017; Hong et al. 2021; Li et al. 2020b; Wan et al. 2018; Wei et al. 2020].

Fig. 3. Object insertion result without appearance decomposition. Note that the direct sunlight on the objects, hard shadows in the sunlight region, and blocked specular reflections are all missing (compare to Fig. 1).

3D model credits: Same as Fig. 1.
We focus on two ubiquitous shading effects that are important for walls. These effects must be located and removed before rendering a panorama. Let $l_{\text{pano}}$ be the original panorama, $F$ be the sampling operator, $f_{\text{tea}}$ be the Teacher Model function, and $f_{\text{situ}}$ be the Student Model function. For example, the locations of windows and other light sources (like lamps) constrain where specular reflections occur. The sunlit areas are constrained by the sun direction and locations of windows. Our key insight is that semantic segmentation provides an easier, discrete supervisory signal for which there are already substantial human annotations and good pre-trained models [Lambert et al. 2020; Wang et al. 2020]. In this section, we describe how to automatically compute semantic panoramic maps and use them to localize these effects coarsely (Fig. 5(b) and (e)). These coarse estimates are used to supervise a GAN-based approach to refine the locations and extract these effects (Fig. 5 (c) and (f)).

3.1 Transfer Learning for Semantic Segmentation

We define seven semantic classes that are of interest to our applications: floor, ceiling, wall, window, door, lamp, and other (see Fig. 5(b) for an example). Most works for semantic segmentation are designed for perspective views. Merging perspective semantics to panorama is non-trivial with problems of view selection, inconsistent overlapping regions, and limited FoV. A few networks [Su and Grauman 2017, 2019] are designed for panoramic images but building and training such networks for our dataset requires annotations and significant engineering effort, which we wish to avoid. Thus, we propose to use a perspective-to-panoramic transfer learning technique, as follows: we obtain an HRNet model [Wang et al. 2020] pre-trained on a perspective image dataset ADE20K [Zhou et al. 2017] (with labels adjusted to our setting) and treat it as the “Teacher Model”. Then we use the same model and weights to initialize a “Student Model”, and adapt it for panoramic image segmentation.

To supervise the Student Model, we sample perspective views from the panorama. Let $l_{\text{pano}}$ be the original panorama, $F$ be the sampling operator, $f_{\text{tea}}$ be the Teacher Model function, and $f_{\text{situ}}$ be the Student Model function. For example, the locations of windows and other light sources (like lamps) constrain where specular reflections occur. The sunlit areas are constrained by the sun direction and locations of windows. Our key insight is that semantic segmentation provides an easier, discrete supervisory signal for which there are already substantial human annotations and good pre-trained models [Lambert et al. 2020; Wang et al. 2020]. In this section, we describe how to automatically compute semantic panoramic maps and use them to localize these effects coarsely (Fig. 5(b) and (e)). These coarse estimates are used to supervise a GAN-based approach to refine the locations and extract these effects (Fig. 5 (c) and (f)).

3 SEMANTICS FOR COARSE LOCALIZATION OF LIGHTING EFFECTS

We focus on two ubiquitous shading effects that are important for virtual staging: (1) specular reflections and (2) direct sunlight. These effects are illustrated in Fig. 4. Dominant specular reflections in empty indoor environments are somewhat sparse and typically due to sources such as lamps, open or transparent windows and doors. The sunlight streaming through windows and doors causes bright and often high-frequency shading on the floor and walls. These effects must be located and removed before rendering

et al. 2019; Zhang et al. 2018b] and achieved significant improvements by carefully designing network architectures. Li et al. [2020b] develop an iterative boost convolutional LSTM with a residual reconstruction loss for single image reflection removal. Also, Hong et al. [2021] propose a two-stage network to explicitly address the content ambiguity for reflection removal in panoramas. However, they mostly use supervised training, requiring large amounts of data with ground truth. Most previous works of both classes do not specifically consider the properties of panoramic floor images.

Our task is in-between these classes, because the reflections are on the floor rather than on glass surfaces but the appearance is similar to that of glass reflection because the floor is flat. To our knowledge, this scenario has not been well studied.

Virtual Staging Services. Some companies provide virtual staging services including Styldod\(^3\), Stuccco\(^4\), and PadStyler\(^5\). Users can upload photos of their rooms and the company will furnish them and provide the rendering results. However, most of them are not free and we are not able to know whether the specular reflections and direct sunlight are handled automatically. Certain applications like changing sun direction are usually not supported. By contrast, our approach is released for free use.

Fig. 4. The geometry of lighting effects considered in this work. (a) For specular reflections, when the camera is upright, the light source point (transparent or open window, door, or indoor lamp), the reflection point, and the camera center lie on a vertical plane (shown as gray), corresponding to a column in the panorama. (b) For direct sunlight, the sun direction establishes the mapping between a window (or door) point and a floor point illuminated by direct sunlight. The sunlit floor area can be back-projected to the window according to the sun direction.

etch et al. 2021] propose a two-stage network to explicitly address the content ambiguity for reflection removal in panoramas. However, they mostly use supervised training, requiring large amounts of data with ground truth. Most previous works of both classes do not specifically consider the properties of panoramic floor images.

Our task is in-between these classes, because the reflections are on the floor rather than on glass surfaces but the appearance is similar to that of glass reflection because the floor is flat. To our knowledge, this scenario has not been well studied.

Virtual Staging Services. Some companies provide virtual staging services including Styldod\(^3\), Stuccco\(^4\), and PadStyler\(^5\). Users can upload photos of their rooms and the company will furnish them and provide the rendering results. However, most of them are not free and we are not able to know whether the specular reflections and direct sunlight are handled automatically. Certain applications like changing sun direction are usually not supported. By contrast, our approach is released for free use.

3 SEMANTICS FOR COARSE LOCALIZATION OF LIGHTING EFFECTS

We focus on two ubiquitous shading effects that are important for virtual staging: (1) specular reflections and (2) direct sunlight. These effects are illustrated in Fig. 4. Dominant specular reflections in empty indoor environments are somewhat sparse and typically due to sources such as lamps, open or transparent windows and doors. The sunlight streaming through windows and doors causes bright and often high-frequency shading on the floor and walls. These effects must be located and removed before rendering

et al. 2019; Zhang et al. 2018b] and achieved significant improvements by carefully designing network architectures. Li et al. [2020b] develop an iterative boost convolutional LSTM with a residual reconstruction loss for single image reflection removal. Also, Hong et al. [2021] propose a two-stage network to explicitly address the content ambiguity for reflection removal in panoramas. However, they mostly use supervised training, requiring large amounts of data with ground truth. Most previous works of both classes do not specifically consider the properties of panoramic floor images.

Our task is in-between these classes, because the reflections are on the floor rather than on glass surfaces but the appearance is similar to that of glass reflection because the floor is flat. To our knowledge, this scenario has not been well studied.

Virtual Staging Services. Some companies provide virtual staging services including Styldod\(^3\), Stuccco\(^4\), and PadStyler\(^5\). Users can upload photos of their rooms and the company will furnish them and provide the rendering results. However, most of them are not free and we are not able to know whether the specular reflections and direct sunlight are handled automatically. Certain applications like changing sun direction are usually not supported. By contrast, our approach is released for free use.

3 SEMANTICS FOR COARSE LOCALIZATION OF LIGHTING EFFECTS

We focus on two ubiquitous shading effects that are important for virtual staging: (1) specular reflections and (2) direct sunlight. These effects are illustrated in Fig. 4. Dominant specular reflections in empty indoor environments are somewhat sparse and typically due to sources such as lamps, open or transparent windows and doors. The sunlight streaming through windows and doors causes bright and often high-frequency shading on the floor and walls. These effects must be located and removed before rendering

etch et al. 2021] propose a two-stage network to explicitly address the content ambiguity for reflection removal in panoramas. However, they mostly use supervised training, requiring large amounts of data with ground truth. Most previous works of both classes do not specifically consider the properties of panoramic floor images.

Our task is in-between these classes, because the reflections are on the floor rather than on glass surfaces but the appearance is similar to that of glass reflection because the floor is flat. To our knowledge, this scenario has not been well studied.

Virtual Staging Services. Some companies provide virtual staging services including Styldod\(^3\), Stuccco\(^4\), and PadStyler\(^5\). Users can upload photos of their rooms and the company will furnish them and provide the rendering results. However, most of them are not free and we are not able to know whether the specular reflections and direct sunlight are handled automatically. Certain applications like changing sun direction are usually not supported. By contrast, our approach is released for free use.

3 SEMANTICS FOR COARSE LOCALIZATION OF LIGHTING EFFECTS

We focus on two ubiquitous shading effects that are important for virtual staging: (1) specular reflections and (2) direct sunlight. These effects are illustrated in Fig. 4. Dominant specular reflections in empty indoor environments are somewhat sparse and typically due to sources such as lamps, open or transparent windows and doors. The sunlight streaming through windows and doors causes bright and often high-frequency shading on the floor and walls. These effects must be located and removed before rendering
be the Student Model function. The transfer learning loss $L_{\text{trans}}$ is defined as the cross entropy loss between $f_{\text{te}}(\Phi(I_{\text{pano}}))$ and $\Phi(f_{\text{stu}}(I_{\text{pano}}))$. To regularize the training, we prevent the Student Model from deviating from the Teacher Model too much by adding a term $L_{\text{reg}}$ defined as the cross entropy loss between $f_{\text{te}}(I_{\text{pano}})$ and $f_{\text{stu}}(I_{\text{pano}})$. The total loss is $L = w_{\text{trans}}L_{\text{trans}} + w_{\text{reg}}L_{\text{reg}}$. $w_{\text{trans}}$ and $w_{\text{reg}}$ are weights given by the confidence of the Teacher Model prediction $f_{\text{te}}(\Phi(I_{\text{pano}}))$ and $f_{\text{te}}(I_{\text{pano}})$.

The semantic map is further improved via HDR map and layout estimation. See Sec. 6.1 for details. Fig. 5(b) shows an example semantic map computed using our approach.

3.2 Coarse Localization of Lighting Effects

Coarse Specular Mask Generation. The panoramic geometry dictates that the possible specular area, on the floor, is in the same columns as the light sources. This assumes that the camera is upright and that the floor is a mirror. See Fig. 4 for an illustration. Thus, we simply treat floor pixels, that are in the same columns as the light sources (windows and lamps in the semantic map), as the coarse specular mask. To handle rough floors, we dilate the mask by a few pixels. See Fig. 5(c) for an example coarse specular mask.

Coarse Sunlight Mask Generation. We first estimate a coarse sun direction and then project the window mask to the floor and walls according to this direction. Since sunlight enters the room through windows (or open doors), the floor texture can be projected back onto the window region based on the sun direction. This projected floor area must be bright and a score is estimated as the intensity difference between window region and projected floor. The projection requires room geometry which can be obtained by calculating wall-floor boundary from the semantic map. Basically, we assume an upright camera at height=1 and the floor plane at Z=-1. Then, the 3D location (XYZ) of a wall boundary point is computed using

$$Z = \frac{1}{\cos(\theta, \phi)} \sqrt{X^2 + Y^2 + Z^2}$$

where $\theta$ is the elevation angle and $\phi$ is the azimuth angle. We do a brute-force search for $\theta$ and $\phi$ with step size $= 5^\circ$. Let $s(\theta, \phi)$ be the matching score. First, top-K elevation angles $\theta_1, ..., \theta_K$ are selected based on score $s'(\theta) = \max_{\phi} s(\theta, \phi)$. Let $\phi^*(\theta) = \arg \max_{\phi} s(\theta, \phi)$ be the optimal azimuth angle for elevation $\theta$. Then the selected sun directions are $(\theta_1, \phi^*(\theta_1)), ..., (\theta_K, \phi^*(\theta_K))$. In practice, we set $K = 5$. Fig. 5(e) for an example coarse direct sunlight mask.

4 GAN-BASED LIGHTING EFFECTS DETECTION

Both the coarse masks estimated above are rough conservative estimates. In this section, we will demonstrate that this inaccurate information can still be used as a supervision signal that can effectively constrain and guide an accurate estimation of specular reflection and sunlight regions. We sketch the key idea in Sec. 4.1, and explain the detailed loss function in Sec. 4.2.

4.1 Key Idea and Overall Architecture

Simply training a lighting effects network with the coarse masks will result in the network predicting only the coarse masks. How do we design a network that is able to produce more accurate masks? Fig. 6 illustrates our architecture that augments the lighting effects network with three local discriminators. The lighting effects network takes a single panoramic image $I$ as an input, and predicts the specular ($I_{\text{spec}}$) and sunlight ($I_{\text{sunl}}$) components. If the prediction is good, a local discriminator trying to locate specular regions should fail on the image with specular component removed ($I - I_{\text{spec}}$). This is the key supervision signal for training the lighting effects network. The local discriminator is supervised by coarse specular masks obtained from semantics. Similar techniques are applied to sunlight regions and regions with overlapping specular and direct sunlight.
should have different levels of sparsity. Lighting effects should be sparser outside the ROI than inside the ROI. They should also be sparser outside the coarse mask than inside the coarse mask. Thus, we apply different levels of sparsity to different regions. Let $I$ be the input image, $R$ be the binary ROI mask, and $M$ be the coarse mask inside the ROI. The sparsity loss is:

$$L_s = \lambda_1 || R \odot I || + \lambda_2 || R \odot (1 - M) || + \lambda_3 || M ||,$$  \hspace{1cm} (1)

where $\lambda_1 \geq \lambda_2 \geq \lambda_3$ are constant weights, and $\odot$ is the element-wise product. We use such sparsity loss for specular reflection and direct sunlight separately, denoted by $L_{sp,a}$ and $L_{sl,a}$.

**Adversarial Loss.** As mentioned in Sec. 4.1, after removing the lighting effects, a discriminator should not be able to locate the effects. Consider specular reflection as an example. Let $I$ be the input image, and $I_{sp}$ be the estimated specular reflection. We use a local discriminator to find reflection areas from $I - I_{spec}$. The discriminator is supervised by coarse specular mask $M_{spec}$. The discriminator loss uses pixel-wise binary cross entropy, $L_{sp,d} = fce(D_{spec}, M_{spec})$. To fool the discriminator, an adversarial loss:

$$L_{sp,a} = fce(D_{spec}, 1 - M_{spec})$$  \hspace{1cm} (2)

is applied to region $M_{spec}$. Similarly, we build adversarial losses for sunlight regions and regions with overlapping specular reflection and direct sunlight, denoted by $L_{sl,a}$ and $L_{ol,a}$, respectively.

**Total Loss.** The total loss $L$ is the sum of the two sparsity losses and the three adversarial losses:

$$L = \lambda_{sp}L_{sp,s} + \lambda_{sl}L_{sl,s} + \lambda_{sp}L_{sp,a} + \lambda_{sl}L_{sl,a} + \lambda_{ol}L_{ol,a},$$ \hspace{1cm} (3)

where $\lambda_{sp}, \lambda_{sl}, \lambda_{sp}, \lambda_{sl}, \lambda_{ol}$ are constant weights.

### 5 LIGHTING EFFECTS REMOVAL

The previous architecture and method focused on estimating the specular and direct sunlight regions. But, naively subtracting the predicted regions from the original image, i.e. $I - I_{spec}, I - I_{spec} - I_{sun}$, may still produce artifacts as shown in Fig. 7. In this section, we instead directly estimate the diffuse image (without specular reflection) and the ambient image (without specular reflection or direct sunlight).

#### 5.1 GAN-Based Specular Reflection Removal

As shown in Fig. 8, a deep network is used to predict a diffuse image. To supervise the training of the network, we adopt the same GAN strategy as Sec. 4.2. Instead of using the coarse mask generated by Sec. 3.2, we obtain a fine specular mask by thresholding the estimated specular reflection for discriminator supervision.

Let $I$ be the original image, $I_{spec}$ be the specular image, $I_{diff}$ be the diffuse image, and $I_{rec} = I_{diff} + I_{spec}$ be the reconstructed image. Inspired by [Zhang et al. 2018b], our loss $L_{diff}$ consists of an adversarial loss $L_{adv}$, a reconstruction loss $L_{recon}$, a perceptual loss $L_{perc}$, and an exclusion loss $L_{excl}$:

$$L_{diff} = \lambda_a L_{adv} + \lambda_r L_{recon} + \lambda_p L_{perc} + \lambda_e L_{excl},$$  \hspace{1cm} (4)

where $\lambda_a, \lambda_r, \lambda_p, \lambda_e$ are constant weights.

**Adversarial Loss.** $L_{adv}$ is the same as the loss defined in Sec. 4.2, replacing the coarse specular mask by the fine specular mask.

**Reconstruction Loss.** $L_{recon} = ||I_{recon} - I||_1$ calculates the L1 loss between the reconstructed image and the original image.

**Perceptual Loss.** Following [Johnson et al. 2016] and [Zhang et al. 2018b], $L_{perc} = ||\Phi(I_{recon}) - \Phi(I)||_1$ calculates the L1 distance between VGG features [Simonyan and Zisserman 2015], where $\Phi$ is the VGG feature function. It is for enhancing perceptual quality.

**Exclusion Loss.** $L_{excl} = ||\tanh(\nabla I_{diff}) \odot \tanh(\nabla I_{spec})||_F$, where $||.||_F$ is the Frobenius norm, and $\odot$ is the element-wise product. This loss minimizes the correlation between diffuse and specular components in the gradient domain to prevent edges from appearing in both diffuse and specular images, following [Zhang et al. 2018b].
5.2 Semantic-Aware Inpainting for Sunlight Removal

A process similar to that used for diffuse image estimation can be used for predicting the ambient image without sunlight. In practice, however, bright sunlit regions are the highly dominant brightness component and are often saturated and the network does not predict the ambient image well. Thus, we use an inpainting approach called SESAME [Ntavelis et al. 2020] that is semantic-aware and preserved the boundaries between different classes (wall versus floor, etc).

6 EXPERIMENTAL ANALYSIS

6.1 Implementation Details

Improving Lamp Semantics with HDR Estimation. The ADE20K definition of lamps includes the whole lamp, while we care about the bright bulb. Thus, we estimate a HDR map via supervised learning to obtain a more accurate bulb segmentation. To obtain a HDR image $I_{hdr}$, we train U-Net [Ronneberger et al. 2015] to predict the HDR-LDR residue in log space $I_{res} = \ln(1 + I_{hdr} - I_{ldr})$, using Laval [Gardner et al. 2017] and HDRI Heaven 6 datasets. A weighted L2 loss is adopted. We set batch size 32, learning rate $10^{-4}$ and epochs 300. Pixels in the lamp segmentation with $I_{hdr} > 2$ are kept.

Improving Room Semantics with Layout Estimation. The floor-wall and ceiling-wall boundaries predicted by semantic segmentation might not be straight in perspective views, causing visual artifacts. Thus, we adopt a layout estimation method LED2-Net [Wang et al. 2021c] to augment the semantic map. The layout estimation predicts floor, wall and ceiling regions for the current room, which is used to improve the floor-wall and ceiling-wall boundaries. We adopt batch size 6, learning rate $3 \times 10^{-4}$, and epochs 14.

Super-Resolution in Post-Processing. Due to limited computing, we run our methods at low-resolution ($256 \times 512$). We enhance the resolution for applications requiring high resolution results with post-processing. We enhance the resolution of the specular image to $1024 \times 2048$ via a pre-trained super-resolution model [Wang et al. 2021b]. To ensure the consistency between the diffuse image and the original image, we adopt Deep Guided Decoder [Uezato et al. 2020] for super-resolution with the original image as guidance. The final resolution is $1024 \times 2048$. For the ambient image, we simply use super-resolution [Wang et al. 2021b] and copy-paste the inpainted region on to the high resolution diffuse image. To preserve high-frequency details, we calculate the difference between the diffuse and ambient components as the high-resolution sunlight image.

Deep Architecture and Hyper-parameters. For semantic transfer learning, we use $w_{trans} = w_{reg} = 1$, batch size 4, learning rate $10^{-3}$, epochs 5, and random perspective crops with FoV 80° to 120°, elevation -60° to 60°, azimuth 0° to 360°. We use U-Net [Ronneberger et al. 2015] for Lighting Effects Network and Diffuse Network, and FCN [Long et al. 2015] for discriminators, optimized with Adam [Kingma and Ba 2015]. See supplementary materials for network architecture. The networks use learning rate $10^{-4}$, weight decay $10^{-5}$, and batch size 16. For effects detection, $\lambda_{sp,5} = \lambda_{al,5} = \lambda_{sp,a} = \lambda_{al,a} = \lambda_{al,g} = 1$. For specular sparsity, $\lambda_{s1} = 10, \lambda_{s2} = 3, \lambda_{s3} = 1$. For sunlight sparsity, $\lambda_{s1} = 5, \lambda_{s2} = \lambda_{s3} = 1$. For specular reflection removal, $\lambda_{r} = \lambda_{p} = \lambda_{e} = 1, \lambda_{a} = 10$. The networks take ~2 days for training on 4 TITAN Xp GPUs and 0.05s for inference. Resolution enhancement using Guided Deep Decoder [Uezato et al. 2020] converges in ~30 min for a single image.

Dataset. We evaluate our approach on the ZInD [Cruz et al. 2021] dataset that includes 54,034 panoramas for training and 13,414 panoramas for testing (we merge their validation and test sets). For better visualization, we inpaint the tripod using SESAME [Ntavelis et al. 2020]. Sec. 6.2 and 6.3 evaluate our approach at $256 \times 512$ resolution. Sec. 6.4 visualizes results at resolution $1024 \times 2048$.

6.2 Ablation Study

Perspective Merging vs. Panoramic Semantics. We compare with merging perspective semantics to panorama, by sampling 14 perspective views and merging the semantics via per-pixel voting. On Structured3D [Zheng et al. 2020] test set (mapped to our labels), the merging method achieves mIoU=44.8%, while ours achieves 68.1%. This shows developing methods directly is meaningful. We also test our method on different emptiness levels: No furniture mIoU: 68.1%; Some furniture: 66.6%; Full furniture: 60.6%. Since ZInD does not contain much furniture, this mIoU reduction is expected.

Lighting Effects Detection. We manually annotate specular and sunlight regions on 1,000 test panoramas sampled from the test set for quantitative analysis. The annotation is coarse because there is no clear boundary of specular regions. For a fair comparison, we report the mean IoU with respect to the best threshold for binarization for each method. The quantitative evaluation is performed on the floor region only.

To show that the GAN-based method introduces priors for refining the coarse mask, we train a network with the same architecture as Lighting Effects Network in a standard supervised manner (without GAN), using the coarse mask as labels. We also compare with the coarse mask itself and the method without the third discriminator for overlap regions. Tab. 1 shows that the full method outperforms the others, demonstrating the effectiveness of our GAN-based approach with three discriminators.

Specular Reflection Removal. We use synthetic data for evaluating specular reflection removal. We render 1,000 images based on micro-facet reflectance model using the Mitsuba [Jakob 2010] renderer with floor textures from AdobeStock. We also define images with peak specular intensity ranking top 5% in the test data as “strong reflection”. PSNR (higher is better) and LPIPS [Zhang et al. 2018a] (lower is better) are reported.

We compare with the naive method by subtracting the specular component from the original image directly, and methods without each loss components. In Tab. 2, the full method outperforms the other variants on all metrics except for the ones without adversarial loss. Although the methods without adversarial loss achieves a high score on "all testdata", it performs worse than the full method on "strong reflection". Besides, as shown in Fig. 9, adversarial and exclusion losses help remove perceivable visual artifacts or specular residues although average metrics don’t reflect clearly. Thus, we adopt the full method with all four losses for the applications.
Table 1. Ablation study for lighting effects detection on real data. mIoU (%) for specular and sunlight regions are reported. Our full method performs better than the others, showing the effectiveness of the GAN-based design.

| Method                              | Specular ↑ | Sunlight ↑ |
|-------------------------------------|------------|------------|
| Coarse mask                         | 6.7        | 3.0        |
| Supervised by coarse mask w/o GAN  | 8.7        | 19.9       |
| No overlap effects discriminator    | 34.6       | 30.0       |
| Our full method                     | **38.9**   | **47.2**   |

Table 2. Ablation study for specular removal on synthetic panoramas. Note that specular reflections are not strongly visible in all images of the dataset and they are sparse when visible. Hence, we also report the metrics on a subset of images with strong reflections. Although “no adversarial loss” performs better on “all testdata”, it does not reach the level of the full method on strong reflections, which significantly affect the visual quality.

| Method                                      | All Testdata | Strong Reflection |
|---------------------------------------------|--------------|-------------------|
| PSNR ↑ | LPIPS ↓ | PSNR ↑ | LPIPS ↓ |
|---------------------------------------------|--------------|-------------------|
| Input subtracts specular image              | 33.2         | 0.0247            | 25.0 | 0.0702 |
| No adversarial loss                         | 33.4         | 0.0233            | 25.2 | 0.0681 |
| No reconstruction loss                      | 33.2         | 0.0248            | 25.1 | 0.0693 |
| No perceptual loss                          | 31.3         | 0.0483            | 25.0 | 0.0903 |
| No exclusion loss                           | 33.3         | 0.0242            | 25.2 | 0.0693 |
| No adversarial/reconstruction               | 33.4         | 0.0252            | 25.1 | 0.0683 |
| No adversarial/perceptual                   | 33.4         | 0.0270            | 25.3 | 0.0749 |
| No adversarial/exclusion                    | 33.3         | 0.0233            | 25.1 | 0.0684 |
| No reconstruction/perceptual                | 33.4         | 0.0252            | 25.1 | 0.0683 |
| No reconstruction/exclusion                 | 33.2         | 0.0242            | 25.2 | 0.0700 |
| No perceptual/exclusion                     | 32.4         | 0.0333            | 25.1 | 0.0800 |
| Our full method                             | **33.3**     | **0.0241**        | **25.3** | **0.0676** |

6.3 Performances of Other Relevant Approaches

Most previous approaches for diffuse-specular separation and inverse rendering were not designed for the setting used in this work to enable direct apples-to-apples comparisons; they require perspective views or supervised (ground truth) training of large scale real or rendered data or are used for a related task. But accurately annotating large data for specular and sunlight effects is non-trivial, because of blurred boundaries or complex texture. The visual and quantitative evaluations below show that previous approaches do not easily bridge the domain gap. These methods have strong value for the domains they were designed for but our approach provides an effective tool for virtual staging with real-world panoramas.

**Specular Reflection Detection.** We evaluate a bilateral filtering based method BF [Yang et al. 2010], two deep reflection removal methods IBCLN [Li et al. 2020b] and LASIRR [Dong et al. 2021], and a deep highlight removal method JSHDR [Fu et al. 2021]. Since these methods are designed for perspective images, we randomly sample 5,000 perspective crops from the 1,000 annotated panoramic images and report the performance in Tab. 3 left column. In the right column, we also evaluate in panoramic domain by converting the panorama to a cubemap, running the baseline methods, and merging the results back to panorama. Among the evaluated relevant methods, JSHDR [Fu et al. 2021] achieves the best performance, likely because it is designed for specular highlight removal, which is closer to our setting than glass reflection removal methods.

Fig. 9. Ablation study for specular reflection removal by not including adversarial, reconstruction, perceptual or exclusion losses. While quantitative metrics are averaged over the image and over the dataset, the visuals illustrate the subtle but noticeable residual specular reflections when the different losses are removed from our method.

Fig. 10. Qualitative results of specular reflection detection and removal. Specular reflections are brightened for visualization. Note that although the walls are shown, only the floor region is evaluated quantitatively. Our method successfully detects and removes the both window and lamp reflections. (d) removes lamp reflections but (b) and (c) do not. (b) (c) (d) leaves residues of window reflections on the floor.
Fig. 11. High resolution appearance decomposition results. The lighting effects are successfully detected on a variety of floor types. For specular reflection removal, the floor texture in (c) is consistent with that in (a). Sunlight regions are inpainted with correct colors. Some floor textures are imperfect due to the limitation of the inpainting algorithm, e.g., the third row, where the camera tripod is inpainted in the original image.

Table 3. Quantitative analysis for specular reflection detection on real data. mIoU (%) for specular regions are reported. “Ours” is inferred in panoramic domain, while the others are inferred in perspective domain. All methods are evaluated in both domains (cross domain via cropping or merging). JSHDR [2021] performs better than the other relevant methods, possibly because of a smaller domain gap.

| Method            | Perspective mIoU | Panoramic mIoU |
|-------------------|------------------|----------------|
| BF [Yang et al. 2010] | 11.3             | 8.9            |
| IBCLN [Li et al. 2020b] | 15.7             | 15.0           |
| LASIRR [Dong et al. 2021] | 20.5             | 17.5           |
| JSHDR [Fu et al. 2021] | 21.4             | 14.1           |
| Ours              | 41.8             | 38.9           |

Direct Sunlight Detection. We are unaware of any work estimating direct sunlight for our setting. Thus, we evaluate an intrinsic image decomposition method USI3D [Liu et al. 2020]. We assume that the sunlight region should have a strong shading intensity and threshold the output shading image as the sunlight prediction. This achieves mean IoU 7.4% on perspective crops and 7.9% on panoramas from cubemap merging while our method achieves 47.2% on

Table 4. Quantitative analysis for specular reflection removal on synthetic data. “Ours” and “IBCLN (re-train)” are inferred in panoramic domain, while the other methods are inferred in perspective domain. All methods are evaluated in both domains (cross domain via cropping or merging). To bridge the domain gap, we re-train IBCLN [2020b] on our data. Although performance is improved, it does not reach the level of our method.

| Evaluation Domain | Method            | All Testdata | Strong Reflection |
|-------------------|-------------------|--------------|--------------------|
|                    |                   | PSNR ↑ | LPIPS ↓ | PSNR ↑ | LPIPS ↓ |
| Perspective        | BF [2010]         | 23.4   | 0.1632 | 17.7   | 0.2326 |
|                    | IBCLN [2020b]     | 27.5   | 0.1013 | 23.7   | 0.1460 |
|                    | LASIRR [2021]     | 20.8   | 0.1984 | 17.5   | 0.2515 |
|                    | JSHDR [2021]      | 21.4   | 0.0354 | 24.1   | 0.1158 |
|                    | IBCLN [2020b] (re-train) | 33.4 | 0.0279 | 25.3   | 0.1016 |
|                    | Ours              | 34.5   | 0.0271 | 27.0   | 0.0935 |
| Panoramic          | BF [2010]         | 22.1   | 0.1526 | 16.5   | 0.1934 |
|                    | IBCLN [2020b]     | 28.8   | 0.0890 | 24.2   | 0.1418 |
|                    | LASIRR [2021]     | 23.0   | 0.1425 | 20.1   | 0.1788 |
|                    | JSHDR [2021]      | 30.6   | 0.0698 | 22.5   | 0.1499 |
|                    | IBCLN [2020b] (re-train) | 32.3 | 0.0250 | 24.7   | 0.0743 |
|                    | Ours              | 33.3   | 0.0241 | 25.3   | 0.0676 |
Fig. 12. Diverse lighting conditions of testing images. Row 1: large sunlight region; Rows 2-3: effects from (partially) occluded light sources (see red boxes in (a)). Our method successfully detects the lighting effects.

perspective crops and 47.2% on panoramas. We conclude that using the semantics and sun direction is important for sunlight estimation.

Specular Reflection Removal. Similar to lighting effect detection, we report performances based on 5,000 perspective crops sampled from the 1,000 synthetic panoramic images. Fig. 10 and the first four rows in Tab. 4 show the performances of pre-trained models on perspective images. Similar to specular reflection detection, JSHDR [Fu et al. 2021] achieves the best performance, likely due to smaller domain differences, since it is trained on a large-scale real dataset. A similar conclusion can be drawn from the evaluation in panoramic domain. We also attempted to bridge the domain gap for a more reasonable quantitative evaluation. For JSHDR, only the compiled code is available online, making it hard to adapt to our domain. Thus, we train IBCLN on our panoramic data. Specifically, let \{I\} be the set of original images and \{I_{spec}\} be the set of estimated specular components. We treat \{I - I_{spec}\} and \{I_{spec}\} as the transmission image set and the reflection image set, respectively. These two sets are then used for rendering the synthetic images for training. Tab. 4 Rows 5 and 11 show that by re-training using the synthetic data based on our specular reflection estimation, the performance of IBCLN improves but still does not reach the level of our method.

6.4 Diverse Lighting and High Resolution Results
Fig. 12 visualizes effects detection results in diverse lighting conditions, including large sunlight area and (partially) occluded light sources. The results show that our detection can handle such cases.

Fig. 11 visualizes high-resolution appearance decomposition results. The specular reflections are well removed and the floor texture is consistent with the original image. The sunlight region is inpainted with the correct colors. Due to the limitation of the inpainting algorithm, the texture is not perfect. As inpainting algorithms improve, we can plug-and-play those within our pipeline.

7 APPLICATIONS

Improvement of Albedo Estimation. Albedo estimation via intrinsic image decomposition suffers from the existence of specular reflection and direct sunlight. As shown in Fig. 13 (b), a pre-trained USI3D [Liu et al. 2020] incorrectly bakes specular reflection and direct sunlight into the albedo component. When we run the same algorithm on the ambient image after removing the lighting effects, the albedo estimation (c) is significantly better.

Improvement of Sun Direction Estimation. With the detected direct sunlight, we can improve the sun direction estimation proposed in Sec. 3.2. There are two modifications: (1) Instead of using the RGB image, we use the direct sunlight image for matching; (2) The wall sunlight is also considered. Previously we do not consider walls for coarse sun direction estimation because walls are usually white, which may mislead the matching score.

To evaluate the sun direction, we obtained sun elevation angles for part of the dataset from the authors of ZInD [Cruz et al. 2021]. The elevation angles are calculated based on timestamps and geolocations. The ground truth azimuth angles are not available because of unknown camera orientations. The elevation angles provided may not be always accurate because the timestamps are based on the time the data were uploaded to the cloud. Thus, we manually filter obviously incorrect values and select images with visible sunlight for evaluation. A total of 257 images are evaluated.

We search for the top-1 direction at step size 1°. Fig. 14 plots the histogram of the angular estimation error. More than half of the estimations are within 10° error. Fig. 14 also visualizes an example of projecting the window mask according to the estimated sun
Fig. 14. Sun direction estimation results. Left: histogram of sun elevation errors for images with visible sunlight; Right: An example of projecting the window mask to the floor and the walls according to estimated sun direction. The fine sun estimations provide more accurate elevation results. 33% of the fine estimations and of the 28% coarse estimations are within 5° error. 53% of the fine estimations and of the 52% coarse estimations are within 10° error. The example shows that the fine estimation produces a reasonable sun direction even with overlapping sunlight and specular reflection. This is an extreme example where coarse estimation completely fails and the fine estimation significantly improves the result.

Fig. 15. Changing floor material. Row 1: original flooring; Row 2: change to (different) wood; Row 3: change to (different) carpet. The specular reflection on row 2 column 2 is rendered using the estimated HDR map (Sec. 6.1) and a uniform surface roughness.

Changing Floor Material. The decomposition is used to change flooring with texture and BRDF parameters. Examples include switching between wood (specular), carpet (diffuse) or tile (specular). Fig. 15 shows four examples, including wood-to-wood (keep specularity), wood-to-carpet (remove specularity), carpet-to-wood (render specularity via HDR map), and carpet-to-carpet changes. Direct sunlight are rendered on the new material by scaling the floor intensity according to the estimated sunlight scale.

Changing Sun Direction. Using the estimated sun direction, we project the sunlit region back to the window position. Then, we re-project sunlight onto the floor and the walls using a new sun direction. Fig. 16 visualizes two examples of changing sun direction.

Furniture Insertion. With the decomposition result, sun direction and the estimated HDR map (Sec. 6.1), we can insert virtual objects into an empty room, by calculating the occluded specular reflection and direct sunlight via ray tracing. Using the computed layout (scale set manually), we render ambient, sunlight and specular effects separately and combine them. Fig. 17 shows several insertion examples, where sunlight is cast on the desk (a) and the bed (b&c), specular reflection is blocked by the chair (a) and bed-side table (c), and sunlight is blocked by the clothing hanger (b) and bed (c).

Fig. 18 shows scenes rendered with Li et al. [Li et al. 2020a], rendered with the estimated HDR map without appearance decomposition, and rendered with our full method. Li et al. [Li et al. 2020a] trains on synthetic data with ground truth supervision and predicts illumination and scene properties for a perspective view. The method estimates spatially-varying lighting for each pixel rather than each 3D voxel. A selected pixel and its lighting is used for rendering. This fails when the object parts are far away from the selected pixel in 3D (e.g., first row in Fig. 18). Besides, the method renders shadows in the specular region by scaling the original floor intensity without explicitly removing blocked specular reflection, leading to insufficient darkening of the floor (e.g., second row).

Combination of Multiple Applications. The applications can be combined to allow more flexible virtual staging effects. Fig. 17 shows
Fig. 17. Appearance decomposition enables several applications such as furniture insertion, and changing flooring and sun direction, making it a useful tool for virtual staging. Top row: a home office and a bedroom staged by inserting chair, table, bed, shelves, plant and other objects. Rows 2, 3, 4: a living room is staged with different furniture, multiple new flooring (dark wood, carpet), and is shown under two different sun directions. Notice the sun light on the surfaces and the blocking of specular reflection that add realism to the scene.

high-quality visualizations that include two or more of changing floor material, sun direction and object insertion.

Credits: (a) landscape painting © adobestock3d/Adobe Stock, ship painting © Rijksmuseum/Adobe Stock, desk © Francesco Milanese/Adobe Stock, computer © Kamil/Adobe Stock, shelf © adobestock3d/Adobe Stock, chair © Kamil/Adobe Stock, bookcase © Cassie Crepa/Adobe Stock, lamp © Maxrus/Adobe Stock, plant © Kids Creative Agency/Adobe Stock; (b) coat rack © Alexandre/Adobe Stock, mirror © Francesco Milanese/Adobe Stock, dresser © Gregory Allen Brown/Adobe Stock, notebook © HQ3DMOD/Adobe Stock, clock © Francesco Milanese/Adobe Stock, bedside table © Jacob Rose/Adobe Stock, bed © shabee/Adobe Stock. (c)-(h): See Fig. 2 and Fig. 15.
8 CONCLUSION

In summary, we present an appearance decomposition method for empty indoor panoramic scenes. Relying on supervision from semantics, our method detects and removes the specular reflection on the floor and the direct sunlight on the floor and the walls. The decomposition result can be applied to multiple virtual staging tasks, including albedo and sun direction estimation, furniture insertion, changing sun direction and floor material.

There are three main limitations. First, we focus on direct lighting effects (first light bounce, no global illumination) on the floor and the walls. However, the specular reflection and direct sunlight could also appear on objects like kitchen countertops or appliances. Fig. 19 shows that our method partially detects the sunlight but the sun direction and object lighting is erroneous since the surface height is unknown. This could potentially be solved by extending our semantics and geometry to include furniture or other objects as long as some coarse mask and geometry can be obtained automatically to allow for accurate ray-tracing. Second, we assume the window producing the sunlight is visible for sun direction estimation. Using multi-view information from different panoramas in a single home can help overcome this limitation. Third, our approach is tailored to panoramas. It is possible to extend our ideas to narrower perspective field of views but there are several alternate methods in that space and we lose the advantages provided by a panorama. Given the increased availability of panoramas of homes, we believe the method is timely and can be an effective tool for virtual staging.

ACKNOWLEDGMENTS

This work was supported by a gift from Zillow Group, USA, and NSF Grants #CNS-2038612, #IIS-1900821.

REFERENCES

Miika Aittala, Timo Weyrich, and Jaakko Lehtinen. 2015. Two-shot SVBRDF Capture for Stationary Materials. ACM Trans. Graph. (Proc. SIGGRAPH) 34, 4, Article 110 (July 2015), 13 pages. https://doi.org/10.1145/2766967

Yasushi Akashi and Takayuki Okatani. 2016. Separation of reflection components by sparse non-negative matrix factorization. Computer Vision and Image Understanding 100, 146 (2016), 77–85.

Nikolaos Arvanitopoulos, Radhakrishna Achanta, and Sabine Susstrunk. 2017. Single image reflection suppression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4498–4506.

Jonathan Barron and Jitendra Malik. 2015. Shape, Illumination, and Reflectance from Shading. IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (08 2015), 1670–1687. https://doi.org/10.1109/TPAMI.2014.2377712

And S Baslamisli, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. 2021. ShadingNet: image intrinsics by fine-grained shading decomposition. International Journal of Computer Vision (2021), 1–29.

Sean Bell, Kavita Bala, and Noah Snavely. 2014. Intrinsic images in the wild. ACM Transactions on Graphics (TOG) 33, 4 (2014), 1–12.

Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. 2021a. Nerf: Neural reflectance decomposition from image collections. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 12684–12694.

Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik Lensch. 2021b. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. Advances in Neural Information Processing Systems 34 (2021).

Mark Boss, Varun Jampani, Kihwan Kim, Hendrik Lensch, and Jan Kautz. 2020. Two-shot spatially-varying brdf and shape estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3982–3991.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 4 (2017), 834–848.

93D model credits: Row 1: See Fig. 17 (b), Row 2: armchair©Dmitrii Ispolatov/Adobe Stock, Row 3: trash can©adobestock3d/Adobe Stock, Row 4: table©Dmitrii Ispolatov/Adobe Stock.

103D model credit: mug©Brandon Westlake/Adobe Stock.
Giavaras Pintore, Marco Agus, and Enrico Gobbetti. 2020. AtlantaNet: Inferring the 3D Indoor Layout from a Single 360 Image Beyond the Manhattan World Assumption. In European Conference on Computer Vision. Springer, 432–448.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention. Springer, 234–241.

S. Sengupta, Jinwei Gu, Kihwan Kim, Gulin Lin, D. Jacobs, and J. Kautz. 2019. Neural Inverse Rendering of an Indoor Scene From a Single Image. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019), 8597–8606.

S. Sengupta, A. Kanazawa, Carlos D Castillo, and D. Jacobs. 2018. SfSNet: Learning Shape, Reflectance and Illuminance of Faces in the Wild. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6296–6305.

Steven A. Shafer. 1985. Using color to separate reflection components. Color Research & Application 10, 4 (1985), 210–218.

Hui-Liang Sheng and Zhi-Huan Huang. 2013. Real-time highlight removal using intensity ratio. Applied optics 52, 19 (2013), 4483–4493.

Jian Shi, Yue Dong, Hao Su, and Stella X. Yu. 2017. Learning non-lambertian object intrinsics across shapenet categories. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3193–3201.

Z. Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, E. Shechtman, and D. Samaras. 2017. Neural Face Editing with Intrinsic Image Disentangling. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017), 5444–5453.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In ICLR.

Gowri Somanath and Daniel Kurz. 2020. HDR Environment Map Estimation for Real-time Augmented Reality. arXiv preprint arXiv:2011.10687 (2020).

Pratul P. Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T. Barron, Richard Tucker, and Noah Snavely. 2020. Lighthouse: Predicting Lighting Volumes for Spatially-Coherent Illumination. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8089–8098.

Yu-Chuan Su and Kristen Grauman. 2017. Learning spherical convolution for fast spherical spherical convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9442–9451.

Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Trong Chen. 2019. HorizonNet: Learning room layout with 1d representation and pano stretch data augmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1047–1056.

Tristan Swedish, Connor Henley, and Ramesh Raskar. 2021. Objects As Cameras: Estimating High-Frequency Illumination From Shadows. In Proceedings of the IEEE International Conference on Computer Vision. 2593–2602.

Robby T Tan and Katsushi Ikeuchi. 2005. Separating Reflection Components of Textured Surfaces Using a Single Image. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 2 (2005), 178–193.

Marshall Tappen, William Freeman, and Edward Adelson. 2003. Recovering Intrinsic Images from a Single Image. In Advances in Neural Information Processing Systems. S. Becker, S. Thrun, and K. Obermayer (Eds.), Vol. 15. MIT Press. https://proceedings.neurips.cc/paper/2002/file/f6eaaed9e05b5e92e56f75b637e26d6-Paper.pdf

Tatsuki Ueato, Danfeng Hong, Naoto Yokoya, and Wei He. 2020. Guided deep deconvolutional networks for large-scale image recognition. In ICLR.

Richard Zhang, Philip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018a. Unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 586–595.

Xuanyuan Zhang, Ren Ng, and Qifeng Chen. 2018b. Single image scene separation with perceptual loss. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4786–4794.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. 2020. Structured3d: A large photo-realistic dataset for structured 3d modeling. In European Conference on Computer Vision. Springer, 519–535.

Bolei Zhou, Hang Zhao, Xavier Puig,Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In Proceedings of the IEEE conference on computer vision and pattern recognition. 633–641.

Chuhang Zou, Alex Colburn, Qi Shans, and Derek Hoiem. 2018. Layoutnet: Recovering spatial layout from a single egp image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2051–2059.