DOI: 10.1111/cgf.13943
EUROGRAPHICS 2020 / U. Assarsson and D. Panozzo
(Guest Editors)

Volume 39 (2020), Number 2

RGB2AO: Ambient Occlusion Generation from RGB Images

N. Inoue[1], D. Ito[2], Y. Hold-Geoffroy[2], L. Mai[2], B. Price[2] and T. Yamasaki[1]

[1]The University of Tokyo, Japan [2]Adobe Research, U.S.

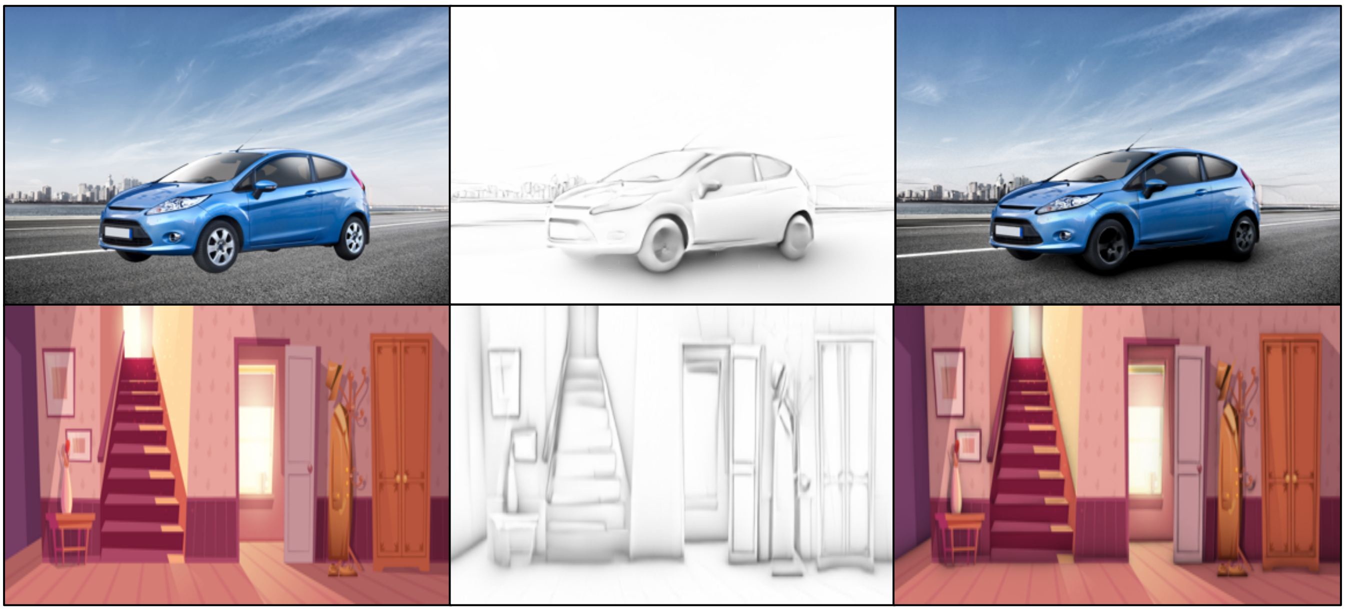

Figure 1: *Ambient Occlusion (AO) generation from a single image with our RGB2AO framework. Given an RGB image (left), our model automatically generates an ambient occlusion map (center). We show two applications of applying the generated AO map effect: image composition (in the first row) and geometry-aware contrast enhancement (in the second row) on the right. The input images are from Adobe Stock. (best viewed with zoom.)*

Abstract

We present RGB2AO, a novel task to generate ambient occlusion (AO) from a single RGB image instead of screen space buffers such as depth and normal. RGB2AO produces a new image filter that creates a non-directional shading effect that darkens enclosed and sheltered areas. RGB2AO aims to enhance two 2D image editing applications: image composition and geometry-aware contrast enhancement. We first collect a synthetic dataset consisting of pairs of RGB images and AO maps. Subsequently, we propose a model for RGB2AO by supervised learning of a convolutional neural network (CNN), considering 3D geometry of the input image. Experimental results quantitatively and qualitatively demonstrate the effectiveness of our model.

CCS Concepts

• ***Computing methodologies*** → ***Image-based rendering;***

1. Introduction

Ambient occlusion (AO) [CT82, ZIK98] is an important rendering technique in 3D computer graphics that significantly improves the visual quality of renders. Ambient occlusion works by darkening the regions in the image where the local scene geometry is concave, where neighboring surfaces shadow or occlude part of the ambient lighting. AO-based rendering is highly effective in adding realistic shading to renders that otherwise often look flat due to the lack of physically accurate global illumination effects. This technique has thus received much attention in the graphics community since its inception. It has benefited from multiple improvements to its efficiency, leading to its widespread use in real-time applications such as video games [Fer04, AMHH08].

Interestingly, the idea of applying ambient occlusion has been appreciated in other areas beyond 3D computer graphics. Professional artists working with 2D content have also developed creative techniques leveraging AO-like information to enhance the realism and impression of their artworks, such as photographs, painting, and illustrations. In particular, it has been demonstrated that applying AO-like shading effect in RGB images in the wild would enable multiple practical applications [Lan09, Sam10], such as the ones shown in Fig. 1:

- 2D image composition: simply pasting an object (e.g., cars, boxes, and bottles) onto a background image (e.g., road and table) would look like the object is floating mid-air in the image. Adding AO around the inserted object would make the composite visually pleasing.
- Geometry-aware contrast enhancement: it can be used to (de)emphasize or exaggerate the geometry in a photograph, providing a user control on perceivable depth.

Conventional techniques for computing AO, however, require 3D information about the scene such as depth buffers, surface normals, or even the whole scene geometry [AMHH08, HSK16, BSD08]. As such, these AO techniques are limited to the rendering of 3D scenes and cannot generate AO from RGB images. To achieve the aforementioned effects for 2D image composition and contrast enhancement, artists often need to manually construct the AO information themselves through a tedious AO map painting process [Sam10, Mar18, Dor16].

This paper introduces the novel task of AO generation from a single RGB image, so that it applies to 2D image editing applications that we have described above. One possible approach to this task would be to compute depth information from the input image with monocular depth estimation methods [CFYD16, LS18, LRSK19] and apply conventional screen-space AO generation from it to obtain the AO map. In our experiments, however, we found this simple approach fails to generate sound AO maps. We observe that while state-of-the-art depth prediction performance has rapidly improved recently, the depth prediction results are still coarse and not accurate enough to be used with existing screen-based approaches that typically assume ground-truth depth values.

In this paper, we present RGB2AO, a learning-based AO generation framework that learns to generate the AO map directly from the input image. Users can then use the generated AO map for editing tasks such as compositing or contrast enhancement. We develop our RGB2AO framework with an image-to-image translation model based on a convolutional neural network. In addition, our model extends a recent image-to-image translation model to account for the 3D geometry of scenes. We also explore data augmentation strategy that is specific and beneficial to AO generation. Our contributions encourage the network to learn to predict an accurate AO map from an RGB input alone.

A key challenge for single image AO generation is the lack of data with ground-truth AO maps for training. To address this challenge, we contribute a large-scale synthetic dataset with thousands of RGB-AO pairs. We construct our dataset from a large number of high-quality 3D scenes that we render realistically. Experimental results show that our model can produce more favorable results compared to existing methods quantitatively.

In summary, our contributions are as follows:

- We introduce the novel task of image-based ambient occlusion generation. To this end, we develop a CNN-based AO generation model considering the 3D geometry of the scenes. To the best of our knowledge, we provide the first approach to infer AO from a single RGB image automatically.
- We contribute a large-scale dataset dedicated to AO generation. This dataset consists of a large number of images with associated accurate ground-truth AO maps.
- We demonstrate the effectiveness of our RGB2AO framework in the application of our AO generation as a filter for 2D image composition and geometry-aware contrast enhancement.

2. Related Work

2.1. Ambient Occlusion

Ambient occlusion (AO) [CT82, ZIK98] is a fast global illumination model that approximates the amount of light reaching a point on a surface considering occlusion by objects and surfaces around them. With reduced computational cost compared to ray-tracing, AO can produce realistic lighting effects such as soft shadows around objects. Real-time AO computation usually requires 2D depth and normal buffer input with respect to the camera viewpoint. AO generation algorithms usually randomly sample nearby pixels and infer AO for each pixel independently. Many algorithms have been proposed to achieve a good tradeoff between accuracy and speed, such as SSAO [Mit07,SA07], HBAO [BSD08], and ASSAO [MOBH11]. AO was generalized to directional occlusion by Ritschel et al. [RGS09], adding directional shadows and color to the original AO darkening effect. Other similar perceptual effects using only the depth buffer were proposed, such as the unsharp masking operator [LCD06].

Recently, data-driven methods using neural networks have been proposed for AO generation, e.g., NNAO [HSK16] and Deep Shading [NAM*17]. They perform better than classical AO computation methods in the same runtime. In NNAO, the authors first collected a large number of paired depth/normal buffers and AO maps and trained a multi-layer perceptron (MLP). Deep Shading [NAM*17] confirms that CNN is better than MLP or classical methods for AO generation, in that it allows larger receptive fields through a stack of convolution and down-sampling layers. Our approach and those approaches are similar in spirit in employing neural networks for predicting the screen-space AO effect. However, those methods assume access to accurate screen-space buffers such as depth and normal. In contrast, our RGB2AO directly generates screen-space AO without an accurate estimation of normal and depth.

2.2. Intrinsic Image Decomposition

Intrinsic image decomposition [LM71] separates an image into a reflectance layer and a shading layer. Recent methods such as [BBS14, YYS*17] show promising results on real-world scenes. However, there is no clear way to extract AO from their shading estimation. Although shading and AO have an almost similar look under spatially uniform and non-directional lighting, our focus is on AO generation in real-world scenes that contain a diverse set of materials and objects lit by complex illumination. [HWBS15] detects AO from multiple images captured with varying illumination. [BM14] detects normal, reflectance, and illumination from shading. In comparison, our proposed method focuses on generating the AO of a whole scene from a single image.

Most related to our work, Innammorati *et al.* [IRWM17] decomposes a single RGB image into diffuse albedo, diffuse illumination, and specular shading and ambient occlusion. Their method aims to *estimate* AO that is already present in an image. In contrast, we *generate* reasonable AO that is not present by inferring geometry and semantics. We emphasize that estimation and generation are entirely different tasks. This difference becomes clear in applications such as image composition, where there is no AO present between the foreground and background in the image, as shown in Sec. 6.1.

2.3. Image Editing

Image Composition Image composition is one of the most common tasks in image editing. In image composition, a foreground region of one image is extracted and pasted to a background region of another image. Generating realistic composited images requires a plausible match for both contexts and appearances. Given a composited image and a mask to identify the foreground, image harmonization methods try to match the appearance of the foreground to that of the background (or vice versa) using global statistics [RAGS01, LE07, SJMP10, XADR12], gradient domain [PGB03, TJP10], or supervised learning [ZKSE15, TSL*17, ZZL19]. However, these approaches only modify inside the foreground region and do not consider the effect of placement, such as occlusion by the placed foreground region itself. For example, those methods cannot produce the soft shadow underneath a car on a sunny day. To the best of our knowledge, our RGB2AO is the first attempt to produce such an effect in image composition.

Image Relighting Lighting estimation from a single image has long been studied [LEN12]. There has been much progress in this field thanks to data-driven approaches for both outdoor scenes [HGSH*17, HGAL19] and indoor scenes [GSY*17, GSH*19]. Estimated lighting condition is used to photo-realistically render 3D objects into background images with many lighting conditions. However, our RGB2AO aims for inserting 2D objects, and these approaches cannot be applied.

2.4. Depth Estimation

Monocular depth estimation from a single RGB image is a fundamental task and has long been studied [SCN05, EPF14, LSLR15, LRB*16]. Recent methods try to encourage smoother gradient changes and sharp depth discontinuities [LS18] or obtain a model that generalizes well on datasets unseen during training [LRSK19]. However, estimating depth that is accurate enough to generate AO on top of it is very hard, as we will later show in Sec. 5.3.

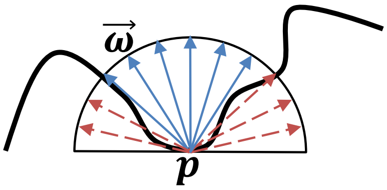

Figure 2: *Overview of how AO is computed. A hemisphere of rays $\vec{\boldsymbol{\omega}}$ is constructed for a given point $\vec{p}$ on a surface. Red (blue) arrows are rays that are (not) occluded by surrounding surfaces.*

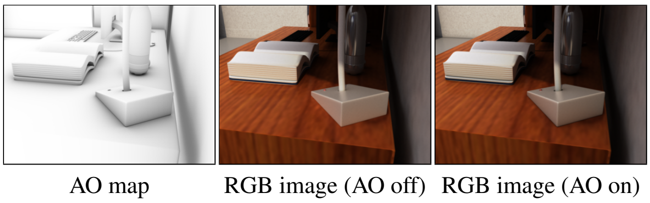

Figure 3: *Ambient occlusion effect applied to an RGB image. Ambient occlusion works by darkening the regions in the image according to the local scene geometry. This effect is achieved by multiplying the AO map (left) to the RGB image (middle) to obtain the enhanced render (right).*

3. AO Formation Model

We briefly summarize how AO is typically computed and used to approximate global illumination in computer graphics. Given a surface point $\boldsymbol{p}$ shown in Fig. 2, ambient occlusion illumination $AO(\boldsymbol{p},\vec{n})$ is defined as follows [SA07]:

$$AO(\boldsymbol{p},\vec{n}) = \frac{1}{\pi}\int_{\Omega} V(\vec{\boldsymbol{\omega}},\boldsymbol{p})\max(0,\vec{\boldsymbol{\omega}}\cdot\vec{n})\,d\vec{\boldsymbol{\omega}}\,, \tag{1}$$

where $\vec{n}$ is the surface normal at point $\boldsymbol{p}$, and $V(\vec{\boldsymbol{\omega}},\boldsymbol{p}) \in \{0,1\}$ is the visibility function over the normal-oriented hemisphere Ω, and $V(\vec{\boldsymbol{\omega}},\boldsymbol{p})$ is one if a ray starting from $\boldsymbol{p}$ intersects an occluder within some fixed distance from $\boldsymbol{p}$ and otherwise zero. The range of $AO(\boldsymbol{p})$ is $0 \leq AO(\boldsymbol{p}) \leq 1$, where $\boldsymbol{p}$ is zero when $\boldsymbol{p}$ is fully visible, and $\boldsymbol{p}$ is one when the whole hemisphere at $\boldsymbol{p}$ is occluded.

Computing integrals in Eq. (1) for each point of a 3D scene is requires excessive computational cost for real-time rendering. To generate a plausible AO efficiently for a specific camera viewpoint, most approaches use information from the screen-space buffers such as depth and normal of neighboring pixels to speed up the computation [Mit07, SA07, BSD08].

Let $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$ be an RGB image, where H and W represent height and width of the image. Applying Eq. (1) on each pixel of $\boldsymbol{x}$, we obtain its corresponding grayscale AO $\boldsymbol{y} \in \mathbb{R}^{H\times W}$. An example of the generated $\boldsymbol{y}$ is shown in Fig. 3. Here, we instead plot the values of $1-\boldsymbol{y}$ for the purpose of intuitive visualization. We call $1-\boldsymbol{y}$ the AO map. When applying the AO effect to create a new image $\boldsymbol{x}' \in \mathbb{R}^{3\times H\times W}$, each pixel in $\boldsymbol{x}'$ is computed by multiplying its color value by the corresponding pixel in the AO map:

$$x'_{ijk} = \left(1 - a\cdot y_{jk}\right)\cdot x_{ijk}\,, \tag{2}$$

where i is the index for the channel dimensions, and j, k are the indices for the pixel dimensions, $0 \leq a \leq 1$ is an arbitrary chosen scaling factor. When $a = 0$, the image is not modified ($x'_{ijk} = x_{ijk}$). When $a = 1$, Eq. (2) is reduced to $x'_{ijk} = (1 - y_{jk}) \cdot x_{ijk}$.

In the following sections, we describe how we approximate Eq. (1) by using only RGB information.

4. Proposed Method

We propose an end-to-end trainable neural network model for AO generation from a single image. First, we describe the baseline approach following recent conditional GAN methods for image-to-image translation in Sec. 4.1. We subsequently describe our AO generation model in Sec. 4.2. We propose two extensions: (i) multi-task learning of AO and depth prediction (in Sec. 4.2.1) and (ii) data augmentation that is specific to AO generation (in Sec. 4.2.2).

4.1. Baseline Model

Here we briefly introduce a recent image-to-image translation model for our baseline. Given a set of images $\{\ldots, (\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, the objective is to obtain a generator G that converts $\boldsymbol{x}$ to $\boldsymbol{y}$. To obtain a better G, conditional GAN methods for image-to-image translation such as Pix2pix [IZZE17] introduce a discriminator D that aims to distinguish real images from generated images. Conditional GANs model the conditional distribution of real images by the following minimax game:

$$\min_G \max_D \mathcal{L}_{GAN}(G, D), \tag{3}$$

where the objective function $\mathcal{L}_{GAN}(G, D)$ is defined as

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}[\log D(\boldsymbol{x}, \boldsymbol{y})] + \mathbb{E}_{\boldsymbol{x}}[\log(1 - D(x, G(x)))], \tag{4}$$

where we use $\mathbb{E}_{\boldsymbol{x}} \triangleq \mathbb{E}_{\boldsymbol{x} \sim p_{data}(x)}$ and $\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})} \triangleq \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}) \sim p_{data}(x,y)}$.

To have larger and better receptive field, Pix2pixHD [WLZ*18] uses multi-scale discriminators $D_1, \ldots, D_N$. An input to D_n is downsampled by a factor of 2^{n-1}. Then, Eq. (3) becomes as follows:

$$\min_G \max_{D_1,\ldots,D_N} \sum_{k=1}^{N} \mathcal{L}_{GAN}(G, D_k). \tag{5}$$

Pix2pixHD [WLZ*18] also introduces a feature matching loss that matches an intermediate representation of a discriminator from the real and the synthesized image. Given the i-th layer feature extractor of discriminator D_k as $D_k^{(i)}$, the feature matching loss $\mathcal{L}_{FM}(G, D_k)$ is as follows:

$$\mathcal{L}_{FM}(G, D_k) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})} \sum_{i=1}^{T} \frac{1}{N_i} \left[||D_k^{(i)}(\boldsymbol{x}, \boldsymbol{y}) - D_k^{(i)}(\boldsymbol{x}, G(\boldsymbol{x}))||_1 \right], \tag{6}$$

where N_i indicates the total number of elements in each layer and T is the total number of layers. Thus, the full objective is as follows;

$$\min_G \left(\left(\max_{D_1,\ldots,D_N} \sum_{k=1}^{N} \mathcal{L}_{GAN}(G, D_k) \right) + \alpha \sum_{k=1}^{N} \mathcal{L}_{FM}(G, D_k) \right), \tag{7}$$

where α is a hyper-parameter to balance the two terms.

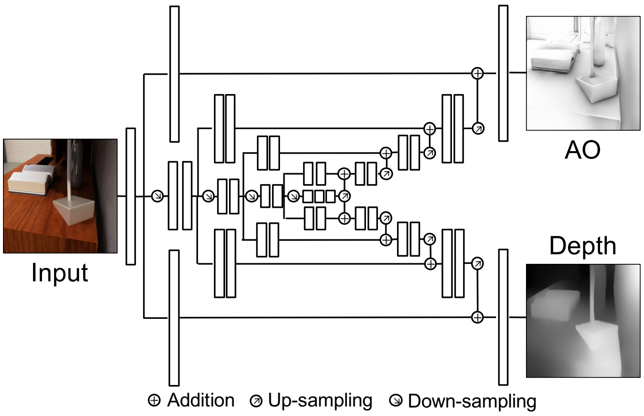

Figure 4: *RGB2AO model overview. We develop a fully convolutional network for AO map generation from an input RGB image. We extend a variant of the Hourglass network to enable multi-task learning, so as to encourage the learned feature to capture geometry-aware information relevant to the AO generation task.*

4.2. AO Generation Model

In this paper, we have a set of triplets $\{\ldots, (\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{d}_i)\}$, where $\boldsymbol{x}_i \in \mathbb{R}^{3 \times H \times W}$, $\boldsymbol{y}_i \in \mathbb{R}^{1 \times H \times W}$, $\boldsymbol{d}_i \in \mathbb{R}^{1 \times H \times W}$ indicate an RGB image, an AO map, and a depth map, respectively. Our objective is to obtain a best generator G that converts $\boldsymbol{x}$ to $\boldsymbol{y}$. We extend the baseline model in Sec. 4.1 in two points that are specific to AO generation: (i) multi-task learning of AO and depth and (ii) data augmentation by randomizing AO distribution.

4.2.1. Multi-task Learning of AO and Depth

Simultaneously inferring both AO and depth is beneficial because ambient occlusion is closely related to 3D geometry of a scene. As shown in Fig. 4, we introduce a depth estimation model F that takes $\boldsymbol{x}$ and generates the depth $F(\boldsymbol{x}) \in \mathbb{R}^{1 \times W \times H}$. For F, we use a similar CNN with G and share the encoder part for multi-task learning. For loss functions, we adopt loss functions used in MegaDepth [LS18].

$$\mathcal{L}_D(F) = \mathcal{L}_{data}(F) + \gamma \mathcal{L}_{grad}(F), \tag{8}$$

where γ is a hyper-parameter and $\mathcal{L}_{data}(F)$ and $\mathcal{L}_{grad}(F)$ are called scale-invariant data term and multi-scale scale-invariant gradient matching term, respectively.

Scale-Invariant Data Term Let r_i be the residual of values between predicted and ground truth log-depth at pixel position i, $\mathcal{L}_{data}(F)$ is defined as follows:

$$\mathcal{L}_{data}(F) = \frac{1}{n} \sum_{i=1}^{n} (r_i)^2 - \frac{1}{n^2} \left(\sum_{i=1}^{n} r_i \right)^2, \tag{9}$$

where n is the number of valid depth values in the ground truth depth maps.

Multi-Scale Scale-Invariant Gradient Matching Term $\mathcal{L}_{grad}(F)$ is defined as follows:

$$\mathcal{L}_{grad}(F) = \frac{1}{n} \sum_k \sum_i (|\nabla_x r_i^k| + |\nabla_y r_i^k|), \tag{10}$$

Table 1: *Dataset statistics.*

Subset	# 3D scenes	# images
Train	21	7188
Validation	3	397
Test	4	1005

where r_i^k is the log-depth residual at position i and scale k.

Therefore, the full objective is as follows:

$$\min_{G,F}\left(\left(\max_{D_1,\ldots,D_N}\sum_{k=1}^{N}\mathcal{L}_{GAN}(G,D_k)\right)+\alpha\sum_{k=1}^{N}\mathcal{L}_{FM}(G,D_k)+\beta\mathcal{L}_D(F)\right). \tag{11}$$

Here, β is a hyper-parameter to balance the multi-task learning.

4.2.2. AO augmentation

Our synthetic dataset used for training contains RGB images devoid of AO-like effects. However, those effects are already present to some unknown extent on real-world images. In order for our method to generalize to real captured images, we propose to augment the input RGB images by adding some AO darkening during training. Formally, we use Eq. (2) to generate a new image, with the scaling factor a taken from the uniform distribution $\mathcal{U}(a_{min}, a_{max})$. We empirically set $(a_{min}, a_{max}) = (0.0, 0.5)$. This AO augmentation is applied on each image with probability $p = 0.75$, leaving 25% of the images without AO effects during training.

4.3. Dataset

To train our data-driven RGB2AO model, triplets of RGB-AO-depth data are required. We have collected a synthetic dataset since there is no dataset available. The dataset consists of 8590 triplets of RGB-AO-depth data in a resolution of 1920×1120. The dataset is rendered from 3D scenes using Maya [Inc19] with Arnold renderer for ray-tracing. Each rendered data has its unique view, and we manually sample the view to cover a broad range of situations, as shown in Fig. 5. Most of the scenes come from typical indoor scenes such as kitchen, living room, and bedroom while we include some outdoor scenes and non-photorealistic scenes. We also manually compose some scenes that contain objects on the floor or cars on the synthetic ground to cover possible situations.

For rendering, we use a perspective camera with the focal length of 18 to make the view wide enough for indoor situations. We set the "falloff" parameter for the AO in Arnold to 0.2. We found that using larger values caused the AO to spread very far from the objects making it harder for the model to infer. In addition to RGB and AO, we also rendered a screen-space depth buffer to see whether simultaneously estimating 3D geometry information benefits the AO generation.

5. Experiments

5.1. Network Architecture

For the generator G, we used a variant of a hourglass network used in monocular depth estimation [CFYD16]. The hourglass network consists of multiple convolutions by a variant of inception [SLJ*15] and down-sampling, followed by multiple convolutions and up-sampling, interleaved with skip connections. We duplicate skip-connection parts and decoder parts for multi-task learning, as shown in Fig. 4. Since it is a fully convolutional network, it can be applied to images with arbitrary sizes during the test phase. For the discriminator D, we used a discriminator which is identical to the one used in Pix2pixHD [WLZ*18].

5.2. Training

G and D were trained for 100 epochs starting with a learning rate of 2.0×10^{-4} and a batch size of 6 with Adam optimizer [KB15]. We kept the learning rate for the first 50 epochs and then linearly decay the rate to zero over the next 50 epochs. It took about a day for the training to finish. For the hyper-parameters we set $(\alpha, \beta, \gamma) = (1.0, 1.0, 0.5)$ in all the experiments. Since the number of samples in the dataset is limited, we also performed massive standard data augmentation to enhance the generalization capability of our model. We changed contrast, brightness, saturation, and hue of each image. All the images were resized to 384×224. During training, the images were randomly flipped and cropped to 352×192. During testing, images were center-cropped to 352×192 for the quantitative evaluation.

5.3. AO Generation Performance

We performed the quantitative evaluation on our synthetic dataset to prove the validity of our model for AO generation. We split the images in our dataset into train, validation, and test subsets so that each does not share the same 3D scenes, as shown in Table 1. The quantitative evaluation was performed on the images with no AO applied in the test subset.

5.3.1. Evaluation Metrics

We evaluated the performance of AO generation methods by comparing the generated AO maps with the corresponding ground-truth AO maps over the whole testing set. In addition to mean absolute error (MAE) and mean squared error (MSE), we used the following evaluation metrics:

- SSIM: Structural similarity (SSIM) index [WBS*04] is widely used in quantifying the perceptual similarity between two images. Higher is better.
- LPIPS: Learned Perceptual Image Patch Similarity (LPIPS) metric [ZIE*18] is a recently developed measure for perceptual similarity assessment. LPIPS uses features extracted from CNN trained on a dataset of human perceptual similarity judgments on image pairs. Lower is better.

5.3.2. Compared Methods

To the best of our knowledge, there is no existing image-based AO generation method in the literature. In this experiment, we evaluated the AO generation performance of our method and compared it with the following methods: (i) screen-space AO methods on top of monocular depth estimation results or (ii) Innamorati *et al.* [IRWM17]'s model originally for AO estimation.

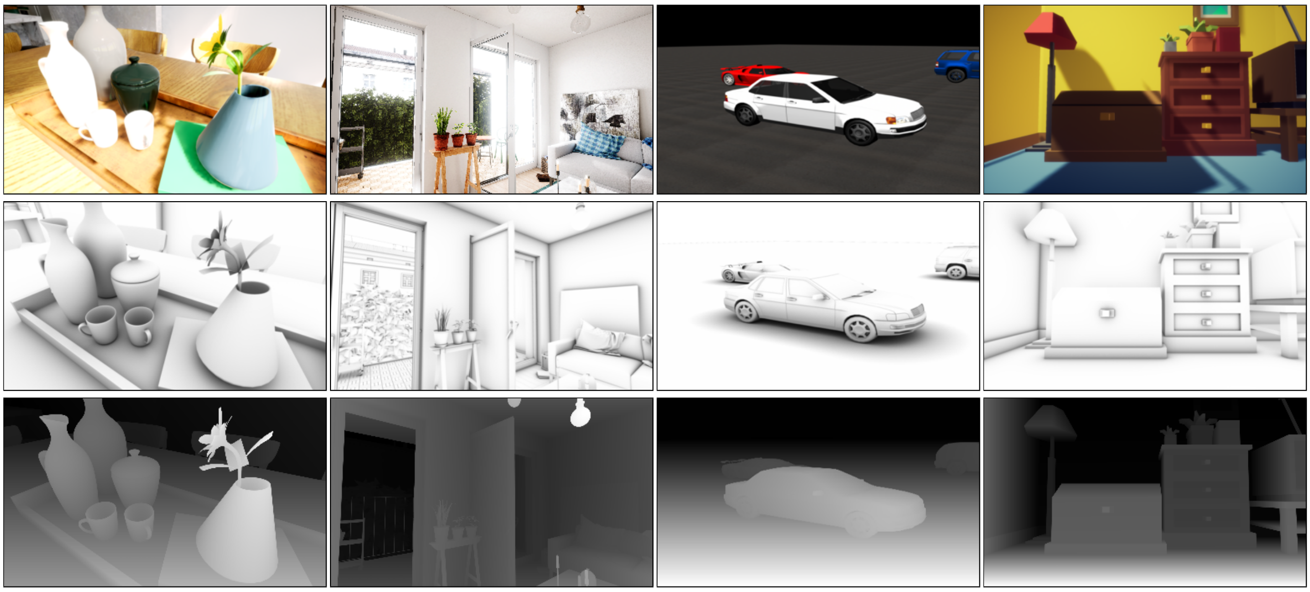

Figure 5: *Dataset for image-based AO generation. We show some samples from our large-scale synthetic data for AO generation. From top to bottom: RGB images, AO maps, and depth maps. The data is created from high-quality 3D scenes covering a wide range of scene content. The images as well as the corresponding AO maps and depth maps are rendered using Maya with the Arnold ray-tracing based renderer.*

Depth Estimation + Screen-Space AO We first trained a monocular depth estimation network, which we call RGB2D for short, using the same dataset. We further improved the quality of the generated depth map by applying bilateral filter [TM98] to smooth the predicted depth values. For training the RGB2D model, we used the same network with F to extract depth. Second, we applied different screen-space AO generation methods on top of the resulting depth estimation results to obtain the final AO map. We experimented with the following three variants.

- **RGB2D+SSAO**: We used a traditional AO generation method, SSAO [SA07].
- **RGB2D+CNN**: We used a CNN that is almost similar to G for a fair comparison. Only the difference is changing the input from a three-channel RGB image to a one-channel depth map.
- **RGB2D (fixed) + CNN**: One may argue that monocular depth estimation methods trained on a mixture of various datasets can be used to extract fine depth without training on our dataset. We tested a pre-trained monocular depth estimation method TM-RDE [LRSK19], which generalizes well to an unseen dataset, without training.

Innamorati *et al.* We compare three methods derived from Innamorati *et al.* [IRWM17]'s model.

- **Innamorati-est**: We directly run publicly available Innamorati *et al.*'s model which is trained on their own dataset.
- **Innamorati-ft-est**: We finetuned Innamorati *et al.*'s model on our dataset for AO *estimation*. Because the task is AO estimation, we used pairs of RGB with AO and AO as the input and the target of the model, respectively.
- **Innamorati-ft-gen**: For a fairer comparison, we finetuned the model on our dataset for AO *generation*. We applied the L2 loss to the occlusion output only. We ignored the other outputs and losses because they are irrelevant or detrimental. For example, the reconstruction loss forces the network to *detect* the AO present and thus unfairly hurt the network's ability to *generate* missing AO.

Table 2: *Experimental results of AO generation on our synthetic dataset. ↓ and ↑ indicate that lower and higher is better, respectively.*

	MAE ↓	MSE ↓	SSIM ↑	LPIPS ↓
RGB2D+SSAO	0.0848	0.0193	0.611	0.524
RGB2D+CNN	0.0785	0.0181	0.687	0.484
RGB2D(fixed)+CNN	0.0797	0.0186	0.689	0.549
Innamorati-est	0.0952	0.0215	0.671	0.423
Innamorati-ft-est	0.0655	0.0120	0.764	0.311
Innamorati-ft-gen	0.0668	0.0107	0.763	0.329
Ours	**0.0589**	**0.0103**	**0.767**	**0.235**

5.3.3. Results

Accuracy Quantitative results are shown in Table 2. Our model outperforms both types of approaches in each metric by a significant margin. We show generated AO on an excerpt of the test set in Fig. 6. The Screen-Space AO approaches fail to capture many of the details on real images, because they rely heavily on monocular depth estimation results, which does not capture the high-frequency details for accurate SSAO computation. This demonstrates the importance of learning a direct RGB-to-AO mapping.

Innamorati *et al.*-based approaches produce very blurry outputs

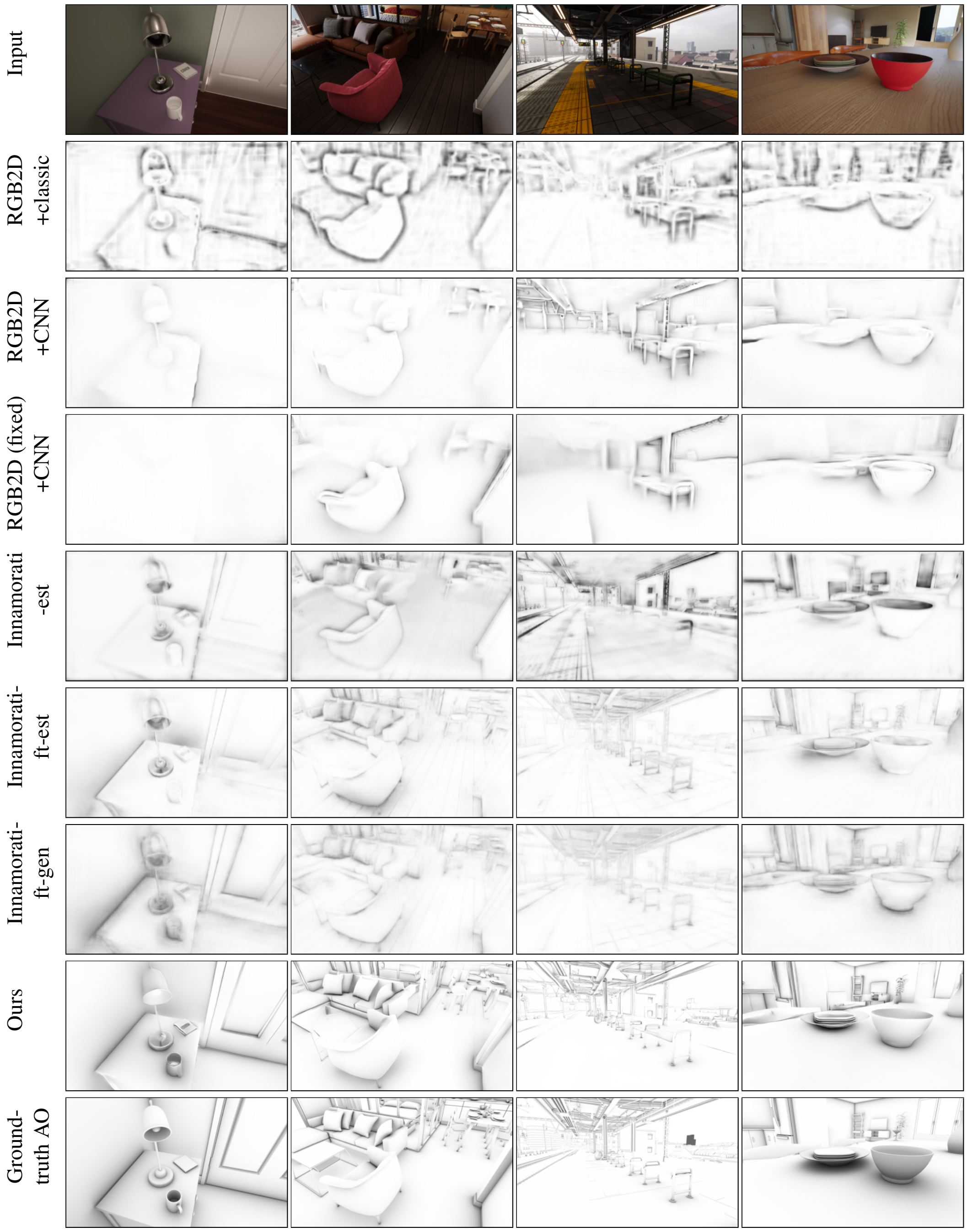

Figure 6: *Comparison of AO generation methods on the test subset of our collected synthetic dataset with no AO present in the input image. (best viewed with zoom.)*

even if the model is finetuned for AO generation. This demonstrates the importance of the choice of the networks and loss functions specialized for AO generation.

Speed We benchmarked our model on several image resolutions. Since our model is a fully convolutional network, it can take images with arbitrary aspect ratios as input. We tested images initialized randomly and show the average time on 100 runs on a TITAN V GPU. Our model runs at about 15 FPS and 8 FPS for 384×224 and 768×448 inputs, respectively.

Table 3: *An ablation study on our proposed components. (i) and (ii) indicate AO augmentation (Sec. 4.2.2) and multi-task learning of AO and depth prediction (Sec. 4.2.1), respectively. ↓ and ↑ indicate that lower and higher is better, respectively.*

(i)	(ii)	MAE ↓	MSE ↓	SSIM ↑	LPIPS ↓
		0.0622	0.0110	0.748	0.243
✓		0.0600	**0.0103**	0.760	**0.235**
	✓	0.0607	0.0107	0.755	0.241
✓	✓	**0.0589**	**0.0103**	**0.767**	**0.235**

Table 4: *An ablation study on changing a_{max} in our AO augmentation (Sec. 4.2.2) during training.*

a_{max}	0.0	0.25	0.5	0.75	1.0
SSIM ↑	0.755	0.762	**0.767**	0.764	0.762

5.4. Ablation Study

5.4.1. Contribution of Proposed Components

We performed quantitative evaluation by changing components of our model and the result is shown in Table 3. Both of the two components are essential to achieve the best result. Surprisingly, the model without both AO augmentation and multi-task learning already clearly surpasses all the compared methods in Table 2. This is due to our choice of networks and loss functions specific to AO generation. We performed an ablation study about these choices in the supplementary material.

5.4.2. AO Augmentation

AO augmentation that we propose in Sec. 4.2.2 has one hyperparameter a_{max}, where $0.0 \leq a_{max} \leq 1.0$, to control the strength of the AO effect that we apply during training. Bigger a_{max} leads to include images with much AO effect during training. We tested different a_{max} and the result is shown in Table 4. As we increase a_{max} from 0.0 (no AO effect applied) to increase the AO effect, the performance was steadily improved until $a_{max} = 0.5$. However, it decreased when $a_{max} > 0.5$. This indicates that including images with excessive AO effect is harmful to improve the performance.

6. Applications

AO generation is useful for many image editing and enhancement tasks. We demonstrate its ability to improve 2D image composition and for increasing contrast in a geometry-aware manner. We edit the input RGB image using the generated AO by the simple element-wise multiplication following Eq. (2). We set the scaling factor $a = 1.0$ in Eq. (2) to show all the results in this section. The experiments are performed on images of size 384×224. Larger input images can be handled by (i) resizing the image to a lower resolution, (ii) estimating AO, (iii) resizing it back to the original resolution, and (iv) multiplying the generated AO with the input. We show some examples of processing high-resolution images in the supplementary material.

6.1. 2D Image Composition

For image composites to look realistic, the lighting of the inserted object must match the background scene. Prior harmonization methods such as [TSL*17] have been developed to address this. Additionally, the object will cast shadows or otherwise affect the lighting of the scene around it. While lighting estimation methods have been used to insert 3D objects, inserting 2D objects is more complicated. With our AO generation, we can address this. Given a composited image, we can apply the generated AO to make it look more realistic as the AO will darken the contact regions around the composited object to simulate the shadows and help ground the object in the scene.

User Study To evaluate the effectiveness of our image composition results on real images quantitatively, we performed a user study. We created a set of 26 images that covers a wide variety of real examples, where some objects are composited (e.g., cups, cars, and chairs). Some of the images were taken from the evaluation set of Deep Image Harmonization (DIH) [TSL*17]. Some example results are shown in Fig. 7. To show that our method is complementary to existing image harmonization methods, we used the harmonized image by DIH as an input to our RGB2AO model. We showed the two images to the subject and asked which one looks more realistic. As a result, a total of 9 subjects participated in this study, with a total of 324 votes. Our method obtained 80% of the votes (260 votes) over the compared method, showing a clear preference for our method on this composition task.

Discussion One limitation of RGB2AO applied to 2D image composition is that our composition sometimes changes some regions of the background image, where compositing the foreground region should not affect. For example, in the third and fourth row of Fig. 7, boundaries between walls and floors were exaggerated apparently, but the composition of objects should not affect those regions. A practical solution is to let users choose the region where the generated AO is multiplied. Extending the current AO generation to automatically propose regions is a promising research direction for future work.

6.2. Geometry-Aware Image Contrast

We can apply the generated AO map as an image filter for RGB images in order to improve image contrast in such a way as to emphasize the depth in the image. We enumerate some usages:

Avoiding Flat Look Images without sufficient contrast appear flat and dull. Applying pixel-level contrast enhancement methods can improve the image appeal, but might break its overall coherence. For example, shadows may become either over- or under-exposed. Instead, our method allows applying contrast in a geometry-aware manner by using AO to darken the right areas based on the scene geometry. Some example results are shown in Fig. 8. For comparison, we also tested an auto-contrast enhancement method, Auto Contrast in Adobe Photoshop. We can see that our geometry-aware image contrast behaves entirely differently from Auto Contrast from the second and third columns in Fig. 8.

One may argue that AO estimation can perform similarly on real photos where no AO is missing. We drop AO augmentation part

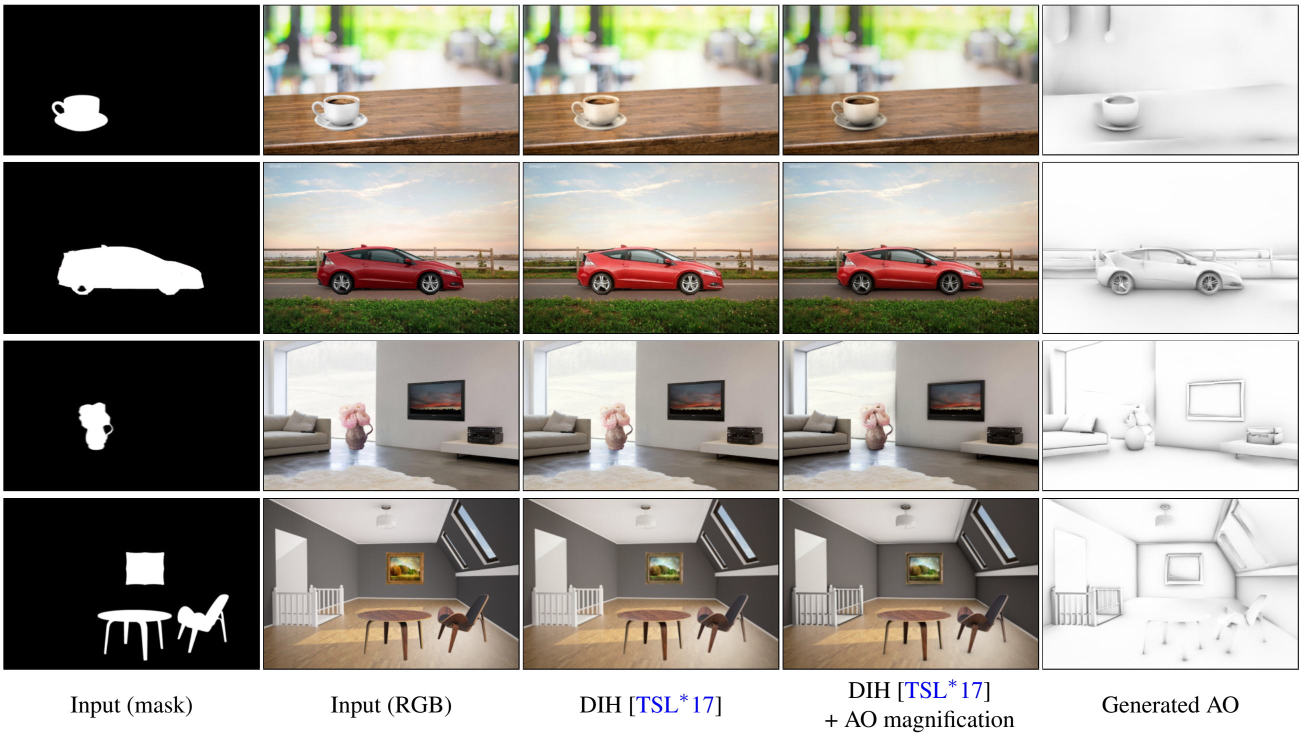

Figure 7: *Example results on 2D image composition. Our method adds a plausible shadow-like effect on both the foreground and background region using the generated AO. Our method is complementary to DIH [TSL*17], which only changes the appearance inside the foreground region. Images in the second and third row are from DIH. Images in the first and fourth row from Adobe Stock. (best viewed with zoom.)*

from our model and train it for AO estimation using pairs of RGB with AO on and AO map. We show the result of AO generation and estimation in the fourth and fifth columns in Fig. 8, respectively. Real photos can contain full AO if the scenes are lit only by spatially uniform and non-directional lighting. However, real scenes are usually lit by complex illumination, making the visible AO reduced and making the AO estimation difficult, as we can see. Thus, we believe that AO generation is different from AO estimation even in real photos, and it is suitable for our downstream application, geometry-aware contrast enhancement.

Manipulating Non-Photorealistic Images The usage of our RGB2AO model is not limited to photorealistic RGB images. Some artists have tried to depict AO-like soft shadows on illustrations manually by brush or some other tools. In contrast, our RGB2AO model can do it in a fully-automatic way. Results of manipulated non-photorealistic illustration images are shown in Fig. 9. We can see that our model can generate plausible AO on everyday objects (e.g., chairs, plants, and pots) and scenes (e.g., room).

7. Limitations

One limitation of our RGB2AO model is that the model struggles on families of objects and scenes that are not under-represented in the dataset for training. Some examples of the failure cases are shown in Fig. 10. Collecting larger datasets with more diverse scenes would alleviate this type of errors.

Another limitation is that our RGB2AO can only handle non-directional shadow-like effects. Augmenting our AO generation framework with lighting prediction in addition to geometry understanding can potentially handle such directional shadow or inter-reflection. We hope our proposed method paves the way for future work in this direction.

8. Conclusion

We present RGB2AO, a novel task that generates AO from a single RGB image with arbitrary size. Our model for RGB2AO is a fully convolutional CNN specially designed for this task by extending an image-to-image translation in two points: data augmentation specific for AO and joint inference of 3D-geometry information by multi-task learning. We show that our model can generate AO adequately. We also apply the generated AO to two image modification tasks, contrast enhancement and 2D image composition, and show remarkable results that would not be possible without our AO-based modification.

9. Acknowledgement

This work was done when N. Inoue was an intern at Adobe Research. N. Inoue is financially supported by The University of Tokyo - NEC AI scholarship. We gratefully acknowledge the support of Adobe with generous gift funding.

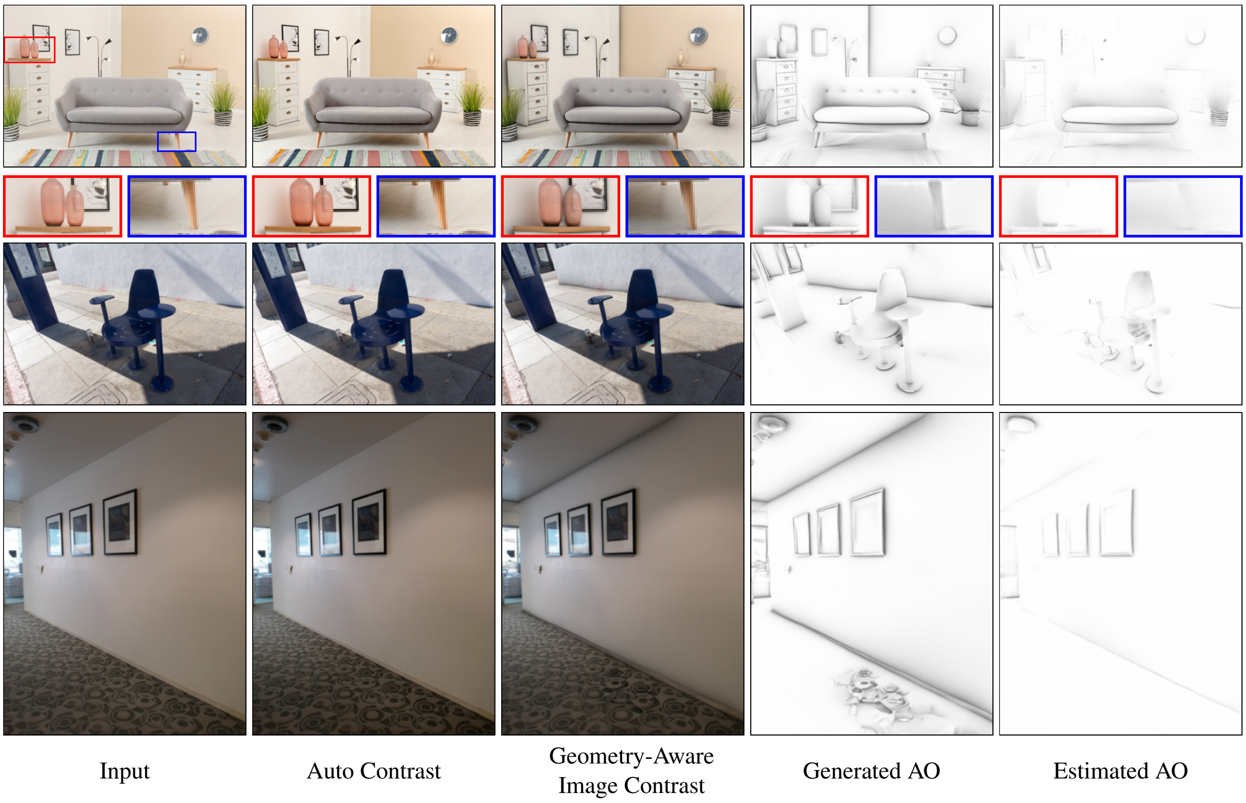

Figure 8: *Comparison between our AO-based contrast enhancement and Auto Contrast in Adobe Photoshop. Note how Auto Contrast changes the global appearance of the image while our technique focuses on darkening object boundaries such as the bottles (red) and under the sofa (blue) in the first row. AO estimation struggles to detect AO under real photos with complex illumination, such as the boundary between the wall and the ceiling in the third row. Images in the first and second rows from Adobe Stock. (best viewed with zoom.)*

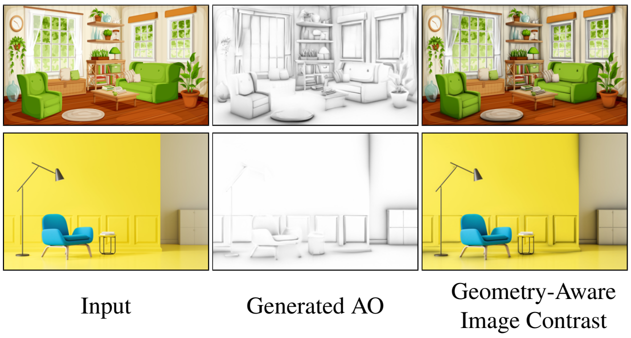

Figure 9: *Results of our RGB2AO on non-photorealistic images. Images from Adobe Stock.*

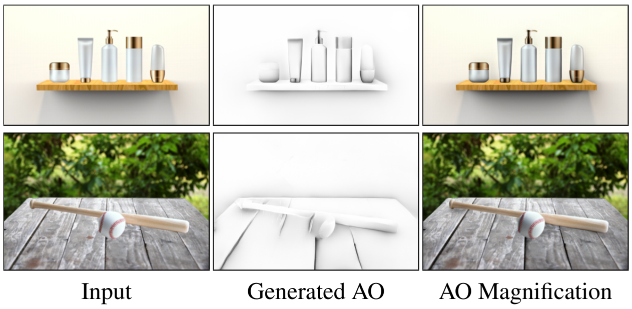

Figure 10: *Failure case examples. Our model failed to obtain AO on the bottom of bottles and a ball in the first and second row, due to the lack of similar content in the training data. Images from Adobe Stock. (best viewed with zoom.)*

References

[AMHH08] Akenine-Moller T., Haines E., Hoffman N.: *Real-Time Rendering*, 3rd ed. A. K. Peters, Ltd., 2008. 2

[BBS14] Bell S., Bala K., Snavely N.: Intrinsic images in the wild. *ACM Transactions on Graphics (TOG) 33*, 4 (2014), 159. doi:10.1145/2601097.2601206. 3

[BM14] Barron J. T., Malik J.: Shape, illumination, and reflectance from shading. *IEEE TPAMI 37*, 8 (2014), 1670–1687. doi:10.1109/TPAMI.2014.2377712. 3

[BSD08] Bavoil L., Sainz M., Dimitrov R.: Image-space horizon-based ambient occlusion. In *ACM SIGGRAPH 2008 talks* (2008), p. 22. doi:10.1145/1401032.1401061. 2, 3

[CFYD16] Chen W., Fu Z., Yang D., Deng J.: Single-image depth perception in the wild. In *Proc. NeurIPS* (2016), pp. 730–738. 2, 5

[CT82] COOK R. L., TORRANCE K. E.: A reflectance model for computer graphics. *ACM TOG 1*, 1 (1982), 7–24. doi:10.1145/357290.357293. 1, 2

[Dor16] DORIAN I.: How to make an ambient occlusion study. https://www.youtube.com/watch?v=WiC4tiOSn_M&feature=youtu.be, 2016. 2

[EPF14] EIGEN D., PUHRSCH C., FERGUS R.: Depth map prediction from a single image using a multi-scale deep network. In *Proc. NeurIPS* (2014), pp. 2366–2374. 3

[Fer04] FERNANDO R.: *GPU Gems: Programming Techniques, Tips and Tricks for Real-Time Graphics*. Pearson Higher Education, 2004. 2

[GSH*19] GARON M., SUNKAVALLI K., HADAP S., CARR N., LALONDE J.-F.: Fast spatially-varying indoor lighting estimation. In *Proc. CVPR* (2019), pp. 6908–6917. 3

[GSY*17] GARDNER M.-A., SUNKAVALLI K., YUMER E., SHEN X., GAMBARETTO E., GAGNÉ C., LALONDE J.-F.: Learning to predict indoor illumination from a single image. *ACM TOG* (2017). doi:10.1145/3130800.3130891. 3

[HGAL19] HOLD-GEOFFROY Y., ATHAWALE A., LALONDE J.-F.: Deep sky modeling for single image outdoor lighting estimation. In *Proc. CVPR* (2019), pp. 6927–6935. 3

[HGSH*17] HOLD-GEOFFROY Y., SUNKAVALLI K., HADAP S., GAMBARETTO E., LALONDE J.-F.: Deep outdoor illumination estimation. In *Proc. CVPR* (2017), pp. 7312–7321. doi:10.1109/CVPR.2017.255. 3

[HSK16] HOLDEN D., SAITO J., KOMURA T.: Neural network ambient occlusion. In *SIGGRAPH Asia Technical Briefs* (2016), p. 9. doi:10.1145/3005358.3005387. 2

[HWBS15] HAUAGGE D., WEHRWEIN S., BALA K., SNAVELY N.: Photometric ambient occlusion for intrinsic image decomposition. *IEEE TPAMI 38*, 4 (2015), 639–651. doi:10.1109/TPAMI.2015.2453959. 3

[Inc19] INC A.: Maya. https://www.autodesk.com/products/maya/overview, 2019. Accessed: 2019-09-19. 5

[IRWM17] INNAMORATI C., RITSCHEL T., WEYRICH T., MITRA N. J.: Decomposing single images for layered photo retouching. In *Computer Graphics Forum* (2017), vol. 36, pp. 15–25. doi:10.1111/cgf.13220. 3, 5, 6

[IZZE17] ISOLA P., ZHU J.-Y., ZHOU T., EFROS A. A.: Image-to-image translation with conditional adversarial networks. In *Proc. CVPR* (2017). doi:10.1109/CVPR.2017.632. 4

[KB15] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. 5

[Lan09] LANIER L.: *Professional Digital Compositing: Essential Tools and Techniques*. SYBEX Inc., Alameda, CA, USA, 2009. 2

[LCD06] LUFT T., COLDITZ C., DEUSSEN O.: Image enhancement by unsharp masking the depth buffer. *ACM TOG 25*, 3 (2006). doi:10.1145/1141911.1142016. 2

[LE07] LALONDE J.-F., EFROS A. A.: Using color compatibility for assessing image realism. In *Proc. ICCV* (2007), IEEE, pp. 1–8. doi:10.1109/ICCV.2007.4409107. 3

[LEN12] LALONDE J.-F., EFROS A. A., NARASIMHAN S. G.: Estimating the natural illumination conditions from a single outdoor image. *Springer IJCV 98*, 2 (2012), 123–145. 3

[LM71] LAND E. H., MCCANN J. J.: Lightness and retinex theory. *Josa 61*, 1 (1971), 1–11. doi:10.1364/JOSA.61.000001. 2

[LRB*16] LAINA I., RUPPRECHT C., BELAGIANNIS V., TOMBARI F., NAVAB N.: Deeper depth prediction with fully convolutional residual networks. In *Proc. 3DV* (2016), pp. 239–248. doi:10.1109/3DV.2016.32. 3

[LRSK19] LASINGER K., RANFTL R., SCHINDLER K., KOLTUN V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341* (2019). 2, 3, 6

[LS18] LI Z., SNAVELY N.: Megadepth: Learning single-view depth prediction from internet photos. In *Proc. CVPR* (2018), pp. 2041–2050. doi:10.1109/CVPR.2018.00218. 2, 3, 4

[LSLR15] LIU F., SHEN C., LIN G., REID I.: Learning depth from single monocular images using deep convolutional neural fields. *IEEE TPAMI 38*, 10 (2015), 2024–2039. doi:10.1109/TPAMI.2015.2505283. 3

[Mar18] MARCO B.: Ambient occlusion (and ambient light) for painters. https://www.youtube.com/watch?v=7fLV5ez064w&feature=youtu.be, 2018. 2

[Mit07] MITTRING M.: Finding next gen: Cryengine 2. In *ACM SIGGRAPH 2007 courses* (2007), pp. 97–121. doi:10.1145/1281500.1281671. 2, 3

[MOBH11] MCGUIRE M., OSMAN B., BUKOWSKI M., HENNESSY P.: The alchemy screen-space ambient obscurance algorithm. In *Proceedings of the ACM SIGGRAPH Symposium on High Performance Graphics* (2011), pp. 25–32. doi:10.1145/2018323.2018327. 2

[NAM*17] NALBACH O., ARABADZHIYSKA E., MEHTA D., SEIDEL H.-P., RITSCHEL T.: Deep shading: convolutional neural networks for screen space shading. 65–78. doi:10.1111/cgf.13225. 2

[PGB03] PÉREZ P., GANGNET M., BLAKE A.: Poisson image editing. *ACM TOG 22*, 3 (2003), 313–318. doi:10.1145/1201775.882269. 3

[RAGS01] REINHARD E., ADHIKHMIN M., GOOCH B., SHIRLEY P.: Color transfer between images. *IEEE Computer graphics and applications 21*, 5 (2001), 34–41. doi:10.1109/38.946629. 3

[RGS09] RITSCHEL T., GROSCH T., SEIDEL H.-P.: Approximating dynamic global illumination in image space. In *Proceedings of the 2009 symposium on Interactive 3D graphics and games* (2009), pp. 75–82. 2

[SA07] SHANMUGAM P., ARIKAN O.: Hardware accelerated ambient occlusion techniques on gpus. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games* (2007), pp. 73–80. doi:10.1145/1230100.1230113. 2, 3, 6

[Sam10] SAM N.: Sam nielson: Painting process. http://theartcenter.blogspot.com/2010/03/sam-nielson-painting-process.html, 2010. 2

[SCN05] SAXENA A., CHUNG S. H., NG A. Y.: Learning depth from single monocular images. In *Proc. NeurIPS* (2005), pp. 1161–1168. 3

[SJMP10] SUNKAVALLI K., JOHNSON M. K., MATUSIK W., PFISTER H.: Multi-scale image harmonization. *ACM TOG 29*, 4 (2010), 125. doi:10.1145/1833349.1778862. 3

[SLJ*15] SZEGEDY C., LIU W., JIA Y., SERMANET P., REED S., ANGUELOV D., ERHAN D., VANHOUCKE V., RABINOVICH A.: Going deeper with convolutions. In *Proc. CVPR* (2015), pp. 1–9. doi:10.1109/CVPR.2015.7298594. 5

[TJP10] TAO M. W., JOHNSON M. K., PARIS S.: Error-tolerant image compositing. In *Proc. ECCV* (2010), pp. 31–44. doi:10.1007/s11263-012-0579-7. 3

[TM98] TOMASI C., MANDUCHI R.: Bilateral filtering for gray and color images. In *Proc. ICCV* (1998), vol. 98, p. 2. doi:10.1109/ICCV.1998.710815. 6

[TSL*17] TSAI Y.-H., SHEN X., LIN Z., SUNKAVALLI K., LU X., YANG M.-H.: Deep image harmonization. In *Proc. CVPR* (2017), pp. 3789–3797. doi:10.1109/CVPR.2017.299. 3, 8, 9

[WBS*04] WANG Z., BOVIK A. C., SHEIKH H. R., SIMONCELLI E. P., ET AL.: Image quality assessment: from error visibility to structural similarity. *IEEE TIP 13*, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861. 5

[WLZ*18] 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. In *Proc. CVPR* (2018), pp. 8798–8807. doi:10.1109/CVPR.2018.00917. 4, 5

[XADR12] Xue S., Agarwala A., Dorsey J., Rushmeier H.: Understanding and improving the realism of image composites. *ACM TOG 31*, 4 (2012), 84. 3

[YYS*17] Yue H., Yang J., Sun X., Wu F., Hou C.: Contrast enhancement based on intrinsic image decomposition. *IEEE TIP 26*, 8 (2017), 3981–3994. doi:10.1109/TIP.2017.2703078. 3

[ZIE*18] Zhang R., Isola P., Efros A. A., Shechtman E., Wang O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR* (2018), pp. 586–595. doi:10.1109/CVPR.2018.00068. 5

[ZIK98] Zhukov S., Iones A., Kronin G.: An ambient light illumination model. In *Rendering TechniquesâĂŹ 98*. 1998, pp. 45–55. 1, 2

[ZKSE15] Zhu J.-Y., Krahenbuhl P., Shechtman E., Efros A. A.: Learning a discriminative model for the perception of realism in composite images. In *Proc. ICCV* (2015), pp. 3943–3951. doi:10.1109/ICCV.2015.449. 3

[ZZL19] Zhan F., Zhu H., Lu S.: Spatial fusion gan for image synthesis. In *Proc. CVPR* (2019), pp. 3653–3662. 3