Shadow Harmonization for Realistic Compositing

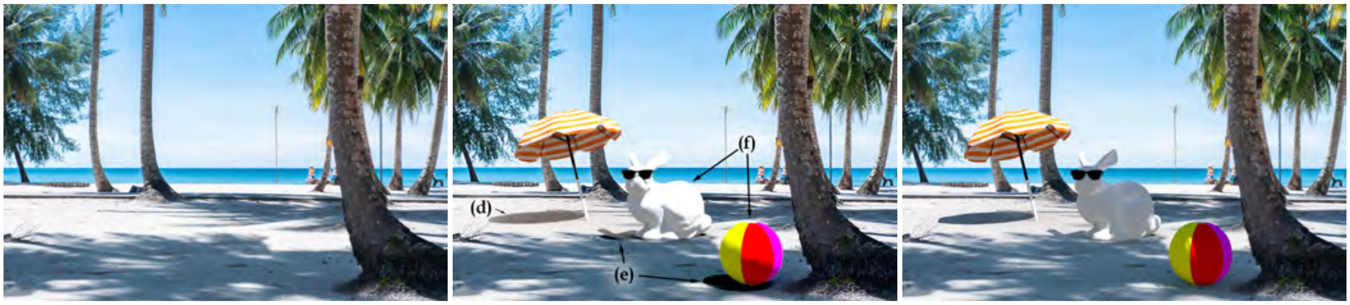
Lucas Valença
Université Laval
Québec, QC, Canada
lucas.valenca@ulaval.ca

Jinsong Zhang
Université Laval
Québec, QC, Canada
jinsong.zhang.1@ulaval.ca

Michaël Gharbi
Adobe
San Francisco, CA, USA
mgharbi@adobe.com

Yannick Hold-Geoffroy
Adobe
San Francisco, CA, USA
holdgeof@adobe.com

Jean-François Lalonde
Université Laval
Québec, QC, Canada
jflalonde@gel.ulaval.ca



(a) input (b) traditional IBL (c) ours

Figure 1: Given a single low dynamic range (LDR) image of an outdoor scene (a), compositing virtual objects using traditional methods such as image-based lighting (IBL, without differential rendering [Debevec 1998]) (b) yields multiple issues such as incorrect shadow intensity and color (d, umbrella), double shadows (e, beach ball, bunny) that do not blend with the background (palm tree), and lack of shadows cast onto virtual objects (f, on the beach ball and bunny). Our method (c) addresses these issues and produces a more realistic composite automatically. Background image by Him Sann TR under free Pexels license.

ABSTRACT

Compositing virtual objects into real background images requires one to carefully match the scene's camera parameters, surface geometry, textures, and lighting to obtain plausible renderings. Recent learning approaches have shown many scene properties can be estimated from images, resulting in robust automatic single-image compositing systems, but many challenges remain. In particular, interactions between real and synthetic shadows are not handled gracefully by existing methods, which typically assume a shadow-free background. As a result, they tend to generate double shadows when the synthetic object's cast shadow overlaps a background shadow, and ignore shadows from the background that should be cast onto the synthetic object. In this paper, we present a compositing method for outdoor scenes that addresses these issues and produces realistic cast shadows. This requires identifying existing shadows, including soft shadow boundaries, then reasoning about the ambiguity of unknown ground albedo and scene lighting to match the color and intensity of shaded areas. Using supervision from shadow removal and detection datasets, we propose a generative adversarial pipeline and improved composition equations that simultaneously handle both shadow interaction scenarios. We evaluate our method on challenging, real outdoor images from multiple distributions and datasets. Quantitative and qualitative comparisons show our approach produces more realistic results than existing alternatives. Our code, datasets, and trained models are publicly available at https://lvsn.github.io/shadowcompositing.

CCS CONCEPTS

• **Computing methodologies** → **Computational photography**; *Neural networks.*

KEYWORDS

compositing, shadows, virtual object insertion, outdoor illumination, generative adversarial networks

ACM Reference Format:
Lucas Valença, Jinsong Zhang, Michaël Gharbi, Yannick Hold-Geoffroy, and Jean-François Lalonde. 2023. Shadow Harmonization for Realistic Compositing. In *SIGGRAPH Asia 2023 Conference Papers (SA Conference Papers '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3610548.3618227

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SA Conference Papers '23, December 12–15, 2023, Sydney, NSW, Australia
© 2023 Association for Computing Machinery.
ACM ISBN 979-8-4007-0315-7/23/12…$15.00
https://doi.org/10.1145/3610548.3618227

1 INTRODUCTION

Compositing virtual objects into real photographs is a routine task in applications ranging from advertising to augmented reality and visual effects. Despite significant progress in automatic scene reconstruction and illumination estimation, realistic composites still require significant manual work, because fully-automated solutions (e.g., [Wang et al. 2022]) often lead to mismatching shadows, the most salient breach to the illusion of realism.

We present a method that harmonizes synthetically rendered objects and their shadows with the real shadows in a background image, to automatically create visually plausible composites for outdoor scenes. We specifically target two kinds of errors that plague most previous work and stem from disregarding existing shadows in the background image. First, shadows cast by a virtual object that overlaps with a background shadow should blend seamlessly with the existing shadow (fig. 1c), rather than create an unrealistic "double shadow", i.e., a non-physical overlap of multiple cast shadows, causing an over-darkening of the background (fig. 1b,e). Second, when the virtual object is composited in a shaded area of the background, it should receive shadows cast by (possibly out-of-frame) occluders from the background scene. Most methods forego these received shadows, which leads to an implausible, overly bright appearance for the composited object (fig. 1b,f). Our algorithm handles both these *object-to-scene* and *scene-to-object* shadow interactions. Additionally, if spatially-varying ground truth scene illumination is not available for differential rendering [Debevec 1998], shadow colors may look inaccurate (fig. 1d). Our algorithm automatically harmonizes such shadow colors after insertion.

Our approach assumes the composite is an outdoor scene, illuminated by a direct sunlight and indirect sky. We start by estimating the scene illumination and coarse geometry using state-of-the-art outdoor illumination [Zhang et al. 2019b] and ground plane estimation [Hold-Geoffroy et al. 2018] techniques. We use these estimates to drive the rendering of the synthetic foreground object, which we then compute a rough composite onto the background image using an IBL approach based on Debevec's differential rendering [Debevec 1998] but with adaptations to handle the unknown spatially-varying scene illumination and ground plane BRDF. Our strategy is to correct errors in this rough composite, using an image-space neural network, informed by a new and improved image formation model and additional input maps generated by the renderer.

Specifically, we refine the foreground object's cast shadow using a multiplicative gain map, estimated by our new network, that corrects both color and intensity discrepancies between the rendered shadow and the real shadows in the background, and seamlessly blends overlapping shadows when they occur. This simplifies the network's task, by relieving it from predicting high-frequency image details already present in the input composite, thus regularizing inference. To synthesize plausible shadows *on* the virtual object, we also train the network to estimate shadowed areas in the background image. This gives us a shadow mask, which we back-project from the estimated ground plane onto the object according to the available sun direction. This results in 3D-consistent shadow cues in image-space, similar to the "shadow displacement map" of Chuang et al. [2003]). We train our proposed network with a large variety of scene layouts, illumination conditions, and shadow and occlusion patterns, using a combination of real datasets and synthetic data rendered using a physically-based path tracer [Developers 2023].

Unlike previous works that only focused on the problem of casting virtual shadows onto the background scene [Chuang et al. 2003; Liu et al. 2020; Sheng et al. 2021], our method is, to the best of our knowledge, the first work to holistically tackle the problem of compositing *all* shadows for virtual object compositing. Compared to several baselines, including image-to-image translation and using a state-of-the-art shadow detector, our new image formation model and conditional generation approach lead to much more realistic, artifact-free composites. We demonstrate state-of-the-art compositing results across a wide range of real photographs.

2 RELATED WORK

We limit our discussion to works on virtual object compositing [Debevec 1998; Nakamae et al. 1986] over a single outdoor image, focusing on shadow estimation issues.

Single image outdoor illumination estimation. Compositing a virtual object such that it realistically blends with the background requires an accurate estimate of backdrop's lighting conditions [Carvalho et al. 2015; Zhu et al. 2015]. For outdoor images, Lalonde et al. [2009] first proposed to estimate the sun visibility and direction by extracting cues such as shadows, sky appearance and shading on vertical surfaces like buildings. Later, Hold-Geoffroy et al. [2017] proposed to learn a mapping between a limited field of view, low dynamic range (LDR) outdoor image and its corresponding illumination, as modeled by the physically-based Hošek-Wilkie sky model [Hošek and Wilkie 2012, 2013]. Follow-up works handle both indoor and outdoor environments [LeGendre et al. 2019], use learned sky models [Hold-Geoffroy et al. 2019; Yu et al. 2021], and can estimate spatially-varying lighting representations [Wang et al. 2022; Zhu et al. 2021b]. Our method draws inspiration from Zhang et al. [2019b], who rely on the Lalonde-Matthews (LM) model [Lalonde and Matthews 2014], which decomposes outdoor illumination in two components: sun and sky. Although simpler than physically-based models, the LM sky model is more expressive, and can thus represent more diverse weather conditions.

Shadow detection and removal. Early shadow detection methods were initially based on photometric and geometric cues [Sanin et al. 2012]. Later, machine learning methods used MRFs [Panagopoulos et al. 2011; Zhu et al. 2010], CRFs [Guo et al. 2013] or SVMs [Lalonde et al. 2010] to estimate shadows. More recently, deep learning methods pushed the accuracy of shadow detection. Vicente et al. [2016] advocates using large-scale approximate datasets during training, while automatically correcting ground truth errors, and post-processing the estimation using a patch CNN. Le et al. [2018] introduced a GAN-based approach for shadow detection, improving robustness by training against an adversarial shadow attenuator. Zhu et al. [2018] paired recurrent attention residual modules with a pyramid feature network to attain state-of-the-art shadow detection accuracy. Wang et al. [2021, 2020a] developed a method to detect individual shadow instances. Beyond detection, methods were also proposed to remove shadows end-to-end using CNNs, either by adding semantic cues [Qu et al. 2017], jointly learning detection and removal [Wang et al. 2018], and decomposing the image using

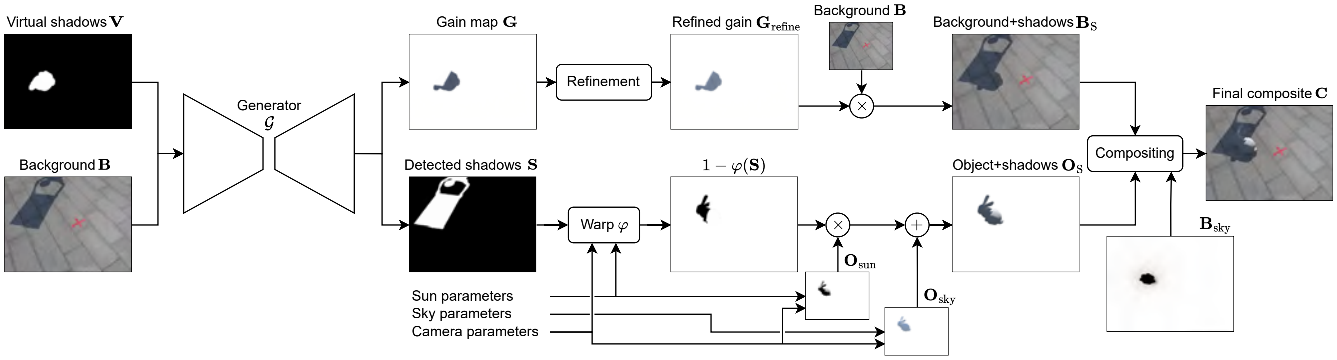

Virtual shadows $\mathbf{V}$
Background $\mathbf{B}$
Generator $\mathcal{G}$
Gain map $\mathbf{G}$
Detected shadows $\mathbf{S}$
Refinement
Warp φ
Refined gain $\mathbf{G}_{\text{refine}}$
$1-\varphi(\mathbf{S})$
Background $\mathbf{B}$
Background+shadows $\mathbf{B}_{\text{S}}$
Object+shadows $\mathbf{O}_{\text{S}}$
$\mathbf{O}_{\text{sun}}$
$\mathbf{O}_{\text{sky}}$
Compositing
Final composite $\mathbf{C}$
$\mathbf{B}_{\text{sky}}$
Sun parameters
Sky parameters
Camera parameters

Figure 2: Overview of our shadow compositing method. We train a generator network that takes as input a background image and a target shadow mask—corresponding to the virtual object shadow region, to produce two outputs: 1) a detected shadow mask, corresponding to the detected shadows in the input background image; and 2) a gain map, to adjust the shadowed regions. These outputs are combined to produce a realistic composite (far right) which realistically blends the rendered shadows with those already present in the scene.

a linear illumination transform [Le and Samaras 2019]. More recently, MTMT [Chen et al. 2020] was proposed for semisupervised shadow detection. This method combines training on unlabeled data with a teacher-student approach to provide state-of-the-art shadow detection accuracy and has become popular due to its robustness to general scenarios and public availability. Since then, many other approaches have also been proposed [Hu et al. 2021; Zhu et al. 2021a, 2022], including a technique to perform shadow removal aided by shadow generation [Liu et al. 2021].

Compositing with deep learning. Deep learning based methods were developed to harmonize composites [Tsai et al. 2017; Zhan et al. 2020], generally by attempting to balance the colors of inserted objects. Recently, GAN-based approaches were developed to determine the ideal location of inserted objects as well as their color [Azadi et al. 2020; Chen and Kae 2019; Lin et al. 2018], but none of these methods deal with cast shadows. Of note, Nicolet et al. [2020] handle shadows but require a multiview image set.

Image relighting and shadow generation. Relighting involves removing, altering and creating new shadows. Facial relighting for example, can enhance a portrait by removing distracting cast shadows on a subject's face, or simulate a more pleasing diffuse illumination [Futschik et al. 2023; Nestmeyer et al. 2020; Pandey et al. 2021; Sun et al. 2019; Wang et al. 2009, 2020b; Zhang et al. 2020; Zhou et al. 2019]. Relighting algorithms for generic scenes typically use multi-view stereo (MVS) to model the scene geometry, using which they estimate existing shadows or render new ones [Duchêne et al. 2015; Philip et al. 2019, 2021]. Relighting approaches have also been demonstrated on single images [Griffiths et al. 2022] and even applied to screen-space shading [Nalbach et al. 2017]. Closest to our work, shadow generation recently emerged as an important relighting sub-problem in the literature [Liu et al. 2020; Sheng et al. 2022, 2021, 2023]. It enables shadow and reflection synthesis directly from 2D composites, without explicit 3D geometry estimation. However, their results are limited to scenarios where the synthetic object is deliberately placed *away* from shadows already present in the background, which sidesteps the problem of virtual and real shadows interaction we seek to solve.

3 BACKGROUND: DIFFERENTIAL IMAGE COMPOSITING

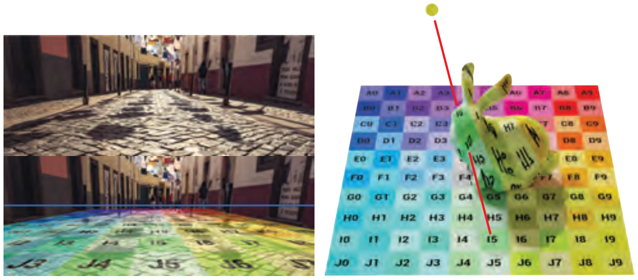
Our goal is to produce a realistic composite of a virtual 3D object onto a background photograph as in fig. 1c. We start by reviewing differential image compositing [Debevec 1998].

To composite a 3D object onto a $h \times w$ color image $\mathbf{B} \in \mathbb{R}^{3hw}$, Debevec [1998] proposed a two-step approach. First, a model of the local scene is constructed (i.e., an approximation of the real scene geometry surrounding the virtual object). This is typically a simple ground plane acting as a shadow receiver. Second, the composite is computed using two renderings of the local scene: one with the virtual object $\mathbf{O} \in \mathbb{R}^{3hw}$, and one without $\mathbf{N} \in \mathbb{R}^{3hw}$. Given a mask $\mathbf{M} \in [0,1]^{hw}$ with 1 indicating the object, the composite $\mathbf{C}$ can be obtained using the differential image rendering equation:

$$\mathbf{C} = \mathbf{M} \cdot \mathbf{O} + (1 - \mathbf{M}) \cdot (\mathbf{B} + c(\mathbf{O} - \mathbf{N})), \qquad (1)$$

where $\cdot$ is the element-wise product of image tensors, and c is a scalar adjusting for the desired shadow strength to compensate for the unknown ground material BRDF (required by Debevec's [1998] original method).

This formulation enables casting a virtual object's shadow onto the background, but it does not model interactions between virtual and real shadows as shown in fig. 1b. This leads to two shortcomings: 1) shadows cast by the virtual object do not blend seamlessly with existing shadows in the background, as even tuning c cannot compensate for the overlap between the virtual and real shadows; and 2) occluders in the background scene (i.e., parts of the scene's geometry whose shadows are visible in the background photo), cannot cast shadows onto the virtual object. Our proposed approach, which we discuss next, addresses these shortcomings and lets us realistically model interactions between virtual and real shadows.



(a) ground plane estimate (b) shadow warping

Figure 3: **Warping ground shadows. Given a horizon line estimate for the background image, we approximate camera extrinsics and a coordinate system for the ground plane (a). We then warp the ground plane, unprojecting pixels along the sun direction onto the virtual 3D object, simulating an occlusion of sun rays (b). Image by Matthias Groeneveld.**

4 SHADOW-AWARE IMAGE COMPOSITING

Our approach has two main goals: 1) matting the shadows cast by the virtual objects with the existing shadows present in the image; and 2) casting shadows onto the virtual object (by modifying shadows cast by real occluders present in the background photo).

Our proposed compositing pipeline achieves this using a generator network that learns to simultaneously: 1) predict a shadow gain map, which we use to blend synthetic shadows with their surroundings; and 2) detect existing shadows on the ground, which can be back-projected to shade the synthetic object.

Like Debevec [1998], our method relies on prior knowledge of the lighting conditions and camera parameters of the input image. As is commonly the case (e.g., [Hošek and Wilkie 2013; Lalonde and Matthews 2014], we assume an outdoor lighting model composed of two light sources: one directional (sun) and one ambient (sky). Additionally, we assume that a ground plane capable of receiving shadows is at least partially visible in the background image.

4.1 Method overview

Our novel compositing pipeline is illustrated in fig. 2. At its core is a generator network $\mathcal{G}$ which accepts as input a low dynamic range RGB background image $\mathbf{B}$ as well as an indicator map of the virtual shadows $\mathbf{V} \in [0,1]^{hw}$ to be matted with the background. The generator outputs a linear-RGB gain map $\mathbf{G} \in [0,1]^{3hw}$ and a single-channel map of the detected soft shadows in the image $\mathbf{S} \in [0,1]^{hw}$ such that $(\mathbf{G}, \mathbf{S}) = \mathcal{G}(\mathbf{B}, \mathbf{V})$. Here, a value of 1 (resp. 0) in $\mathbf{V}$ and $\mathbf{S}$ indicate a fully shadowed (resp. fully sunny) pixel.

The shadow gain $\mathbf{G}$ is then used to blend the virtual shadow with the real shadows already present in the background image (see sec. 4.2, fig. 2 top row) to obtain $\mathbf{B}_\text{S} \in \mathbb{R}^{3hw}$. In addition, the detected shadows are back-projected onto the virtual 3D object to cast real shadows onto the virtual object (see sec. 4.3, fig. 2 bottom row) to obtain $\mathbf{O}_\text{S} \in \mathbb{R}^{3hw}$. One last render of the virtual scene's shadow catcher is performed with the sky-only illumination $\mathbf{B}_\text{sky}$, to capture shading effects such as ambient occlusion onto the ground (caused by the virtual object), further darkening ground regions close to the object. Finally, these intermediate images are merged to obtain the final composite

$$\mathbf{C} = \mathbf{M} \cdot \mathbf{O}_\text{S} + (1 - \mathbf{M}) \cdot (\mathbf{B}_\text{S} \cdot \mathbf{B}_\text{sky}) \,. \tag{2}$$

This equation effectively replaces the user-defined shadow intensity gain c from eq. (1) with a spatially-varying shadow map, which our model generates without user interaction.

4.2 Blending virtual shadows with the background image

Ideally, we could obtain the background with blended shadows by simply applying our predicted gain $\mathbf{G}$ to the input background image, i.e., $\mathbf{B}_\text{S} = \mathbf{B} \cdot \mathbf{G}$. However, we observed that a single forward pass through $\mathcal{G}$ still often results in a mismatch in overall shadow intensity. We attribute this to the fact that for the results to "look right", the network must produce pixel-perfect shadows, i.e., even subtle deviations are immediately visible to the human eye. To alleviate this issue, we use a refinement procedure to adjust the predicted gain map, as illustrated in fig. 2 (top row).

Gain map refinement. We correct the gain map $\mathbf{G}$ using a global scale factor $f \in \mathbb{R}$, so that the average intensity of the generated shadows matches that of the background shadows on the region where they overlap:

$$\mathbf{G}_\text{refine} = (1 - \mathbf{S}) \cdot f\mathbf{G} + \mathbf{S} \cdot \mathbf{G} \,, \tag{3}$$

$$f = \frac{\mu(\mathbf{S} \cdot \mathbf{V} \cdot \mathbf{B})}{\mu((1 - \mathbf{S}) \cdot \mathbf{G} \cdot \mathbf{V} \cdot \mathbf{B}) + \epsilon} \,, \tag{4}$$

where $\mu(\cdot)$ denotes the mean operator over non-zero pixels of the *value channel* after conversion from RGB to HSV, and ϵ is a small value to prevent division by zero. Pixels with intensity below 0.1 are discarded in eq. (4) to avoid noise.

Our final background with blended shadows is thus given by $\mathbf{B}_\text{S} = \mathbf{B} \cdot \mathbf{G}_\text{refine}$. Alternatively, this ratio could also be computed on each color channel separately to account for color mismatches, in which case $f \in \mathbb{R}^3$. A similar procedure was used to perform shadow removal [Le and Samaras 2022], relighting [Griffiths et al. 2022], and intrinsic image decomposition [Duchêne et al. 2015].

4.3 Casting real shadows on the virtual object

In order to cast real shadows on the virtual object, it would be intuitive to train a network on the appearance of shaded objects. However, such a dataset would be very hard to capture. To bridge the reality gap, we instead leverage shadow removal data (see sec. 5.3) on a ground-only formulation. In this case, to obtain cast shadows, we first warp the detected shadows $\mathbf{S}$ onto the 3D object according to the sun direction (obtained from the lighting model) and the ground plane equation (obtained from the camera parameters). We obtain the shadow warping operator φ by following Chuang et al. [2003], illustrated in fig. 3, to obtain the warped shadow map $\varphi(\mathbf{S})$. Real shadows are then projected onto the object using

$$\mathbf{O}_\text{S} = \mathbf{O}_\text{sun} \cdot (1 - \varphi(\mathbf{S})) + \mathbf{O}_\text{sky} \,, \tag{5}$$

where $\mathbf{O}_\text{sun}$ (resp. $\mathbf{O}_\text{sky}$) are renderings of the virtual object using the sun (resp. sky) lighting model only (see fig. 2, bottom row).

It is worth noting that in our current formulation only the ground plane is used, so cast shadows do not get darker as they get closer

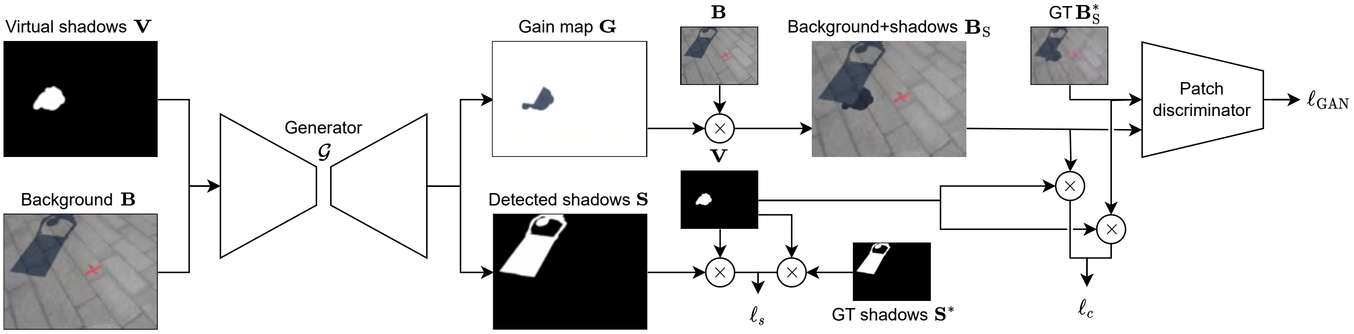

Virtual shadows $\mathbf{V}$
Background $\mathbf{B}$
Generator $\mathcal{G}$
Gain map $\mathbf{G}$
Detected shadows $\mathbf{S}$
$\mathbf{B}$
Background+shadows $\mathbf{B}_\mathrm{S}$
GT $\mathbf{B}^*_\mathrm{S}$
Patch discriminator
ℓ_GAN
$\mathbf{V}$
ℓ_s
GT shadows $\mathbf{S}^*$
ℓ_c

Figure 4: Overview of our training procedure. This formulation allows us to train our approach on real images from existing shadow removal datasets. Our network is trained using a combination of 3 loss functions (ℓ_*, see text).

to (unknown) scene geometry (see fig. 6, last row). On the known geometry of the ground plane, this effect is achieved through $\mathbf{B}_\mathrm{sky}$ (c.f. sec. 4.1).

5 TRAINING THE SHADOW GENERATOR

We now describe our model architecture and training procedure.

5.1 Network architecture

For the generator network, we use a UNet [Ronneberger et al. 2015] with fixed update initialization [Zhang et al. 2019a] (implementation from [Griffiths et al. 2022]) and a simple patch discriminator $\mathcal{D}$ [Isola et al. 2017]. The network takes as input a 4-channel image concatenating the virtual shadow mask $\mathbf{V}$ and the background $\mathbf{B}$. It produces a 4-channel output representing the shadow gain $\mathbf{G}$ and the mask of detected background shadows S. All images are processed with patches of 128 × 128 spatial resolution.

5.2 Training objective

As illustrated in fig. 4, we train our model to minimize the sum of three loss functions: $\ell_s + \ell_c + \ell_\mathrm{GAN}$. First, ℓ_s compares the detected shadows S to a ground truth background shadow mask S^*:

$$\ell_s = ||\mathbf{V} \cdot (\mathrm{S} - \mathrm{S}^*)||_1. \quad (6)$$

Both shadows are masked by the virtual shadow mask $\mathbf{V}$ because, to handle interactions, we only need to accurately detect real shadows which overlap with the virtual ones. The second loss, ℓ_c, compares the resulting matting $\mathbf{B}_\mathrm{S}$ with the ground truth shaded image $\mathbf{B}^*_\mathrm{S}$:

$$\ell_c = ||\mathbf{V} \cdot (\mathbf{B}_\mathrm{S} - \mathbf{B}^*_\mathrm{S})||_1 \, . \quad (7)$$

As before, images are masked by the virtual shadow mask $\mathbf{V}$. Lastly, we use a conventional GAN loss ℓ_GAN on $\mathbf{B}_\mathrm{S}$ and $\mathbf{B}^*_\mathrm{S}$. The generator and discriminator are trained simultaneously.

5.3 Datasets

We leverage different datasets including a mix of synthetic and real data to train our model. For all previously published datasets, we employ the provided train/test splits. During training, 10% of the training set is further separated as a validation set. The test sets were only used for benchmarking the final models (as seen in tab. 1 and tab. 2). We also augment data by flipping the images horizontally.

Shadow removal. We leverage shadow removal datasets to train our shadow generation network, as illustrated in fig. 5. These datasets are composed of pairs of images with and without shadows $\mathbf{I}_\mathrm{w}$ and $\mathbf{I}_\mathrm{wo}$ resp., combined with a shadow mask $\mathbf{I}_\mathrm{m}$. We generate the background $\mathbf{B}$ and virtual shadow $\mathbf{V}$ images by generating a binary pixel mask $\mathbf{Z}$, consisting of a random subset of the shadow region in $\mathbf{I}_\mathrm{w}$. This random subset is obtained by subtracting two randomly-generated masks, each containing 15 overlapping ellipsoids with random translation, rotation, and scale. These masks are subtracted one from the other and the result is blurred with a Gaussian kernel, then masked by $\mathbf{I}_\mathrm{m}$. The resulting mask $\mathbf{Z}$ is then used to generate $\mathbf{B} = \mathbf{Z} \cdot \mathbf{I}_\mathrm{w} + (1 - \mathbf{Z})\mathbf{I}_\mathrm{wo}$, and we set $\mathbf{V} = \mathbf{I}_\mathrm{m}$.

During training, these augmentations are made online. In total, we randomly mixed: the originals (the task of full shadow generation, enforcing shadow color), the augmented subset of shadows (as described above, for the task of shadow matting), border-less subsets (the original mask eroded with a square kernel then blurred with a Gaussian kernel, to emphasize soft edge generation), shadow edges (the inverse of a border-less mask, limited to the original shaded mask, for matting the entire contour), no-insertions (a blank mask with a shadow-less input and target, to punish insertions out of the desired region), and no-changes (a subset mask and a fully-shaded input and target, to further punish double shadows). Examples can be found in the supplemental.

Specifically, we use the following datasets: ISTD [Wang et al. 2018] (adjusted by [Le and Samaras 2019]), DESOBA [Hong et al. 2022], and SRD [Qu et al. 2017] (binary shadow masks from Cun et al. [2020], which we refined with a median filter).

Shadow detection. We also leverage the SBU [Vicente et al. 2016] and UCF [Zhu et al. 2010] shadow detection datasets. These datasets contain images with labeled shadow regions, but no non-shadow counterparts, so they cannot be used for matting evaluations. Nonetheless, SBU's over 4000 samples and UCF's roughly 100 provide great camera and scene variation, which we used to train for shadow detection and punish double shadows. To do this, we generate network input masks as for the shadow removal datasets and use the shaded images as input and target.

Synthetic soft shadows dataset. Even though the approaches described above allow us to train our network using real data, it remains the case that high-quality shadow matting data samples

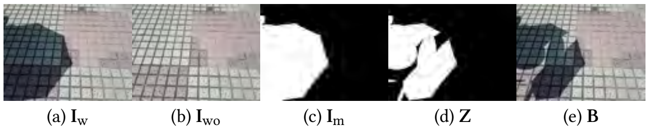

(a) $\mathrm{I_w}$ (b) $\mathrm{I_{wo}}$ (c) $\mathrm{I_m}$ (d) $\mathbf{Z}$ (e) $\mathbf{B}$

Figure 5: Adapting a real shadow removal sample for shadow compositing. From input images (a) with shadows $\mathrm{I_w}$, (b) without shadows $\mathrm{I_{wo}}$ and (c) mask $\mathrm{I_m}$ available in shadow removal datasets, we generate (d) a random mask Z which is used to create (e) a background image B that can be used for training. The virtual shadows image is simply $\mathrm{V} = \mathrm{I_m}$. This example is taken from the ISTD dataset [Wang et al. 2018].

such as these are very few (the only datasets with truly accurate, masked shadow and shadow-less pair annotations are DESOBA and ISTD, totaling a little over 2000 training samples). Therefore, we supplement our model with 3000 samples of synthetic data, introducing a new dataset of synthetically-generated composite images. To generate this dataset, we follow a procedure similar to [Zhu et al. 2021b] and leverage the Blender SceneCity plugin [Couturier 2023] which generates high-quality urban scenarios. To augment the diversity in ground textures, we randomly sample physics-based textures from Polyhaven [Tuytel et al. 2023]. A Stanford bunny object is inserted in the camera field of view with random rotation, translation, and scale, creating a wide variety of shadow patterns when combined with different sun directions. The scene is illuminated by a random HDR environment map from the Laval Outdoor HDR Dataset [Hold-Geoffroy et al. 2019], which contains synthetic annotations for Lalonde-Matthews [2014] sun and sky parameters. A random amount of flying occluders, in the form of multiple geometric primitives, is rendered outside of the camera field of view, in random distances along the line between the object and the sun. The occluders are rotated at random and scaled so as to not completely cover the object. This creates multiple soft shadow interactions, unique to every frame. The cast shadow ground truths are obtained through a shadow catcher plane, rendered after disabling the sky element in the LM HDRI, leaving the sun as the only illuminant. For this dataset, flipping was the only augmentation performed.

To our knowledge, our real-augmented and synthetic datasets are the only ones to contain soft shadow annotations, as well as shadow intersections and overlaps. All our augmented, adjusted, and generated datasets, annotations, and augmentation code are publicly available. Samples are also available in the supplemental.

5.4 Implementation details

Handling multiple resolutions. Training GANs on higher resolutions may lead to instabilities and incur both prohibitive training times and memory requirements. However, to insert virtual objects convincingly we need considerable resolution—a common limitation for deep learning approaches. We circumvent this issue by using a sliding window approach, dubbed *patch-based local averaging*, over high-resolution images. The network $\mathcal{G}$ is executed on 128×128 patches, and the final output for each pixel is the average over that pixel's estimation value in all overlapping patches. To evenly sample all pixels, we pad the image with 128 pixels on each border, using a *reflect* pattern. While this approach scales linearly with the number of pixels to be detected, the patches are independent, so the algorithm can be sped up by batching the patches on the GPU (e.g. we use a batch size of 256 on a 12GB RAM NVIDIA RTX 2080 Ti GPU, with an inference time of approximately 20ms per batch). When doing so, we obtain shadow generation results at resolutions over full HD (1920×1080), whereas state-of-the-art shadow detection tends to shrink the image to fit the network, resulting in loss of detail (see fig. 7). We further propose to speed up the algorithm by adding a stride to the patch-averaging sliding window. We have found a stride of 16 to obtain sufficient shadows while reducing the computing time from minutes to seconds per image.

Lighting and camera parameters. We use off-the-shelf automatic algorithms to estimate lighting (Zhang et al. [2019b]) and camera parameters (Hold-Geoffroy et al. [2018]) from the background image.

6 EVALUATION

6.1 Compositing

Qualitative results. Fig. 6 shows qualitative compositing results compared against traditional compositing [Debevec 1998] and a strong baseline based on the state-of-the-art MTMT shadow detector [Chen et al. 2020]. This baseline utilizes the MTMT detection $\mathrm{S_{MT}}$, creates a gain map $\mathbf{G}_{\mathrm{MT}} = \mathrm{S_{MT}}/2 + 1/2$, then uses the same refinement, warping, and compositing as our main approach.

We observe that traditional compositing exhibits issues when virtual and real shadows should interact. In contrast, the MTMT baseline produces much improved results, thanks to our compositing equations (sec. 4), but still yields visible shadow boundaries since it was not trained to take the penumbra into account. Our results (rightmost column) show better overall matting and fewer visible shadow seams. Moreover, our gain map preserves texture details more accurately than the baseline plain multiplier, e.g., the darker lines on brick textures (fig. 6, last row).

Quantitative results. To compare our method, we extend our shadow removal augmentation procedure (see fig. 5) to the test sets of DESOBA and ISTD. An example of a resulting augmented sample (from the test set) can be seen in fig. 5. All augmented inputs must have at least 20% of the original shadow and no more than 80%, to guarantee blending. On DESOBA, 9 samples were discarded for the total shadows covering less than 0.5% of the image. As for SRD, the masks are not meant for evaluations, as the considerable inaccuracies would overly punish correct detections and mattings (more details in the supplemental). This evaluation includes two additional baselines: a variation of our model that aims to generate the RGB ground composite directly ("comp. net") and a variation that takes MTMT detection as input and tries to estimate just the gain map to blend the shadows ("gain net"), as opposed to learning the tasks jointly ("ours"). We report the SSIM, PSNR, and L1 metrics.

Tab. 1 shows that the refinement process benefits all approaches and confirms that learning joint detection and matting is the most promising path. Due to the lower resolution and overall larger shadow areas in ISTD, our model without local averaging performs best. However, for the larger images in DESOBA, our local averaging configuration outperforms every other approach by a considerable

Table 1: Quantitative results for shadow compositing. We quantitatively evaluate our method on shadow renders on SSIM, PSNR (dB), and L1 error. Our method is compared to 3 strong baselines ("MTMT", "comp. net" and "gain net"), see text. Variants on each method are also evaluated. Colors indicate best , medium and worst performance for each method independently, and bold-underlined the best performance across all.

	ISTD [2018]			DESOBA [2022]			averages		
	SSIM$_\uparrow$	PSNR$_\uparrow$	L1$_\downarrow$	SSIM$_\uparrow$	PSNR$_\uparrow$	L1$_\downarrow$	SSIM$_\uparrow$	PSNR$_\uparrow$	L1$_\downarrow$
MTMT	0.925	27.596	6.505	0.979	32.184	1.569	0.952	29.890	4.037
+ crf	0.925	27.582	6.457	0.979	32.148	1.564	0.952	29.865	4.010
+ val. scale	0.925	29.523	5.436	0.981	34.434	0.940	0.953	31.978	3.188
+ rgb scale	0.925	30.120	5.173	0.981	34.402	0.923	0.953	32.261	3.048
comp. net	0.873	23.459	10.054	0.967	32.061	1.703	0.920	27.760	5.878
+ pbla	0.925	29.031	5.816	0.981	34.564	1.149	0.953	31.797	3.482
+ val. scale	0.928	30.047	5.364	0.982	35.310	0.965	0.955	32.679	3.165
+ rgb scale	0.928	30.229	5.284	0.982	35.304	0.960	0.955	32.767	3.122
gain net	0.924	29.263	5.695	0.982	34.351	1.007	0.953	31.807	3.351
+ pbla	0.920	28.237	6.086	0.980	33.931	1.094	0.950	31.084	3.590
+ val. scale	0.924	29.682	5.436	0.981	34.297	0.935	0.953	31.990	3.185
+ rgb scale	0.925	29.923	5.318	0.981	34.325	0.916	0.953	32.124	3.117
ours	**0.930**	**30.342**	**5.008**	0.982	34.784	0.994	**0.956**	32.563	3.001
+ pbla	0.928	29.857	5.347	0.982	35.074	0.958	0.955	32.465	3.152
+ val. scale	0.929	30.220	5.226	**0.983**	35.707	0.818	**0.956**	32.964	3.022
+ rgb scale	0.929	30.321	5.166	**0.983**	**35.720**	**0.804**	**0.956**	**33.020**	**2.985**

Table 2: Quantitative results for shadow detection. We compare our method (without and with the patch-based local averaging, *pbla*) to the MTMT baseline (without and with the CRF refinement) on two datasets: ISTD [Wang et al. 2018] and DESOBA [Hong et al. 2022], the last set of columns shows the average over the two datasets. We report the precision on shadow (S) and non-shadow (NS) regions as well as the BER metric. Colors indicate best and worst performance for each method independently, and bold-underlined the best performance across all.

	ISTD [2018]			DESOBA [2022]			averages		
	S$_\downarrow$	NS$_\downarrow$	BER$_\downarrow$	S$_\downarrow$	NS$_\downarrow$	BER$_\downarrow$	S$_\downarrow$	NS$_\downarrow$	BER$_\downarrow$
MTMT	12.114	0.162	6.138	17.676	0.066	8.871	14.895	0.114	7.505
+ crf	13.360	**0.119**	6.740	21.136	**0.056**	10.596	17.248	**0.088**	8.668
ours	**3.742**	0.671	**2.207**	9.273	0.207	4.740	6.508	0.439	3.473
+ pbla	7.718	0.389	4.053	**4.785**	0.143	**2.464**	**6.251**	0.266	**3.258**

margin. It is worth noting that despite the large amount of ground texture variation present in DESOBA, the RGB scaling refinement step (eq. (4)) outperforms other approaches in most cases.

Further validation can be found in the supplemental, including per-loss and per-dataset ablations, cross-dataset validations, and samples of softer shadow detection and overcast scenarios.

6.2 Shadow detection

Our compositing model needs a predicted mask of the background shadows, which effectively gives us a shadow detector as a byproduct of our training strategy. While generic shadow detection is complex and out of our scope, our network is trained to perceive detailed shadows (including soft edges) on a wide variety of outdoor ground plane textures, so we evaluate this capability here.

Fig. 7 shows that our method tends to produce much more finer-grained results than the current state-of-the-art in shadow detection [Chen et al. 2020], correctly identifying several small regions in the shade. However, regions outside the ground plane (e.g., the sky) are often misclassified since those receive little to no supervision signal during training. This is not a problem for our use case, but limits the applications of our soft shadow detector. Quantitatively, tab. 2 shows our method outperforms the MTMT baseline significantly in the true shadow regions ("S") and on the Balanced Error Rate (BER) metric while assigning slightly more false positives, i.e., fewer true negatives ("NS"). This evaluation was performed on the same images as in sec. 6.1, measuring how well the shadows were detected so that matting could be performed. This result from MTMT is expected because shadow detection datasets rarely have such a level of detail in their annotations, and none possess soft shadow edge annotations. Moreover, MTMT's resizing operations to the network resolution (416×416) further decrease detail. Traditional detection also emphasizes connected bodies of shadow to reduce noise and better match the simpler dataset annotations, at the expense of detail (e.g., CRF post-processing in MTMT). We also observe the patch-based local averaging ("pbla", see sec. 5.4) having a mixed impact on shadow detection. However, it allows the method to be applied to high-resolution images.

6.3 Limitations

Although it proposes a substantial improvement in shadow compositing over previous techniques, our method still bears some limitations. Soft shadow detection (penumbra estimation) and ground reflectance estimation are very ill-posed problems to solve from a single low dynamic range outdoor image with completely unknown lighting, geometry, material, and camera properties. Notably, our method expects mostly diffuse ground planes. More complex geometries or specular materials would likely not be handled well. The same applies to glass objects and caustics, which could result in incoherent matting. Additionally, shadows back-projected onto the object using the warp operator φ might be abruptly cut if they extend beyond the edges of the background image. This could be resolved using in-painting networks to extend the image. Another potential weakness is that our shadow gain map refinement step assumes the shadow intersection region to be reliable, but this might not hold in rarer situations where the ground texture changes abruptly. Finally, the visual quality of our results depends on the camera and lighting estimation techniques we receive as input (though they can be replaced by manual inputs).

7 CONCLUSION

In this paper, we introduce a deep learning based method that performs virtual object compositing in a real background from a single image, focusing on complex shadow interactions in outdoor scenes. In our method, we consider both the shadows cast by the virtual object onto the real scene and vice-versa. Our key insight is to leverage a neural network to estimate spatially-varying corrective maps for newly-proposed compositing equations. Namely, we estimate soft shadow edges by jointly learning detection and matting, also leveraging image queues to compensate for the missing knowledge of the ground's reflectance or the scene's indirect lighting, all aspects which determine the shadow color. Another key point of our approach is to recognize that the sun, as the main light source, can be used to warp the shadow detection into a direct light mask for the virtual object, shading it instead. In addition, our method can be applied as-is at large resolutions, obtaining finer-grained ground shadows than the state-of-the-art. Our approach circumvents the lack of available data for training by re-purposing and augmenting existing shadow removal and detection datasets, as well as our own synthetic dataset, resulting in large amounts of automatically-generated, detailed annotated samples—both real and synthetic. Though no existing approaches propose to solve this specific problem, we have shown through several experiments that our proposed method performs the proposed task with greater success than baselines based on existing methods. We hope our method can help pave the way to more pleasant virtual compositing and more immersive AR experiences. In the future, we hope our method can be extended to video inference to increase its accuracy and temporal consistency. We also hope it can be expanded to different types of challenging ground materials, non-planar geometries, and more complex light effects such as caustics and indoor scenes. Finally, we hope our GAN-based shadow detection through matting and our re-purposing of scarce real data can serve as inspiration for future works on shadow detection, removal, and matting.

ACKNOWLEDGMENTS

This work was supported by NSERC grants ALLRP 557208-20, RGPIN-2020-04799, Adobe, and the Digital Research Alliance Canada. The authors thank NVIDIA for donating GPUs. We also thank Kalyan Sunkavalli for his insights in early phases of the project.

REFERENCES

Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. 2020. Compositional GAN: Learning Image-Conditional Binary Composition. *Int. J. Comput. Vis.* 128, 10 (2020), 2570–2585. https://doi.org/10.1007/s11263-020-01336-9

Tiago J. Carvalho, Hany Farid, and Eric Kee. 2015. Exposing photo manipulation from user-guided 3D lighting analysis. In *Media Watermarking, Security, and Forensics (SPIE Proceedings, Vol. 9409)*, Adnan M. Alattar, Nasir D. Memon, and Chad Heitzenrater (Eds.). SPIE, San Francisco, CA, USA, 940902. https://doi.org/10.1117/12.2075544

Bor-Chun Chen and Andrew Kae. 2019. Toward Realistic Image Compositing With Adversarial Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Vision Foundation / IEEE, 2019, Long Beach, CA, USA, 8415–8424. https://doi.org/10.1109/CVPR.2019.00861

Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. 2020. A Multi-Task Mean Teacher for Semi-Supervised Shadow Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Vision Foundation / IEEE, Seattle, WA, USA, 5610–5619. https://doi.org/10.1109/CVPR42600.2020.00565

Yung-Yu Chuang, Dan B. Goldman, Brian Curless, David Salesin, and Richard Szeliski. 2003. Shadow matting and compositing. *ACM Trans. Graph.* 22, 3 (2003), 494–500. https://doi.org/10.1145/882262.882298

Arnaud Couturier. 2023. SceneCity blender plugin. https://www.cgchan.com/store/scenecity.

Xiaodong Cun, Chi-Man Pun, and Cheng Shi. 2020. Towards Ghost-Free Shadow Removal via Dual Hierarchical Aggregation Network and Shadow Matting GAN. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*. AAAI Press, New York, NY, USA, 10680–10687. https://doi.org/10.1609/aaai.v34i07.6695

Paul E. Debevec. 1998. Rendering Synthetic Objects into Real Scenes: Bridging Traditional and Image-based Graphics with Global Illumination and High Dynamic Range Photography. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH*, Steve Cunningham, Walt Bransford, and Michael F. Cohen (Eds.). ACM, Orlando, FL, USA, 189–198. https://doi.org/10.1145/280814.280864

Cycles Developers. 2023. Cycles Open Source Production Rendering. https://www.cycles-renderer.org.

Sylvain Duchêne, Clément Riant, Gaurav Chaurasia, Jorge Lopez-Moreno, Pierre-Yves Laffont, Stefan Popov, Adrien Bousseau, and George Drettakis. 2015. Multiview Intrinsic Images of Outdoors Scenes with an Application to Relighting. *ACM Trans. Graph.* 34, 5 (2015), 164:1–164:16. https://doi.org/10.1145/2756549

David Futschik, Kelvin Ritland, James Vecore, Sean Fanello, Sergio Orts-Escolano, Brian Curless, Daniel Sýkora, and Rohit Pandey. 2023. Controllable Light Diffusion for Portraits. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE, Vancouver, BC, Canada, 8412–8421. https://doi.org/10.1109/CVPR52729.2023.00813

David Griffiths, Tobias Ritschel, and Julien Philip. 2022. OutCast: Outdoor Single-image Relighting with Cast Shadows. *Comput. Graph. Forum* 41, 2 (2022), 179–193. https://doi.org/10.1111/cgf.14467

Ruiqi Guo, Qieyun Dai, and Derek Hoiem. 2013. Paired Regions for Shadow Detection and Removal. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 12 (2013), 2956–2967. https://doi.org/10.1109/TPAMI.2012.214

Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. 2019. Deep Sky Modeling for Single Image Outdoor Lighting Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Vision Foundation / IEEE, Long Beach, CA, USA, 6927–6935. https://doi.org/10.1109/CVPR.2019.00709

Yannick Hold-Geoffroy, Kalyan Sunkavalli, Jonathan Eisenmann, Matt Fisher, Emiliano Gambaretto, Sunil Hadap, and Jean-François Lalonde. 2018. A Perceptual Measure for Deep Single Image Camera Calibration. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Vision Foundation / IEEE Computer Society, Salt Lake City, UT, USA, 2354–2363. https://doi.org/10.1109/CVPR.2018.00250

Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. 2017. Deep Outdoor Illumination Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, Honolulu, HI, USA, 2373–2382. https://doi.org/10.1109/CVPR.2017.255

Yan Hong, Li Niu, and Jianfu Zhang. 2022. Shadow Generation for Composite Image in Real-World Scenes. In *Thirty-Sixth AAAI Conference on Artificial Intelligence*. AAAI Press, Virtual Event, 914–922. https://doi.org/10.1609/aaai.v36i1.19974

Lukáš Hošek and Alexander Wilkie. 2012. An analytic model for full spectral skydome radiance. *ACM Trans. Graph.* 31, 4 (2012), 95:1–95:9. https://doi.org/10.1145/2185520.2185591

Lukáš Hošek and Alexander Wilkie. 2013. Adding a Solar-Radiance Function to the Hošek-Wilkie Skylight Model. *IEEE Computer Graphics and Applications* 33, 3 (2013), 44–52. https://doi.org/10.1109/MCG.2013.18

Xiaowei Hu, Tianyu Wang, Chi-Wing Fu, Yitong Jiang, Qiong Wang, and Pheng-Ann Heng. 2021. Revisiting Shadow Detection: A New Benchmark Dataset for Complex World. *IEEE Trans. Image Process.* 30 (2021), 1925–1934. https://doi.org/10.1109/TIP.2021.3049331

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE Computer Society, Honolulu, HI, USA, 5967–5976. https://doi.org/10.1109/CVPR.2017.632

Jean-François Lalonde, Alexei A. Efros, and Srinivasa G. Narasimhan. 2009. Estimating natural illumination from a single outdoor image. In *IEEE 12th International Conference on Computer Vision, ICCV*. IEEE Computer Society, Kyoto, Japan, 183–190. https://doi.org/10.1109/ICCV.2009.5459163

Jean-François Lalonde, Alexei A. Efros, and Srinivasa G. Narasimhan. 2010. Detecting Ground Shadows in Outdoor Consumer Photographs. In *11th European Conference on Computer Vision, ECCV Proceedings, Part II (Lecture Notes in Computer Science, Vol. 6312)*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.). Springer, Heraklion, Crete, Greece, 322–335. https://doi.org/10.1007/978-3-642-15552-9_24

Jean-François Lalonde and Iain A. Matthews. 2014. Lighting Estimation in Outdoor Image Collections. In *2nd International Conference on 3D Vision, 3DV, Volume 1*. IEEE Computer Society, Tokyo, Japan, 131–138. https://doi.org/10.1109/3DV.2014.112

Hieu Le and Dimitris Samaras. 2022. Physics-Based Shadow Image Decomposition for Shadow Removal. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 12 (2022), 9088–9101. https://doi.org/10.1109/TPAMI.2021.3124934

Hieu M. Le and Dimitris Samaras. 2019. Shadow Removal via Shadow Image Decomposition. In *IEEE/CVF International Conference on Computer Vision, ICCV*. IEEE, Seoul, Korea (South), 8577–8586. https://doi.org/10.1109/ICCV.2019.00867

Hieu M. Le, Tomas F. Yago Vicente, Vu Nguyen, Minh Hoai, and Dimitris Samaras. 2018. A+D Net: Training a Shadow Detector with Adversarial Shadow Attenuation. In *15th European Conference on Computer Vision, ECCV Proceedings, Part II (Lecture Notes in Computer Science, Vol. 11206)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer, Munich, Germany, 680–696. https://doi.org/10.1007/978-3-030-01216-8_41

Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul E. Debevec. 2019. DeepLight: Learning Illumination for Unconstrained Mobile Mixed Reality. In *IEEE Conference on Computer Vision and Pattern*

Recognition, CVPR. Computer Vision Foundation / IEEE, Long Beach, CA, USA, 5918–5928. https://doi.org/10.1109/CVPR.2019.00607

Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. 2018. ST-GAN: Spatial Transformer Generative Adversarial Networks for Image Compositing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR.* Computer Vision Foundation / IEEE Computer Society, Salt Lake City, UT, USA, 9455–9464. https://doi.org/10.1109/CVPR.2018.00985

Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. 2020. ARShadowGAN: Shadow Generative Adversarial Network for Augmented Reality in Single Light Scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR.* Computer Vision Foundation / IEEE, Seattle, WA, USA, 8136–8145. https://doi.org/10.1109/CVPR42600.2020.00816

Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. 2021. From Shadow Generation To Shadow Removal. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR.* Computer Vision Foundation / IEEE, Virtual, 4927–4936. https://doi.org/10.1109/CVPR46437.2021.00489

Eihachiro Nakamae, Koichi Harada, Takao Ishizaki, and Tomoyuki Nishita. 1986. A montage method: the overlaying of the computer generated images onto a background photograph. In *Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH*, David C. Evans and Russell J. Athay (Eds.). ACM, Dallas, Texas, USA, 207–214. https://doi.org/10.1145/15922.15909

Oliver Nalbach, Elena Arabadzhiyska, Dushyant Mehta, Hans-Peter Seidel, and Tobias Ritschel. 2017. Deep Shading: Convolutional Neural Networks for Screen Space Shading. *Comput. Graph. Forum* 36, 4 (2017), 65–78. https://doi.org/10.1111/cgf.13225

Thomas Nestmeyer, Jean-François Lalonde, Iain A. Matthews, and Andreas M. Lehrmann. 2020. Learning Physics-Guided Face Relighting Under Directional Light. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR.* Computer Vision Foundation / IEEE, Seattle, WA, USA, 5123–5132. https://doi.org/10.1109/CVPR42600.2020.00517

Baptiste Nicolet, Julien Philip, and George Drettakis. 2020. Repurposing a Relighting Network for Realistic Compositions of Captured Scenes. In *Symposium on Interactive 3D Graphics and Games, I3D*, Dan Casas, Eric Haines, Sheldon Andrews, Natalya Tatarchuk, and Zdravko Velinov (Eds.). ACM, San Francisco, CA, USA, 4:1–4:9. https://doi.org/10.1145/3384382.3384523

Alexandros Panagopoulos, Chaohui Wang, Dimitris Samaras, and Nikos Paragios. 2011. Illumination estimation and cast shadow detection through a higher-order graphical model. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR.* IEEE Computer Society, Colorado Springs, CO, USA, 673–680. https://doi.org/10.1109/CVPR.2011.5995585

Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Häne, Sofien Bouaziz, Christoph Rhemann, Paul E. Debevec, and Sean Ryan Fanello. 2021. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.* 40, 4 (2021), 43:1–43:21. https://doi.org/10.1145/3450626.3459872

Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei A. Efros, and George Drettakis. 2019. Multi-view relighting using a geometry-aware network. *ACM Trans. Graph.* 38, 4 (2019), 78:1–78:14. https://doi.org/10.1145/3306346.3323013

Julien Philip, Sébastien Morgenthaler, Michaël Gharbi, and George Drettakis. 2021. Free-viewpoint Indoor Neural Relighting from Multi-view Stereo. *ACM Trans. Graph.* 40, 5 (2021), 194:1–194:18. https://doi.org/10.1145/3469842

Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W. H. Lau. 2017. DeshadowNet: A Multi-context Embedding Deep Network for Shadow Removal. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR.* IEEE Computer Society, Honolulu, HI, USA, 2308–2316. https://doi.org/10.1109/CVPR.2017.248

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *18th International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI Proceedings, Part III (Lecture Notes in Computer Science, Vol. 9351)*, Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi (Eds.). Springer, Munich, Germany, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Andres Sanin, Conrad Sanderson, and Brian C. Lovell. 2012. Shadow detection: A survey and comparative evaluation of recent methods. *Pattern Recognit.* 45, 4 (2012), 1684–1695. https://doi.org/10.1016/j.patcog.2011.10.001

Yichen Sheng, Yifan Liu, Jianming Zhang, Wei Yin, A. Cengiz Öztireli, He Zhang, Zhe Lin, Eli Shechtman, and Bedrich Benes. 2022. Controllable Shadow Generation Using Pixel Height Maps. In *17th European Conference on Computer Vision, ECCV Proceedings, Part XXIII (Lecture Notes in Computer Science, Vol. 13683)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, Tel Aviv, Israel, 240–256. https://doi.org/10.1007/978-3-031-20050-2_15

Yichen Sheng, Jianming Zhang, and Bedrich Benes. 2021. SSN: Soft Shadow Network for Image Compositing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR.* Computer Vision Foundation / IEEE, Virtual, 4380–4390. https://doi.org/10.1109/CVPR46437.2021.00436

Yichen Sheng, Jianming Zhang, Julien Philip, Yannick Hold-Geoffroy, Xin Sun, He Zhang, Lu Ling, and Bedrich Benes. 2023. PixHt-Lab: Pixel Height Based Light Effect Generation for Image Compositing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR.* IEEE, Vancouver, BC, Canada, 16643–16653. https://doi.org/10.1109/CVPR52729.2023.01597

Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E. Debevec, and Ravi Ramamoorthi. 2019. Single image portrait relighting. *ACM Trans. Graph.* 38, 4 (2019), 79:1–79:12. https://doi.org/10.1145/3306346.3323008

Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. 2017. Deep Image Harmonization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR.* IEEE Computer Society, Honolulu, HI, USA, 2799–2807. https://doi.org/10.1109/CVPR.2017.299

Rob Tuytel, Rico Cilliers, Dario Barresi, and Dimitrios Savva. 2023. Poly Haven textures library. https://polyhaven.com/textures.

Tomás F. Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. 2016. Large-Scale Training of Shadow Detectors with Noisily-Annotated Shadow Examples. In *14th European Conference on Computer Vision, ECCV Proceedings, Part VI (Lecture Notes in Computer Science, Vol. 9910)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer, Amsterdam, The Netherlands, 816–832. https://doi.org/10.1007/978-3-319-46466-4_49

Jifeng Wang, Xiang Li, and Jian Yang. 2018. Stacked Conditional Generative Adversarial Networks for Jointly Learning Shadow Detection and Shadow Removal. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR.* Computer Vision Foundation / IEEE Computer Society, Salt Lake City, UT, USA, 1788–1797. https://doi.org/10.1109/CVPR.2018.00192

Tianyu Wang, Xiaowei Hu, Chi-Wing Fu, and Pheng-Ann Heng. 2021. Single-Stage Instance Shadow Detection With Bidirectional Relation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR.* Computer Vision Foundation / IEEE, Virtual, 1–11. https://doi.org/10.1109/CVPR46437.2021.00007

Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, and Chi-Wing Fu. 2020a. Instance Shadow Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR.* Computer Vision Foundation / IEEE, Seattle, WA, USA, 1877–1886. https://doi.org/10.1109/CVPR42600.2020.00195

Yang Wang, Lei Zhang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang, and Dimitris Samaras. 2009. Face Relighting from a Single Image under Arbitrary Unknown Lighting Conditions. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 11 (2009), 1968–1984. https://doi.org/10.1109/TPAMI.2008.244

Zian Wang, Wenzheng Chen, David Acuna, Jan Kautz, and Sanja Fidler. 2022. Neural Light Field Estimation for Street Scenes with Differentiable Virtual Object Insertion. In *17th European Conference on Computer Vision, ECCV Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13662)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, Tel Aviv, Israel, 380–397. https://doi.org/10.1007/978-3-031-20086-1_22

Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. 2020b. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Trans. Graph.* 39, 6 (2020), 220:1–220:13. https://doi.org/10.1145/3414685.3417824

Piaopiao Yu, Jie Guo, Fan Huang, Cheng Zhou, Hongwei Che, Xiao Ling, and Yanwen Guo. 2021. Hierarchical Disentangled Representation Learning for Outdoor Illumination Estimation and Editing. In *IEEE/CVF International Conference on Computer Vision, ICCV.* IEEE, Montreal, QC, Canada, 15293–15302. https://doi.org/10.1109/ICCV48922.2021.01503

Fangneng Zhan, Shijian Lu, Changgong Zhang, Feiying Ma, and Xuansong Xie. 2020. Adversarial Image Composition with Auxiliary Illumination. In *15th Asian Conference on Computer Vision, ACCV Revised Selected Papers, Part II (Lecture Notes in Computer Science, Vol. 12623)*, Hiroshi Ishikawa, Cheng-Lin Liu, Tomás Pajdla, and Jianbo Shi (Eds.). Springer, Kyoto, Japan, 234–250. https://doi.org/10.1007/978-3-030-69532-3_15

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. 2019a. *Fixup Initialization: Residual Learning Without Normalization.* arXiv. arXiv:1901.09321 http://arxiv.org/abs/1901.09321

Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenmann, and Jean-François Lalonde. 2019b. All-Weather Deep Outdoor Lighting Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR.* Computer Vision Foundation / IEEE, Long Beach, CA, USA, 10158–10166. https://doi.org/10.1109/CVPR.2019.01040

Xuaner Cecilia Zhang, Jonathan T. Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E. Jacobs. 2020. Portrait shadow manipulation. *ACM Trans. Graph.* 39, 4 (2020), 78. https://doi.org/10.1145/3386569.3392390

Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David Jacobs. 2019. Deep Single-Image Portrait Relighting. In *IEEE/CVF International Conference on Computer Vision, ICCV.* IEEE, Seoul, Korea (South), 7193–7201. https://doi.org/10.1109/ICCV.2019.00729

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. 2015. Learning a Discriminative Model for the Perception of Realism in Composite Images. In *IEEE International Conference on Computer Vision, ICCV.* IEEE Computer Society, Santiago, Chile, 3943–3951. https://doi.org/10.1109/ICCV.2015.449

Jiejie Zhu, Kegan G. G. Samuel, Syed Zain Masood, and Marshall F. Tappen. 2010. Learning to recognize shadows in monochromatic natural images. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR.* IEEE Computer Society, San Francisco, CA, USA, 223–230. https://doi.org/10.1109/CVPR.2010.5540209

Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. 2018. Bidirectional Feature Pyramid Network with Recurrent Attention Residual Modules for Shadow Detection. In *15th European Conference on Computer Vision, ECCV Proceedings, Part VI (Lecture Notes in Computer Science, Vol. 11210)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer, Munich, Germany, 122–137. https://doi.org/10.1007/978-3-030-01231-1_8

Lei Zhu, Ke Xu, Zhanghan Ke, and Rynson W. H. Lau. 2021a. Mitigating Intensity Bias in Shadow Detection via Feature Decomposition and Reweighting. In *IEEE/CVF International Conference on Computer Vision, ICCV*. IEEE, Montreal, QC, Canada, 4682–4691. https://doi.org/10.1109/ICCV48922.2021.00466

Yurui Zhu, Xueyang Fu, Chengzhi Cao, Xi Wang, Qibin Sun, and Zheng-Jun Zha. 2022. Single Image Shadow Detection via Complementary Mechanism. In *The 30th ACM International Conference on Multimedia, MM*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, Lisbon, Portugal, 6717–6726. https://doi.org/10.1145/3503161.3547904

Yongjie Zhu, Yinda Zhang, Si Li, and Boxin Shi. 2021b. Spatially-Varying Outdoor Lighting Estimation From Intrinsics. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Vision Foundation / IEEE, Virtual, 12834–12842. https://doi.org/10.1109/CVPR46437.2021.01264

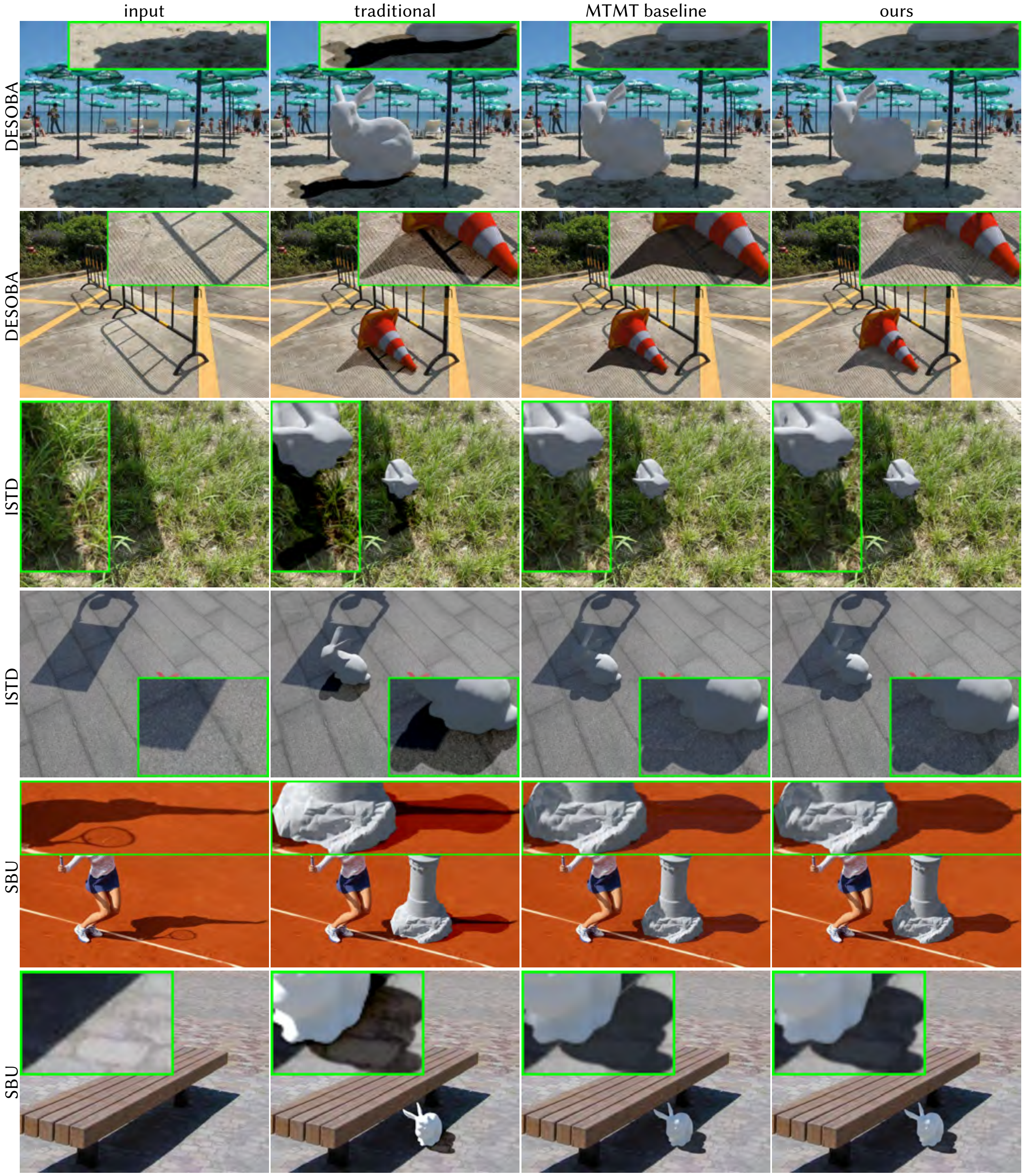


Figure 6: Qualitative results on test dataset images. From an input image (left col.) never seen in training, we composite a virtual object using the traditional pipeline [Debevec 1998] (2nd col.), which exhibits issues like double shadows on the ground and absent shadows on the virtual objects. Using our spatially-varying compositing equation (3rd, 4th column) yields much more plausible results. However, MTMT [Chen et al. 2020] (3rd col.) results have visible shadow boundaries in the composite (see insets). In contrast, our results (rightmost) display the most plausible shadow interactions.

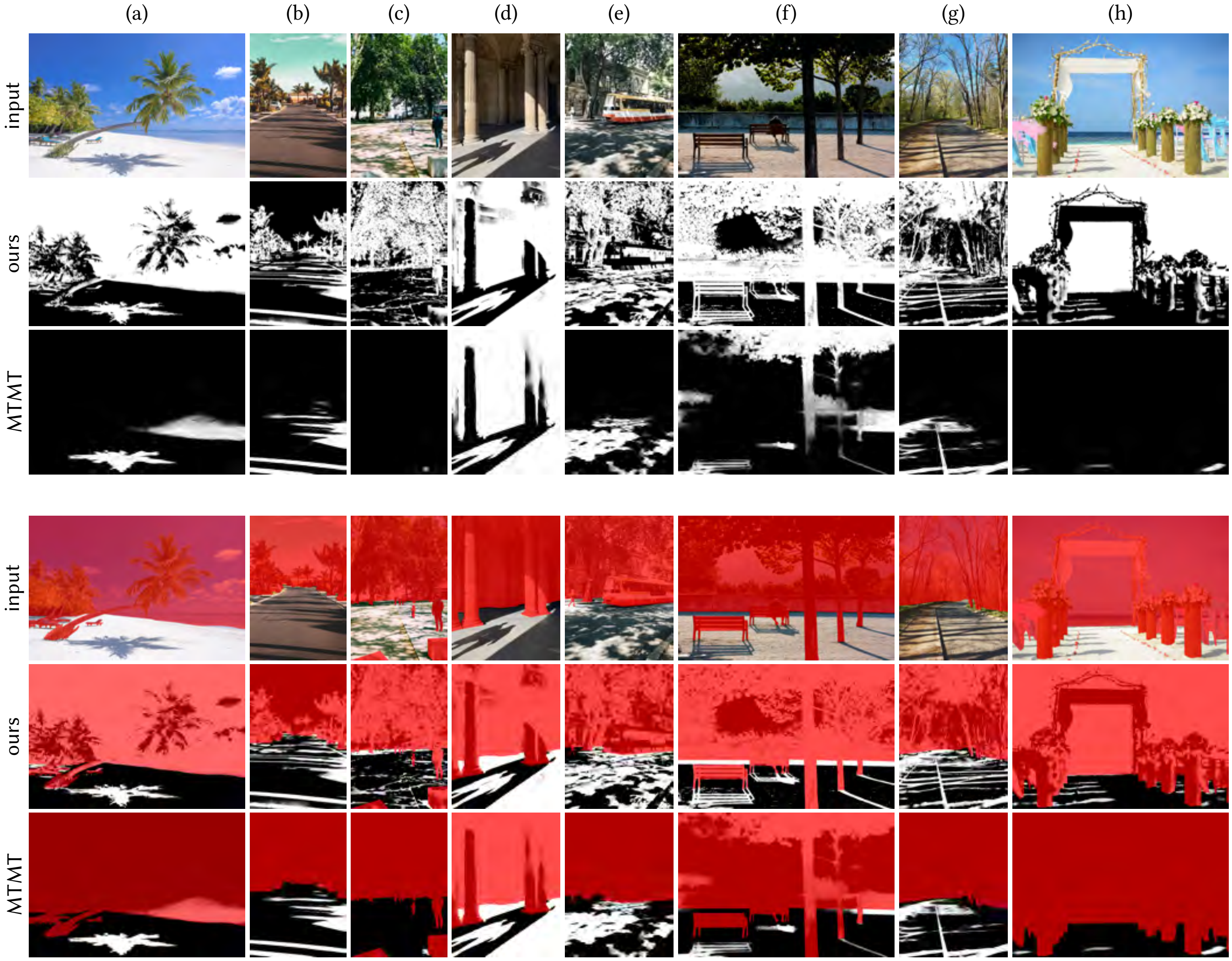


Figure 7: Shadow detection. Our model (2nd row) produces a more detailed soft shadow map than the current state-of-the-art shadow detection method [Chen et al. 2020]. Note that our method does not provide training signals for above-ground pixels such as the sky, so the method will produce unpredictable results in those regions. This is not an issue for shadow compositing on ground surfaces, as can be seen in the last 3 rows, with non-ground regions segmented in red. Moreover, in all results, we observe that our method extracts more shadow details. Our algorithm with stride 16 was used (see sec. 5.4). All samples were taken from our "in the wild" set of images found online (unseen during training) available under free Pexels license. Author credits: Asad Photo Maldives (a, h), Myburgh Roux (b), Adrien Olichon (c), Polina Chistyakova (d), Marta Dzedyshko (e), Jasper de Vreede (f), and Milica Vitor (g).