SIMBAR: Single Image-Based Scene Relighting For Effective Data Augmentation For Automated Driving Vision Tasks

Xianling Zhang[1*], Nathan Tseng[2*], Ameerah Syed[1], Rohan Bhasin[1], Nikita Jaipuria[1]
[1]Ford Greenfield Labs, Palo Alto [2]University of Michigan
{xzhan258, asyed17, rbhasin, njaipuri}@ford.com, tsnathan@umich.edu

Abstract

Real-world autonomous driving datasets comprise of images aggregated from different drives on the road. The ability to relight captured scenes to unseen lighting conditions, in a controllable manner, presents an opportunity to augment datasets with a richer variety of lighting conditions, similar to what would be encountered in the real-world. This paper presents a novel image-based relighting pipeline, SIMBAR, that can work with a single image as input. To the best of our knowledge, there is no prior work on scene relighting leveraging explicit geometric representations from a single image. We present qualitative comparisons with prior multi-view scene relighting baselines. To further validate and effectively quantify the benefit of leveraging SIMBAR for data augmentation for automated driving vision tasks, object detection and tracking experiments are conducted with a state-of-the-art method, a Multiple Object Tracking Accuracy (MOTA) of 93.3% is achieved with CenterTrack on SIMBAR-augmented KITTI - an impressive 9.0% relative improvement over the baseline MOTA of 85.6% with CenterTrack on original KITTI, both models trained from scratch and tested on Virtual KITTI. For more details and sample relit datasets, please visit our project website (https://simbarv1.github.io).

1. Introduction

A lack of diversity in lighting conditions is a known issue with manually collected real-world autonomous driving datasets [1, 3, 14, 18]. For example, KITTI [18] has video sequences captured only during noon, with similar lighting and shadow conditions across different sequences. More recent datasets [32, 37, 59], one such as BDD100K [59], are comparatively better in terms of diversity and have images captured during multiple times of the day. Still, between images collected from the same drive, there are minimal changes in lighting conditions. Furthermore, attempting to acquire data for all types of lighting conditions is implausible both in terms of time and money.

This lack of diversity in lighting conditions and by extension, the shadows present within a scene, often serves as a crucial roadblock in successful real-world deployment of perception models for safety-critical automated driving applications. Models trained with limited lighting conditions are unable to generalize to the plethora of lighting conditions encountered in the real-world [26, 28]. The ability to relight existing datasets in a controllable manner presents an opportunity to develop improved perception models.

Figure 1. Input images (Left) are shown against SIMBAR-relit outputs (Middle, Right). SIMBAR synthesized two lighting variations for (a)(b) Div2k, (c) BDD100K and (d) KITTI.

However, scene relighting, in the absence of depth sensors, is an extremely difficult vision task. It implicitly comprises of three main sub-tasks: shadow detection [10, 25, 53], removal [23, 24, 53] and insertion [61]. Of these, shadow removal and insertion are most challenging because shadows blend tightly with the source object geometry [2, 15]. This coupling makes it difficult to separate the shadow from its parent object without a strong 3D ge-

ometric understanding of the scene [4, 8, 20]. To address this, most prior scene relighting methods rely on multiple camera views of the source lighting condition to estimate the 3D scene geometry [42, 49, 62]. The relatively few prior methods that can work with a single image are based on Generative Adversarial Networks (GANs) [6]. GANs are known to be difficult to train [31, 38], limited in controllability [52], and often produce results that are physically inconsistent with scene geometry [16]. To the best of our knowledge, there is no prior work on *controllable* scene relighting using a single input image.

This paper presents a novel **S**ingle **IM**age-**BA**sed scene **R**elighting pipeline, SIMBAR. It takes a single image as input and produces relit versions for a wide variety of sun positions and sky zeniths, as shown in Fig. 1. The top two rows show relit results from the Div2k [1]. Div2k is an internet-scraped dataset with images of a wide variety of object classes, that SIMBAR is able to effectively relight. The first row shows realistic variations in sky colors, shadow orientations, and consistent cast shadow locations and light intensities for an outdoor scene with complex structures. The second row is a challenging low-light desert scene. SIMBAR cleanly removes existing hard cast shadows of the rock in the foreground and realistically recasts geometrically consistent shadows for the provided sun angle. Additionally, the mountainous landscape in the horizon has also been effectively relit. The third and fourth rows also show geometrically consistent and visually realistic relit versions of a KITTI road driving scene and a tunnel/underpass scene from BDD100K respectively. Most notable is the variation in hard cast shadows of the tunnel in the BDD100K example and the two cars in the KITTI example.

SIMBAR consists of two main modules: (i) geometry estimation and (ii) image relighting. The geometry estimation module is responsible for computing scene mesh proxy and illumination buffers. We are inspired by WorldSheet [22] to use external depth networks to obtain a scene mesh. Note that WorldSheet is a novel view synthesis pipeline that does not have relighting purpose. The image relighting module is inspired by prior work on multi-view scene relighting using a geometry aware network [42], referred to as MVR for brevity. Section 3.1 provides a short overview of Single Image-Based Scene Geometry Estimation and MVR, followed by a detailed description of SIMBAR's pipeline description in Section 3.2. Our work is closest in terms of goals and overall pipeline structure to MVR. Therefore, scene relighting comparisons are provided with both out-of-the-box MVR and its improved version, MVR-I, where we refined MVR for autonomous driving datasets with limited views, in Section 3.4. Across the board, SIMBAR provides significantly more realistic and geometrically consistent relit images, even though it takes as input a single image, as compared to MVR/MVR-I that take as input multiple images of the same scene.

Another major limitation of all prior works on scene relighting is the lack of a quantitative evaluation of the effectiveness of scene relighting in augmenting vision datasets. In the absence of such a metric, the real-world applicability and usefulness of any scene relighting methodology cannot be established. To address this, in Section 4, we perform image relighting-based data augmentation experiments with a state-of-the-art object detection and tracking network, CenterTrack [64]. Section 4.1 provides a detailed overview of our experiment setup. We train three different CenterTrack models on: (i) original KITTI tracking dataset with 21 real-world sequences captured at noon; (ii) augmented KITTI with MVR-I relit sequences; and (iii) augmented KITTI with SIMBAR relit sequences. All models are tested on Virtual KITTI (vKITTI) [17], which consists of clones of real KITTI sequences in a variety of lighting conditions. Section 4.2 shows that CenterTrack models augmented with relit KITTI images (from either MVR-I or SIMBAR) consistently outperform the baseline CenterTrack model. Specifically, the CenterTrack model trained on KITTI augmented with SIMBAR achieves the highest Multiple Object Tracking Accuracy (MOTA) of 93.3% - a 9.0% relative improvement over the baseline MOTA of 85.6%. This model also achieves the highest Multiple Object Detection Accuracy (MODA) of 94.1% - again an impressive 8.9% relative improvement over the baseline MODA of 86.4%.

To summarize, the main contributions of this paper are:

1. A novel single-view image-based scene relighting pipeline, called SIMBAR, that offers lighting controllability without the need for multi-perspective images.
2. Single image-based geometry estimation via adapting dense prediction transformer monodepth model and better representation of far-away background objects.
3. An improved version of MVR [42], called MVR-I, with fewer artifacts and smoother surfaces in the generated mesh for road driving scenes with limited views, resulting in more realistic relit images.
4. Qualitative evaluation and comparison of scene relighting results using MVR, MVR-I and SIMBAR, on multiple automated driving datasets, such as KITTI [18] and BDD100K [59].
5. Quantitative evaluation of the effectiveness of augmenting the popular KITTI 2D tracking dataset using SIMBAR and MVR-I for simultaneous object detection and tracking using CenterTrack.

2. Related Work

Our work is closely related to the fields of novel view synthesis [34, 47, 54], 3D reconstruction [9, 56, 57], and physics-based differentiable rendering [29, 41]. Given the direct connection between the relighting task and scene geometry [12, 60, 65], we split the related work into two broad

categories: (i) implicit approaches learning geometric priors and encoding them into a model; and (ii) explicit approaches leveraging multiple views of the input scene to generate a 3D mesh to apply rendering and image processing techniques upon. While explicit approaches provide better controllability and geometrically consistent shadows, their multi-view prerequisite inhibits their application to most automated driving datasets. This is due to the unique challenge of limited views from a front-facing car camera, compounded by high scene complexity of ever-moving cars and pedestrians. Our work falls in the explicit category, while leveraging insights from the implicit approaches.

2.1. Using Implicit Geometric Representations

Both Generative Adversarial Networks (GANs) [21] and Neural Radiance Fields (NeRFs) [35] have explored scene relighting. As is typical of GANs, the shadow manipulation network from [6] struggles to maintain geometric consistency and is difficult to train, thus resulting in conservative relighting effects. This also occurs for GANs that focus on image-to-image translation and ignore geometric priors [11, 16]. The recent success of NeRF-based methods for novel view synthesis has naturally resulted in their application to the scene relighting task as well. Rather than querying an explicit scene geometry, NeRFs encode the scene into a multilayer perceptron (MLP) [33], which takes as input a viewing direction and location to output color and density values, that are then used for volumetric rendering [39, 40]. At training time, many different views of a static scene are given to the network to learn the scene geometry. At test time, the input viewing direction and location are used to render the scene with accurate lighting and shadows. Recent works have repurposed NeRFs for scene relighting by modeling the surface material and reflectance properties [5, 49, 62]. However, such methods face a significant computational roadblock in their application to automated driving datasets with dynamic scenes, since each scene requires training a different model.

2.2. Using Explicit Geometric Representations

Combining Structure-from-Motion with Multi-View Stereo (SFM+MVS) is a common way of modeling scene geometry. It relies on feature matching across images captured from different views of a single scene of interest. After the application of SFM+MVS, bundle adjustment [51] can be used to generate a 3D point cloud, as is the case in COLMAP [45, 46]. The point cloud allows for application of traditional mesh reconstruction techniques, such as Delaunay [7] or Poisson [30] reconstruction, to generate an explicit geometric representation of the scene. Vision tasks that utilize geometric priors, such as novel view synthesis, can take advantage of such an explicit scene representation [44, 58]. The mesh can also be applied towards scene relighting tasks, as explored by [42]. In their work, physics-based rendering is used to approximate shadow locations using the generated mesh, with an additional network for shadow refinement. The relighting results are realistic and geometrically consistent. However, this method is severely limited in its application to a wide variety of datasets. For example, limited views and dynamic scenes result in failed mesh reconstruction [27]. In the case of relatively simpler and restricted datasets, such as human portraits, image relighting using a single view has been successful, owing to the high similarity in structure across facial data [36, 63]. However, the same is not true for datasets of outdoor scenes, which contain a wider variety of structure and content [13].

3. Single Image-Based Scene Relighting

Our proposed pipeline, SIMBAR, models the scene as a 3D mesh to explicitly represent scene geometry. Physics-based rendering is then used in conjunction with a shadow refinement network to produce realistic shadow maps. The original image can be composited with the target shadow maps to form the final relit output. Such an approach addresses the limitations posed by prior works on multi-view scene relighting and can generalize across scenes.

3.1. Preliminaries

3.1.1 Single Image-Based Scene Geometry Estimation

To solve the multi-view limitations of SFM+MVS-based mesh reconstruction, we have been inspired by WorldSheet [22] to use external depth for scene geometry estimation in order to perform single image-based mesh reconstruction. Note that the underlying ideas for the overall WorldSheet and SIMBAR pipelines are completely different. WorldSheet is a differentiable rendering pipeline, trained end-to-end for novel view synthesis, while SIMBAR is designed to manipulate existing views with various shadows cast.

For scene mesh formation, external depth predictions are treated as ground truth, thus requiring no predictions for grid offsets in the x and y direction. Let $z_{w,h}$ be the depth prediction at the corresponding sheet coordinate (w, h) and $x_{w,h}$ and $y_{w,h}$ are simply linearly spaced samples in the Normalized Device Coordinates (NDC) space from $[0, 1]$, with the camera placed at the origin. Given a fixed size of the mesh sheet of 129×129, depth predictions are grid-sampled to account for differences in resolution. With FoV angle, θ_F, this gives the following equation for forming the vertex coordinates:

$$V_{w,h} = \begin{bmatrix} z_{w,h}\, x_{w,h}\, tan(\theta_f/2) \\ z_{w,h}\, y_{w,h} tan(\theta_f/2) \\ z_{w,h} \end{bmatrix} \quad (1)$$

Grid edges that connect neighboring vertices form the mesh faces [22]. The faces are then smoothed with a Laplacian function [48] for the final output mesh.

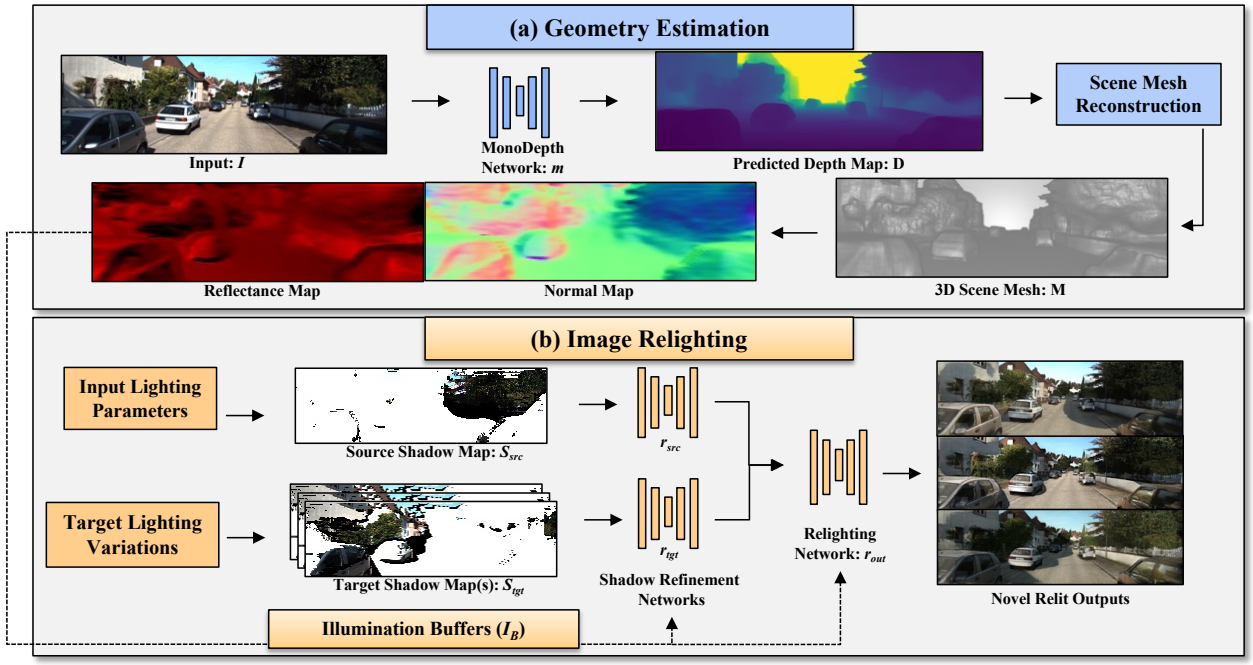

Figure 2. **(a) Geometry Estimation Component**: a single input image, I, is fed to monocular depth estimation networks (m). The predicted depth map, D, is used to form scene mesh using the vertex coordinates in Eq. 1. The resulting set of vertices and faces forms the 3D mesh, M. A set of input buffers, I_B, are rendered with respect to the camera pose using M. **(b) Image Relighting Component**: With estimated input lighting parameters and demanded target lighting variations, source shadow map S_{src} and target shadow map(s) S_{tgt} are generated. The shadow refinement networks r_{src} and r_{tgt} refine the shadow maps S_{src} and S_{tgt} respectively. Finally, relighting network r_{out} takes refined shadow maps with I_B, to generate the final relit images.

3.1.2 Geometry Aware Multi-View Relighting

Encoding the scene geometry priors and the relationship between scene geometry and lighting effects is an established method of providing strong signals to shadow removal and synthesis networks [35, 42, 62]. The image relighting networks in SIMBAR follow MVR [42], in which a set of geometric priors are leveraged as inputs in addition to the source image. A set of input buffers, I_B, are generated which consists of normal maps, reflectance maps, and refined shadow maps. The normal map encodes the surface normals at each pixel. The reflectance map is a dot product between the surface normals and sun directions. To obtain refined shadow maps, a set of coarse RGB shadow maps are used as inputs to two shadow refinement networks - one each for the source and target lighting condition. These coarse RGB shadow map are created from rays cast onto a 3D mesh of the scene to generate shadow locations. For each ray that intersects the mesh and casts a shadow, let m_i represent the point of intersection. The coordinates of m_i can be re-projected to find the corresponding 2D image pixel and its RGB value. The latter is encoded in the shadow maps to create RGB shadow maps. Encoding the RGB value that corresponds to the object that cast the shadow can help the shadow refinement networks correct the errors made by the 3D mesh reconstruction, in order to produce final refined shadow maps for the relighting network.

To finish the relighting process, a third network is used in combination with the shadow refinement networks. All of them are pre-trained on synthetically rendered data. Given the input image and RGB shadow maps for the source and target lighting conditions, the source and target shadow refinement networks attempt to refine the shadow maps to correct for errors in the mesh construction. This is followed by the final relighting network that takes in both the scene priors and refined shadow maps to produce the relit output.

3.2. Method Description: SIMBAR

Most prior scene relighting methods [42, 49, 62] require multiple images with different viewpoints. In contrast, SIMBAR leverages monocular depth estimation to obtain geometry approximation. SIMBAR is modular with two distinct components, geometry estimation and image relighting. The full pipeline is shown in Fig. 2. The geometry estimation module (a) represents the scene as a 3D mesh, which allows for a variety of informative priors to be generated for the image relighting module (b). This allows for a novel system design of a single image-based scene relighting that leverages explicit geometric scene representations.

3.2.1 Geometry Estimation Component

The geometry estimation module in SIMBAR generates a 3D scene mesh M, from a single input image I, as shown in Fig. 2. This is in direct contrast to MVR, which relies on SFM+MVS [45,46] for multi-view scene reconstruction. The steps taken to generate the mesh M from a single image I are inspired by WorldSheet (refer Section 3.1.1), but with additional modifications for improved mesh reconstruction.

In SIMBAR, an external pre-trained monocular depth estimation network is used to provide depth information for generating the scene mesh. This is because higher-quality meshes are given for outdoor driving scenes when leveraging the Worldsheet variant that uses an external depth prediction rather than the full end-to-end pipeline that predicts depth and grid offsets. This observation makes sense as with WorldSheet trained models, there is no direct loss on the mesh M in the end-to-end training regime. The supervision is instead obtained only via rendering losses on the final relit image. Thus, the predicted grid offset may not be as geometrically accurate as the one obtained using an external depth network. In addition, we have adapted new monodepth backbones for improved scene geometry estimation for relighting purpose.

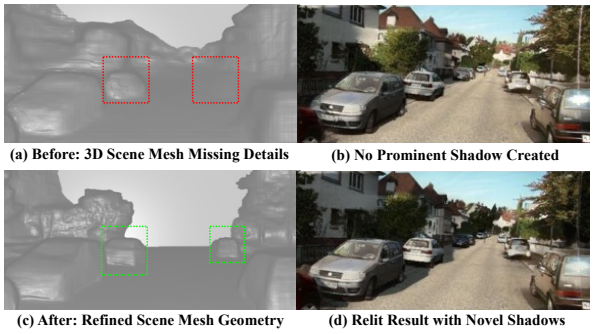

Figure 3. (a) With MiDaS v2.1, the 3D scene mesh misses details, resulting in (b) no prominent shadow created. (c) Our improvements with DPT Hybrid which leverages dense vision transformers captures far-away car objects, (d) creating realistic shadows.

Improved Monocular Depth Estimation: While WorldSheet utilizes MiDaS v2.1 as the external depth backbone, we have experimented with Dense Prediction Transformer (DPT) monodepth models [43]). Fig. 3 shows that the generated mesh M misses faraway car objects with the MiDaS v2.1 depth prediction, thus missing out on encoding structural details that can potentially cast shadows. This issue is particularly visible in the case of the KITTI scene on the top row, where the faraway car objects are not well relit. To address this limitation, we find that using the improved, dense vision transformers in DPT Hybrid-Kitti (finetuned on KITTI), helps produce more detailed meshes.

Foreground/Background Scene Separation: As shown in Fig. 2, for a given input image I, a pre-trained monocular depth estimation network is used to obtain the pixel-wise inverse depth values D. These values are then used to inform the deformation of a planar scene mesh. We observe that thresholding the inverse depth at different scales allows us to focus on different levels of detail.

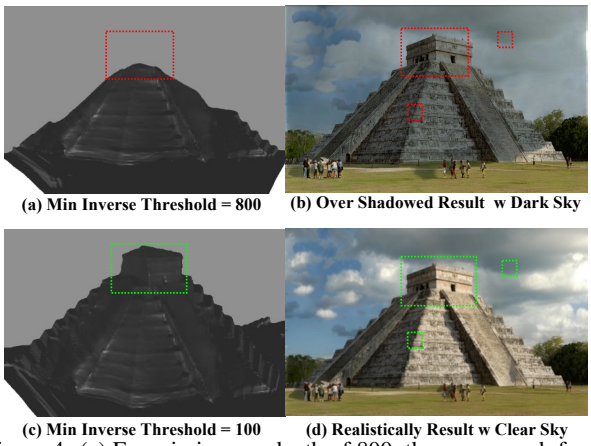

Figure 4. (a) For min inverse depth of 800, the scene mesh forms a flat vertical surface at the corresponding threshold distance. This phenomenon is observed as a flat, light gray wall mistakenly cuts off the top part of pyramid geometry. (b) This wall artifact casts a large dark shadow in the relit image labeled "over shadowed". (c) Using min inverse depth of 100 effective pushes the wall boundary further away, which gives a greater level of detail to the scene mesh, resulting in (d) a more realistic clear shadowed result.

Experiments with different levels of inverse depth thresholds are shown in Fig. 4. For the high inverse depth threshold of 800, a wall surface is generated fairly close to the camera and scene content. This set up could work for scenes with low depth range, but fail at diverse outdoor scenes with various depth boundaries. This results in over-shadowed results where the fake surface casts its own shadow over the scene. We opt for a lower inverse depth threshold, since this corresponds to a distance further away from the camera position. This allows the mesh to extend further back and produces cleaner shadows. Both the sky and surfaces far away in the horizon are better represented in the mesh M with a lower inverse depth threshold.

3.2.2 Image Relighting Component

As shown in Fig. 2, given the scene mesh M from the geometry estimation module, a set of priors or input buffers, as described in Section 3.1.2, are generated. They are fed as inputs to the shadow refinement networks (r_{src}, r_{tgt}) and the subsequent image relighting network (r_{out}). We choose to use MVR's pre-trained networks for r_{src}, r_{tgt} and r_{out} since they performed well despite imperfect mesh constructions across different datasets. Furthermore, obtaining a large and diverse set of high-resolution synthetic data for re-training the relighting networks is both time and cost intensive. Therefore, in SIMBAR, we focus on the novel adaption to single-view geometry-aware scene relighting.

3.3. Improved MVR Medthod as Baseline: MVR-I

The out-of-the-box MVR method fails at single-view collected autonomous driving dataset. To allow for comparisons with a strong baseline, we optimize MVR for road driving scenes with limited views, which we call MVR-I. We use MVR-I as a baseline for all qualitative (Section 3.4) and quantitative comparisons (Section 4.2).

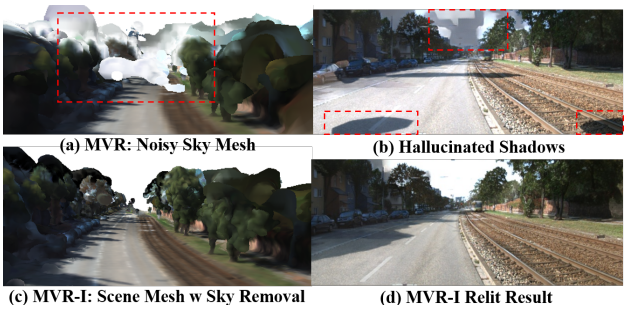

Figure 5. RGB point cloud overlayed on top of the scene mesh generated for a KITTI scene visualization. Using out-of-the-box MVR, hallucinates surfaces in the sky (a) resulting in phantom shadows (b), which we improve with MVR-I (c) leading to a more realistic image relighting result (d).

Removal of Hallucinated Mesh Surfaces: Firstly, we find that running MVR on KITTI scenes results in hallucinated sky surfaces in the generated mesh, thus casting corresponding phantom shadows on the ground. This is because SFM+MVS reconstruction triangulates selected 3D feature points in the input images with low re-projection error across images. In Fig. 5, note that the triangulated points leading to surface reconstruction in the sky in (a). These hallucinated surfaces cast prominent shadows in the sky and also on the foreground corner in the relit image in (b). While minor inaccuracies in the mesh can be addressed by the shadow refinement networks [42], the major inaccuracies shown lead to unrealistic scene relighting effects. To solve this issue, we implement a simple yet highly effective fix. We exclude confounding factors that appear in the sky in (c), such as clouds, as well as the sky itself, through segmentation using Detectron2 [55] on the input multi-view images. This addresses the issue of hallucinated mesh surfaces in the sky and corresponding phantom shadows (d).

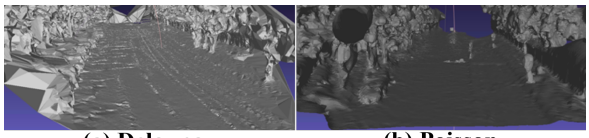

Figure 6. (a) Delaunay surface reconstruction is sensitive to noise, causing triangular artifacts. (b) Poisson reconstructed mesh has much smoother surfaces.

Improved Surface Reconstruction: The second improvement is replacing the Delaunay surface reconstruction algorithm [7] for mesh generation with the Poisson surface reconstruction algorithm [30]. Fig. 6 (left) shows that the Delaunay algorithm results in a noisy mesh, especially for the ground surface. Poisson surface reconstruction (right) for the same scene results in fewer angled edges and overall smoother road and tree surfaces.

A natural outcome of both these fixes is more realistic relighting results, as shown in Fig. 5 (d).

3.4. Scene Relighting Results

Both MVR and MVR-I require multiple viewpoints of a scene to generate an approximate 3D mesh using SFM+MVS. Such an approach fails in video sequences captured by a stationary ego vehicle because of the lack of multiple view-points within the captured sequence. This is a known limitation of SFM+MVS, which leads to many hallucinated shadows rendered in a KITTI frame relit using MVR-I. This can be observed in the top row in Fig. 7.

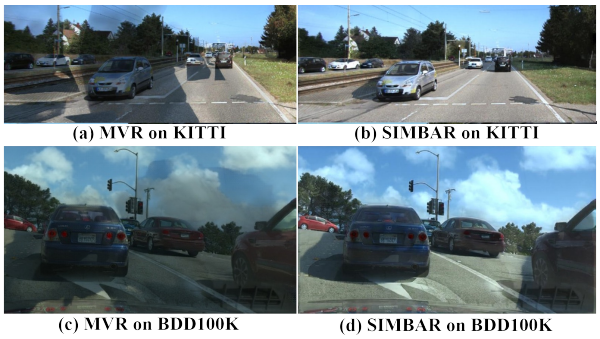

Figure 7. Relighting results from MVR-I (a)(c) and SIMBAR (b)(d) on KITTI and BDD100K respectively.

In contrast, SIMBAR provides significantly more realistic and geometrically consistent relighting results, as shown in Fig. 7 (b) and (d). While MVR-I fails to realistically relight images of road driving scenes from both KITTI (top) and BDD100K (bottom), SIMBAR's relighting results are consistently more realistic in terms of target shadow orientations and sky colors. However, there are some strong cast shadow residual that cannot be removed cleanly.

3.5. Limitations

Full Occlusion: With our proposed improvements in the geometry estimation module (refer Section 3.2.1), there are significant improvements in the generated mesh leading to more surface details of foreground objects and better inclusion of background objects. However, the natural drawback of a monocular depth approach is the exclusion of fully-occluded objects. While partially-occluded objects mesh errors can be corrected by shadow refinement networks, fully occluded objects currently present issues with shadow removal. The mesh is unable to represent the object without an additional view containing the object, yet in the real input image, the object can still contribute shadows. We find this to occasionally result in shadow residue from shadow

removal due to the lack of context on the object when utilizing single-view sources.

Scene Mesh Manipulation: Using low inverse threshold generates sky objects as wall surface further away in the horizon (see Fig. 4), and ideally we hope to remove the flat wall surface via scene mesh manipulation for more robust scene mesh separation. To achieve a better geometric understanding of individual objects in the scene and more granular control for scene relighting and shadow manipulation, another optimization could be leveraging 3D mesh predictions using neural networks such as Mesh R-CNN [19]. We currently use the 3D mesh as a geometric representation, and do not model the specific surface properties. Further modeling could allow for realistic lighting effects that account for specular reflection.

4. Data Augmentation Using Scene Relighting For Object Detection & Tracking

A serious limitation of all prior works on scene relighting is the lack of quantitative metrics to validate the effectiveness of scene relighting as a useful data augmentation methodology for vision tasks. In the absence of such a metric, the efficacy of real-world applicability of any scene relighting pipeline cannot be assessed. Therefore, to validate the effectiveness of scene relighting as a data augmentation strategy for vision tasks, we present real-world application results by integrating a state-of-the-art simultaneous object detection and tracking model, CenterTrack, with SIMBAR-augmented datasets. Our goal is to evaluate the enhanced generalization capability of vision models trained on data augmented using SIMBAR.

4.1. Experiment Setup

Train & Test Datasets: The KITTI tracking dataset consists of 21 sequences of road scenes, collected during daytime, with minimal variation in lighting conditions. Vision models trained on such a limited dataset cannot generalize well to the wide variety of lighting conditions that might be encountered in the real-world. To approximate this real-world generalization challenge, we train CenterTrack models on KITTI and test on vKITTI (only contains 'car' annotations) [17] which comprises of 'morning' and 'sunset' lighting variations. Prior work has also shown that testing on vKITTI is a useful strategy for evaluating data augmentations [50]. A visual of the domain gap between the training and test sets is shown in Fig. 8. Such an experiment setup is important in highlighting that vision models trained on limited datasets are susceptible to failure when encountering a seen scene in unseen lighting conditions.

Data Augmentation Using Scene Relighting: To compare the data augmentation effectiveness of SIMBAR with that of MVR-I (see Section 3.3), we compare the performance of CenterTrack models trained on two types of augmented KITTI datasets. Both use the full training set of ground-truth KITTI sequences, and an augmented versions of sequence numbers: 0001, 0002, 0006. The two augmented datasets differ in how the images are relit, one using MVR-I and the other using SIMBAR, where input parameters such as sun direction and sky zenith are randomly initialized. For our experiments, 4 different relit versions were generated for each frame in the 3 KITTI sequences. However, up to 120 different lighting conditions can be generated for each frame. We perform this augmentation offline. The rest of the training procedure follows the original CenterTrack implementation as-is. For brevity, we will refer to the CenterTrack models trained on the original 21 KITTI sequences without any image relighting-based augmentation as **(K)**; **(K+M)** and **(K+S)** denote the models trained with KITTI augmented with MVR-I relit sequences and SIMBAR relit sequences respectively.

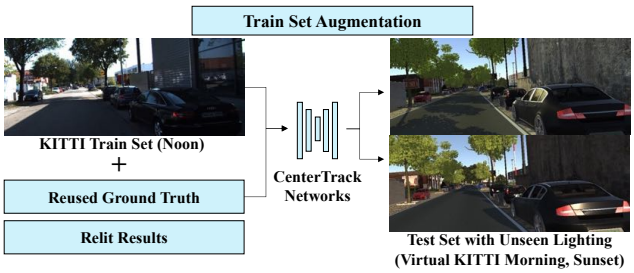

Figure 8. KITTI images taken at noon augmented with MVR-I/SIMBAR relit results used for training CenterTrack models, and vKITTI 'morning' and 'sunset' images used for testing.

Metrics: To quantify the effectiveness of data augmentation using scene relighting for object detection and tracking, we report the Multiple Object Tracking Accuracy (MOTA), MOT Precision (MOTP), Multiple Object Detection Accuracy (MODA), MOD Precision (MODP), complemented with Precision (P), Recall(R), F1 score, False Positives (FP) and False Negatives (FN).

4.2. Evaluation Results

A summary of the quantitative results is shown in Table. 1. All models trained from scratch, and the best checkpoint for each training run is chosen based on MOTA on the real KITTI validation set. CenterTrack models trained on KITTI augmented with relit KITTI, from either MVR-I (**K+M**) or SIMBAR (**K+S**), consistently outperform the baseline CenterTrack model trained on KITTI (**K**), on all metrics except for MODP. Specifically, the CenterTrack model trained on KITTI augmented with SIMBAR (**K+S**) has the highest MOTA of 93.3% - a 9.0% relative improvement over the baseline MOTA of 85.6% from **K**. Similarly, the highest MODA of 94.1% is also achieved by **K+S** - again an impressive 8.9% relative improvement over the

baseline MODA of 86.4% from **K**. In addition, **K+S** has the least amount of false positives and false negatives.

	K	K+M	K+S
MOTA ↑	85.6%	92.0%	**93.3%**
MOTP ↑	83.1%	83.5%	**83.5%**
MODA ↑	86.4%	92.7%	**94.1%**
MODP ↑	87.6%	87.6%	87.4%
Recall ↑	94.0%	96.5%	**96.9%**
Precision ↑	94.4%	97.4%	**98.1%**
F1 ↑	94.2%	96.9%	**97.5%**
False Positives ↓	283	133	**95**
False Negatives ↓	302	179	**157**

Table 1. Compared to baseline CenterTrack, models trained with data augmented using both MVR-I and SIMBAR provide consistently better performance.

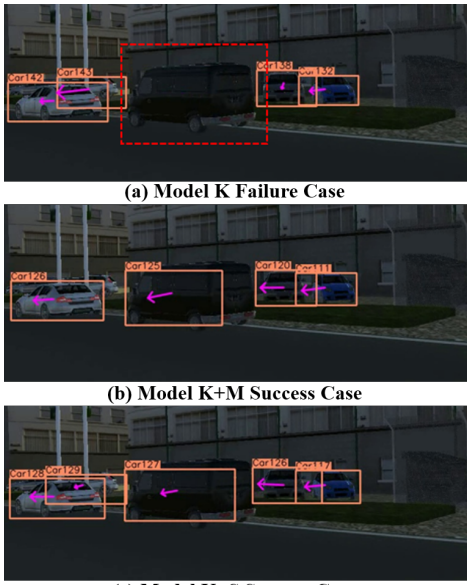

Figure 9. Object detection and tracking results on vKITTI from CenterTrack models **K** (a), **K+M** (b), and **K+S** (c).

Fig. 9 shows a qualitative downstream task performance comparison of detection and tracking results from **K**, **K+M** and **K+S** on vKITTI. The top result shows that model **K**, trained on original KITTI, fails to detect and track a black van obscured by a dense, dark shadow. Even though **K** was trained on the exact same scene from KITTI, it fails in this scenario because the training set, limited to images captured at noon, does not contain diverse lighting and shadow variations. Thus, model **K** performs properly in this seen scene with an unseen lighting condition. In contrast, both the models **K+M** and **K+S** perform well on this edge case with challenging lighting conditions.

To investigate the reliability of the obtained improvements, we ran 5 different training instances for each of the three models **K**, **K+M** and **K+S**. Each training instance was run for 100 epochs, which took 9 hours with 8 NVIDIA A100 GPUs. Fig. 10 shows consistent improvement in MOTA and MODA, averaged across the 5 training runs, with **K+S** achieving the best overall performance. Note that as compared to **K**, **K+M** also performs relatively well on the unseen lighting conditions, with a small variance across different training jobs.

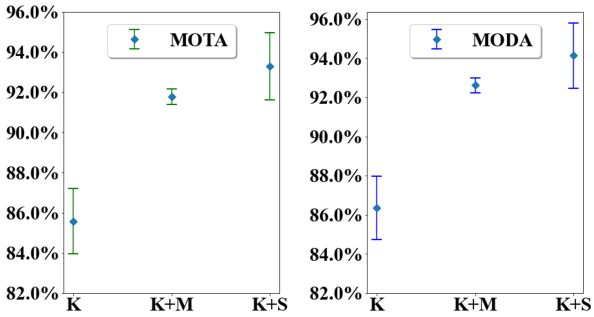

Figure 10. MOTA (left) and MODA (right) variance across 5 different training instances of models **K**, **K+M** and **K+S**.

5. Conclusion

We present a novel single-image based scene relighting pipeline, SIMBAR, for time and cost effective diversification of real-world datasets to include a plethora of lighting conditions. SIMBAR consists of two main modules. The geometry estimation module, inspired by 3D scene geometry estimation from a single image using WorldSheet, exploits various inverse depth thresholds and monocular depth networks to improve the scene mesh. The image relighting module re-purposes the relighting networks from prior art MVR and further relaxes the application-prohibitive requirement of multiple input images with different camera views. An improved version of MVR (MVR-I) is also provided for benchmark purposes. MVR-I leverages segmentation pre-processing to remove confounding classes, and is refined for road driving scenes.

Additionally, a comprehensive quantitative evaluation of CenterTrack models trained on KITTI augmented with relit data is used to demonstrate the effectiveness of scene relighting as a data augmentation strategy for object detection and tracking. Our results show an impressive MOTA of 93.3% on the vKITTI dataset with CenterTrack trained on KITTI augmented using SIMBAR - a 9.0% relative improvement over the baseline MOTA of 85.6% with CenterTrack trained on original KITTI. These results present a strong case for using SIMBAR as an effective data augmentation technique for vision tasks in automated driving.

Acknowledgements: The authors would like to thank Ronghang Hu and Deepak Pathak for sharing WorldSheet source code, and Julien Philip for the inspiring discussions on his MVR relighting method.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 1, 2

[2] Nijad Al-Najdawi, Helmut E Bez, Jyoti Singhai, and Eran A Edirisinghe. A survey of cast shadow detection algorithms. *Pattern Recognition Letters*, 33(6):752–764, 2012. 1

[3] Mohamed Aly. Real time detection of lane markers in urban streets. *CoRR*, abs/1411.7113, 2014. 1

[4] Iro Armeni, Sasha Sax, Amir Roshan Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *CoRR*, abs/1702.01105, 2017. 2

[5] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. Nerd: Neural reflectance decomposition from image collections. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3

[6] Alexandra Carlson, Ram Vasudevan, and Matthew Johnson-Roberson. Shadow transfer: Single image relighting for urban road scenes. *arXiv preprint arXiv:1909.10363*, 2019. 2, 3

[7] Frédéric Cazals and Joachim Giesen. Delaunay triangulation based surface reconstruction. In *Effective computational geometry for curves and surfaces*, pages 231–276. Springer, 2006. 3, 6

[8] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. *CoRR*, abs/1709.06158, 2017. 2

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5939–5948. Computer Vision Foundation / IEEE, 2019. 2

[10] Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 10680–10687. AAAI Press, 2020. 1

[11] Sourya Dipta Das, Nisarg A Shah, and Saikat Dutta. Msr-net: Multi-scale relighting network for one-to-one relighting. *arXiv preprint arXiv:2107.06125*, 2021. 3

[12] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In John Fujii, editor, *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996*, pages 11–20. ACM, 1996. 2

[13] Farshad Einabadi, Jean-Yves Guillemaut, and Adrian Hilton. Deep neural models for illumination estimation and relighting: A survey. In *Computer Graphics Forum*. Wiley Online Library, 2021. 3

[14] Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2013. 1

[15] Qingxu Fu, Xiaoguang Di, and Yu Zhang. Learning an adaptive model for extreme low-light raw image processing, 2020. 1

[16] Paul Gafton and Erick Maraz. 2d image relighting with image-to-image translation. *arXiv preprint arXiv:2006.07816*, 2020. 2, 3

[17] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016. 2, 7

[18] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 2

[19] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh R-CNN. *CoRR*, abs/1906.02739, 2019. 7

[20] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. *CoRR*, abs/1806.01260, 2018. 2

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[22] Ronghang Hu, Nikhila Ravi, Alexander C Berg, and Deepak Pathak. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12528–12537, 2021. 2, 3

[23] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(11):2795–2808, 2020. 1

[24] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2472–2481. IEEE, 2019. 1

[25] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7454–7462. Computer Vision Foundation / IEEE Computer Society, 2018. 1

[26] Manuel José Ibarra-Arenado, Tardi Tjahjadi, and Juan Pérez-Oria. Shadow detection in still road images using chrominance properties of shadows and spectral power distribution of the illumination. *Sensors*, 20(4):1012, 2020. 1

[27] Matthias Innmann, Kihwan Kim, Jinwei Gu, Matthias Nießner, Charles Loop, Marc Stamminger, and Jan Kautz. Nrmvs: Non-rigid multi-view stereo. In *Proceedings of the*

IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2754–2763, 2020. 3

[28] Nikita Jaipuria, Xianling Zhang, Rohan Bhasin, Mayar Arafa, Punarjay Chakravarty, Shubham Shrivastava, Sagar Manglani, and Vidya N Murali. Deflating dataset bias using synthetic data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 772–773, 2020. 1

[29] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1248–1258. Computer Vision Foundation / IEEE, 2020. 2

[30] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006. 3, 6

[31] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017. 2

[32] Seokju Lee, Junsik Kim, Jae Shin Yoon, Seunghak Shin, Oleksandr Bailo, Namil Kim, Tae-Hee Lee, Hyun Seok Hong, Seung-Hoon Han, and In So Kweon. Vpgnet: Vanishing point guided network for lane and road marking detection and recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[33] Popescu Marius, Valentina Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8, 07 2009. 3

[34] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *CoRR*, abs/1905.00889, 2019. 2

[35] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 3, 4

[36] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5124–5133, 2020. 3

[37] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *International Conference on Computer Vision (ICCV)*, 2017. 1

[38] Behnam Neyshabur, Srinadh Bhojanapalli, and Ayan Chakrabarti. Stabilizing gan training with multiple random projections. *arXiv preprint arXiv:1705.07831*, 2017. 2

[39] Michael Niemeyer, Lars M. Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3501–3512. Computer Vision Foundation / IEEE, 2020. 3

[40] Michael Niemeyer, Lars M. Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3501–3512. Computer Vision Foundation / IEEE, 2020. 3

[41] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10707–10716. Computer Vision Foundation / IEEE, 2020. 2

[42] Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei A Efros, and George Drettakis. Multi-view relighting using a geometry-aware network. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2, 3, 4, 6

[43] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 5

[44] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer Vision*, pages 623–640. Springer, 2020. 3

[45] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 5

[46] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 3, 5

[47] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. *CoRR*, abs/2004.04727, 2020. 2

[48] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, SGP '04, page 175–184, New York, NY, USA, 2004. Association for Computing Machinery. 3

[49] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 2, 3, 4

[50] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 969–977, 2018. 7

[51] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthe-

sis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999. 3

[52] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, pages 9786–9796. PMLR, 2020. 2

[53] Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1877–1886. Computer Vision Foundation / IEEE, 2020. 1

[54] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. *CoRR*, abs/1912.08804, 2019. 2

[55] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 6

[56] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomír Mech, and Ulrich Neumann. DISN: deep implicit surface network for high-quality single-view 3d reconstruction. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 490–500, 2019. 2

[57] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 204–213. Computer Vision Foundation / IEEE, 2021. 2

[58] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5336–5345, 2020. 3

[59] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018. 1, 2

[60] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas A. Funkhouser. 3dmatch: Learning local geometric descriptors from RGB-D reconstructions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 199–208. IEEE Computer Society, 2017. 2

[61] Shuyang Zhang, Runze Liang, and Miao Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Comput. Vis. Media*, 5(1):105–115, 2019. 1

[62] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *arXiv preprint arXiv:2106.01970*, 2021. 2, 3, 4

[63] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7194–7202, 2019. 3

[64] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, pages 474–490. Springer, 2020. 2

[65] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon A. J. Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM Trans. Graph.*, 23(3):600–608, 2004. 2