This CVPR 2021 workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Shadow Neural Radiance Fields for Multi-view Satellite Photogrammetry

Dawa Derksen
European Space Agency - ESTEC
Keplerlaan 1, 2201 AZ Noordwijk, Netherlands
dawa.derksen@esa.int

Dario Izzo

Abstract

We present a new generic method for shadow-aware multi-view satellite photogrammetry of Earth Observation scenes. Our proposed method, the Shadow Neural Radiance Field (S-NeRF) follows recent advances in implicit volumetric representation learning. For each scene, we train S-NeRF using very high spatial resolution optical images taken from known viewing angles. The learning requires no labels or shape priors: it is self-supervised by an image reconstruction loss. To accommodate for changing light source conditions both from a directional light source (the Sun) and a diffuse light source (the sky), we extend the NeRF approach in two ways. First, direct illumination from the Sun is modeled via a local light source visibility field. Second, indirect illumination from a diffuse light source is learned as a non-local color field as a function of the position of the Sun. Quantitatively, the combination of these factors reduces the altitude and color errors in shaded areas, compared to NeRF. The S-NeRF methodology not only performs novel view synthesis and full 3D shape estimation, it also enables shadow detection, albedo synthesis, and transient object filtering, without any explicit shape supervision.

1. Introduction

Many people think of satellite images as being from straight above, but they are almost always taken at an oblique angle. So-called *off-nadir* images provide information regarding vertical structure, and can be useful for estimating 3D shape. Moreover, the combination of images from different viewpoints reveals aspects that are in most cases impossible to capture with only one image.

One of the goals of multi-view imagery missions (SPOT6-7, WorldView-3) is to estimate the topography of the Earth's land cover. This knowledge can be relevant for several applications in Remote Sensing, including terrain mapping for flood risk mitigation [19], biomass estimation [29], land cover classification [7] and change detection [24].

While it is possible to use an active sensor such as a Light Detection And Ranging (LiDAR) to directly measure the distance from the satellite to the surface, these require significant amounts of energy compared to passive cameras.

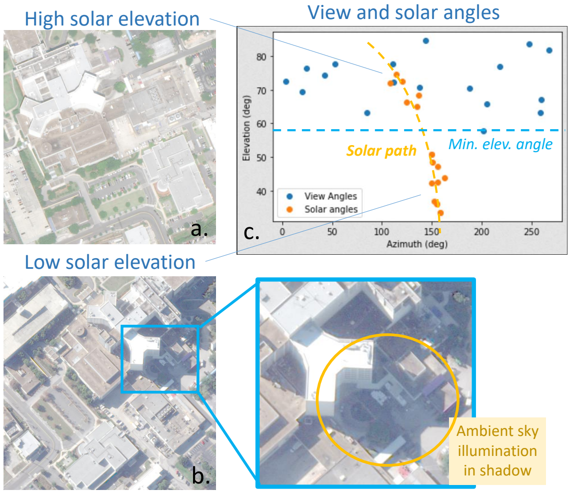

Figure 1. Two off-nadir WorldView-3 images with high (a.) and low (b.) solar elevation (θ_s). Strong temporal non-correlation principally manifests as changing shadows. The zoom on image b. shows that shadows are weakly illuminated by diffuse blue light coming from the sky. The graph (c.) displays the distribution of viewing angles (blue) and solar angles (orange). The solar directions are concentrated around the solar path, illustrated by the dotted orange line. They are clustered in two groups with no images in between.

In this research we focus on the performances achievable by optical sensors alone, motivated by applications with strong restrictions on available energy. These could include micro/nano-satellites for Earth Observation, or probes for space exploration around other planets [9], comets or asteroids.

In this paper, we rework one of the latest methods in 3D computer graphics, Neural Radiance Fields [20] (NeRF) to perform photogrammetry of urban scenes using Very High

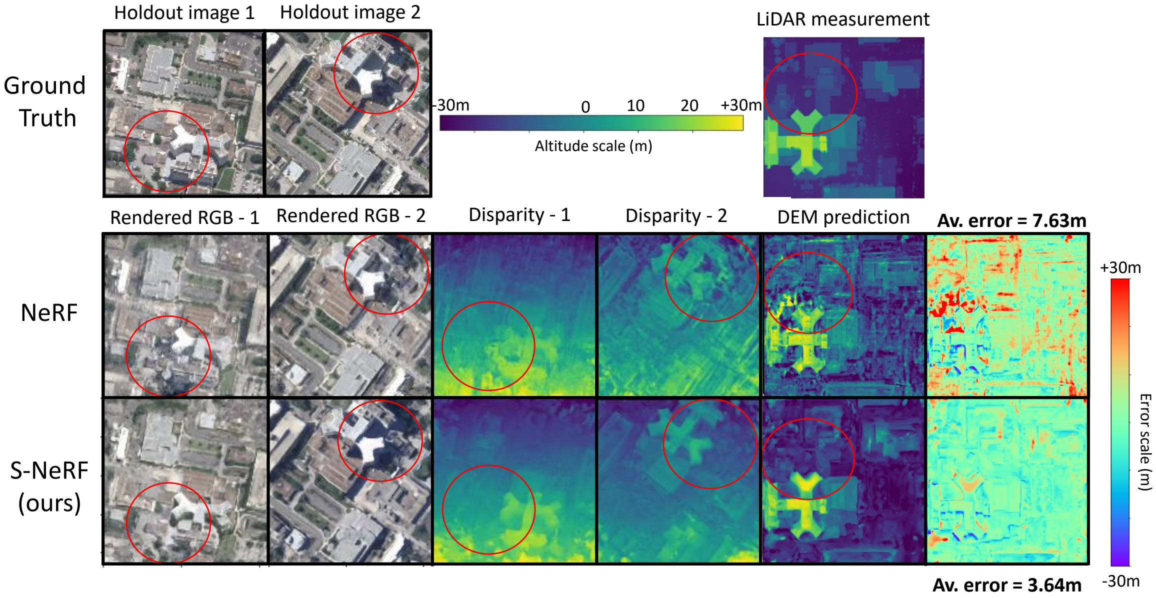

Figure 2. Comparison between NeRF [20] and S-NeRF. The first line displays two sources of ground truth information; test images in unseen viewing and lighting conditions, and a LiDAR measurement of the surface altitude. The second line demonstrates the performance of NeRF on both novel view synthesis and surface altitude prediction (DEM). The circled area, susceptible to containing non-correlation due to shadows, is where most of the errors are found both in shape and color. Our results are shown in the third row. S-NeRF allows for a higher quality estimation of the shadows, and produces a more accurate altitude prediction than NeRF. This is evident in the last column which shows the difference between the measured and estimated altitudes as well as the average absolute altitude error.

Spatial Resolution (VHSR) optical images from off-nadir viewing angles. Our problem differs from on-ground computer vision applications in several ways.

First of all, the images exhibit strong *non-correlation* because they are taken at different moments in time. Figure 1 shows an example of non-correlation due to varying lighting conditions on two WorldView-3 images [13]. Shadows moving in between images can cause errors to manifest when using altitude estimation methods based on pixel or feature matching [26, 25, 12]. Moreover, shadows cannot be easily detected due to ambient illumination from the sky. A further discussion about the different non-correlation sources and their impact on photogrammetry is made in Section 2.1.

Second, the number and distribution of viewing angles is restricted. A yearly WorldView-3 data set such as the one used in our experiments contains in average 10-20 images with an off-nadir angle lower than 35°. Figure 1 also reveals how the distribution of solar angles is relatively sparse, because the data was originally captured with stereo pair/triplet processing in mind.

The latest research in image-based photogrammetry has seen the arrival of a new family of methods, known as Neural Radiance Fields (NeRF) [20]. These have recently been explored for computer vision tasks such as novel view synthesis, which is the process of creating images of an object from an unseen viewing angle, given a set of 2D images of the object. By learning an internal representation of the volume as a continuous 3D function, NeRFs succeed in generating photo-realistic images from previously unseen viewing directions, using 20-100 posed training images. More details are provided in Section 2.2.

We demonstrate that the NeRF approach shows errors when applied to images with inconsistent lighting sources in Figure 2. An explicit model of the shadow effects is necessary in order to properly learn the shape of the scene, when faced with non-correlation due to changing lighting conditions.

This paper presents Shadow Neural Radiance Fields (S-NeRF), an extension of NeRF [20], that enables their application to multi-view satellite imagery with varying light conditions. Our changes to the NeRF model and rendering scheme are detailed in Section 3.

> **Section** 3.1 - Incoming light from a directional white light source (the Sun) and a diffuse colored light source (the sky) are learned by the model for each scene, in order to render realistic shadows with ambient light effects.
>
> **Section** 3.2 An additional training step that traces rays from the light source into the scene is proposed to im-

prove the quality of the learned light source visibility function at each point.

Section 3.3 An altitude-based sampling scheme is employed to better fit the overall flat shape of Earth Observation scenes.

In Section 4, we demonstrate the performance of S-NeRF on two tasks : novel view synthesis in unseen light conditions, and altitude estimation. We quantitatively show that S-NeRF outperforms NeRF on both tasks. We also show how using an explicit shadow model allows us to detect shadows in the training and test images, and estimate the albedo on the extracted 3D surface, even in areas that are persistently in the shadow in the images.

2. Related work

Learning the three dimensional properties of a scene based on two dimensional image information alone is an important challenge for Computer Vision and Remote Sensing alike. This task is often referred to as inverse rendering. Generally speaking, the objective is to estimate physical aspects such as opacity, spectral albedo, roughness, metallicity, or others, in the 3D volume or on a 2D subspace at the surface of the object. In theory, inverse rendering appears as a fundamentally under-constrained problem, an illustration is the monochrome scene, where the absence of visual cues means an infinity of possible shapes could be equally valid explanations for any set of images. In most natural images, however, the presence of texture and sharp details are sufficient to infer three dimensional structure with a reasonable number of observations.

Automatic inverse rending methods that reason about the 3D nature of the scene can be divided into the following groups, which differ on the kind of data that they require to operate.

Methods of the first group require prior knowledge of the shape of the observed scene. With this, it becomes relatively straightforward to project the different images onto the shape, to learn aspects such as surface specularity and shadows, to perform high-quality relighting [16, 22], or to refine the 3D geometry [3].

The second group seek to match visual patterns in the images with 3D shape, by training on examples of images where the scene shape is known. Once training has been accomplished, the models are able to infer the 3D shape of a new scene using a stereo pair [34], or a single image [21, 5]. However, this requires large data sets of images with a known shape for training.

The third group, which are the focus of this paper, are "unsupervised" or rather *self-supervised*: they infer the shape and surface properties from the images only. For Earth Observation with optical sensors alone, and for many other space-based applications, this appears to be the most interesting approach as it does not require any external form of data regarding surface shape. These methods are particularly relevant for space exploration where very limited prior shape information is available.

Section 2.1 elaborates upon methods that explicitly match different image elements together [27, 10]. Section 2.2 shows recent progress made in computer vision based on light transport through an implicit volumetric representation [15, 20, 1].

2.1. Stereo Matching

Many efforts in the direction of satellite-based photogrammetry have been focused on Stereo Matching [6, 11, 25, 23]. In this paradigm, the surface is represented by a function of the type $f(x, y) = h$, called the Digital Elevation Map (DEM), where (x, y) are spatial coordinates on the Earth (latitude, longitude for example) and h the surface height. These methods explicitly match pairs or triplets of pixels, with the idea that a group of matching rays must intersect at the surface of an object in the scene. This is posed as an optimization problem which maximizes multi-view consistency through mutual information pixel matching, while minimizing a global energy function to encourage regular 3D structures.

In its basic form, Stereo Matching lacks a way of modeling inconsistencies in the captured scenes, due to the fact that the visual cost is based on feature matching. For example, the movement of shadows between images at different moments of the day can make it difficult to accurately match spatial features. Other sources of temporal non-correlation include specular effects, transient objects (e.g. cars), vegetation growth, land cover change, and weather phenomena (e.g. snow). The impact of these inconsistencies is reduced by using pairs or triplets of images in similar conditions: roughly at the same time of day, and of the year. Studies by [26, 25] adapt the cost function of the semi-global optimization to tolerate non-correlation, but such approaches would most likely fail on strongly non-correlated images, for example, a set of morning and afternoon images mixed together.

2.2. Neural Radiance Fields

Neural Radiance Fields (NeRF) [20], simultaneously model color and geometry using a volumetric representation written as:

$$f(\mathbf{x}) = (\mathbf{c}_\lambda, \sigma)$$

The representation function f is defined on 3D space by the coordinate vector $\mathbf{x} \in \mathbb{R}^3$ The color of outgoing light $\mathbf{c}_\lambda$, also called *radiance*, is defined at discrete wavelengths: for visible images, $\mathbf{c}_\lambda \in [0, 1]^3$. The shape is modeled by $\sigma \in \mathbb{R}^+$, the density, which expresses how transparent or opaque a medium can be. In this study, we focus only on the effects of shadows on a Lambertian scene, and therefore omit the dependency of the radiance on the viewing angles. The combination of these factors is discussed in Section 5.

NeRFs construct f using a Multi-Layer Perceptron (MLP) network, thereby achieving a continuous "implicit representation" which has more interesting scaling properties than a voxel grid [15, 2]. The differentiable nature of the neural network is also key to solving the inverse rendering problem using gradient descent.

2.2.1 Fully emissive rendering model

Optical volume rendering works by simulating light transport through a 3D representation, to infer what image a camera would see from a certain position in space. Each of the pixels in a desired image defines a ray, which accumulates light as it traverses the scene.

Previous successful neural volume rendering approaches [15, 20] are based on the fully emissive light transport model [18]. In practice, the continuous rendering integral is estimated as a discrete sum, by performing a Monte-Carlo integration [20]. This means sampling the volume properties $(\mathbf{c}, \sigma)$ predicted by the neural network at N_s locations along each ray. The discretized form of the rendering integral is shown in Equation 1.

$$\begin{aligned} \hat{\mathbf{I}} &= \sum_{i=1}^{N_s} w_i \mathbf{c_i} \\ w_i &= T_i \alpha_i \\ \alpha_i &= 1 - e^{-\sigma_i \delta x_i} \\ T_i &= \prod_{j=0}^{i-1} (1 - \alpha_j) \end{aligned} \tag{1}$$

The approximated perceived light color $\hat{\mathbf{I}}$ is written as a weighted sum of the color vectors $\mathbf{c_i}$, with weights w_i. This numerical integration scheme defines α_i as the opacity along ray segment i of length δx_i, by definition $\alpha_i \in [0, 1]$. The transparency T_i is written as the cumulative product of the inverse opacity, from the origin to the current segment i. A high w_i value indicates that ray segment i is either emitting or reflecting light (high α_i), and not blocked earlier along the ray (high T_i).

2.2.2 Training

Solving the inverse rendering problem means learning how much each individual element of the scene contributes to the final perceived color.

To train a neural volumetric representation to fit with a data set of images, a batch of N_b pixels is randomly sampled from all of the images. The perceived color $\hat{\mathbf{I}}(\mathbf{r})$ is then estimated along the corresponding rays. This results in a loss value for each pixel, which is back-propagated back to the weights of the neural network. In this way, the neural volumetric representation is modified to better fit with what is observed in the training images. It should be noted that such gradient descent optimization is only possible because both the representation and the rendering functions are differentiable.

2.2.3 Neural Reflectance Fields

For 3D scene relighting, Neural Reflectance Fields [1] were recently introduced. These methods actively reason about the quantity of incoming light at a spatial position, by computing the visibility between each query point and the light source on the fly during inference. Training a Neural Reflectance Field using images with a known non-collocated light source, or multiple light sources, would imply a *quadratic* or even *cubic* increase in the number of sample points required to estimate the light transport function, depending on the complexity of the lighting effects that are modeled. This would be prohibitive for practical purposes, especially in terms of memory. Therefore, Neural Reflectance Fields can only be trained on images taken with a white light source collocated with the viewing axis (camera flash only). Nonetheless, they show success in extracting surface parameters such as the albedo, and perform accurate relighting on objects that contain complex structure such as fine elements.

Our work proposes a different strategy than Neural Reflectance Fields [1], which is necessary because in multi-view satellite images (and in many other cases), the light source is non-collocated with the camera axis. For this reason, we propose to restrict the lighting conditions to the two sources that illuminate the Earth: the Sun and the sky. Explicit knowledge of the solar direction is used to simplify the full optical volume rendering formulation. In our method we put high priority on maintaining linear complexity with respect to the number of sample points along the rays, $O(N_s)$ which is *the same complexity as NeRF*.

3. Method

3.1. Irradiance model

We define the spectral irradiance vector $\ell_\lambda(\mathbf{x}, \omega_\mathbf{s})$, as the intensity of incoming light at a 3D location, for the solar direction $\omega_\mathbf{s}$, at wavelengths $\lambda \in \{R, G, B\}$.

Instead of explicitly calculating the irradiance at each point [2, 1], we propose to implicitly model this quantity in the neural network. The main goal of this is to avoid numerically estimating a full 2D or 3D integral during training and inference.

In our method, we model variations in incoming light by introducing two new outputs to the network, $s(\mathbf{x}, \omega_\mathbf{s})$ and $\mathbf{sky}(\omega_\mathbf{s})$, which both depend on one new 2D input, the solar direction $\omega_\mathbf{s} = (\theta_s, \phi_s)$.

Physically, $s(\mathbf{x}, \omega_\mathbf{s})$ represents the ratio of incoming solar light with respect to the diffuse sky light. This quantity can also be loosely interpreted as the visibility of the directional light source at 3D location $\mathbf{x}$, along the direction $\omega_\mathbf{s}$. A value of 0 means no solar visibility and 1 means full solar visibility.

The other quantity, $\mathbf{sky}(\omega_\mathbf{s})$ is a learned vector that expresses the color of the illumination incoming from the sky (3D for RGB images). The network predicts these values based on the solar direction alone, and not on spatial coordinates, to model how the sky acts as a diffuse light source that only visibly contributes to lighting up areas in the shadow.

In order to take both of these quantities into account in the rendering, a new term is added to the alpha-compositing formula from Equation 1.

We define the total irradiance ℓ as a weighted sum of the known and learned light sources, using the network-predicted values of s and $\mathbf{sky}$. The light mixing model, shown in Equation 2 poses ℓ as a weighted sum of a white light source $\mathbb{1}_3$ and the learned sky color using s and $(1-s)$ as weights. This follows from the assumption that the Sun emits white light in the visible bands.

Instead of directly modeling the radiance $\mathbf{c}(\mathbf{x})$ as was done previously [15, 20] we explicitly model the albedo $\mathbf{a}(\mathbf{x})$ as a network output (Figure 4). We then use a simple Lambertian reflectance model that dictates that the radiance (noted $\mathbf{c}$ in Equation 1) is the point-wise product of the irradiance ℓ and the albedo $\mathbf{a}$ vectors.

$$\begin{aligned} \ell &= s\mathbb{1}_3 + (1-s)\mathbf{sky} \\ \mathbf{c}(\mathbf{x}, \omega_\mathbf{s}) &= \mathbf{a}(\mathbf{x}) \cdot \ell(\mathbf{x}, \omega_\mathbf{s}) \end{aligned} \tag{2}$$

The new alpha-compositing model to estimate the perceived color *with shading*, $\hat{\mathbf{I}}_\mathbf{s}$ is expressed in Equation 3, using the same definition for w_i as Equation 1.

$$\hat{\mathbf{I}}_\mathbf{s} = \sum_{i=1}^{N_s} w_i \mathbf{a_i} \ell_\mathbf{i} \tag{3}$$

A visual example of the S-NeRF rendering procedure that is described in Equations 2 and 3 is shown in Figure 3.

This simplification of the full light rendering model avoids extra sampling during training iterations, but does not account for specular effects and makes heavy simplifications on the indirect illumination. In our model, every area in the shadow receives an equal amount of light from the sky, meaning that occlusion of sky light by the scene itself is not taken into consideration. Furthermore, it does not allow for indirect light coming from other parts of the scene, also known as double-bounce indirect illumination. These are very challenging to take into account without increasing the complexity of the integration.

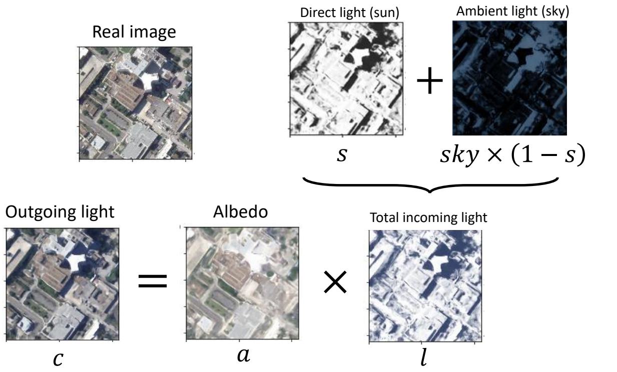

Figure 3. Schematic of the S-NeRF shading model (Equation 2). The outgoing light (radiance) $\mathbf{c}$ at each point along a ray is modeled as a point-wise product between the albedo $\mathbf{a}$ and the total incoming light (irradiance) ℓ. The latter is a mixture between the direct white light from the Sun and the indirect light from the sky, which is weighted by the solar visibility function. We see how the faint sky light is learned as a bluish tone, and how the albedo image contains no visible shadows.

3.2. Solar correction rays

To accurately learn s, the solar visibility function, we propose to draw a second set of rays at each training iteration. We call these Solar Correction (SC) rays, as they aim to correct the solar visibility function based on the current estimated scene geometry. Along such a ray we compute the transparency T_i from Equation 1. We then add a term to the optimization loss function to penalize differences between T_i and s_i. The idea of this loss is that along a solar ray, the transparency at each point along the ray is equivalent to the ratio of incoming solar light at that point.

Equation 4 shows the loss function L for a batch of rays b, which contains three terms.

$$\begin{aligned} L(b) = &\sum_{\mathbf{r} \in b} ||\mathbf{I}(\mathbf{r}) - \hat{\mathbf{I}}_\mathbf{s}(\mathbf{r})||_2^2 + \\ &\lambda_s \sum_{\mathbf{r} \in SC} (\sum_{i=1}^{N_s} (T_i - s_i)^2 + 1 - \sum_{i=1}^{N_s} w_i s_i) \end{aligned} \tag{4}$$

1. The pixel-based RGB loss, a mean squared error between the network predicted pixels $\hat{\mathbf{I}}_\mathbf{s}$ and the image pixels $\mathbf{I}$, of a batch of rays b.

2. The L_2 loss between T and s, on the solar correction rays (SC), using the same definition for T as Equation 1. To encourage the s values to become close to the current T values, we stop the gradient on T.

3. The L_1 norm of ws subtracted from 1. This term expresses the idea that the entirety of the solar light should be absorbed by the visible surface. The gradient of w is stopped.

λ_s is a weight parameter to balance the participation of the different losses. This parameter represents the trade-off between color accuracy and shadow validity. We found $\lambda_s = 0.05$ to work best in our experiments.

Solar correction rays can be randomly drawn from the upper hemisphere, or focused on the positions of certain requested angles. For the experiments, we used solar correction rays generated from directions along an interpolation axis that follows the solar path (Figure 1). The result of this interpolation, Figure 5, demonstrates the importance of solar correction when generalizing to lighting conditions that are significantly different from the conditions present in the data set.

3.3. Altitude-based sampling

We define a new sampling scheme for selecting integration points along the rays, to better fit Earth Observation scenes where the bounds of the X and Y dimensions often exceed the bounds of the Z axis. The samples are selected in regular intervals along the absolute Z axis rather than along the sensor axis.

We also sample between minimum and maximum altitude values rather than near and far distances, to avoid introducing parameters which would depend on the satellite configuration. Selecting the minimal and maximal altitude parameters for a given zone can be done in various ways, for example, from a large-scale Digital Terrain Map based on a low-resolution data source. In our case we selected values based on the minimum and maximum values of the airborne LiDAR maps (Table 1, 3rd row).

We also use importance sampling as proposed by [20] on the visual rays only, as it appears to increase the effective depth resolution for a fixed sampling budget.

3.4. Network architecture

These experiments are based on the network architecture shown in Figure 4. The internal representation network can be divided into four parts.

First, the density output is connected to the spatial inputs through a 8 fully-connected SIREN layers of width 100, using the initialization procedure described by [30] instead of positional encoding. Importantly, like in [20], the final activation function that provides σ is a ReLU, so as to provide higher values than 1. We also add noise to the unactivated output with a standard deviation of 10, which we decrease to 0 throughout the training [28].

Second, the RGB albedo is connected to the last layer of the density network, activated by a sigmoid. Next, the solar visibility s is connected to the last layer of the density network, concatenated to the solar direction. Here, 4 layers are used, with initial output noise standard deviation of 1. Finally, the sky color is estimated using only the solar direction, through 1 ReLU layer.

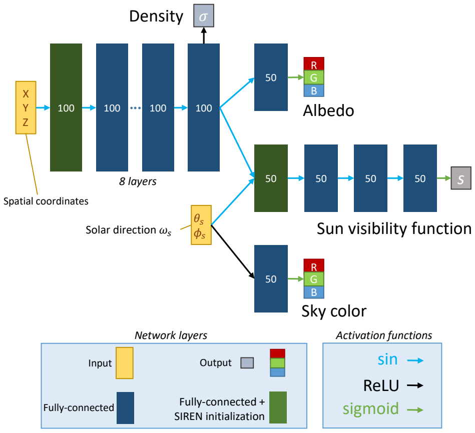

Figure 4. S-NeRF network architecture. Spatial coordinates are used as inputs to the density and albedo layers in the same way as the original NeRF [20]. Our major changes are the addition of 1. a solar visibility network which takes as an input both the (unactivated) outputs of the last density layer and the solar direction ω_s, and 2. a sky color estimation layer.

Explicitly separating the contributions of these four factors enables an independent synthesis of each property, which can help in correcting them individually (solar correction rays), or to render certain properties without others (shadow-free rendering).

4. Experiments

We evaluate the performance of S-NeRF on satellite imagery from the WorldView-3 optical sensor [13], training the model using only the visible RGB bands. Instead of using the images at full resolution (0.3m), we downsample the images with anti-aliasing by a factor of 2, to obtain images with a resolution of 0.6m. This is done because the sampling distance of the airborne LiDAR is 0.5m, meaning that we cannot evaluate the benefit of finer spatial resolution images.

We perform two sets of experiments on four different urban scenes of Jacksonville, shown in Figure 6. Each area covers 300×300m.

The first experiment (Section 4.1) aims to evaluate how well S-NeRF can perform novel view synthesis in previously unseen conditions: both viewing and lighting.

The second experiment (Section 4.2) examines the accuracy of the surface shape by computing the expected altitude of the scene (Equation 5), evaluated with respect to the airborne LiDAR measurement.

Area index	004	068	214	260
# train/test	8/2	16/2	21/2	14/2
Alt. bounds (m)	0/-30	30/-30	80/-30	30/-30
SSIM (test set)				
NeRF	**0.364**	**0.471**	0.377	0.409
S-NeRF no SC	0.352	0.322	0.360	0.401
S-NeRF + SC	0.344	0.459	**0.384**	**0.416**
Altitude MAE (m)				
NeRF	5.607	7.627	8.035	11.97
S-NeRF no SC	**3.342**	4.799	**4.499**	10.18
S-NeRF + SC	4.418	**3.644**	4.829	**7.173**

Table 1. Evaluation metrics on NeRF and S-NeRF. The upper table contains the number of training and test images used for each area, as well as the altitude bounds of the scene. Area indices correspond to the images in Figure 6. The middle table demonstrates the results of the novel view synthesis experiment. The best SSIM values are shown in bold. Overall, S-NeRF provides similar image quality to NeRF. The third table shows that S-NeRF always provides a lower altitude Mean Average Error (MAE) than NeRF. However, solar correction is not clearly beneficial for both metrics, meaning that the shading model and correction scheme could be improved further.

4.1. Novel view synthesis

In Figure 2 we compare novel views from unseen viewing and lighting conditions to real images in those conditions. Without the shadow model (NeRF), the shaded areas are smoothed together and form a grayish area without crisp edges. With our proposed S-NeRF model, the shadows are well positioned, and illuminated by a faint blue ambient sky light.

Figure 5 pushes this conclusion further by showing how S-NeRF allows for a smooth interpolation between two known lighting conditions. Because the RGB albedo values do not depend on the lighting conditions, the model maintains consistent colors when interpolating between different solar positions. The solar correction rays further allow for the shadows to be well placed and coherent with the estimated shape in previously unseen conditions.

In terms of quantitative results, Table 1 shows that NeRF and S-NeRF produce statistically equivalent image quality scores, as measured by the SSIM. It appears as though in certain scenes, an overall darkening effect can counteract the increase in quality in the shadowed areas, which is visible in Figure 2. This table also demonstrates that solar correction maintains or increases the SSIM values, compared to when this scheme is not used, on 3/4 of the evaluated scenes. Area 004 (upper left image of Figure 6) is the only region where we observe a decrease in SSIM when using solar correction. We believe this is due to the small variations in altitude in this scene, which means that cast shadows occupy smaller parts of the images. Further statistical analysis is required to draw conclusions on the best way to apply solar correction.

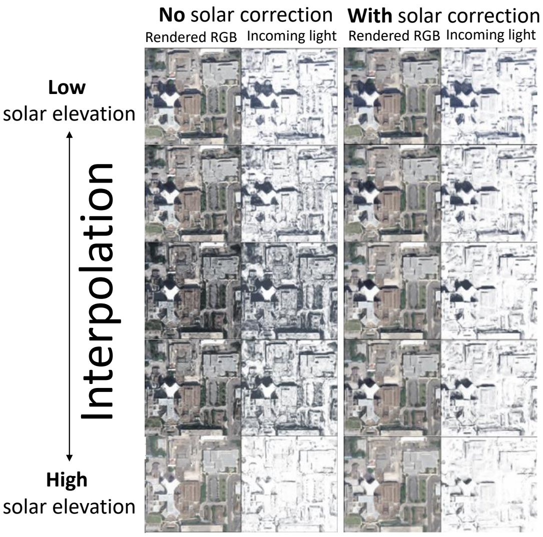

Figure 5. Interpolation of the solar direction vector ω_s between two solar positions that are present in the training data. We compare two shading schemes. Left: with no solar correction, the incoming light is well estimated on training angles, but poorly estimated on unseen angles. Right: solar correction fixes darkening effects by learning the light source visibility based on the estimated scene geometry. These views are rendered at nadir to show how S-NeRF can produce synthetic ortho-images.

Next, we show the estimated albedo at the surface for various scenes in Figure 6. This result shows how S-NeRF can be used to provide a 3D estimate of the albedo, even in areas that are entirely in the shadow in the training image set. This is because the model is able to learn the color of the ambient sky light, and its local contribution to the total perceived light. The performance cannot easily be quantified on real-world data as we have no widespread measure of the surface albedo.

We also observe that the albedo contains no or very few transient objects (specifically cars) as the model tries to learn a coherent representation for all of the training images. This is visible in area 214, on the lower left of Figure 6. The cars in the parking lots are mostly filtered out entirely, or at least blurred out. These outputs could be useful to detect transient objects in a new unseen image, or to provide a more robust semantic segmentation of the scene in the case where such objects are not of interest.

4.2. Altitude extraction

S-NeRF provides a full 3D model of the density from which [20, 30] extract a surface boundary by selecting an iso-σ surface and applying marching cubes. However, we lack a ground truth shape of the full 3D structure, so for

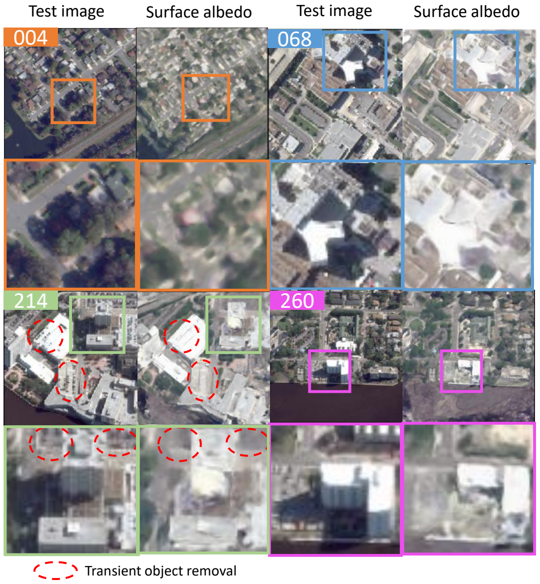

Figure 6. Demonstration of S-NeRF shadow removal and transient object removal. The first and third column show unseen test images and the second and last column show the estimated albedo at the surface. The approach is mostly successful in restoring color to the persistent shadows. Remaining errors are particularly found at the intersection between the ground and vertical facades and along their outer edges. We can also note the transient object removal in the three parking lots circled in area 214 (lower left).

this experiment we only evaluate the quality of the shape on its *maximal altitude*. For this we render a vertical view and compute the Mean Average Error (MAE) between the expected altitude and the altitude measured by an airborne LiDAR [13].

$$\hat{h} = \sum_{i=1}^{N_s} w_i h_i \qquad h_i \in [h_{max}, h_{min}] \tag{5}$$

Equation 5 demonstrates how we estimate the surface altitude $\hat{h}$ from the volumetric representation. For each ray at a desired ground spatial resolution (0.5m), we sample altitudes h_i between h_{max} and h_{min} (ray traveling downwards) using the sampling strategy described in Section 3.3. The estimated altitude is written as a weighted sum of the sample altitudes h_i with weights w_i. By definition (Equation 1), w is close to 1 at the location where the first object surface is visible.

Figure 2 shows how S-NeRF corrects most of the shape inaccuracies that the original NeRF causes in shadowed areas. This is confirmed by Table 1, which indicates that S-NeRF produces a lower altitude MAE than NeRF on four different areas. In terms of comparison to Stereo Matching, we report that the average error of the predicted DEMs is in the same order of magnitude as the ones measured by [8], which use a fusion of Stereo Matching predictions on 47 WorldView-3 images.

The training takes around 8 hours on a NVIDIA GeForce RTX GPU with 12GB RAM. We use an Adam optimizer with learning rate decaying from $1e^{-4}$ to $1e^{-5}$ over 100k iterations. The batch size is set to 256 pixel rays and 256 solar correction rays (when used). The number of samples and importance samples are 64 and 64, respectively.

5. Discussion

One of the limitations of NeRF and S-NeRF alike is the need to retrain a neural network for each new scene. Recent efforts have been made towards generalizing neural volumetric representations to multiple scenes [28, 32, 33], and in making them more compact [14]. This may be relevant for multi-view satellite imagery in the future.

More robust transient object filtering and deeper temporal analysis could be attempted using latent embedding optimization [4], taking inspiration from NeRF in the Wild [17]. This method can handle severe changes in illumination, but does not provide a solution to general relighting.

Neural Reflectance and Visibility Fields (NeRV) [31] were very recently developed to address the issue of varying non-collocated light source conditions. This research was made in parallel to ours, so we did not perform a comparison with their method. They account for all sources of indirect illumination (including the scene itself) and model bidirectional specular effects, but require prior knowledge of the light conditions (sky color) and add a linear term to the complexity, which becomes $O(N+d)$ instead of $O(N)$.

6. Conclusion

This study proposes a novel adaptation of Neural Radiance Fields able to harness non-correlation effects (such as that of shadows) in satellite images and use them as a source of information rather than a source of perturbation with regards to 3D shape estimation.

To this end, an explicit light transport model is used to simultaneously take into account occlusion and varying illumination effects (shading). We propose to implicitly model the spectral irradiance with a learned light source visibility and ambient sky color, to resolve the ambiguities that this inverse problem implies while maintaining linear complexity.

Our experiments show that S-NeRF succeeds in generating realistic images in previously unseen viewing and lighting conditions, and in estimating the surface altitude with an accuracy of a few meters. Moreover, our methodology shows promise in detecting shadows, in recovering the albedo of surfaces with persistent shadows, and in localizing and removing transient objects.

References

[1] Sai Bi, Zexiang Xu, Pratul Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020. 3, 4

[2] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. *arXiv preprint arXiv:2007.09892*, 2020. 4

[3] Ksenia Bittner and Marco Korner. Automatic large-scale 3d building shape refinement using conditional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1887–1889, 2018. 3

[4] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. *arXiv preprint arXiv:1707.05776*, 2017. 8

[5] M. Carvalho, B. Le Saux, P. Trouvé-Peloux, F. Champagnat, and A. Almansa. Multitask learning of height and semantics from aerial images. *IEEE Geoscience and Remote Sensing Letters*, 17(8):1391–1395, 2020. 3

[6] Carlo De Franchis, Enric Meinhardt-Llopis, Julien Michel, Jean-Michel Morel, and Gabriele Facciolo. An automatic and modular stereo pipeline for pushbroom images. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2014. 3

[7] Valérie Demarez, Florian Helen, Claire Marais-Sicre, and Frédéric Baup. In-season mapping of irrigated crops using landsat 8 and sentinel-1 time series. *Remote Sensing*, 11(2):118, 2019. 1

[8] Gabriele Facciolo, Carlo De Franchis, and Enric Meinhardt-Llopis. Automatic 3d reconstruction from multi-date satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 57–66, 2017. 8

[9] Klaus Gwinner, Ralf Jaumann, Ernst Hauber, Harald Hoffmann, Christian Heipke, Jürgen Oberst, Gerhard Neukum, Veronique Ansan, J Bostelmann, Alexander Dumke, et al. The high resolution stereo camera (hrsc) of mars express and its approach to science analysis and mapping for mars and its satellites. *Planetary and Space Science*, 126:93–138, 2016. 1

[10] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007. 3

[11] Patrick Knobelreiter, Christian Reinbacher, Alexander Shekhovtsov, and Thomas Pock. End-to-end training of hybrid cnn-crf models for stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2339–2348, 2017. 3

[12] Thomas Krauß, Pablo d'Angelo, and Lorenz Wendt. Cross-track satellite stereo for 3d modelling of urban areas. *European Journal of Remote Sensing*, 52(sup2):89–98, 2019. 2

[13] Bertrand Le Saux, Naoto Yokoya, Ronny Hänsch, and Myron Brown. 2019 ieee grss data fusion contest: Large-scale semantic 3d reconstruction. *IEEE Geoscience and Remote Sensing Magazine (GRSM)*, 7(4):33–36, 2019. 2, 6, 8

[14] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *arXiv preprint arXiv:2007.11571*, 2020. 8

[15] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 3, 4, 5

[16] Ricardo Martin-Brualla, Rohit Pandey, Sofien Bouaziz, Matthew Brown, and Dan B Goldman. Gelato: Generative latent textured objects. In *European Conference on Computer Vision*, pages 242–258. Springer, 2020. 3

[17] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. *arXiv preprint arXiv:2008.02268*, 2020. 8

[18] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 4

[19] Fergus McClean, Richard Dawson, and Chris Kilsby. Implications of using global digital elevation models for flood risk analysis in cities. *Water Resources Research*, 56(10), 2020. 1

[20] 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. 1, 2, 3, 4, 5, 6, 7

[21] Lichao Mou and Xiao Xiang Zhu. Im2height: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network. *arXiv preprint arXiv:1802.10249*, 2018. 3

[22] Michael Oechsle, Michael Niemeyer, Christian Reiser, Lars Mescheder, Thilo Strauss, and Andreas Geiger. Learning implicit surface light fields. In *2020 International Conference on 3D Vision (3DV)*, pages 452–462. IEEE, 2020. 3

[23] Roland Perko, Hannes Raggam, and Peter M Roth. Mapping with pléiades—end-to-end workflow. *Remote Sensing*, 11(17):2052, 2019. 3

[24] Rongjun Qin, Jiaojiao Tian, and Peter Reinartz. 3d change detection–approaches and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 122:41–56, 2016. 1

[25] Ewelina Rupnik and Marc Deseilligny. *More surface detail with One-Two-Pixel Matching*. PhD thesis, IGN-Laboratoire MATIS, 2019. 2, 3

[26] Ewelina Rupnik, Marc Pierrot-Deseilligny, and Arthur Delorme. 3d reconstruction from multi-view vhr-satellite images in micmac. *ISPRS Journal of Photogrammetry and Remote Sensing*, 139:201–211, 2018. 2, 3

[27] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 3

[28] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *arXiv preprint arXiv:2007.02442*, 2020. 6, 8

[29] Marc Simard, Keqi Zhang, Victor H Rivera-Monroy, Michael S Ross, Pablo L Ruiz, Edward Castañeda-Moya, Robert R Twilley, and Ernesto Rodriguez. Mapping height and biomass of mangrove forests in everglades national park with srtm elevation data. *Photogrammetric Engineering & Remote Sensing*, 72(3):299–311, 2006. 1

[30] Vincent Sitzmann, Julien NP Martel, Alexander W Bergman, David B Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *arXiv preprint arXiv:2006.09661*, 2020. 6, 7

[31] 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. *arXiv preprint arXiv:2012.03927*, 2020. 8

[32] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d scene representation and rendering. *arXiv preprint arXiv:2010.04595*, 2020. 8

[33] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. *arXiv preprint arXiv:2012.02190*, 2020. 8

[34] Jure Zbontar, Yann LeCun, et al. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(1):2287–2318, 2016. 3