# Sat-NeRF: Learning Multi-View Satellite Photogrammetry With Transient Objects and Shadow Modeling Using RPC Cameras

Roger Marí Gabriele Facciolo Thibaud Ehret
Université Paris-Saclay, CNRS, ENS Paris-Saclay, Centre Borelli, 91190, Gif-sur-Yvette, France

https://centreborelli.github.io/satnerf

## Abstract

*We introduce the Satellite Neural Radiance Field (Sat-NeRF), a new end-to-end model for learning multi-view satellite photogrammetry in the wild. Sat-NeRF combines some of the latest trends in neural rendering with native satellite camera models, represented by rational polynomial coefficient (RPC) functions. The proposed method renders new views and infers surface models of similar quality to those obtained with traditional state-of-the-art stereo pipelines. Multi-date images exhibit significant changes in appearance, mainly due to varying shadows and transient objects (cars, vegetation). Robustness to these challenges is achieved by a shadow-aware irradiance model and uncertainty weighting to deal with transient phenomena that cannot be explained by the position of the sun. We evaluate Sat-NeRF using WorldView-3 images from different locations and stress the advantages of applying a bundle adjustment to the satellite camera models prior to training. This boosts the network performance and can optionally be used to extract additional cues for depth supervision.*

## 1. Introduction

High-resolution satellite imagery is a valuable resource for countless economic activities, many of them based on knowledge of the geometry of the Earth's surface and its changes. This has triggered the development of a number of pipelines capable of highly accurate depth estimation from disparity using multiple satellite views [4, 11, 12, 15, 19, 44, 45]. The output large-scale 3D models are usually represented using discrete point clouds or digital surface models (DSMs) of a certain resolution.

The latest works in 3D modeling from multiple views show that it is possible to achieve a superior representation of a 3D object or scene by *learning* it as a continuous function or field $\mathcal{F}$ [49]. Neural rendering methods learn $\mathcal{F}$ by integrating differentiable rendering techniques into a neural network. The tasks of novel view synthesis and 3D reconstruction are then solved implicitly, as the network is trained to figure out which geometry and color radiances fit the camera projection mappings of the different views.

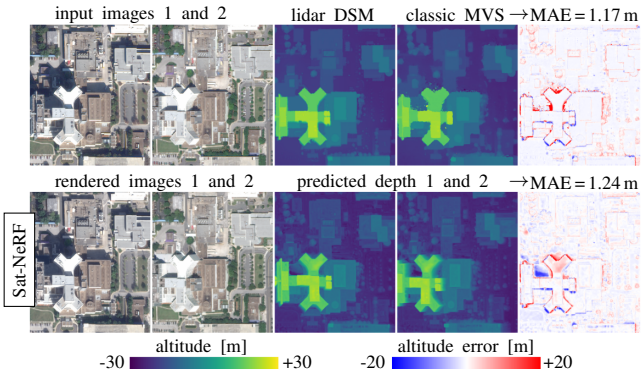


Figure 1. Images 1 and 2 exhibit color inconsistencies (e.g. shadows, cars, vegetation), hindering the direct use of NeRF. Sat-NeRF overcomes these problems and learns to render realistic views and underlying geometry. The digital surface model (DSM) derived from the network depth predictions is compared with a lidar equivalent, obtaining a mean absolute altitude error (MAE) similar to that of multi-view stereo (MVS) relying on handcrafted features.

Neural radiance fields (NeRFs) have gained great popularity in the field of neural rendering [35, 49]. In this paper, we introduce a NeRF variant architecture that achieves state-of-the-art results in novel view synthesis and 3D reconstruction from high-resolution satellite imagery in the wild. We refer to our variant as Satellite NeRF or Sat-NeRF. The original NeRF approach is not adapted to satellite images, e.g. because of the specificities of the camera models, the large distance between the cameras and the scene or the appearance inconsistencies of multi-date collections [14, 33]. Sat-NeRF addresses these challenges using some of the latest advances in NeRFs [13, 14, 33] and adapting well-known tools for satellite image processing. As a result, the model learns highly accurate 3D geometry, similar to that obtained with state-of-the-art stereo pipelines relying on handcrafted features (Figure 1).

Our contributions consist of:

- A NeRF architecture and loss function robust to the radiometric inconsistencies of multi-date satellite imagery, comprising shadows caused by a single non-static light source (the sun) and small transient objects (mainly trees or cars in open-air parkings).
- A point sampling strategy adapted to satellite camera models. The rational polynomial camera (RPC) model [16,20] of each input image is directly used to cast rays in the object space. This RPC-based strategy provides independence to the satellite system and improves the results obtained with approximate pinhole cameras.
- A study of the advantages of correcting RPC inconsistencies before training, e.g. by means of a bundle adjustment [31]. We show that eluding this step leads to a drop in the performance of the model. In addition, we detail how to reuse the sparse point cloud employed in the bundle adjustment to improve geometry learning.

We evaluate Sat-NeRF on different areas of interest covering $256 \times 256$ m each, using $\sim$10-20 RGB crops from multi-date WorldView-3 images for training [8, 27]. A lidar digital surface model (DSM) of resolution 0.5 m/pixel is used as ground truth model to assess the geometry. Sat-NeRF is compared to other NeRF variants [14, 35] as well as a state-of-the-art traditional satellite stereo pipeline [12]. We also publish the code and data used for this article.

## 2. Related work

Current state-of-the-art 3D reconstruction pipelines for satellite images typically follow multi-view stereo approaches, which can outperform sophisticated true multi-view software [18, 37]. Due to the complexity of the task, satellite stereo pipelines can still be improved in a number of aspects. Some of the most important limitations are:

- The 3D reconstruction usually follows the estimation of a dense disparity map using matching strategies derived from the semi-global matching algorithm [4, 11, 12, 19, 24]. Therefore, human-crafted features and cost functions are at the core of the methodology.
- The selection of suitable stereo pairs to estimate disparity is another major challenge. Criteria based on image metadata (e.g. acquisition dates, incidence angles, etc.) have proven to be useful, but do not guarantee the best choice [15, 19, 22].
- The lack of consensus on how the geometry derived from multiple stereo pairs should be refined or aggregated. Local point-wise operations are common to merge altitude values derived from different pairs, e.g. median [11, 19, 32] or k-medians [15]. However, recent work has shown that deep learning approaches can greatly improve the result, e.g. by exploiting geometric priors related to urban areas. [6, 7, 28, 48].
- Very often, it is necessary to make adjustments or parameter tuning to handle different sources or types of satellite images [37, 55].

Neural rendering represents an opportunity to find a natural solution to the previous issues, as it automatically learns the optimal features and operations adapted to each individual 3D scene. The main advantage of traditional pipelines is preserved as no explicit geometry supervision is required: the learning is self-supervised and based solely on the color of the input images. This is a key difference with respect to other state-of-the-art deep learning methods for DSM generation from satellite imagery [7, 17, 18, 48], which depend on ground truth geometry models.

### 2.1. Neural Radiance Fields

NeRF [35] represents a static scene as a continuous volumetric function $\mathcal{F}$, encoded by a fully-connected neural network. $\mathcal{F}$ predicts the emitted RGB color $\mathbf{c} = (r, g, b)$ and a non-negative scalar volume density $\sigma$ at a 3D point $\mathbf{x} = (x, y, z)$ of the scene seen from a viewing direction $\mathbf{d} = (d_x, d_y, d_z)$, i.e.

$$\mathcal{F} : (\mathbf{x}, \mathbf{d}) \mapsto (\mathbf{c}, \sigma) \tag{1}$$

Multi-view consistency is encouraged by restricting the network to predict the volume density $\sigma$ based only on the spatial coordinates $\mathbf{x}$, while allowing the color $\mathbf{c}$ to be predicted as a function of both $\mathbf{x}$ and the viewing direction $\mathbf{d}$. The dependency of $\mathbf{c}$ on the viewing direction allows to recreate specular reflections caused by to static light sources.

Given a set of input views and their camera poses, the training strategy is based on rendering the color of individual rays traced across the scene and projected onto the known pixels. Individual rays are chosen randomly, encouraging gradient flow at those ray intersections where the surface of the scene is susceptible of being located. Each ray $\mathbf{r}$ is defined by a point of origin $\mathbf{o}$ and a direction vector $\mathbf{d}$. The color $\mathbf{c}(\mathbf{r})$ of a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ is computed as

$$\mathbf{c}(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i \mathbf{c}_i. \tag{2}$$

The rendered color $\mathbf{c}(\mathbf{r})$ results from the weighted integration of the colors $\mathbf{c}_i$ predicted at different points of the ray $\mathbf{r}$, which is discretized into $N$ 3D points $\mathbf{x}_i$ between the near and far bounds of the scene, $t_n$ and $t_f$. Each point $\mathbf{x}_i$ in $\mathbf{r}$ is obtained as $\mathbf{x}_i = \mathbf{o} + t_i\mathbf{d}$, where $t_i \in [t_n, t_f]$.

Following (2), the weight given to the color predicted for each point $\mathbf{x}_i$ of $\mathbf{r}$ is defined by a transmittance factor $T_i$ representing the probability that light reaches the point

without hitting any other particle, and an alpha compositing value $\alpha_i$ encoding the opacity. Both $T_i$ and $\alpha_i$ are set according to the volume density $\sigma_i$ predicted for $\mathbf{x}_i$:

$$\alpha_i = 1 - \exp(-\sigma_i \delta_i); \quad T_i = \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{3}$$

where $\delta_i$ is the distance between two consecutive points along the ray, i.e. $\delta_i = t_{i+1} - t_i$. Higher values of $\sigma_i$ will result in larger opacity $\alpha_i$, indicating that $\mathbf{x}_i$ possibly belongs to a non-transparent surface. Occlusions are handled by the transmittance $T_i$, equal to the cumulative product of the inverse opacity. Even if $\mathbf{x}_i$ is given a large $\sigma_i$, $T_i$ only allows it to contribute decisively to the rendered color if it is not preceded by previous opaque points in the ray.

Given (3), the depth $d(\mathbf{r})$ observed in the direction of a ray $\mathbf{r}$ can be rendered in a similar manner to (2) [13, 43] as

$$d(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i t_i. \tag{4}$$

NeRF is optimized by minimizing the mean squared error (MSE) between the rendered color and the real color of the input images, at the positions where the rays project:

$$\sum_{\mathbf{r} \in \mathcal{R}} \| \mathbf{c}(\mathbf{r}) - \mathbf{c}_{\text{GT}}(\mathbf{r}) \|_2^2, \tag{5}$$

where $\mathbf{c}_{\text{GT}}(\mathbf{r})$ is the observed color of the pixel intersected by the ray $\mathbf{r}$, and $\mathbf{c}(\mathbf{r})$ is the color predicted by the NeRF using (2). $\mathcal{R}$ is the set of rays in each input batch.

### 2.2. NeRF variants

NeRF assumes that the density, radiance and illumination of the target 3D scene is constant. This is a strong limitation, as these conditions are rarely encountered outside laboratory settings. Many variants have been proposed to address this problem. In this section we briefly review three models that inspired our work.

**NeRF-W** [33] or *NeRF in the Wild* gains robustness to radiometric variation and transient objects by learning to separate transient phenomena from the static scene. An extra head of fully-connected layers is used to predict a transient color $\mathbf{c}^\tau$ and volume density $\sigma^\tau$ for each input point, in addition to the usual $\mathbf{c}$ and $\sigma$. The transient outputs are linearly combined with the static ones to render the color of each ray. NeRF-W also uses the transient head to emit an uncertainty coefficient $\beta$, which measures the confidence of the network that a point belongs to a transient object. The value of $\beta$ is used in the loss function to reduce the impact of transient/unreliable pixels in the learning process.

**S-NeRF** [14] or *Shadow NeRF* is, to the best of our knowledge, the first attempt to apply NeRF for multi-view satellite photogrammetry. S-NeRF showed the benefits in geometry estimation of simultaneously exploiting the direction of solar rays to learn the amount of sunlight $s_i$ that reaches each point $\mathbf{x}_i$ of the scene. The direction of solar rays is a common metadata of satellite images. Our work can be seen as an extension of S-NeRF that incorporates a modeling of transient objects similar to [33] and a representation of the camera models more adapted to satellite data.

**DS-NeRF** [13] or *Depth Supervised NeRF* incorporates a depth supervision term to the loss function to accelerate the learning and reduce the amount of input images. The depth supervision term exploits a sparse set of 3D points that belong to the surface of the scene, which can be easily retrieved using structure-from-motion (SfM) pipelines. SfM is a common pre-processing step in NeRF frameworks, as it can estimate the camera poses needed to cast the input rays. A similar strategy to DS-NeRF is used in [43], which converts the sparse point clouds into dense depth priors.

Other recent NeRF variants are yet to be investigated in the context of satellite imagery. Some works are focused on achieving smoother scene representations or reducing the number of input views: e.g. DietNeRF [25] introduces an auxiliary semantic loss to maximize similarity between high-level features instead of RGB colors; Mip-NeRF [3] prevents blurring and aliasing in collections of images with different resolutions; PixelNeRF [54] describes a framework that is trained across multiple scenes and learns priors that can generalize to unseen scenes with few available images. Recent undergoing research is also progressing to extend NeRFs to dynamic scenes [40–42, 51], to gain efficiency and reduce the training time [23, 36, 52, 53] or to handle complex illumination settings, under arbitrary, multiple light sources [5, 47] or near-darkness conditions [34].

## 3. Method

Sat-NeRF represents the scene as a static surface with an albedo color, i.e. the intrinsic color of static objects. The model learns to predict the geometry and the albedo color simultaneously with a set of additional outputs, which seek to explain the transient phenomena observed in the input images without inducing changes in the scene geometry.

We train the model following the ray casting strategy of NeRF (Section 2.1). Unlike the original NeRF (1), we assume a Lambertian surface and omit the color dependence on viewing angles, but add two new cues. The inputs are

- $\mathbf{x}$: 3-valued vector with the spatial coordinates of points located in the scene volume. $\mathbf{x}$ is part of a ray $\mathbf{r}$.
- $\boldsymbol{\omega}$: 3-valued direction vector encoding the direction of solar rays. For each input image, $\boldsymbol{\omega}$ is extracted from the azimuth and elevation angles $(\theta, \phi)$ that indicate the position of the sun in the satellite image metadata.
- $\mathbf{t}_j$: $N^{(\text{t})}$-valued embedding vector, learned as a function of the image index $j$. The objective of $\mathbf{t}_j$ is to featurize the transient elements in the $j$-th view that

cannot be explained by the position of the sun given by $\boldsymbol{\omega}$. We manually set $N^{(\mathrm{t})} = 4$.

The volumetric function of Sat-NeRF then writes $\mathcal{F} : (\mathbf{x}, \boldsymbol{\omega}, \mathbf{t}_j) \mapsto (\sigma, \mathbf{c}_\mathrm{a}, s, \mathbf{a}, \beta)$, where the outputs are

- $\sigma$: scalar encoding the volume density at location $\mathbf{x}$.
- $\mathbf{c}_\mathrm{a}$: albedo RGB color, which depends exclusively on the geometry, i.e. the spatial coordinates $\mathbf{x}$.
- $s$: shadow-aware shading scalar, learned as a function of $\mathbf{x}$ and the solar rays direction vector $\boldsymbol{\omega}$.
- $\mathbf{a}$: ambient RGB color, independent of scene geometry, that defines a global hue bias according to the position of the sun given by $\boldsymbol{\omega}$.
- $\beta$: uncertainty coefficient related to the probability that the color of $\mathbf{x}$ is explained by a transient object.

### 3.1. Shadow-aware irradiance model

This section describes how Sat-NeRF predicts the color $\mathbf{c}(\mathbf{r})$ of a ray $\mathbf{r}$ projected onto a certain pixel. We keep the rendering as in (2) and (4), with the transmittance and opacity factors as defined in (3), but adopt the shadow-aware irradiance model proposed in S-NeRF [14] to compute the color $\mathbf{c}$ at each point $\mathbf{x}$ of a ray $\mathbf{r}$:

$$\mathbf{c}(\mathbf{x}, \boldsymbol{\omega}, \mathbf{t}_j) = \mathbf{c}_\mathrm{a}(\mathbf{x}) \cdot s(\mathbf{x}, \boldsymbol{\omega}) + (1 - s(\mathbf{x}, \boldsymbol{\omega})) \cdot \mathbf{a}(\boldsymbol{\omega}), \quad (6)$$

where $\mathbf{c}(\mathbf{x}, \boldsymbol{\omega}, \mathbf{t}_j)$ substitutes $\mathbf{c}_i$ in (2). The shading scalar $s(\mathbf{x}, \boldsymbol{\omega})$ takes values between 0 and 1 and is used to add shadows by darkening the albedo (Figure 2). Ideally, $s \approx 1$ in those 3D points directly illuminated by the sun, whose color should be entirely explained by the albedo $\mathbf{c}_\mathrm{a}(\mathbf{x})$.

In addition, (6) attempts to capture the bluish hues of shadows [2, 30] by means of the ambient color $\mathbf{a}(\boldsymbol{\omega})$, which contributes to the points where $s$ takes values closer to 0. In practice, we find that the direction of the solar rays $\boldsymbol{\omega}$ is narrowly related to the acquisition date (especially if the satellite passes at the same hours of the day), as shown in Figure 2. Thus, $\mathbf{a}(\boldsymbol{\omega})$ ends up capturing ambient irradiance due to a mixture of phenomena, which is related to $\boldsymbol{\omega}$ but also date-specific conditions like weather or seasonal changes.

As observed in S-NeRF [14], the shading scalar $s(\mathbf{x}, \boldsymbol{\omega})$ in (6) can produce unrealistic results for solar rays directions that are not seen in the training data. This can be minimized by adding a *solar correction* term to the loss:

$$L_\mathrm{SC}(\mathcal{R}_\mathrm{SC}) = \sum_{\mathbf{r} \in \mathcal{R}_\mathrm{SC}} \left( \sum_{i=1}^{N_\mathrm{SC}} (T_i - s_i)^2 + 1 - \sum_{i=1}^{N_\mathrm{SC}} T_i \alpha_i s_i \right), \quad (7)$$

where $\mathcal{R}_\mathrm{SC}$ is a secondary batch of solar correction rays. Note that the rays in $\mathcal{R}_\mathrm{SC}$ follow the direction of solar rays $\boldsymbol{\omega}$, while the rays in $\mathcal{R}$, in the main term of the loss (5), follow the viewing direction of the camera.

The solar correction term (7) uses the learned geometry, encoded by the transmittance $T_i$ and opacity $\alpha_i$ (3), to further supervise the learning of the shadow-aware shading $s(\mathbf{x}, \boldsymbol{\omega})$. The first part of (7) enforces that, for each ray $\mathbf{r}$ in $\mathcal{R}_\mathrm{SC}$, the $s_i$ predicted at the $i$-th point should resemble $T_i$, i.e. high values before reaching the visible surface, low values afterwards (both $s_i$ and $T_i$ take values between 0 and 1). The second part of (7) encourages that the integration of $s$ over $\mathbf{r}$ reaches 1, since non-occluded and non-shadow areas have to be mostly explained by the albedo in (6).

### 3.2. Uncertainty weighting for transient objects

Similarly to NeRF-W [33], we use the task-uncertainty learning approach introduced in [26] to gain robustness to transient objects by means of $\beta$. In our context, transient objects are punctual local changes across the input images that cannot be explained by the static surface or the available metadata, like the position of the sun. The irradiance model (6) does not handle transient objects explicitly. As a result, we observe that $s$ and $\sigma$ usually try to account for them, leading to wrong depth predictions, as shown in Figure 4. Thanks to $\beta$, Sat-NeRF is given some margin to ignore the color inconsistencies caused by these objects.

The uncertainty prediction $\beta$ weights the contribution of each ray to the MSE between rendered and known colors:

$$L_\mathrm{RGB}(\mathcal{R}) = \sum_{\mathbf{r} \in \mathcal{R}} \frac{\|\mathbf{c}(\mathbf{r}) - \mathbf{c}_\mathrm{GT}(\mathbf{r})\|_2^2}{2\beta'(\mathbf{r})^2} + \left( \frac{\log \beta'(\mathbf{r})}{2} + \eta \right), \quad (8)$$

where $\beta'(\mathbf{r}) = \beta(\mathbf{r}) + \beta_\mathrm{min}$. In (8), we use $\beta_\mathrm{min} = 0.05$ and $\eta = 3$ to avoid negative values in the logarithm. The role of the logarithm in $L_\mathrm{RGB}$ is to prevent $\beta$ from converging to infinity to solve the problem. In this way the model is forced to find a compromise between the uncertainty coefficients $\beta$ and the differences of colors.

The $\beta(\mathbf{r})$ associated to a ray $\mathbf{r}$ is obtained by integrating the uncertainty predictions across the $N$ points of $\mathbf{r}$:

$$\beta(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i \beta(\mathbf{x}_i, \mathbf{t}_j), \quad (9)$$

where $\beta(\mathbf{x}_i, \mathbf{t}_j)$, is the uncertainty coefficient predicted at the $i$-th point of $\mathbf{r}$. Sat-NeRF learns to predict the uncertainty $\beta$ at each point of the scene based on its spatial coordinates $\mathbf{x}$ (some areas are more likely to exhibit transient objects, e.g. open-air parkings in Figure 2) and on the transient embedding vector $\mathbf{t}_j$ of each input training image. Depending on each view, the areas typically affected by transient objects will arbitrarily differ to a greater or lesser extent with respect to the albedo. Note that the embedding vector $\mathbf{t}_j$ is learned from the image index $j$ during training.[1]

We find that it is better to start using $\beta$ after the second epoch, when the shadow-aware shading $s$ is already well initialized. Otherwise the model may use $\beta$ to overlook shadow areas instead of trying to explain them with $s$. Thus, we replace (8) with (5) in the first two epochs.

[1] At test time, we use an arbitrary $\mathbf{t}_j$ selected from the training set.

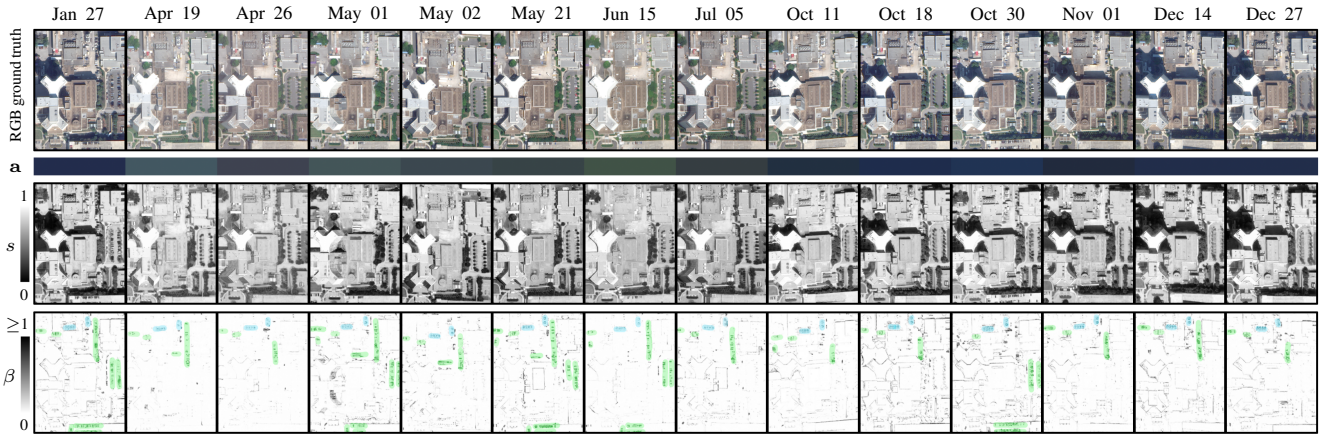


Figure 2. The shading scalar $s$ related to the solar rays direction $\boldsymbol{\omega}$ learns shadows and material roughness. We observe that $\boldsymbol{\omega}$ is narrowly related to the acquisition date, causing the ambient color $\mathbf{a}$ associated with the low values of $s$ (see (6)) to capture a mixture of phenomena, including seasonal changes reflected in the vegetation. The uncertainty prediction $\beta$ does not affect shadows and concentrates on small color inconsistencies, mostly caused by cars in open-air parkings (green marks), large fans in rooftops (blue marks) or building edges.

### 3.3. Point sampling from satellite RPC models

Sat-NeRF casts rays directly using the RPC camera models of a set of satellite images. The RPC model is widely used for optical satellite imagery, as it allows to describe complex acquisition systems independently of satellite-specific physical modeling [1, 20]. Each RPC is defined by a projection function (to project 3D points onto image pixels) and its inverse, the localization function.

The use of RPCs in a NeRF framework represents an improvement with respect to previous work with satellite data. In S-NeRF [14] the RPC model of each input view is replaced with a custom simplified pinhole camera matrix, which is the common representation used in NeRF for close-range imagery [35]. The RPC-based sampling described here corresponds to a more general approach, which also leads to better results (see Section 4).

We denote the minimum and maximum altitudes of the scene as $h_{\text{min}}$ and $h_{\text{max}}$, respectively.[2] The ray that crosses the scene and intersects the pixel $\mathbf{p}$ of the $j$-th image is modeled as a straight line between an initial and a final 3D point, i.e. $\mathbf{x}_{\text{start}}$ and $\mathbf{x}_{\text{end}}$. These boundary points are obtained by localizing the pixel $\mathbf{p}$ at $h_{\text{min}}$ and $h_{\text{max}}$, using the RPC localization function $\mathcal{L}_j$ of the $j$-th image:

$$\mathbf{x}_{\text{start}} = \mathcal{L}_j(\mathbf{p}, h_{\text{max}})_{\text{ECEF}}; \quad \mathbf{x}_{\text{end}} = \mathcal{L}_j(\mathbf{p}, h_{\text{min}})_{\text{ECEF}}, \tag{10}$$

where the subindex ECEF indicates that the 3D points returned by the localization function $\mathcal{L}_j$ are converted to the Earth-centered, Earth-fixed coordinate system (or geocentric system), to work in a cartestian system of reference.

Given $\mathbf{x}_{\text{start}}$ and $\mathbf{x}_{\text{end}}$, the origin $\mathbf{o}$ and direction vector $\mathbf{d}$ of the ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ that intersects the pixel $\mathbf{p}$ of the $j$-th image are expressed as

$$\mathbf{o} = \mathbf{x}_{\text{start}}; \quad \mathbf{d} = \frac{\mathbf{x}_{\text{end}} - \mathbf{x}_{\text{start}}}{\|\mathbf{x}_{\text{end}} - \mathbf{x}_{\text{start}}\|_2}. \tag{11}$$

The point of maximum altitude, $\mathbf{x}_{\text{start}}$, which is the closest to the camera, is taken as the origin $\mathbf{o}$ of the ray. The boundaries of the ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, i.e. $[t_{\text{min}}, t_{\text{max}}]$, are set as $t_{\text{min}} = 0$ and $t_{\text{max}} = \|\mathbf{x}_{\text{end}} - \mathbf{x}_{\text{start}}\|_2$. Since working with ECEF coordinates is impractical, due to large values used in the representation, we normalize all ray points in the interval $[-1, 1]$ using an offset subtraction and scaling procedure similar to the one used in the RPC functions [20]. The set of 3D points resulting from localizing all pixels in the input images at $h_{\text{min}}$ and $h_{\text{max}}$ is used to compute the offset and scale in each spatial dimension.

### 3.4. RPC refinement for improved performance

Bundle adjustment approaches are a common good practice in remote sensing to correct inconsistencies between a collection of RPC models observing the same scene [21, 31, 32, 38]. In particular, bundle adjustment methods correct the RPCs by minimizing the reprojection error of a set of reference points seen across the images [50].

In absence of a prior RPC refinement, a 3D point projected with different raw RPC functions often falls on non-coincident image points, by a distance of up to tens of pixels [21]. This would cause a systematic loss of accuracy in any NeRF methodology for satellite imagery, because rays traced from corresponding pixels of different views would not intersect at an exact point in the object space. To prevent this situation, before training Sat-NeRF, we apply the bundle adjustment method described in [31], which performs a relative correction of the RPC models of all input images. The reference points used by the bundle adjustment are derived from correspondences of SIFT keypoints [29].

[2] The altitude bounds $[h_{\text{min}}, h_{\text{max}}]$ can be selected in various ways, e.g. from a large-scale elevation model extracted from a low-resolution data.

While the refined RPC models directly increase the accuracy of the point sampling strategy described in Section 3.3, a prior bundle adjustment can also improve the Sat-NeRF performance in other ways. DS-NeRF [13] discussed how the training of a NeRF can benefit from a sparse set of previously known 3D points, under the idea that such points can be easily produced using SfM pipelines. In the case of satellite imagery, the bundle adjustment produces an equivalent set of sparse 3D points derived from image features [10, 31, 39, 45]. Based on this idea, we explore the benefits of adding the depth-supervision term proposed in [13] to the loss of our Sat-NeRF model:

$$L_{\text{DS}}(\mathcal{R}_{\text{DS}}) = \sum_{\mathbf{r}\in\mathcal{R}_{\text{DS}}} w(\mathbf{r}) \left(d(\mathbf{r}) - \|\mathbf{X}(\mathbf{r}) - \mathbf{o}(\mathbf{r})\|_2\right)^2, \tag{12}$$

where $d(\mathbf{r})$ is the depth (4) predicted for a ray $\mathbf{r}$, whose origin point is $\mathbf{o}(\mathbf{r})$. If $\mathbf{r}$ intersects $\mathbf{X}(\mathbf{r})$, a known 3D point, then $\|\mathbf{X}(\mathbf{r}) - \mathbf{o}(\mathbf{r})\|_1$ is the target depth that should be learned. $\mathcal{R}_{\text{DS}}$ denotes a batch of rays that intersects known 3D points. Since the pixel coordinates associated to these 3D points are already provided by the bundle adjustment, all rays in $\mathcal{R}_{\text{DS}}$ can be defined as explained in Section 3.3. Normalized coordinates between $[-1, 1]$ are used in (12) to represent points in the object space, for consistency with Section 3.3.

Similarly to [13], we only use $L_{\text{DS}}$ in the initial 25% of training iterations. In our experience, this proportion is usually enough to gain accuracy in the learned geometry. Observe that the contribution of each depth-supervision ray $\mathbf{r}$ in $\mathcal{R}_{\text{DS}}$ is weighted by $w(\mathbf{r})$ in (12), where $w(\mathbf{r})$ is a scalar set according to the reprojection error of each point $\mathbf{X}(\mathbf{r})$ provided by the bundle adjustment.

### 3.5. Multi-task loss and network architecture

The main term of the Sat-NeRF loss function is the $L_{\text{RGB}}$ defined in (8), which is complemented by the solar correction term $L_{\text{SC}}$ (7) and the depth-supervision term $L_{\text{DS}}$ (12). The complete loss function can be expressed as

$$L = L_{\text{RGB}}(\mathcal{R}) + \lambda_{\text{SC}} L_{\text{SC}}(\mathcal{R}_{\text{SC}}) + \lambda_{\text{DS}} L_{\text{DS}}(\mathcal{R}_{\text{DS}}), \tag{13}$$

where $\lambda_{\text{SC}}$ and $\lambda_{\text{DS}}$ are an arbitrary weight given to each secondary term. For feasibility reasons, to keep training time below 20 h, we do not use $\lambda_{\text{SC}} > 0$ and $\lambda_{\text{DS}} > 0$ simultaneously and train using only one of the secondary terms at a time, as shown in Section 4. We empirically find $\lambda_{\text{SC}} = 0.1$ and $\lambda_{\text{DS}} = 1000$ to provide good results, to keep the secondary terms sufficiently relevant but below the magnitude of $L_{\text{RGB}}$. For depth supervision, we used ∼2k-10k bundle adjustment points depending on the output of [31] for each area of interest. $\mathcal{R}$, $\mathcal{R}_{\text{SC}}$ and $\mathcal{R}_{\text{DS}}$ have the same batch size.

The architecture of Sat-NeRF is shown in Figure 3. The main block of fully-connected layers, with $h$ channels per layer, is dedicated to the prediction of the static properties of the scene: the volume density $\sigma$ and the albedo color $\mathbf{c}_{\text{a}}$. A secondary head is added with fewer layers and half as many channels per layer to estimate the shading scalar $s$ based on the direction of solar rays $\boldsymbol{\omega}$ and the vector of $h$ geometry-related features learned by the main block. Lastly, two single-layer heads are used to predict the uncertainty coefficient $\beta$ and the ambient color $\mathbf{a}$, from the transient embedding vector $\mathbf{t}_j$ and $\boldsymbol{\omega}$, respectively.

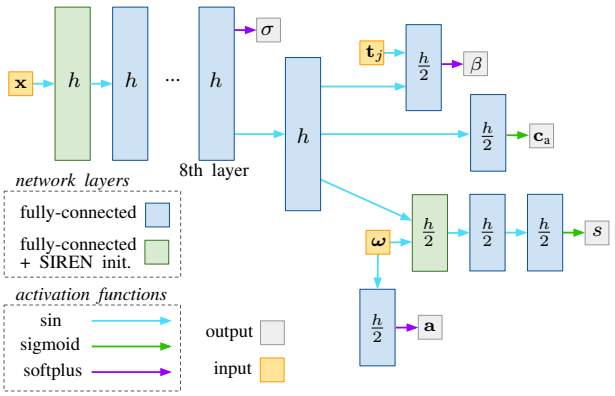


Figure 3. Sat-NeRF network architecture, where $\mathbf{x}$ are the input spatial coordinates, $\boldsymbol{\omega}$ is the direction of solar rays and $\mathbf{t}_j$ is the learned transient embedding of image $j$. The model predicts the volume density $\sigma$, the components of the irradiance model (6), i.e. albedo color $\mathbf{c}_{\text{a}}$, shading scalar $s$, ambient color $\mathbf{a}$, and an uncertainty coefficient $\beta$ to weight the impact of transient objects.

We employ SIREN layers using the initialization proposed in [46], as suggested in [14], and found the use of a softplus activation function to predict $\sigma$ to be essential to achieve satisfactory results. The uncertainty $\beta$ is also produced by a softplus function [33], which yields a smoother optimization problem compared to the usual ReLU [3]. The rest of outputs result from sigmoid functions, since they are directly related to normalized RGB values and have to be in the interval $[0, 1]$. In this work we set $h = 512$, but the value of $h$ should be adjusted according to the resolution and the area observed in the input images.

## 4. Evaluation

We evaluate Sat-NeRF on different areas of interest (AOI) of the *2019 IEEE GRSS Data Fusion Contest* [8, 27], which provides 26 Maxar WorldView-3 images collected between 2014 and 2016 over the city of Jacksonville, Florida, US. From this data, we take as input a set of RGB crops of varying size, around $800 \times 800$ pixels, with a resolution of 0.3 m/pixel at nadir, covering $256 \times 256$ m for each AOI. The indices of the selected AOIs and the number of training and test images that we used are listed in Table 1.

In all conducted experiments we use a single NeRF

| Area index | 004 | 068 | 214 | 260 |
|---|---|---|---|---|
| # train/test | 9/2 | 17/2 | 21/3 | 15/2 |
| Alt. bounds [m] | [-24, 1] | [-27, 30] | [-29, 73] | [-30, 13] |

Table 1. Number of training and test images used for each area, and the altitude bounds of the scene considered in each case.

model, trained with an Adam optimizer starting with a learning rate of $5e^{-4}$, which is decreased at every epoch by a factor $\gamma = 0.9$ according to a step scheduler. The batch size is 1024 rays, and each ray $\mathbf{r}$ is discretized into 64 uniformly distributed 3D points. Training takes 300k iterations to converge, resulting in $\sim$10 h if a single batch of rays is used at each training iteration, or $\sim$20 h if a secondary term for solar correction or depth supervision is added to the loss (trained on a GPU with 16 GB RAM). We used bundle adjusted RPCs (Section 3.4) unless otherwise noted.

### 4.1. Ablation study

We evaluated the Sat-NeRF model starting from a simple NeRF and gradually adding new components. To this end, we propose three categories of experiments that are discussed below. Table 2 shows the quantitative results.

**Category 1.** Rows 0-3 are an ablation study dedicated to the irradiance model and the solar correction term described in Section 3.1. We verify that the S-NeRF irradiance model outperforms a basic NeRF and is strengthened by the solar correction term. Comparing our results with the ones reported in the original S-NeRF work [14] reveals the impact of the proposed RPC-based point sampling and bundle adjustment detailed in Section 3, to which we attribute the difference between the metrics of row 0 and row 3.

**Category 2.** Rows 4-5 assess our Sat-NeRF model, which incorporates the uncertainty prediction of $\beta$ and employs (8) as main term of the loss function. These rows show that the uncertainty modeling improves both the learned geometry and the novel view synthesis, as illustrated in Figure 4. Compared to the best S-NeRF results (row 3), Sat-NeRF (row 4) provides higher PSNR/SSIM and similar or even smaller altitude MAE without requiring a solar correction term. This insight could be exploited in settings that cannot afford additional training time to process a secondary batch of rays for solar correction. If we add the solar correction term (7) to the Sat-NeRF loss (8), the altitude MAE decreases even more: row 5 outperforms all the previous configurations across all AOIs.

**Category 3.** Rows 6-7 were added to demonstrate the benefits derived from using a prior bundle adjustment to refine the RPC models of the satellite images, as explained in Section 3.4. Comparing rows 6 and 5 illustrates how the use of the original raw RPCs induces a performance drop: both PSNR/SSIM and altitude MAE are worse. Lastly, row 7 adds the depth supervision term (12) to the Sat-NeRF loss

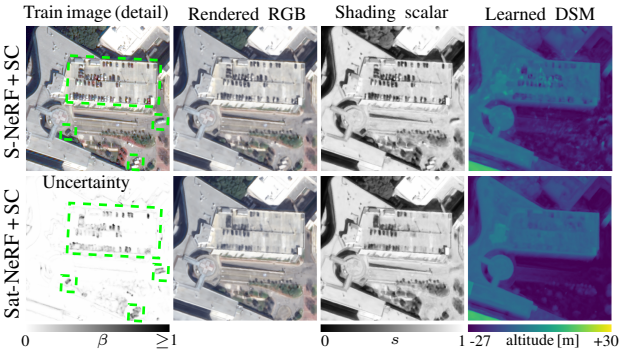


Figure 4. The uncertainty coefficient $\beta$ learned by Sat-NeRF helps to improve the geometry learning with respect to S-NeRF [14]. In addition to shadows and textures, the shading scalar $s$ of the irradiance model (6) usually attempts to account for transient objects (e.g. cars, marked in green). Sat-NeRF uses $\beta$ to minimize the contribution of transient objects to the loss function (8), thus allowing the geometry and $s$ to learn to ignore their color. In practice, we notice that $s$ still retains some transient objects, but the learned geometry is much better, as shown in the corresponding DSM.

(8), to leverage the sparse point cloud generated by the bundle adjustment. We observe that this strategy improves the metrics in a similar measure to the solar correction term (row 5), even attaining better altitude MAE in some cases (e.g. areas 214 and 260).

### 4.2. Comparison to traditional stereo pipelines

Sat-NeRF learns high quality 3D models, similar in accuracy to those obtained with satellite stereo pipelines relying on traditional algorithms for stereo matching [4, 12]. In this work, we compare the DSMs produced by Sat-NeRF with a multi-view stereo DSM of the same area generated with S2P [12, 15], the satellite stereo pipeline that won the *2016 IARPA Multi-View Stereo 3D Mapping Challenge* [9].

We follow the methodology described in [15] to produce the S2P DSMs. For each AOI, we manually select 10 stereo pairs for disparity estimation. The selection criterion prioritizes pairs with an angle between views of 5 to 45 degrees, with a maximum incidence angle of 40 degrees for each view. Within this set, we take the 10 pairs with closer acquisition dates and run S2P. The RPCs used by S2P were the same used to train Sat-NeRF, i.e. all RPCs are bundle adjusted using [31]. The 10 pairwise models are fused into a single DSM by taking the median altitude at each cell. To maximize the quality of the S2P DSMs we used the panchromatic product of the WorldView-3 images, instead of the RGB crops employed to train Sat-NeRF. Considering that the RGB images have a compressed dynamic (integer values in $[0, 255]$), i.e with less texture and more saturated areas, the Sat-NeRF DSMs are very encouraging compared to the state of the art with manual pair selection.

As shown in Figure 5 and 6, structures are more detailed in Sat-NeRF DSMs, but S2P provides more regular

| Area index | PSNR ↑ 004 | 068 | 214 | 260 | SSIM ↑ 004 | 068 | 214 | 260 | Altitude MAE [m] ↓ 004 | 068 | 214 | 260 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0. S-NeRF + SC [14] | — | — | — | — | 0.344 | 0.459 | 0.384 | 0.416 | 4.418 | 3.644 | 4.829 | 7.173 |
| 1. NeRF | 20.72 | 20.99 | 18.42 | 20.08 | 0.640 | 0.826 | 0.808 | 0.773 | 3.562 | 5.157 | 8.654 | 5.131 |
| 2. S-NeRF | 25.86 | 24.29 | 24.16 | 21.37 | 0.864 | 0.897 | 0.936 | 0.816 | 1.790 | 1.550 | 4.118 | 3.251 |
| 3. S-NeRF + SC | 26.29 | 24.67 | 24.67 | 21.05 | 0.874 | 0.899 | 0.942 | 0.819 | 1.387 ● | 1.372 | 3.036 | 2.534 |
| 4. Sat-NeRF | 26.32 | 25.11 | 24.99 | 21.79 | 0.877 | 0.912 | 0.946 | 0.842 | 1.402 | 1.348 | 2.848 | 2.488 |
| 5. Sat-NeRF + SC | 26.58 | 25.05 | 25.31 | 21.93 | 0.881 | 0.909 | 0.949 | 0.845 | **1.323** ● | **1.244** ● | 2.029 ● | 1.853 ● |
| 6. Sat-NeRF + SC (no BA) | 21.63 | 23.12 | 24.34 | 21.30 | 0.578 | 0.881 | 0.940 | 0.822 | 1.559 | 1.390 | 2.220 | 1.895 |
| 7. Sat-NeRF + DS | 26.30 | 24.99 | 25.36 | 21.86 | 0.877 | 0.904 | 0.951 | 0.837 | 1.416 | 1.321 ● | **1.670** ● | **1.627** ● |
| S2P (10 pairs) [15] | — | — | — | — | — | — | — | — | 1.370 ● | 1.174 ● | 1.811 ● | 1.640 ● |

Table 2. Numerical results using the test images (unseen in training). Experiments 1-7 use RPC-based sampling (Section 3.3). $\lambda_{SC} = 0.05$ in experiments 0 and 3, for coherence with [14], otherwise $\lambda_{SC} = 0.1$ and $\lambda_{DS} = 1000$. The methods with least altitude MAE are given gold ●, silver ● and bronze ● medals. Sat-NeRF + solar correction (SC) or depth supervision (DS) are the most awarded NeRF variants.

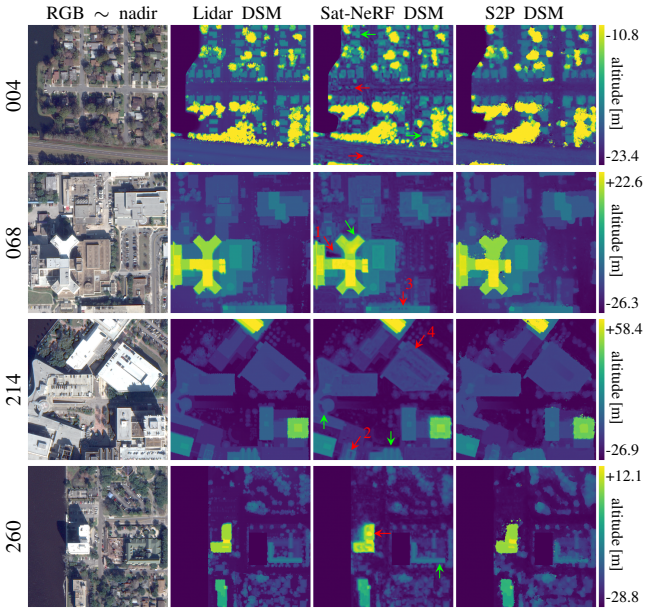


Figure 5. Left to right: ground truth lidar DSM, Sat-NeRF DSM and S2P DSM [12]. The Sat-NeRF DSM corresponds to the model with lowest altitude MAE, in bold in Table 2. Compared to S2P, structures are sharper and more detailed in Sat-NeRF DSMs (green arrows), which are also free of single-point outliers. However, Sat-NeRF produces more local irregularities: roofs and roads are less flat (red arrows). Certain roofs exhibit holes, that can be explained by their constant changes across the training sequence. The uncertainty coefficient $\beta$ can only absorb occasional inconsistencies, which does not include roofs under construction (arrows 1-2) or roofs that are unusually free of parked cars (arrows 3-4). For clarity, we provide an RGB view of the area from a near-nadir perspective. Water bodies are masked in the DSMs.

surfaces. Numerically, the global altitude MAE obtained with Sat-NeRF can be slightly better compared to the S2P DSMs (Table 2, last row), which are affected by single-point outliers. Future work points to hybrid methods or the aggregation of contour-preserving regularization techniques.

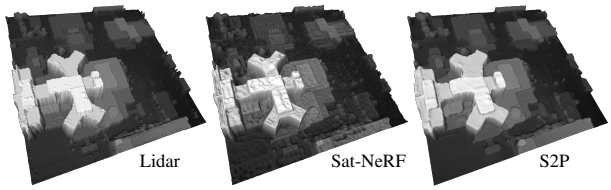


Figure 6. 3D visualization of the lidar, Sat-NeRF and S2P DSMs shown in Figure 5 (068). Compared to S2P, Sat-NeRF provides finer details and sharper edges but exhibits local irregularities.

## 5. Conclusion

We introduced Sat-NeRF, a NeRF variant adapted for collections of multi-view satellite images in the wild. The geometry and appearance of permanent structures are simultaneously learned using a main backbone, while shadows and transient objects are learned in parallel using secondary heads.

The proposed method achieves state-of-the-art results in novel view synthesis and 3D modeling from satellite imagery. It also highlights the benefits of incorporating well-known techniques for satellite image processing into a NeRF framework. In particular, we show how to represent the input cameras using the RPC models characteristic of satellite images, instead of the pinhole cameras commonly used in NeRF for close-range imagery. We also demonstrate the advantages of applying a bundle adjustment step before training time, to improve reconstruction quality and, optionally, to provide additional cues for depth supervision.

## 6. Acknowledgements

This work was supported by a grant from Région Île-de-France. It was also partly financed by Office of Naval research grant N00014-17-1-2552, DGA Astrid project « filmer la Terre » n° ANR-17-ASTR-0013-01, MENRT, and Kayrros. This work was performed using HPC resources from GENCI–IDRIS (grants 2021-AD011012453

and 2022-AD011011801R1) and from the "Mésocentre" computing center of CentraleSupélec and ENS Paris-Saclay supported by CNRS and Région Île-de-France (http://mesocentre.centralesupelec.frs).

## References


[1] Roland Akiki, Roger Marí, Carlo De Franchis, Jean-Michel Morel, and Gabriele Facciolo. Robust rational polynomial camera modelling for SAR and pushbroom imaging. In *2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 7908–7911, 2021. 5

[2] Vicente Arévalo, Javier González, and Gregorio Ambrosio. Shadow detection in colour high-resolution satellite images. *International Journal of Remote Sensing*, 29(7):1945–1963, 2008. 4

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5855–5864, 2021. 3, 6

[4] Ross A Beyer, Oleg Alexandrov, and Scott McMichael. The Ames Stereo Pipeline: NASA's open source software for deriving and processing terrain data. *Earth and Space Science*, 5(9):537–548, 2018. 1, 2, 7

[5] 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

[6] Ksenia Bittner, Pablo d'Angelo, Marco Körner, and Peter Reinartz. DSM-to-LoD2: Spaceborne stereo digital surface model refinement. *Remote Sensing*, 10(12):1926, 2018. 2

[7] Ksenia Bittner, Marco Körner, and Peter Reinartz. Late or earlier information fusion from depth and spectral data? Large-scale digital surface model refinement by hybrid-cGAN. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1471–1478, 2019. 2

[8] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D Hager, and Myron Brown. Semantic stereo for incidental satellite images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1524–1532, 2019. 2, 6

[9] Marc Bosch, Zachary Kurtz, Shea Hagstrom, and Myron Brown. A multiple view stereo benchmark for satellite imagery. In *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–9, 2016. 7

[10] Michael J Broxton, Ara V Nefian, Zachary Moratto, Taemin Kim, Michael Lundy, and Aleksandr V Segal. 3D lunar terrain reconstruction from Apollo images. In *International Symposium on Visual Computing*, pages 710–719, 2009. 6

[11] Pablo d'Angelo and Georg Kuschk. Dense multi-view stereo from satellite imagery. In *2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 6944–6947, 2012. 1, 2

[12] Carlo De Franchis, Enric Meinhardt-Llopis, Julien Michel, Jean-Michel Morel, and Gabriele Facciolo. An automatic and modular stereo pipeline for pushbroom images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2(3):49–56, 2014. 1, 2, 7, 8

[13] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3, 6

[14] Dawa Derksen and Dario Izzo. Shadow neural radiance fields for multi-view satellite photogrammetry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1152–1161, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[15] 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 (CVPR) Workshops*, pages 57–66, 2017. 1, 2, 7, 8

[16] Clive S Fraser, Gene Dial, and Jacek Grodecki. Sensor orientation via RPCs. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60(3):182–194, 2006. 2

[17] Jian Gao, Jin Liu, and Shunping Ji. Rational polynomial camera model warping for deep learning based satellite multi-view stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6148–6157, 2021. 2

[18] Alvaro Gómez, Gregory Randall, Gabriele Facciolo, and Rafael Grompone von Gioi. An experimental comparison of multi-view stereo approaches on satellite images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 844–853, 2022. 2

[19] Ke Gong and Dieter Fritsch. DSM generation from high resolution multi-view stereo satellite imagery. *Photogrammetric Engineering & Remote Sensing*, 85(5):379–387, 2019. 1, 2

[20] Jacek Grodecki. IKONOS stereo feature extraction–RPC approach. In *ASPRS Annual Conference*, 2001. 2, 5

[21] Jacek Grodecki and Gene Dial. Block adjustment of high-resolution satellite images described by rational polynomials. *Photogrammetric Engineering & Remote Sensing*, 69(1):59–68, 2003. 5

[22] Yilong Han, Shugen Wang, Danchao Gong, Yue Wang, and X Ma. State of the art in digital surface modelling from multi-view high-resolution satellite images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 5(2):351–356, 2020. 2

[23] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5875–5884, 2021. 3

[24] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008. 2

[25] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting NeRF on a diet: Semantically consistent few-shot view synthesis.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5885–5894, 2021. 3

[26] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7482–7491, 2018. 4

[27] Bertrand Le Saux, Naoto Yokoya, Ronny Hansch, Myron Brown, and Greg Hager. 2019 data fusion contest [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 7(1):103–105, 2019. 2, 6

[28] Lukas Liebel, Ksenia Bittner, and Marco Körner. A generalized multi-task learning approach to stereo DSM filtering in urban areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:213–227, 2020. 2

[29] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 5

[30] Haijian Ma, Qiming Qin, and Xinyi Shen. Shadow segmentation and compensation in high resolution satellite images. In *2008 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, volume 2, pages 1036–1039, 2008. 4

[31] Roger Marí, Carlo de Franchis, Enric Meinhardt-Llopis, Jérémy Anger, and Gabriele Facciolo. A generic bundle adjustment methodology for indirect RPC model refinement of satellite imagery. *Image Processing On Line*, 11:344–373, 2021. 2, 5, 6, 7

[32] Roger Marí, Carlo de Franchis, Enric Meinhardt-Llopis, and Gabriele Facciolo. To bundle adjust or not: A comparison of relative geolocation correction strategies for satellite multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019. 2, 5

[33] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7210–7219, 2021. 1, 3, 4, 6

[34] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul Srinivasan, and Jonathan T Barron. NeRF in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[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, 2020. 1, 2, 5

[36] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 3

[37] Ozge C Ozcanli, Yi Dong, Joseph L Mundy, Helen Webb, Riad Hammoud, and Victor Tom. A comparison of stereo and multiview 3-D reconstruction using cross-sensor satellite imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 17–25, 2015. 2

[38] Ozge C Ozcanli, Yi Dong, Joseph L Mundy, Helen Webb, Riad Hammoud, and Tom Victor. Automatic geo-location correction of satellite imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 307–314, 2014. 5

[39] Hongbo Pan, Tao Huang, Ping Zhou, and Zehua Cui. Self-calibration dense bundle adjustment of multi-view Worldview-3 basic images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176:127–138, 2021. 6

[40] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5865–5874, 2021. 3

[41] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. HyperNeRF: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), 2021. 3

[42] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10318–10327, 2021. 3

[43] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[44] 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. 1

[45] David E Shean, Oleg Alexandrov, Zachary M Moratto, Benjamin E Smith, Ian R Joughin, Claire Porter, and Paul Morin. An automated, open-source pipeline for mass production of digital elevation models (DEMs) from very-high-resolution commercial stereo satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116:101–117, 2016. 1, 6

[46] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 6

[47] 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 (CVPR)*, pages 7495–7504, 2021. 3

[48] Corinne Stucker and Konrad Schindler. ResDepth: Learned residual stereo reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 184–185, 2020. 2

[49] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-

Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, volume 39, pages 701–727, 2020. 1

[50] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International Workshop on Vision Algorithms*, pages 298–372, 1999. 5

[51] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-NeRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[52] Guo-Wei Yang, Wen-Yang Zhou, Hao-Yang Peng, Dun Liang, Tai-Jiang Mu, and Shi-Min Hu. Recursive-NeRF: An efficient and dynamically growing NeRF. *arXiv preprint arXiv:2105.09103*, 2021. 3

[53] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5752–5761, 2021. 3

[54] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, 2021. 3

[55] Kai Zhang, Noah Snavely, and Jin Sun. Leveraging vision reconstruction pipelines for satellite imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019. 2