



**HAL**  
open science

# **SparseSat-NeRF: Dense Depth Supervised Neural Radiance Fields for Sparse Satellite Images**

Lulin Zhang, Ewelina Rupnik

► **To cite this version:**

Lulin Zhang, Ewelina Rupnik. SparseSat-NeRF: Dense Depth Supervised Neural Radiance Fields for Sparse Satellite Images. ISPRS Annals 2023, 2023. <hal-04205387>

**HAL Id: hal-04205387**

**<https://hal.science/hal-04205387v1>**

Submitted on 12 Sep 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

# SPARSESAT-NeRF: DENSE DEPTH SUPERVISED NEURAL RADIANCE FIELDS FOR SPARSE SATELLITE IMAGES

Lulin Zhang<sup>1,2</sup>, Ewelina Rupnik<sup>2</sup>

<sup>1</sup> Université de Paris, Institut de Physique du Globe de Paris, CNRS, Paris, France - lzhang@ipgp.fr

<sup>2</sup> Université de Gustave Eiffel, IGN-ENSG, LaSTIG, Saint-Mandé, France - ewelina.rupnik@ign.fr

Commission II, WG II/1

**KEY WORDS:** neural radiance fields, depth supervision, multi-view stereo matching, satellite images, sparse views

**ABSTRACT:**

Digital surface model generation using traditional multi-view stereo matching (MVS) performs poorly over non-Lambertian surfaces, with asynchronous acquisitions, or at discontinuities. Neural radiance fields (NeRF) offer a new paradigm for reconstructing surface geometries using continuous volumetric representation. NeRF is self-supervised, does not require ground truth geometry for training, and provides an elegant way to include in its representation physical parameters about the scene, thus potentially remedying the challenging scenarios where MVS fails. However, NeRF and its variants require many views to produce convincing scene’s geometries which in earth observation satellite imaging is rare. In this paper we present SparseSat-NeRF (SpS-NeRF) – an extension of Sat-NeRF adapted to sparse satellite views. SpS-NeRF employs dense depth supervision guided by cross-correlation similarity metric provided by traditional semi-global MVS matching. We demonstrate the effectiveness of our approach on stereo and tri-stereo Pléiades 1B/WorldView-3 images, and compare against NeRF and Sat-NeRF. The code is available at <https://github.com/LulinZhang/SpS-NeRF>

## 1. INTRODUCTION

Satellite imagery and 3D digital surface models (DSM) derived from them are used in a wide range of applications, including urban planning, environmental monitoring, geology, disaster rapid mapping, etc. Because in many of those applications the quality of the DSMs is essential, a vast amount of research has been undertaken in the last few decades to enhance their precision and fidelity.

Classically, DSMs are derived from images with semi-global dense image matching (Hirschmuller, 2005, Pierrot-Deseilligny and Paparoditis, 2006) (SGM) followed by a depth map fusion step (Rupnik et al., 2018) or more recently with hybrid (Hartmann et al., 2017) or end-to-end (Chang and Chen, 2018) deep learning based approaches. A new way of solving the dense image correspondence problem is proposed by Neural Radiance Fields (NeRF) (Mildenhall et al., 2020). Unlike the traditional methods, NeRF leverages many views to learn to represent the scene as a continuous volumetric representation (i.e., 3D radiance field). This representation is defined by a neural network and has the unique capacity to incorporate different aspects of the physical scene, e.g., surface radiance or illumination sources.

Despite the tremendous *hype* around the neural radiance fields, the *state-of-the-art* results remain conditioned on a rather large number of input images. With few input images, NeRF has the tendency to fit incorrect geometries, possibly because it does not know that the majority of the scene is composed of empty space and opaque surfaces. In a space-borne setting, it is rare to have many images of a given scene acquired under multiple viewing angles within a defined time window. With the exception of the Pléiades *persistent surveillance* collection mode, the most common configuration includes a stereo pair or a stereo-triplet of images. Previous works have attempted to apply NeRF on satellite images, including S-NeRF (Derksen and Izzo, 2021) and Sat-NeRF (Marí et al., 2022), but they bypassed

the problem of sparse input views by using multi-date images.

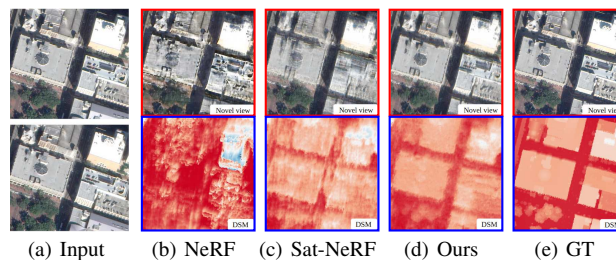


Figure 1. **SpS-NeRF (Ours) and competitors.** NeRF variants trained on 2 images. Our network leverages dense depth information calculated by stereo-matching on downsampled images. Compared to NeRF and Sat-NeRF, SpS-NeRF renders sharper novel views (□), reconstructs more reliable 3D geometries (□).

**Contributions.** In this paper, we present a NeRF variant that attains *state-of-the-art* results in novel view synthesis and 3D reconstruction using sparse single-date satellite images. Inspired by the architecture proposed in (Marí et al., 2022), we lay down its extension adapted to sparse satellite views and refer to it as SparseSat-NeRF, or SpS-NeRF. Precisely

- we adopt low resolution dense depths generated with traditional MVS for supervision and consequently enable the generation of novel views and 3D surfaces from sparse satellite images. We demonstrate the efficiency of this method on as few as two and three input views;
- we increase the robustness of the predicted views and surfaces by incorporating correlation-based uncertainty into the guidance of NeRF using depth information;

- we provide in-depth analysis of the benefits of adding dense depth supervision into the NeRF architecture.

## 2. RELATED WORK

**Image matching vs NeRF** Traditional stereo or multi-view stereo (MVS) matching approaches (Hirschmuller, 2005, Gallup et al., 2007, Bleyer et al., 2011, Bulatov et al., 2011, Furukawa and Ponce, 2009) establish correspondences between pixels in different images by calculating patch-based similarity metrics such as correlation coefficient or mutual information. Although these methods often produce impressive results in favourable matching conditions, they tend to struggle with images lacking texture, at discontinuities or in the presence of non-Lambertian surfaces such as forest canopies or icy surfaces. Learning-based MVS methods (Bittner et al., 2019, Stucker and Schindler, 2020, Gao et al., 2021, Gómez et al., 2022, Huang et al., 2018) attempt and often succeed in overcoming those challenges, however, they require very precise and up-to-date ground truth depth maps for training and those are difficult to obtain in a satellite setting. In contrast, NeRF offers a self-supervised deep learning approach without resorting to ground truth geometry, and relying exclusively on images at input. Because it operates on a truly single-pixel level, it overcomes the shortcomings of traditional patch-based methods (Buades and Facciolo, 2015). Furthermore, NeRF defined as a function of radiance accumulated along image rays opens up the possibility to model physical parameters of the scene such as reflectance of scene’s materials.

**NeRF variants towards fewer input views.** Vanilla NeRF relies exclusively on RGB values to maintain consistency between training images. Consequently, it requires a large number of images to resolve the ambiguity embedded within the modelled volumetric fields. This greediness of NeRF has been addressed across several research works, which focus on adding priors through incorporating semantic information, or sparse/ dense depth supervision. The latter is particularly interesting because *Structure from Motion* (SfM) or the subsequent MVS matching provide reliable depth information. Additionally, in satellite imaging, the dense depth information is available without extra processing through, e.g., the global SRTM elevation model.

**Learning priors with semantics.** PixelNeRF demonstrates excellent results in novel view synthesis over an unknown scene with only one view. To this end, (Yu et al., 2021) extend the canonical NeRF with deep features and pre-train the entire architecture enabling its generalization to new scenes. Analogously, DietNeRF (Jain et al., 2021) adopts a pre-trained visual transformer (ViT) and enforces consistent semantics across all views (including the novel view). SinNeRF (Xu et al., 2022) extends further this idea by combining global semantics using the self-supervised Dino ViT, then instead of using image feature embeddings leverages the classification token representation, thus making their approach less susceptible to pixel misalignments between views. SinNeRF also employs local texture regularization and depth supervision through depth warping to novel views. MVSNeRF (Chen et al., 2021) borrows from multi-view stereo matching in projecting 2D convolutional neural networks (CNN) features to planes sweeping through the scene. 3D CNNs are then used to extract a neural encoding volume, which once regressed translate to RGB and density.

**Sparse depth supervision** DS-NeRF (Deng et al., 2022) was the first to propose sparse depth supervision using 3D points

obtained from SfM. The authors propose an adapted ray sampling strategy and a depth termination loss weighted by the 3D point’s reprojection error. Sat-NeRF (Marí et al., 2022) applied the same sparse depth supervision in multi-date satellite images, reducing the number of training images to approximately 15. Interestingly, Sat-NeRF architecture includes scene’s physical parameters specific to earth observation satellites such as albedo and solar correction (for asynchronous acquisitions).

**Dense depth supervision.** NerfingMVS (Wei et al., 2021) combines learning-based multi-view stereo with NeRF for indoor mapping. Starting from a set of sparse 3D points output from SfM, NerfingMVS first trains a monocular dense depth prediction network. Consistency checks between per-view predicted depths serve as error maps and guide the following ray sampling in the final NeRF optimization. In their most view-sparse scenario 35 images are available for training. Similarly, Roessle et al. (Roessle et al., 2022) (referred to in the following as DDPNeRF) incorporate dense depth supervision in their NeRF variant. However, unlike in NerfingMVS where dense depths are guessed from single views, DDPNeRF learns a depth completion network from sparse depth maps. This, together with an explicit depth loss, makes it a better performing method. Experiments demonstrate good performance with as few as 18 train images.

The above methods resort to learning-based dense depth prediction because their focus is on indoor scenes, with textureless surfaces where traditional MVS might fail. In our real world satellite scenario this is, in general, less of an issue and we demonstrate that dense image matching with SGM is good enough to guide the NeRF optimization.

## 3. METHODOLOGY

Our method builds on top of Sat-NeRF (Marí et al., 2022) and DDPNeRF (Roessle et al., 2022). We borrow from Sat-NeRF the general architecture save for the transient objects and solar correction modelling as we deal with synchronous acquisitions. We add a dense depth supervision and depth loss similar to the one proposed in DDPNeRF, but we replace the depth loss distance metric and define an uncertainty based on SGM’s correlation maps. The workflows of NeRF, Sat-NeRF and SpS-NeRF are illustrated in Figure 2.

### 3.1 Neural Radiance Fields Preliminaries

NeRF (Mildenhall et al., 2020) learns a continuous volumetric representation of the scene from a set of images characterised by the sensor position and the viewing direction. This representation is defined by a fully-connected (non-convolutional) deep network. It samples  $N$  query points along each camera ray through the 3D field and integrate the weighted radiance to render each pixel, and optimize the NeRF network  $F_{\Theta}$  by imposing the rendered pixel values to be close to the training images. For each query point, NeRF simultaneously models the volume density  $\sigma$  and the emitted radiance  $\mathbf{c} = (r, g, b)$  at that 3D point  $\mathbf{x} = (x, y, z)$  from the viewing angle  $\mathbf{d} = (d_x, d_y, d_z)$ :

$$F_{\Theta}(\mathbf{x}, \mathbf{d}) = (\mathbf{c}, \sigma). \quad (1)$$

Each camera ray  $\mathbf{r}$  is defined by a point of origin  $\mathbf{o}$  and a direction vector  $\mathbf{d}$  as  $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ . Each query point in  $\mathbf{r}$  is defined as  $\mathbf{x}_i = \mathbf{o} + t_i\mathbf{d}$ , where  $t_i$  locates between the near and

far bounds of the scene,  $t_n$  and  $t_f$ . The rendered pixel value  $\mathbf{C}(\mathbf{r})$  of ray  $\mathbf{r}$  is calculated as:

$$\mathbf{C}(\mathbf{r}) = \sum_{i=1}^N T_i \alpha_i c_i ,$$

$$\alpha_i = 1 - e^{-\sigma_i \delta_i} , \quad T_i = \prod_{j=1}^{i-1} (1 - \alpha_j) , \quad \delta_i = t_{i+1} - t_i ,$$
(2)

where  $\alpha_i$  represents the opacity of the current query point  $\mathbf{x}_i$ ,  $T_i$  stands for the probability that  $\mathbf{x}_i$  reaches the ray origin  $\mathbf{o}$  without being blocked. In other words, the color  $c_i$  of the current query point  $\mathbf{x}_i$  contributes to the accumulated color  $\mathbf{C}(\mathbf{r})$  only if it is highly opaque (i.e., large value of  $\alpha_i$ ) and there are no opaque particles in front of it (i.e., high value of  $T_i$ ).

### 3.2 SparseSat-NeRF

**Pre-processing.** Following the Sat-NeRF’s pipeline, the RPC-poses of our input images are first refined in a bundle adjustment. Then, for  $N$  input images, we run  $N$  independent SGMs to obtain a low-resolution depth map for each image (i.e., scale factor of  $2^{-2}$ ). We choose to rely on low resolution depths to (i) avoid biasing our SpS-NeRF towards the SGM solution; and (ii) because high resolution depths might provide incomplete depth information at challenging surfaces (e.g., low texture). The depth maps are accompanied by similarity metrics that will further act as depth prediction quality measures in supervising the SpS-NeRF. In our case, the metric is the cross-correlation map. If low-resolution depth maps are not available, the SGM depths can be replaced by coarse global DEM such as SRTM (with the similarity metric globally set to a constant value).

**Depth supervision.** Our goal is to include the depth prior in SpS-NeRF optimization. Analogously to the formulation presented in (Roessle et al., 2022), three ingredients are necessary for that end: (i) a way to predict the depth of a given ray by accumulating radiance fields throughout the optimized volume; (ii) description of the sample distribution along a given ray; and finally (iii) input depth maps and their associated uncertainty metrics. The depth prediction along a ray  $D(\mathbf{r})$  can be calculated as:

$$D(\mathbf{r}) = \sum_{i=1}^N T_i \alpha_i t_i ,$$
(3)

where the depth  $t_i$  of the current sample point  $i$  would contribute to the accumulated depth  $D(\mathbf{r})$  if it is opaque, ignoring the sample points in front of  $t_i$ . To characterise the samples’ distribution along the ray we follow the standard deviation equation (Roessle et al., 2022):

$$S(\mathbf{r})^2 = \sum_{i=1}^N T_i \alpha_i (t_i - D(\mathbf{r}))^2 .$$
(4)

Here, lower standard deviation values indicate samples located around the estimated depths and lead to sharper edges at object surfaces. We now define an equivalent uncertainty driven by our input data, i.e., the similarity metrics produced by SGM:

$$\Sigma(\mathbf{r}) = \gamma \cdot (1 - \text{corr}(\mathbf{r})) + m ,$$
(5)

where  $\text{corr}(\mathbf{r})$  is the cross-correlation similarity for a ray sample at the input depth,  $\gamma$  and  $m$  are the normalizing scaling and

shift parameters, in our experiments empirically set to 1.0 and  $10e^{-4}$ . The uncertainty measure (Equation (5)) intervenes three times during the optimization: (i) as a weight applied to the final depth loss; (ii) as a threshold to determine whether the loss should be activated; and (iii) in guided ray sampling (see next paragraph).

All ingredients combined constitute the depth loss encouraging depths’ predictions  $D(\mathbf{r})$  to be close to the input dense depths  $\bar{D}(\mathbf{r})$ , guided by the input uncertainty:

$$\mathcal{L}_{depth}(\mathbf{r}) = \sum_{\mathbf{r} \in R_{sub}} (\text{corr}(\mathbf{r})(D(\mathbf{r}) - \bar{D}(\mathbf{r}))^2) .$$
(6)

The  $R_{sub}$  is defined as a ray’s subregion where either of the two conditions are satisfied: (1)  $S(\mathbf{r}) > \Sigma(\mathbf{r})$ ; (2)  $|(D(\mathbf{r}) - \bar{D}(\mathbf{r}))| > \Sigma(\mathbf{r})$ . Those bounds favour ray termination within  $(1 \cdot \Sigma)$  from our depth priors (Roessle et al., 2022). Outside this region, the depth loss is inactive or clipped. The depth loss participates in all training iterations.

**Total loss.** Our SpS-NeRF is supervised with the ground truth pixel color  $\bar{\mathbf{C}}(\mathbf{r})$  and the dense depth information  $\bar{D}(\mathbf{r})$  weighted by the quality metric  $\text{corr}(\mathbf{r})$ . Following Equation (2), the color (RGB) of a pixel is rendered through the accumulation of the RGB values of samples along the casted ray. The color loss encourages the predicted pixel colors  $\mathbf{C}(\mathbf{r})$  to be as close as possible to the ground truth colors and is defined on a set  $R$  containing all ray samples (there is no clipping unlike in the depth loss):

$$\mathcal{L}_{color}(\mathbf{r}) = \sum_{\mathbf{r} \in R} \|\mathbf{C}(\mathbf{r}) - \bar{\mathbf{C}}(\mathbf{r})\|_2^2 .$$
(7)

The SpS-NeRF’s total loss is thus a combination of Equation (7) and Equation (6):

$$\mathcal{L} = \mathcal{L}_{color}(\mathbf{r}) + \lambda \mathcal{L}_{depth}(\mathbf{r}) ,$$
(8)

where  $\lambda$  is a weight balancing the color and depth contributions. We empirically found that  $\lambda = \frac{1}{3}$  performs best in urban areas and  $\lambda = \frac{50}{3}$  in rural areas.

**Ray sampling** We adopt guided sampling from (Roessle et al., 2022), whose approach takes advantage of depth cues to efficiently query samples. It substitutes the hierarchical sampling coarse network in the original NeRF. More specifically, the ray samples are divided into two groups queried sequentially. The points of the first group are sampled randomly within the entire scene’s envelope, while the second group of points is concentrated around the known input (train) or predicted (test) surface. The points around the surface are spread following a Gaussian distribution determined by (1) the input depth  $N(\bar{D}(\mathbf{r}), \Sigma(\mathbf{r}))$  for the pixels with input depth information during training; or (2) the estimated depth  $N(D(\mathbf{r}), S(\mathbf{r}))$  for all the pixels during testing, as well as the pixels without input depth during training (e.g., SGM provides no depth in occluded areas). We illustrated the distribution of the rays sampled by this strategy in Figure 3.

## 4. EXPERIMENTS

We conduct experiments on two datasets:

- **Djibouti dataset** located in the Asal-Ghoubbet rift, Republic of Djibouti, introduced in (Labarre et al., 2019) and

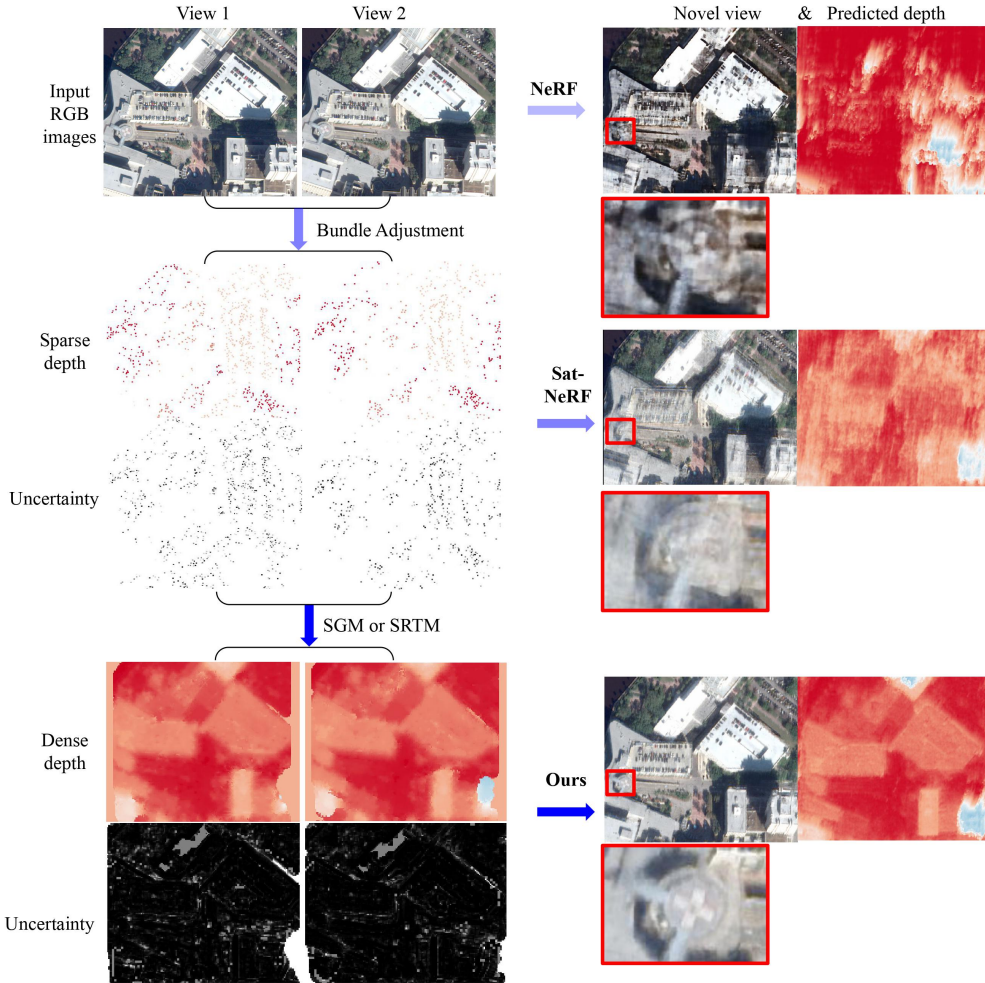


Figure 2. **Workflows of SpS-NeRF (Ours), Sat-NeRF and NeRF.** In our experimental setting we use 2 or 3 satellite images to optimize the neural radiance fields for photo-realistic novel view rendering, and for DSM recovery. Without any depth supervision, NeRF fails to render high quality novel views and DSM. Sat-NeRF incorporates sparse depth information and uses the bundle adjustment re-projection errors as uncertainties to weigh the depth loss; it improves the results, but the artifact remain present due to the insufficient number of training views. SpS-NeRF further employs low resolution dense depth maps from traditional methods such as SGM, and uses the  $(1 - correlation)$  score as uncertainty, and takes advantage of the dense depth to guide sampling along the casted ray, leading to improved performance.

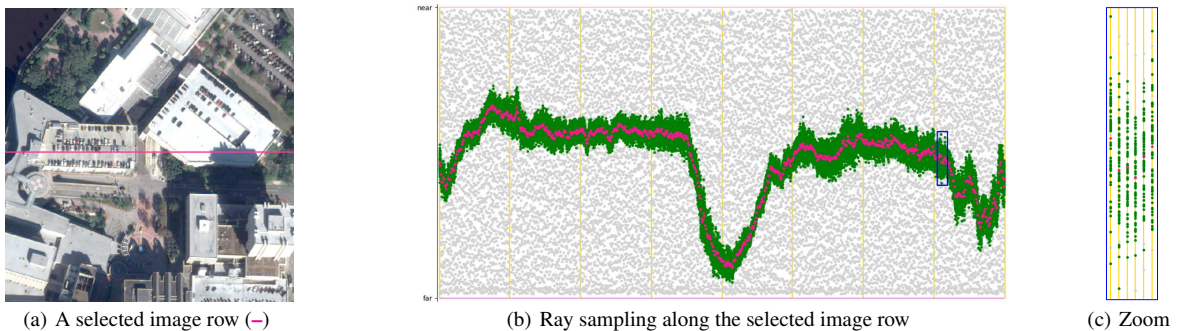


Figure 3. **Ray sampling.** The samples in (b) correspond to the selected image row in (a), while in (c) we zoom over a few ray samples. Similarly to Roessle et al., we divide ray samples in two groups of the same cardinality (i.e.,  $2 \times 64$ ). The first group draws samples (- -) within the near and far planes. At inference, the second group draws samples (- -) following a Gaussian distribution around the estimated dense depths  $D(\mathbf{r})$  (- -) (see Equation (3)), their upper and lower bounds are defined by the estimated standard deviation  $S(\mathbf{r})$  (see Equation (4)). At train time we use the input depths and their corresponding uncertainties  $\{\bar{D}, \Sigma\}$ . The yellow lines (|) represent the rays.

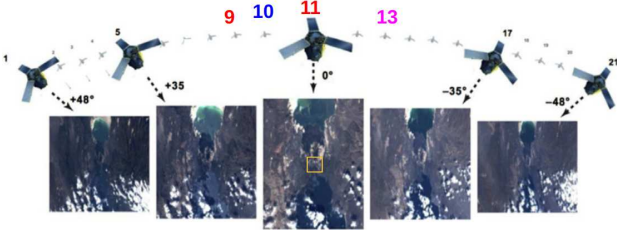


Figure 4. **Djibouti dataset**. The images labeled  $\{9, 11\}$  are used for training 2-views scenario, and the images labeled  $\{9, 11, 13\}$  are used for training 3-views scenario. The image labeled  $\{10\}$  is used for testing both scenarios. The remaining images are ignored. The yellow rectangle ( $\square$ ) represents the area of interest cropped for our experiments.

Image name	Sun elevation	Sun azimuth	Acquisition date y-m-d
007	33.5	158.9	14-12-27 <sup>16:11:09</sup>
008	36	155.0	15-01-21 <sup>16:12:43</sup>
009	36	155.1	15-01-21 <sup>16:12:53</sup>
010	36	155.2	15-01-21 <sup>16:13:08</sup>

Table 1. **DFC2019 dataset, AOI 214**. During training, we use the following subsets of images:  $\{009, 010\}$ ,  $\{008, 009, 010\}$ .

illustrated in Figure 4. It represents a series of 21 multi-angular Pléiades images collected in a single flyby on January 26, 2013. During training we use only two or three RGB cropped images ( $\sim 800 \times 800$  px), with 2m Ground Sampling Distance (GSD).

- **DFC2019 dataset** The 2019 IEEE GRSS Data Fusion Contest (Le Saux et al., 2019) contains different areas of interest (AOI) in the city of Jacksonville, Florida, USA, providing in total 26 WorldView-3 images collected between 2014 and 2016. We choose the AOI 214 as it contains 3 images taken at the same time and use it to train two independent networks: with 2 and 3 views used in the training images. For novel view generation, we choose another image from the dataset and consider it the ground truth. Because SpS-NeRF does not model transient objects, our goal was to minimize the acquisition time gap and respect the seasonality in choosing the novel views. The sun elevation, azimuth and the acquisition time of the 4 selected images are displayed in the table.

#### 4.1 Implementation details

We use Sat-NeRF as the backbone architecture ( $\text{lr}=1e^{-5}$ ,  $\text{decay}=0.9$ ,  $\text{batch\_size}=1024$ ). Our focus is on sparse views captured synchronously from the same orbit thus we disable the uncertainty weighting for transient objects and the solar correction. We also disable the two components for Sat-NeRF because our experiments are conducted on single-epoch images. In contrast to NeRF and Sat-NeRF, SpS-NeRF uses only the coarse architecture (no fine model) with 64 initial samples and 64 guided samples (— and - - in Figure 3). For a fair comparison the number of samples and *importance* samples (i.e., fine model) in NeRF and Sat-NeRF are also 64 each. We optimize SpS-NeRF for 30k iterations, which takes  $\sim 2$  hours on NVIDIA GPU with 40GB RAM. The input low resolution DSMs were computed from images downsampled by a factor of 4 ( $SGM_{scl4}$ ).

#### 4.2 Evaluation

Tests are carried out using 2 and 3 views leading to 4 scenarios:

1.  $DFC_{2v}$ , test on 008 and train on  $\{009, 010\}$ ;
2.  $DFC_{3v}$ , test on 007 and train on  $\{008, 009, 010\}$ ;
3.  $Dji_{2v}$ , test on 10 and train on  $\{9, 11\}$ ;
4.  $Dji_{3v}$ , test on 10 and train on  $\{9, 11, 13\}$ .

We evaluate the performance of SpS-NeRF qualitatively and quantitatively on 2 tasks: (1) novel view synthesis and (2) altitude extraction. Precision metrics are Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index measure (SSIM) (Wang et al., 2004) for view synthesis, and Mean Altitude Error (MAE) for altitude extraction. We differentiate between  $MAE_{in}$  and  $MAE_{out}$  for errors computed on valid pixels and invalid pixels (e.g., due to low correlation or occlusions). The classification into valid and invalid pixels is produced by SGM. Ground truth (GT) images are *true* images not seen during training, while GT DSMs are a LiDAR acquisition for the DFC2019 dataset, and a photogrammetric DSM generated with 21 high-resolution panchromatic Pléiades images ( $GSD=0.5m$ ) for Djibouti dataset. SpS-NeRF is also compared with competitive vanilla NeRF, Sat-NeRF, and DSMs generated with SGM using full-resolution images (i.e.,  $SGM_{scl1}$ ).

#### 4.3 Results & discussion

**Novel view synthesis** Qualitative and quantitative results are given in Figure 5 and Table 2. In the urban DFC2019 dataset NeRF’s and Sat-NeRF’s novel views are poorly rendered. SpS-NeRF provides better quality synthetic views with 2 input images (Figure 5(k)), and further improves the result with 3 input images (Figure 5(l)). In the rural Djibouti dataset, the performance gap between NeRF, Sat-NeRF and SpS-NeRF is less significant, however, in Figure 5 *ghost* artifacts are revealed by NeRF (c), which are attenuated by Sat-NeRF (g) and are not present in SpS-NeRF (o).

**Altitude extraction** The qualitative and quantitative results are in Figure 6 and Table 2. NeRF fails to recover reasonable DSM geometries for all 4 scenarios (a, b, c, d). This is because using only RGB consistency between input images is insufficient to recover the scene’s surface with 2 or 3 images. Adding sparse depth supervision in Sat-NeRF helps to recover rough buildings’ shapes in  $DFC_{3v}$  scenario (f). Nevertheless, it fails at the remaining three scenarios (e, g, h), indicating that sparse depths are not enough to complete the missing information with 2 or 3 input images.

Our SpS-NeRF takes as input dense depths computed with SGM using downsampled images (factor 4). The input depth maps are incomplete (due to occlusions) and imprecise ( $\times 4$  bigger GSD), but SpS-NeRF is able to complete and refine the depth information. We attribute this to the jointly optimized RGB and depth losses. Compared to the SGM result obtained with full-resolution images ( $SGM_{scl1}$ ), SpS-NeRF behaves better close to the outlines of buildings and is free of outliers, but lacks regularization on flat surfaces (see Figure 7). Such local irregularities are a common problem in NeRF (Marí et al., 2022). Adding semantic information to the framework might be a possible solution. Interestingly, SpS-NeRF with 3 views is capable of recovering trees’ canopy surface (see Figure 6 (n)), a task traditionally challenging for traditional patch-based methods such

	PSNR $\uparrow$				SSIM $\uparrow$				MAE <sub>in</sub> $\downarrow$				MAE <sub>out</sub> $\downarrow$	
	DFC <sub>2v</sub>	DFC <sub>3v</sub>	Dji <sub>2v</sub>	Dji <sub>3v</sub>	DFC <sub>2v</sub>	DFC <sub>3v</sub>	Dji <sub>2v</sub>	Dji <sub>3v</sub>	DFC <sub>2v</sub>	DFC <sub>3v</sub>	Dji <sub>2v</sub>	Dji <sub>3v</sub>	DFC <sub>2v</sub>	DFC <sub>3v</sub>
NeRF	12.89	14.56	27.8	35.22	0.65	0.67	0.8	0.94	9.51	6.56	9.72	14.44	13.2	11.98
Sat-NeRF	17.72	18.46	32.3	36.17	0.8	0.83	0.9	<b>0.95</b>	5.89	4.63	9.51	10.11	11.75	7.53
SpS-NeRF	<b>20.2</b>	<b>19.06</b>	<b>32.85</b>	<b>36.26</b>	<b>0.87</b>	<b>0.86</b>	<b>0.92</b>	<b>0.95</b>	<b>3.02</b>	<b>2.86</b>	<b>1.57</b>	<b>1.35</b>	<b>7.77</b>	<b>5.62</b>
SGM <sub>scl1</sub>	/	/	/	/	/	/	/	/	2.77	2.05	1.15	0.81	9.82	6.68

Table 2. **Quantitative metrics.** Best performing metrics in PSNR and SSIM are in bold, while best and second best performing metrics in MAE<sub>in</sub> and MAE<sub>out</sub> are in blue and magenta. SpS-NeRF outperformed NeRF and Sat-NeRF in all the scenarios. SpS-NeRF is less good than SGM<sub>scl1</sub> in altitude extraction on valid pixels (MAE<sub>in</sub>) which we attribute to the lack of regularization. However, SpS-NeRF is better than SGM<sub>scl1</sub> in occluded and poorly textured areas (MAE<sub>out</sub>). Note that no invalid pixels were identified for the Djibouti dataset.

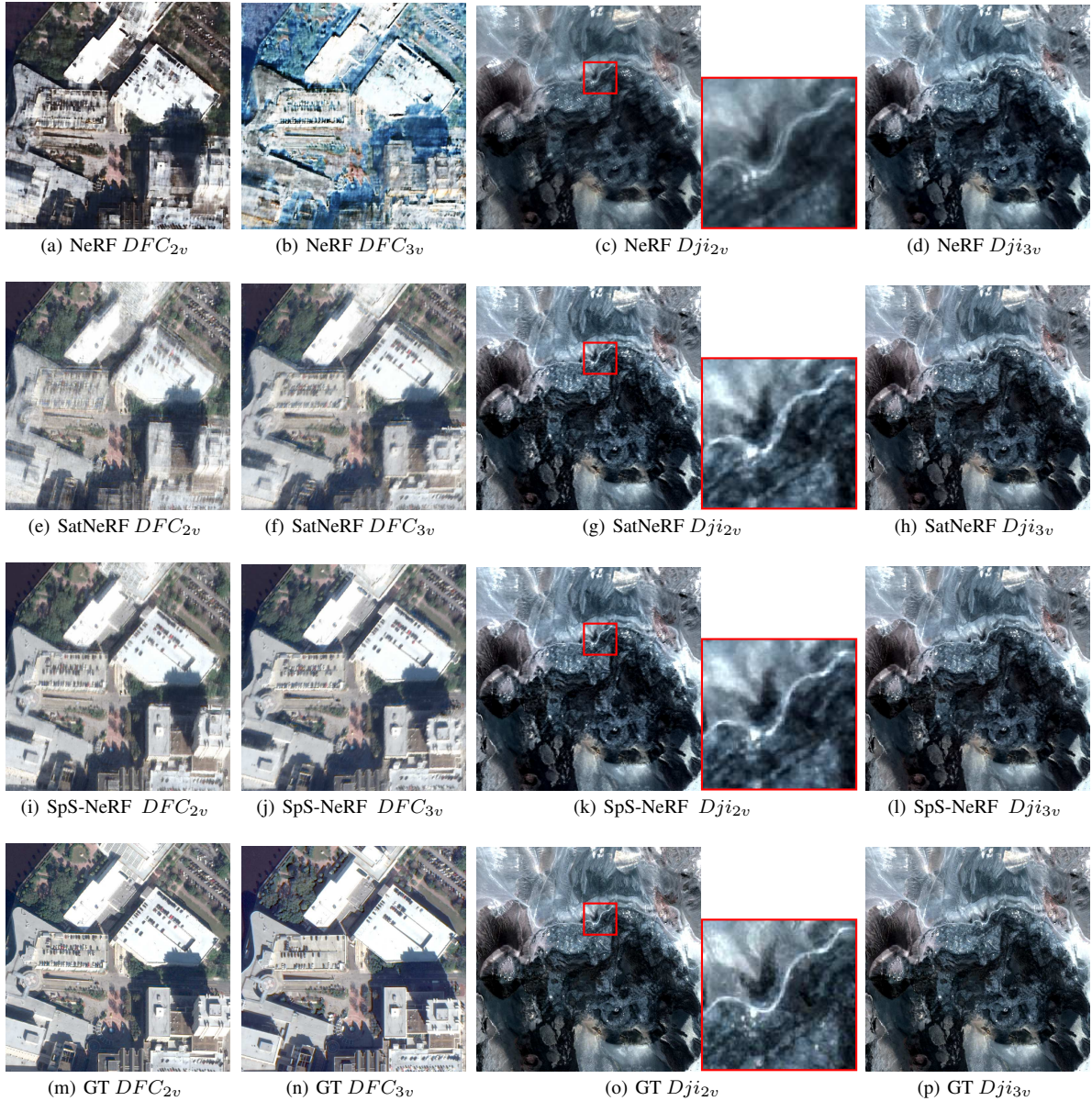


Figure 5. **Novel view synthesis.** Qualitative evaluation is performed on DFC2019 (DFC) and Djibouti (Dji) datasets using 2-views (<sub>2v</sub>) and 3-views (<sub>3v</sub>) for training. NeRF renders blurry images, Sat-NeRF reduces the blur thanks to sparse depth supervision, SpS-NeRF renders sharpest images of all.

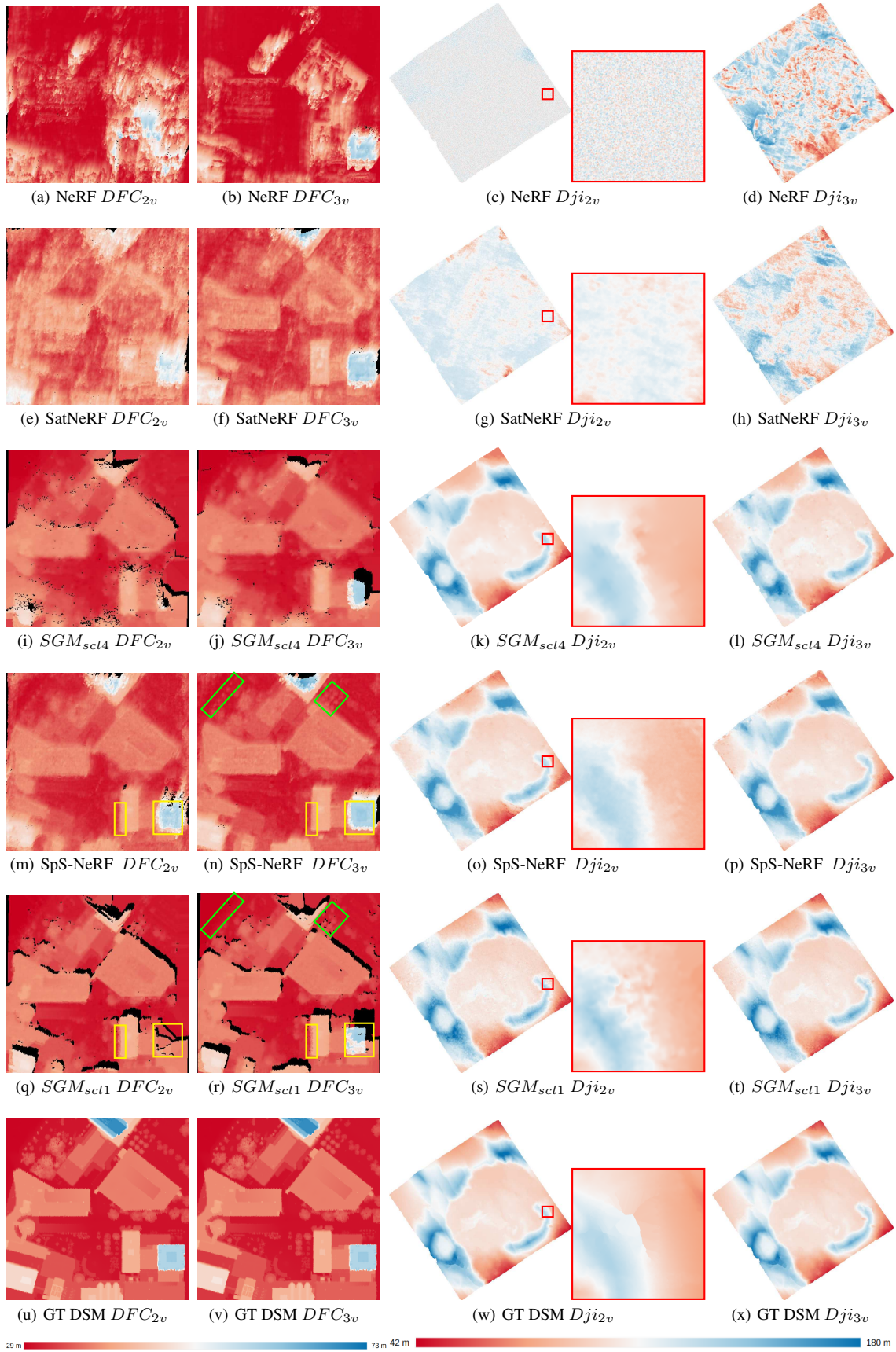


Figure 6. **Altitude extraction.** SpS-NeRF outruns all tested NeRF variants, and reconstructs 3D geometry comparably to  $SGM_{scl1}$ . In urban DFC2019 dataset, SpS-NeRF is better at reconstructing vegetation ( $\square$ ) and at handling building outlines near occlusions ( $\square$ ) but the surface is generally less smooth than that of  $SGM_{scl1}$ . In rural Djibouti dataset, notice the more detailed and coherent reconstruction of SpS-NeRF in (o) compared to  $SGM_{scl1}$  result in (s).

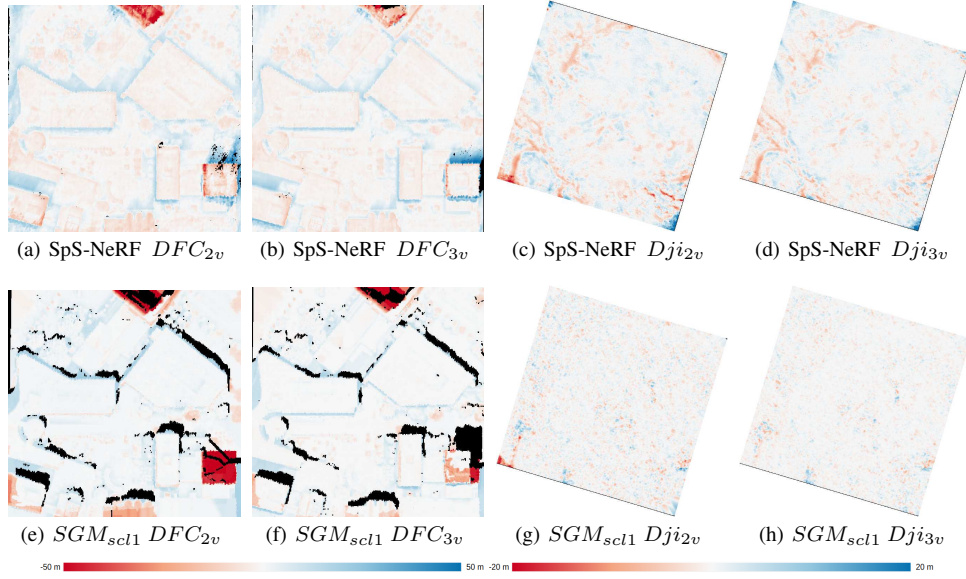


Figure 7. **Difference of DSMs.** We compute the differences w.r.t. GT DSMs for the two best performing methods. Although SpS-NeRF behaves better near discontinuities in urban DFC dataset, it is unable to recover high frequency details in rural Djibouti. Notice that the difference maps for SGM (g,h) carry a repetitive signal typical for aliasing due to image resampling. Such artefacts are not present in SpS-NeRF.

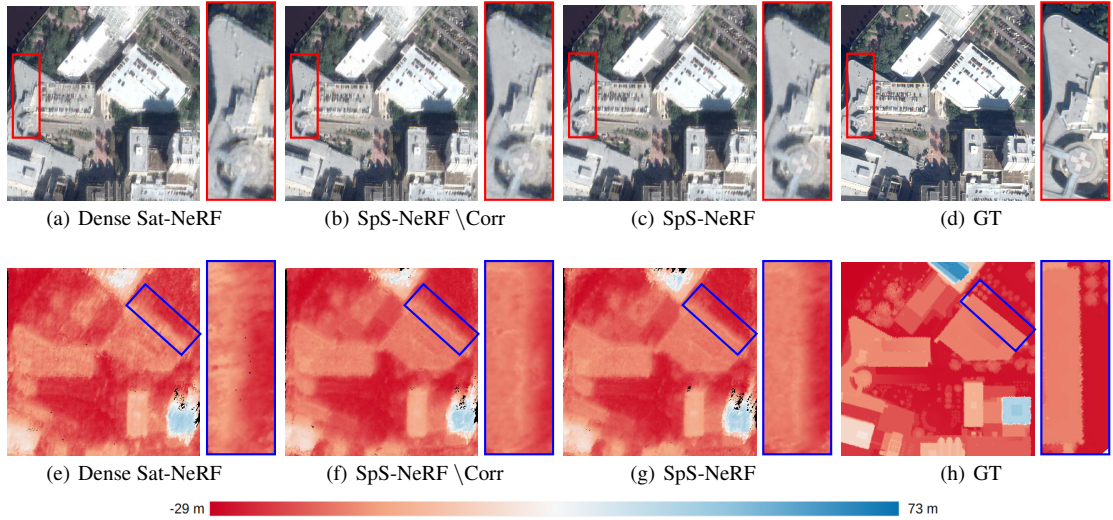


Figure 8. **Ablation experiment.** Qualitative result on NeRF variants trained with 2 views (DFC2019). The top row (a-d) represents the novel views, while the bottom row (e-h) shows DSMs. Adding dense supervision (a,e), guided ray sampling (b,f) and uncertainty measures (c,g) contribute to visually better surface geometries and sharper novel views.

Method	PSNR $\uparrow$	SSIM $\uparrow$	MAE <sub>in</sub> $\downarrow$	MAE <sub>out</sub> $\downarrow$
Dense Sat-NeRF	19.39	0.86	3.58	7.91
SpS-NeRF \ Corr	19.67	0.86	3.21	8.03
SpS-NeRF	<b>20.2</b>	<b>0.87</b>	<b>3.02</b>	<b>7.77</b>

Table 3. **Ablation experiment.** Quantitative metrics on NeRF variants trained with 2 views from DFC2019. Adding dense supervision (Dense Sat-NeRF), guided ray sampling (SpS-NeRF \ Corr) and uncertainty measures (SpS-NeRF) improve the novel view generation and surface recovery metrics.

as SGM.

It should be mentioned that the GT DSM in the Djibouti dataset Figure 6(w, x) was generated with the very same SGM as the best performing  $SGM_{scl1}$ . This correlation might potentially bias the comparison. Additionally, SGM is susceptible to outliers, as shown in the zoom-in view of GT DSM in Figure 6(w). Hence, our GT DSM is likely corrupt with some erroneous depth estimations.

**Ablation study.** We perform two experiments training different variants of NeRF with 2 views from the DFC2019 dataset: (i) *Dense Sat-NeRF* where we train the vanilla Sat-NeRF and replace the sparse depth supervision with our dense depths; (ii) *SpS-NeRF \ Corr* where we train our SpS-NeRF and set the  $corr(\mathbf{r})=1$  for every pixel in Equation (5) and Equation (6) thus we deactivate the uncertainty metric but maintain the ray sampling strategy.

In Figure 8 we compare the novel view and depths generated by *Dense Sat-NeRF*, *SpS-NeRF \ Corr* with our full SpS-NeRF. Without the guided ray sampling, *Dense Sat-NeRF* struggles to recover a high contrast image (a) and sharp buildings' outlines (e). The performance improves in *SpS-NeRF \ Corr* (b and f), where the network is encouraged to estimate the depth within the  $m$  margin (Equation (5)) of the input depth while balancing the color loss. The performance is further enhanced by adding  $corr(\mathbf{r})$  (Figure 8(c, g)). Quantitative results in Table 3 show the same tendencies.

## 5. CONCLUSION

We presented SparseSat-NeRF (SpS-NeRF) – an extension of Sat-NeRF adapted to novel view generation and 3D geometry reconstruction from sparse satellite image views. The adaptation consists of including dense depth supervision with low resolution surfaces obtained with traditional dense image matching, and a suitable ray sampling borrowed from (Roessle et al., 2022). To add robustness to our supervision we incorporate uncertainty metrics based on dense image matching cross-correlation maps. We demonstrate that SpS-NeRF performs better than NeRF and Sat-NeRF in sparse view scenarios. It is also competitive with the traditional semi-global matching.

## 6. ACKNOWLEDGEMENT

This research was funded by CNES (Centre national d'études spatiales). The Djibouti dataset was obtained through the CNES ISIS framework. The numerical computations were performed on the SCAPAD cluster processing facility at the Institute de Physique du Globe de Paris. We thank Stéphane Jacquemoud and Tri Dung Nguyen for familiarizing us with the Djibouti dataset.

## REFERENCES

Bittner, K., Reinartz, P., Korner, M., 2019. Late or earlier information fusion from depth and spectral data? large-scale digital surface model refinement by hybrid-cgan. *CVPRW*.

Bleyer, M., Rhemann, C., Rother, C., 2011. Patchmatch stereo-stereo matching with slanted support windows. *Bmvc*, 11, 1–11.

Buades, A., Facciolo, G., 2015. Reliable multiscale and multiwindow stereo matching. *SIAM Journal on Imaging Sciences*, 8(2), 888–915.

Bulatov, D., Wernerus, P., Heipke, C., 2011. Multi-view dense matching supported by triangular meshes. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(6), 907–918.

Chang, J., Chen, Y., 2018. Pyramid stereo matching network. *CVPR*, 5410–5418.

Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H., 2021. MVSNerF: Fast generalizable radiance field reconstruction from multi-view stereo. *ICCV*, 14124–14133.

Deng, K., Liu, A., Zhu, J.-Y., Ramanan, D., 2022. Depth-supervised nerf: Fewer views and faster training for free. *CVPR*, 12882–12891.

Derksen, D., Izzo, D., 2021. Shadow neural radiance fields for multi-view satellite photogrammetry. *CVPRW*, 1152–1161.

Furukawa, Y., Ponce, J., 2009. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8), 1362–1376.

Gallup, D., Frahm, J.-M., Mordohai, P., Yang, Q., Pollefeys, M., 2007. Real-time plane-sweeping stereo with multiple sweeping directions. *CVPR*, IEEE, 1–8.

Gao, J., Liu, J., Ji, S., 2021. Rational polynomial camera model warping for deep learning based satellite multi-view stereo matching. *ICCV*, 6148–6157.

Gómez, A., Randall, G., Facciolo, G., von Gioi, R. G., 2022. An experimental comparison of multi-view stereo approaches on satellite images. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 844–853.

Hartmann, W., Galliani, S., Havlena, M., Schindler, K., Gool, L. V., 2017. Learned multi-patch similarity. *ICCV*, 1586–1594.

Hirschmuller, H., 2005. Accurate and efficient stereo processing by semi-global matching and mutual information. *CVPR*, 2, IEEE, 807–814.

Huang, P.-H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.-B., 2018. Deepmvs: Learning multi-view stereopsis. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2821–2830.

Jain, A., Tancik, M., Abbeel, P., 2021. Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis. *CVPR*, 5885–5894.

Labarre, S., Jacquemoud, S., Ferrari, C., Delorme, A., Derrien, A., Grandin, R., Jalludin, M., Lemaître, F., Métois, M., Pierrot-Deseilligny, M., 2019. Retrieving soil surface roughness with the Hapke photometric model: Confrontation with the ground truth. *Remote Sensing of Environment*, 225, 1–15.

Le Saux, B., Yokoya, N., Hansch, R., Brown, M., Hager, G., 2019. 2019 data fusion contest [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 7(1), 103–105.

Marí, R., Facciolo, G., Ehret, T., 2022. Sat-NeRF: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using RPC cameras. *CVPRW*, 1311–1321.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2020. NeRF: Representing scenes as neural radiance fields for view synthesis. *ECCV*, 99–106.

Pierrot-Deseilligny, M., Paparoditis, N., 2006. A multiresolution and optimization-based image matching approach: An application to surface reconstruction from SPOT5-HRS stereo imagery. *Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 36(1/W41), 1–5.

Roessle, B., Barron, J. T., Mildenhall, B., Srinivasan, P. P., Nießner, M., 2022. Dense depth priors for neural radiance fields from sparse input views. *CVPR*, 12892–12901.

- Rupnik, E., Pierrot-Deseilligny, M., Delorme, A., 2018. 3D reconstruction from multi-view VHR-satellite images in MicMac. *ISPRS Journal of Photogrammetry and Remote Sensing*, 139, 201–211.
- Stucker, C., Schindler, K., 2020. Resdepth: Learned residual stereo reconstruction. *CVPRW*, 184–185.
- Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600–612.
- Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J., Zhou, J., 2021. Nerfing-MVS: Guided optimization of neural radiance fields for indoor multi-view stereo. *CVPR*, 5610–5619.
- Xu, D., Jiang, Y., Wang, P., Fan, Z., Shi, H., Wang, Z., 2022. SinNeRF: Training neural radiance fields on complex scenes from a single image. *ECCV*, Springer, 736–753.
- Yu, A., Ye, V., Tancik, M., Kanazawa, A., 2021. PixelNeRF: Neural radiance fields from one or few images. *CVPR*, 4578–4587.