

# Semantic Neural Radiance Fields for Multi-Date Satellite Data

Valentin Wagner, Sebastian Bullinger, Christoph Bodensteiner, and Michael Arens  
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation

{firstname.lastname}@iosb.fraunhofer.de

## Abstract

In this work we propose a satellite specific Neural Radiance Fields (NeRF) model capable to obtain a three-dimensional semantic representation (neural semantic field) of the scene. The model derives the output from a set of multi-date satellite images with corresponding pixel-wise semantic labels. We demonstrate the robustness of our approach and its capability to improve noisy input labels. We enhance the color prediction by utilizing the semantic information to address temporal image inconsistencies caused by non-stationary categories such as vehicles. To facilitate further research in this domain, we present a dataset comprising manually generated labels for popular multi-view satellite images. Our code and dataset are available at <https://github.com/wagnva/semantic-nerf-for-satellite-data>.

## 1. Introduction

As the number of satellites in orbit continues to rise, high-resolution satellite imagery is more accessible than ever. Image-based reconstruction for creating large-scale environmental models has gained popularity due to its cost-effectiveness compared to dedicated systems like LiDAR. There is also an increasing emphasis on assigning semantic information to 3D environments, which enhances the analysis of urban areas and natural resources, leading to better decision-making in fields such as urban planning, environmental monitoring, and disaster management.

Traditional methods have focused on extracting explicit representations like pointclouds or meshes from the satellite images by matching image features [2, 4, 5, 17]. Subsequent works [13, 15, 16] assign the semantic information to the extracted representation as additional step. *Neural Radiance Fields (NeRF)* [12] take a different approach to 3D-reconstruction by leveraging *Multi-Layer-Perceptrons (MLP)*. The scene structure is represented with the weights of the network. Aggregation of *MLP* outputs along visual rays allows rendering of novel, unseen views. *NeRF* are capable to learn and reproduce additional modalities be-

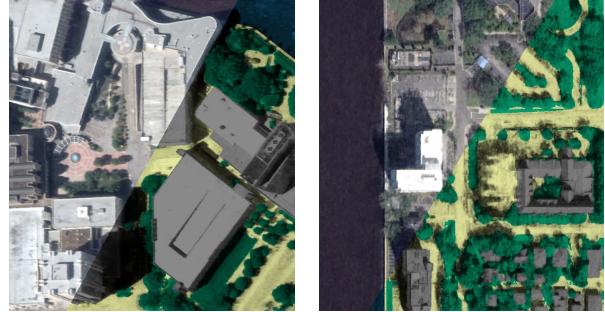


Figure 1. Example output of our proposed satellite-domain-adapted semantic *NeRF* model. Fusion of RGB and semantics allows to render the scene in two distinct modalities for novel, during training unseen views. The semantic visualization is enriched with a three-dimensional structure through combination with a learned lighting component.

side color using a shared three-dimensional representation [18, 19]. Handling multi-date satellite image data introduces additional challenges such as domain-specific camera models, variable lighting, moving shadows and transient objects such as vehicles. Existing works [1, 6, 10, 11] present *NeRF* adaptions to tackle these challenges. As of our best knowledge this is the first work to present a dual modality *NeRF* to the satellite domain.

### 1.1. Contributions

Our contributions are as following: (A) A satellite-domain-adapted semantic *NeRF* capable to fuse semantic and color information onto a unified 3D representation. (B) Improved color consistency for transient areas such as vehicles in inconsistent images by exploiting semantic information. (C) Demonstration of the robustness/feasibility of our approach and its capability to improve quality of semantic training data through enforcement of multi-view consistency for the satellite domain. (D) Publicly released dataset with manually generated pixel-wise annotations for 71 satellite images across four multi-view scenes covering five semantic classes: *Ground*, *Water*, *Vegetation*, *Buildings*, *Vehicles*. (E) Publicly available code base to reproduce our results and allow further research.

## 2. Related Works

*Semantic-NeRF* [19] shows that *NeRF* architectures are capable to simultaneously learn a combined appearance and semantic 3D model for a scene given a set of input color images with corresponding semantic labels. This allows to render novel, unseen views as color image with a corresponding pixel-wise segmentation image. The two modalities share the same scene structure, leading to a high cohesion in rendering quality. Additionally, they show that the inherent multi-view consistency provided by *NeRF* is able to smooth noisy semantic training labels.

*Shadow-NeRF* [6] adapts *NeRF* to the satellite domain by introducing an altitude-based ray sampling and an irradiance lighting model. They separate the static scene color from variable shadows and ambient lighting using the sun direction as input. This allows the network to learn highly variant lighting in its own component. *SatNeRF* [11] introduces transient-object filtering and adapts the domain-specific *Rational Polynomial Coefficient (RPC)* [14] camera model for ray generation, increasing accuracy. As satellite datasets are usually multi-date, transient objects are moving inbetween image captures. To reduce artifacts *SatNeRF* [11] introduces an uncertainty term, learned based on an image-specific embedding. *EO-NeRF* [10] further iterates on handling the varying lighting conditions. Instead of learning shadows they are dynamically rendered based on the scene geometry and sun direction. *SUNDIAL* [1] models complex illumination to improve dynamic shadow rendering of vegetation and water areas. Whereas the results for *EO-NeRF* [10] and *SUNDIAL* [1] are promising, no code is publicly available. We therefore base our work on *Shadow-NeRF* [6] and *SatNeRF* [11].

## 3. Semantic Satellite NeRF

An overview of our proposed satellite-domain-adapted semantic *NeRF* can be seen in Fig. 2. We utilize satellite images with their corresponding semantic pixel-wise annotations to train a dual-modality *NeRF* specifically adapted for satellite-imagery. Thus, in contrast to previous approaches, our pipeline renders pixel-wise semantic labels for novel, unseen views of scenes captured by satellite data.

### 3.1. Satellite-Domain-Specific Camera Model

During the training and testing phase of a *NeRF*, rays are constructed for each pixel of the given images and their corresponding camera poses, traversing a pre-defined, bounded three-dimensional volume. As proposed by *SatNeRF* [11] we use the *RPC* [14] camera model instead of the commonly used pinhole [7] camera model. This domain-specific camera model is an approximation of the physical satellite camera. We extract it from the satellite image metadata as it is typically used for georegistration of the image.

To improve accuracy, *SatNeRF* [11] performs *Bundle Adjustment* on the *RPC* models. This process optimizes relative camera poses to increase ray accuracy.

### 3.2. Learning Semantic in 3D for Satellite Data

*NeRF* [12] represent a static three-dimensional scene as a continuous volumetric function  $\mathcal{F}_{NeRF}$  encoded with a *MLP* network.

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

Given a 3D point  $\mathbf{x} = (x, y, z)$  and a viewing direction  $\mathbf{d} = (d_x, d_y, d_z)$ , the network predicts a density scalar  $\sigma$  and a color  $\mathbf{c} = (r, g, b)$ . The scene structure is represented by the density  $\sigma$  and used to weigh the influence of the color  $\mathbf{c}$ . *SatNeRF* [11] proposes an expanded satellite-domain-specific architecture.

$$\mathcal{F}_{SatNeRF} : (\mathbf{x}, \boldsymbol{\omega}, \mathbf{t}_j) \mapsto (\sigma, \mathbf{c}_a, sun, \mathbf{a}, \beta) \quad (2)$$

The scene color  $\mathbf{c}$  is split up into three separate lighting components: the albedo color  $\mathbf{c}_a$ , the sun shading scalar  $sun$  and the ambient color  $\mathbf{a}$ . To predict the variable lighting components  $sun$  and  $\mathbf{a}$ , the sun direction  $\boldsymbol{\omega}$  is used as auxiliary input. The uncertainty  $\beta$  is used to reduce artifacts caused by transient objects such as vehicles. It is learned with the per-image embedding  $\mathbf{t}_j$  as auxilliary input. Our proposed semantic satellite-adapted network represents the following function  $\mathcal{F}_{SemSatNeRF}$ .

$$\mathcal{F}_{SemSatNeRF} : (\mathbf{x}, \boldsymbol{\omega}, \mathbf{t}_j) \mapsto (\sigma, \mathbf{c}_a, sun, \mathbf{a}, \beta, s) \quad (3)$$

We extend the satellite-domain-adapted architecture from *SatNeRF* [11] with the additional semantic output  $s$ . Similar to *Semantic-NeRF* [19] the semantic output  $s$  is represented as a view-invariant function  $\mathcal{F}_s(\mathbf{x}) \mapsto s$  of the world coordinates  $\mathbf{x}$  and describes a distribution over  $C$  semantic labels as pre-softmax logits. In addition, we introduce a *sigmoid* activation function with the intention to increase the generalization across views, using the *sigmoid* activation function as form of normalization. This reduces the ability of single samples to dominate the semantic prediction.

$$s(\mathbf{x}) = sigmoid(\mathcal{F}_s(\mathbf{x})) \quad (4)$$

To produce the final result for a specific pixel position, *NeRF*-based pipelines aggregate the outputs of the *MLP* for individual points  $\mathbf{x}$  along a ray  $\mathbf{r}(t) = \mathbf{o} + t \cdot \mathbf{d}$  representing the corresponding line of sight. Here,  $\mathbf{o}$  and  $\mathbf{d}$  denote the origin and direction vectors.

$$s(\mathbf{r}) = \sum_{i=1}^N T_i \alpha_i s(\mathbf{x}_i) \quad (5)$$

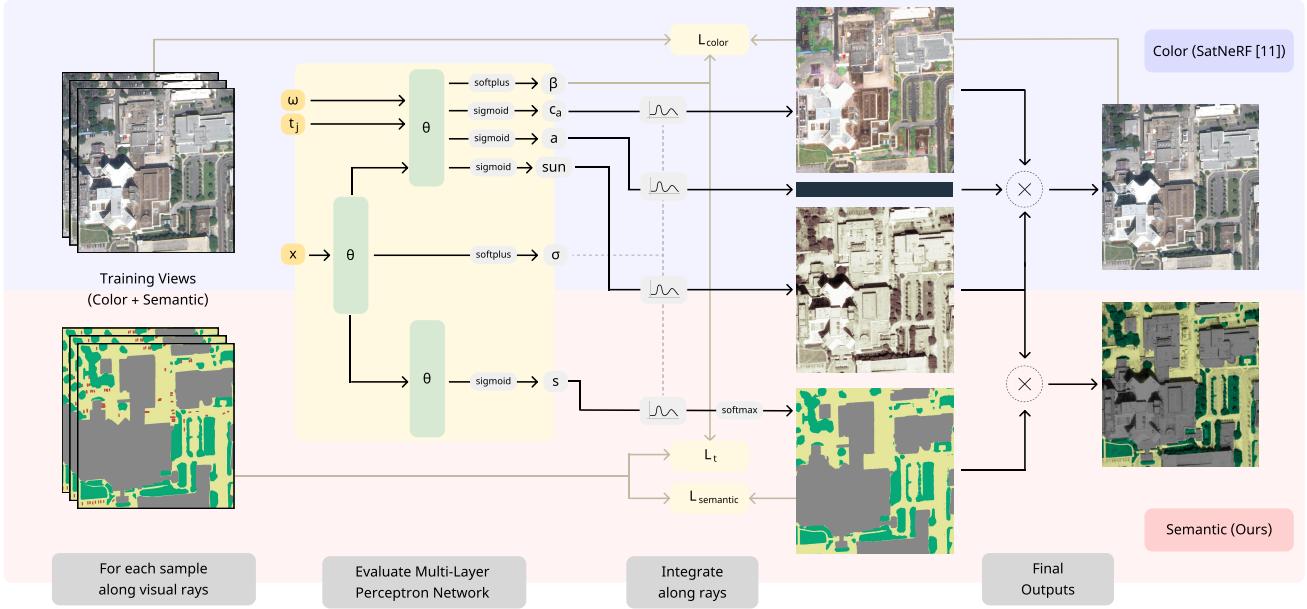


Figure 2. Overview of our proposed model. The satellite-domain-adapted outputs (*i.e.* elements in the blue area) are combined using an irradiance lighting model to produce the color rendering as originally proposed by *SatNeRF* [11]. Using an additional semantic head (*i.e.* elements in the red area) our proposed method is able to produce a corresponding semantic pixel-wise labeling. We combine this with the learned lighting scalar to create a three-dimensional semantic visualization. We introduce a transient regularization loss  $L_t$  to reduce artifacts in the learned appearance based on the semantic input data.

The influence of each sample  $\mathbf{x}_i$  along a ray is based on its opacity  $\alpha_i$  and transmittance  $T_i$ .

$$\alpha_i = 1 - \exp(-\sigma_i \delta_i) \text{ and } T_i = \prod_{j=1}^{i-1} (1 - \alpha_j) \quad (6)$$

The distance  $\delta_i = t_i - t_{i-1}$  describes the length traveled along the ray between the previous and current sampled position  $\mathbf{x}$ . Using softmax normalization the combined logits  $s(\mathbf{r})$  are converted into probabilities  $\hat{p}(\mathbf{r})$ . The class with the highest probability is chosen to represent the semantic class of the ray.

### 3.3. Semantic Structural Visualization

To visualize the semantic labels for a given image, usually a simple color mapping operation  $f_c : s \mapsto RGB$  is used. This visualization is sub-optimal for large areas with the same class, as the structure of the scene is lost. For large structures consisting of multiple buildings with large height differences this is especially noticeable, such as in the scenes JAX\_068 and JAX\_214 in Fig. 3b.

To handle variance in lighting caused by different sun positions in training images, we use a satellite-domain-adapted irradiance lighting model as proposed by *Shadow-NeRF* [6]. The sample color  $c(\mathbf{x}, \omega)$  is split up into three components: scene color  $c_a(\mathbf{x})$ , lighting scalar  $sun(\mathbf{x}, \omega)$

and ambient sky color  $a(\omega)$ .

$$c(\mathbf{x}, \omega) = c_a(\mathbf{x}) \cdot (sun(\mathbf{x}, \omega) + (1 - sun(\mathbf{x}, \omega)) \cdot a(\omega)) \quad (7)$$

By referencing the visualization of the  $sun(\mathbf{x}, \omega)$  lighting component in Fig. 3d we can see that it resembles a grayscale relief map of the scene structure. We use this to provide a three-dimensional depth to the semantic output. We aggregate the  $sun(\mathbf{x}, \omega)$  component along rays  $\mathbf{r}$  analog to Eq. (5) and combine it with the mapped semantic color  $f_c(s(\mathbf{r}))$ .

$$c_s(\mathbf{r}) = f_c(s(\mathbf{r})) * \sum_{i=1}^N T_i \alpha_i sun(\mathbf{x}_i, \omega) \quad (8)$$

The resulting visualization of the structural semantic prediction is shown in Fig. 3e. The commonly flat color visualization is enriched with a three-dimensional structure. Finer details are made visible and objects such as the distinct main building can now be easily identified.

### 3.4. Handling Transient Objects

A major issue by processing multi-date satellite images are temporal inconsistencies. Besides changes in the scene geometry, transient objects pose a substantial challenge for the reconstruction process. As transient objects change positions across training images, the network receives contra-

dictory information. Since we are interested in reconstructing a consistent model reflecting the stationary scene geometry, transient objects represent a type of noise. To improve the model consistency, we use available transient category information contained in the semantic labels to reduce the impact of such objects on the semantic as well as the appearance prediction. We separately consider the problem for both modalities as semantic and appearance are learned with their own distinct loss function.

### 3.4.1 Handling Transient Objects in RGB Data

*NeRF* are optimized by minimizing the  $L_2$ -Loss between the predicted ray color  $\mathbf{c}(\mathbf{r})$  and the ground truth pixel color  $\mathbf{c}_{GT}(\mathbf{r})$ .

$$L_2(\mathcal{R}) = \sum_{\mathbf{r} \in \mathcal{R}} \|\mathbf{c}(\mathbf{r}) - \mathbf{c}_{GT}(\mathbf{r})\|_2^2 \quad (9)$$

To reduce the effect of transient objects in the color prediction *SatNeRF* [11] introduces an uncertainty mechanism. Based on a per-image learned embedding  $t_j$  the network predicts an uncertainty factor  $\beta$  for each location  $\mathbf{x}$ . The color-loss is decreased depending on the uncertainty, giving the network the ability to selectively reduce the loss of specific rays.

$$L_{color}(\mathcal{R}) = \sum_{\mathbf{r} \in \mathcal{R}} \frac{\|\mathbf{c}(\mathbf{r}) - \mathbf{c}_{GT}(\mathbf{r})\|_2^2}{2\beta'(\mathbf{r})^2} + \left( \frac{\log \beta'(\mathbf{r}) + \eta}{2} \right) \quad (10)$$

The second term forces the network to use the uncertainty sparingly preventing the naive solution of  $\beta$  converging towards infinity. To prevent negative values in the logarithm we set  $\beta'(\mathbf{r}) = \beta(\mathbf{r}) + 0.05$  and  $\eta = 3$  analog to *SatNeRF* [11].

The uncertainty  $\beta$  gives the network an mechanism to reduce the impact of local outliers in the color training images. Although many transient objects such as vehicles move inbetween images, locations such as parking lots contain vehicles in majority of the images. To achieve our goal of learning a static, transient free representation of the scene we introduce a transient regularization loss. Instead of a purely statistical approach we propose guiding the uncertainty towards transient objects using existing information from the semantical ground truth data. To this end, we introduce a transient regularization loss as seen in Eq. (11). For rays  $\mathbf{r} \in \mathcal{R}^t$  that are associated with transient ground truth labels, the following loss guides the uncertainty  $\beta$  towards 1.

$$L_t(\mathcal{R}^t) = \sum_{\mathbf{r} \in \mathcal{R}^t} \|1.0 - \beta(\mathbf{r})\|_2^2 \quad (11)$$

We do not make any assumptions about rays outside of the transient classes, allowing the network to still learn its own

uncertainty values. This ensures that the network still handles other occurring phenomena such as chromatic aberration. Whereas the uncertainty  $\beta$  is theoretically unbounded, the second term of the color-loss in Eq. (10) punishes high values. The uncertainty scalar  $\beta$  naturally resides in the interval from 0 to around 1. We therefore guide the network towards 1 for rays  $\mathbf{r} \in \mathcal{R}^t$ .

*SatNeRF* [11] proposes using the non-modified color-loss as seen in Eq. (9) for the first two epochs. This gives the network time to learn an initial understanding of the scene geometry and color. We enable our proposed transient regularization loss after three epochs. This gives the network one additional epoch to initialize the uncertainty  $\beta$ .

### 3.4.2 Handling Transient Objects in Semantic Data

Common *NeRF* methods only derive a 3d appearance model of the scene. By introducing a semantic representation, the network now has a secondary goal, requiring an additional loss term. *Semantic-NeRF* [19] uses a cross-entropy-loss  $\sum_{c=1}^C p^c(\mathbf{r}) \log \hat{p}^c(\mathbf{r})$  comparing the ground truth label  $p^c$  with the predicted semantic probability  $\hat{p}^c$  for each class  $c$  across all rays  $\mathbf{r} \in \mathcal{R}$ . We define the semantic loss to only include rays from static classes  $\mathbf{r} \in \mathcal{R} \setminus \mathcal{R}^t$ .  $\mathcal{R}^t$  hereby describes all rays  $\mathbf{r}$  with a semantic class defined as transient.

$$L_{semantic}(\mathcal{R} \setminus \mathcal{R}^t) = - \sum_{\mathbf{r} \in (\mathcal{R} \setminus \mathcal{R}^t)} \sum_{c=1}^C p^c(\mathbf{r}) \log \hat{p}^c(\mathbf{r}) \quad (12)$$

This effectively equals in ignoring parts of the semantic training labels, leaving holes in the pixel-wise label map. The network does not receive any learning guidance for these positions and has to rely on information given by different training views to fill in the missing information. This is made possible by the inherent multi-view consistency mechanism provided by *NeRFs*.

## 4. Semantic Multi-View Satellite Dataset

The publicly available *Urban Semantic 3D* [3] dataset provides automatically generated labels for the *Data-Fusion-Contest 2019 Track-3 (DFC2019)* [9] dataset. The generation process fuses information from LiDAR source data with limited supervision, leading to a high degree of noise and poor accuracy. Transient objects such as *Vehicles* are not captured in the provided labels. We therefore chose to produce our own annotations and ensure high quality annotations through a high degree of manual refinement. We make it publicly available to further advance research in the field of semantic 3D reconstruction.

Our semantic multi-view satellite dataset contains manually generated annotations for 71 satellite images across four scenes. Analog to previous work [6, 10, 11] we select a popular subset containing four scenes (004, 068, 214, 260)



Figure 3. Qualitative results of our proposed *NeRF* model for the popular scenes of the JAX dataset. The pixel-wise semantic annotations in (b) are part of our dataset. The domain-adapted *NeRF* model is able to learn and reproduce a representation of the scene containing color and semantic information. By combining the predicted semantic class with the learned lighting component we enrich the visualization with three-dimensional depth cues.

of the Jacksonville (Florida, USA) area. The provided Maxar WorldView-3 satellite images are captured between 2014 and 2016 with a 0.3m/pixel resolution at nadir.

The annotations are generated as pixel-wise segmentation maps and cover the main  $256 \times 256$ m area-of-interest of each scene. This area is defined by the LiDAR *Digital Surface Model (DSM)* contained in the *DFC2019* [9] dataset. To create the annotations we manually refine initial rough estimates of a user guided, class agnostic foundation model for semantic segmentation [8]. We manually verify and potentially correct all annotations to ensure high accuracy and consistency across all images. The five following classes

were selected based on scene content: *Ground, Water, Vegetation, Buildings* and *Vehicles*. We show one satellite image per scene in Fig. 3a and its accompanying pixel-wise segmentation in Fig. 3b.

## 5. Implementation and Training Details

Analog to *SatNeRF* [11] we use 8 layers of 512 features for the main feature backbone of the MLP. The additional semantic head consists of a single hidden layer with 256 features and uses a *sigmoid* activation function for semantic logit prediction. The architecture of the other heads such as

Split	Semantic Accuracy ↑				
	004	068	214	260	Mean
Train	0.936	0.950	0.947	0.934	0.942
Test	0.930	0.948	0.944	0.916	0.935

Table 1. Quantitative evaluation of our domain-adapted semantic *NeRF*. The high semantic accuracy for test views shows the ability of the network to render a correct semantic segmentation of the scene for novel, during training unseen views.

the lighting scalar *sun* or the uncertainty  $\beta$  are identical to *SatNeRF* [11].

The semantic loss weight is set to  $\lambda_s = 0.04$  as suggested by *Semantic-NeRF* [19]. A transient regularization loss value of  $\lambda_t = 0.1$  shows in our experiments a good compromise between guiding the network to filter transient objects while simultaneously applying the uncertainty  $\beta$  for other color inconsistencies. The transient embedding  $t_j$  has a dimension of 4. As proposed by *SatNeRF* [11] we use additional depth supervision rays for the first 25% of training iterations with a weight of  $\lambda_{DS} = 1000$  and solar correction rays with a weight of  $\lambda_{SC} = 0.05$ . Depth supervision refers to using a sparse pointcloud created during the initial *Bundle Adjustment* on the *RPC* camera models for depth guidance. Solar correction rays are used as part of the satellite-domain-adapted lighting model to align shadows with the scene structure based on the density.

We trained all models in this work for  $300k$  iterations with an Adam optimizer starting with a learning rate of  $5e^{-4}$ , which is decreased by a factor of 0.9 every epoch. Each ray is discretized into 64 points and we process 1024 rays each training iteration. All models were trained using a NVIDIA RTX 4090 with 24 GB of RAM.

## 6. Evaluation and Experiments

We evaluate the ability of our proposed *NeRF* pipeline to learn a generalized semantic understanding of satellite scenes. Since the *Vehicle* category consists of transient objects, we treat this class according to Sec. 3.4 when computing the semantic and color representation

In Sec. 6.1 we evaluate the general quality of the semantic rendering in comparison to ground truth data. In Sec. 6.2 we show the impact of our transient regularization term for color representation and how it reduces artifacts. Lastly, we evaluate the capability of our pipeline to improve the semantic segmentation quality of input images by leveraging the inherent multi-view-consistency provided by *NeRF* in Sec. 6.3.

### 6.1. Learned Semantic Representation

To evaluate the quality of the learned semantic representation, we use the semantic accuracy metric, *i.e.* the ratio

Transient Regularization	Transient Uncertainty ↑				
	004	068	214	260	Mean
✗	0.121	0.092	0.144	0.027	0.096
✓	0.976	0.900	0.877	0.915	0.917

Table 2. Quantitative evaluation of the proposed transient regularization loss. The results shows a great increase in uncertainty for locations marked as transient in the semantic ground truth data. The model learns a static, transient free appearance of the scene.

of correctly to incorrectly identified pixels. We create alternative transient-free ground truth annotations to assess whether our proposed network is able to accurately replace the missing semantic information for transient locations with the correct static class (in most cases either *Ground* or *Building*). The evaluation results are reported in Tab. 1. Our domain-adapted semantic *NeRF* achieves an accuracy of over 90% for all scenes, proving its ability to generalize the learned semantic understanding of the scene for both training and testing viewpoints.

We visualize one view of each scene in Fig. 3 with the corresponding ground truth color image (Fig. 3a), the semantic annotations (Fig. 3b), and the outputs learned by our domain-adapted semantic *NeRF*. Comparing the learned semantic prediction in Fig. 3c with the ground truth annotations shows a high degree of accuracy. By combining the predicted semantic class with the learned shadow scalar in Fig. 3d we produce a three-dimensional visualization in Fig. 3e. Comparing this visualization to the ground truth color image shows how the network is able to transfer the scene structure on the semantic understanding.

### 6.2. Transient Regularization

We evaluate the impact of our proposed transient regularization loss in increasing uncertainty of known transient locations by measuring the *Transient Uncertainty*. We define this metric as the mean uncertainty for all transient locations  $\mathcal{R}^t$ , *i.e.* the *Vehicles* class, in the semantic ground truth data. The uncertainty  $\beta(\mathbf{x}, t_j)$  is aggregated along rays  $r$  analog to Eq. (5) with the per-image embedding  $t_j$  as additional input.

$$\beta_t(\mathcal{R}^t) = \frac{1}{|\mathcal{R}^t|} \sum_{\mathbf{r} \in \mathcal{R}^t} \sum_{i=0}^N \alpha_i T_i \beta(\mathbf{x}_i, t_j) \quad (13)$$

Since the uncertainty  $\beta$  is only relevant during training, we do not consider test images during this evaluation. In Tab. 2 we can see the impact of our transient regularization loss. The baseline *SatNeRF* [11] model is not able to adequately capture the transient objects, resulting in a low *Transient Uncertainty* of 0.096 across all four scenes. This causes the *SatNeRF* model to integrate the corresponding appearance

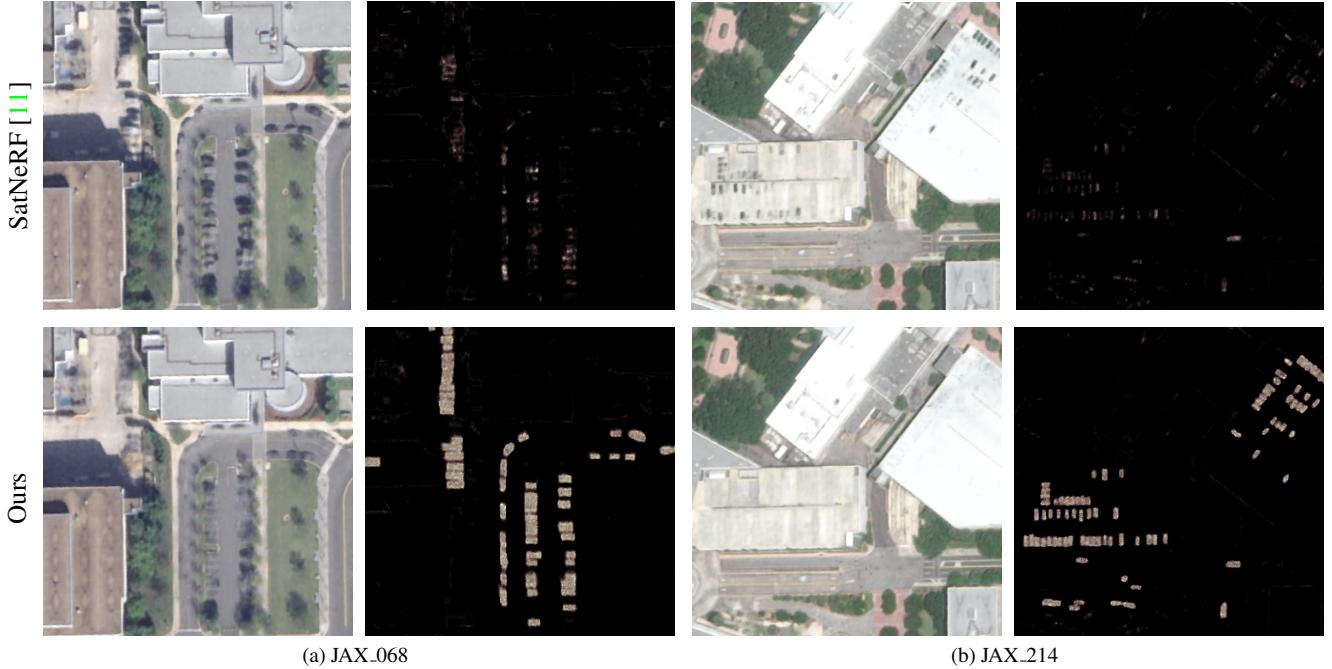


Figure 4. Qualitative comparison of the proposed transient regularization loss. Using this loss, our model is able to substantially reduce blurry artifacts by guiding the uncertainty  $\beta$  on locations of transient objects. Shown here are closeups of the rendered color image and visualization of the uncertainty scalar  $\beta$  for training views of the scenes JAX\_068 and JAX\_214. The Baseline *SatNeRF* [11] model is unable to filter the cars out, leading to visible remnants. Our method is able to render a static, transient free representation of the scene.

information into the color representation. Our proposed transient regularization loss greatly increases the *Transient Uncertainty* to 0.917, showing the benefit from the additional guidance.

We provide close up visualizations of the uncertainty  $\beta$  and the corresponding rendered color image for two scenes in Fig. 4. Both contain parking lots with large numbers of vehicles parked across multiple training views. For both scenes the uncertainty  $\beta$  shows increased focus using our transient regularization loss, resulting in a drastic decrease in visible vehicles in the rendered color image. Only the static elements of the scene are rendered by the model.

### 6.3. Multi-View Consistency

*NeRF* learn a generalized scene understanding through maximizing consistency across all training views. This multi-view consistency reflects an averaging of information between input images. The semantic prediction potentially benefits from this process, improving semantic segmentation quality through enforcing multi-view consistency.

*Semantic NeRF* [19] conducts multiple experiments regarding the effect of partial or corrupted labels. They show that the network is able to learn a meaningful semantic representation, even if pixel-wise noise is added or entire instances are flipped to a different class. With these experiments they demonstrate that the underlying multi-view con-

sistency enforced by *NeRF* enables the generation of coherent semantic renderings even when training with partial and/or faulty labels. The *NeRF* allows to denoise and improve the initial semantic segmentations by rendering the same viewpoint. We conduct additional experiments researching this use case for the domain of segmenting satellite data.

We degrade the quality of our training data to more closely match the predicted output of automatic segmentation methods. We aim to simulate the effect of automatic segmentation methods correctly recognizing and segmenting most of an object, with smaller areas being incorrectly classified. We therefore explicitly refrain from pixel-wise noise and instance specific manipulation in contrast to *Semantic-NeRF* [19]. To corrupt our labels we generate a normal noise map for each semantic class. We blur and rescale the noise maps to create larger regions. Based on a specified accuracy loss, thresholds are determined to identify regions in each noise map that should be corrupted. Each noise region is reassigned to a different class. We release the corrupted labels alongside the code for the automated label corruption as part of the dataset. To evaluate the impact of training our pipeline with corrupted labels, we degrade the annotations for each scene by 20%, resulting in an overall semantic accuracy of 80%. The impact on the pixel-wise labels is illustrated in Figs. 5a and 5b.

Labels	Activation Function	Semantic Accuracy (Train) $\uparrow$				
		JAX_004	JAX_068	JAX_214	JAX_260	mean
Corrupted	-	0.800	0.800	0.800	0.800	0.800
Fused Labels	None	0.857	0.930	0.924	0.902	0.903
Fused Labels	Sigmoid	0.894	0.936	0.933	0.920	0.921

Table 3. Quantitative evaluation of the multi-view consistency on imperfect labels. Enforced multi-view consistency is able to improve initial semantic segmentations used as training input data. By using a sigmoid activation function for semantic logit prediction we can increase the semantic accuracy by a mean of 12.1%.

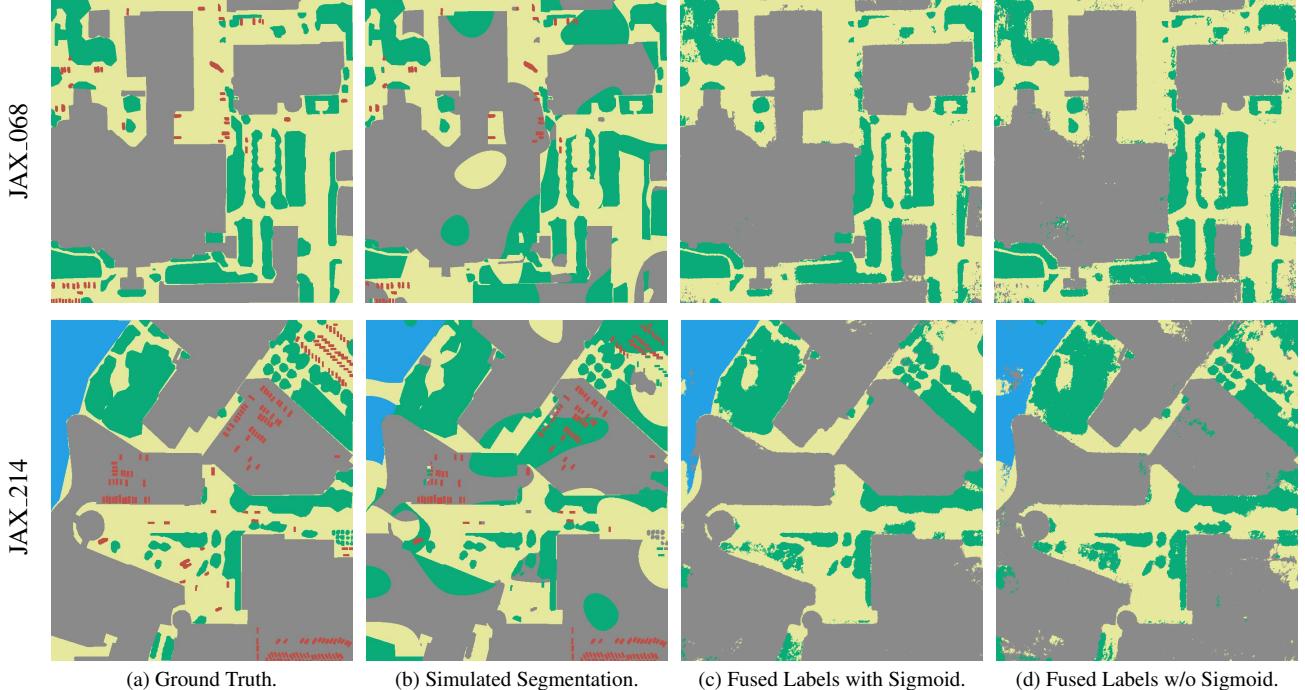


Figure 5. Effect of the multi-view consistency on local segmentation errors. The semantic information across all training views is merged in order to improve accuracy. Using a *sigmoid* activation function as normalization decreases noise in the reconstructed segmentation.

Rendering the same training views with our fully trained *NeRF* model leads to a considerable improvement in semantic accuracy. The results are reported in Tab. 3. Across all four scenes the semantic accuracy increases by a mean of 12.1% when using a sigmoid activation function for the semantic logits  $s$ . If using no activation function, as originally proposed by *Semantic-NeRF* [19], the semantic accuracy increases only by 10.3%. Comparing the predicted semantic output of the *NeRF* in Fig. 5c with the corrupted labels in Fig. 5b shows the ability of the model to largely replace and fix the holes caused by the automatic label degradation. Using no activation function leads to a higher amount of noise as visible in Fig. 5d.

## 7. Conclusions

This paper proposes a novel *NeRF* architecture for generating a unified color and semantic 3D representation of scenes derived from satellite images and their corresponding pixel-wise labels. By leveraging semantic information of transient categories during the training process we are able to improve the appearance representation of corresponding areas. This allows the *NeRF* to learn a more generalized and artifact-free representation of the scene. Our method allows to enhance the quality of noisy input labels by exploiting the multi-view consistency offered by modern *NeRF* models. To support future research in this area, we provide a dataset featuring manually generated labels for four multi-view scenes. Additionally, we release the code for this work, including scripts for model training and evaluation.

## References

- [1] Nikhil Behari, Akshat Dave, Kushagra Tiwary, William Yang, and Ramesh Raskar. Sundial: 3d satellite understanding through direct, ambient, and complex lighting decomposition, 2023. [1](#), [2](#)
- [2] 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](#)
- [3] Marc Bosch, Kevin Foster, Gordon Christie, Sean Wang, Gregory D Hager, and Myron Brown. Semantic stereo for incidental satellite images, 2018. [4](#)
- [4] Sebastian Bullinger, Christoph Bodensteiner, and Michael Arens. 3d surface reconstruction from multi-date satellite images, 2021. [1](#)
- [5] C. de Franchis, E. Meinhardt-Llopis, J. Michel, J.-M. Morel, and G. Facciolo. An automatic and modular stereo pipeline for pushbroom images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3:49–56, 2014. [1](#)
- [6] Dawa Derksen and Dario Izzo. Shadow neural radiance fields for multi-view satellite photogrammetry. *CoRR*, abs/2104.09877, 2021. [1](#), [2](#), [3](#), [4](#)
- [7] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*, chapter 6: Camera Models. Cambridge University Press, ISBN: 0521540518, second edition, 2004. [2](#)
- [8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. [5](#)
- [9] Bertrand Le Saux, Naoto Yokoya, Ronny Hänsch, and Myron Brown. Data fusion contest 2019 (dfc2019), 2019. [4](#), [5](#)
- [10] Roger Marí, Gabriele Facciolo, and Thibaud Ehret. Multi-date earth observation nerf: The detail is in the shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2034–2044, June 2023. [1](#), [2](#), [4](#)
- [11] Roger Marí, Gabriele Facciolo, and Thibaud Ehret. Sat-nerf: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using rpc cameras, 2022. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)
- [12] 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, 2020. [1](#), [2](#)
- [13] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017. [1](#)
- [14] C. Vincent Tao and Yong Hu. A comprehensive study of the rational function model for photogrammetric processing. *Photogrammetric Engineering and Remote Sensing*, 67:1347–1357, 2001. [2](#)
- [15] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds, 2019. [1](#)
- [16] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xi-hui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger, 2024. [1](#)
- [17] Kai Zhang, Jin Sun, and Noah Snavely. Leveraging vision reconstruction pipelines for satellite imagery, 2019. [1](#)
- [18] Mingtong Zhang, Shuhong Zheng, Zhipeng Bao, Martial Hebert, and Yu-Xiong Wang. Beyond rgb: Scene-property synthesis with neural radiance fields, 2022. [1](#)
- [19] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation. *CoRR*, abs/2103.15875, 2021. [1](#), [2](#), [4](#), [6](#), [7](#), [8](#)