SatSynth: Augmenting Image-Mask Pairs through Diffusion Models for Aerial Semantic Segmentation
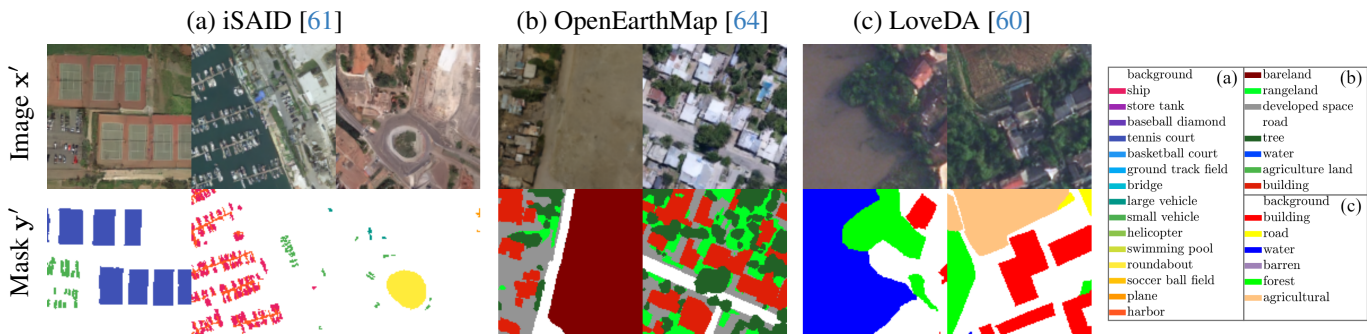
Aysim Toker[1]   Marvin Eisenberger[1]   Daniel Cremers[1]   Laura Leal-Taixé[2]

[1]Technical University of Munich   [2]NVIDIA

Figure 1. We leverage generative image diffusion to synthesize novel training data instances $(\mathbf{x}', \mathbf{y}') \sim p(\mathbf{x}, \mathbf{y})$ for a given labeled earth observation dataset [60, 61, 64]. In our experiments, we demonstrate that integrating such synthetic pairs as training data for downstream semantic segmentation yields significant quantitative improvements.

Abstract

In recent years, semantic segmentation has become a pivotal tool in processing and interpreting satellite imagery. Yet, a prevalent limitation of supervised learning techniques remains the need for extensive manual annotations by experts. In this work, we explore the potential of generative image diffusion to address the scarcity of annotated data in earth observation tasks. The main idea is to learn the joint data manifold of images and labels, leveraging recent advancements in denoising diffusion probabilistic models. To the best of our knowledge, we are the first to generate both images and corresponding masks for satellite segmentation. We find that the obtained pairs not only display high quality in fine-scale features but also ensure a wide sampling diversity. Both aspects are crucial for earth observation data, where semantic classes can vary severely in scale and occurrence frequency. We employ the novel data instances for downstream segmentation, as a form of data augmentation. In our experiments, we provide comparisons to prior works based on discriminative diffusion models or GANs. We demonstrate that integrating generated samples yields significant quantitative improvements for satellite semantic segmentation – both compared to baselines and when training only on the original data.

1. Introduction

Satellite imagery is a powerful tool to monitor the earth's surface, both in terms of specific events and global trends in land use. This has direct implications for humanitarian challenges such as disaster response, food security, and quantifying the impact of climate change. The United Nations summarizes a number of landmark objectives in its Sustainable Development Goals (SDGs)[1], in which they include general long-term goals such as the accessibility of clean water (SDG-6), reducing carbon emissions (SDG-13), or maintaining forests and combating desertification (SDG-15). In this context, satellite data can help provide crucial insights for monitoring progress, facilitating more targeted interventions. At the beginning of August 2023, the extent of Antarctic sea ice was observed to be 2.4 million km^2 less than the mean value from previous records (1979 to 2022), an area larger than Greenland. Monitoring this alarming trend, now commonly referred to as a five-sigma event among experts, was largely made possible through the wide-spread use of satellite observations.

While raw satellite data is readily available from various sources, obtaining corresponding semantic labels is challenging and costly due to the need for extensive manual

---

[1]https://www.un.org/sustainabledevelopment/sustainable-development-goals/

annotation. Such labels are imperative for numerous applications, since they enable us to reason about the semantic content of satellite scenes, particularly through supervised learning. For instance, the relative extent of Antarctic sea ice is directly reflected in respective land-cover annotations. A common solution is to leverage data augmentation to increase the sample diversity and make optimal use of existing labels. However, conventional image augmentation techniques designed for object-centric data – such as flipping, rotating, and rescaling – are often insufficient to emulate the large sample diversity of satellite imagery. Individual scenes typically contain a multitude of different object instances and land-cover categories. Semantic correlations within these images are mostly local. This vast diversity implies that most existing datasets are sparse, covering only a fraction of potential earth observation scenes.

In this work, we advocate for enhancing semantic segmentation of satellite data by harnessing recent advances in generative diffusion models [27]. Such models approximate the distribution of an existing dataset, to produce novel samples. We demonstrate that this can be leveraged to train a model that imitates the joint distribution of images and semantic segmentation labels in a given satellite dataset. Sampling from this distribution effectively enables us to generate additional training data as a form of data augmentation, see Fig. 1 for several such sample pairs. The enhanced training set can then be utilized for downstream semantic segmentation. In a broader context, our work serves as a study of the potential of image diffusion for data synthesis when annotations are scarce and costly.

Contributions.
1. For a given earth observation dataset, we propose to learn the joint data distribution $p(\mathbf{x}, \mathbf{y})$ of images $\mathbf{x}$ and labels in bit space $\mathbf{y}$ via a diffusion model $\mathcal{G}$.
2. We employ $\mathcal{G}$ to generate novel training data instances as a form of data augmentation to enhance downstream semantic segmentation.
3. We demonstrate that integrating the synthesized pairs yields significant quantitative improvements on three satellite benchmarks [60, 61, 64].

2. Related work

Denoising diffusion models. In recent times, diffusion models have emerged as a central technique for image generation. Their main advantage is the ability to synthesize high-quality samples, on par with GANs, while being less prone to suffer from mode collapse [19, 27, 55, 56]. Besides unconditional generation of novel data instances, common applications of image diffusion models include inpainting [39, 45], super-resolution [29], style transfer [70], depth prediction [54], and general x-to-image tasks [47, 49, 52].

In the literature, there is a specific interest in generating images conditioned on textual input. Some well-known approaches include: Stable diffusion [49], DALLE-2 [47], Imagen [52], Imagen Video [28], and GLIDE [45].

Semantic segmentation. Segmenting images, *i.e.*, assigning semantic labels to each pixel in 2D space, is a key challenge in computer vision. Supervised learning has emerged as the central paradigm for this task, as evidenced by numerous approaches [4, 8, 10–12, 35, 50, 57, 71] and benchmark datasets [16, 36, 42, 74]. For a comprehensive review, we refer the reader to a recent survey [17].

Compared to mainstream computer vision datasets, satellite imagery is subject to unique challenges. Due to the limited resolution of individual objects, distinct classes often exhibit a high visual similarity [67, 73]. Moreover, scale variations between different object categories can result in class imbalances. The relevant foreground classes are often dominated by a much larger background class. For instance, the ratio of foreground pixels is 29.75% in VOC2012 [20], whereas it is only 2.85% in the iSAID satellite dataset [61, 73]. Common solutions to these challenges involve specific architectures that emphasize salient foreground object features and increase robustness to background noise [34, 67, 73].

Despite the significant progress, supervised learning approaches still require dense ground-truth labels, which are both costly and difficult to obtain. To decrease the demand for annotated data, several works [2, 40, 41] explore self-supervised learning. Most such approaches harness some of the unique characteristics inherent to satellite imagery, such as time-series observations [59] or geolocation metadata [2]. Similarly, SatMAE [15] proposes large-scale pre-training for spectral and temporal satellite images using masked autoencoders [24]. Rather than leveraging self-supervision for representation learning and fine-tuning for segmentation, we instead propose to synthesize labeled data points directly via a generative image diffusion model.

Synthesizing training data. Building on the success of text-to-image diffusion models, a number of works employ them to enhance discriminative tasks. For instance, [25] explores whether such models are suitable for image recognition in the data-scarce regime. Specifically, they conduct zero-shot and transfer learning experiments, formulating text prompts derived from class label names. In a similar vein, [3, 6] generate text-image pairs, which are then used as additional training data for image classification.

Beyond diffusion models, generating synthetic datasets is a well-established technique to produce vast amounts of labeled data with minimal human input. Several works [21, 46, 48, 51, 72] use 3D graphic engines to emulate real-world scenes.

Other approaches [33, 69] utilize generative adversarial networks (GANs) [23] to synthesize high-quality datasets. Similar to ours, SemGAN [32] fits the training distribution for both images and masks. They then apply test-time optimization for a given query image and extract semantic labels by aligning the learned feature embeddings. In contrast, we propose to directly extract joint pairs of data instances and employ them as training data for downstream semantic segmentation.

Image diffusion for segmentation. There are several ways in which diffusion models can improve segmentation tasks. A common strategy is to obtain masks conditioned on input images, adopting the standard image-to-image translation paradigm [1]. This approach was subsequently extended to synthesize videos and panoptic masks [13] for the discrete data representation introduced in Analog-Bits [14]. Another related line of works leverage pre-trained text-to-image models, such as stable diffusion, with tailored text prompts to address open vocabulary segmentation [30, 62, 66] and object-centric segmentation [44, 58]. Alternatively, some methods leverage diffusion models for self-supervised pre-training. The resulting feature embeddings are subsequently fine-tuned for downstream applications, such as image segmentation [7, 9] or binary change detection [5]. To the best of our knowledge, ours is the first approach that employs diffusion models to jointly generate satellite scenes and the corresponding labels to augment a given training dataset for semantic segmentation.

3. Preliminaries

We provide a brief overview of denoising diffusion probabilistic models (DDPM) as presented in [27]. The main idea is to devise a generative model capable of synthesizing images by reversing a stochastic Gaussian noising process

$$\mathbf{x}_t := \sqrt{1 - \beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \quad (1)$$

Here, the hyperparameter $\beta_t > 0$ specifies the noise variance at each timestep. The model's forward process follows a predefined schedule of progressive noising steps $\mathbf{x}_0 \rightarrow \mathbf{x}_1 \rightarrow \cdots \rightarrow \mathbf{x}_T$. For sufficiently large T, we effectively obtain a Gaussian random sample $\lim_{T \to \infty} \mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$. This noising process is subsequently reversed $\mathbf{x}_T \rightarrow \cdots \rightarrow \mathbf{x}_0$ by utilizing a U-Net backbone [50] to predict the noise vectors from the noisy samples $\mathbf{x}_t$. Through appropriate reparameterization, it effectively learns a mapping $\mathbf{x}_t \mapsto \mathbf{x}_{t-1}$. Starting from a random sample $\mathbf{x}_T$, this results in an iterative inverse process

$$\mathcal{G} : \mathbb{R}^L \rightarrow \mathbb{R}^{H \times W \times C}. \quad (2)$$

This stochastic generative model $\mathcal{G}$ maps from a predefined noise distribution $\mathbf{z} \sim \mu := \mathcal{N}(0, \mathbf{I}_L)$ to synthesized images

$\mathbf{x} := \mathcal{G}(\mathbf{z})$ that follow the input training distribution. In the most general setting, we typically have $L = THWC$. For further technical details and in-depth explanations of DDPM, we refer the reader to the original publication [27].

4. Method

4.1. Problem statement

In this work, we address the task of semantic segmentation for earth observation data. Specifically, we consider a dataset

$$\mathcal{D} := \big\{(\mathbf{x}_i, \mathbf{y}_i) \big| \mathbf{x}_i \in \mathbb{R}^{H \times W \times 3},$$
$$\mathbf{y}_i \in \{0, \dots, K-1\}^{H \times W}, 1 \leq i \leq N\big\}, \quad (3)$$

of N satellite images $\mathbf{x}_i$ and corresponding semantic maps $\mathbf{y}_i$ with a spatial size of $H \times W$, and K distinct semantic classes.

We further assume that $\mathcal{D}$ is sampled from an underlying latent data manifold $\mathcal{M}$, which we can only access indirectly through the given instances $(\mathbf{x}_i, \mathbf{y}_i) \sim \mathcal{M}$. The goal of semantic segmentation approaches is to devise a discriminative model for the inference task specified by $p(\mathbf{y}|\mathbf{x})$. Standard supervised learning approaches estimate this conditional probability by training a model on the discrete set of samples $\mathcal{D} \subset \mathcal{M}$. This approach, however, requires that $\mathcal{D}$ yields an adequate coverage, *i.e.*, is sufficiently large and diverse. This poses a significant challenge for earth observation, where labeled data is often limited.

Motivation. In our approach, we circumvent this classical data bottleneck by leveraging recent advances in generative modeling. Instead of predicting $p(\mathbf{y}|\mathbf{x})$ directly, we first approximate the joint distribution $p(\mathbf{x}, \mathbf{y})$ with an unconditional image diffusion model. The resulting network $\mathcal{G}$ subsequently enables us to generate novel samples $(\mathbf{x}'_i, \mathbf{y}'_i) \sim \mathcal{M}$. We then train a segmentation model on the joint dataset $\mathcal{D} \cup \mathcal{D}'$ comprising both synthetic $\mathcal{D}' := \big\{(\mathbf{x}'_1, \mathbf{y}'_1), \dots, (\mathbf{x}'_{N'}, \mathbf{y}'_{N'})\big\}$ and real $\mathcal{D}$ samples. For a graphical representation of our approach, refer to Fig. 2.

4.2. Discrete labels in bit-space

Most conventional generative models for image data focus on synthesizing instances from the source image distribution $p(\mathbf{x})$. Instead, we propose to generate new training instances according to the joint probability $p(\mathbf{x}, \mathbf{y})$ of images $\mathbf{x}$ and corresponding labels $\mathbf{y}$. Extending standard generative models from $\mathbf{x}$ to $(\mathbf{x}, \mathbf{y})$ is not straightforward, since each pixel in $\mathbf{y}$ is associated with a discrete label $\{0, \dots, K-1\}$, as opposed to continuous values in $\mathbb{R}$. We circumvent this issue by modeling the discrete values $\mathbf{y}$ in terms of their binary code, specified as

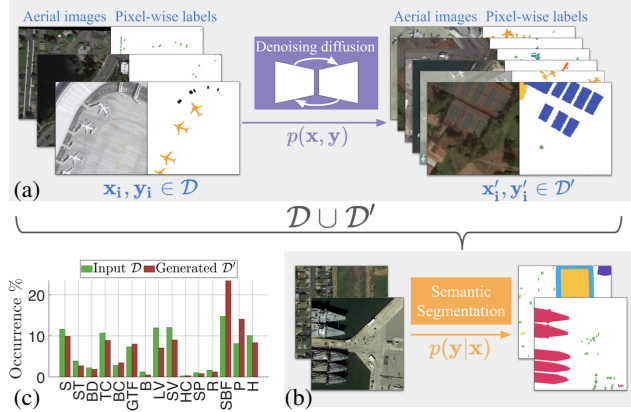$$\text{bin} : \{0, \dots, K-1\} \rightarrow \{0, 1\}^{\lceil \log_2 K \rceil}. \quad (4)$$

Figure 2. **Approach overview.** (a) We train a generative image diffusion model $\mathcal{G}$ on the joint data instances $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}$ of images $\mathbf{x}_i$ and corresponding labels $\mathbf{y}_i$. We then employ $\mathcal{G}$ to generate a dataset $\mathcal{D}'$ of novel training samples $(\mathbf{x}'_i, \mathbf{y}'_i)$. (b) Both the real $\mathcal{D}$ and generated $\mathcal{D}'$ pairs are integrated and leveraged for the downstream semantic segmentation task. (c) Moreover, we compare the resulting distributions of foreground classes, highlighting that the set of generated labels in $\mathcal{D}'$ closely matches the original distribution $\mathcal{D}$. For a legend of label acronyms, refer to Fig. 1 (a).

Compared to a standard one-hot encoding, this was shown to yield an improved stability for generative models in prior work [14]. It also leads to only a small (logarithmic) increase in dimensionality when considering a high number of classes K, see Tab. 3 for a comparison. In practice, we normalize the RGB values of the input images $\mathbf{x}_i$ to the same range $[0, 1]$ for compatibility with the binary values $\text{bin}(\mathbf{y}_i)$. This facilitates synthesizing both quantities jointly, which we describe in the next section.

4.3. Synthesizing satellite segmentation data

The transformation in Eq. (4) maps the discrete labels to the domain of binary values $\text{bin}(\mathbf{y}_i) \in \{0, 1\}^{H \times W \times \lceil \log_2 K \rceil}$. This allows us to simply concatenate them with the (normalized) RGB values as additional input channels. In order to synthesize pairs of novel data points $\mathbf{x}'_i$ and $\mathbf{y}'_i$, we follow a simple strategy:

1. Train a generative model $\mathcal{G}$ on all joint data instances $\left(\mathbf{x}_i, \text{bin}(\mathbf{y}_i)\right) \in [0, 1]^{H \times W \times (3 + \lceil \log_2 K \rceil)}$ in $\mathcal{D}$.
2. Generate novel, synthetic samples $(\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i) \sim \mathcal{G}$.
3. Threshold and transform $(\mathbf{x}'_i, \mathbf{y}'_i) := (\hat{\mathbf{x}}_i, \text{bin}^{-1}(\hat{\mathbf{y}}_i))$.
4. Employ $\mathcal{D}' := \left\{ (\mathbf{x}'_i, \mathbf{y}'_i) | 1 \le i \le N' \right\}$ for downstream tasks, *e.g.*, train a segmentation model.

While other choices are possible, in this work we leverage recent advances in state-of-the-art diffusion models $\mathcal{G}$. The exact architecture and training schedule is based on DDPM [27], see Sec. 3 for a brief overview.

4.4. Image super-resolution

In the context of diffusion models, many architectures that operate in image-space focus on relatively coarse resolutions $H = W \le 128$. However, for many earth observation tasks, a sufficient level of detail is crucial. Although in theory, existing models can be trained for higher resolutions, there are significant practical limitations [49]. For once, the training cost of such models, which are already computationally intensive, increases further. Additionally, it often leads to unstable training behaviour and inferior samples, see Fig. 10 in the appendix for a comparative analysis.

Instead of training our model for higher resolutions directly, we leverage recent advances in image super-resolution [29]. Rather than introducing a new architecture, we simply employ a conditional variant of the architecture detailed in Sec. 4.2. Specifically, we train a DDPM image-to-image translation model that jointly generates images and labels in bit space, while being conditioned on the corresponding low-resolution samples:

$$\mathcal{G}_{\text{SR}} : \mathbb{R}^L \times \mathbb{R}^{H \times W \times C} \to \mathbb{R}^{2H \times 2W \times C}. \quad (5)$$

The first input is a random noise vector $\mathbf{z} \in \mathbb{R}^L$ analogous to Eq. (2), whereas the second input is the conditional low-resolution image-mask pair with a channel size of $C = 3 + \lceil \log_2 K \rceil$. At each timestep, the current prediction of the denoising U-Net is concatenated with the low-resolution pair. During test time, we generate higher-resolution samples in two consecutive steps: $(\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i) := \mathcal{G}_{\text{SR}}(\mathbf{z}_1, \mathcal{G}(\mathbf{z}_0))$. We provide visualizations of several obtained super-resolution pairs in Fig. 6.

4.5. Implementation details

Diffusion. As described above, we employ DDPM [27] to learn the joint training distribution $\mathcal{D}$. Specifically, we apply $T = 1000$ consecutive denoising steps with a linear noise schedule $\beta_t \in \{1e-4, \ldots, 2e-2\}$. We consistently generate images with a spatial size of 128×128 and optionally upsample them to 256×256 as specified in Sec. 4.4.

In Sec. 4.3, we denote the inverse binary transformation with a slight abuse of notation as bin^{-1}. In practice, this involves a combination of thresholding the continuous values $\hat{\mathbf{y}}_i$, as well as mapping the resulting binary values to the original index domain $\{0, \ldots, K-1\}$. Since K is not necessarily a power of 2, we use a simple nearest-neighbor assignment of $\hat{\mathbf{y}}_i$ to $\left\{ \text{bin}(0), \ldots, \text{bin}(K-1) \right\} \subset \{0, 1\}^{\lceil \log_2 K \rceil}$ to ensure bijectivity.

Prior to querying the denoising U-Net backbone, an additional linear transformation $z \mapsto 2z - 1$ is applied to the input pairs for an improved numerical stability. At test time, the generated samples are clipped to the range $[-1, 1]$, before the inverse transformation $z \mapsto 0.5(z + 1)$ maps the values back to the original interval $[0, 1]$.

	iSAID			LoveDA			OpenEarthMap		
	FID (↓)	sFID (↓)	IS (↑)	FID (↓)	sFID (↓)	IS (↑)	FID (↓)	sFID (↓)	IS (↑)
SemGAN [32]	10.21	3.63	1.11	37.47	7.16	3.29	16.20	4.89	3.13
DDPM [27]	17.50	7.07	1.03	22.70	2.10	3.44	15.35	4.87	**3.31**
Ours	**8.66**	**3.10**	**1.17**	**13.87**	**1.87**	**3.88**	**12.09**	**2.76**	3.16

(a) Visual sample quality.

	iSAID		LoveDA		OpenEarthMap	
	IoU (↑)	F1 (↑)	IoU (↑)	F1 (↑)	IoU (↑)	F1 (↑)
SemGAN [32]	13.01	18.08	31.75	42.74	40.43	54.75
SegDiff [1]	41.25	54.77	36.60	49.93	51.23	65.70
Ours	**52.13**	**66.13**	**48.97**	**64.83**	**62.24**	**76.10**

(b) Semantic segmentation.

Table 1. **Comparative analysis of generative models.** We present quantitative comparisons on three distinct satellite benchmarks [60, 61, 64]. (a) For once, we assess the visual quality of the synthetic satellite images $\mathbf{x}'_i \in \mathbb{R}^{128 \times 128 \times 3}$ obtained with our approach, compared to images generated by SemGAN [32] and DDPM [27]. (b) We further quantify the segmentation accuracy of our approach with the FPN segmentation backbone, comparing it to the generative semantic segmentation approaches SemGAN [32] and SegDiff [1]. In both tables, (↓) indicates lower metric values are better, whereas (↑) denotes higher values are better.
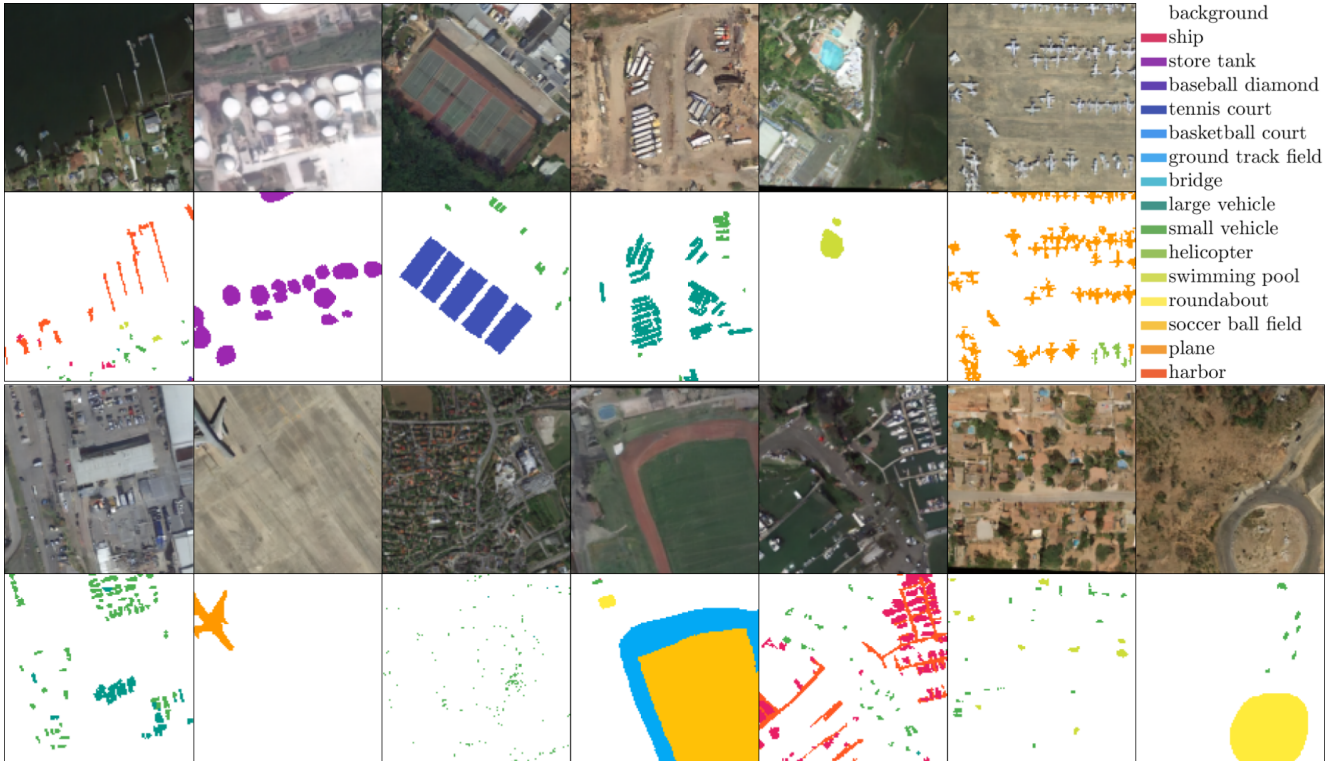

Figure 3. **Generated samples, iSAID [61].** We visualize several pairs $(\mathbf{x}'_i, \mathbf{y}'_i)$ sampled from the diffusion model $\mathcal{G}$ detailed in Sec. 4.3. Color coding for the semantic masks $\mathbf{y}'_i$ is indicated by the corresponding palette legend (top right). The generated scenes are of high quality and the semantic layout is coherent – for instance, soccer ball fields are frequently surrounded by ground track fields (bottom, 4th).

Segmentation. By default, we utilize a standard feature pyramid network (FPN) [31, 37] with a ResNet50 backbone to perform multi-class semantic segmentation. The hierarchical design of FPN has proven to be particularly effective in state-of-the-art earth observation approaches [34, 73]. It allows for predicting accurate masks in the presence of large scale variations and detailed fine-scale structures. For a more complete picture, we additionally include a recent vision transformer segmentation model SegFormer [65] and the state-of-the-art satellite segmentation models PFSeg-Net [34] and FarSeg [73] for specific settings.

5. Experiments

We evaluate the impact of synthesized training samples $\mathcal{D}'$ on the task of semantic segmentation for earth observation data. We provide direct comparisons to prior techniques on generative models, thorough analyses of our generated data, and segmentation results for diverse settings.

5.1. Datasets

We consider three popular earth observation benchmarks that address object-centric segmentation [61] and land-cover classes [60, 64], respectively.

Figure 4. **Generated samples, LoveDA [60].** We display pairs $(\mathbf{x}'_i, \mathbf{y}'_i)$ generated by $\mathcal{G}$ on LoveDA [60]. The obtained satellite scenes consist of visually plausible images $\mathbf{x}'_i$ and corresponding semantic masks $\mathbf{y}'_i$ for general land-cover classes.

iSAID. The iSAID [61] dataset focuses on semantic segmentation of individual object categories such as cars, bridges, or tennis courts. Overall, it contains 2,806 high-resolution satellite images with 655,451 object instances from 15 classes. The original source of the data is the DOTA dataset [63], while the semantic labels were annotated specifically for iSAID. Since the main goal of the dataset is to segment individual objects, it is prone to large scale variation and class imbalances. Certain object categories occupy only a small number of pixels ($\approx 0.3\%$ pixels for 'small vehicles') compared to the dominant background class ($> 97\%$ pixels).

LoveDA. LoveDA [60] consists of 5,987 high resolution images of both rural and urban scenes, along with 166,768 individual land-cover annotations. A notable challenge are similarities in appearance of distinct categories across different geographical contexts. To increase the number of input images, we extract non-overlapping 256×256 patches from each image, resulting in a total of 40,352 training and 26,704 validation images.

OpenEarthMap. The OpenEarthMap [64] dataset integrates high-resolution satellite imagery from several different sources to create a unified benchmark for land-use and land-cover mapping. In particular, they assemble 5,000 images spanning 97 regions across 44 countries from 6 continents. Individual images have a resolution of 1024×1024 with a pixel granularity ≤ 0.5m. Similar to LoveDA, we extract 256×256 non-overlapping patches as a pre-processing step. The pixel-wise annotations encompass 8 distinct land-cover categories, namely: bareland, rangeland, developed space, road, tree, water, agriculture land, and buildings.

5.2. Comparisons to generative approaches

We evaluate our method in two distinct settings. First, we compare our synthesized images $\{\mathbf{x}'_1, \ldots, \mathbf{x}'_{N'}\}$ to alternative generative approaches in terms of visual sample quality. Second, we assess the impact of our synthesized image-mask pairs on semantic segmentation, contrasting our approach with methods employing generative models for segmentation.

Visual sample quality. We compare the sample quality of generated satellite images $\mathbf{x}'_i \in \mathbb{R}^{128 \times 128 \times 3}$ to two baseline approaches [27, 32] in Tab. 1a. The first baseline, SemGAN [32], synthesizes image-mask pairs through adversarial training. Additionally, we consider vanilla DDPM [27], trained solely on the input images $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. In each case, we report the Fréchet inception distance (FID) [26], spatial FID (sFID) [43], and the inception score (IS) [53], confirming the superior visual quality of our synthesized images. Specifically, we compute the IS in terms of the segmentation logits of a pretrained ResNet-50 U-Net segmentation model [50]. For the FID and sFID scores, we compare the generated feature distribution of the encoder to the distribution of real images $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. We further provide qualitative samples of generated training pairs on iSAID and LoveDA in Fig. 3 and Fig. 4, and on Open-EarthMap in Fig. 9 in the appendix.

Generative segmentation. Our proposed approach is based on synthesizing novel training data instances $\mathcal{D}'$. Further, we integrate the augmented training set $\mathcal{D} \cup \mathcal{D}'$ for downstream segmentation. While this approach is straightforward, there exist other potential strategies for employing generative models to predict segmentation masks directly. To provide context, we consider two baseline approaches. SemGAN [32] trains a joint generator and a separate encoder, applying test-time optimization to extract semantic labels conditioned on the latent embedding of a given query image. Conversely, SegDiff [1] directly models binary segmentation as an image-to-image regression task utilizing
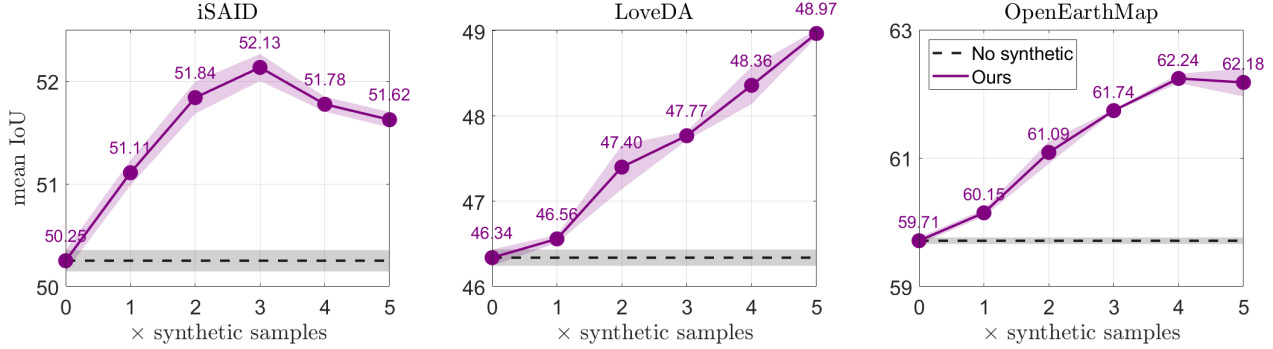
Figure 5. **Analysis of synthetic data.** We assess the impact of generated samples $\mathcal{D} \cup \mathcal{D}'$ on the mIoU segmentation score for iSAID [61], LoveDA [60], and OpenEarthMap [64], with a spatial size of 128×128. Different resampling ratios are applied, defined as sampling $R \in \{0, \dots, 5\}$ synthetic pairs per original instance, *i.e.*, $|\mathcal{D}'| = R \cdot |\mathcal{D}|$ pairs in total. In each case, error bars are provided which denote the standard error (SE). We separately plot the accuracies without synthetic samples $R = 0$ (gray dashed lines) for ease of comparison.

Train on	PFSegNet [34]		FarSeg [73]		SegFormer [65]		FPN [31]		PSPNet [71]	
256×256	IoU (↑)	F1 (↑)	IoU (↑)	F1 (↑)	IoU (↑)	F1 (↑)	IoU (↑)	F1 (↑)	IoU (↑)	F1 (↑)
$\mathcal{D}$	60.93±0.62	74.10±0.55	62.28±0.27	75.16±0.22	60.95±0.61	74.18±0.55	59.52±0.19	72.82±0.15	48.95±2.91	63.13±2.83
$\mathcal{D} \cup \mathcal{D}'$	**63.71±0.21**	**76.37±0.17**	**62.95±0.25**	**75.72±0.28**	**62.13±0.25**	**75.10±0.21**	**60.65±0.38**	**73.69±0.33**	**56.54±1.32**	**70.16±1.17**

Table 2. **Object-centric segmentation**. We demonstrate that integrating our generated training pairs $\mathcal{D} \cup \mathcal{D}'$ improves the performance over the original data $\mathcal{D}$ on the iSAID [61] benchmark. To this end, we consider the recent state-of-the-art approaches PFSegNet [34] and FarSeg [73] that specialize on satellite segmentation, as well as three generic segmentation models [31, 65, 71]. In each setting, we utilize super-resolution pairs with a spatial size of 256×256 as defined in Sec. 4.4.

DDPM. Although it is not primarily designed for multi-class scenarios, we can extend it in a straightforward manner by predicting labels in bit space, as detailed in Eq. (4).

We report the resulting accuracies in Tab. 1b. For our approach, we train a standard FPN backbone on the augmented training set $\mathcal{D} \cup \mathcal{D}'$. Compared to SemGAN [32] and SegDiff [1], our approach yields significantly more accurate predictions, both in terms of the mIoU and F1 score. A visualization of the predicted semantic maps is provided in Fig. 11 in the appendix.

5.3. Analysis of synthetic training data

We assess the impact of additional synthetic samples $\mathcal{D}'$ on satellite semantic segmentation with an FPN backbone. To this end, we experiment with various resampling ratios $|\mathcal{D}'| = R \cdot |\mathcal{D}|$, where $R \in \mathbb{N}$. In principle, we can sample an arbitrary number of pairs $(\mathbf{x}'_i, \mathbf{y}'_i)$ from the diffusion model $\mathcal{G}$ specified in Sec. 4.3. However, as we generate more samples, they become increasingly correlated and redundant – yielding diminishing returns.

We report the resulting mIoU scores on different benchmarks [60, 61, 64] for values of $R \in \{0, \dots, 5\}$ in Fig. 5. In each scenario, we quantify how training on the joint dataset $\mathcal{D} \cup \mathcal{D}'$ compares to the original samples $\mathcal{D}$. We apply oversampling to ensure a balanced ratio between real and synthetic samples. In each setting, we observe

that adding synthetic pairs consistently enhances the performance. This confirms our assertion that integrating generated pairs serves as a form of data augmentation. Moreover, the optimal performance varies on specific benchmarks, for instance, a resampling ratio of $R = 3$ is ideal for iSAID.

5.4. Object-centric segmentation

We evaluate our approach in terms of object-centric satellite segmentation, see Tab. 2 for a summary of the resulting accuracies. Specifically, we report the mIoU and F1 scores on the iSAID [61] benchmark for several state-of-the-art satellite [34, 73] and general [31, 65, 71] segmentation approaches. These results demonstrate that training on the augmented dataset $\mathcal{D} \cup \mathcal{D}'$ consistently improves the performance over $\mathcal{D}$. For instance, the mIoU score increases by 7.59% for PSPNet [71]. Since most considered segmentation methods expect high input resolutions, we leverage our super-resolution model $\mathcal{G}_{\text{SR}}$ specified in Sec. 4.4 to generate pairs with a spatial size of 256×256. While many satellite baselines consider even higher resolutions > 256 in their original publications [34, 73], we found this to be a reasonable trade-off due to the substantial computational demand of high-resolution diffusion models. For qualitative visualizations of different super-resolution image-mask pairs, see Fig. 6.
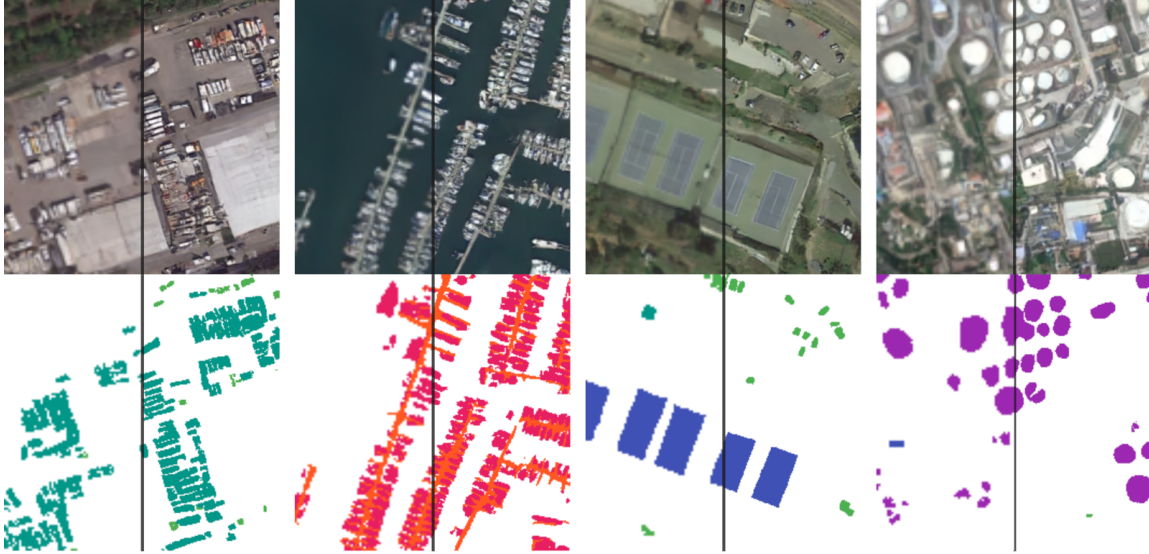
Figure 6. **Super-resolution, qualitative.** We provide four super-resolution samples obtained on the iSAID dataset. These generated pairs are obtained in two steps, querying $\mathcal{G}$ to produce low-resolution images with a spatial size of 128×128 (left), before upsampling them via $\mathcal{G}_{\mathrm{SR}}$ to 256×256 (right) as defined in Eq. (5).

5.5. Ablation study

In Sec. 4.3, we propose to jointly generate images and labels $(\mathbf{x}, \mathbf{y})$ for earth observation data via a denoising diffusion model $\mathcal{G}$. We now assess the impact of specific assumptions on the quality of the generated samples, see Tab. 3 for a summary. Specifically, we compare the binary encoding $\mathrm{bin}(\mathbf{y})$ defined in Eq. (4) to conventional one-hot encoding $\mathrm{OH}(\mathbf{y})_k := \mathrm{I}[\mathbf{y} = k]$. The findings in Tab. 3 indicate that the binary coding yields superior samples. We attribute this to the more compact encoding of $\lceil \log_2 K \rceil$ channels, as opposed to K for one-hot. For instance, for the $K = 16$ classes in iSAID, the binary codes require only 4 channels, resulting in well-balanced joint samples $(\mathbf{x}, \mathbf{y})$ with a total of $C = 7$ channels.

We further investigate the prediction type targeted by the diffusion U-Net backbone. There exist two prevalent choices, where in each denoising step $\mathbf{x}_t$ the network either predicts the total noise vector ϵ, or the original sample $\mathbf{x}_0$. From both quantities, the noise vector between two timesteps $\mathbf{x}_{t-1} \rightarrow \mathbf{x}_t$ can be derived. On the other hand, we find that, for our purposes, predicting ϵ leads to a superior segmentation performance – especially for the higher resampling ratio $R = 3$.

6. Conclusion

Our work serves as a showcase for the potential of image diffusion models for data generation in domains where ground truth labels are scarce and costly. For satellite imagery, specifically, obtaining such labels requires extensive manual annotation by human experts. In a broader context,

Label encoding		Predict		$R = 1$		$R = 3$	
$\mathrm{bin}(\mathbf{y})$	$\mathrm{I}[\mathbf{y} = k]$	ϵ	$\mathbf{x}_0$	IoU (↑)	F1 (↑)	IoU (↑)	F1 (↑)
	✓	✓		47.98	62.08	49.38	63.71
✓			✓	51.00	65.05	51.45	65.54
✓		✓		**51.11**	**65.13**	**52.13**	**66.13**

Table 3. **Ablation study.** We investigate how two central design choices for the joint data generation, outlined in Sec. 4.3, relate to the downstream segmentation performance on iSAID. Specifically, we compare two variants of the network $\mathcal{G}$, predicting the noise component ϵ or the initial sample $\mathbf{x}_0$, and we contrast the binary label embedding with standard one-hot encoding. These results confirm that both components are crucial (lower row, bin+ϵ) for obtaining optimal results.

we anticipate that approaches similar to ours will become increasingly ubiquitous for data synthesis and augmentation tasks in different areas. For instance, an average of 1.5 hours was spent to annotate a single image from Cityscapes, according to the original publication [16].

Given the high visual fidelity of the obtained instances, we plan to explore extensions to other related earth observation tasks in future work. Potential applications include sample data fusion of different observations, image inpainting for cloud removal, or change detection.

27702

References

[1] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 3, 5, 6, 7, 2

[2] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 2

[3] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 2

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 2

[5] Wele Gedara Chaminda Bandara, Nithin Gopalakrishnan Nair, and Vishal M Patel. Ddpm-cd: Remote sensing change detection using denoising diffusion probabilistic models. *arXiv preprint arXiv:2206.11892*, 2022. 3

[6] Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. *arXiv preprint arXiv:2302.02503*, 2023. 2

[7] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3

[8] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Semantic segmentation with boundary neural fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3602–3610, 2016. 2

[9] Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4175–4186, 2022. 3

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[11] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[12] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2

[13] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. *arXiv preprint arXiv:2210.06366*, 2022. 3

[14] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022. 3, 4

[15] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022. 2

[16] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2, 8

[17] Gabriela Csurka, Riccardo Volpi, Boris Chidlovskii, et al. Semantic image segmentation: Two decades of research. *Foundations and Trends® in Computer Graphics and Vision*, 14(1-2):1–162, 2022. 2

[18] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2

[19] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[20] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 2

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

[22] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021. 2

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[25] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 2

[26] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 4, 5, 6, 1

[28] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[29] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 2, 4

[30] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. *arXiv preprint arXiv:2306.09316*, 2023. 3

[31] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 5, 7, 1, 3

[32] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8300–8311, 2021. 3, 5, 6, 7, 2

[33] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21330–21340, 2022. 3

[34] Xiangtai Li, Hao He, Xia Li, Duo Li, Guangliang Cheng, Jianping Shi, Lubin Weng, Yunhai Tong, and Zhouchen Lin. Pointflow: Flowing semantics through points for aerial image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4217–4226, 2021. 2, 5, 7, 1, 3

[35] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 2

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[37] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 5

[38] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 5

[39] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 2

[40] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5261–5270, 2023. 2

[41] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021. 2

[42] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 2

[43] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021. 6

[44] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[45] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[46] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018. 2

[47] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[48] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 2

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 3, 6

[51] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large

collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 2

[52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[53] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 6

[54] Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023. 2

[55] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2015. 2

[56] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[57] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5229–5238, 2019. 2

[58] Weimin Tan, Siyuan Chen, and Bo Yan. Diffss: Diffusion model for few-shot semantic segmentation. *arXiv preprint arXiv:2307.00773*, 2023. 3

[59] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21158–21167, 2022. 2

[60] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 1, 2, 5, 6, 7

[61] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. 1, 2, 5, 6, 7

[62] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *arXiv preprint arXiv:2303.11681*, 2023. 3

[63] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 6

[64] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6254–6264, 2023. 1, 2, 5, 6, 7, 3

[65] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 5, 7, 1, 3

[66] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 3

[67] Fengyu Yang and Chenyang Ma. Sparse and complete latent organization for geospatial semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1809–1818, 2022. 2

[68] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 2

[69] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021. 3

[70] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10156, 2023. 2

[71] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2, 7, 1, 3

[72] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 519–535. Springer, 2020. 2

[73] Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4096–4105, 2020. 2, 5, 7, 1, 3

[74] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2