SatSynth: Augmenting Image-Mask Pairs through Diffusion Models for Aerial Semantic Segmentation

Supplementary Material

A. Object-centric segmentation analysis

Our results of object-centric segmentation on iSAID in Sec. 5.4 demonstrate substantial improvements for five separate baseline approaches, including general segmentation models [31, 65, 71] and approaches tailored for satellite imagery [34, 73]. We additionally provide a per-class analysis of the results reported in Tab. 2 of the main paper, summarized in Tab. 6. A key insight is that, beyond overall improvements of the average scores, a majority of individual classes benefit from the augmentation. In the extreme case, our approach yields a 15.26% gain in the IoU score (SBF, PSPNet [71]), whereas the most significant drop in performance is 2.79% (HC, SegFormer [65]). For the BC class, the mean and median increase over all baselines is 5.55% and 4.75%, respectively.

Remarkably, even for the overall best performing baseline PFSegNet [34], our approach still yields significant improvements for all but one classes, with an increased IoU score of up to 8.98% (basketball court). We conclude that the observed improvements of leveraging our synthesized data are homogeneous and consistent throughout all considered settings and for most individual classes.

We further provide an analysis of the impact of additional synthetic samples on rare classes, see Fig. 7. For each of the 15 foreground classes, we consider the absolute improvement of the mean IoU score on PSPNet [71]. This is contrasted with the relative class occurrence, defined as the fraction of images that contain any such instances. We observe a negative Pearson correlation coefficient of -0.47, which indicates that the generated samples help mitigate class imbalances.

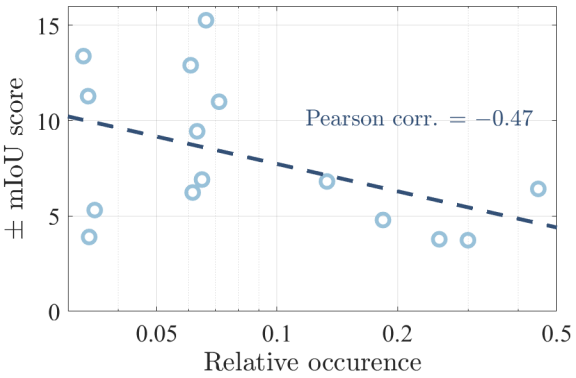

Figure 7. **Class imbalance.** We analyze the per-class IoU score for the results from Tab. 2 in the main paper. Specifically, we contrast the relative occurrence of the 15 foreground classes, with their absolute improvement in the mean IoU score on PSPNet [71]. The resulting negative correlation confirms that our approach implicitly mitigates class imbalances, since rare classes disproportionally benefit from additional generated samples.

B. Super-resolution discussion

In Sec. 4.4, we devise a super-resolution approach that allows us to upsample generated images to a resolution of 256×256. Specifically, we utilize a diffusion-based super-resolution model $\mathcal{G}_{\text{SR}}$ that takes generated images with a size of 128×128 as a conditional input to the denoising U-Net. While it is conceivable to extend this approach to even higher resolutions > 256, considered by some existing satellite segmentation baselines [34, 73], we leave such investigations for future work due to the substantial computational demand of high-resolution diffusion models.

As a straightforward alternative to our super-resolution approach, we employ DDPM [27] to directly generate samples with a spatial size of 256×256. The resulting accuracies of both approaches on the iSAID dataset are summarized in Tab. 4, considering two standard backbones FPN [31] and SegFormer [65]. The experimental setting is analogous to Tab. 2 in the main paper.

These results indicate that our super-resolution approach yields more consistent results compared to the direct DDPM generations. To investigate this effect, we additionally provide visualizations of the resulting pairs in Fig. 10. While both approaches yield comparable generations in terms of fine-scale details, the images and masks obtained with DDPM-256 are less coherent in the overall semantic layout. Moreover, the convergence behavior of DDPM-256 is less stable, producing erroneous image contrasts and saturations. While both approaches yield improvements for FPN [31], the lack of semantic coherence slightly decreases the performance of SegFormer [65]. Our super-resolution strategy effectively decouples the challenges of creating semantically consistent images through $\mathcal{G}$ and recovering fine-scale details through $\mathcal{G}_{\text{SR}}$, leading to a superior downstream performance in Tab. 4.

C. Advanced data augmentation

In Tab. 1b of the main paper, we demonstrate consistent quantitative improvements in downstream segmentation tasks compared to existing generative approaches. Here, we

	FPN		*SegFormer*	
	IoU (↑)	F1 (↑)	IoU (↑)	F1 (↑)
No synthetic	59.52	72.82	60.95	74.18
DDPM-256	60.30	73.22	60.61	73.85
Ours	**60.65**	**73.69**	**62.13**	**75.10**

Table 4. **Super-resolution.** We compare our super-resolution approach to directly generating synthetic samples with DDPM, analogous to our approach in Sec. 4.3. We find that DDPM exhibits unstable training behaviour for resolutions $H = W \geq 256$ which results in a suboptimal downstream segmentation performance on iSAID. The obtained joint samples display notable artifacts, particularly in terms of the saturation and contrast of the generated images, refer to Fig. 10 for a qualitative comparison.

No additional samples	Ours	Cutout [18]	CutMix [68]	Copy-Paste [22]
50.25	**51.11**	50.47	50.60	50.51

Table 5. **Quantitative comparison of augmentation methods.** We compare our method to the recent augmentation techniques Cutout [18], CutMix [68], and Copy-Paste [22]. Across all experiments, we generate additional training pairs with a resampling ratio of $R = 1$. The experimental setup is equivalent to the results on iSAID reported in Fig. 5 of the main paper.

evaluate the effectiveness of our approach against state-of-the-art augmentation techniques such as Cutout [18], CutMix [68], and Copy-Paste [22].

Cutout [18] applies regional dropout in the input space for image classification. We adapt this to semantic segmentation by masking out random squares from both the image and its corresponding semantic mask. CutMix [68] crops random regions from one image and pastes them onto another image, along with the corresponding masks. The instance segmentation augmentation approach Copy-Paste [22] copies connected semantic regions from one image to another. Compared to CutMix [68], such regions correspond to object instances instead of squares.

We revisit the quantitative results from Fig. 5 of the main paper, and report the resulting accuracies in Tab. 5. Specifically, we consider the iSAID dataset with a resampling ratio of $R = 1$, and apply an FPN backbone. While all augmentation techniques enhance the performance, our approach yields the most significant quantitative improvements.

D. Additional qualitative

For a complete picture, we provide several additional qualitative samples of different settings. For once, we visualize the joint denoising process proposed in our approach in Fig. 8. We further show visualizations of generated samples on OpenEarthMap [64] in Fig. 9, analogous to Fig. 3 for iSAID and Fig. 4 for LoveDA in the main paper. The semantic labels of OpenEarthMap are associated with land-cover classes, comparable to LoveDA.

In Fig. 11 and Fig. 12, we visualize the predicted semantic masks for iSAID and OpenEarthMap, respectively. Compared to the two baselines SemGAN [32] and SegDiff [1], our approach yields the most consistent results – both in terms of accuracy and mask quality.

Finally, we provide 49 random samples from LoveDA [60] in Fig. 13 for detailed insights into the obtained samples quality of our generative approach.

	mIoU (↑)	F1 (↑)	*per class IoU* (↑)															
			BG	S	ST	BD	TC	BC	GTF	B	LV	SV	HC	SP	R	SBF	P	H
PFSegNet + $\mathcal{D}$	60.93	74.10	98.84	61.74	63.21	76.33	84.25	48.98	54.13	34.95	61.66	41.60	26.52	**50.43**	67.95	66.11	81.41	56.80
PFSegNet + $\mathcal{D} \cup \mathcal{D}'$	**63.71**	**76.37**	**98.93**	**63.88**	**67.58**	**77.11**	**87.70**	**57.96**	**58.01**	**40.22**	**62.91**	**44.70**	**27.63**	50.42	**70.28**	**69.59**	**82.75**	**59.71**
FarSeg + $\mathcal{D}$	62.28	75.16	98.84	**62.45**	**68.63**	**76.76**	86.15	57.14	54.39	**38.95**	**61.49**	**41.08**	27.50	**45.55**	71.53	70.45	81.09	54.53
FarSeg + $\mathcal{D} \cup \mathcal{D}'$	**62.95**	**75.72**	**98.87**	62.33	68.16	74.83	**86.60**	**57.73**	**57.22**	38.56	61.35	40.20	**30.10**	45.53	**74.23**	**72.69**	**81.69**	**57.09**
SegFormer + $\mathcal{D}$	60.95	74.18	98.83	61.91	63.58	**74.35**	84.72	54.01	57.74	**40.37**	58.20	34.32	**32.48**	42.27	68.25	72.67	78.53	52.83
SegFormer + $\mathcal{D} \cup \mathcal{D}'$	**62.13**	**75.10**	**98.88**	**64.17**	**64.86**	74.22	**85.86**	**58.76**	**58.01**	40.19	**59.53**	**35.93**	29.69	**46.20**	**69.92**	**72.73**	**79.89**	**55.21**
FPN + $\mathcal{D}$	59.52	72.82	98.78	58.66	63.72	76.39	84.97	55.32	58.05	36.15	56.82	34.40	**23.82**	44.52	63.60	70.45	76.57	50.14
FPN + $\mathcal{D} \cup \mathcal{D}'$	**60.65**	**73.69**	**98.82**	**59.17**	**64.87**	**76.53**	**85.97**	**55.35**	**58.28**	**36.71**	**57.96**	**34.54**	23.64	**46.66**	**69.79**	**71.34**	**77.70**	**53.05**
PSPNet + $\mathcal{D}$	48.95	63.13	98.35	46.33	46.43	68.27	77.92	38.63	46.76	21.45	48.28	18.03	22.50	35.96	55.80	55.24	66.36	36.88
PSPNet + $\mathcal{D} \cup \mathcal{D}'$	**56.54**	**70.16**	**98.60**	**51.12**	**59.33**	**73.58**	**84.15**	**52.01**	**56.20**	**32.44**	**52.01**	**24.45**	**26.40**	**42.87**	**67.08**	**70.50**	**70.14**	**43.69**

Table 6. **Per-class segmentation scores on iSAID** 256×256**.** We provide a detailed analysis of the per-class segmentation scores on iSAID. Specifically, we report mean IoU scores for approaches tailored for high-resolution satellite imagery [34, 73] and the general-purpose segmentation models SegFormer [65], FPN [31], and PSPNet [71]. Each model is trained on the combined dataset of original and generated samples $\mathcal{D} \cup \mathcal{D}'$, and compared against models trained solely on the original data $\mathcal{D}$. For a majority of classes, the synthesized data yields marked improvements in performance. We abbreviate the 16 semantic classes with the following acronyms: background (BG), ship (S), store tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground track field (GTF), bridge (B), large vehicle (LV), small vehicle (SV), helicopter (HC), swimming pool (SP), roundabout (R), soccer ball field (SBF), plane (P), harbour (H).

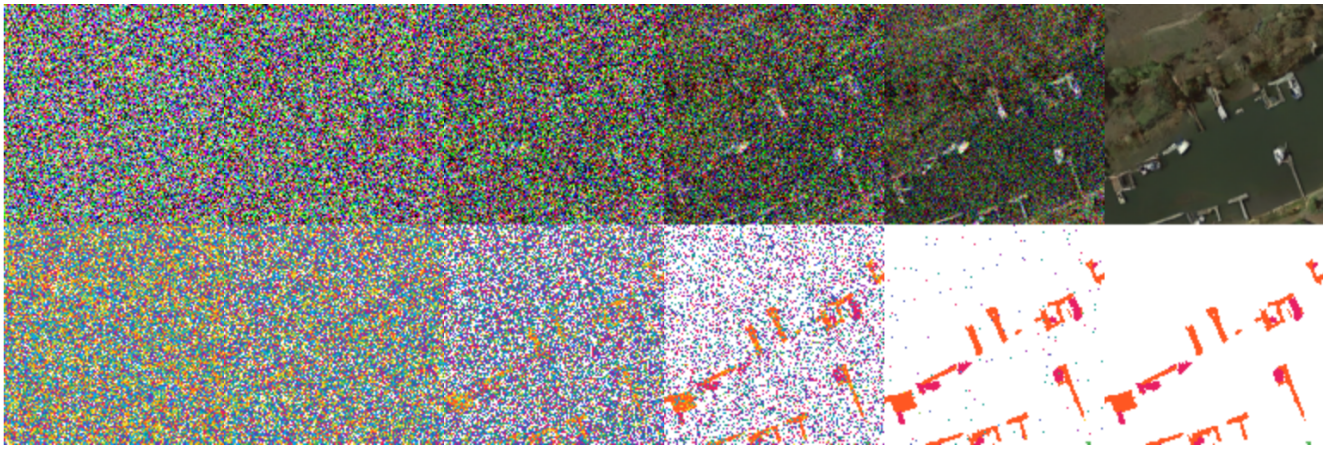

Figure 8. **Denoising process, qualitative.** We provide a qualitative example of the coupled denoising proposed in our approach. Similar to DDPM [27], the novel training samples $(\mathbf{x}'_i, \mathbf{y}'_i)$ emerge through an iterative scheme, reversing the forward Gaussian noising steps.

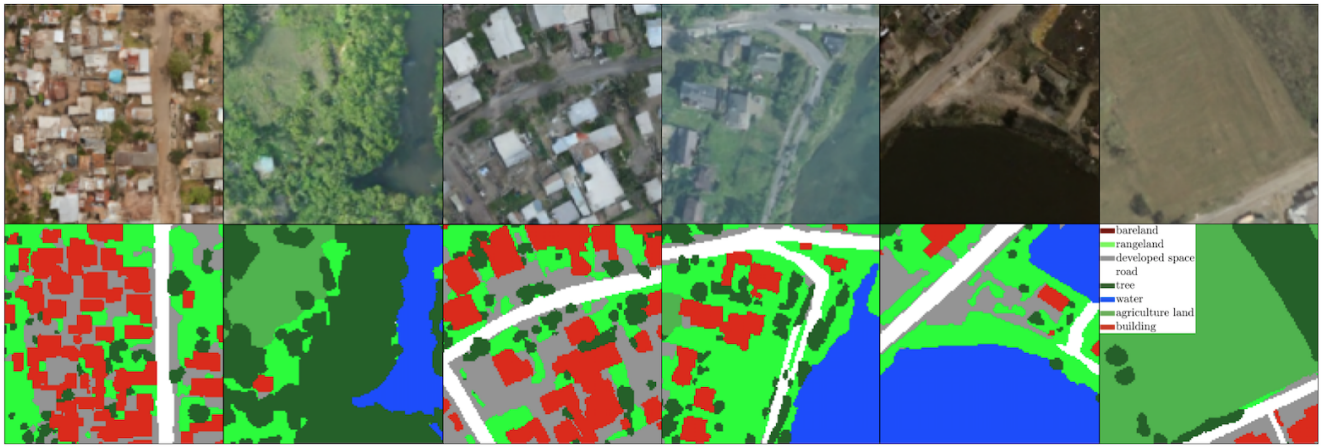

Figure 9. **Generated samples, OpenEarthMap [64].** We display several generated joint instances $(\mathbf{x}'_i, \mathbf{y}'_i)$ on OpenEarthMap [64], obtained by the diffusion model $\mathcal{G}$ detailed in Sec. 4.3 of the main paper.

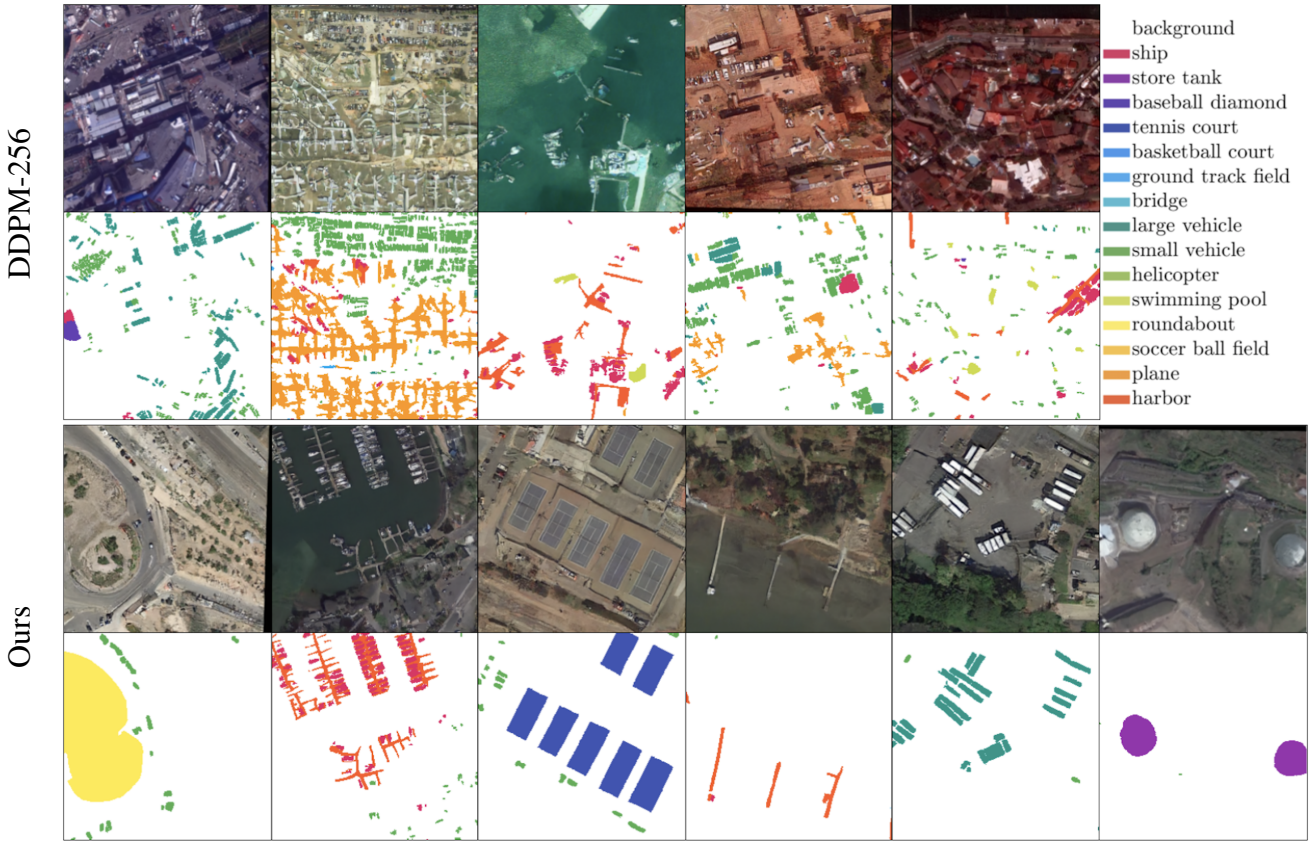

Figure 10. **Super-resolution comparison.** We provide a qualitative comparison of image super-resolution to standard DDPM generations. At resolutions ≥ 256, DDPM exhibits unstable training behaviour, leading to severe artifacts – both in terms of the saturation and contrast of obtained samples, as well as the overall semantic layout. In contrast, our super-resolution approach, outlined in Sec. 4.4 of the paper, generates coherent and high-quality scenes (lower row).

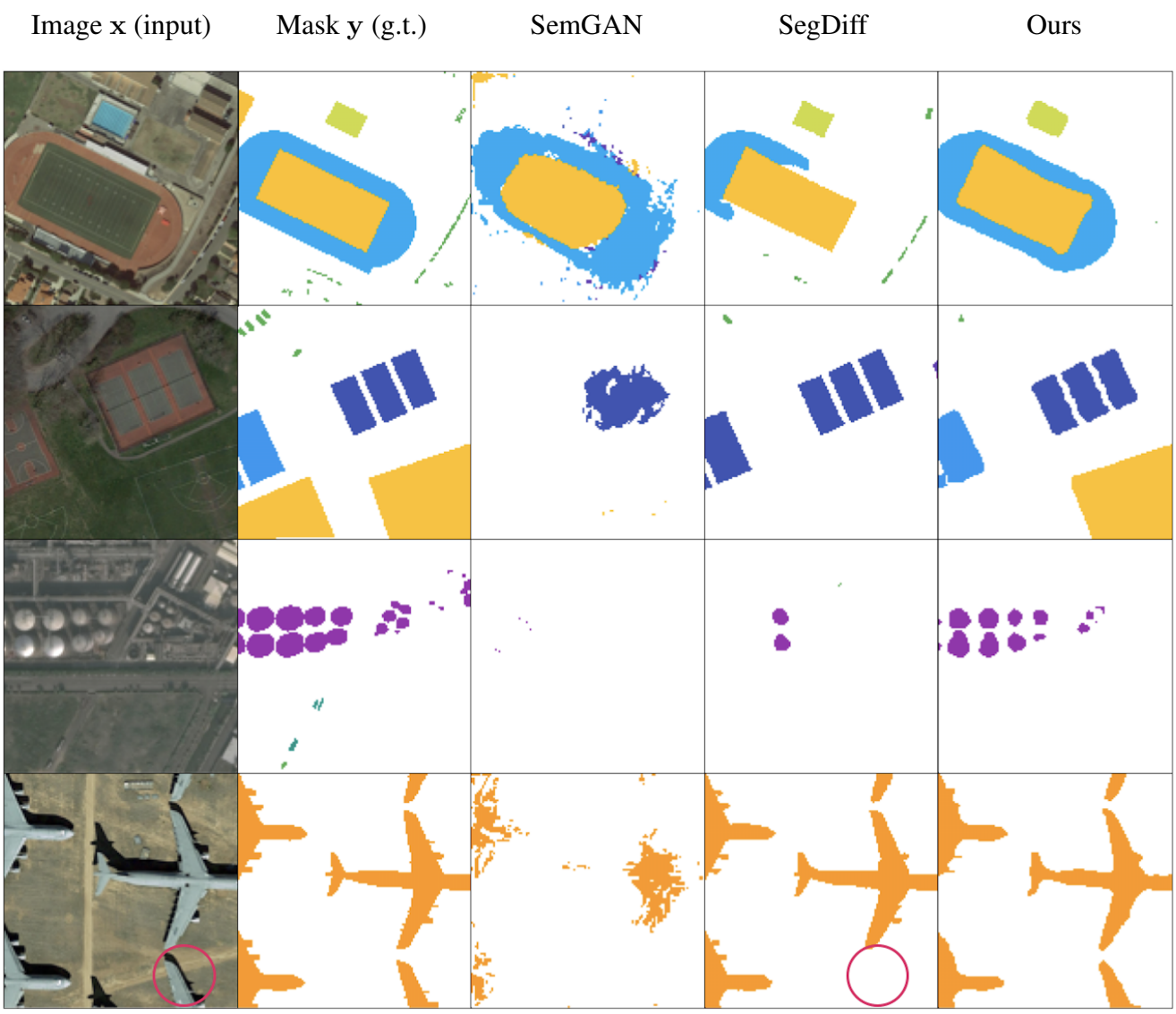

Figure 11. **iSAID baseline comparison.** We contrast the semantic masks obtained with our approach to our two considered baselines [1, 32]. These correspond to the results presented in Tab. 1b in the main paper. SemGAN is primarily designed for conventional segmentation benchmarks such as CelebA [38], whereas the generalization to imbalanced earth observation datasets is limited. Like ours, SegDiff yields high quality masks but individual regions are mislabeled more frequently (*e.g.* red marker), as indicated by the quantitative results in Tab. 1b.

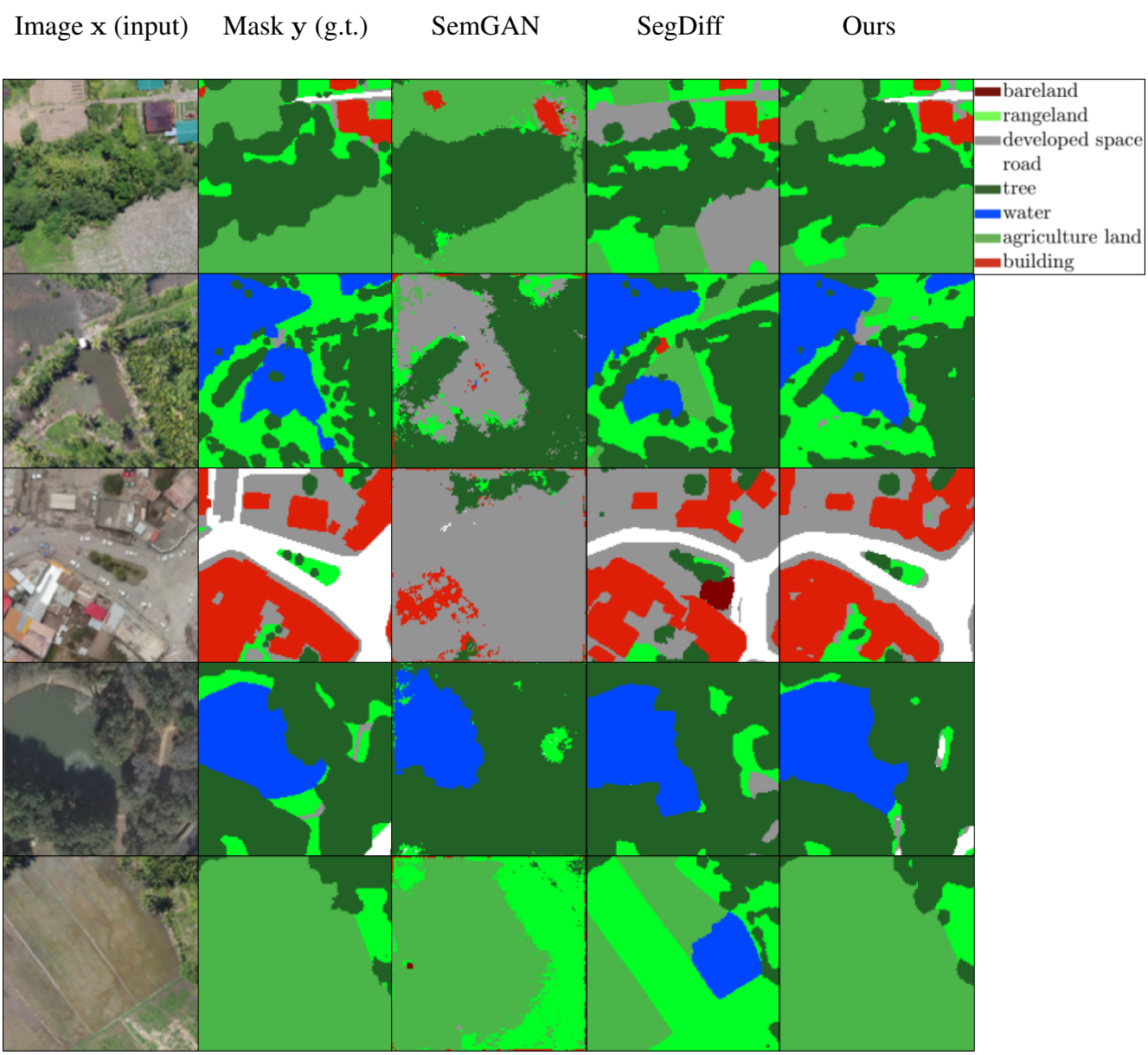

Figure 12. **OpenEarthMap baseline comparison.** Analogously to Fig. 11, we show a number of qualitative comparisons of our approach to our considered baselines [1, 32] on OpenEarthMap.

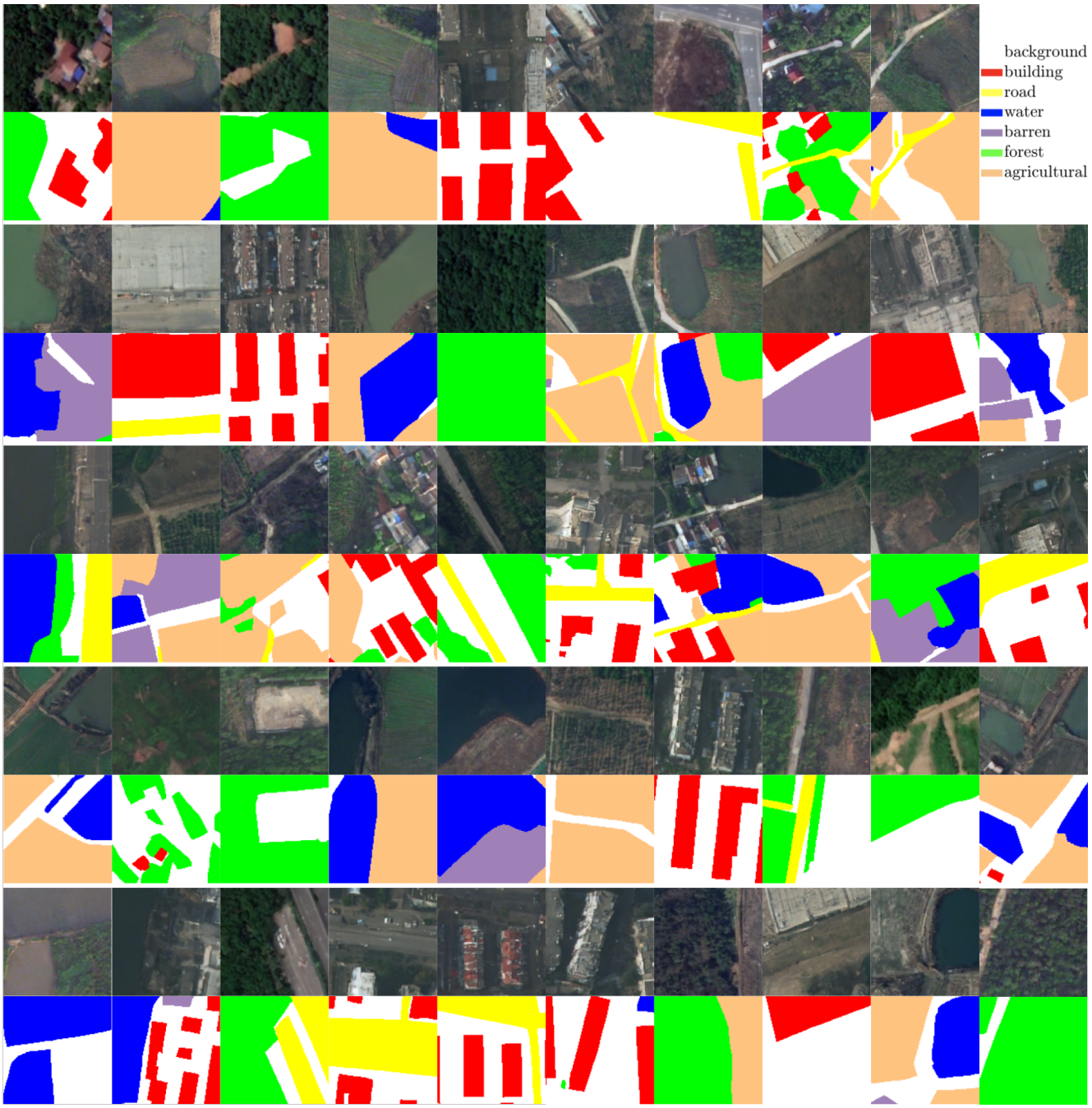

Figure 13. **LoveDA, qualitative.** We provide 49 random samples generated on LoveDA [60], for an in-depth understanding of the quality of obtained samples. As usual, we show pairs of synthesized images $\mathbf{x}$ and corresponding synthesized masks $\mathbf{y}$.