Supplementary Materials: Unsupervised Semantic Segmentation Through Depth-Guided Feature Correlation and Sampling

A. Training Details

A.1. General Hyperparameters

We provide the hyperparameters used to train our models. While all models share some common parameters, there are many that can vary for each dataset. The hyperparameters in Table 1 are identical for all models:

Component	Value
Training Size	224×224
Test Size	320×320
Learning Rate	$5e^{-4}$

Table 1. **General hyperparameters.**

A.2. Dataset-specific Hyperparameters

Table 2 shows the collection of hyperparamers that are specific for our results on various datasets and for various model sizes. We note that Potsdam-3 already undergoes a preprocessing where the images are cropped into 200×200 sized crops, following [2]. Further, as detailed in the main text, for Cityscapes the entire feature map is used for learning.

Dataset	COCO-Stuff 27		Cityscapes	Potsdam
Model	ViT-S	ViT-B	ViT-B	ViT-S
↓ Component				
λ_{DepthG}	0.19	0.16	0.09	0.13
λ_{self}	0.58	0.23	0.95	0.61
λ_{knn}	0.36	1.05	1.02	0.34
λ_{random}	0.70	0.24	0.57	0.72
b_{DepthG}	0.03	0.03	0.03	0.14
b_{self}	0.07	0.12	0.39	0.2
b_{knn}	0.02	0.21	0.25	0.09
b_{random}	0.76	0.97	0.26	0.63
Training Steps	7000	7000	7000	7000
Pointwise Sampling	✓	✓	✗	✓
N	9	12	All	11
Decay Step	250	300	400	None
Decay Factor	0.6	0.64	0.8	None
Cropping	Five-Crop	Five-Crop	Five-Crop	None

Table 2. **Dataset-specific hyperparameters.**

B. Further Ablations

B.1. Guidance Variations

To explore the functionality of our guidance mechanism, we present further ablations in Table 3 where we also explore the use of feature maps from the penultimate layer of the monocular depth estimator (MDE) and using image and perspective planes. We further try plugging the ground-truth segmentation maps into the guidance mechanism. We show unsupervised metrics and use the ViT-S/8 config. Our experiments show guidance with depth maps exceeds the placebo effect of using image or perspective planes, and is most effective when utilizing depth maps.

	COCO-Stuff		Potsdam
Guidance	Accuracy	mIoU	Accuracy
STEGO	48.3	24.5	77.0
Depth Map	**56.3**	**25.6**	**80.4**
MDE Features	55.9	25.4	69.3
Image Plane	53.3	22.8	71.1
Perspective Plane	52.1	23.7	67.4
SemSeg Map	52.8	23.2	<u>80.4</u>

Table 3. **Guidance Variations.** We experiment with different ways of guiding our model. In addition to the depth map, we show results for use MDE features, an image and perspective plane, as well as the semantic maps. Our experiments show that depth maps are the most effective guidance modality.

B.2. Number Of Feature Samples

N	6	7	8	9	10	11	12
U. Accuracy	52.6	52.6	54.2	**56.3**	53.4	54.3	54.1
U. mIoU	22.2	23.0	23.8	**25.6**	24.2	24.2	24.0

Table 4. **Different number of sampled features N^2 .**

We ablate varying the number of sampled features N^2 for the ViT-S backbone on COCO-Stuff 27. Table 4 shows the results. For $N = 9$, our method obtains the best result.

Generally, more samples work better than fewer samples. For $N < 8$, our method shows a significant drop in performance. We further find that for the ViT-S model for COCO-Stuff 27, reducing the number of samples during training can lead to a slight gain in performance. There, N is reduced by 1 at every 3000 steps.

B.3. Guidance Scheduling

We evaluate the effect of scheduling the impact of our *Depth-Feature Correlation* loss. As detailed in the main text, with our method, we enable to model to get a head start and learn about the rough structure in the scene, to then shift the focus on learning representation from the images as training progresses. Our experiments in Table 5 confirm this. When disabling guidance scheduling, our model's performance deteriorates.

Guidance Scheduling	✗	✓
Unsupervised Accuracy	49.4	**56.3**
Unsupervised mIoU	18.8	**25.6**

Table 5. **STEGO + Ours with and without guidance scheduling.**

B.4. NYUv2 With Ground Truth Depth

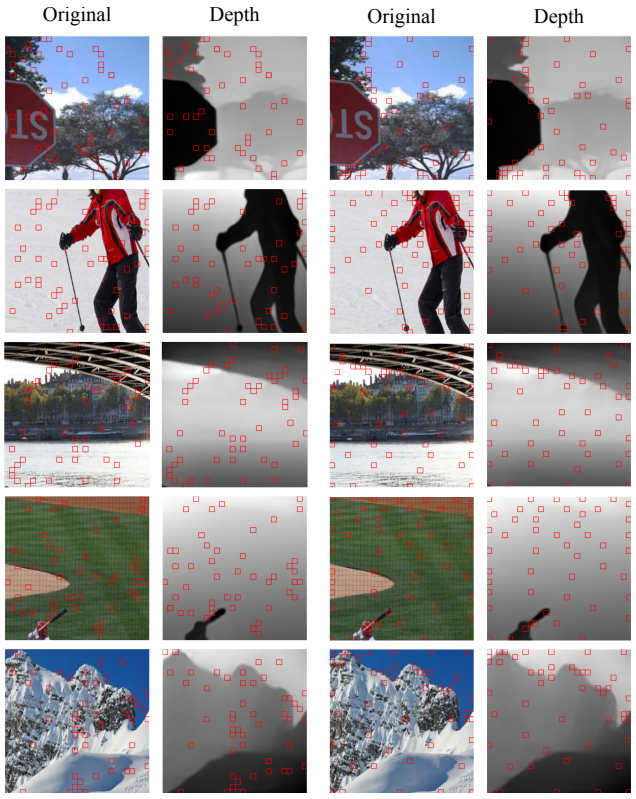
Depth Source	mIoU
ZoeDepth	26.1
Sensor GT	26.2

Table 6. **Results On NYUv2.** We compare using predicted depth to using ground-truth depth

To compare how our method performs with ground-truth depth vs. predicted depth from ZoeDepth [1], we evalute our model on the NYUv2 [4] semantic segmentation dataset. We report results in Table 6 and observe that using predicted depth yields similar results as using ground-truth depth.

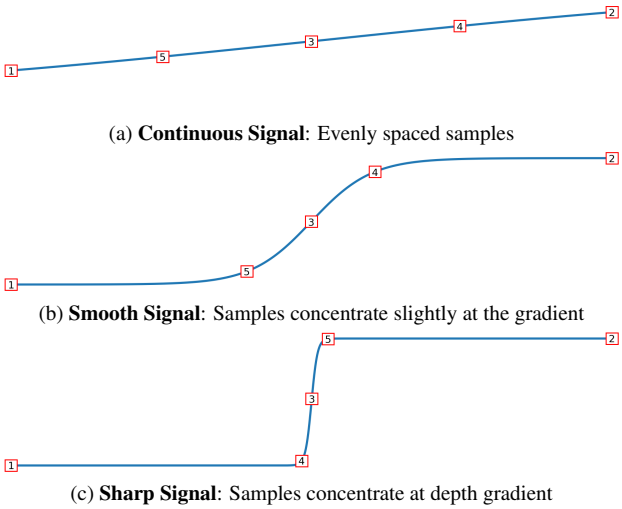
B.5. Farthest Point Sampling Visualizations

To further underline the importance of FPS for our method, and to provide an intuition for how it samples feature, we show additional sampling visualizations. In Figure 3, we display FPS along a sampled axis in the image and depth map. The depth gradient is displayed below in 1D, along with the samples visualized. Figure 2 provides an intuition how FPS selects sampling locations along different gradients. For a continuous signal, FPS samples to space evenly, but the sharper the surface becomes, the more samples are concentrated around the gradient. These 3D



(a) Random Sampling (b) Farthest Point Sampling

Figure 1. **Further examples of Random vs. Farthest Point Sampling.**



(a) **Continuous Signal**: Evenly spaced samples

(b) **Smooth Signal**: Samples concentrate slightly at the gradient

(c) **Sharp Signal**: Samples concentrate at depth gradient

Figure 2. **Visualization of FPS.** We show that samples in FPS converge towards a depth gradient in the signal. The stronger the gradient the more samples are drawn at this region, resulting in meaningful samples for our Depth-Feature Correlation Loss. Numbers denote the order in which the points are sampled.

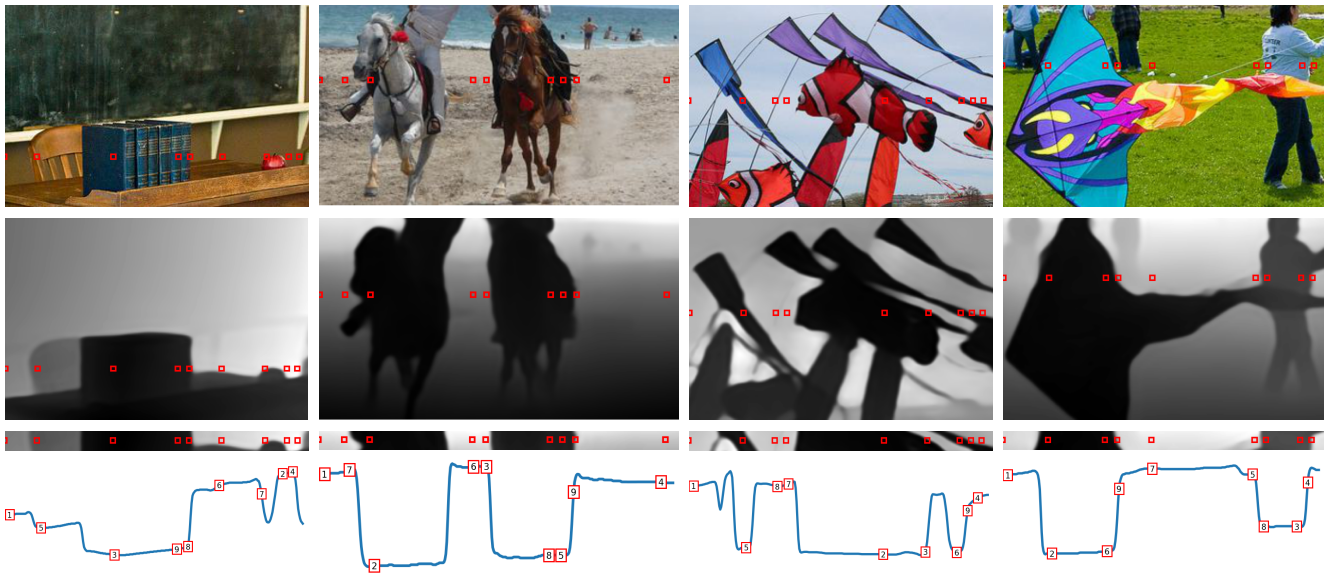


Figure 3. **Visualization of FPS.** We show how FPS samples the depth space on a selection of images. To show the sapling process, we apply FPS along a line in the image to show how it behaves for sampling the depth space. We project the sampled locations along with the depth gradients onto a 1D plot in the bottom row. It can be observed that FPS samples along the edges of objects and encourages depth diversity in the chosen samples.

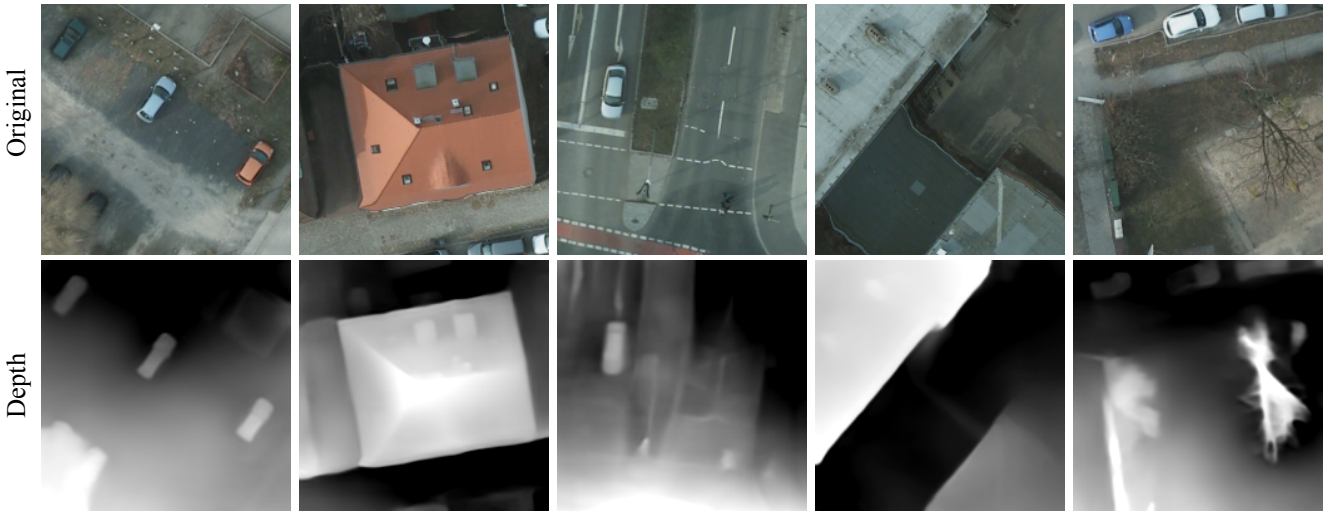


Figure 4. **Predicted Depth for Potsdam dataset.** We use ZoeDepth [1] to predict on the Potsdam aerial images. As can be observed from the visualization, the predictions can overemphasize the foreground and display a bias of predicting depth gradients towards the bottom of the depth map, despite no significant change in depth.

sampling locations are consequently converted to 2D samples and after conversion, they appear around the edges of objects. We also show further random sampling vs. FPS examples in Figure 1.

C. Qualitative Results

C.1. More Results

We show additional qualitative results in Figure 6. All results were generated by ViT-S models, also for competitive methods. Throughout all examples, our depth guidance is effective at enabling our model to segment the scene nicely with more consistent surfaces.

C.2. Comparison to HP

As part of Figure 6, we also add qualitative comparisons to Hidden Positives. While their approach significantly improves the performance of STEGO on the shown examples, our method often produces more consistent segmentations for surfaces. For example, in the top row, Hidden Positives fails to segment the boat at all, while our method produces the correct segmentation.

C.3. Potsdam Depth Predictions

As mentioned in the main text, we show examples of depth precitions from ZoeDepth [1] in Figure 4. While the predictions have sharp boarders around houses, there are a few cases displayed where the model struggles. For example, in the most left column, it produces a depth gradient towards the bottom of the images, despite the entire parking lot having the same depth. In the center column, the part of the road in the top right-hand corner is predicted as further away, while the red cyclist path appears much closer. Further, in the most right column, the model is irritated by the trees.

C.4. Source Of Depth Maps

Figure 5 provides a qualitative comparison of the depth maps produced by ZoeDepth [1], MiDaS [3] and Kick Back & Relax [5]. All maps are Min-Max normalized. Out of all models, ZoeDepth produces the most consistent depth surfaces, even for complex COCO-Stuff scenes such as crowds. The depth maps from MiDaS are similar but lack detail. Kick Back & Relax shows impressive results for a self-supervised method, but fails to capture details such as the right zebra's ears in the center column.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2, 3, 4, 5

[2] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9865–9874, 2019. 1

[3] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 4, 5

[4] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 2

[5] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Kick back & relax: Learning to reconstruct the world by watching slowtv. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 4, 5

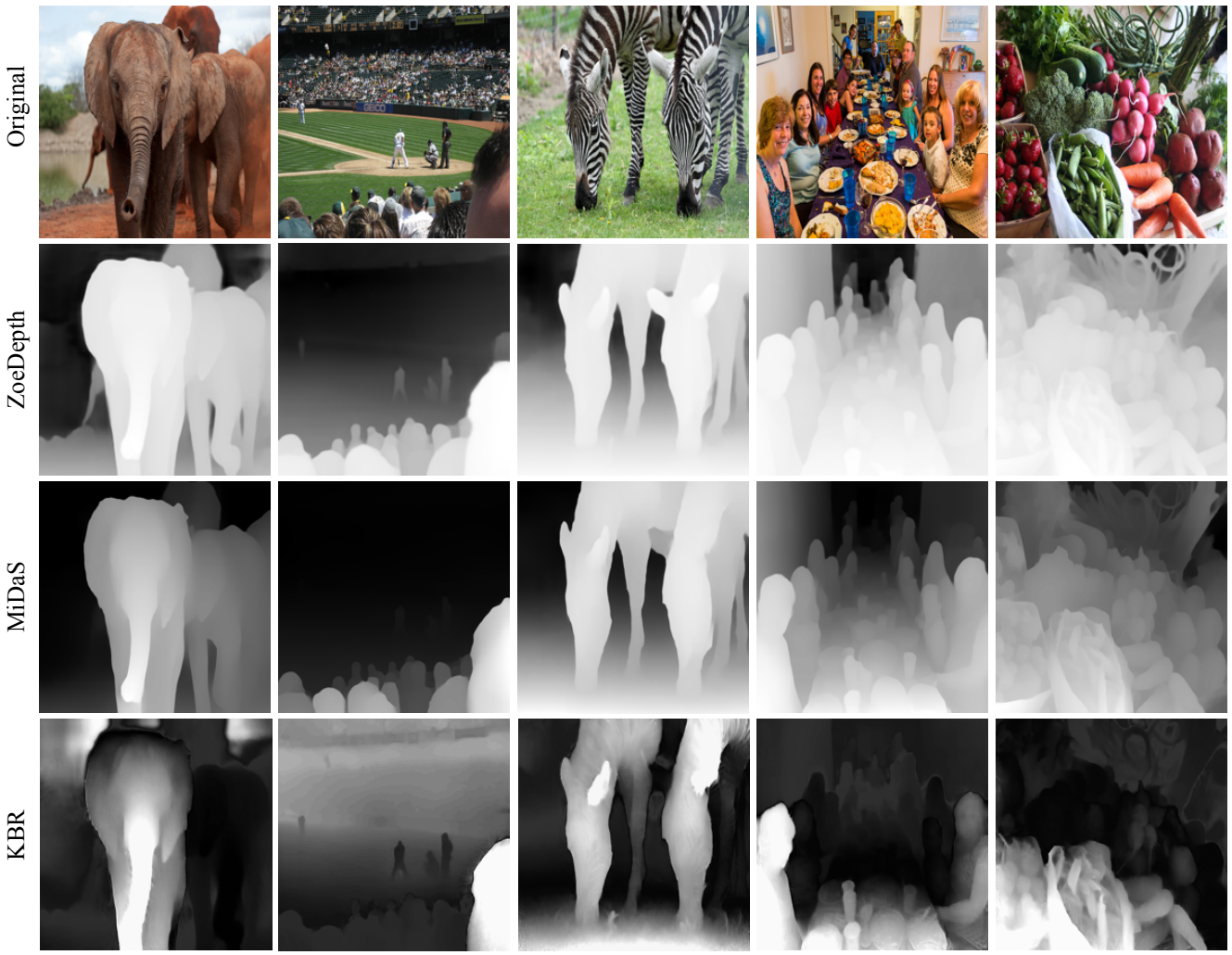


Figure 5. **Comparison of different monocular depth estimators on COCO-Stuff 27.** Our visualizations qualitatively compares the depth maps predicted by ZoeDepth [1], MiDaS [3], and Kick Back & Relax [5].

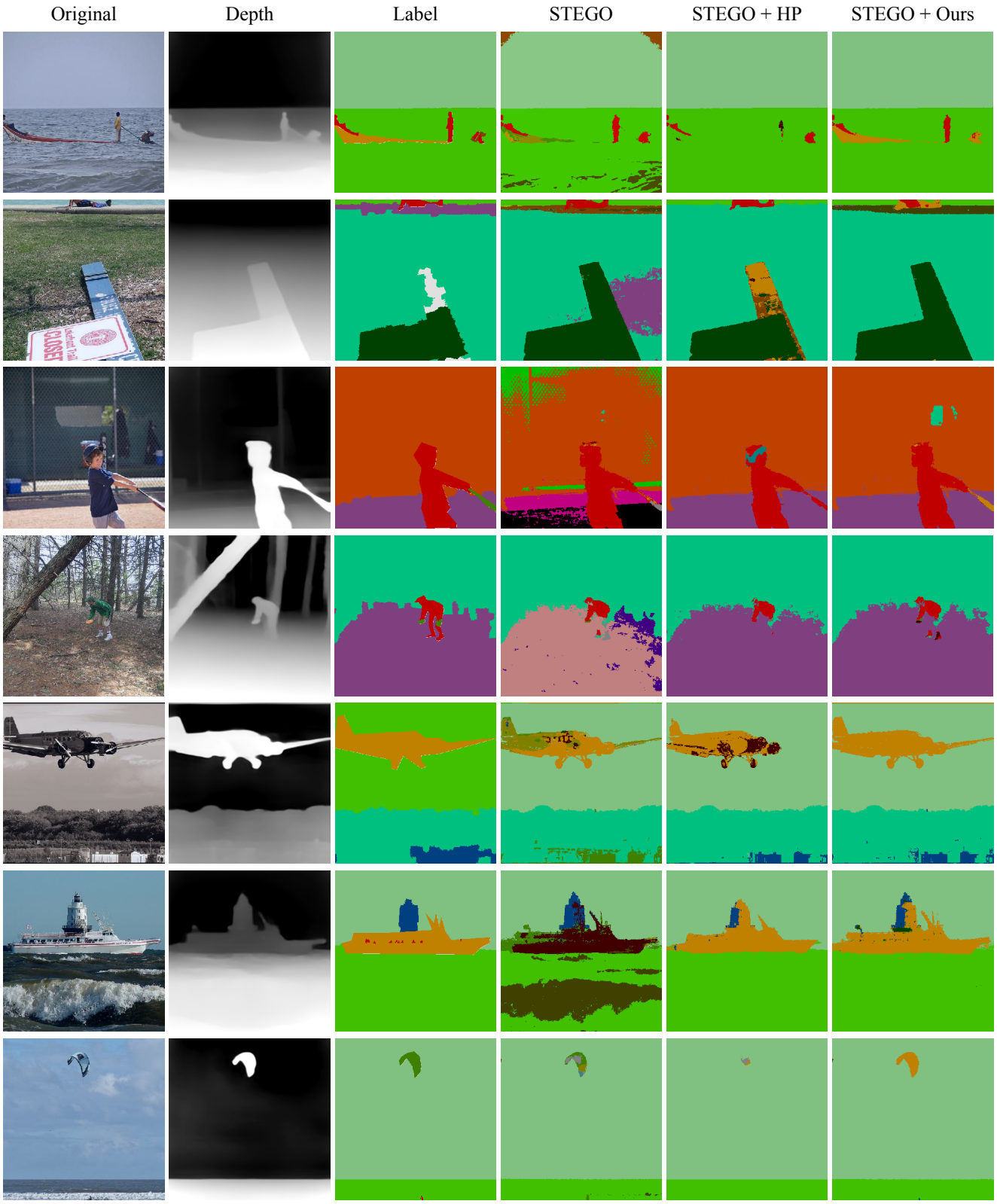


Figure 6. **More Qualitative Results on COCO-Stuff 27.** We show further qualitative results, adding also a comparison to Hidden Positives.