CutS3D: Cutting Semantics in 3D for 2D Unsupervised Instance Segmentation

Supplementary Material

A. CutS3D Qualitative Results

We show more qualitative examples of predictions from our CutS3D detector and compare it to other competitive methods in a zero-shot manner in Figure 1 and Figure 3. Our approach shines for challenging examples with instances that are connected in 2D, such as the person holding the child or the baseball players standing together. The CutS3D CAD is also able to detect more instances, such as the additional zebra or the additional human at the bottom.

B. Further Ablations

B.1. Spatial Importance Lower Bound

We further ablate the effect for lower bound β for our Spatial Importance maps for the performance of our method. For this, we adopt the same evaluation protocol as in the main paper, i.e. we train our model only once on the generated ImageNet [9] pseudo-masks. To isolate the effect of β, we train our model without Spatial Confidence. Table 1 reports the results on COCO val2017 [6] for β variations.

β	0.3	0.45	0.6
AP_{mask}	**8.5**	**8.5**	8.3

Table 1. Different values for β.

SC_{ij}^{min}	0.5	0.67	0.83
AP_{mask}	**9.1**	9.0	9.0

(a) SC Lower Bound.

	With	Without
AP_{mask}	9.1	9.1

(b) SC Maps Mask-Alignment.

Table 2. Further Ablations of our Spatial Confidence Components.

Contribution	AP^{mask}
CutLER	9.4
+ SIS	9.6
+ LocalCut	9.8

(a) Spatial Importance Sharpening

Confidence	AP^{mask}
No Conf.	9.8
CRF	10.0
Depth	10.2

(b) Depth vs CRF Confidence

Table 3. Applying Spatial Importance Sharpening without LocalCut & CRF scores as confidences.

B.2. Spatial Confidence Lower Bound

In Table 2a, we explore setting different values as lower bound for our spatial confidence map. As minimum, we set 0.5 and consecutively add $1/6 \approx 0.17$ since we make cuts at 6 different thresholds. We find that our approach yields the best results at $SC_{ij}^{min} = 0.5$.

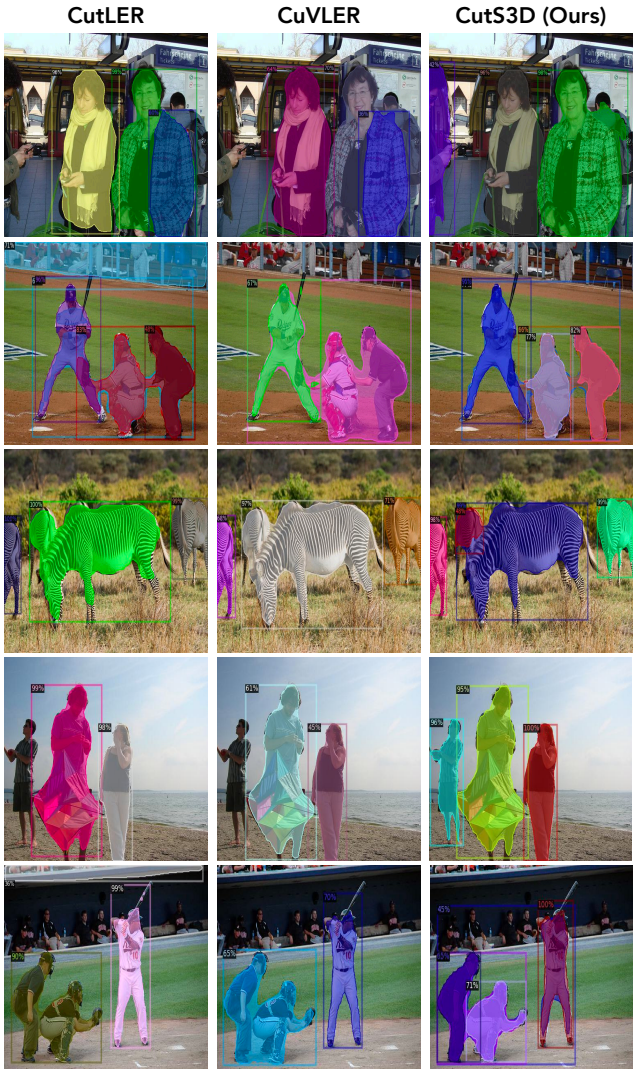

Figure 1. **More Qualitative Results.** We show COCO val2017 [6] predictions of our CutS3D zero-shot model and compare to zero-shot competitors, namely CutLER [11] and CuVLER [1]. Overall, we observe that the CutS3D Cascade Mask R-CNN [3] is able to better differentiate instances that are connected in 2D, e.g. located together in groups. On the other hand, the other two models often fail to separate such instances.

B.3. Spatial Confidence Mask Alignment

In an additional experiment, we investigate the effect of aligning the spatial confidence maps to pixel precision with further refinement instead of patch resolution. Using the two different resolutions for our spatial confidence soft target loss results in two different learning principles: While

	Datasets	AP^{box}_{50}	AP^{box}_{75}	AP^{box}	AP^{box}_{S}	AP^{box}_{M}	AP^{box}_{L}	AR^{box}_{1}	AR^{box}_{10}	AR^{box}_{100}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}	AP^{mask}_{S}	AP^{mask}_{M}	AP^{mask}_{L}	AR^{mask}_{1}	AR^{mask}_{10}	AR^{mask}_{100}
Zero-Shot	COCO	24.3	12.5	13.3	3.7	13.9	31.0	6.9	20.8	30.2	20.8	9.8	10.7	2.3	9.5	27.0	6.1	17.8	24.2
	COCO20K	24.6	12.6	13.4	4.0	14.0	30.7	6.9	20.9	30.4	21.3	9.9	10.9	2.4	9.9	26.7	6.1	17.8	26.7
	LVIS	8.8	4.2	4.8	2.3	8.8	15.7	2.4	9.9	17.9	7.1	3.6	3.9	1.4	5.7	13.8	2.2	8.5	14.6
	VOC	40.8	19.8	21.4	1.2	7.6	33.7	16.5	33.9	41.7	-	-	-	-	-	-	-	-	-
	Objects365	23.6	11.2	12.4	2.7	11.0	22.5	3.2	16.5	30.8	-	-	-	-	-	-	-	-	-
	KITTI	21.1	7.6	9.7	0.1	5.7	22.9	7.0	17.5	25.3	-	-	-	-	-	-	-	-	-
In-Domain	COCO	24.7	12.5	13.3	3.6	13.4	29.8	6.7	21.1	31.5	21.8	10.4	11.4	2.5	10.5	27.5	6.1	19.0	27.0
	COCO20K	25.1	12.5	13.3	3.9	13.6	29.6	6.7	21.3	31.7	22.4	10.7	11.6	2.9	10.8	27.3	6.2	19.1	27.2
	LVIS	9.6	4.6	5.2	2.5	8.9	15.3	2.4	10.4	29.5	8.1	4.1	4.5	1.9	6.5	14.2	2.3	9.5	17.2

Table 4. **Full Results.** We report performance metrics for "Zero-Shot" and "In-Domain" settings for all datasets.

Model	AP^{mask}_{50}	AP^{mask}
ZoeDepth [2]	20.81	10.70
Kick Back & Relax [10]	20.72	10.69

Table 5. **3-Round Self-Training with Different Depth Models.** Our model fully trained with depth from ZoeDepth or Kick Back & Relax.

Method	AP^{box}	AP^{box}_{50}	AP^{box}_{75}
MaskCut† [11]	8.4	15.0	8.0
+ Ours	**11.0**	**21.4**	**10.0**

Table 6. **Pseudo-Mask Evaluation on ImageNet.** We evaluate the generated pseudo-masks on the ImageNet validation split [9] for our baseline, MaskCut, and with our pseudo-mask contributions added (+ Ours). †Results reproduced using the authors' official implementation. Since they do not provide pseudo-mask evaluation code, we use our own implementation only for this.

the coarse spatial confidence map encodes boundary confidence within a region, the refined map specifies exact borders with confidence. The detector either discovers or adjusts mask borders based on confidence. In Table 2b, we find both approaches result in equal performance, rendering further refinement an unnecessary computational.

B.4. Effect of Spatial Importance Sharpening

In our method, we apply spatial importance sharpening (SIS) on the fully-connected feature graph before anything is cut. The goal of SIS is to sharpen the feature similarities of this fully-connected graph in areas where the edges are in the depth map. The intuition behind this is to increase the discriminatory effect of the feature similarities where they are the most important. Only adding SIS improves performance, as shown in Table 3a with a model trained on masks with only SIS applied for the semantic cut. We see LocalCut benefits from SIS on the semantic cut. Adding LocalCut in 3D now adds $+0.2$ (Table 3a) instead $+0.1$ without SIS (as shown in the main paper ablation).

B.5. Confidence from CRF

We experiment with using soft masks from an alternative sources. Table 3b compares confidence from the CRF output vs. Spatial Confidence from depth. We find that using the CRF only slightly improves over not using any confidence. The results underline the value of deriving Spatial Confidence from 3D. A standalone object has high confidence, whereas for a candidate mask in an object group, the optimal 3D cut is less certain.

B.6. Extended Depth Sources Ablation

We extend our depth sources ablation in the main paper of employing a self-supervised depth estimator, i.e. Kick Back & Relax [10] that is trained on videos without any depth data. We now perform 3-round self-training on the model already shown in the main paper, and, in Table 5, present that it also achieves SOTA performance on COCO val2017, quasi-matching the model trained with ZoeDepth.

B.7. Pseudo Mask Evaluation

To train our CutS3D models, we first extract pseudo-masks on the ImageNet [9] training split. Since ImageNet is a dataset that is mainly used for classification tasks, it lacks precise annotations for instance masks. Nevertheless, it comes with bounding box annotations, but those are constrained to one box per image. In an attempt to capture the abilities of our approach to extract a useful instance signal on ImageNet, we evaluate the our pseudo-mask extraction pipeline on the ImageNet validation split and report unsupervised object detection results in Table 6. To produce the numbers for our baseline, we use the official author implementation for CutLER [11]for their pseudo-mask process called MaskCut. Both approaches use DiffNCuts [7] for feature extraction. As can be observed, our method scores higher across several metrics. This pseudo-mask advantage is also reflected in our presented results in the paper, where our trained CAD outperforms the baseline, CutLER [11], with fewer self-training iterations.

C. Full Results

In addition to our results presented in the main paper, we report all metrics for the evaluated datasets in Table 4. This also includes instance-size specific and recall metrics.

D. Further Visualizations

D.1. Pseudo-Mask Failure Cases

While many of our generated pseudo-masks provide a reasonable segmentation of the instances in the scene, in some cases the predicted masks can be faulty or imprecise. Common cases are when objects are positioned next to each other with no discernible 3D boundary or simply when the initial semantic cut fails to find an instance. We therefore show examples of failure cases in Figure 2.

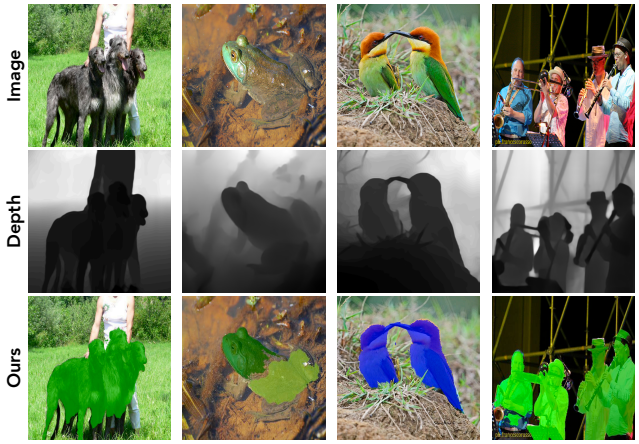

Figure 2. **Pseudo-Mask Failure Cases.** Our CutS3D pseudo-mask approach can struggle for objects with no discernable 3D boundary, such as the two birds sitting next to each other.

D.2. Depth Map Comparison

Our ablations in the main paper show that all evaluated zero-shot monocular depth estimators are suitable for our approach. Therefore, as part of Figure 4, we show examples of predicted depth maps for all three models, namely ZoeDepth [2], Marigold [5], and Kick Back & Relax [10]. Similar to the quantative evaluation, we observe that the depth maps from all three models are of similar high quality across a variety of scenes.

D.3. Spatial Importance Maps

As the contribution ablation reveals, sharpening the semantic affinity graph with Spatial Importance maps greatly improves the performance of our method. Therefore, we show further examples of Spatial Importance maps as part of Figure 5. As can be observed, our Spatial Importance maps extract areas of high-frequency depth changes from the depth maps across various scenes.

E. Method Details

E.1. Pseudo-Mask Extraction

We detail our hyperparameters for pseudo-mask and Spatial Confidence map extraction in Table 7. We perform 3 iterations to identify instances. To extract Spatial Confidence, we linearly sample 6 variations of τ_{knn}. For our main results, we extract depth from ZoeDepth [2] and features from DINO [4] or DiffNCuts [7].

Parameter	Value
N	3
τ_{NCut}	0.13
τ_{knn}	0.115
$\tau_{\text{knn}}^{\text{min}}$	0.05
β	0.45
T	6
Depth Model	ZoeDepth [2]
Backbones	ViT-B/8 (DINO) ViT-S/8 (DiffNCuts)

Table 7. **Pseudo-Mask Extraction Hyperparameters.** We report the hyperparameters used for our LocalCut, Spatial Importance and Spatial Confidence processes.

E.2. Initial Pseudo-Mask Training

We report the hyperparameters used for the initial training of the Cascade Mask R-CNN on the pseudo-masks generated from ImageNet in Table 8. For training the model, we largely follow the standard settings from CutLER [11] and train for 160K iterations. Due to additional memory needs from the Spatial Confidence maps, we reduce our batch size to 4. For our ablations without spatial confidence, we increase it to 8. Like CutLER [11], we scale the copy-pasted masks between 0.3 and 1.0 to vary the resulting size of copied instances. We also initialize the model backbone from DINO [4] weights.

Component	Value
Detector	Cascade Mask R-CNN [3]
Batch Size	4
Base Learning Rate	$1e^{-2}$
Optimizer	SGD
Momentum	0.9
Weight Init.	DINO [4]
Warmup Iterations	1K
Total Iterations	160K
Copy-Paste Min. Ratio	0.3
Copy-Paste Max. Ratio	1.0

Table 8. **Initial Pseudo-Mask Training Cascade Mask R-CNN Hyperparameters.** We detail hyperparameters used for training the CAD on the generated pseudo-masks.

E.3. Self-Training

We further conduct self-training with the predicted masks from the initially trained Cascade Mask R-CNN and report our hyperparameters in Table 9. Since the CAD trained on the initial pseudo-masks cannot predict Spatial Confidence maps for self-training, we no longer have additional memory needs and hence increase the batch size to 8. Further, we find the model converges after 80K iterations, partly due to its weights being initialized from the previously trained CAD. We further increase the minimum scale for copy-paste augmentation to 0.5. Different from CutLER [11], we only conduct 1 round of self-training, saving computational costs. For further in-domain self-training on COCO, we keep our settings largely the same and mainly reduce the total iterations to 14K since COCO is considerably smaller in size than ImageNet. We will provide configuration files for all our trainings as part of the code release after acceptance.

Component	Value
Detector	Cascade Mask R-CNN [3]
Batch Size	8
Base Learning Rate	$5e^{-3}$
Optimizer	SGD
Momentum	0.9
Weight Init.	Previous Training
Self-Training Rounds	1
Warmup Iterations	1K
Total Iterations	80K
Copy-Paste Min. Ratio	0.5
Copy-Paste Max. Ratio	1.0

Table 9. **IN1K Self-Training Cascade Mask R-CNN Hyperparameters.** We report hyperparameters used for performing self-training of our CAD.

References

[1] Shahaf Arica, Or Rubin, Sapir Gershov, and Shlomi Laufer. Cuvler: Enhanced unsupervised object discoveries through exhaustive self-supervised transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23105–23114, 2024. 1

[2] 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, 6, 7

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019. 1, 3, 4

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[5] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 3, 6

[6] 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. 1

[7] Yanbin Liu and Stephen Gould. Unsupervised dense prediction using differentiable normalized cuts. In *ECCV*, 2024. 2, 3

[8] 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. 6

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1, 2

[10] 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*, pages 15768–15779, 2023. 2, 3, 6

[11] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3124–3134, 2023. 1, 2, 3, 4

Figure 3. **More Qualitative Results.** We show further qualitative results on COCO val2017 from our zero-shot model and compare them to other zero-shot competitors for a fair comparison.

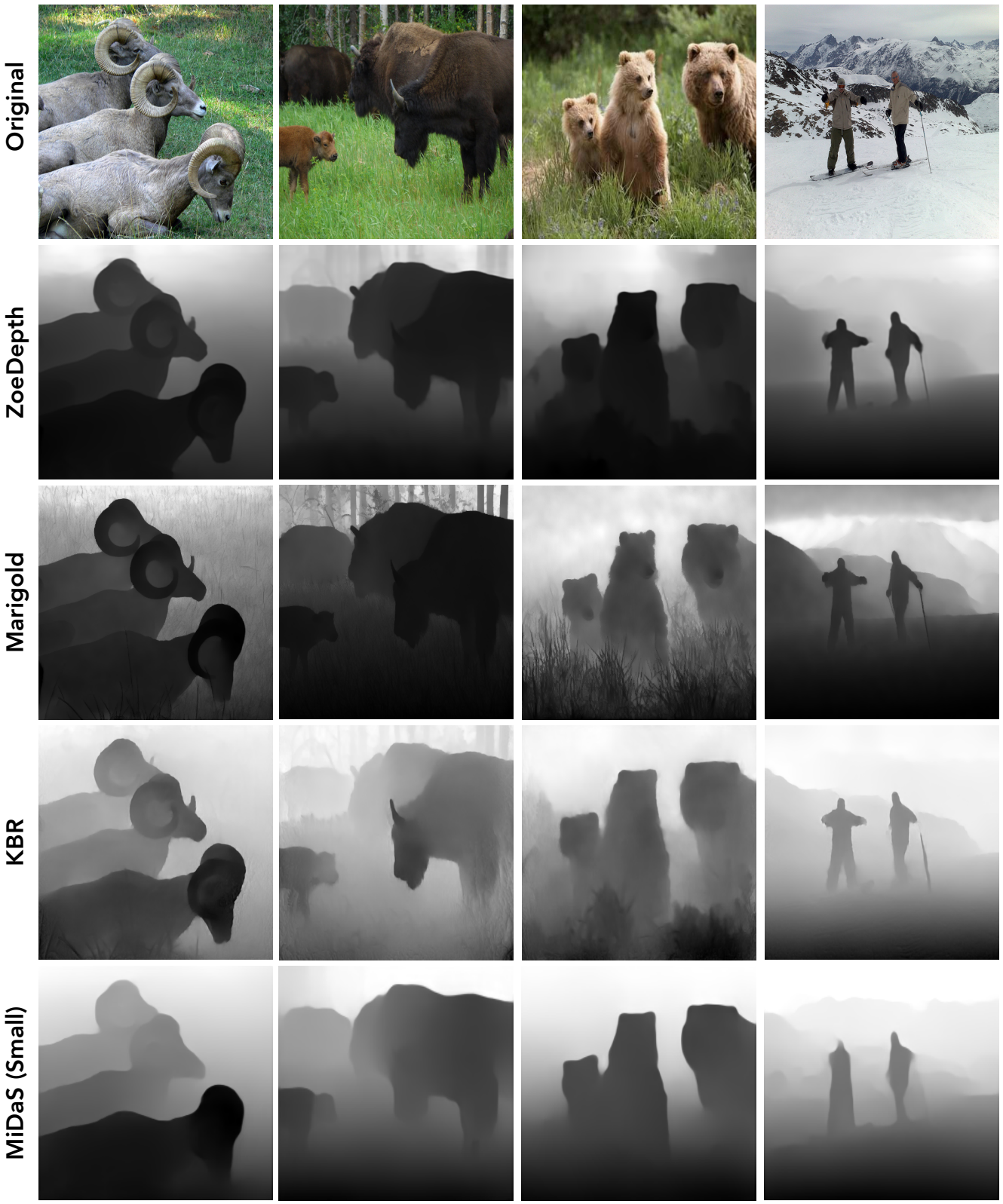

Figure 4. **Comparison of Different Monocular Depth Estimators.** Our visualizations qualitatively compare the depth maps predicted by ZoeDepth [2], Marigold [5], Kick Back & Relax [10] and MiDaS [8].

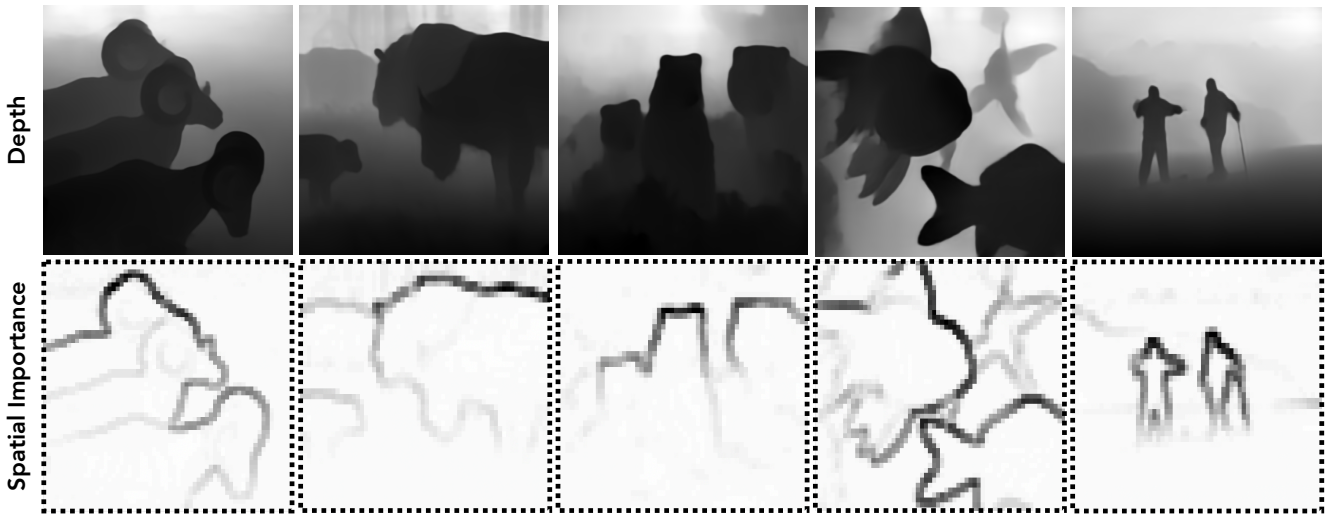

Figure 5. **Spatial Importance Examples.** We show Spatial Importance maps generated from depth maps predicted by ZoeDepth [2].