BiPA: Bilevel Prompt Adaptation for Underwater Instance Segmentation

Supplementary Material

1. Visual comparisons on the UIIS

In this section, we present additional visual comparisons for the task of underwater instance segmentation. We compare results from several representative methods, including PointRend [3], QueryInst [2], Mask2Former [1], ConvNeXt-V2 [8], and PolarNeXt [7], as well as specialized methods WaterMask [5] and USIS-SAM [6]. We further compare these methods on the UIIS [5] dataset to provide more comprehensive qualitative evidence.

In the first two examples of Figure 1, which depict dense reef scenes, generic methods (PointRend, QueryInst, ConvNeXt-V2, PolarNeXt, and Mask2Former) often merge nearby instances or miss small objects, while WaterMask and USIS-SAM produce cleaner yet still incomplete masks; BiPA yields segmentations closest to the labels, with clearer separation of fish and coral. In addition, a similar trend appears in the remaining examples of Figure 1 and the third example of Figure 2, where competing methods either fail to segment distant reef structures or break thin coral branches into fragments, whereas BiPA better preserves these fine details. Overall, across both figures, BiPA produces more complete instances with sharper boundaries and fewer artifacts, demonstrating stronger robustness to clutter, low contrast, and typical underwater degradations than competing approaches.

2. Visual comparisons on the USIS10K

In this section, we provide more qualitative results for the underwater salient instance segmentation task. We include visual comparisons of several generic and task-specific methods, namely ConvNeXt-V2 [8], Mask2Former [1], WaterMask [5], and USIS-SAM [6], evaluated on the USIS10K [6] dataset to offer richer visual evidence.

Figure 3 showcases several challenging cases from USIS10K, including a close-up fish, low-contrast circular targets on a green seabed, a clam with intricate edges, and a small fish in cluttered background. Across these rows, ConvNeXt-V2, Mask2Former, WaterMask, and even USIS-SAM tend to erode thin parts, blur boundaries, or leave fragmented regions, whereas BiPA more tightly adheres to the object contours and preserves complete shapes. Figure 4 presents complementary scenes with single fish, dense coral mounds, submerged trash in turbid water, and large coral fans; here, competing methods frequently under- or over-segment the targets, missing fine branches or introducing spurious pieces, while BiPA maintains coherent masks that are closer to the labels. Overall, these examples highlight BiPA's stronger ability to handle scale variation, low contrast, and structural complexity on USIS10K compared with existing approaches.

3. Generalization Evaluation on USIS10K

As stated in the main paper, we include additional results in this section to further verify the generalization ability of existing methods. Specifically, we focus on two representative task-specific approaches, WaterMask [5] and USIS-SAM [6], and conduct all comparisons on the UIIS10K dataset [4], explicitly evaluating both seen and unseen categories.

Cross-scene generalization within seen categories. Figure 5 and Figure 6 illustrate cross-scene generalization within seen categories on the UIIS10K dataset. For the same categories, appearance and background vary substantially across rows: in several examples of Figure 5, WaterMask either truncates limbs or misses parts of the diver and turtle, while USIS-SAM improves coverage but still leaks into the background or leaves holes around the objects; BiPA, however, follows the silhouettes more faithfully and produces masks closest to the labels. Similar patterns can be observed in Figure 6, where, for instance, starfish, reef fish, and sharks under different viewpoints and lighting conditions cause competing methods to merge instances with the seabed or break fine boundaries, whereas BiPA maintains coherent, well-localized masks. These results indicate that BiPA generalizes more reliably across diverse scenes even when the semantic categories are seen during training.

Table 1. **Quantitative results for zero-shot generalization**.

Train→Test	WaterMask	USIS-SAM	BiPA
UIIS→UIIS10K	42.5	49.2	**50.6**

Zero-shot generalization to unseen categories. Given that the label spaces are incompatible, mAP is inapplicable and quantitative evaluation on seen categories is unreliable. We simplify the problem by treating all unseen-class instances as foreground and reporting the foreground-background mask IoU to quantify zero-shot performance. Table 1 shows results on the unseen UIIS10K categories (Artiodactyla, Mollusk, Garbage; 2,279 images), where BiPA achieves the best IoU.

Figure 7 and Figure 8 evaluate zero-shot generalization to unseen categories on UIIS10K, focusing on Artiodactyla, Mollusk, and Garbage. In Figure 7, which mainly contains Artiodactyla and Mollusk, WaterMask and USIS-SAM often shrink the masks or leak into the background, failing to follow the smooth but complex body and shell contours,

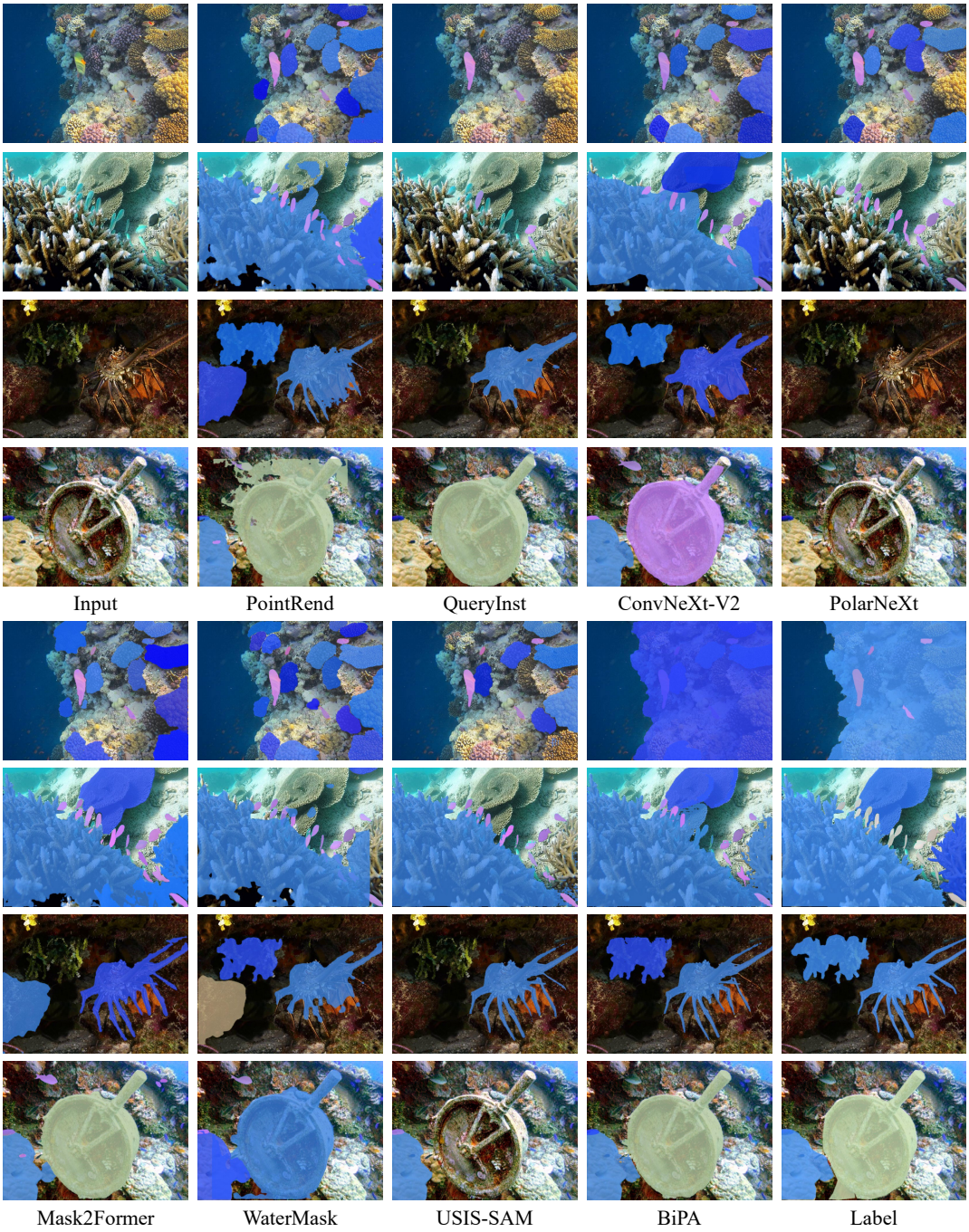

Figure 1. **Visual comparisons on the UIIS dataset [5].**

whereas BiPA produces segmentations much closer to the labels. Figure 8 further shows diverse Garbage scenes (e.g., cans, rings, coins, and mechanical parts) under strong color casts and cluttered seafloors; competing methods frequently fragment objects or confuse them with the background, while BiPA preserves complete shapes and better separates adjacent instances. Together, these results demonstrate that BiPA generalizes more effectively to novel underwater categories not seen during training, even when object appearance and geometry differ significantly from those in the training set. This suggests that BiPA captures more transferable visual cues and is less reliant on category-specific priors than existing methods.

Figure 2. **Visual comparisons on the UIIS dataset [5].**

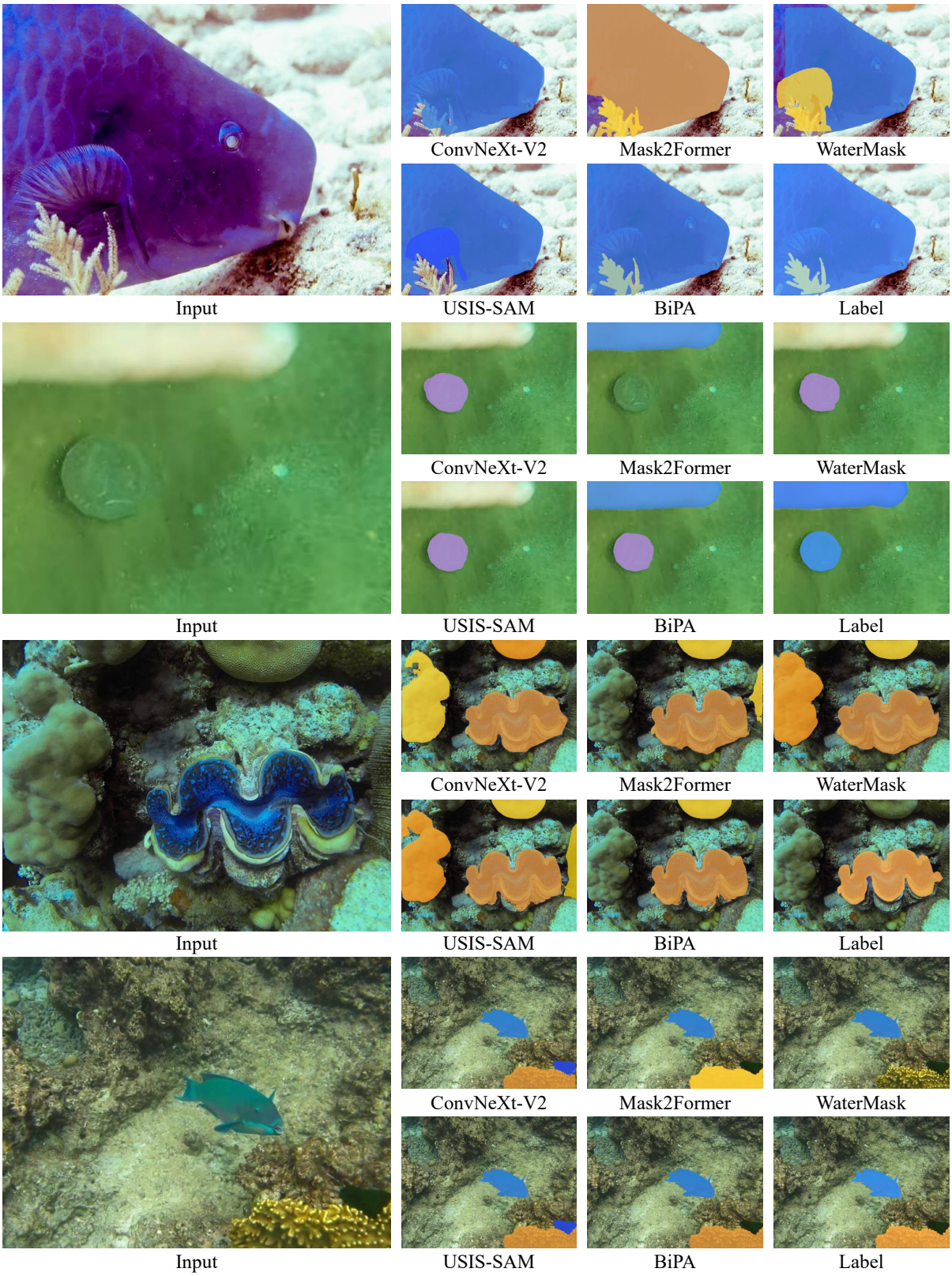

Figure 3. **Visual comparisons on the USIS10K dataset [6].**

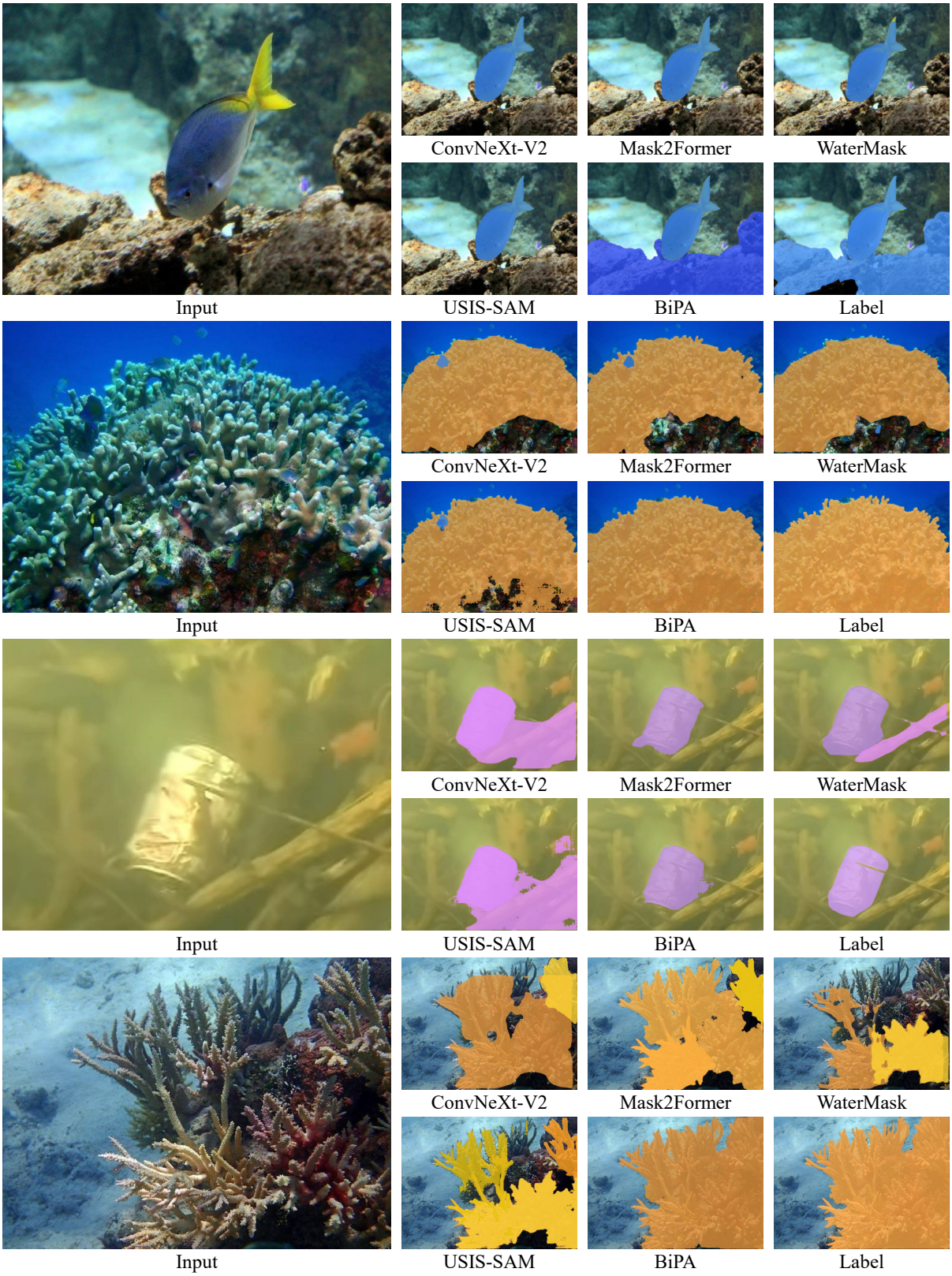

Figure 4. **Visual comparisons on the USIS10K dataset [6].**

Figure 5. **Visual comparisons of *seen categories* on the UIIS10K dataset [4].**

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1

[2] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6910–6919, 2021. 1

[3] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020. 1

[4] Hua Li, Shijie Lian, Zhiyuan Li, Runmin Cong, and Sam Kwong. Advancing marine research: Uwsam framework and uiis10k dataset for precise underwater instance segmentation. *arXiv preprint arXiv:2505.15581*, 2025. 1, 6, 7, 8, 9

[5] Shijie Lian, Hua Li, Runmin Cong, Suqi Li, Wei Zhang, and Sam Kwong. Watermask: Instance segmentation for underwater imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1305–1315, 2023. 1, 2, 3

[6] Shijie Lian, Ziyi Zhang, Hua Li, Wenjie Li, Laurence Tianruo

Figure 6. **Visual comparisons of *seen categories* on the UIIS10K dataset [4].**

Yang, Sam Kwong, and Runmin Cong. Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset. In *International Conference on Machine Learning*, pages 29545–29559, 2024. 1, 4, 5

[7] Jiacheng Sun, Xinghong Zhou, Yiqiang Wu, Bin Zhu, Jiaxuan Lu, Yu Qin, and Xiaomao Li. Polarnext: Rethink instance segmentation with polar representation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19315–19324, 2025. 1

[8] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142, 2023. 1

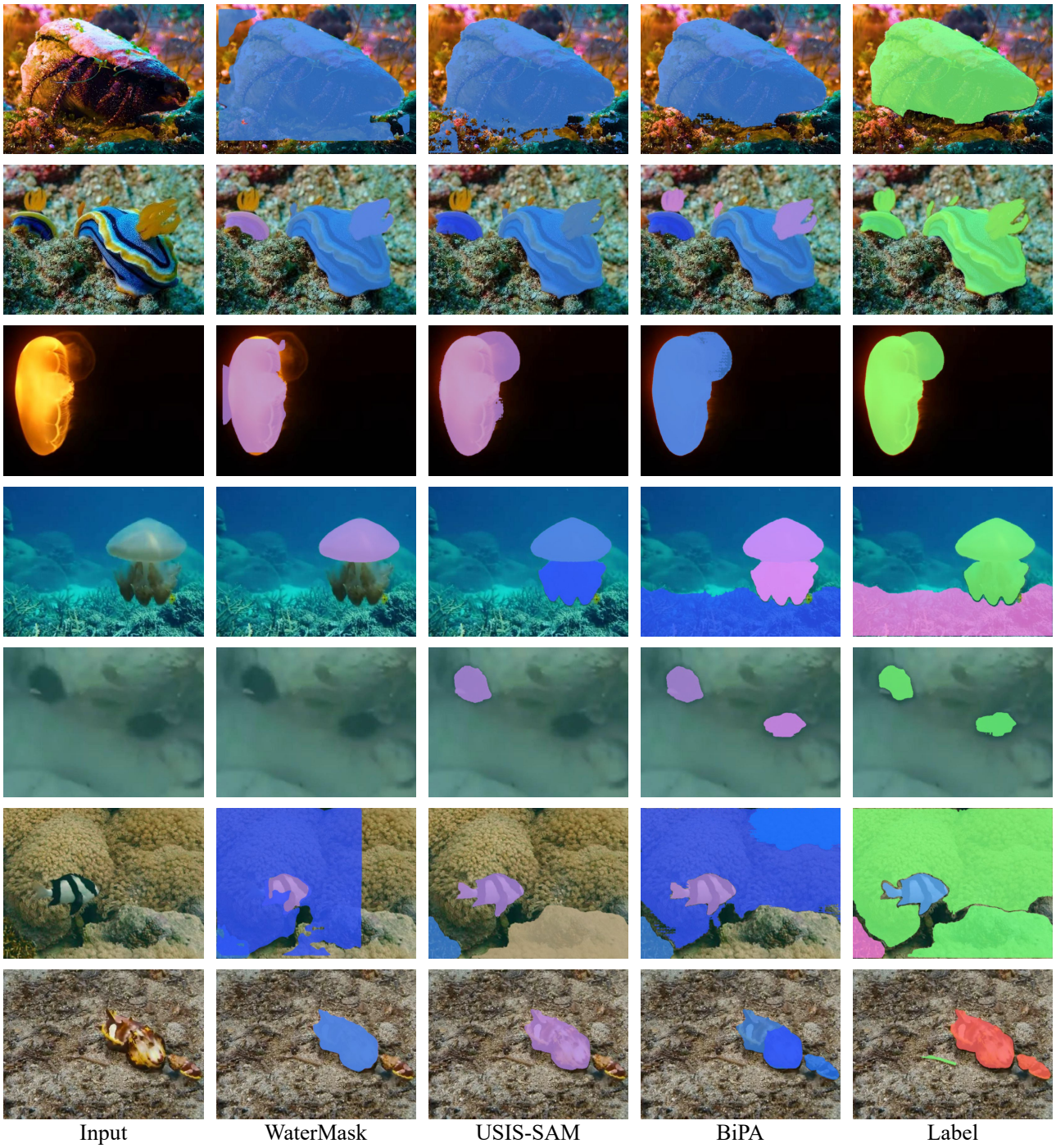

Figure 7. **Visual comparisons of *unseen categories* on the UIIS10K dataset [4].**

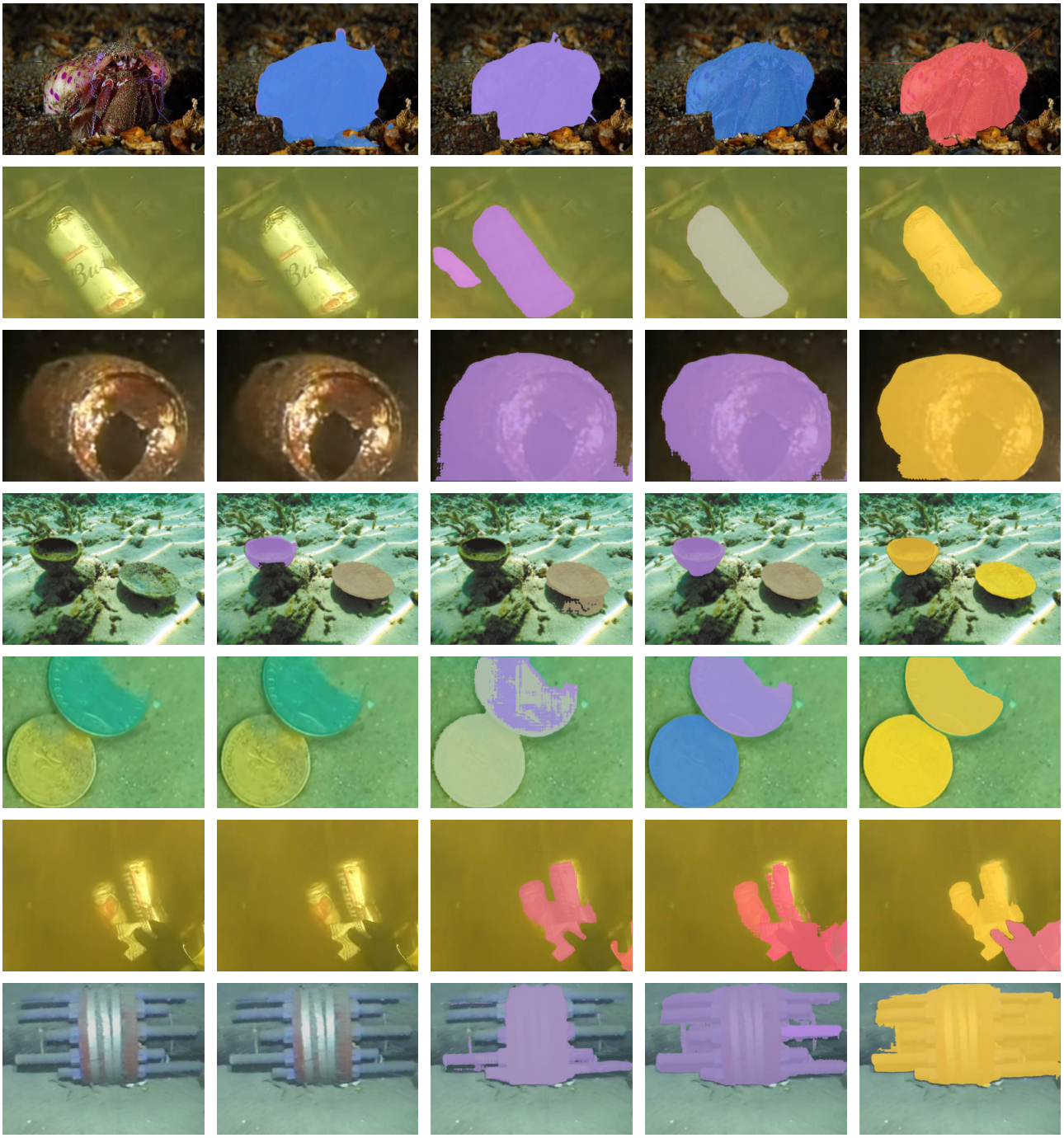

Figure 8. **Visual comparisons of *unseen categories* on the UIIS10K dataset [4].**