Catch Me If You Can Describe Me: Open-Vocabulary Camouflaged Instance Segmentation with Diffusion

Tuan-Anh Vu[1,2], Duc Thanh Nguyen[3], Nhat Minh Chung[2,4], Qing Guo[2], Binh-Son Hua[5], Ivor W. Tsang[2], and Sai-Kit Yeung[1]

[1] The Hong Kong University of Science and Technology, Hong Kong SAR
[2] CFAR & IHPC, A*STAR, Singapore
[3] Deakin University, Australia
[4] Vietnam National University Ho Chi Minh City, Vietnam
[5] Trinity College Dublin, Ireland

Abstract. Text-to-image diffusion techniques have shown exceptional capabilities of producing high-quality, dense visual predictions from open-vocabulary text. This indicates a strong correlation between visual and textual domains in open concepts and that diffusion-based text-image discriminative models can capture richly diverse information for effective segmentation in the wild. However, we found that those advantages are more difficult to hold true for camouflaged individuals because of the significant blending between their visual boundaries and their surroundings. In this paper, while leveraging the benefits of diffusion-based text-to-image models for open-vocabulary performance, we aim to address a challenging problem in computer vision: camouflaged instance segmentation. Specifically, we propose a method built upon a state-of-the-art diffusion model empowered by open-vocabulary to learn multi-scale textual-visual features for camouflaged object representations. Such cross-domain representations are desirable in segmenting camouflaged objects where visual cues are subtly to distinguish the objects from the background, especially in segmenting novel objects not seen in training. We also develop technically supportive components to fuse cross-domain features effectively and engage relevant features towards respective foreground objects. We validate our method and compare it with existing ones on several benchmark datasets of camouflaged instance segmentation and generic open-vocabulary instance segmentation. Experimental results confirm the advances of our method over existing ones.

Keywords: Camouflaged object · text-to-image diffusion · open vocabulary · instance segmentation

1 Introduction

Camouflage is a powerful biological mechanism for avoiding detection and identification. In nature, camouflage tactics are employed to deceive the sensory and cognitive processes of both preys and predators. Wild animals utilise these tactics in various ways, ranging from blending themselves into the surrounding environment to employing disruptive patterns and colouration [19]. Identifying

camouflage is pivotal in many wildlife surveillance applications [8, 25], as it assists in locating hidden individuals for study and protection.

In fact, localisation of camouflaged objects [7, 11], such as Camouflaged Instance Segmentation (CIS), is an important research topic in computer vision, whose challenge lies in the need to learn discriminative features that can be used to discern camouflaged target objects from their surroundings. CIS operates under the conditions where object features closely resemble each other, resulting in class-independent segmentation masks [20]. However, the diversity of camouflages within a single scene can lead to complex intertwining patterns, making the task especially more challenging in severe environmental conditions, *e.g.* terrestrial and aquatic environments, and poor imaging quality, *e.g.* occlusions, image blurriness, and low-light conditions in underwater-applications.

Meanwhile, as humans look at the world and can recognise a limitless number of target categories, open-vocabulary recognition has been developed to mimic human intelligence at unbounded understanding, yet current endeavors have only focused on generic objects and individuals [5, 9, 10, 13, 18, 23, 26]. For example, while Xu *et al.* [23] found that Internet-scale text-to-image diffusion models can be utilised to create a state-of-the-art open-vocabulary segmenter for many concepts, our investigations show that they demonstrate inconsistency and a lack of discernment when it comes to camouflaging effects. Existing works in open-vocabulary segmentation [3, 23, 24, 27, 30, 31] share the similar traits as the ability to detect camouflages are not primary to their designs.

In order to overcome the aforementioned hurdles, we propose a method that leverages text-to-image diffusion to address the problem of open-vocabulary CIS. Our method is inspired by the advanced representation learning ability of diffusion techniques and language-vision transferability of text-image models. Text-to-image diffusion models, *e.g.* the stable diffusion model by [22], are designed to learn essential object features against noise, so they can be useful in extracting features relevant to the target objects in noisy and cluttered backgrounds. While we observed that features learnt solely from the visual domain are weak to distinguish camouflaged objects from their surroundings, the features learnt by text-image discriminative models, *e.g.* CLIP [21], still contain rich information about the real world thanks to the variety of concepts in open-vocabulary training data. We hypothesise that an effective combination of features learnt from both the textual and visual domains would benefit representation learning of camouflaged objects. To the best of our knowledge, such a cross-domain combination with open-vocabulary for CIS is *novel*, and ours is the *first framework* for localizing camouflaged object instances at such a scale.

Our method assimilates an input image and a text prompt about the objects included in the input image, so the input image and its implicit caption (generated by a captioner) are integrated into a text-to-image diffusion model to extract visual features. While our method shares a similar high-level perspective with the works by [23, 28], our proposed textual-visual pipeline aggregates textual and visual features in a mask-out manner to recognise the masks of the target objects. The diffusion model utilises a cross-attention mechanism to link textual features

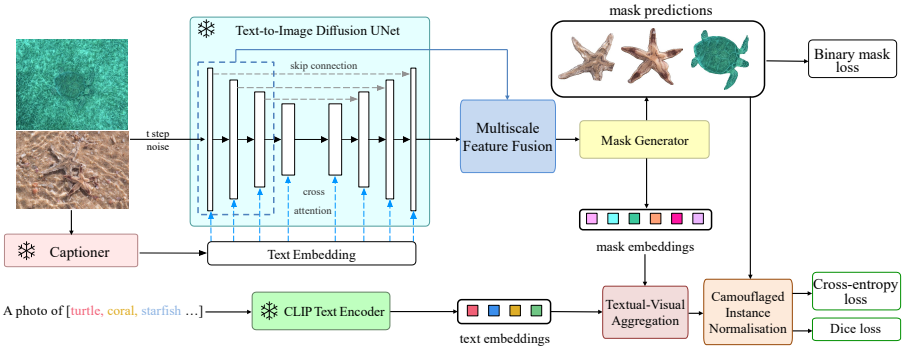

Fig. 1: Pipeline of our proposed method for Camouflaged Instance Segmentation. Inputs include an image and a text prompt of target objects (novel or unseen in the training data). Outputs include instance masks of the target objects. We leverage state-of-the-art text-to-image diffusion and vision-language models to learn textual-visual features that facilitate representation learning for segmenting camouflaged objects.

with visual features and condition the feature learning process, so the learnt features are likely to be distinct and connected to high/mid-level semantic notions that may be expressed in the language part. Our pipeline is more specialised to CIS by designing camouflage-specialised modules. In summary, we make the following contributions to our work.

- We propose a first method for CIS, which is built upon text-to-image diffusion and text-image transfer techniques with open-vocabulary utilization.
- We propose an open-vocabulary-based object representation learning paradigm through a Multi-scale Features Fusion (MSFF), a Textual-Visual Aggregation (TVA), and a Camouflaged Instance Normalisation (CIN) modules.
- We conduct extensive experiments and ablation studies that demonstrate the advantages of our method over existing works.

2 Proposed Method

We aim to build and train an instance segmentation model with a set of pre-defined object categories, referred to as $\mathbf{C}_{\text{train}}$. The instance segmentation model can work on a new domain with $\mathbf{C}_{\text{test}}$ object categories, where $\mathbf{C}_{\text{test}}$ and $\mathbf{C}_{\text{train}}$ may or may not share common object categories. In other words, $\mathbf{C}_{\text{test}}$ may include object categories previously unseen during the training of the instance segmentation model. Throughout the training process, it is presumed that binary mask annotations for target objects in each training image are available. Moreover, each mask is either associated with a category name or a caption presented in the text form. During the testing phase, however, neither the category label nor the caption is accessible for any test image. Only the names of the test categories in $\mathbf{C}_{\text{test}}$ are provided.

2.1 Our Pipeline

Figure 1 illustrates our pipeline that inputs an image and a text prompt about target objects, thereby producing instance masks with matching target categories.

The visual image is first passed to the SD model, which is pre-trained and frozen, to extract latent features. The input image is also fed to the pre-trained and frozen CLIP model to calculate its implicit caption embedding. The caption embedding is inserted into the SD model at various scales (layers) and fused with the SD model's last layer to form image-guided features. These features are "image-guided features" with textual information, as the textual features from the implicit caption embedding are driven by the input image. These features are then combined at different scales by our proposed Multi-scale Features Fusion (MSFF) module, whose outputs are coupled with annotated training masks, and serve as inputs to train a mask generator capable of producing instance masks for all potential categories within the input image. The instance masks are then used to locate object-relevant features in a mask-out manner. This step results in mask embeddings (*i.e.*, features extracted within masked regions).

The textual prompt is processed by CLIP to obtain its corresponding embeddings. These text embeddings are transferable to visual features yet extracted from the textual input, hence considered as "text-guided features".

The textual-visual representations are extracted under an aggregation process of text embeddings (text-guided features) and mask embeddings (image-guided features) using the Textual-Visual Aggregation (TVA) module. It aims to emphasise the learnt features towards foreground objects defined in the input text prompt, resulting in a textual-visual representation of the input image and text prompt. Through our proposed Camouflaged Instance Normalisation (CIN) module, the representations are normalised concerning the instance masks segmented by the mask generator and classified into object categories by a mask classifier.

Open-vocabulary capabilities are facilitated as the entire pipeline is trained with object categories in $\mathbf{C}_{\text{train}}$, while the frozen SD and CLIP models were pre-trained at Internet scales. The training of the entire pipeline is equivalent to learning parameters in modules specialised for camouflage instance segmentation (multi-scale feature fusion, mask generator, textual-visual aggregation, camouflaged instance normalisation). Once the training is completed, the inference process performs open-vocabulary instance segmentation, *i.e.*, instance segmentation of novel object categories in $\mathbf{C}_{\text{test}}$.

2.2 Training

We train the entire pipeline of our method by optimising the loss functions used in the mask generator and the CIN module with supervision. Specifically, we adopt a binary cross-entropy loss as our binary mask loss $\mathcal{L}_{bce}$ and a dice loss $\mathcal{L}_{dice}$ [17] for supervising binary mask predictions in the mask generator to remedy class imbalance.

The training of the CIN module is carried out under the conventional close-vocabulary training approach. Suppose that we can access the ground-truth category label for each object mask during the training phase. For each mask

embedding z_i^{pred} produced by the mask generator, let $y_i^{cate} \in \mathbf{C}_{\text{train}}$ be the corresponding ground-truth category of z_i^{pred}. We invoke the text encoder $\mathcal{T}$ in the pre-trained CLIP model to encode the names of all categories in $\mathbf{C}_{\text{train}}$. This results in a set of text embeddings $\mathcal{T}(\mathbf{C}_{\text{train}}) = \left\{\mathcal{T}(c_1), \ldots, \mathcal{T}\left(c_{|\mathbf{C}_{\text{train}}|}\right)\right\}$ where $c_k \in \mathbf{C}_{\text{train}}$ represents a category name. Thus, the loss for embedding classification (*i.e.*, associating mask embeddings m_i^{pred} with their categories y_i^{cate}) is calculated as:

$$\mathcal{L}_{\text{ce}} = \frac{1}{N}\sum_{i=1}^{N} \text{CE}\left(\text{Softmax}\left(\frac{z_i^{pred}\mathcal{T}(\mathbf{C}_{\text{train}})}{\tau}\right), y_i^{cate}\right), \tag{1}$$

where τ is a learnable temperature parameter and CE is the cross-entropy loss for the classification of each training embedding.

The total loss for the training of our pipeline is finally defined as,

$$\mathcal{L} = \alpha\mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{ce}}, \tag{2}$$

where α is a hyper-parameter, empirically set to 0.4. Furthermore, in line with work done by [1], we apply the Hungarian matching [12] to match predicted masks with ground-truth masks and compute the loss between matching pairs.

3 Experiments

3.1 Datasets

Following previous studies [4, 23, 28, 29], we used the instance segmentation part of the MS-COCO dataset [14] with 80 object categories to pre-train our model. Pre-training our proposed framework on the MS-COCO dataset helps to emphasise the discernment of prompted objects from their backgrounds in the wild, specifically for open-vocabulary segmentation performance. For closed-set validation, fine-tuning the model on the 3,040 images from the training set of the COD10K-v3 dataset [6] further adapts the model to camouflaged objects and significantly boosts up the performance of our method.

We tested our method on two benchmark camouflaged object datasets: the test set of the COD10K-v3 (including 2,026 images) and the NC4K [16] (including 4,121 images). The NC4K dataset contains only test images. The training sets (for both pre-training and fine-tuning) and the test sets (for both the COD10K-v3 and NC4K) share only 6 common object categories (out of 80 and 69 object categories from the MS-COCO and COD10K-v3/NC4K, respectively). This setting, *i.e.*, cross-dataset training-testing, has been used widely in the evaluation of the generalisation ability of CIS models. It reflects the practicality of CIS, thus ensuring the reliability of evaluations.

3.2 Results

We evaluated our method and existing works using the average precision (AP) measured at different intersection-over-union (IOU) thresholds. In particular, we

Table 1: Comparison of our method with existing instance segmentation methods on the test set of the COD10K-v3 and the NC4K datasets. We denote "Ours" and "Ours (task-specific)" for two variants of our method without and with fine-tuning on the training set of the COD10K-v3 dataset. Params (M) denotes the number of *trainable* parameters. The best results are **bold**, and the second best results are underline.

Method		COD10K-v3 Test			NC4K			Params (Milions)
		AP	AP50	AP75	AP	AP50	AP75	
closed-set supervised learning	Mask2Former [2]	39.4	67.7	38.5	45.8	73.6	47.5	43.9
	OSFormer [20]	41.0	71.1	40.8	42.5	72.5	42.3	46.6
	DCNet [15]	**45.3**	70.7	**47.5**	**52.8**	**77.1**	**56.5**	53.4
	Ours (task-specific)	44.9	**70.9**	47.2	52.7	76.6	55.8	28.7
open-vocab text-to-image (*w/o finetuning*)	ODISE [23]	21.1	37.8	20.5	22.9	37.2	21.4	28.1
	X-Decoder [30]	7.7	12.9	7.5	3.9	8.1	3.4	38.3
	Ours	**23.4**	**43.8**	**22.6**	**24.3**	**43.7**	**23.5**	28.7

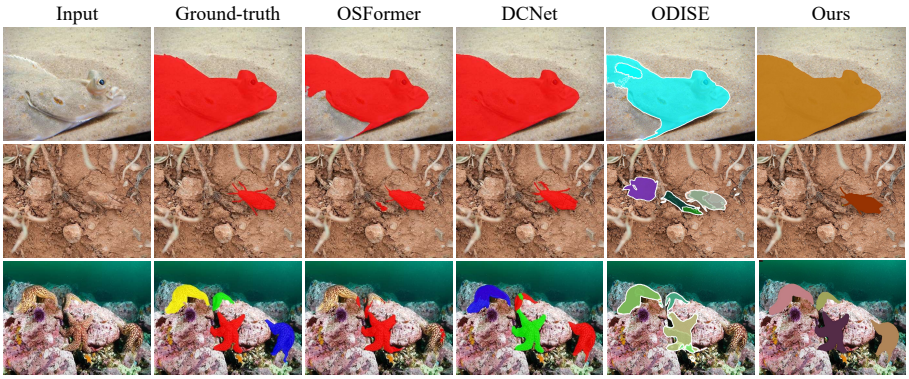

Fig. 2: Qualitative comparison on COD10K-v3 and NC4K. Our method outperforms others and has competitive performance as DCNet [15] at a twice smaller model size.

calculated the overall AP in the range $[50\%, 95\%]$ for the IOU thresholds (*i.e.*, for a threshold within the above range, a predicted instance is considered as true positive if there exists a true instance in the ground-truth such that their IOU is equal or greater than that threshold). We also measured detailed AP for the IOU thresholds of 50% (AP50) and 75% (AP75).

We report the performance of our method on camouflaged object datasets (the COD10K-v3 and NC4K) in Table 1 (last row). Recall that, following the conventional setting in CIS, *e.g.*, [4, 23, 28, 29], we pre-trained our model on the MS-COCO dataset and then fine-tuned it on the training set of the COD10K-v3 dataset. To show the effectiveness of this strategy, we experimented with a variant of our method by skipping the fine-tuning phase. In particular, we pre-trained our method on the MS-COCO dataset and then evaluated it directly on the test set of the COD10K-v3 and the NC4K datasets.

We compare our method with existing instance segmentation methods on the CIS task in Table 1. We group existing methods into two groups: "closed-set

supervised learning approach", which follows the traditional fashion of supervising an instance segmentation model on a training set and tests the model on a test set, and "open-vocab text-to-image approach", which includes methods using text-to-image diffusion techniques with open-vocabulary. Most existing instance segmentation methods lie in the first group. Our method and related works [23,30] belong to the second group.

In Table 1, we show that our method with the pre-training setting significantly outperforms ODISE on all evaluation metrics, making a new state-of-the-art for open-vocabulary CIS. With pre-training and fine-tuning, our method also performs on par with DCNet [15] (best method of the "closed-set supervised learning approach"), while requiring much fewer parameters in Table 1.

In summary, with regard to both the segmentation accuracy and memory usage, our method is more advanced, compared with existing ones. Recall that only 6 object categories are shared between MS-COCO (80 object categories) and COD10K-v3/NC4K (69 object categories). This challenge shows the ability of our method to handle open-vocabulary tasks. We visualise several results of our methods and existing ones in Figure 2, where our method excels at accurately delineating camouflaged objects along their blurry boundaries in cluttered backgrounds at significant proficiency.

4 Conclusion

This work advances the computer vision research for camouflaged instance segmentation by leveraging text-to-image diffusion and text-image transfer techniques. We aim to raise people's awareness about the possible lack of transfer effectiveness in open-vocabulary segmentation regarding camouflages. Furthermore, we propose a method that effectively integrates textual information learnt from open-vocabulary into the visual domain to enrich the representations of camouflaged objects. We evaluate our method and compare it with existing methods in CIS open-vocabulary segmentation benchmark datasets. On the one hand, the method struggles with segmenting occluded objects, and under severe occlusions, a camouflaged object can be over-segmented into non-semantic fragments. Nevertheless, our experimental results show the effectiveness and advantages of our method over existing baselines in both tasks.To the best of our knowledge, our work segments camouflaged object instances in an open-vocabulary manner for the first time. We believe it will open up new avenues for research and developments in surveillance, wildlife monitoring, and military reconnaissance.

Acknowledgement. This research is supported by an internal grant from HKUST (R9429), the National Research Foundation, Singapore, and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-GC-2023-008), Career Development Fund (CDF) of Agency for Science, Technology and Research (A*STAR) (No.: C233312028), National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative (No. DTC-RGC-04), and an MAAP Discovery funding (2022-2025) from Deakin University. This work is partially done during Tuan-Anh Vu's internship at CFAR, A*STAR, Singapore.

References

1. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1280–1289 (2022)
2. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299 (2022)
3. Ding, Z., Wang, J., Tu, Z.: Open-vocabulary universal image segmentation with maskclip. In: Proceedings of the International Conference on Machine Learning (2023)
4. Ding, Z., Wang, J., Tu, Z.: Open-vocabulary universal image segmentation with MaskCLIP. In: Proceedings of the International Conference on Machine Learning. pp. 8090–8102 (2023)
5. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: Learning to prompt for open-vocabulary object detection with vision-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14064–14073 (2022)
6. Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 6024—-6042 (2022)
7. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2777–2787 (2020)
8. Fleming, P.J.S., Meek, P.D., Ballard, G., Banks, P.B., Claridge, A.W., Sanderson, J.G., Swann, D.E.: Camera Trapping: Wildlife Management and Research. CSIRO Publishing (2014). https://doi.org/10.1071/9781486300402
9. Gao, M., Xing, C., Niebles, J.C., Li, J., Xu, R., Liu, W., Xiong, C.: Open vocabulary object detection with pseudo bounding-box labels. In: Proceedings of the European Conference on Computer Vision. pp. 266–282 (2022)
10. Ghiasi, G., Gu, X., Cui, Y., Lin, T.: Scaling open-vocabulary image segmentation with image-level labels. In: Proceedings of the European Conference on Computer Vision. pp. 540–557 (2022)
11. He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y., Guo, Z., Li, X.: Camouflaged object detection with feature decomposition and edge reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22046–22055 (2023)
12. Kuhn, H.W.: The Hungarian Method for the Assignment Problem. Naval Research Logistics Quarterly **2**(1), 83–97 (March 1955)
13. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A.: F-vlm: Open-vocabulary object detection upon frozen vision and language models. In: Proceedings of the International Conference on Learning Representations (2023)
14. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755 (2014)
15. Luo, N., Pan, Y., Sun, R., Zhang, T., Xiong, Z., Wu, F.: Camouflaged instance segmentation via explicit de-camouflaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17918–17927 (2023)
16. Lyu, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., Fan, D.P.: Simultaneously localize, segment and rank the camouflaged objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11586–11596 (2021)

17. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the International Conference on 3D Vision. pp. 565–571 (2016)
18. Minderer, M., Gritsenko, A.A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N.: Simple open-vocabulary object detection with vision transformers. In: Proceedings of the European Conference on Computer Vision. pp. 728–755 (2022)
19. Nguyen, T.T.T., Eichholtzer, A.C., Driscoll, D.A., Semianiw, N.I., Corva, D.M., Kouzani, A.Z., Nguyen, T.T., Nguyen, D.T.: Sawit: A small-sized animal wild image dataset with annotations. Multimedia Tools and Applications pp. 1–26 (2023)
20. Pei, J., Cheng, T., Fan, D.P., Tang, H., Chen, C., Van Gool, L.: Osformer: One-stage camouflaged instance segmentation with transformers. In: Proceedings of the European Conference on Computer Vision. pp. 19–37 (2022)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning. pp. 8748–8763 (2021)
22. Robin, R., Andreas, B., Dominik, L., Patrick, E., Björn, O.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10674–10685 (2022)
23. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., Mello, S.D.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2955–2966 (2023)
24. Xu, X., Xiong, T., Ding, Z., Tu, Z.: Masqclip for open-vocabulary universal image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 887–898 (October 2023)
25. Yan, J., Le, T., Nguyen, K., Tran, M., Do, T., Nguyen, T.V.: Mirrornet: Bio-inspired camouflaged object segmentation. IEEE Access **9**, 43290–43300 (2021)
26. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary detr with conditional matching. In: Proceedings of the European Conference on Computer Vision. pp. 106–122 (2022)
27. Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., Zhang, L.: A simple framework for open-vocabulary segmentation and detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1020–1031 (October 2023)
28. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5729–5739 (2023)
29. Zheng, Y., Wu, J., Qin, Y., Zhang, F., Cui, L.: Zero-shot instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2593–2602 (2021)
30. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., Peng, N., Wang, L., Lee, Y.J., Gao, J.: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15116–15127 (June 2023)
31. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)