CE-OST: Contour Emphasis for One-Stage Transformer-based Camouflage Instance Segmentation

Thanh-Danh Nguyen[1,2], Duc-Tuan Luu[1,2], Vinh-Tiep Nguyen[†1,2], and Thanh Duc Ngo[1,2]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{*danhnt, tuanld, tiepnv, thanhnd*}*@uit.edu.vn*, [†]*corresponding author*

***Abstract*—Understanding camouflage images at instance level is such a challenging task in computer vision. Since the camouflage instances have their colors and textures similar to the background, the key to distinguish them in the images should rely on their contours. The contours seperate the instance from the background, thus recognizing these contours should break their camouflage mechanism. To this end, we address the problem of camouflage instance segmentation via the Contour Emphasis approach. We improve the ability of the segmentation models by enhancing the contours of the camouflaged instances. We propose the CE-OST framework which employs the well-known architecture of Transformer-based models in a one-stage manner to boost the performance of camouflaged instance segmentation. The extensive experiments prove our contributions over the state-of-the-art baselines on different benchmarks, i.e. CAMO++, COD10K and NC4K.**

***Index Terms*—Camouflage Image Understanding, One-Stage Instance Segmentation, Transformer-based model**

I. Introduction

In nature, animals conceal themselves from their enemies by blending in with their environment [1]. That is the so-called camouflaged animals or camouflaged instances. Various applications can be considered via autonomously distinguishing those instances e.g., search-and-rescue missions, wild species discovery, preservation campaign [2] and media forensics (manipulated image/video detection and segmentation [3]); or even in the medical domain where the tumors, polyps or cells seems to be similar to other cells [4]–[6]. Despite the long development history of image segmentation methods, general segmentation models cannot deal well with camouflaged instances [7]–[10]. At first, traditional approaches [11]–[18] utilized low-level handcrafted features for simple camouflaged images. Recent models with a learning-based approach rely on large-scale datasets [2], [19] to address camouflaged object segmentation.

The concealing mechanism of camouflage animals aims to merge their texture and color with the surroundings. Therefore, human vision fails to recognize those instances at a glance. However, such boundaries of the camouflaged animals can hardly be disappeared. In this work, based on the biological assumption, we propose a method to enhance the boundary of the camouflaged instance with the purpose of supporting the segmentation models to segment the instances. Inspired by the work on boundary detection [20]–[23], we enhance the boundary of the camouflaged instances in the images so that the segmentation model can better differentiate the instances from the background.

Furthermore, recent work on computer vision has achieved a significant explosion in performance since the work [24] on Transformer architecture. Lately, the methods are applied commonly to the computer vision domain [25], [26]. Specifically, the performance of instance segmentation models has also improved thanks to transformer architecture [27]–[30]. In this work, we follow OSFormer [30] which is the first one-stage transformer-based framework designed for this task.

To summarize, in this work, we propose a simple yet effective boundary enhancement module in a plug-and-play manner. We focus on addressing camouflaged image instance segmentation via a one-stage transformer-based model. To this end, we propose the **C**ontour **E**mphasis for **O**ne-**S**tage **T**ransformer-based Camouflage Instance Segmentation, dubbed CE-OST framework. To prove our method, we conduct experiments on the three well-known benchmarks in this camouflage domain, i.e. CAMO++ [31], COD10K [19], and NC4K [32]. The reported results demonstrate the performance of our proposed method over state-of-the-art baselines.

The rest of this paper is organized as follows. Section II reviews related work on camouflage research, Section III presents our proposed method - CE-OST. In Section IV, extensive experiments and ablation studies prove the effectiveness of our proposal. Finally, Section V concludes our work.

II. Related work

A. Camouflaged Research

Early approaches mainly exploit low-level handcrafted features, including color features, edge, texture, and brightness [33]–[35]. Recent studies have taken advantage of the large capacity of deep networks to recognize more intricate characteristics of camouflage, enhancing the performance of detecting camouflaged objects. Zhai *et al.* [36] utilized a mutual graph learning technique to interactively train the boundaries and regions of camouflaged objects. PFNet [37] was designed to simulate the natural process of predation. Le *et al.* [2] introduced Anabranch network that combines both object classification and segmentation. SINet [19] tried to mimic the predators' hunting behavior by containing two main modules for locating and identifying the camouflaged objects. Lyu *et al.* [32] designed a network that ranks concealed

TABLE I
COMPARISON AMONG CAMOUFLAGE DATASETS (WITHOUT NON-CAMOUFLAGED IMAGES).

Dataset	#Annot. Camo. Img.	#Meta-Cat.	#Obj. Cat.	Bbox. GT	Obj. Mask GT	Ins. Mask GT
CAMO [2]	1,250	2	8	×	✓	×
COD10K [19]	5,066	5	69	✓	✓	✓
NC4K [32]	4,121	5	69	✓	✓	✓
CAMO++ [31]	2,695	10	47	✓	✓	✓

objects while simultaneously localizing and segmenting them to enhance prediction accuracy. A dual-stream MirrorNet [38] was proposed to capture various scene layouts while TINet [39] interactively refined texture and segmentation features at multiple levels. Le *et al.* [31] presented CFL approach which integrates various models via learning image contexts.

B. Instance Segmentation

Instance segmentation in computer vision involves identifying individual objects and generating their corresponding masks. Existing methods can be divided into two categories: **Two-stage** and **One-stage approach**. Methods in the first group employ a traditional detect-then-segment scheme that initially identifies Regions of Interest (ROIs) using bounding boxes and subsequently generates local pixel-level instance segmentation [40]. Mask RCNN [41], built upon Faster RCNN [42], is a well-known approach that incorporates an additional mask-prediction branch at the instance level. Mask Scoring RCNN [43] includes a MaskIOU head on top of Mask RCNN to evaluate the quality of the predicted instance masks. Cascade Mask RCNN [44] is a multi-stage architecture including a series of detectors trained with increasing IOU thresholds to filter out false positives more effectively. PANet [45] was proposed to shorten information flow and enhance the feature extractor by designing a bottom-up path augmentation. Moreover, Chen *et al.* [46] presented the HTC to combine both detection and segmentation features for joint processing. Recently, DCNet was introduced [47] with a de-camouflaging mechanism to extract the camouflage characteristics.

In the second category, single-stage methods adopt the anchor-free object detection approach. YOLACT [48] is the first method that attempts real-time instance segmentation by combining the results of two parallel tasks: generating a set of non-local prototype masks and predicting per-instance mask coefficients. BlendMask [49], based on YOLACT, initially generates dense yet shallow positional sensitive instance features for each pixel. Then, a blender module will merge the features for each instance to produce an attention map. SOLO [50], [51] redefines instance segmentation as predicting categories then generating masks. This method utilizes semantic categories to locate the center of the instances and separates mask prediction into dynamic kernel feature learning. Consequently, the output masks are generated without the need to compute bounding boxes. CondInst [40] can solve instance segmentation with fully convolutional networks while eliminating the ROI cropping and feature alignment.

C. Camouflaged Datasets

CHAMELEON [52] and Camouflaged Animals [53] are the first two camouflage datasets providing mask annotations. However, the sizes of the test datasets are less than 300, which is insufficient for deep learning methods. Regarding object detection task, MoCA dataset [54] was introduced which contains only bounding box ground-truth. In terms of suitable instance segmentation datasets, i.e. [2], [19], [31], [55], we carefully describe in Section IV. Table I provides a comprehensive comparison on our chosen datasets.

III. PROPOSED METHOD

A. Overview our CE-OST framework

Our proposed framework of Contour Emphasis for One-Stage Transformer-based Camouflage Instance Segmentation, dubbed **CE-OST**, is illustrated in Figure 1. There are two main blocks: Contour Emphasis Block and Transformer Block. A camouflaged input image should go through the two blocks before meeting the Fusion Module at the end of the framework to return the segmentation mask. Our proposed Contour Emphasis Block can be considered a portable plug-and-play component. The image passing through this block has its boundary enhanced. Then, the enhanced image continues its journey to perform feature extraction via a CNN model. The extracted features join the One-Stage Transformer Block to conduct instance segmentation masks. The details are explained in the following.

B. Contour Emphasis Method

Boundary plays an important role in supporting our vision to recognize the whole shape of an arbitrary instance or object. Since the work on handcrafted features like Canny Edge Detection to work applied deep methods like HED [22], or COB [20], [21] set the very outstanding performance on edge detection. In this work, we propose a Contour Emphasis approach to enhance the visual features of camouflaged instances to improve the segmentation model. In Figure 1, we present the **C**ontour **E**mphasis (CE) Block that takes the responsibility of fusing the boundary to the original image.

Originating from HED [22], we adopt a multi-scale convolutional network with a pre-trained VGG-16 backbone [56] to detect the instance boundary. First, the whole CE Block is trained on an edge detection dataset, dubbed BSD500 dataset [57]. The losses are computed at multi-scale with pixel-wise cross-entropy loss and fused into a total loss in the end. To reduce the edge annotation cost, we utilize the trained model to predict the camouflaged image boundary. The contours are then added to the original images to enhance their appearances.

Accordingly, there are several ways of contour combination. Therefore, we empirically choose the two methods, i.e *color contrast* (1) and *brightness addition* (2). In the brightness addition, we straight forward add the boundary result to the image (pixel-wise value addition). In the color contrast method, we perform addition on the compensation value of the pixel at the corresponding pixel-wise location. In Figure 3, readers can compare the visual differences between the two methods of our contour fusion. To overcome some intensive cases where the texture is too complex, we conduct a simple *grid-condition* procedure (Figure 2). The edge detector may

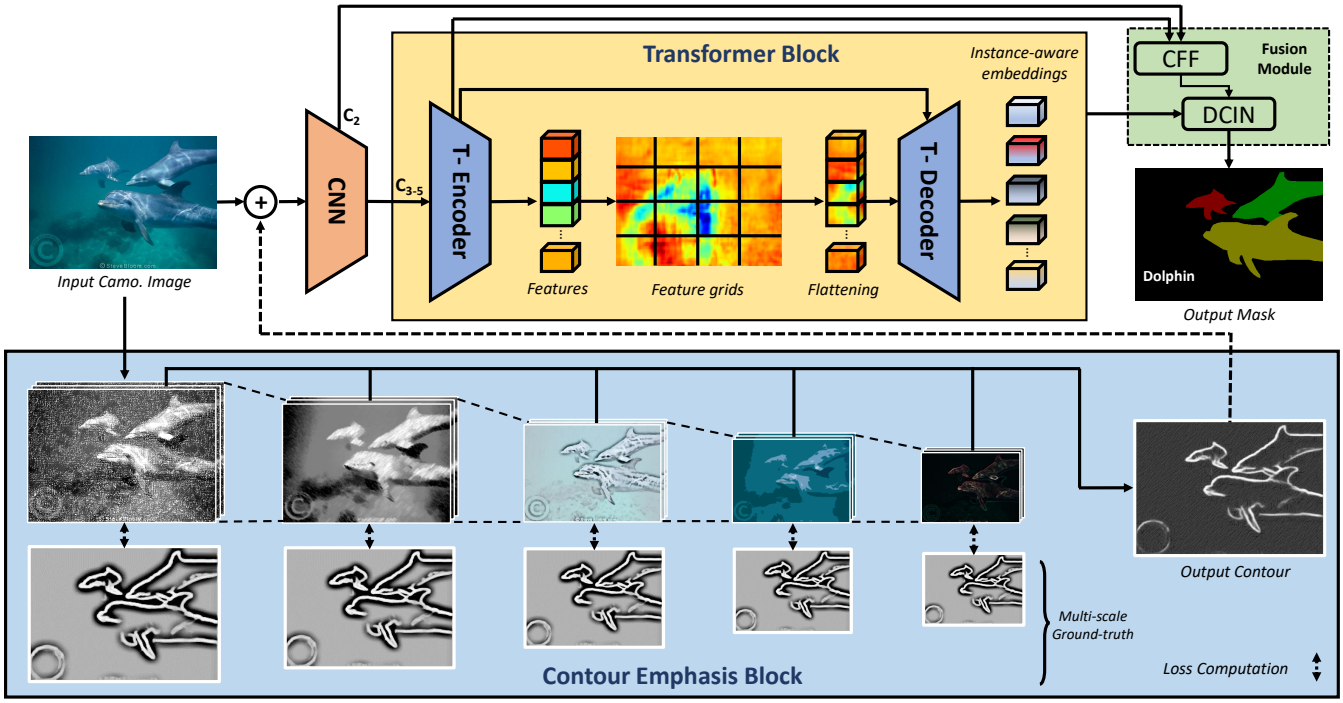

Fig. 1. Overall our CE-OST framework: Contour Emphasis for One-Stage Transformer-based Camouflage Instance Segmentation.

fail, whose results are further noisy than enhancing the camouflaged instances. In this case, we apply a grid 5×5 to each image and decide not to apply boundary fusion if the number of eliminated cells is over half (which is equal to approx. 12). A cell is eliminated if the number of pixels in the detected boundary area covers over a half area of the cell. This proposal is adaptive to every single image size of the camouflage dataset.

C. One-Stage Transformer-based Camouflage Instance Segmentation Model

Feature Extractor. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, we adopt a CNN to release multi-scale features C_{2-5}. The input for the Transformer Block is the flattened features from C_{3-5} while C_2 is fed into the Fusion Module for a low-level feature enhancement. The details of the backbones are in Section IV.

Transformer-based Instance Segmentation Model. We employ the structure of an Encoder-Decoder model [24], [30] since previous work done on Transformers proved the effectiveness of the self-attention layers in extracting global information from the image. In this architecture, the network focuses on the precise location, which is cruel for instance segmentation. Notably, the input of this block is multi-scale, compared to the limited single-scale of DETR [58].

Fusion Module. In this Fusion Module, we follow [30] while having Dynamic Camouflaged Instance Normalization (DCIN) and Coarse-to-Fine Fusion (CFF). The CNN feature C_2 and middle features of the T-Encoder are sent to the CFF to create comprehensive features. Then, features from the final layer of the Transformer Block and CFF are inputs of the DCIN. In DCIN, there is a fully-connected layer used to gain the location label. At the same time, a multi-layer perception is employed to gain the instance-aware parameters. These parameters are then used for establishing the segmentation mask. Please visit [30] for more implementation details.

IV. Experiments

A. Dataset and Settings

In our experiments, we utilize COD10K [19], NC4K [32] and CAMO++ [31] for camouflage instance segmentation. Please see Table I for a more detailed comparison among these datasets. The following paragraphs are their brief reviews.

COD10K Dataset. [19] comprises around $10,000$ images divided into 5 meta-categories. However, in terms of camouflage, the training set of COD10K contains approximately $3,040$ and the test set includes $2,026$ images.

NC4K Dataset. The NC4K is a testing dataset [32] with $4,121$ images of camouflaged instances collected from online resources. We utilize this benchmark to evaluate our proposed method applied to the one-stage transformer-based model.

CAMO++ Dataset. The original CAMO++ dataset contains both images of camouflaged and non-camouflaged instances with a total of 5,500 images corresponding to 32,756 instances [31]. There are 47 fine-grained camouflaged classes designed with a hierarchical structure and assigned into 10 coarse-grained classes. CAMO++ contributes 2,695 camouflage images including 1,250 existing images in CAMO [2] and 1,450 newly collected images.

Experimental Settings. To select the base models, we employed ResNet-50 [59], ResNet-50 (with input size of 550×550 for real-time manner) [59], ResNet-101 [59], Pyramid Vision Transformer (PVT) [60], and Swin Transformer (Swin-T) [61] to apply our proposed method. The framework was built on top of Detectron2 [62] and other models originated from their own publications. Our Contour Emphasis

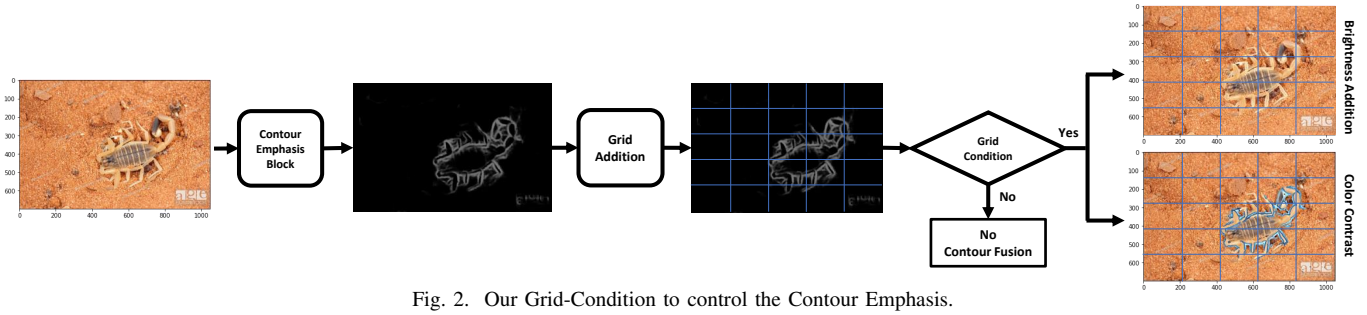

Fig. 2. Our Grid-Condition to control the Contour Emphasis.

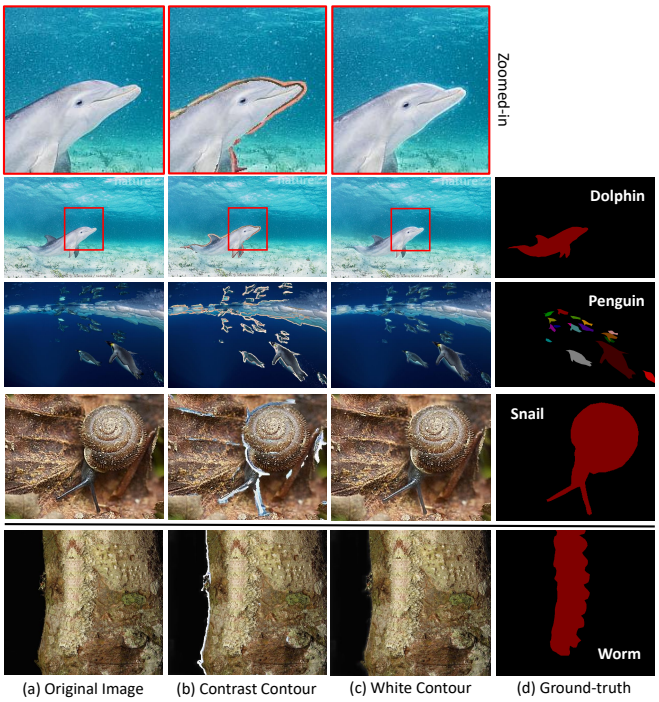

Fig. 3. Exemplary contour emphasized images on CAMO++ [31]. The first rows are the zoomed-in regions. Best viewed online with colors and zoom-in.

framework follows the original configurations [30]. In detail, we adopted a single GeForce RTX 2080Ti GPU and trained with Stochastic Gradient Descent. To initialize the models, we utilized the well-known pre-trained weights of ImageNet [63]. Our training process occurred with $90K$ iterations with a batch size of 1 and base learning rate of $2.5e^{-4}$ heading by $1K$ iterations of warming-up. We also use a learning rate reduction of 0.1 after $60K$ and $80K$ iterations. The weight decay and the momentum values are 10^{-4} and 0.9, respectively.

Evaluation metrics. To report our results, we use average precision (AP). To be detailed, we report AP, AP@50, and AP@75. Readers can reach this site https://cocodataset.org/#detection-eval for more details on the evaluation metrics.

B. State-of-the-art Comparison

To prove the performance of our proposed method - the Contour Emphasis approach, we reported the established experiments in Table II. With this setting, we utilize the results on COD10K and NC4K datasets to compare the performance among models. To this end, we employed models from two-stage (i.e. [41], [43], [44], [46], [49], [64]) and one-stage approaches (i.e. [30], [48], [51], [65]–[67]). For a fair comparison, ResNet-101 [59], which is the popular backbone utilized by other publications, is utilized. The reported results recognized our improvement on all AP, AP50, and AP75 evaluation metrics. On COD10K [19], we achieved $43.2\%, 72.2\%$, and 44.1% on AP, AP50, and AP75, respectively. On NC4K [32], the three respective values were $45.1\%, 74.0\%$, and 46.4%. To this end, we present the state-of-the-art results over the baseline methods on both one-stage and two-stage approaches. Please find the next ablation section for more empirical details of our Contour Emphasis method.

TABLE II
STATE-OF-THE-ART COMPARISON ON COD10K [19] AND NC4K [32] DATASET. THE CHOSEN BACKBONE IS THE COMMON RESNET-101 [59].

Method		COD10K			NC4K		
		AP	AP50	AP75	AP	AP50	AP75
Two-Stage	Mask R-CNN [41]	28.7	60.1	25.7	36.1	68.9	33.5
	MS R-CNN [43]	33.3	61.0	32.9	35.7	63.4	34.7
	Cascade R-CNN [44]	29.5	61.0	25.9	34.6	66.3	31.5
	HTC [46]	30.9	61.0	28.7	34.2	64.5	31.6
	BlendMask [49]	31.2	60.0	28.9	31.4	61.2	28.8
	Mask Transfiner [64]	31.2	60.7	29.8	34.0	63.1	32.6
One-Stage	YOLACT [48]	29.0	60.1	25.3	37.8	70.6	35.6
	CondInst [65]	34.3	67.9	31.6	38.0	71.1	35.6
	QueryInst [66]	32.5	65.1	28.6	38.7	72.1	37.6
	SOTR [67]	32.0	63.6	29.2	34.3	65.7	32.4
	SOLOv2 [51]	35.2	65.7	33.4	37.8	69.2	36.1
	OSFormer [30]	42.0	71.3	42.8	44.4	73.7	45.1
	CE-OST (Ours)	**43.2**	**72.2**	**44.1**	**45.1**	**74.0**	**46.4**

C. Ablation Study

In our CE-OST framework, the Contour Emphasis can be applied in two ways. In Table III, we present the effectiveness of color contrasting and brightness addition. We also conducted experiments on different base models including the 5 aforementioned methods in the Experimental Settings section. In general, Transformer-based models such as Swin-T or PVT give the best results among methods. The CAMO++ [31] is the most intensive dataset following by NC4K [32] and COD10K [19]. This ablation study also proves the generality of our CE-OST over the base models when improving almost every result in comparison with the state-of-the-art baselines. Especially to CAMO++ [31] dataset, the PVT backbone stably holds the best performance. The results can be explained as the multi-scale feature extractor of PVT can well handle the various scales of CAMO++. In Figure 4, we present our best results on the PVT backbone (left) and some failure cases (right) of over-segmentation or mislabeling. The instance scale and too complex background cause this phenomenon.

TABLE III
ABLATION STUDY ON DIFFERENT BASE MODELS ON COD10K [19], NC4K [32], AND CAMO++ [31].

Method	Base-Model	COD10K			NC4K			CAMO++		
		AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75
OSFormer	ResNet-50 [59]	41.0	71.1	40.8	42.5	72.5	42.3	19.0	33.8	18.3
	ResNet-50-550 [59]	-	-	-	-	-	-	20.1	36.3	19.3
	ResNet-101 [59]	42.0	71.3	42.8	44.4	73.7	45.1	20.6	34.4	20.2
	PVTv2-B2-Li [60]	47.2	74.9	49.8	-	-	-	27.7	44.7	27.9
	Swin-T [61]	47.7	78.6	49.3	-	-	-	22.3	36.6	21.8
CE-OST (Color Contrast)	ResNet-50 [59]	41.6	70.7	42.3	42.4	71.4	42.6	20.1	34.2	19.6
	ResNet-50-550 [59]	35.9	65.2	34.3	41.1	70.9	41.1	20.6	35.7	20.0
	ResNet-101 [59]	43.2	72.2	44.1	45.1	74.0	46.4	21.7	36.6	21.3
	PVTv2-B2-Li [60]	48.4	75.7	51.3	51.4	77.9	55.0	28.5	45.3	29.9
	Swin-T [61]	49.1	78.0	52.1	50.5	78.9	53.1	22.7	37.6	22.4
CE-OST (Brightness Addition)	ResNet-50 [59]	41.2	69.0	41.6	42.4	71.1	42.9	20.2	34.8	19.5
	ResNet-50-550 [59]	35.9	65.2	34.6	40.8	71.1	40.3	21.0	37.1	20.3
	ResNet-101 [59]	42.4	70.8	43.7	44.2	73.1	45.0	21.1	34.4	20.9
	PVTv2-B2-Li [60]	47.9	74.6	50.5	51.1	77.3	54.9	27.9	45.1	29.2
	Swin-T [61]	49.0	78.5	51.4	50.8	79.3	53.9	22.7	38.4	23.1

**The first, second, and third best results are marked in red, blue, and green, respectively.*

D. Discussion

In Figure 3, we present several examples of camouflaged instances with their enhanced boundaries. From left to right, we show the original (a), contrast contour (b), bright contour (c), and ground-truth (d) images, respectively. Both kinds of contours are generated by the Contour Emphasis, and we present these contours under two appearances. The bright contours are the brightness addition to the boundary lines of the instances. While, the contrast contours shift the color values to another value in the contrast range, bringing a better-distinguished contour view compared to bright color contours. Thus, we can observe the first best results major in the Color Contrast approach. The last row illustrates a case where the recognized boundary fails to enhance the visual appearance of the considered instance, i.e. worm. Our future work should focus on camouflage boundary recognition.

V. CONCLUSION

In this work, we proposed the CE-OST framework - a Contour Emphasis approach for One-Stage Transformer-based model to address the instance segmentation task on camouflaged images. We have demonstrated the improvement of our proposed method over the three well-known camouflage benchmarks of COD10K, NC4K, and CAMO++. In the future, we plan to extend our idea to other specific domains of medical imaging where the instances carry camouflaged features.

VI. ACKNOWLEDGEMENT

This research is funded by University of Information Technology - Vietnam National University Ho Chi Minh City under grant number D1-2023-20. Thanh-Danh Nguyen was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2022.ThS.104. We also would like to acknowledge Tam V. Nguyen, Assoc. Prof. for his inspiration for camouflage research.

REFERENCES

[1] S. Singh, C. Dhawale, and S. Misra, "Survey of object detection methods in camouflaged image," *IERI Procedia*, vol. 4, pp. 351 – 357, 2013.

[2] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *CVIU*, vol. 184, pp. 45–56, 2019.

[3] T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild," in *ICCV*, 2021.

[4] Q. Zhangli, J. Yi, D. Liu, X. He, Z. Xia, Q. Chang, L. Han, Y. Gao, S. Wen, H. Tang *et al.*, "Region proposal rectification towards robust instance segmentation of biological images," in *MICCAI*. Springer, 2022, pp. 129–139.

[5] X. Liu, B. Hu, W. Huang, Y. Zhang, and Z. Xiong, "Efficient biomedical instance segmentation via knowledge distillation," in *MICCAI*. Springer, 2022, pp. 14–24.

[6] C. Li, D. Liu, H. Li, Z. Zhang, G. Lu, X. Chang, and W. Cai, "Domain adaptive nuclei instance segmentation and classification via category-aware feature alignment and pseudo-labelling," in *MICCAI*. Springer, 2022, pp. 715–724.

[7] C. Kervrann and F. Heitz, "A markov random field model-based approach to unsupervised texture segmentation using local and global spatial statistics," *IEEE TIP*, vol. 4, no. 6, pp. 856–862, 1995.

[8] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient nd image segmentation," *IJCV*, vol. 70, no. 2, pp. 109–131, 2006.

[9] X. Li and H. Sahbi, "Superpixel-based object class segmentation using conditional random fields," in *ICASSP*, 2011, pp. 1101–1104.

[10] L. Sulimowicz, I. Ahmad, and A. Aved, "Superpixel-enhanced pairwise conditional random field for semantic segmentation," in *ICIP*, 2018.

[11] M. Galun, E. Sharon, R. Basri, and A. Brandt, "Texture segmentation by multiscale aggregation of filter responses and shape elements," in *ICCV*, Oct 2003, pp. 716–723.

[12] L. Song and W. Geng, "A new camouflage texture evaluation method based on wssim and nature image features," in *ICMT*, Oct 2010.

[13] F. Xue, C. Yong, S. Xu, H. Dong, Y. Luo, and W. Jia, "Camouflage performance analysis and evaluation framework based on features fusion," *MTAP*, vol. 75, pp. 4065–4082, 2016.

[14] Y. Pan, Y. Chen, Q. Fu, P. Zhang, and X. Xu, "Study on the camouflaged target detection method based on 3d convexity," *Modern Applied Science*, vol. 5, no. 4, p. 152, 2011.

[15] Z. Liu, K. Huang, and T. Tan, "Foreground object detection using top-down information based on em framework," *IEEE TIP*, vol. 21, no. 9, pp. 4204–4217, Sept 2012.

[16] A. W. P. Sengottuvelan and A. Shanmugam, "Performance of decamouflaging through exploratory image analysis," in *ICETET*, 2008.

[17] J. Yin, Y. Han, W. Hou, and J. Li, "Detection of the mobile object with camouflage color under dynamic background based on optical flow," *Procedia Engineering*, vol. 15, pp. 2201 – 2205, 2011.

[18] J. Gallego and P. Bertolino, "Foreground object segmentation for moving camera sequences based on foreground-background probabilistic models and prior probability maps," in *ICIP*, Oct 2014, pp. 3312–3316.

[19] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *CVPR*, 2020, pp. 2777–2787.

[20] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. V. Gool, "Convolutional oriented boundaries," in *ECCV*, 2016.

[21] ——, "Convolutional oriented boundaries: From image segmentation to high-level tasks," *IEEE TPAMI*, vol. 40, no. 4, pp. 819 – 833, 2018.

Fig. 4. Qualitity visualization results on the CAMO++ [31] testing set on our CE-OST-PVT. The confidence threshold is 0.5.

[22] S. Xie and Z. Tu, "Holistically-nested edge detection," in *ICCV*, 2015, pp. 1395–1403.

[23] S. Candemir and S. Antani, "A review on lung boundary detection in chest x-rays," *IJCARS*, vol. 14, pp. 563–576, 2019.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[25] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[26] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE TPAMI*, vol. 45, no. 1, pp. 87–110, 2022.

[27] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," in *CVPR*, Oct 2023.

[28] H. Zhang, F. Li, H. Xu, S. Huang, S. Liu, L. M. Ni, and L. Zhang, "Mp-former: Mask-piloted transformer for image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 074–18 083.

[29] F. Li, H. Zhang, H. xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," in *CVPR*, Oct 2023.

[30] J. Pei, T. Cheng, D.-P. Fan, H. Tang, C. Chen, and L. Van Gool, "Osformer: One-stage camouflaged instance segmentation with transformers," in *ECCV*. Springer, 2022.

[31] T.-N. Le, Y. Cao, T.-C. Nguyen, M.-Q. Le, K.-D. Nguyen, T.-T. Do, M.-T. Tran, and T. V. Nguyen, "Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite," *IEEE TIP*, vol. 31, pp. 287–300, 2022.

[32] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan, "Simultaneously localize, segment and rank the camouflaged objects," in *CVPR*, 2021, pp. 11 591–11 601.

[33] F. Xue, C. Yong, S. Xu, H. Dong, Y. Luo, and W. Jia, "Camouflage performance analysis and evaluation framework based on features fusion," *MTAP*, vol. 75, no. 7, pp. 4065–4082, 2016.

[34] S. Li, D. Florencio, Y. Zhao, C. Cook, and W. Li, "Foreground detection in camouflaged scenes," in *ICIP*. IEEE, 2017, pp. 4247–4251.

[35] T. W. Pike, "Quantifying camouflage and conspicuousness using visual salience," *Methods in Ecology and Evolution*, vol. 9, no. 8, 2018.

[36] Q. Zhai, X. Li, F. Yang, C. Chen, H. Cheng, and D.-P. Fan, "Mutual graph learning for camouflaged object detection," in *CVPR*, Jun 2021.

[37] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged object segmentation with distraction mining," in *CVPR*, 2021.

[38] J. Yan, T.-N. Le, K.-D. Nguyen, M.-T. Tran, T.-T. Do, and T. V. Nguyen, "Mirrornet: Bio-inspired camouflaged object segmentation," *IEEE Access*, vol. 9, pp. 43 290–43 300, 2021.

[39] J. Zhu, X. Zhang, S. Zhang, and J. Liu, "Inferring camouflage objects by texture-aware interactive guidance network," in *AAAI*, 2021.

[40] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *ECCV*, 2020.

[41] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2980–2988.

[42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *IEEE TPAMI*, 2016.

[43] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring r-cnn," in *CVPR*, 2019.

[44] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *CVPR*, 2018.

[45] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *CVPR*, 2018.

[46] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang *et al.*, "Hybrid task cascade for instance segmentation," in *CVPR*, 2019, pp. 4974–4983.

[47] N. Luo, Y. Pan, R. Sun, T. Zhang, Z. Xiong, and F. Wu, "Camouflaged instance segmentation via explicit de-camouflaging," in *CVPR*, 2023, pp. 17 918–17 927.

[48] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *ICCV*, 2019.

[49] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, "Blendmask: Top-down meets bottom-up for instance segmentation," in *CVPR*, 2020, pp. 8573–8581.

[50] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "SOLO: Segmenting objects by locations," in *ECCV*, 2020.

[51] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic, faster and stronger," in *NeurIPS*, 2020.

[52] P. Skurowski, H. Abdulameer, J. Baszczyk, T. Depta, A. Kornacki, and P. Kozie, "Animal camouflage analysis: Chameleon database," 2018.

[53] E. L.-M. Pia Bideau, "It's moving! a probabilistic model for causal motion segmentation in moving camera videos," in *ECCV*, 2016.

[54] H. Lamdouar, C. Yang, W. Xie, and A. Zisserman, "Betrayed by motion: Camouflaged object discovery via motion segmentation," in *ACCV*, Nov 2020.

[55] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan, "Simultaneously localize, segment and rank the camouflaged objects," in *CVPR*, Jun 2021.

[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[57] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE TPAMI*, vol. 26, no. 5, pp. 530–549, 2004.

[58] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*. Springer, 2020, pp. 213–229.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, Jun 2016, pp. 770–778.

[60] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.

[61] Z. Liu, Y. Lin, Y. Cao, H. Han, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Hierarchical vit using shifted windows," in *CVPR*, 2021.

[62] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009.

[64] L. Ke, M. Danelljan, X. Li, Y.-W. Tai, C.-K. Tang, and F. Yu, "Mask transfiner for high-quality instance segmentation," in *CVPR*, 2022.

[65] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *ECCV*. Springer, 2020, pp. 282–298.

[66] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu, "Instances as queries," in *ICCV*, 2021, pp. 6910–6919.

[67] R. Guo, D. Niu, L. Qu, and Z. Li, "Sotr: Segmenting objects with transformers," in *ICCV*, 2021, pp. 7157–7166.