Indiscernible Object Counting in Underwater Scenes
Supplementary Material

Guolei Sun[1] Zhaochong An[1] Yun Liu[2] Ce Liu[1]
Christos Sakaridis[1] Deng-Ping Fan[1*] Luc Van Gool[1,3]
[1] CVL, ETH Zurich, [2] I2R, A*STAR, [3] VISICS, KU Leuven
guolei.sun@vision.ee.ethz.ch

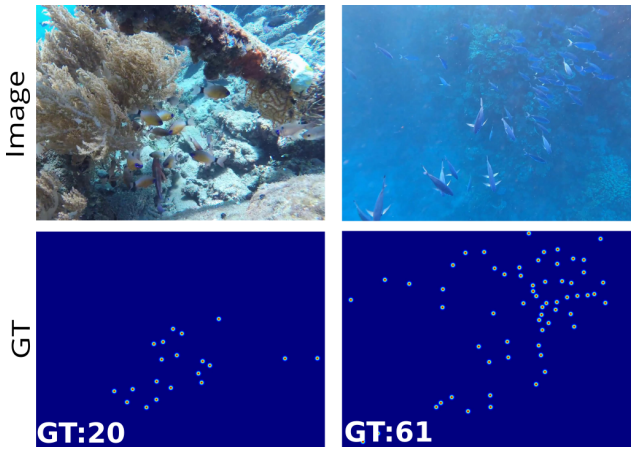

Figure 1. Ground-truth annotations, for example images shown in the *introduction* of the main paper. Best viewed with zooming.

1. More Details about IOCfish5K

Answers. The ground-truth point maps of the IOC examples shown in the *introduction* of the main paper are presented in Fig. 1. For a normal visual system, it is difficult to find all the object (fish) instances in the given examples because some objects are highly blended in their environment. Surprisingly, there are a total of 20 and 61 objects in the two examples, respectively.

More examples. We show more example images of IOCfish5K in Fig. 2. Our dataset contains plenty of high-quality images with various indiscernible levels and object densities. It would contribute to research areas such as indiscernible scene understanding and dense object counting.

Annotation details. As mentioned in our main paper, the annotation process is split into *three* steps. For the first step, we studied books and watched Youtube videos to learn the background knowledge about sea animals before instructing annotators. The used books include *Wonders of the Red Sea* and *Fishing guide Mediterranean + Atlantic*, which are available on Amazon. We also took some examples/pictures from the above materials to instruct annotators.

*The corresponding author

2. More Experimental Details

As mentioned, we evaluated 14 popular algorithms for DOC on IOCfish5K. Their codes are all publicly available. The details of these methods are as follows.

MCNN [16]: It proposes a multi-column convolutional neural network that contains different convolution branches with different receptive fields. The ground-truth density map is calculated using geometry-adaptive kernels.

CSRNet [4]: It aims at conducting crowd counting under highly congested scenes. CSRNet exploits dilated convolutions in this task and achieve promising results.

LCFCN [3]: This method predicts a blob for each object instance by using only point supervision. It achieves excellent performance in crowd counting as well as generic object counting.

CAN [8]: CAN processes encoded features (VGG-16) with different receptive fields, which are then combined using the learned weights. The final context-aware features are passed to estimate the density map.

DSSI-Net [7]: It focuses on tackling the problem of large-scale variation in crowd counting and proposes structured feature enhancement and dilated multi-scale structural similarity loss to generate better density maps.

BL [9]: Different from previous works which adopt L_1 or L_2 loss for supervising the learning of density maps, BL proposes a Bayesian loss which directly uses point annotations to learn density probability.

NoisyCC [11]: NoisyCC explicitly models the annotation noise in crowd counting with a joint Gaussian distribution. A low-rank covariance approximation is derived to improve the efficiency [11].

DM-Count [14]: This method proposes to exploit distribution matching for crowd counting. The optimal transport

Figure 2. More example images from the proposed IOCfish5K. From *left* column to *right* column: typical samples, indiscernible & dense samples, indiscernible & less dense samples, less indiscernible & dense samples, less indiscernible & less dense samples.

algorithm is used to minimize the gap between the predicted density map and the ground-truth point map.

GL [12]: GL proposes a perspective-guided optimal transport cost function for crowd counting. It is currently the most powerful loss for crowd counting and achieves state-of-the-art performance on mainstream DOC datasets compared to other loss functions.

P2PNet [10]: It directly predicts a number of point proposals (location and confidence score). Then Hungarian algorithm [1] is used to match proposals and point annotations. It is a purely point-based algorithm for crowd counting [10] and achieves impressive performance on DOC

datasets.

KDMG [13]: Different from previous density-based methods, which generates ground-truth density map by convolving the point map with a/an (adaptive) Gaussian kernel, KDMG proposes a density map generator that is jointly trained with counting model.

MPS [15]: This method generates multi-scale features for the crowd image and benefits from the joint learning of crowd counting as well as localization.

MAN [6]: It deals with the problem of large-scale variations in crowd counting by integrating global attention, local attention, and instance attention in a unified framework. MAN achieves state-of-the-art performance on mainstream datasets such as JHU++ and NWPU.

CLTR [5]: It directly predicts the point locations by adopting a transformer encoder and decoder structure to process the features. The trainable embeddings are used to extract object locations from the encoded features.

For the above methods, CAN, CSRNet and MCNN use the SGD optimization algorithm for training the network, while others use Adam optimizer [2]. For IOCFormer, the initial learning rate is set as 1e-5 and the weight decay is 5e-4. Following [5], our approach is trained by 1500 epochs.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[2] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3

[3] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Where are the blobs: Counting by localization with point supervision. In *ECCV*, 2018. 1

[4] Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In *IEEE CVPR*, 2018. 1

[5] Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *ECCV*, 2022. 3

[6] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *IEEE CVPR*, 2022. 3

[7] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *IEEE ICCV*, 2019. 1

[8] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *IEEE CVPR*, 2019. 1

[9] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *IEEE ICCV*, 2019. 1

[10] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *IEEE ICCV*, 2021. 2

[11] Jia Wan and Antoni Chan. Modeling noisy annotations for crowd counting. In *NeurIPS*, 2020. 1

[12] Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *IEEE CVPR*, 2021. 2

[13] Jia Wan, Qingzhong Wang, and Antoni B. Chan. Kernel-based density map generation for dense object counting. *IEEE TPAMI*, 44(3):1357–1370, 2022. 3

[14] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai. Distribution matching for crowd counting. In *NeurIPS*, 2020. 1

[15] Mohsen Zand, Haleh Damirchi, Andrew Farley, Mahdiyar Molahasani, Michael Greenspan, and Ali Etemad. Multi-scale crowd counting and localization by multitask point supervision. In *IEEE ICASSP*, 2022. 3

[16] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *IEEE CVPR*, 2016. 1