Supplementary Material:
CrowdCLIP: Unsupervised Crowd Counting via Vision-Language Model

Table 1. The comparisons of Frames Per Second (FPS) between our method and other methods. The results are conducted on an NVIDIA 3090 GPU.

Method	Label	Resolution	FPS
CSRNet [2]	Point	1024×768	18.4
BL [3]	Point	1024×768	21.3
CSS-CCNN [1]	None	1024×768	37.4
CrowdCLIP	None	1024×768	[24.0, 50.8]

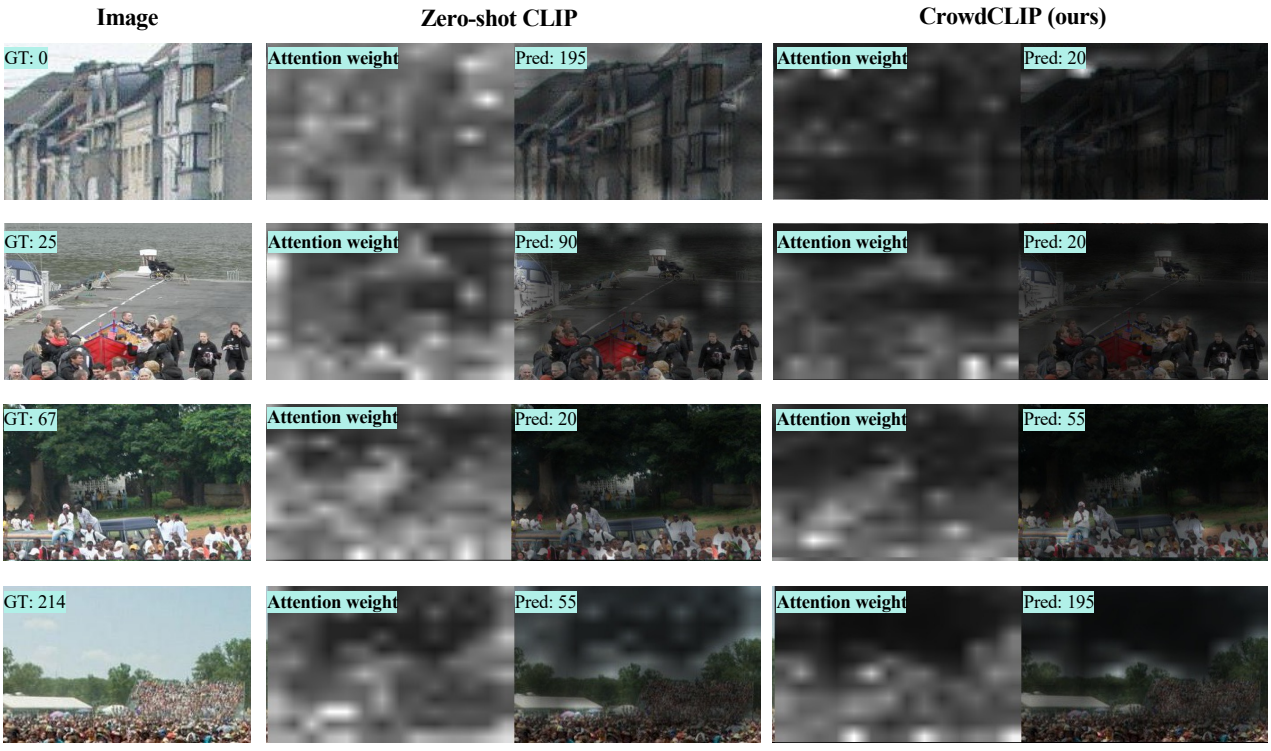

Figure 1. Qualitative visualizations of zero-shot CLIP and our proposed CrowdCLIP. From left to right, there are ground truth, results from zero-shot CLIP, and results from CrowdCLIP.

Figure 2. Evaluations of CSRNet [2] and the proposed CrowdCLIP on the Seoul Halloween crowd crush scenes. The two models are trained on the UCF-QNRF dataset.

1. Inference speed

We provide the comparison of inference speed, as shown in Tab. 1. Note that the run time of our CrowdCLIP is dynamic, as the proposed progressive filtering strategy aims to choose high-confidence crowd patches, while the number of selected crowd patches from different images is different. To this end, we report the range of inference speed, *i.e.*, [24.0, 50.8], where the former means all patches of a given image contain human heads, and the latter means there are no crowd patches. Specifically, the fully supervised methods need to maintain high-resolution features to generate high-quality density maps (*e.g.*, $1/8$ size of the input in CSRNet [2] and $1/16$ size of the input in BL [3]), leading to low inference speed. Compared with the unsupervised SOTA CSS-CCNN [1], our method is highly competitive in terms of inference speed.

2. More visualizations

Qualitative comparisons. We provide qualitative comparisons to further demonstrate the effectiveness of our method, as shown in Fig. 1. Specifically, we can observe that the zero-shot CLIP[1] can not understand crowd semantics well, leading to poor performance. In contrast, the proposed CrowdCLIP can generate more reasonable attention through the proposed ranking-based contrastive fine-tuning, resulting in better performance.

Testing on Seoul Halloween crowd crush scenes. On the night of 29 October 2022, a crowd crush occurred during Halloween festivities in the Itaewon neighborhood of Seoul, South Korea. At least 158 people were killed, and 196 others were injured. One of the reasons is that hundreds of people simultaneously appear in the narrow alley. A good crowd counting algorithm will help relieve the crowd crush event. In this part, we test the CrowdCLIP on the Seoul Halloween crowd crush scenes, as shown in Fig. 2. Note that there is no ground truth for these images, so we choose a representative fully-supervised counting method CSRNet [2] as a comparison. We can find that the prediction of our method is close to the CSRNet in most cases.

References

[1] Deepak Babu Sam, Abhinav Agarwalla, Jimmy Joseph, Vishwanath A Sindagi, R Venkatesh Babu, and Vishal M Patel. Completely self-supervised crowd counting via distribution matching. In *European Conference on Computer Vision*, pages 186–204. Springer, 2022. 1, 2

[2] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pages 1091–1100, 2018. 1, 2

[3] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Porc. of IEEE Intl. Conf. on Computer Vision*, pages 6142–6151, 2019. 1, 2

[1]Zero-shot CLIP means directly adopting the original non-fine-tuned image encoder of CLIP.