Single Domain Generalization for Few-Shot Counting via Universal Representation Matching

Supplementary Material

Method	MAE	RMSE
CLIP Enc.	27.32	43.28
URM-V	26.54	42.82

Table 1. **Comparison with generating visual prototypes by using CLIP vision encoder directly.**

Metric	MAE	MSE	MAE	MSE
Task	A → B		A → A	
CLIP(RN50×16)	28.8(-1.4)	45.7(-2.3)	11.7(+0.9)	61.4(+4.2)

Table 2. **Result of URM using CLIP as backbone and without universal knowledge distillation.**

1. More Analysis

Comparison with prototypes generated by CLIP. URM learns universal knowledge through distillation only during the training phase, making it as efficient as other methods during inference. In this section, we further compare our method with generating prototypes by CLIP vision encoder directly, which introduces more parameters. Specifically, the input image is encoded by the CLIP vison tower, and prototypes are obtained via global average pooling from the 2D visual tokens. Since obtaining category names during testing is unrealistic, we rely solely on the vision prototype. For fair comparison, we compare this approach with URM using only vision prototype, termed URM-V. The results in Table 1 demonstrate that our method can learn useful representation as well as achieve better performance.

CLIP as Backbone *E*. The result of URM w.o. distillation with the backbone changed to CLIP in Table 2 shows that this setting is inferior. This is mainly because CLIP with global supervision lacks fine-grained spatial details that are crucial for object counting. While the universal representation in CLIP does provide some improvements, it is not sufficient to compensate for the loss of spatial information. And the improvements primarily stem from the language representation.

2. Results of Domain Adaptation Methods

In this section, we further conduct experiments on domain adaptation methods. The results are shown is Table 3 and are provided only for reference as a part of data from target domain is visible during adaptation. We select SE CycleGAN [3] and DAOT [5], both of which are domain adaptation methods for crowd counting and have released code. Specifically, SE CycleGAN learns to translate scenes from one domain to another, aiming to reduce low-level differences. DAOT aligns domain-agnostic factors between domains via an aligned optimal transport strategy. It is worth noting that even part of the target data is provided, these DA methods cannot achieve results as good as those in other fields [1, 2], when data categories are disjoint in the context of class-agnostic setting of FSC, highlighting the importance the domain generalized FSC. Additionally, the poor performance of SE CycleGAN suggests that low-level image style is not the influencing factor contributing to domain shift.

Source → Target	A → B		B → A	
Metric	MAE	RMSE	MAE	RMSE
SE CycleGAN [3]	32.48	50.77	34.63	115.97
DAOT [5]	22.79	39.33	20.82	73.44

Table 3. **Results of the domain adaptation methods.**

3. Details of the Prompt Generator

In this section, we first detail the hand-written templates and prompts for LLMs we used in Table 4. Then, we perform an ablation study on the prompts for language representation in Table 5. Although using the naive template only achieves satisfactory performance, we find that applying all the Imagenet prompts used in CLIP leads to poor performance, due to the dataset distribution gap discussed above. Thus, we only use the general templates which yield better performance. Additionally, it is worth noting that we append "within 20 words" to some prompts for LLMs, as GPT tends to refuse to answer questions that describe low-detail objects. Finally, URM-L with language representation learning alone achieves the best result with our prompt generator.

4. Visualization

The visualization results obtained by MaskCLIP [4] are shown in Figure 1. We show that the method yields compelling open set segmentation results and is robust to data augmentation.

Templates	A photo of a {category name}. A photo of {number} {category name}. A bad photo of a {category name}. A photo of many {category name}. A low resolution photo of the {category name}. A photo of a hard to see {category name}. A cropped photo of a {category name}. A blurry photo of a {category name}. A good photo of a {category name}.
LLM-prompts	Describe what a {category name} looks like? How can you identify a {category name}? Describe what a {category name} in a distance look like within 20 words. Describe what a {category name} in a low resolution photo look like within 20 words.

Table 4. **The templates and prompts for LLM used in our prompt generator.**

Metric	MAE	RMSE
naive template	25.44	42.59
80 ImageNet templates	26.54	43.82
general templates	24.65	40.34
customized prompts	24.84	41.68
URM-L	23.83	41.03

Table 5. **Ablation of the prompts for language representation.**

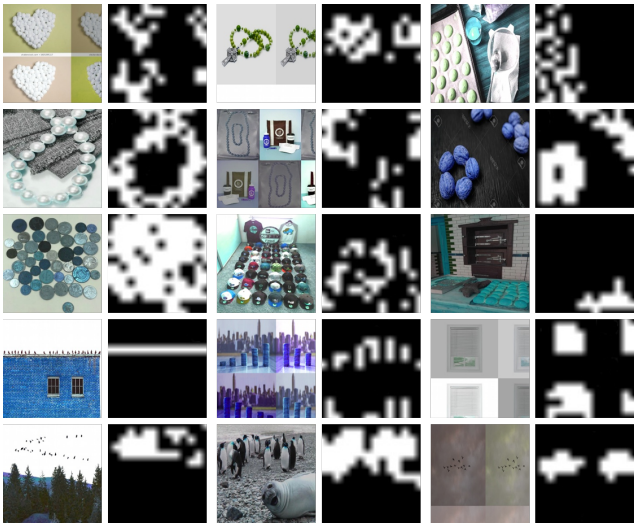

Figure 1. **Visualization of the segmentation.**

References

[1] Shikun Liu, Tianchun Li, Yongbin Feng, Nhan Tran, H. Zhao, Qiu Qiang, and Pan Li. Structural re-weighting improves graph domain adaptation. In *International Conference on Machine Learning*, 2023. 1

[2] Zhuoxuan Peng and S.-H. Gary Chan. Single domain generalization for crowd counting. *ArXiv*, abs/2403.09124, 2024. 1

[3] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8190–8199, 2019. 1

[4] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, 2021. 1

[5] Huilin Zhu, Jingling Yuan, Xian Zhong, Zhengwei Yang, Zheng Wang, and Shengfeng He. Daot: Domain-agnostically aligned optimal transport for domain-adaptive crowd counting. *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 1