Appendix

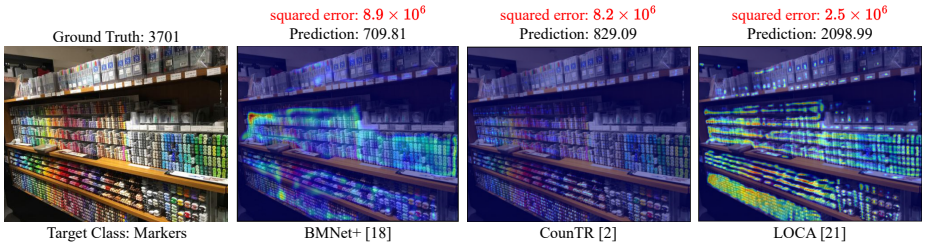

Fig. A: Comparison of CAC models evaluated on 1123.jpg from FSC-147 dataset. Both BMNet+ and CounTR exhibit a considerable difference in squared error compared to LOCA.

1 Counting error of 1123.jpg on FSC-147 dataset

In Section 3.3 of the main paper, we demonstrate that images with a high object count significantly impact MAE and RMSE scores. Specifically, we show that most CAC models produce large counting errors for 1123.jpg (Figure A), which is an outlier in this dataset.

2 Test-Time Normalization

We construct a variant of our MGCAC (denoted as MGCAC*) by adopting test-time normalization on MGCAC. Specifically, we utilize the divide-and-conquer technique in [1], where each query image is cropped by $M \times M$ pieces if the average reference to query area ratio is less than a threshold T. For each cropped sub-image, we first interpolate it to the query's original size followed by feeding the resized image into the model. The predicted count of the entire image is the sum of the predicted count of M^2 cropped images. We set $M = 8$ and $T = 0.0002$. As shown in Figure B, MGCAC* exhibits improvement on 1123.jpg (i.e. the outlier with the largest object count from the test dataset) and achieves SOTA on test MAE and test RMSE (cf. Table 2 in the main paper). Given the bias of MAE and RMSE towards extremely large-count query images (cf. Section 3.3 in the main paper), we demonstrate that our model is able to easily hack existing metrics by applying Test-Time Normalization on large-count query images (e.g., 1123.jpg). To provide a more balanced evaluation, we suggest utilizing metrics like NAE or RMSE, which can better reflect the overall performance of the model across the entire dataset.

3 Model Architecture

In this section, we provide the details of our proposed MGCAC as shown in Figure D, which is another strong baseline candidate for CAC. Deviating from the standard CAC pipeline that exclusively matches high-level features, the MGCAC model follows MixFormer [2] architecture design, which simultaneously extracts visual features from the reference images and augments query features by matching a similarity map among the references and the query, and resulting in finer matching results. We also aggregate features across model stages to capture the nuance of multi-scale features. More explanations of specific components are described as follows.

3.1 Extract-And-Match Module

To capture reference-specific features and enhance cross-branch interactions, we adopt the MAM (Mixed-Attention Module) from [2]. As shown in Figure D(2), both reference and query branches are first performed with convolutional projection to preserve spatial context. Then, the reference feature strengthens its representation through self-attention, and the query branch captures the query's pixel-wise correlation through cross-attention with reference features. By refining and matching within the same block, the model excels at shaping discriminative features that generate robust cross-attention similarity maps. Unlike previous CAC models that perform feature extraction separately on query and reference and then interact through matching, MAM facilitates frequent cross-branch information exchange.

3.2 Multi-Scale Enhancement for Matching and Density Map Estimation

After feature extraction and matching, most of the prior works adopt a simple density head consisting of several upsampling and convolution operations

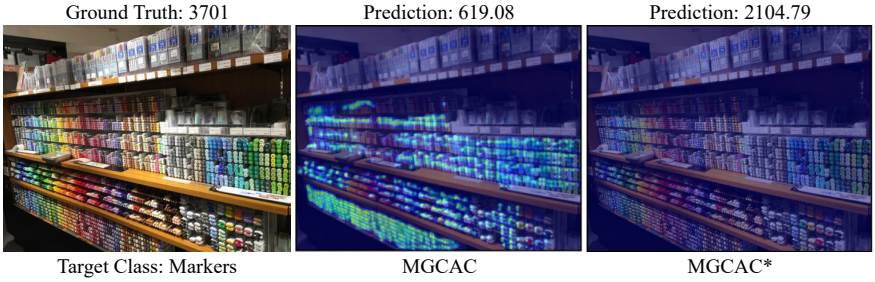

Fig. B: It shows the influence of applying test-time normalization on 1123.jpg in the FSC-147 dataset using our MGCAC, where * indicates applying test-time normalization.

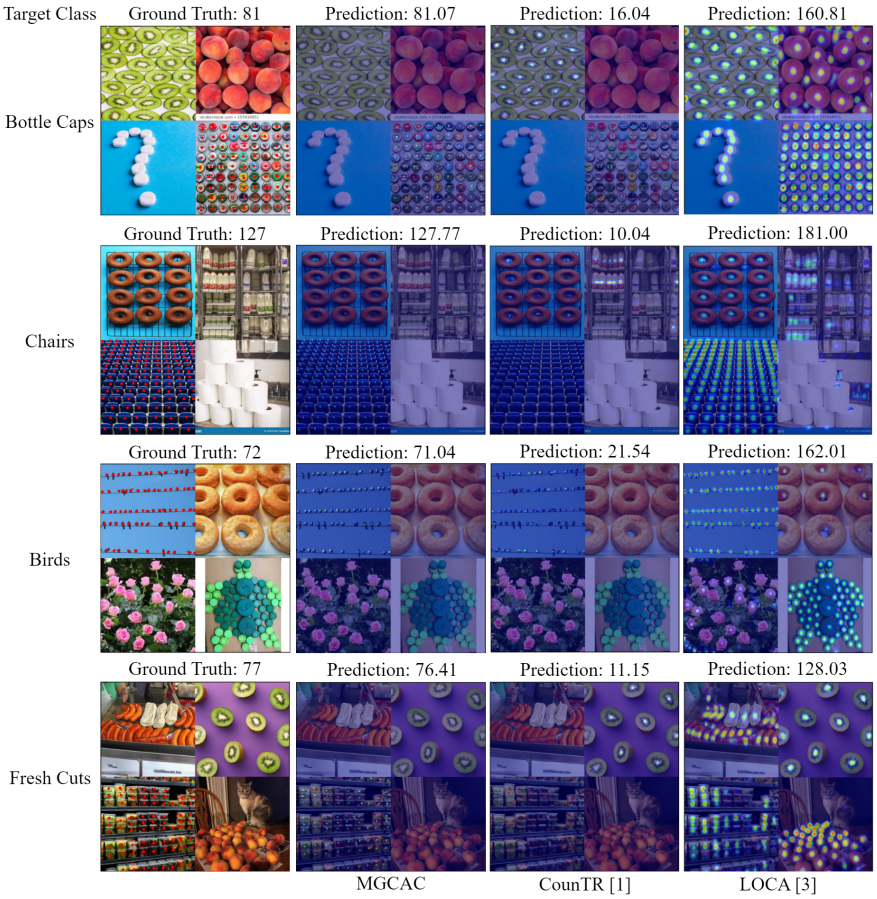

Fig. C: Qualitative Comparisons on Mosaic Dataset based on FSC-147. The first column indicates the target classes, the second column displays queries with ground truth annotated by red dots and the following columns present the predictions from the respective CAC models (please kindly zoom in to better visualize the highlighted regions of heat/response maps).

on single-scale feature maps and achieve acceptable performance. To better preserve the information of fine features from early layers/stages across scales, we respectively perform matching on features across stage 1, 2, and 3 of the model, where the feature dimensions are decreased by half as the stage progresses. Then, we utilize a U-Net-like expansion path as the density head to fuse query and similarity-related features across scales. After fusing feature maps, we utilize 1×1 convolutions to compress multi-channel feature maps into a single final density map as shown in Figure D(3).

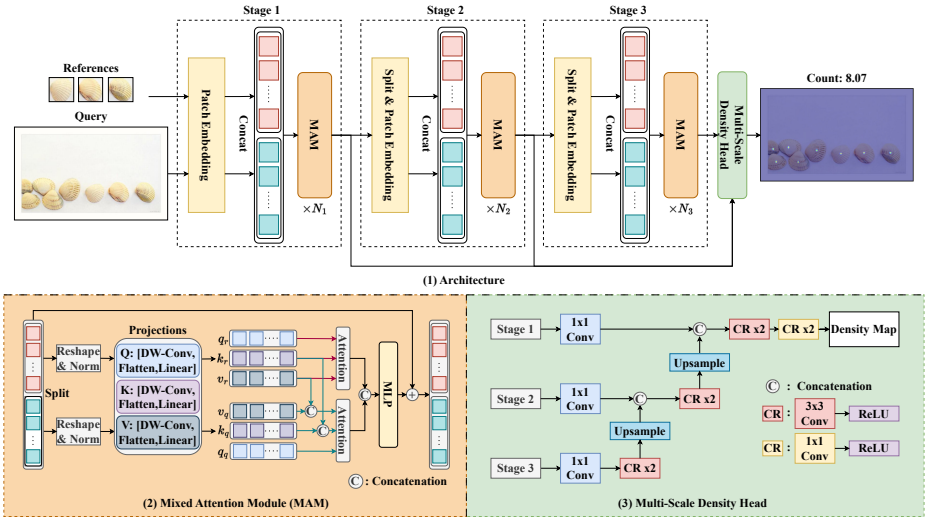

Fig. D: MGCAC Architecture. We employ a multi-stage extract-and-match module from [2] and a density head as an additional strong CAC baseline candidate.

4 More Evaluation Results

4.1 Qualitative Comparisons on Mosaic Dataset

In Figure C, we also show qualitative comparisons of our MGCAC and recent SOTA CAC models (e.g. LOCA and CounTR) on the proposed mosaic dataset. LOCA [4] fails to distinguish the corresponding objects and localize all objects in the queries. CounTR [1] either captures wrong objects (row 1, 2, and 4) or gives inaccurate counts (row 3). In contrast, our MGCAC effectively localizes the right objects as the one provided in the reference and gives accurate count predictions. The results are consistent with Figure 2 in the main paper, which utilized real image inputs. Therefore, our proposed mosaic dataset can effectively reflect the performance of the model under multi-class scenarios of the real world, enabling a more comprehensive evaluation of CAC models.

4.2 Cross-Dataset Generalization.

A strong CAC model should excel in counting objects on scenarios that differ from the training dataset in terms of perspectives, scales, illumination, etc. To evaluate the ability to generalize on other datasets, we train the proposed method on the FSC-147 dataset and then evaluate on the CARPK testing set without the use of fine-tuning on the CARPK training set. Specifically, since the CARPK dataset mainly consists of car images, we remove the car category from the FSC-147 dataset during the training stage. In Table A, our models outperform the non-fine-tuned baselines, which demonstrates MGCAC's generalizability in diverse counting scenarios.

4.3 Impact of Number of References Images (i.e. n)

We investigate our model performance in relation to the number of reference images, where $n \leq 3$. In Table B, the results indicate that using our training protocol, our robust model architecture effectively captures reference representative features and improves performance as the number of reference images increases.

Table A: Cross-dataset evaluation results compared with other SOTA models where 'fine-tuned' means the model is further fine-tuned using the CARPK training data.

Model	Fine-tuned	MAE	RMSE
BMNet [3]	✓	8.05	9.7
BMNet+ [3]	✓	5.76	7.83
CounTR [1]	✓	**5.75**	**7.45**
BMNet [3]	×	14.61	24.60
BMNet+ [3]	×	10.44	13.77
LOCA [4]	×	9.97	12.51
MGCAC (Ours)	×	**6.41**	**8.81**

Table B: Evaluation results of MGCAC using different numbers (i.e. n) of reference images on FSC-147.

n	VAL				TEST			
	MAE	RMSE	NAE	SRE	MAE	RMSE	NAE	SRE
1	17.78	68.9	0.19	3.05	14.55	111.04	0.16	2.79
2	12.80	58.83	0.13	2.34	11.03	107.09	**0.15**	**4.24**
3	**11.00**	**51.42**	**0.12**	**2.05**	**10.46**	**96.60**	0.16	6.17

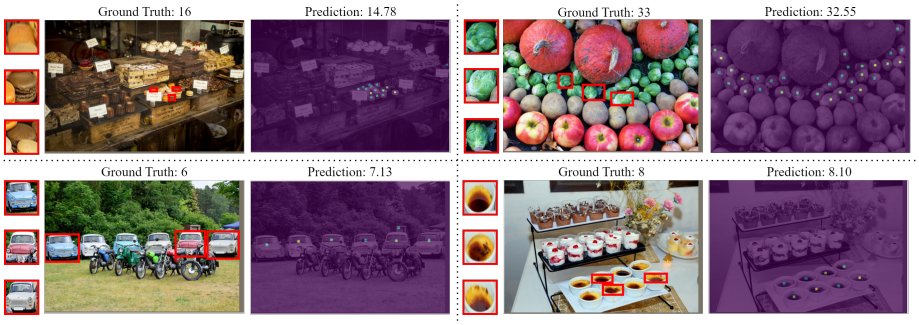

Fig. E: More real-world examples of MGCAC. The references are annotated by bounding boxes.

References

1. Chang, L., Yujie, Z., Andrew, Z., Weidi, X.: Countr: Transformer-based generalised visual counting. In: British Machine Vision Conference (BMVC) (2022)
2. Cui, Y., Jiang, C., Wang, L., Wu, G.: Mixformer: End-to-end tracking with iterative mixed attention. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13608–13618 (2022)
3. Shi, M., Lu, H., Feng, C., Liu, C., Cao, Z.: Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9529–9538 (2022)
4. Đukić, N., Lukežič, A., Zavrtanik, V., Kristan, M.: A low-shot object counting network with iterative prototype adaptation. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 18872–18881 (October 2023)