Supplementary Material: CounTR: Transformer-based Generalised Visual Counting

Chang Liu[1]
liuchang666@sjtu.edu.cn
Yujie Zhong[2]
jaszhong@hotmail.com
Andrew Zisserman[3]
az@robots.ox.ac.uk
Weidi Xie[1]
weidi@sjtu.edu.cn

[1] Coop. Medianet Innovation Center, Shanghai Jiao Tong University, China
[2] Meituan Inc., China
[3] Visual Geometry Group (VGG), Department of Engineering Science University of Oxford, UK

In the supplementary material, we first give additional experiment results on Val-COCO, Test-COCO [7], and CARPK [5] in Section 1; then we show the qualitative visualisation of CounTR's results on the FSC-147 dataset in Section 2; and introduce the annotation error of 7171.jpg in the FSC-147 test set in Section 3.

1 Additional Experiments

In this section, we further evaluate the model on several other datasets: Val-COCO, Test-COCO, and CARPK.

Methods	Year	Method	Val-COCO		Test-COCO	
			MAE	RMSE	MAE	RMSE
Faster-RCNN [9]	NIPS2015	Detection	52.79	172.46	36.20	79.59
RetinaNet [4]	ICCV2017	Detection	63.57	174.36	52.67	85.86
Mask-RCNN [2]	ICCV2017	Detection	52.51	172.21	35.56	80.00
FamNet [7]	CVPR2021	Regression	39.82	108.13	22.76	45.92
CounTR (ours)	2022	Regression	**24.66**	**83.84**	**10.89**	**31.11**

Table 1. Comparison with state-of-the-art on the FSC-147 subsets.

Val-COCO and Test-COCO. Val-COCO and Test-COCO [7] are FSC-147 subsets collected from COCO, and they are often used as evaluation benchmarks for detection-based object counting models. Here we compare our CounTR model with several counting models based on detection, including: Faster-RCNN [9], RetinaNet [4], and Mask-RCNN [2]. As shown in Table 1, our model has a huge improvement even compared to the best-performing Mask-RCNN [2], halving its error on both Val-COCO and Test-COCO. We also compared our model with the few-shot counting SOTA method FamNet [7], and our model outper-

forms it significantly (15.16 MAE and 24.29 RMSE on Val-COCO and 11.87 MAE and 14.81 RMSE on Test-COCO), which demonstrates the superiority of our model.

CARPK. CARPK [3] is a class-specific car counting benchmark with 1448 images of parking lots from a bird's view. We also fine-tuned our model on the CARPK train set and test on it with Non-Maximum Suppression (NMS). We compared our CounTR model with several detection-based object counting models and regression-based few-shot counting models. As shown in Table 2, even compared with the existing class-specific counting models, *i.e.*, the models that can only count cars, our CounTR still shows comparable performance.

Methods	Year	Method	Type	CARPK	
				MAE	RMSE
YOLO [3]	CVPR2016	Detection	Generic	48.89	57.55
Faster-RCNN [9]	NIPS2015	Detection	Generic	47.45	57.39
S-RPN [3]	ICCV2017	Detection	Generic	24.32	37.62
RetinaNet [4]	ICCV2017	Detection	Generic	16.62	22.30
LPN [3]	ICCV2017	Detection	Generic	23.80	36.79
One Look [3]	ECCV2016	Detection	Specific	59.46	66.84
IEP Count [11]	TIP2018	Detection	Specific	51.83	-
PDEM [1]	CVPR2019	Detection	Specific	6.77	8.52
GMN [5]	CVPR2021	Regression	Generic	7.48	9.90
FamNet [3]	CVPR2021	Regression	Generic	18.19	33.66
BMNet+ [11]	CVPR2022	Regression	Generic	5.76	7.83
CounTR (ours)	2022	Regression	Generic	**5.75**	**7.45**

Table 2. Comparison with state-of-the-art on the CARPK dataset.

2 Qualitative Results

We show qualitative results from our few-shot counting setting in Figure 2. As we can see from the first five images from FSC-147, our model can easily count the objects' numbers and locate their position. In the last image, the model mistakenly chose the smallest self-similarity unit of the spectacle lenses instead of the sunglasses for counting due to the ambiguity of the bounding boxes, which can be corrected by test-time normalisation.

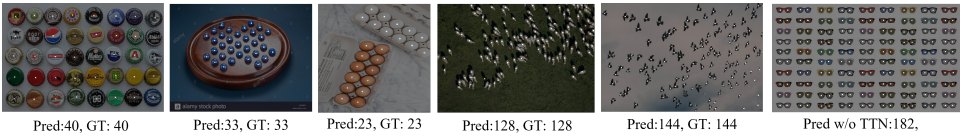


Figure 1. Qualitative results of CounTR on FSC-147. For visualisation purpose, we have overlaid the predited density map on the original image. TTN stands for test-time augmentation.

3 On the potential annotation error

We discover that the 7171.jpg in the FSC147 dataset maybe annotated wrongly, the annotated object count is inconsistent with the given exemplars, which tends to lead to a significant

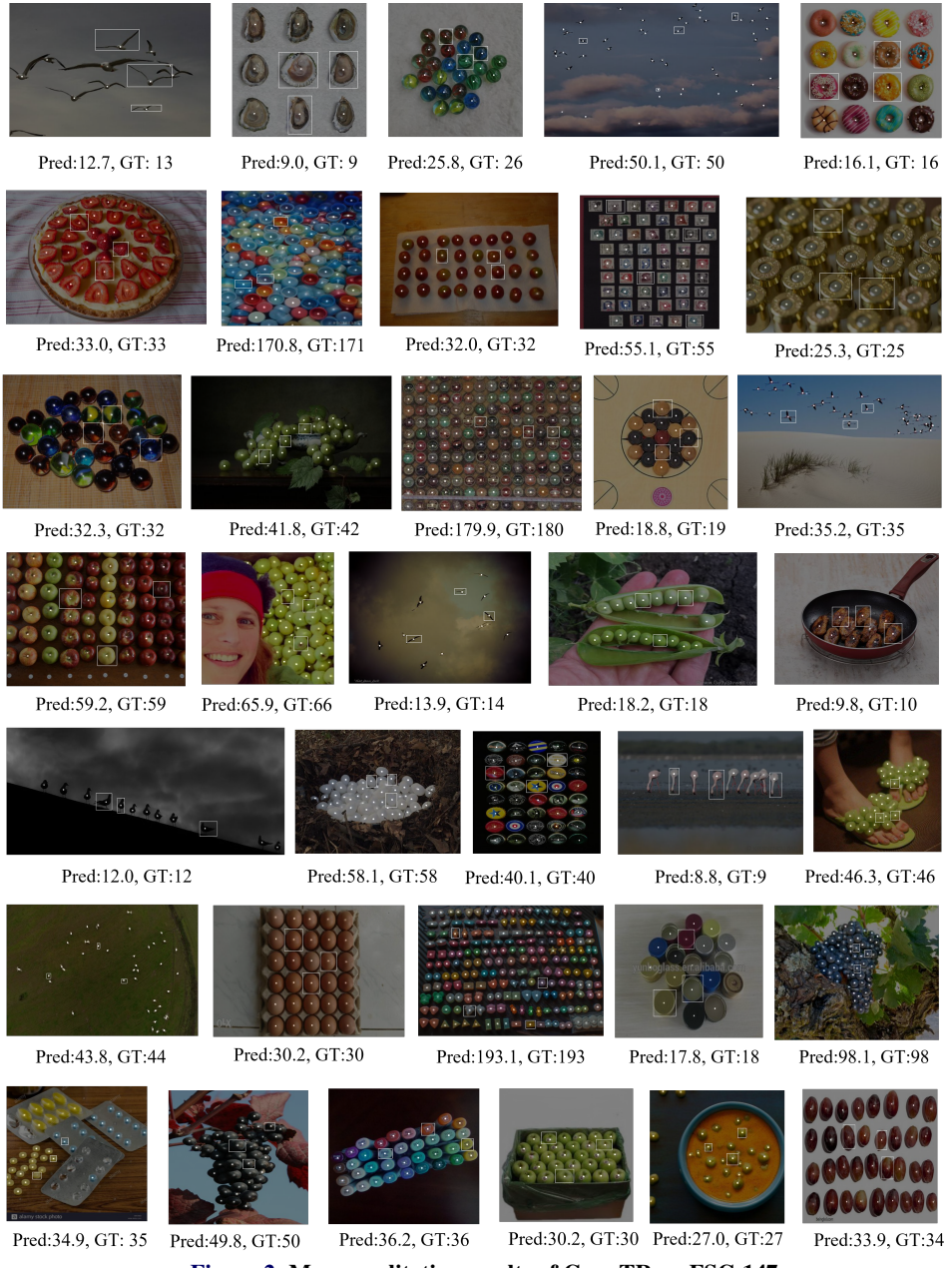
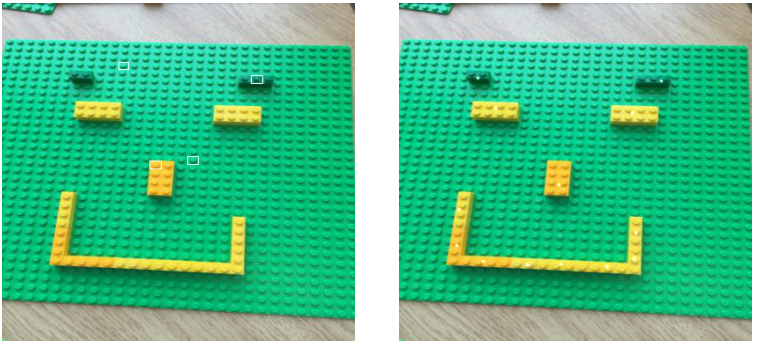


Figure 2. More qualitative results of CounTR on FSC-147.

error during evaluation. The ground truth annotation and exemplar annotation are shown in Figure 3.



(a) The exemplar annotation of 7171.jpg. (b) The ground truth annotation is 14.

Figure 3. The ground truth annotation and exemplar annotation of 7171.jpg, and we can easily figure out the inconsistency.

References

[1] Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2017.

[3] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, 2017.

[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2017.

[5] Erika Lu, Weidi Xie, and Andrew Zisserman. Class-agnostic counting. In *Asian Conference on Computer Vision*, 2018.

[6] T Nathan Mundhenk, Goran Konjevod, Wesam A Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *European conference on computer vision*. Springer, 2016.

[7] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[8] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 2015.

[10] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhi-Guo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[11] Tobias Stahl, Silvia L Pintea, and Jan C Van Gemert. Divide and count: Generic object counting by image divisions. *IEEE Transactions on Image Processing*, 2018.