Semantic Generative Augmentations for Few-Shot Counting Supplementary Material

1. Benchmark Comparison

In Tab. 3, we present the benchmark results for Few-shot Counting on FSC147 [4]. When applied, our augmentation method allows to improve the performances of the models.

	Val		Test	
	MAE	RMSE	MAE	RMSE
BMNet+ (CVPR'22) [5]	15.74	58.53	14.62	91.83
SAFECount (WACV'23) [6]	15.28	47.20	14.32	85.54
CounTR (BMVC'22) [1]	13.13	49.83	11.95	91.23
LOCA* (arXiv'23) [2]	10.24	32.56	10.79	56.97
SAFECount (Ours)	12.59	44.95	12.74	89.90
CounTR (Ours)	12.31	49.47	11.32	77.50

Table 3. Benchmark for 3-shot counting on FSC147. $(*)$ We do not apply our approach to LOCA as the code was not available.

2. Result Reproduction

We give more details on the reproduction of the results for SAFECount and CounTR trained solely on real images from FSC147. Reproductions are based on the official implementations[1] [2].

SAFECount First, we reproduce the model using the same parameters and number of epochs. As shown in Tab. 4, we obtain similar results to the ones reported in the authors' github. The models trained on our synthetic data are trained for more epochs to account for the larger number of training images. Specifically, we increase the number of training epochs from 200 to 300. Consequently, we modify the learning rate schedule. We use a learning rate of $2e^{-5}$ reduced by 0.25 every 160 epochs. For a fair comparison with these models, we also train the model without synthetic augmentations in the same setting. We find that we achieve slightly better results than the reported ones for 200 epochs.

[1]SAFECount: github.com/zhiyuanyou/SAFECount
[2]CounTR: github.com/Verg-Avesta/CounTR

	Val		Test	
	MAE	RMSE	MAE	RMSE
SAFECount (paper)	15.28	47.20	14.32	91.30
SAFECount (github)	14.42	51.72	13.56	91.30
SAFECount (repro†)	14.48	54.80	13.86	91.71
SAFECount (repro‡)	13.95	51.73	13.73	91.85
CounTR (paper)	13.13	49.83	11.95	91.23
CounTR (repro†)	14.45	51.28	13.03	91.89
CounTR (repro‡)	14.25	50.15	13.13	88.21

Table 4. Results reproduction for SAFECount and CounTR trained only on real images from FSC147. (†) same parameters and number of epochs, (‡) same parameters with more epochs (see text).

CounTR On CounTR, the reproduction with the same parameters and number of epochs gives slightly different results than the reported ones (cf. Tab. 4). In a github issue[3] the author indicates that he has kept the model that gave the best results and that an MAE between 12 and 13 should be expected for his model. The models trained with synthetic augmentations are trained for 1200 epochs instead of 1000 epochs. We employ the same cyclic learning rate scheduler with $1e^{-5}$ as the initial learning rate. CounTR is also used for zero-shot counting (counting with no exemplars). In the official implementation, the number of shots is randomly chosen between 0 and 3 during training. We focus on the 3-shot case, thus we fixed the number of shots to 3. In Sec. 3.1 we report results for the model trained in the various shot settings.

3. More Quantitative Results

3.1. Counting with Fewer Shots

SAFECount SAFECount is a 3-shot counting method but it can generalize to 1-shot counting without retraining [6]. We evaluate our model trained on diversified synthetic augmentations in the 1-shot case. As shown in Tab. 5, our model also exhibits good performances in the 1-shot set-

[3]https://github.com/Verg-Avesta/CounTR/issues/26

	3-shot				*1-shot*			
	Val		Test		Val		Test	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Traditional Augmentation	13.95	51.73	13.73	91.85	19.92	67.63	18.08	**104.32**
+ Diverse Generation (Ours)	**12.59**	**44.95**	**12.74**	**89.90**	**18.55**	**61.22**	**17.60**	106.47

Table 5. Quantitative results: 3-shot and 1-shot evaluation for SAFECount [6] on FSC147.

	3-shot				*0-shot*			
	Val		Test		Val		Test	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Traditional Augmentation	14.25	50.15	13.13	88.21	-	-	-	-
+ Diverse Generation (Ours)	12.31	49.47	11.32	77.50	-	-	-	-
Traditional Augmentation	14.61	51.33	13.24	94.01	**18.82**	**69.64**	17.51	122.01
+ Diverse Generation (Ours)	**13.09**	**48.29**	**11.59**	**83.23**	18.84	69.90	**15.70**	**112.25**

Table 6. Quantitative results: 3-shot and 0-shot evaluation for CounTR [1] on FSC147. Top: 3-shot training. Bottom: [0,3]-shots training.

ting, outperforming the traditional augmentation model.

CounTR We retrained the models with a number of shots randomly chosen between 0 and 3. In Tab. 6, we evaluate the models trained with and without synthetic augmentations in both the 3-shot and 0-shot settings. We observe a small degradation of the performances for both models in the 3-shot case in comparison with the models trained with 3 shots fixed (first two lines). The performances of our model nevertheless remain higher. In the 0-shot case, we find that both models perform similarly on the validation test, while our model is significantly better on the test set.

3.2. Counting Accuracy per Object Count

In Tab. 7, we report the test MAE and RMSE per range of object counts. We compare our model with the model trained without synthetic data (Traditional Augmentation). Our model increases the counting performances for all ranges. We also compare with the Real Guidance [3] synthetic augmentations. We outperform Real Guidance for all ranges except for the range with a very high number of objects ($[301, 3701]$). As shown in Tab. 7, this range of object counts contains few images (they represent 1% of the total number of test images) but they dominate the global RMSE (80.69 for Real Guidance and 89.90 for our model). In particular, there are 2 outliers images with respectively 2560 and 3701 objects. In comparison, the maximum number of objects in the training set is 1912 objects. Real Guidance performs better on one of these outlier images as shown in the first row of Fig. 12. For other images with a high object count, both models are on par (last two rows of Fig. 12).

4. More Qualitative Results

4.1. Synthetic Augmentations

In Figs. 13 and 14, we show additional qualitative results of our synthetic augmentations. We observe that our diverse strategy leads to modifying the semantics, size or shape of the objects as well as the background and sometimes the viewpoint. These modifications allow to expose the network to unseen data that can improve generalization.

4.2. Counting Results

In Fig. 15, we show some qualitative results of the density maps predicted by the models. In the illustrated cases, our model seems to produce fewer false positives for a low number of objects (first two rows) and fewer false negatives for a high number of objects (last three rows).

References

[1] Liu Chang, Zhong Yujie, Zisserman Andrew, and Xie Weidi. Countr: Transformer-based generalised visual counting. In *British Machine Vision Conference (BMVC)*, 2022. 1, 2

[2] Nikola Djukic, Alan Lukezic, Vitjan Zavrtanik, and Matej Kristan. A low-shot object counting network with iterative prototype adaptation. *arXiv preprint arXiv:2211.08217*, 2022. 1

[3] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 2, 3

[4] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403, 2021. 1

Nb. Objects	Nb. Images	Traditional Augmentation		+ Diverse Generation (Ours)		+ Real Guidance [3]	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
$[1, 10]$	60	4.76	10.22	3.75	**6.42**	**3.24**	7.47
$[11, 20]$	268	6.27	47.81	**6.05**	**45.5**	6.22	50.3
$[21, 50]$	413	6.61	12.22	**5.38**	**10.21**	6.05	11.47
$[51, 100]$	254	12.14	18.26	**10.29**	**15.84**	12.30	18.39
$[101, 300]$	172	23.58	35.47	**21.65**	**29.91**	26.08	38.03
$[301, 3760]$	23	197.18	637.04	205.4	604.42	**186.38**	**538.68**

Table 7. Quantitative results: Test counting accuracy (3-shot) per range of number of objects for SAFECount [6] on FSC147.

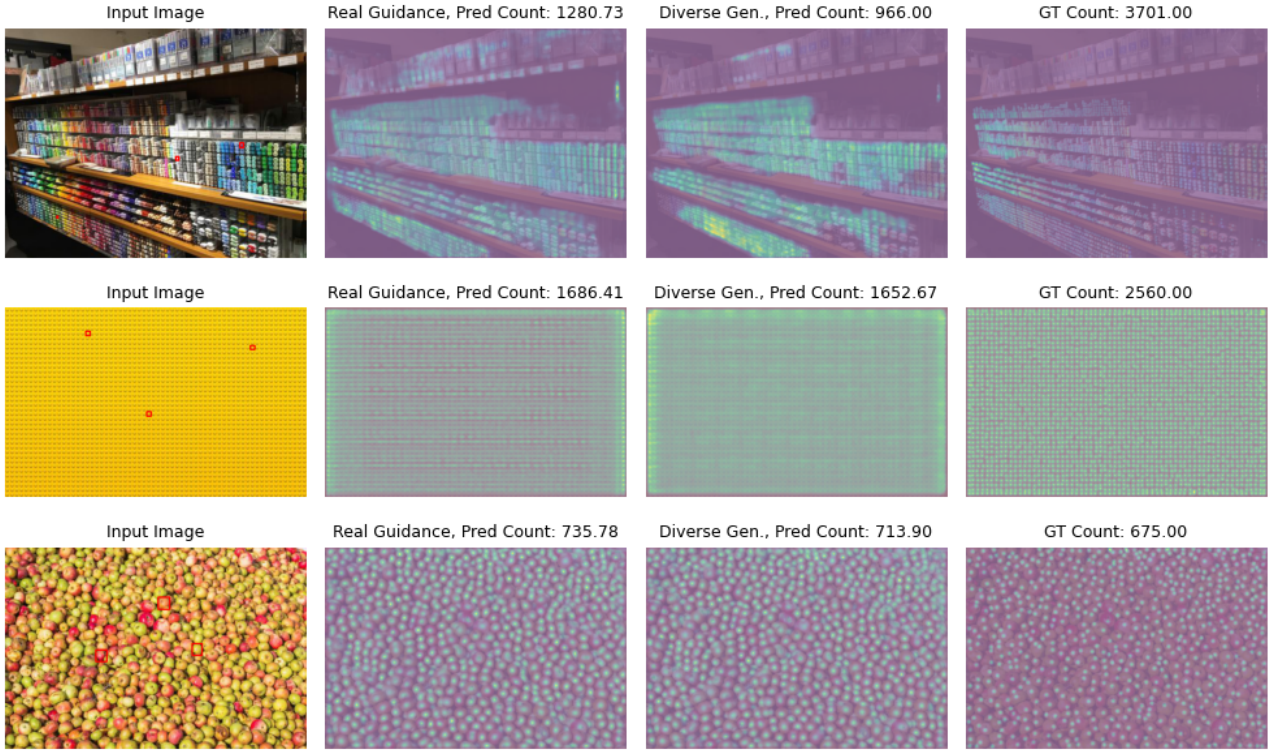

Figure 12. Qualitative counting results on FSC147 test images. We compare the model trained with Real Guidance's augmentations (2^{nd} column) vs. our augmentations (3^{rd} column) for images with a high number of objects. Predicted and ground-truth density maps are overlapped with the images.

[5] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhiguo 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*, pages 9529–9538, 2022. 1

[6] Zhiyuan You, Kai Yang, Wenhan Luo, Xin Lu, Lei Cui, and Xinyi Le. Few-shot object counting with similarity-aware feature enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6315–6324, January 2023. 1, 2, 3

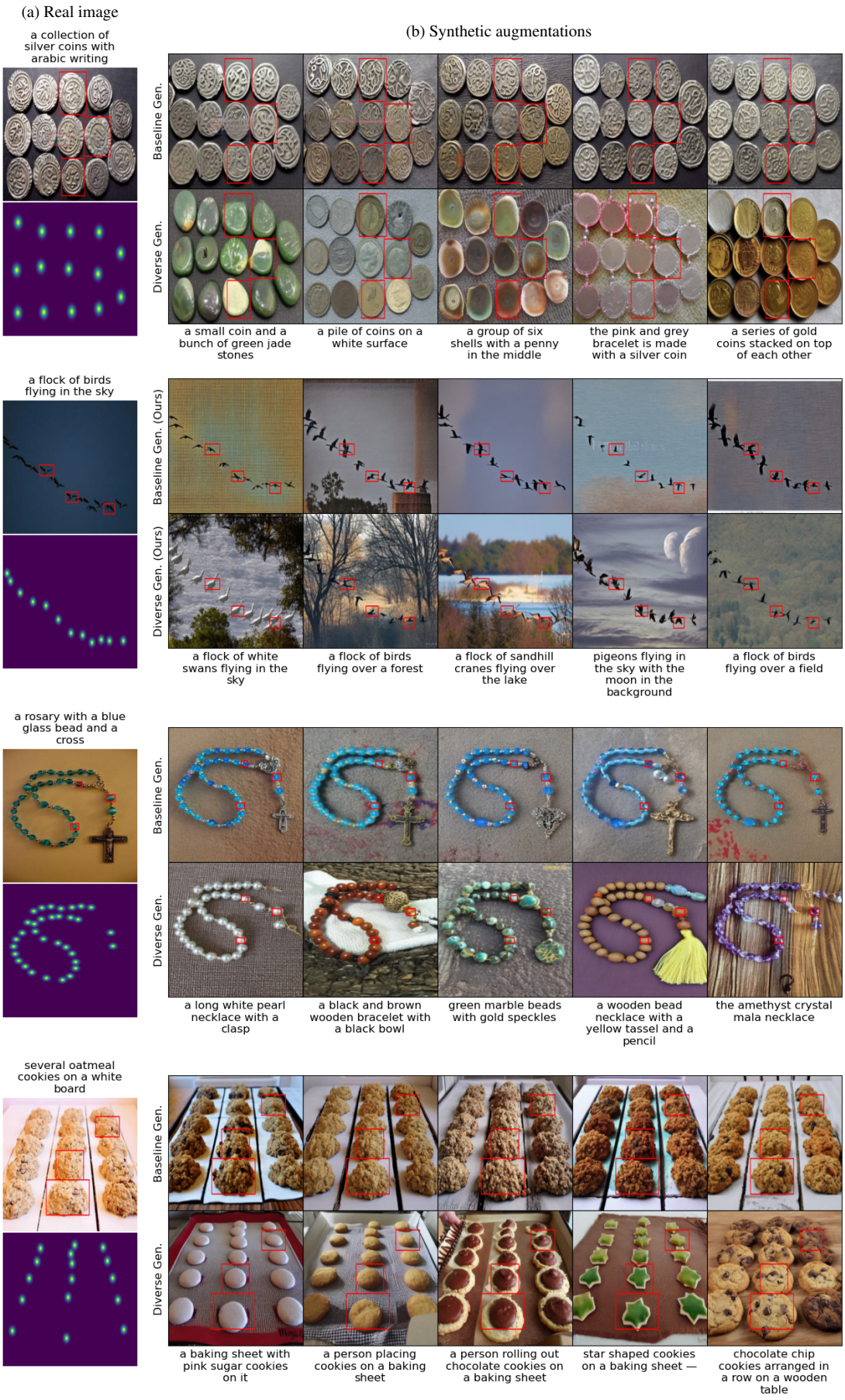

Figure 13. Qualitative results of synthetic augmentations of FSC147. We compare our Baseline vs. Diverse augmentations.

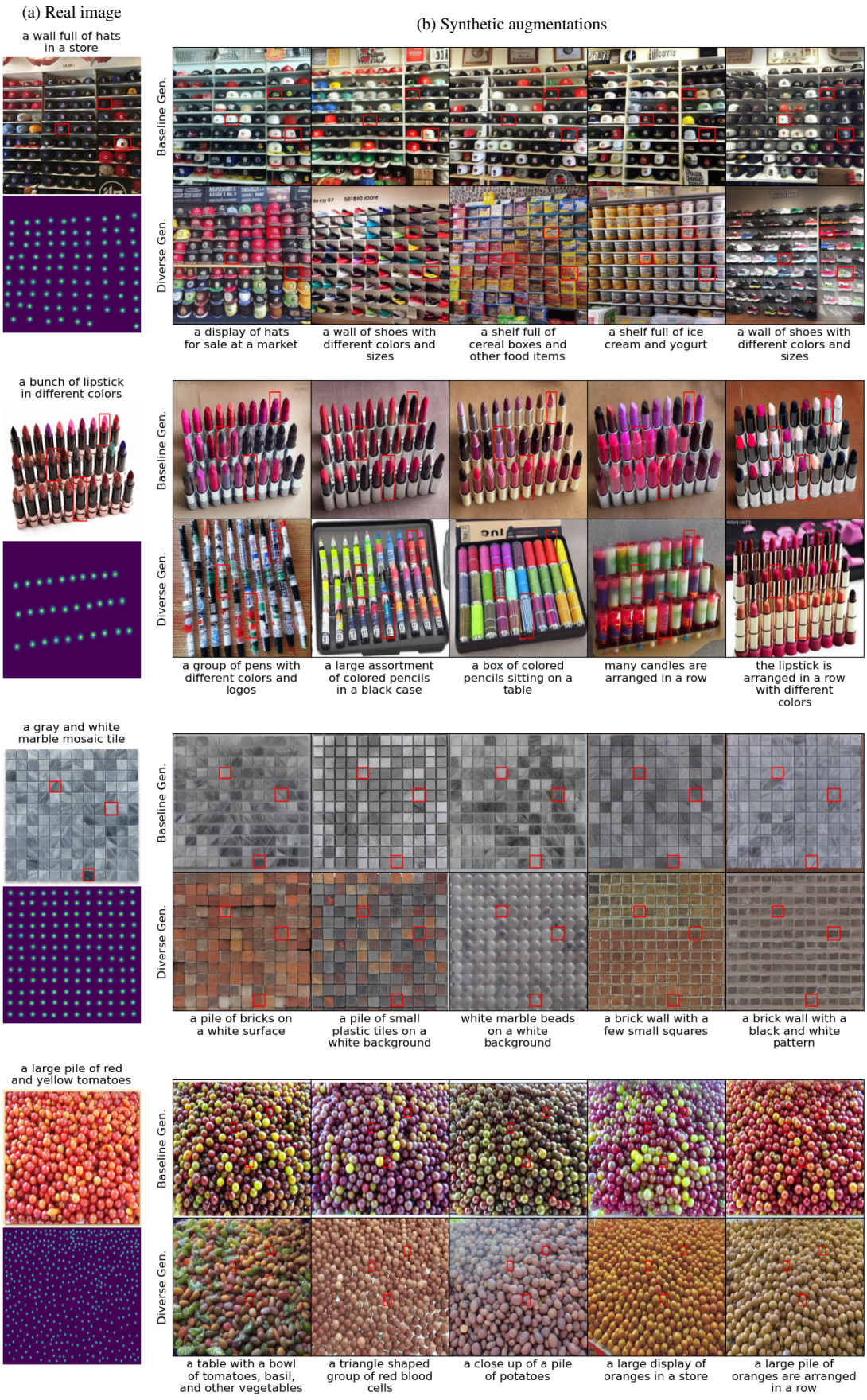

Figure 14. Qualitative results of synthetic augmentations of FSC147. We compare our Baseline vs. Diverse augmentations.

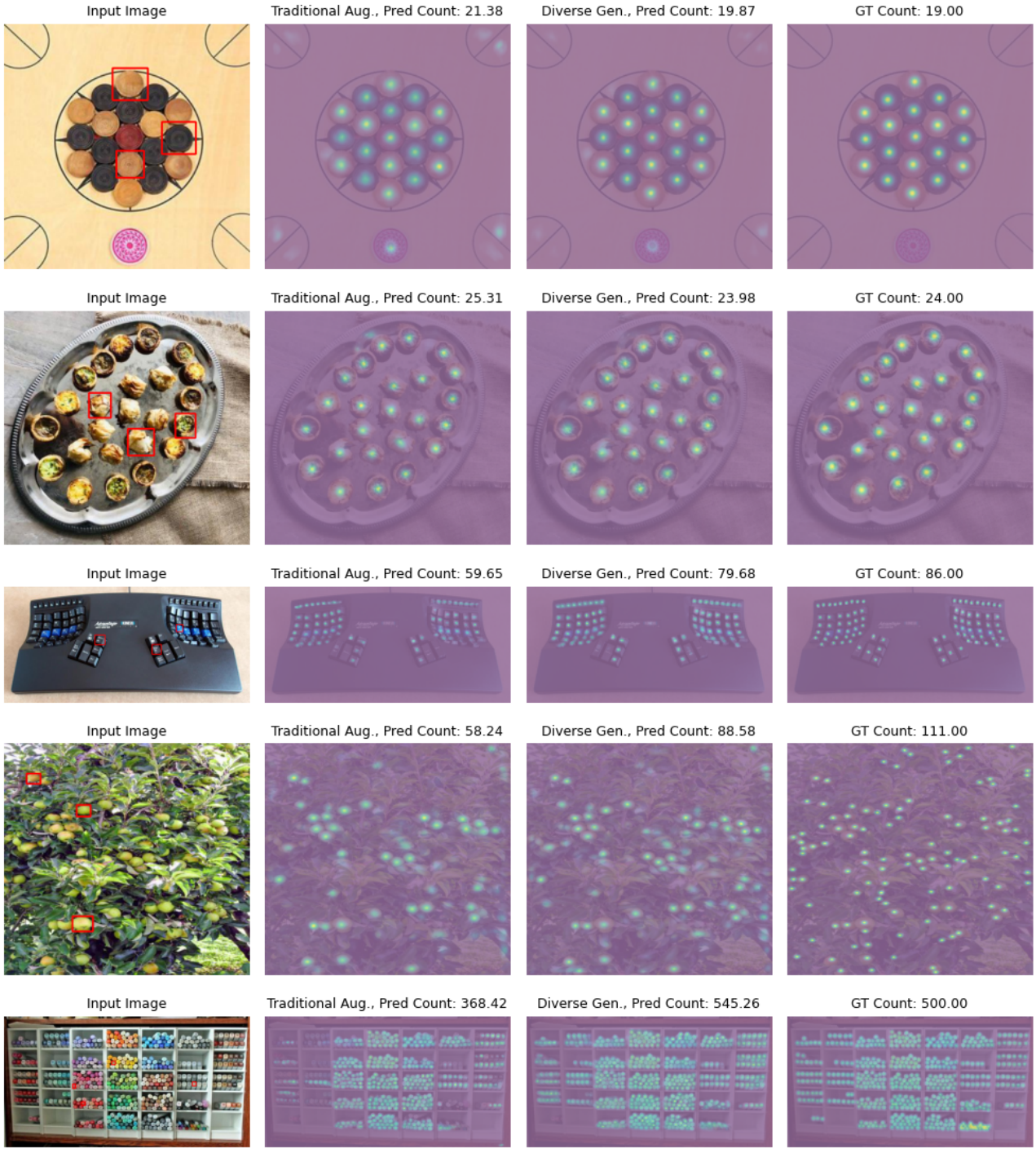

Figure 15. Qualitative counting results on FSC147 test images. We compare the model trained without synthetic augmentations (2nd column) vs. with our augmentations (3rd column). Predicted and respectively ground-truth density maps are overlapped to the images.