AFreeCA: Annotation-Free Counting for All
Supplementary Materials

1 Evaluating Synthetic Dataset Quality

Our goal is to assess the quality of the synthetic datasets we use to train our unsupervised counting model. This section provides a thorough analysis of these datasets, aiming to deepen the understanding of our method's strengths and limitations.

1.1 Synthetic Counting Data

Understanding Prompt Label Noise. In the beginning of our paper, we introduced the concept of label noise when using a latent diffusion model (LDM), such as Stable Diffusion, to generate images with a specified number of objects. Often, the actual number of objects in these images doesn't align with the requested counts. To understand this discrepancy better, we manually annotated 40 synthetic examples for counts ranging from 1 to 40. Our in-depth analysis, shown in Tab. 1, highlights how the error between the requested and actual counts grows with the increase in the prompt count. Although these findings may not be universally applicable across all ranges or object categories, they provide useful insights into how synthetic counting data affected by label noise can still be valuable for learning. Notably, we observed that the average for the true underlying counts is often close to the requested count, despite significant variations. Often, the average count for these distributions is within 15% of the prompt label, indicating that Stable Diffusion produces a distribution of images with true counts centered near the desired amounts

Table 1: Prompt Count Noise. we manually annotate 40 synthetic examples for each of several prompt count categories. We evaluate the statistics of the true underlying counts for each prompt count category.

Prompt Count	Mean	Std.	MAE	rMAE
1	1.02	0.34	0.07	0.07
5	4.26	0.79	0.88	0.18
10	10.0	3.48	2.57	0.26
15	14.86	5.73	4.86	0.32
20	23.07	11.10	8.98	0.45
40	49.29	39.00	24.38	0.61

Table 2: Outlier Removal Ablation Study. We explore the performance implications of filtering likely outliers in the noisy synthetic counting dataset.

	SHB		JHU		SHA		QNRF	
Strategy	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
No Outlier Removal	48.2	67.2	189.0	582.4	188.3	292.7	366.8	558.3
Outlier Removal	**35.0**	**50.7**	**173.8**	**519.4**	**152.7**	**219.0**	**283.1**	**453.2**

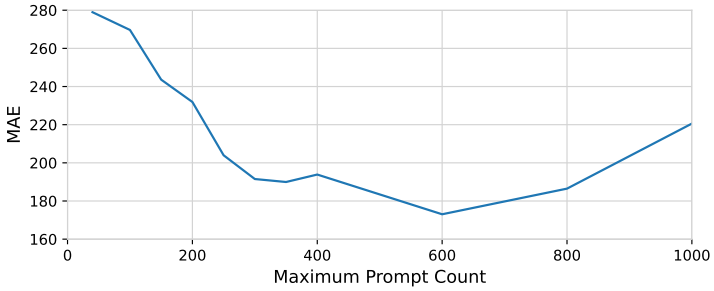

Fig. 1: Impact of Maximum Prompt Count. We overview the performance impact of changing the maximum prompt count used when training the counting network. Our method is remains accurate across a range of values.

Impact of Outliers. Within our methodology, we introduced a straightforward outlier removal technique to help reduce label noise. Tab. 2 evaluates the effectiveness of this method in the context of crowd counting. The results indicate that this basic form of outlier removal significantly boosts performance across different crowd counting datasets, highlighting its importance in enhancing model accuracy. These results suggest that label noise hampers performance, especially in datasets with larger average counts, such as SHA and QNRF. Nonetheless, our analysis also indicates that identifying and mitigating this noise is feasible, and further investigation into noise reduction strategies could be highly advantageous.

Impact of the Maximum Prompt Count. In our approach to generating noisy synthetic counting data, we employ a wide range of prompt counts to create a rich dataset. This naturally raises the question of whether the range of counts used during training influences the training and performance of our counting network. Specifically, in Fig. 1, we explore the effect of setting different maximum prompt counts (c_{max}) on the model's accuracy using the JHU++ test set. This involves using all synthetic counting images with a prompt count equal to or lower than c_{max}. Our analysis reveals an optimal range for this maximum count—specifically, between 250 and 800 for the JHU dataset. This optimal range underscores our method's adaptability, showing it can handle a wide variety of scenarios effectively.

Fig. 2: Synthetic Crowd Count Qualitative. We highlight samples of synthetic counting images from a wide range of prompt count categories. We demonstrate that images contain realistic crowds which are organized in a natural way.

Qualitative. In Fig. 2, we provide synthetic counting samples for a crowd counting problem across a wide range of crowd densities. Despite the presence of label noise, synthetic images remain reasonably consistent within expected ranges, even at high prompt counts. For instance, images prompted with a count of 1000 exhibit significant noise, often featuring several thousand objects. Nonetheless, these images consistently depict large, dense crowds, suggesting that Stable Diffusion retains an understanding of quantity to some extent, even at these higher numbers. This observation aligns with our findings in Sec. 1.1 and Tab. 1, where the average true count if often similar to the prompted count.

Moreover, our analysis sheds light on scene biases within the synthetic counting images. Predominantly, the images showcase clear, daylight settings, frequently set in expansive outdoor areas with greenery. Despite this tendency, the dataset also exhibits variety; some images mimic historical photographs, while others depict indoor scenes, such as concert halls, indicating a diversity in the visual contexts of the generated crowds.

1.2 Synthetic Sorting Data

Removal Accuracy. To substantiate the reliability of the synthetic sorting data, we examine 50 object removal examples from each dataset. We only analyze

Table 3: Object Removal. We manually annotate 50 synthetic sorting examples from each dataset to determine how frequently Stable Diffusion successfully removes objects from a reference image. We corroborate this by using a fully-supervised crowd counting model (DM-Count [8]) to estimate the synthetic image counts for *all* examples.

Estimate Type	SHB	JHU	SHA	QNRF
Manual	100.0%	96.0%	98.0%	96.0%
DM-Count [8]	99.6%	90.2%	99.2%	97.2%

Fig. 3: Synthetic Object Count Qualitative. Stable Diffusion can produce synthetic counting data for a wide range of objects, including sheep and penguins.

object removal, due to the fact that outpainting inherently preserves original objects, and thus these images always have at least as many objects as the reference image. This inspection revealed minimal discrepancies: ShanghaiTechB produced no incorrectly ranked examples, while ShanghaiTechA exhibited only one. QNRF and JHU++ presented slightly higher instances of two incorrectly ranked examples each.

Moreover, to corroborate our manual assessment, we employed a fully supervised DM-COUNT [8] model to evaluate crowd count estimations across 500 real and synthetic removal pairs from each dataset. The resultant accuracy rates were high: 99.6% for ShanghaiTechB, 99.2% for ShanghaiTechA, 97.2% for QNRF, and 90.2% for JHU++. However, these results are dependent on the accuracy of the model, and are only meant to compliment the manually collected annotations. These findings, summarized in Tab. 3, affirm the credibility of the synthetic object removal strategy used during the data generation process.

1.3 Synthetic Density Classification Data

We observe that Stable Diffusion excels in creating density classification data, which plays a crucial role in training our density classification network. In Fig. 4, we showcase synthetic images crafted specifically for crowd counting scenarios. These examples demonstrate Stable Diffusion's capability to generate diverse and accurate representations for such tasks.

Fig. 5 presents the classification maps generated by our network, trained on this synthetic data. The maps reveal the network's proficiency in accurately identifying areas with dense and sparse crowds, as well as empty spaces devoid

Fig. 4: Synthetic Density Qualitative. Stable Diffusion excels in producing diverse and high-quality images across various density categories, encompassing dense, sparse, and empty scenes with broad coverage.

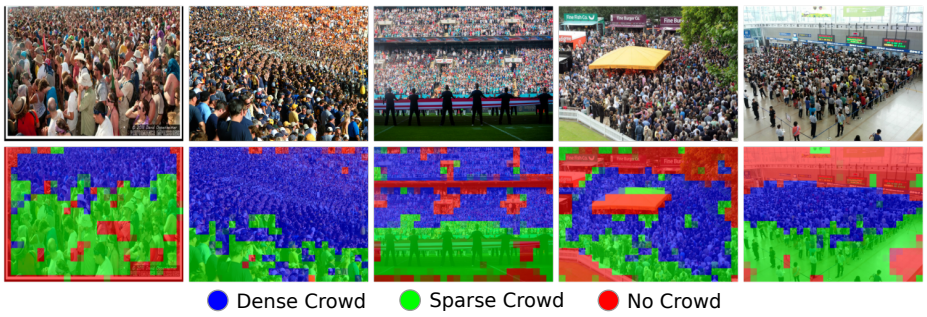

Fig. 5: Density Class Maps. We overview the class maps generated by passing the sorting network features through the density classifier. Our method does not utilize any localization information, and yet it accurately localizes crowded regions within images.

of pedestrians. This underscores the significant impact and utility of utilizing synthetic data for density classification

2 Expanded Qualitative Analysis

Crowd Counting. In Fig. 6, we delve deeper into the crowd counting challenge with an expanded qualitative analysis. This section sheds light on our method's process, which integrates whole image estimates with those from densely populated patches identified within the image. We illustrate this approach by presenting the initial count map estimate, $C^{(0)}$, for the entire image, alongside $C^{(1)}$, which represents the concatenated estimates for partitioned patches from a 3×3 grid. Additionally, we show the combined count map resulting from our density classifier guided partitioning technique. These examples are valuable for understanding how our model calculates various estimates to achieve a precise overall

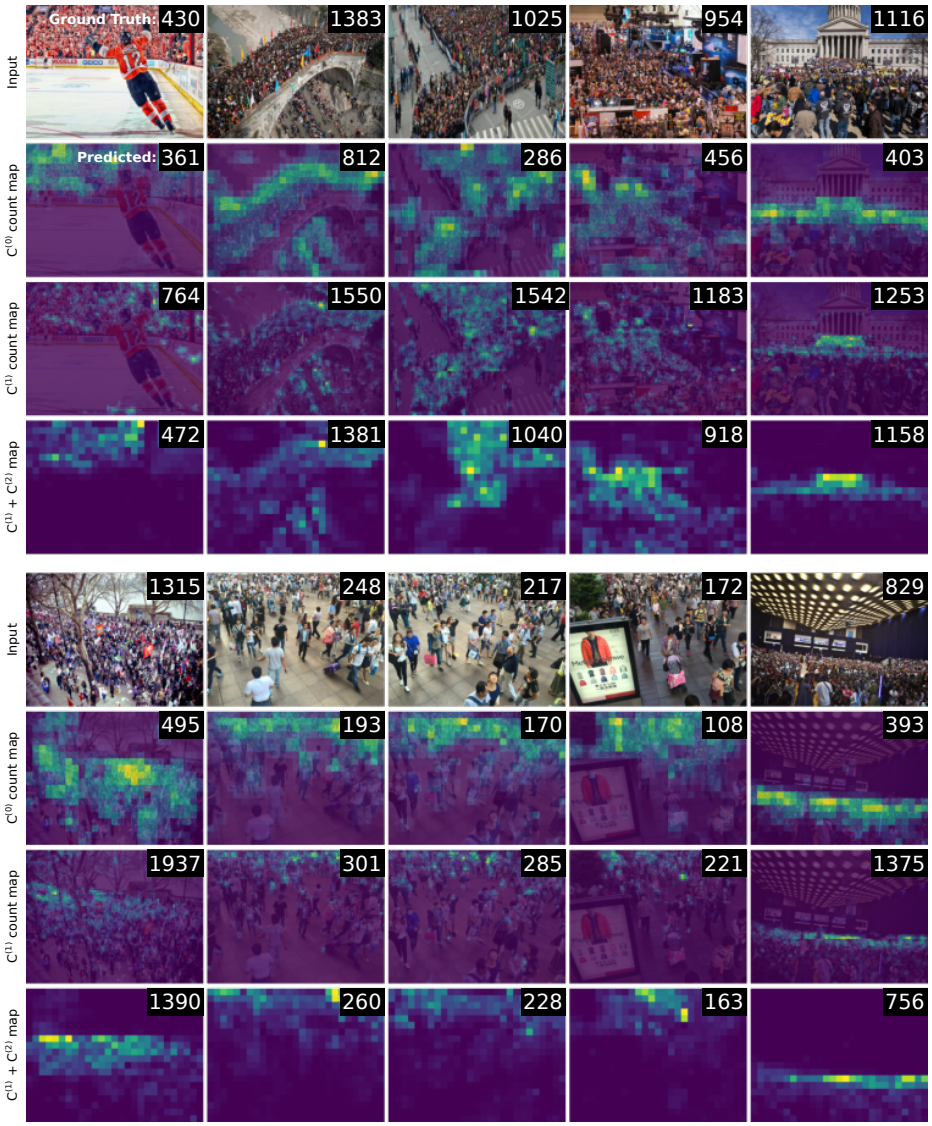

Fig. 6: Crowd Counting Qualitative. We explore the quality of the model output for the $c^{(0)}$, $c^{(1)}$, and density guided count maps which join $c^{(0)}$, $c^{(1)}$. Input images are annotated with the ground truth count, and count maps are annotated with the estimated count for that map.

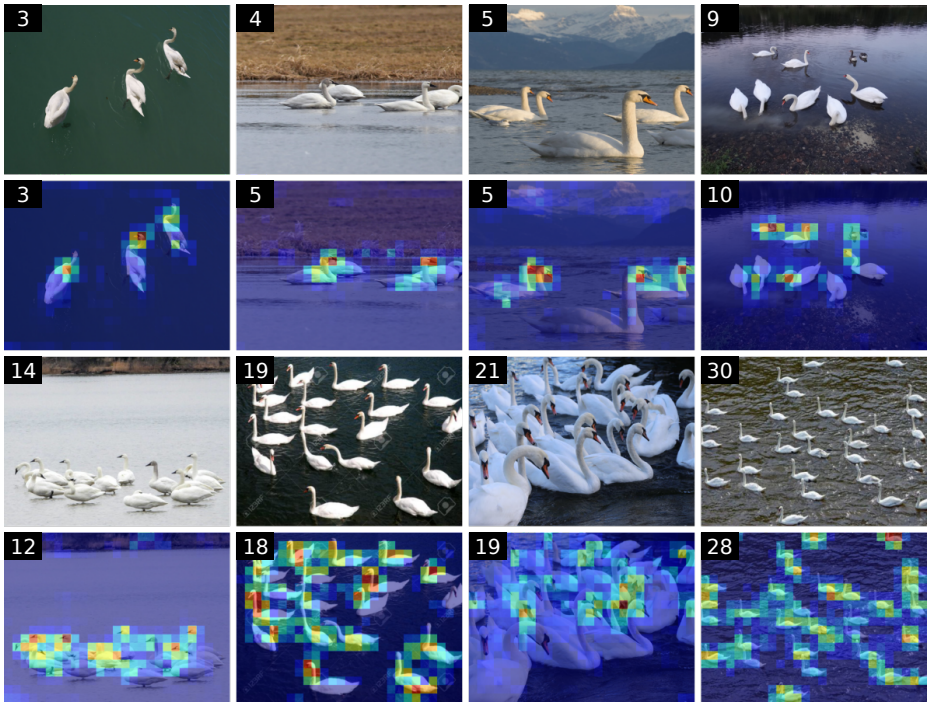

Fig. 7: Extension to Similar Objects. We investigate an emergent capability of our system to generalize to similar object categories. For example, a model trained to identify flamingos is also capable of identifying waterfowl.

count. Moreover, they demonstrate our method's proficiency in accurately identifying crowd locations, whether analyzing the complete image or specific patches. Importantly, our model proves resilient against a wide array of non-target elements within a scene, such as buildings, roads, and natural features, underscoring its robustness in complex environments.

Extension to Similar Objects. In this section, we examine our model's ability to recognize objects outside its training set but within related categories. Fig. 7 demonstrates that a model trained on flamingos can also accurately identify waterfowl, showcasing impressive generalization to similar objects. This versatility is significant, indicating that a model trained on a specific object category can be effectively applied to analogous categories, enhancing its utility and broadening its applicability.

3 Failure Cases

In Fig. 8, we highlight instances where our model does not perform as expected, categorizing these into underestimation and overestimation failures. Underesti-

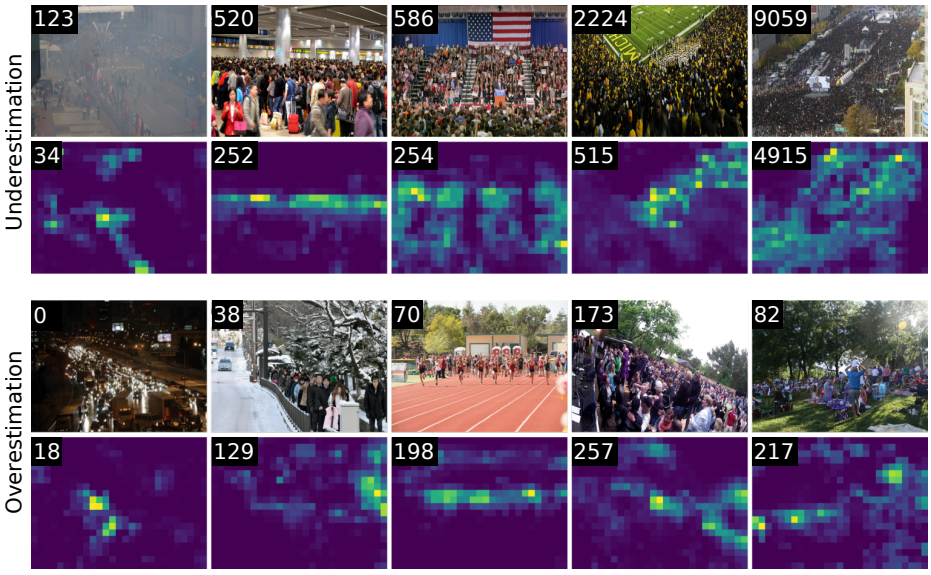

Fig. 8: Failure Cases. We highlight instances of underestimation and overestimation by our model, including challenges posed by environmental factors like fog and clutter, and the model's underestimation in some dense scenes.

mation can occur for several reasons. For instance, in the top left-most image, fog obscures pedestrians, making it difficult for our model to recognize them. Similarly, in other cases, people are hidden behind one another, or there simply isn't enough visible detail due to the object size or image resolution for accurate detection. On the other hand, overestimation issues often arise from the model mistaking unrelated elements for the target object. An example includes a cluttered traffic scene at night being misinterpreted as a group of pedestrians. Similarly, image regions with trees and dense foliage sometimes confuse the model. Despite these challenges, our method generally performs well across a range of scenarios. Nevertheless, these limitations highlight areas for improvement in future research.

4 Implementation Details

Implementation. We employ ResNet50 [3] as the underlying architecture. We train f_θ for 5 epochs, utilizing the Adam optimizer with a learning rate set to $5e^{-5}$. We resize all images to a uniform size of (640, 853, 3). During inference, we use a partition rate of 3 for all datasets. For the data generation process, we rely on Stable Diffusion 2.1. When performing image-to-image generation, we set the strength parameter to 0.45. Throughout all image generation procedures, we maintain a fixed guidance scale of 7.5 and carry out optimization for 50 steps.

For crowd counting problems, we set the prompt labels to:

$$N = [0, 1, 5, 10, 15, 20, 40, 60, 80, 100, 150, 200, 250, 300, 350, 400, 600, 800, 1000].$$

And for all other object categories, which have significantly lower average counts, we set the prompt labels to:

$$N = [0, 1, 2, 3, 5, 10, 15, 20, 30, 60].$$

Further, we set aside 15% of the synthesized sorting data as a validation set for performing count model selection. We use this data to set the maximum prompt count, c_{max}, in the set N when training the counting network g_{Φ}. We do this by performing inference on these validation sorting examples with g_{Φ} and selecting the model with the highest accuracy. This provides a c_{max} of 150 for SHB, 600 for SHA, 600 for JHU, and 1000 for QNRF, which approximately follows the mean of each dataset.

Table 4: Prompt List. This list showcases the straightforward prompts utilized for data generation, emphasizing the simplicity of the generation process.

Category	Usecase	Prompt	Negative
Crowd	Count	A group of {N} people.	-
Vehicle	Count	{N} vehicles. Overhead view.	-
Penguin	Count	{N} Penguins.	-
Crowd	Remove	An empty outside space. Nobody around.	people. crowds. pedestrians. humans.
Penguins	Remove	An empty outdoor arctic space. Nobody around.	penguins, birds, animals, fowl, avian.
Vehicles	Remove	An empty parking lot. Nobody around.	vehicles, cars, automobiles, trucks, jeeps, suvs, vans.
Crowd	Add	A crowd of people.	-
Penguins	Add	A large group of penguins.	-
Vehicles	Add	A busy parking lot with many cars. Overhead view.	-

Prompt Selection. In Tab. 4, we present a selection of prompts utilized for various synthesis tasks and object categories. It is not meant to be an exhaustive list, but rather to highlight the simplicity of the prompts used. The most complex specification involves directing that vehicle counting data be generated from an overhead perspective to better align the synthetic images with the aerial drone photography found within the CARPK dataset. All datasets utilize the following negative prompt list to ensure realism: *artistic, painting, vector art, graphic design, watercolor, text, writing, anime.*

Inference Throughput. This section assesses the inference speed of our model, conducted on a Nvidia Titan X Pascal GPU. Our findings reveal that the model processes images with dimensions (640, 853, 3) at an average rate of 49.5 frames per second (FPS). However, the model's speed varies when handling images with high-density areas, necessitating subdivision into sub-patches for detailed analysis. Specifically, in scenarios requiring the image to be divided into a 3×3 grid due to dense regions, the throughput decreases to an average of 5.5 FPS. This variation outlines the range of our model's inference speed, providing insights into its performance under different conditions.

5 Negative Social Impacts & Human Subjects Data

The modern deep learning approaches to crowd counting emerged with a key paper published in 2010 [5]. This area of research, crucial for tasks like event management, disaster response, and public safety enhancement, has seen substantial developments over the years. For instance, advanced crowd counting techniques played a pivotal role in analyzing crowd behavior during significant events, such as the January 6th Capitol riot [1], which underscored its societal importance. However, key datasets like ShanghaiTech A and B [10], JHU++ [6,7], QNRF [4], and NWPU [9] have relied heavily on images from public surveillance and the internet. This raises concerns about privacy and the potential for misuse in surveillance by various entities.

it is important to note that crowd counting datasets do not contain information related to facial recognition or individual identification; they merely mark the location of persons with dots or bounding boxes without revealing any personal details which somewhat mitigates privacy concerns. However, concerns remain since individuals might unknowingly appear in these datasets. Some dataset creators, like the authors of JHU++, offer a removal process for those depicted in images, but this process often lacks transparency. We advocate for clearer and more efficient opt-out procedures to protect individual privacy.

Moreover, the potential misuse of crowd counting in surveillance applications cannot be overlooked. Although crowd counting is distinct from facial recognition and not intended for invasive monitoring, its misuse remains a concern. Dataset licenses usually restrict use to academic and non-commercial purposes, yet these licenses may still be too permissive to prevent downstream harms. Recent proposals like the Open Responsible AI License (OpenRAIL) [2] have been introduced to ensure AI's ethical use, especially concerning applications that could infringe on personal rights or safety. We propose that these more restrictive licenses should be more widely adopted in the field.

We argue that the potential of crowd counting methods to serve societal good outweighs the limited scope for misuse. Nevertheless, the ethical implications of these technologies demand continuous vigilance from researchers. It is imperative for contributors in this domain to be conscientious of how and by whom their work is utilized, maintaining an awareness of the broader societal implications of their contributions.

References

1. Cheng, Z.Q., Dai, Q., Li, H., Song, J., Wu, X., Hauptmann, A.G.: Rethinking spatial invariance of convolutional networks for object counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19638–19648 (June 2022)
2. Contractor, D., McDuff, D., Haines, J.K., Lee, J., Hines, C., Hecht, B., Vincent, N., Li, H.: Behavioral use licensing for responsible ai. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 778–788 (2022)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
4. Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M.: Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European conference on computer vision (ECCV). pp. 532–546 (2018)
5. Lempitsky, V., Zisserman, A.: Learning to count objects in images. Advances in neural information processing systems **23** (2010)
6. Sindagi, V.A., Yasarla, R., Patel, V.M.: Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1221–1231 (2019)
7. Sindagi, V.A., Yasarla, R., Patel, V.M.: Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(5), 2594–2609 (2020)
8. Wang, B., Liu, H., Samaras, D., Nguyen, M.H.: Distribution matching for crowd counting. Advances in neural information processing systems **33**, 1595–1607 (2020)
9. Wang, Q., Gao, J., Lin, W., Li, X.: Nwpu-crowd: A large-scale benchmark for crowd counting and localization. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020). `https://doi.org/10.1109/TPAMI.2020.3013269`
10. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 589–597 (2016)