Semantic Generative Augmentations for Few-Shot Counting

Perla Doubinsky[1]   Nicolas Audebert[1]   Michel Crucianu[1]   Hervé Le Borgne[2]

[1] CEDRIC (EA4329), Cnam Paris, France [2] Université Paris-Saclay, CEA List, Palaiseau, France

{perla.doubinsky,nicolas.audebert,michel.crucianu}@lecnam.net, herve.le-borgne@cea.fr

Figure 1. Diversified augmentation generation by joint conditioning on density map and semantically similar prompt.

Abstract

With the availability of powerful text-to-image diffusion models, recent works have explored the use of synthetic data to improve image classification performances. These works show that it can effectively augment or even replace real data. In this work, we investigate how synthetic data can benefit few-shot class-agnostic counting. This requires to generate images that correspond to a given input number of objects. However, text-to-image models struggle to grasp the notion of count. We propose to rely on a double conditioning of Stable Diffusion with both a prompt and a density map in order to augment a training dataset for few-shot counting. Due to the small dataset size, the fine-tuned model tends to generate images close to the training images. We propose to enhance the diversity of synthesized images by exchanging captions between images thus creating unseen configurations of object types and spatial layout. Our experiments show that our diversified generation strategy significantly improves the counting accuracy of two recent and performing few-shot counting models on FSC147 and CARPK.

1. Introduction

Counting objects is a task with applications in many domains *e.g.* manufacturing, medicine, monitoring, that involve different types of objects. While earlier works focused on learning specialized networks [2, 7, 14, 16], Few-Shot object Counting (FSC) [31] was recently introduced to train models that can count any object, including from categories outside the training data. Methods tackling FSC rely on *exemplar* objects annotated with bounding boxes (cf. Fig. 2), in an *extract-then-match* manner [21]. The features of the exemplars and query image are compared using *e.g.* correlation maps [31, 41] or attention [6, 9]. Matched features are then transformed into a *density* map indicating at each location in the image the density of the objects of interest. The density map is then summed to obtain the predicted count.

The reference dataset for FSC, namely FSC147 [31], contains a limited amount of data (3659 train images) thus bounding the performances of counting networks [30]. Expanding such a dataset is costly as the annotation process requires pinpointing the center of each object present in a query image, with a potentially high number of occurrences. To overcome the small dataset size, Ranjan *et al.* [30] augment FSC147 using a GAN to diversify the image styles. Diffusion models have now surpassed GANs owing to their training stability and lower sensitivity to mode collapse. These models produce more effective and diverse augmentations [12, 37, 39]. Recent works mostly aim at augmenting classification datasets *e.g.* ImageNet [8], where augmentations are generated by prompting the models with the image labels.
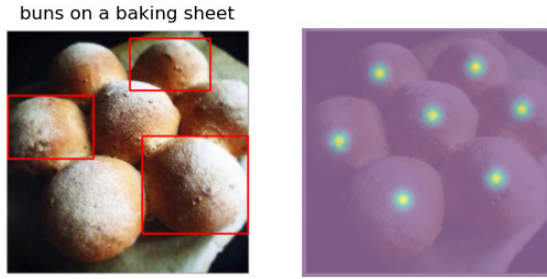
buns on a baking sheet

Figure 2. **Left**: FSC147 image with BLIP2 caption (above) and exemplar boxes (in red). **Right**: Ground-truth density map.

This fails to produce satisfying images for counting datasets as text-to-image models struggle to generate the correct number of objects [26]. Some works tackle improving compositionality in vision-language models [19, 25, 27] but are limited to small numbers of objects. Other works add more control to pre-trained text-to-image models [15, 23, 42].

To tackle few-shot counting, we propose to synthesize unseen data with Stable Diffusion conditioned by both a textual prompt and a density map. We thus build an augmented FSC dataset that is used to train a deep counting network. The double conditioning, implemented with ControlNet [42], allows us to generate novel synthetic images with a precise control, preserving the ground truth for the counting task. It deals well with large numbers of objects, while current methods fail in such cases [19, 27]. To increase the diversity of the augmented training set, we swap image descriptions between the n available training samples, leading to $\frac{n(n-1)}{2}$ novel couples, each being the source of several possible synthetic images. However, we show that some combinations do not make sense and lead to poor quality samples. Therefore, we only select plausible pairs, resulting in improved augmentation quality. We evaluate our approach on two class-agnostic counting networks, namely SAFECount [41] and CounTR [6]. We show that it significantly improves the performances on the benchmark dataset FSC147 [28] and allow for a better generalization on the CARPK dataset [14].

2. Related work

Learning with Generated Data   Improvements in image synthesis using generative models have sparked great interest in generating fake images to train deep neural networks. GANs were the first popular models to synthesize data for image classification [1, 5, 17], crowd counting [40] and image segmentation [43]. Nowadays, diffusion models such as DDPM [13] or Latent Diffusion [32] seem to outperform GANs, demonstrating more stable training, better coverage of the training distribution and higher image quality. The availability of powerful text-conditioned diffusion models [24, 29, 32, 33] has led to many works exploring how to leverage synthetic data for computer vision, e.g. im-

age classification in low-data regime [12], zero/few-shot learning [37, 39], ImageNet classification [3, 4, 34] and self-supervised learning [38]. These works focus on how to reduce domain gap [12], improve the prompts using e.g. text-to-sentence model [12] or WordNet [34] and increase diversity by optimizing the guidance scale [3, 34, 37]. This body of literature consistently demonstrates how generated data allow deep networks to learn more robust representations and improve generalization for image classification. We focus more specifically on few-shot class-agnostic object counting. Compared to image classification, this task involves small datasets and local spatial understanding, as objects can be small and follow complex layouts. The generated data needs a level of compositionality that current generative models, including diffusion models, struggle to achieve. To bring the power of synthetic data to counting, we propose to condition diffusion models not only on text prompts but also on counting density maps to generate images with the correct number of objects in the desired spatial configuration. We exploit this double control to generate diversified unseen data by prompting the model with novel combinations of the controls.

Few-shot Object Counting   The goal of few-shot class-agnostic object counting is to count how many instances of objects of *any* arbitrary category there are in a given image, by leveraging only *a few* exemplars of the category of interest. This was initially formulated as matching exemplars and image patch features [21]. FSC147 [31] was later put forward as the main dataset for this task, with an open set train and test split to evaluate generalization to unseen object categories. Its authors introduced FamNet, a deep net trained to infer density maps from feature similarities. In the same lineage, BMNet [36] refines the similarity map by learning the similarity metric jointly with the counting network. In SAFECount [41], the similarities are used to fuse exemplars features into the query image features. The density map is then predicted from the enhanced features. Other works e.g. CounTR [6] and LOCA [9] focus on improving the feature representations using a Transformer backbone as the visual encoder and injecting information about the exemplars' shape in the network [9]. The closest comparison to our work is the Vicinal Couting Network from Rajan *et al.* [30]. It augments FSC147 with generated data by training a conditional GAN jointly with the counting network, producing augmentations that preserve the image content while modifying its visual appearance. While outperformed by later models, it introduced the idea that well-chosen augmentations can significantly boost counting accuracy. In this work, we leverage large pre-trained text-to-image diffusion models to produce diverse augmentations that not only alter the appearance, but are also able to change the content, to synthesize augmentations with a variety of object semantics.
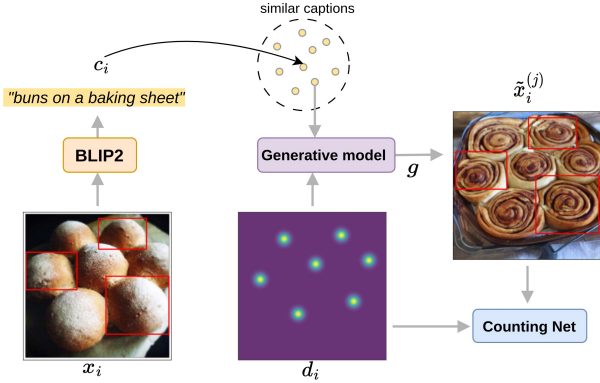
Figure 3. Overview of our approach. We condition a pre-trained diffusion model on both text prompts and density maps and perform swaps with similar captions. The density and original exemplars boxes are used as ground-truth for the generated augmentation.

3. Latent diffusion

Rombach *et al.* [32] introduced Latent Diffusion Models (LDMs) that are diffusion models applied in the latent space of generative models such as VQGAN [10]. These models operate in a compressed latent space thus reducing training and inference time. For more controlled generation, LDMs can be conditioned with *e.g.* text, images or semantic maps. Stable Diffusion is a popular open-source LDM, trained on the large-scale LAION-2B dataset [35] and enabling high-quality text-to-image generation. Prompt-based control heavily relies on the capabilities of the underlying text encoder, typically CLIP [28] that is known to poorly integrate compositional concepts such as counting [25].
ControlNet Recent works have added ancillary control signals to LDMs. Such a popular approach is ControlNet [42], which extends available pre-trained diffusion models by creating a trainable copy of the original network with an additional control input such as a semantic, edge or depth map. This trainable copy is linked to a locked copy to preserve the capacities of the original model. The training objective is an LDM objective similar to [32]:

$$\mathcal{L} = ||\epsilon_t - \epsilon_\theta(z_t, t, c_t, c_f)||^2 \qquad (1)$$

where $\epsilon_t \sim \mathcal{N}(0, 1)$, $\epsilon_\theta(,)$ is a model of the diffusion noise, z_t is the latent representation of x_t, c_t the text prompt and c_f the task-specific control.

4. Semantic Generative Augmentations

4.1. Few-Shot Counting

The goal of few-shot class-agnostic counting is to learn to count objects regardless of their category. To achieve this, the query image $x \in \mathbb{R}^{H \times W \times 3}$ is annotated with $n \in \{0, 1, 2, 3...\}$ *exemplar* boxes of coordinates $b \in \mathbb{R}^4$. The counting network takes as input both the query image and

the set of n boxes. It predicts a density map $d \in \mathbb{R}^{H \times W}$ of same size as the image. As shown in Fig. 2, this ground-truth density map has zero values where there are no objects, and a Gaussian kernel of fixed variance at the center of every object. The final count is obtained by summing across all positions of the density map. The model is typically trained with an L_2 loss between the predicted and ground-truth densities. To evaluate class-agnostic models, object categories from the test set are disjoint from those in the validation and train sets. This open set evaluation allows us to measure the network's ability to count objects from *unseen* categories.

4.2. Text-and-Density Guided Augmentations

To synthesize new images that can effectively augment a few-shot counting dataset, we need to have control over the number of objects and how they are laid out. Indeed, we need to ensure that we know the density maps of the synthetic samples so that they can be used to train the model. In addition, being able to control object type and spatial configuration also constitutes a lever to diversify the dataset by generating new combinations of categories and densities. As few-shot counting datasets are generally limited in size, we take advantage of available pre-trained diffusion models to synthesize diversified augmentations of the training samples, reducing overfitting and improving generalization. However, large pre-trained generative models such as Stable Diffusion are usually conditioned through textual prompts.

To finetune these models, we first have to pair textual captions to the training images. We obtain diverse and descriptive captions using an off-the-shelf captioning model, *e.g.* BLIP2 [20]. This produces richer captions than plain object categories such as "a photo of {class}". However, two shortcomings remain. First, generated captions may not contain any information about the number or arrangement of the objects. Second, text-conditioned LDMs poorly respect prompts regarding compositional constraints. Even adding this information in the caption does not guarantee that generated images would follow them. This is especially problematic as the correctness of the layout is a prerequisite to generate images for which we know the ground-truth. Therefore, we further condition the generative model directly on the density maps as an additional input, using the ControlNet fine-tuning strategy. To summarize, our generative model is now conditioned on a text prompt, obtained by an automated captioning of the training image, and its ground truth density map to enforce the spatial layout of the objects. This allows us to synthesize new samples that augment the original image, while keeping the ground truth intact, making the augmentation amenable to supervised learning.

4.3. Diversity-Enhanced Augmentations

To formalize the augmentation process, let $\mathcal{D}_{\text{train}} = \{x_i, b_i, d_i\}_{i=1}^N$ be an annotated counting dataset, with x_i
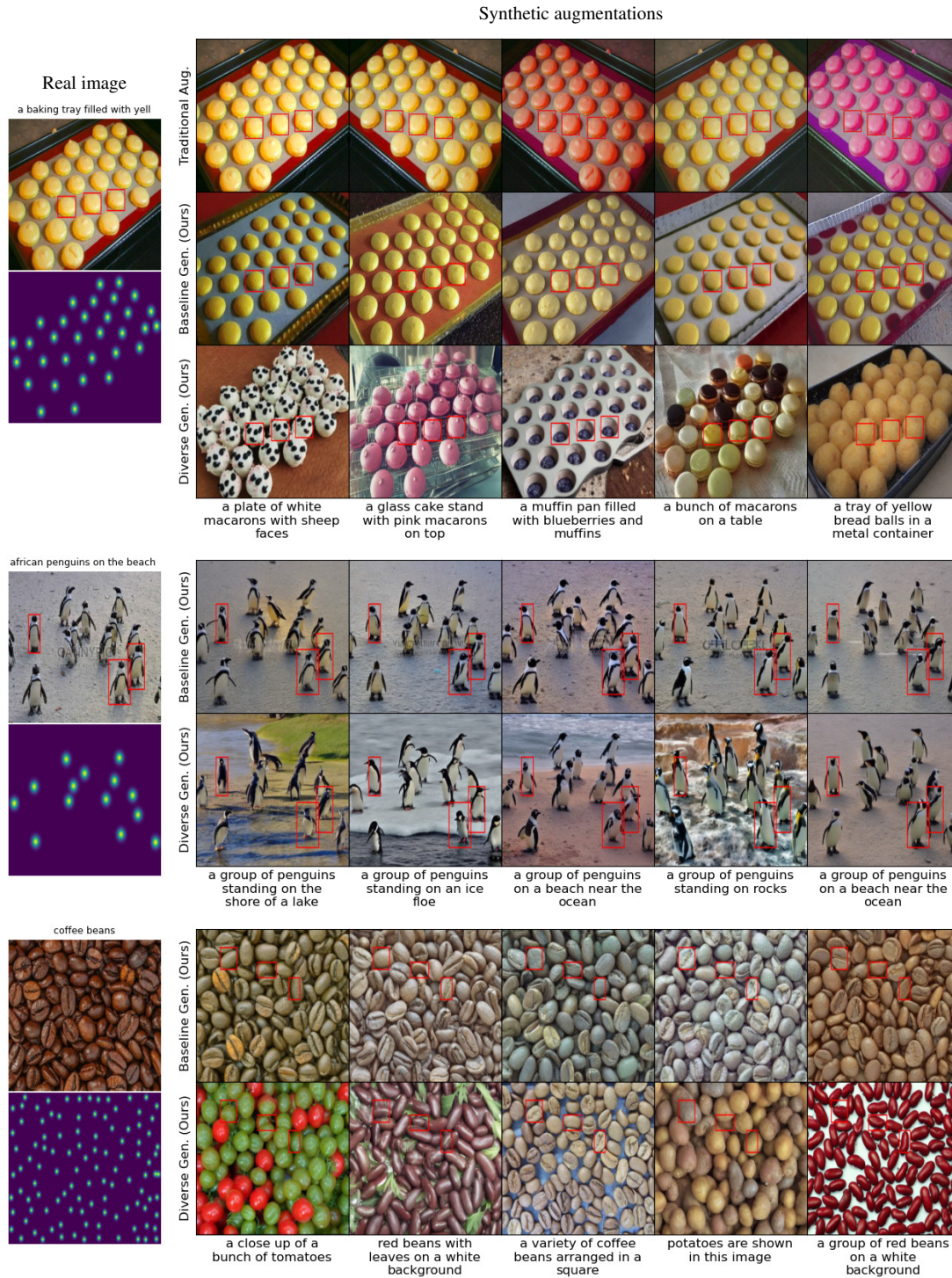
Synthetic augmentations

Real image

a baking tray filled with yell

Traditional Aug.

Baseline Gen. (Ours)

Diverse Gen. (Ours)

a plate of white macarons with sheep faces

a glass cake stand with pink macarons on top

a muffin pan filled with blueberries and muffins

a bunch of macarons on a table

a tray of yellow bread balls in a metal container

african penguins on the beach

Baseline Gen. (Ours)

Diverse Gen. (Ours)

a group of penguins standing on the shore of a lake

a group of penguins standing on an ice floe

a group of penguins on a beach near the ocean

a group of penguins standing on rocks

a group of penguins on a beach near the ocean

coffee beans

Baseline Gen. (Ours)

Diverse Gen. (Ours)

a close up of a bunch of tomatoes

red beans with leaves on a white background

a variety of coffee beans arranged in a square

potatoes are shown in this image

a group of red beans on a white background

Figure 4. Qualitative results for the Baseline vs. Diverse augmentations. At the bottom of each diverse sample we show the caption used to generate the image. Our strategy allows to diversify the type of objects and/or the background.

an image, b_i the exemplar bounding boxes for each image, and d_i its ground-truth density map. Let $\mathcal{C} = \{c_i\}_{i=1}^N$ be the set of corresponding captions. For each image x_i, we aim at generating M augmentations using our text-density

conditional generative model $g(d_i, c_i)$.

Baseline We sample augmentations from the LDM by taking advantage of the non-deterministic *reverse* diffusion process and the expressiveness of the pre-trained model. For

an image x_i we produce M augmentations $\tilde{x}_i^{(j)}$ that share its caption and density map:

$$\tilde{x}_i^{(j)} = g(d_i, c_i), \quad j = 1, ..., M \quad (2)$$

These augmentations preserve both the number and layout of objects – because of the density conditioning – and the semantics *e.g.* object category and type of background – because of the text prompt. This already augments the number of samples available for training.

Diverse We can however go further and *diversify* the augmentations by altering either the text description or the spatial organisation of the objects. To do so, we take advantage of dual conditioning on both densities and captions. We mix the two sets to create new combinations (density map, caption), producing augmentations that are semantically and geometrically more diverse than the original dataset. Yet, this mixing of the conditionings should be done carefully, to avoid low quality augmentations. Indeed, not all combinations make sense, *e.g.* "a herd of cows" and "a pearl necklace" exhibit very different spatial layouts. To prompt the generative model with realistic (*density*, *text*) pairs, we rely on caption similarity to find new associations between images that share some semantics, *e.g.* "cows" and "bisons".

We swap captions at random between pairs of *compatible* images. Two images are said to be compatible if their captions are more similar than some threshold t_c, *i.e.*:

$$\text{sim}(c_i, c_k) = \frac{\Psi(c_i)^\top \Psi(c_k)}{||\Psi(c_i)||_2 ||\Psi(c_k)||_2} > t_c$$

where Ψ is a suitable text encoder. We then sample new images using the initial density map, but replacing the original caption with the caption $c_k \in \mathcal{C}$ from a compatible training observation chosen at random:

$$\tilde{x}_i^{(j)} = g(d_i, c_k), \quad j = 1, ..., M \quad (3)$$

This process results in more diverse augmentations compared to the baseline and alters more the images than traditional augmentations (color jitter, crops, etc.), as shown in Fig. 4.

Synthetic and Diverse Balance We follow the training strategy from Trabucco *et al.* [39], where the synthetic augmentations are used as a regular data augmentation with a probability p_0 when training the counting model. As a way to balance baseline and diversified augmentations, we set a probability p_c that defines the fraction of the M augmentations that use a swapped caption instead of the original one. Typically, $p_c = 0.5$ means that 50% of the generated augmentations employ the original (caption, density) pair and that the remaining 50% use new (caption, density) combinations. For each augmentation, we keep the density used to condition the image generation and the original exemplar boxes as ground truth to train the model.[1]

---

[1] Note that if the caption changes the object category, bounding boxes

5. Experiments

5.1. Datasets and metrics

FSC147 FSC147 [31] is a 3-shot counting dataset with 147 object categories. It is the *de facto standard* of class-agnostic counting benchmarking. 89 categories are used for the training set, 29 are included in the validation set the remaining 29 constitute the test set. Note that the categories from the three sets are completely disjoint. In total, the dataset contains 6135 images, from which 3659 are used for training. The number of objects in the images varies from 7 to 3731 with an average of 56. Every image is annotated with 3 exemplar bounding boxes and an object density map.

CARPK CARPK [14] is a class-specific dataset for counting cars in parking lots based on overhead imagery. The dataset contains 1448 images from a UAV in 4 parking lots: 3 for training and the last one for testing. There are 5 exemplar objects in total that are randomly extracted from the training set and employed during both training and testing. Following [6, 41], we evaluate on CARPK the generalization ability of our class-agnostic models on a new dataset.

Metrics We follow the standard evaluation of the counting accuracy through the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

5.2. Implementation details

Augmentation We train ControlNet on the training images and density maps from FSC147. Text prompts are obtained by captioning the images with BLIP2 [20]. The underlying pre-trained diffusion model is Stable Diffusion $v1.5$ trained on LAION 2B. We use the default settings and train for 350 epochs. After training, we employ a guidance scale of 2.0 and 20 denoising steps to generate an image. For each augmentation strategy, we generate $M = 10$ augmentations per training sample unless specified otherwise. We swap the original caption with another one with a probability $p_c = 0.5$. Compatible captions to swap with are obtained by extracting caption features with the BLIP2 text encoder and filtering the captions with a similarity higher than $t_c = 0.7$.

Counting networks We demonstrate the effectiveness of our augmentation strategies on two state-of-the-art counting networks: SAFECount [41] and CounTR [6]. SAFECount is a CNN while CounTR is Transformer-based. In CounTR, training is done in two phases. First, the network is pre-trained using a self-supervised masked auto-encoder [11], then it is fine-tuned in a supervised fashion with the usual L_2 loss on the densities. For CounTR, we employ the pre-trained model released by the authors and only retrain the fine-tuning phase. We use the hyperparameters reported in the original papers to train both networks, except training

---

for the exemplars might not be accurate anymore (*e.g.* "pens" are narrow and elongated, while "erasers" are closer to squares, see Sec. 6).

| | (a) SAFECount [41] | | | | (b) CounTR [6] | | | |
| | Val | | Test | | Val | | Test | |
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Traditional Augmentation*,†	13.95	51.73	13.73	91.85	14.25	50.15	13.13	88.21
+ Real Guidance [12]	14.94	53.09	13.48	**80.69**	15.37	49.47	13.37	96.44
+ Baseline Generation (Ours)	13.30	49.38	13.22	92.47	12.60	43.53	11.83	87.97
+ Diverse Generation (Ours)	**12.59**	**44.95**	**12.74**	89.90	**12.31**	**41.65**	**11.32**	**77.50**
Traditional Augmentation (reported)	15.28	47.5	14.25	85.54	13.13	49.83	11.95	91.23

Table 1. Quantitative results on FSC147. (*) Traditional augmentations include color jitter, random cropping. (†) [41] and [6] are reproduced, details are provided in the supplementary material.

is 100 epochs longer to account for the higher number of training images. During training, we replace an image x_i with one of its augmentations $\tilde{x}_i^{(j)}$ with probability $p_0 = 0.5$. This balances the ratio of real vs. synthetic data in a single batch. We also employ traditional data augmentation strategies *e.g.* flips, color jitter, random cropping, that are applied to every image, both real and synthetic, as done in the original models. Both networks are trained and evaluated in the 3-shot setting. All results are averages over two runs.

5.3. Few-Shot Counting on FSC147

Comparison with Traditional Augmentation We report in Tab. 1 the improvement in counting accuracy on FSC417 with our augmentation strategies when training SAFECount and CounTR. Consistent with the literature on synthetic data augmentation, baseline augmentations improve the results for both networks: MAE decreases by respectively 5% and 10% for SAFECount and CounTR on the val set. Nonetheless, diversifying the augmentations allows us to reduce the MAE even further, by 10% and 11% on the same val set and by 7% (SAFECount) and 13% (CounTR) on the test set. We attribute this to ControlNet overfitting the training data due to the small dataset size. The low guidance employed to generate the images (2.0) aims at promoting diversity [34] but, as shown in Fig. 4 (Baseline Gen.), the generated images remain close to the original image in terms of visual appearance of the objects and background. However, ControlNet generalizes to different captions. In Fig. 4 (Diverse Gen.), we observe that swapping captions allows us to create more diverse data, altering the size and texture of objects and their background. Such features cannot be altered with traditional data augmentation. When mixing baseline and diverse augmentations, the performances for both networks improve significantly with respect to the model without synthetic augmentation, or with naive augmentations only.

Comparison with Real Guidance We compare our approach with Real Guidance, an augmentation strategy for image classification by He *et al.* [12]. Augmentations are
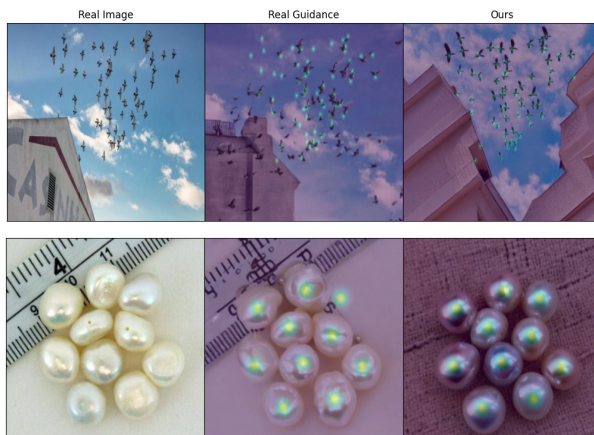

Figure 5. Qualitative comparison with Real Guidance [12]. Our augmentations preserve the layout while creating more diverse backgrounds. Ground-truth density maps overlap with the generated images (last 2 columns).

generated by prompting a pre-trained text-to-image diffusion model with the image classes. To reduce the domain gap, the synthetic images are generated from the real images with added noise as proposed in SDEdit [22]. Tab. 1 shows that our augmentation strategy outperforms Real Guidance[2]. Starting from the real image with added noise is generally insufficient to preserve the number of objects and their positions (Fig. 5, $2^{nd}col.$). It shows that the density map conditioning ensures the preservation of object positions and number without requiring to start from the real image, which can limit the diversity of the generated images.

5.4. Ablation Study

We study here the influence of the hyperparameters of our approach: impact of the caption similarity threshold, t_c, influence of introducing diversified augmentations by varying

---

[2]Except on test RSME with SAFECount, where Real Guidance performs better, due to two outlier test images with more than 2500 objects that dominate the average error (see supplementary material).
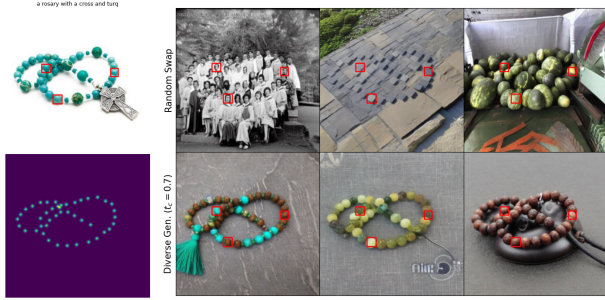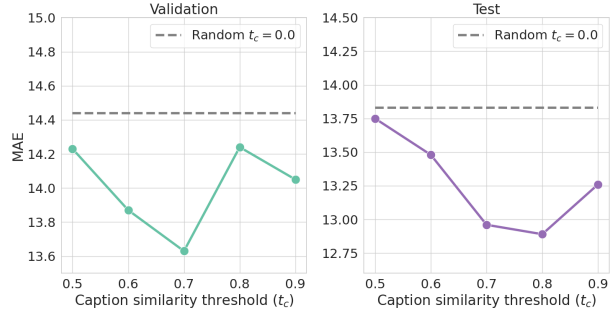
Figure 6. Caption swap at random (top) vs. similarity-based swap (bottom, $t_c = 0.7$). Random swapping results in a mismatch between the layout and the semantics.

(a) Validation and test MAE on FSC147 w.r.t. t_c (SAFECount).

(b) Distribution of object categories in the set of captions to swap from w.r.t t_c for a sample of category "bread rolls". Lower thresholds result in more diverse augmentations, while objects still belong to similar classes.

Figure 7. Impact of caption similarity threshold t_c on SAFECount.

the p_c parameter, the effect of the number of augmentations M, and of the ratio of synthetic samples during training, p_0. All ablations are conducted on SAFECount trained for 200 epochs to reduce training time.

Caption Similarity Threshold We swap captions based on caption similarity to form novel but plausible (*density*, *text*) combinations. As shown in Fig. 6, associating a completely unrelated caption to a given density map results in generated images that do not correspond to the input density map or are of poor quality, as it is harder for the model to generalize. In Fig. 7a we evaluate different similarity thresholds to the naive approach where all captions can be swapped freely at random ($t_c = 0.0$). The performances are improved compared to random swaps with all thresholds between 0.5 and 0.9. However, there is a quality-diversity tradeoff shown in Fig. 7b. Setting the threshold too high ($t_c = 0.8, 0.9$) swap captions between images of objects belonging to the same category, thus limiting diversity. With a lower threshold, *e.g.* $t_c = 0.7$, new captions can also belong to objects from different similar categories, *e.g.* swapping "bread rolls" and "macarons".

Rate of Diverse Samples In Fig. 8, we vary the rate of diverse augmentations among $M = 10$ augmentations. We compare with SAFECount trained solely with baseline non-diverse augmentations. More diverse samples overall increase the counting accuracy. We further find that adding 70% of diverse samples gives better performances than 50%. This suggests that the diverse augmentations are more beneficial than the baseline ones. The augmentations using the original captions yet remain useful to the model as we observe a slight increase in MAE when more than 90% of the augmentations are obtained on new combinations.

Number of Augmentations In Fig. 9, we vary the number of augmentations generated for each image. With a single augmentation, the performances already improve. For low values of $M = 1, 3, 5$, the performances are comparable, then a stronger increase is observed for $M = 10$. With twice as many augmentations ($M = 20$), performances degrade on the validation set. This might be due to an insufficient con-
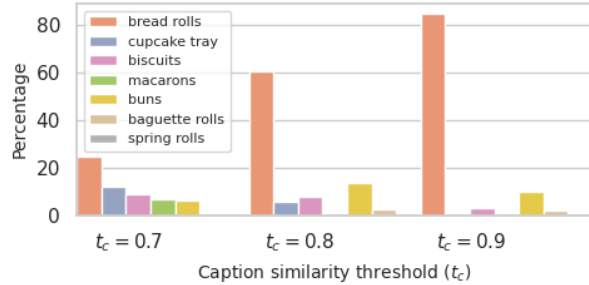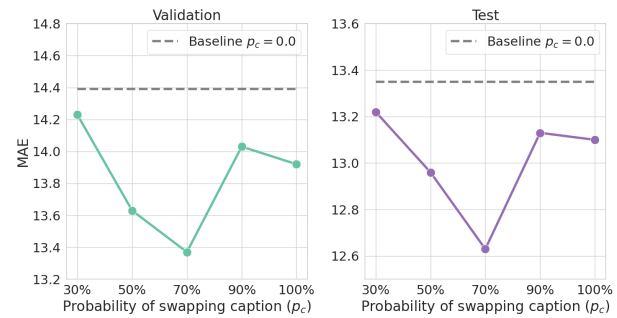

Figure 8. Impact of the percentage of diverse augmentations p_c on SAFECount. MAE is reported for the val and test sets of FSC147.

vergence, as the model is trained with many more different data points but for the same number of iterations.

Rate of Synthetic Samples Fig. 10 shows the counting accuracy w.r.t. ratio p_0 of synthetic samples vs. real samples in a batch. We observe that equally balancing the synthetic and real data gives the best performances, which is consistent with what has been observed in previous works generating synthetic data for image classification [4, 12].

5.5. Generalization on CARPK

CARPK [14] was introduced to train networks that can count cars in aerial views of parking lots. It is also used to evaluate the ability of class-agnostic models to count in a class-specific setting. Given a model trained on FSC147
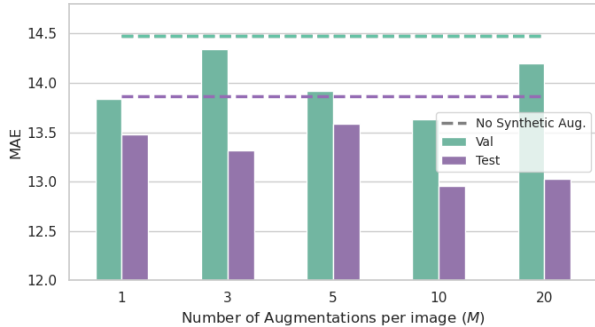
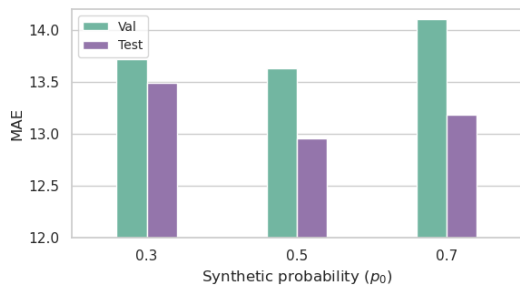Figure 9. Impact of the number of augmentations M on SAFE-Count (val. and test MAE on FSC147).

Figure 10. Impact of the synthetic data ratio p_0 on SAFECount (val. and test MAE on FSC147).

	Aug.	MAE	RMSE
Pre-trained on FSC147	Trad.	17.65	23.83
	Div.	**16.49**	**19.05**
Fine-tuned on CARPK	Trad.	5.44	6.94
	Div.	**4.87**	**6.17**
SAFECount [41] (WACV'23)	Trad.	5.33	7.04
CounTR [6] (BMVC'22)	Trad.	5.75	7.45
BMNet+ [36] (CVPR'22)	Trad.	5.76	7.83

Table 2. Counting performance on CARPK with SAFECount.

(without the examples of the "car" category), the model is first evaluated without any fine-tuning, then with fine-tuning on CARPK. We evaluate our model trained on FSC147 with diverse augmentations in the same setting. Tab. 2 reports improved counting performances in both the pre-trained and fine-tuned settings in comparison to the models trained without synthetic augmentations. In the fine-tuning setting, we reach state-of-the-art counting accuracy (4.87 MAE/6.17 RMSE) on CARPK amongst class-agnostic models.

6. Limitation

Our synthetic data needs a ground truth and exemplars to train the counting network. Conditioning on densities makes it possible to reuse both the original density and the exemplar
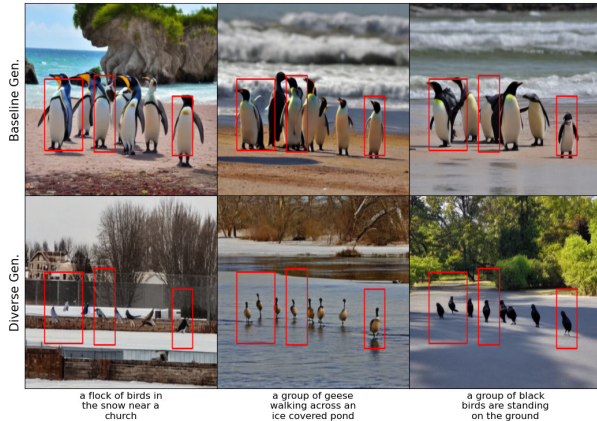

Figure 11. Limitation: our diverse generation strategy can change the size and shape of generated objects, leading to exemplar boxes (in red) that do not fit perfectly.

bounding boxes. However, changing the caption can affect the object category, and in turn its shape. In some rare cases, exemplar boxes do not fit the generated objects anymore, as illustrated in Fig. 11. We explored to what extent refining these boxes could improve our model. We segmented objects using SAM in zero-shot [18] prompted with object centers. Preliminary results showed no improvement with box refinement, possibly due to inaccurate segmentation.

7. Conclusion

We show that synthetic data generated by diffusion models improve deep models for few-shot counting. We adapt a pretrained text-to-image model with a density map conditioning and we propose a diversification strategy that exploits caption similarities to generate unseen but plausible data that mixes the semantics and the geometry of different training images. We show that selecting compatible images improves synthetic image quality with beneficial effects on model performance. We demonstrate that learning with our diverse synthetic data leads to improved counting accuracy on FSC147 and state of the art generalization on CARPK. This strategy could be adapted to other tasks requiring fine-grained compositionality, such as object detection and semantic segmentation. Our diversification scheme could be further extended by swapping both the captions and the density controls, by introducing a suitable similarity metric that operates on the density maps.

References

[1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017. 2

[2] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 483–498, Cham, 2016. Springer International Publishing. 1

[3] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 2

[4] Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. In *ICLR Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023. 2, 7

[5] Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. This dataset does not exist: training models from generated images. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2020. 2

[6] Liu Chang, Zhong Yujie, Zisserman Andrew, and Xie Weidi. Countr: Transformer-based generalised visual counting. In *British Machine Vision Conference (BMVC)*, 2022. 1, 2, 5, 6, 8

[7] Jiwei Chen, Wen Su, and Zengfu Wang. Crowd counting with crowd attention convolutional neural network. *Neurocomputing*, 382:210–220, 2020. 1

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1

[9] 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, 2

[10] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878, 2021. 3

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 5

[12] 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. 1, 2, 6, 7

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[14] 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*, pages 4145–4153, 2017. 1, 2, 5, 7

[15] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 13753–13773. PMLR, 23–29 Jul 2023. 2

[16] Zhongyi Huang, Yao Ding, Guoli Song, Lin Wang, Ruizhe Geng, Hongliang He, Shan Du, Xia Liu, Yonghong Tian, Yongsheng Liang, S. Kevin Zhou, and Jie Chen. Bcdata: A large-scale dataset and benchmark for cell detection and counting. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 289–298, Cham, 2020. Springer International Publishing. 1

[17] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation. In *International Conference on Learning Representations*, 2022. 2

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 8

[19] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 2

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 23–29 Jul 2023. 3, 5

[21] Erika Lu, Weidi Xie, and Andrew Zisserman. Class-agnostic counting. In *Asian conference on computer vision*, pages 669–684. Springer, 2018. 1, 2

[22] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 6

[23] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[24] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 17–23 Jul 2022. 2

[25] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. *arXiv preprint arXiv:2302.12066*, 2023. 2, 3

[26] Vitali Petsiuk, Alexander E Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A Plummer, Ori Kerret, et al. Human

evaluation of text-to-image models on a multi-task benchmark. In *NeurIPS Workshop on Human Evaluation of Generative Models*, 2022. 2

[27] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. *arXiv preprint arXiv:2306.05427*, 2023. 2

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[30] Viresh Ranjan and Minh Hoai. Vicinal counting networks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4220–4229, 2022. 1, 2

[31] 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, 2, 5

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[34] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learningtocount transferable representations from synthetic imagenet clones. In *CVPR 2023–IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 6

[35] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc., 2022. 3

[36] 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. 2, 8

[37] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 769–778, 2023. 1, 2

[38] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. StableRep: Synthetic images from text-to-image models make strong visual representation learners. *arXiv preprint arXiv:2306.00984*, 2023. 2

[39] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023. 1, 2, 5

[40] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8198–8207, 2019. 2

[41] 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, 5, 6, 8

[42] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 3

[43] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. DatasetGAN: Efficient labeled data factory with minimal human effort. In *CVPR*, 2021. 2