

Represent, Compare, and Learn: A Similarity-Aware Framework for Class-Agnostic Counting

Min Shi Hao Lu Chen Feng Chengxin Liu Zhiguo Cao*

Key Laboratory of Image Processing and Intelligent Control, Ministry of Education
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China
{min.shi, hlu, chen.feng, cx.liu, zgcao}@hust.edu.cn

Abstract

Class-agnostic counting (CAC) aims to count all instances in a query image given few exemplars. A standard pipeline is to extract visual features from exemplars and match them with query images to infer object counts. Two essential components in this pipeline are feature representation and similarity metric. Existing methods either adopt a pretrained network to represent features or learn a new one, while applying a naive similarity metric with fixed inner product. We find this paradigm leads to noisy similarity matching and hence harms counting performance. In this work, we propose a similarity-aware CAC framework that jointly learns representation and similarity metric. We first instantiate our framework with a naive baseline called Bilinear Matching Network (BMNet), whose key component is a learnable bilinear similarity metric. To further embody the core of our framework, we extend BMNet to BMNet+ that models similarity from three aspects: 1) **representing** the instances via their self-similarity to enhance feature robustness against intra-class variations; 2) **comparing** the similarity dynamically to focus on the key patterns of each exemplar; 3) **learning** from a supervision signal to impose explicit constraints on matching results. Extensive experiments on a recent CAC dataset FSC147 show that our models significantly outperform state-of-the-art CAC approaches. In addition, we also validate the cross-dataset generality of BMNet and BMNet+ on a car counting dataset CARPK. Code is at tiny.one/BMNet

1. Introduction

Object counting aims to infer the number of objects from an image. Most existing methods focus on a specific category, e.g., crowd [42], animal [2], or car [26], while requiring numerous training data to learn a good model. In contrast, given only one exemplar of a novel category, e.g., a

*Corresponding author

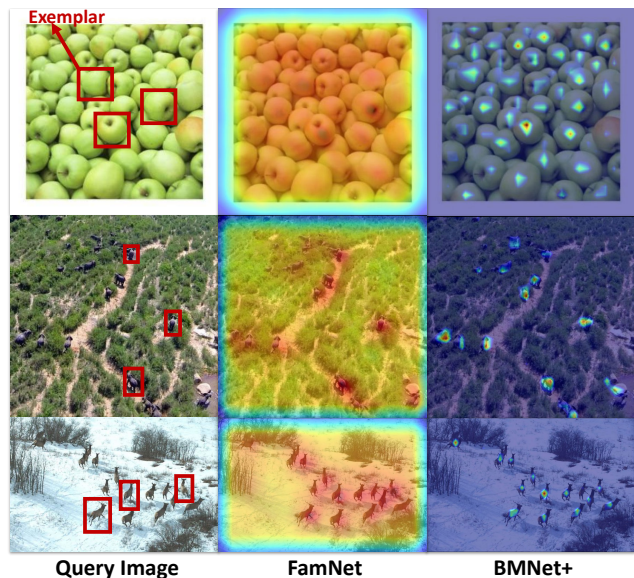


Figure 1. Visualizations of intermediate similarity maps in class-agnostic counting. Compared with the state-of-the-art FamNet [29], our model (BMNet+) generates high-fidelity results.

car, even a child can easily capture its visual properties and count cars in new scenes. Recently, CAC (Class Agnostic Counting) [21, 29, 40], which counts objects of arbitrary categories given only few exemplars, is proposed to reduce the reliance on training data. CAC points out a promising direction for object counting, i.e., from learning to count objects to learning the way to count.

Generally, existing CAC methods [21, 29, 40] work in an *extract-and-match* pipeline. They first extract visual features from exemplars and match these features with those of query images. Similarity matching results are then used as intermediate representations to infer object counts. Intuitively, two factors play critical roles: feature representation and similarity metric. Existing methods either use a learnable [21, 40] or a fixed feature extractor [29], but apply a similarity metric with some pre-defined rules, e.g., in-

ner product [29, 40]. We find this can yield unsatisfactory matching results. From Fig 1, by examining a recent model FamNet [29], we observe obvious noise on background and weak responses on target positions. The resulting density map may be erroneous given such ambiguity.

In this work, we present a generic similarity-aware framework for CAC, which jointly learns representation and similarity metric in an end-to-end manner. Our goal is to seek better similarity modeling that can generalize well to novel categories. First, we instantiate a bilinear matching network (BMNet), which extends the fixed inner product to a learnable bilinear similarity metric and also allows learnable representation through back-propagation. Unlike fixed inner product, the bilinear similarity metric captures flexible interactions among feature channels to measure similarity. Then, we extend BMNet to BMNet+ to embody the core motivation of our framework from three aspects: *representing* instances via self-similarity, *comparing* the similarity dynamically, and *learning* with explicit, similarity-aware supervision. In particular, we apply self-attention [43] to represent self-similarity among features to mitigate intra-class variations. It augments the feature of each instance with information from other intra-class instances such that complementary clues like scales or viewpoints can be offered. The dynamic similarity metric applies a feature selection module to the exemplars to find key patterns and hence embraces both dynamism and selectivity. Then, inspired by metric learning [25], the similarity loss imposes an explicit supervision on the intermediate similarity map to pull the exemplar and the target close but to push the exemplar and background away.

Experiments on the public benchmark FSC147 [29] show that our method outperforms the previous best approaches by large margins, with a relative improvement of +33.72% and +33.79% on the validation and test sets in terms of mean absolute error. According to Fig. 1, our method outputs better intermediate similarity results and presents generality over different categories. The ablation study validates the three main components within BMNet+. And we further show the cross-dataset generality of our models on a car counting dataset CARPK [13].

Our contributions are two-fold:

- A generic CAC framework that includes the existing pipeline and also generalizes it with joint representation learning and similarity learning;
- BMNet and BMNet+: two CAC models instantiated from our framework, which models packed similarity.

2. Related Work

2.1. Class-Specific Object Counting

According to how the counting problem is formulated, existing methods can be categorized into counting by de-

tection [10], regression [7, 36, 42, 44], classification [17], and localization [1, 9, 31]. The most-studied regression-based approaches formulate counting as a dense prediction [22, 23] task, which learns to predict density maps [15]. Under this paradigm, most methods focus on designing network architectures [44], multi-scale strategies [32, 39], or new loss functions and learning targets [24, 36]. Recently, new paradigms are developed such as reinforcement learning [18] and counting by localization [1, 9, 31]. The key difference between class-specific counting and class-agnostic one lies in that the latter requires a more generic representation and a more discriminative similarity metric.

2.2. Class-Agnostic Counting

Lu *et al.* [21] first address CAC and propose a general matching network. One convolutional neural network is shared to extract feature maps for both query images and exemplars. These features are then concatenated to regress the object count. Considering that direct regression from concatenated features may cause overfitting, recent methods start to model similarity explicitly. CFOCNet [40] uses the feature map of exemplar as a 2D kernel to convolve over the query feature map, following the spirit of Siamese network in object tracking [4]. They also design a multi-scale matching framework to improve robustness. FamNet [29] also adopts siamese way to model similarity and further proposes test-time adaptation given test exemplars. To alleviate the shortage of training data, Ranjan *et al.* [29] propose the first and only CAC dataset FSC147 that covers challenges like occlusion and scale variation. The above methods report promising results for CAC. However, they typically focus on multi-scale strategy, data amplification, or test-time adaptation, but neglect a fundamental problem – similarity modeling. In this work, we show the importance of similarity modeling and also present a generic framework that jointly learns both representation and similarity metric.

2.3. Metric Learning

Metric learning aims to embed data into a space where similar samples are pulled close and dissimilar ones are pushed away [25]. The similarity is measured in a fixed [3, 16] or learned [33, 37] manner. One common way constrains the similarity between features in a pair [11] or triplet [38]. Another way adds constraints based on signal-to-noise ratios [30, 34, 41], where the similarity between positive sample pairs is considered as the signal, and the similarity between negative pairs as the noise. The idea is to strengthen the signal and weaken the noise. We repurpose this idea into CAC, *i.e.*, pulling close the features between the exemplar and target instances, while pushing away the features between exemplars and background patches. We further design a similarity loss based on this idea to supervise the similarity matching results.

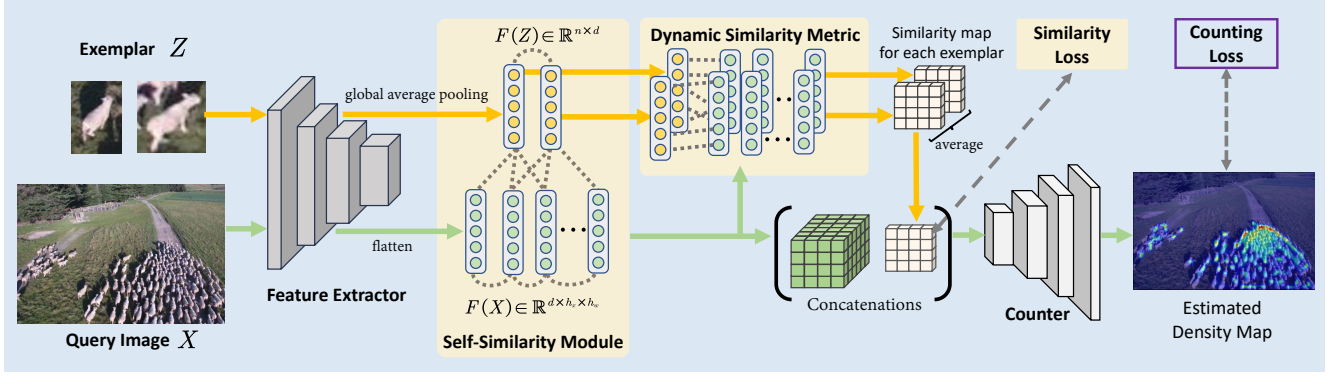


Figure 2. **The pipeline of BMNet and BMNet+.** BMNet follows the *extract-then-match* paradigm but learns representation and similarity metric jointly in an end-to-end manner. BMNet+ is an improved version whose differences from BMNet are highlighted in colored blocks.

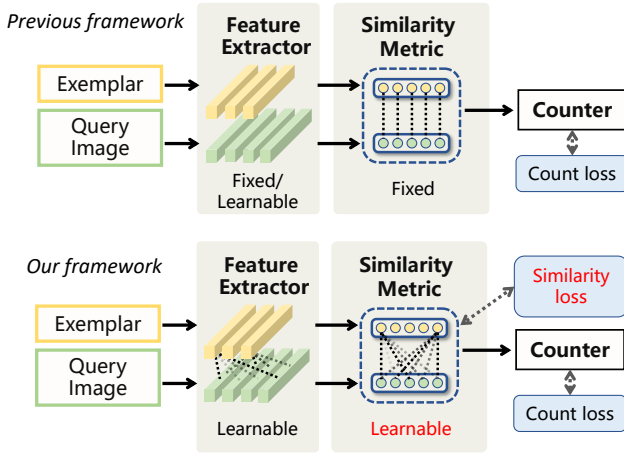


Figure 3. **The comparison between the previous framework and ours.** Ours can learn both representation and similarity metric jointly with more flexibility and generality.

3. A Similarity-Aware Framework for Class-Agnostic Counting

This section presents our framework for CAC, which jointly learns representation and similarity metric in an end-to-end manner (Fig. 3). We first instantiate this framework with a naive baseline, termed Bilinear Matching Network (BMNet), and then propose an extended BMNet+ to exemplify our idea on how to represent, dynamize, and learn similarity for both representation and similarity metric. The detailed pipeline of our methods is in Fig. 2.

3.1. Bilinear Matching Network

Differing from previous CAC methods, BMNet allows simultaneous optimization of representation and similarity metric. The core of BMNet is the bilinear similarity metric that captures flexible interactions among feature channels to model similarity.

Given a query image X and an exemplar Z of arbitrary

category c , CAC aims to count all the instances of category c within X . Without loss of generality, we use one exemplar to explain our pipeline (we will also note how to operate with multiple exemplars).

Feature Extractor. The feature extractor consists of layers of convolutional operations that map the input into d -channel features. For the query X , it outputs a downsampled feature map $F(X) \in \mathbb{R}^{d \times h_x \times w_x}$. For the exemplar Z , the output feature map is further processed with global average pooling to form a feature vector $F(Z) \in \mathbb{R}^d$.

Learning Bilinear Similarity Metric. Previous methods apply fixed inner product to compute the similarity between two feature vectors. We argue that such fixed one-to-one interactions may be insufficient in modeling class-agnostic similarity. Inspired by neural similarity learning [19] and bilinear models [28], we propose to extend the original inner product to a learnable bilinear similarity, which establishes flexible connections between two vectors. Specifically, let $F_{ij}(X) \in \mathbb{R}^d$ be the channel feature at spatial position (i, j) . By redefining $\mathbf{x}_{ij} = F_{ij}(X)$ and $\mathbf{z} = F(Z)$, the similarity map S can be obtained by

$$S_{ij}(\mathbf{x}, \mathbf{z}) = (P\mathbf{x}_{ij} + \mathbf{b}_x)^T (Q\mathbf{z} + \mathbf{b}_z), \quad (1)$$

where $P, Q \in \mathbb{R}^{d \times d}$ are learnable matrices, and $\mathbf{b}_x, \mathbf{b}_z \in \mathbb{R}^{d \times 1}$ are learnable biases. The initial bilinear metric is in the form $\mathbf{x}^T W \mathbf{z}$. We decompose W into P, Q specific to the query image and the exemplar, respectively. In practice, we find that this can yield better performance (refer to supplementary material for more details).

Given n exemplars, one can use Eq. 1 repetitively to compute n similarity maps, and then output their averaged similarity as the final similarity map S .

Counter. The counter receives the channel-wise concatenation of the query feature map $F(X)$ and the similarity map S , and then predicts a density map D_{pr} . The final count is the integral of D_{pr} . In practice, the counter consists of convolutional and bilinear upsampling layers.

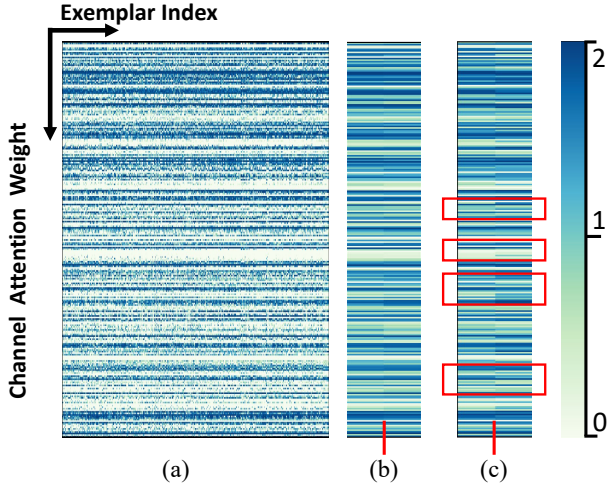


Figure 4. **Visualization of channel attention weights for exemplars from the same and different categories.** We visualize the attention weights (each vertical line) for exemplars from (a) the same category *apple*, (b) visually similar categories *apple* vs. *strawberry*, and (c) dissimilar ones *apple* vs. *stamp*. For (b) and (c), the red short line splits the samples into two categories. By focusing on the horizontal lines, we can observe that, channel attention weights for exemplars of the same or similar categories (cf. (a) and (b)) shows more consistency than those for different categories (cf. (c), especially in red boxes). Better zoom in for details. More visualizations can be found in supplementary material.

Supervision Signal. We adopt a conventional ℓ_2 loss as the counting loss $\mathcal{L}_{\text{count}}$:

$$\mathcal{L}_{\text{count}} = \|D_{\text{pr}}(X, Z) - D_{\text{gt}}(X, Z)\|_2^2, \quad (2)$$

where D_{gt} denotes the ground truth density map.

3.2. Learning Dynamic Similarity Metric

The bilinear similarity in Sec. 3.1 increases flexibility to model similarity. However, the learned similarity metric stays fixed once trained and treats all categories equally during inference. Considering that humans may learn to recognize a category based on category-specific patterns, *e.g.*, if told something is furry with four legs and pointy ears, one may suppose it to be a cat. We therefore think it is better to develop a dynamic similarity metric that can adaptively learn to focus on the key patterns of exemplars. Inspired by this intuition, we integrate a feature selection module over the exemplars to generate an exemplar-specific metric. Specifically, we regard each channel in $Qz + b_z$ as a pattern. Similar to SENet [14], we learn the dynamic channel attention weight a conditioned on $Qz + b_z$ such that similarity S can be computed by

$$S_{ij}(\mathbf{x}, \mathbf{z}) = [(P\mathbf{x}_{ij} + \mathbf{b}_x)]^T [a \circ (Q\mathbf{z} + \mathbf{b}_z)], \quad (3)$$

where \circ denotes the Hadamard product.

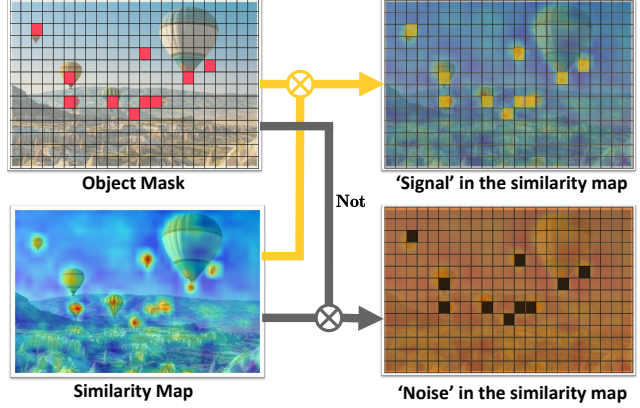


Figure 5. **An illustration on how to compute the similarity loss.** We consider the similarity between exemplars and target instances as signals (positive labels), while similarity between exemplars and background as noises (negative labels).

We exemplify the learned dynamic attention weights in Fig. 4. For exemplars of the same category (cf. Fig. 4(a)), the generated dynamic attention weights turn similar. Similar phenomena can be observed given two visually close categories (cf. Fig. 4(b)). This validates our intuition that the dynamic similarity metric learns to focus on similar visual patterns for similar categories. In contrast, given two visually different categories (cf. Fig. 4(c)), our method learns to extract different key patterns with clear distinction. Note that whatever the cases are, there exist common patterns between different categories. This accords with the way we humans recognize objects: first use general visual clues like shapes and colors, then focus on category-specific details.

3.3. Supervising the Similarity Map

Both existing CAC methods and our baseline BMNet only use the counting loss as supervision during training. In practice, we find that direct supervision on similarity matching results can help to guide similarity modeling. To this end, we start by posing a fundamental question: *what makes an ideal similarity metric for CAC?* In our opinion, it should output high similarity between the two features of the same category and low one for differing categories. This accords with the idea of metric learning [25].

Here we present a simple way to achieve this. Suppose the size of S is $1/r$ of that of X , *i.e.*, each position in similarity map corresponds to a $r \times r$ block within query image. For each position in S , we assign a positive label if its corresponding $r \times r$ block contains more than one target, and assign a negative one if it contains no target. We then derive the similarity loss \mathcal{L}_{sim} with signal-to-noise ratio:

$$\mathcal{L}_{\text{sim}} = -\log \frac{\sum_{i \in \text{pos}} \exp(S_i)}{\sum_{i \in \text{pos}} \exp(S_i) + \sum_{j \in \text{neg}} \exp(S_j)}. \quad (4)$$

Here *pos*, *neg* denotes positive and negative positions in S .

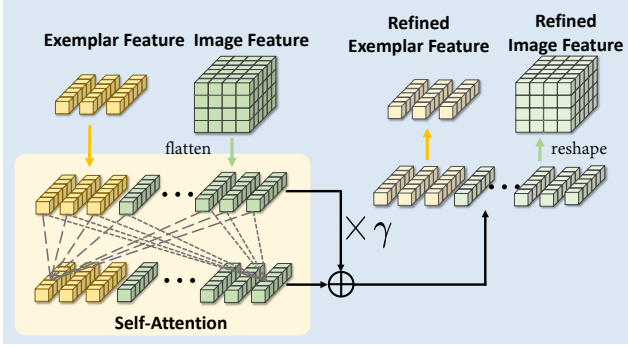


Figure 6. Self-similarity module.

With the counting loss $\mathcal{L}_{\text{count}}$ and the similarity loss \mathcal{L}_{sim} , the final training loss can be written as

$$\mathcal{L} = \mathcal{L}_{\text{count}}(D_{\text{pr}}, D_{\text{gt}}) + \alpha \cdot \mathcal{L}_{\text{sim}}(S), \quad (5)$$

where α balances the two component loss items.

3.4. Self-Similarity Module

The core of our framework also includes improving the representation suitable for similarity matching. Here we present a feasible way to address this. As in Fig. 7, in reality, instances of the same category often appear with different attributes like poses and scales. Such intra-class variations impose great challenges on similarity matching. Accordingly, we propose to augment each instance feature with complementary information from other instances of the same category but with different attributes.

Technically, we first collect the exemplar feature $F(Z)$ and each feature vector $F_{ij}(X)$ from the query feature map into a feature set. Then each vector in the feature set is updated via a self-attention mechanism [43] (Fig. 6). The updated features are added back to the original ones with a learnable ratio γ . The resulting feature set is then re-split and re-shaped to obtain the final $F(Z)$ and $F(X)$.

We remark that, [40] also applies self-attention over the feature maps similar to our work; hence the self-similarity module does not constitute our contribution. However, here we attempt to explain how self-attention works in our task. We start by visualizing the self-attention maps given the query points as in Fig. 7. It can be observed that each query point mainly focuses on instances of the same category. This differs from self-attention in object detection [5] where the query point mainly focuses on a single instance. This indicates that, the self-similarity module in CAC tends to aggregate same-category information and hence enhances representations with robustness towards intra-class variations.

Scale Embedding. Inspired by the positional embedding in Transformer [35], we wonder if we could similarly embed the scale information of the exemplars to improve the representation. Note that two factors cause the exemplars to

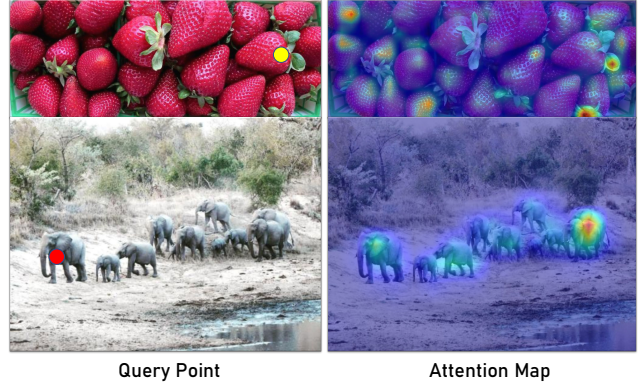


Figure 7. Exemplified attention maps using self-similarity map.

lose scale information in our method: one is the resizing of exemplars and the other is the pooling operation during feature extracting. To compensate for this loss, we propose to augment the exemplar’s feature with its corresponding scale embedding. We discretize the scale space into l_{total} levels. Each scale level is assigned with a d -dimensional embedding vector, yielding an embedding set whose cardinality equals l_{total} . Given an exemplar Z and query image X , we first derive Z ’s scale level $l(Z)$ by

$$l(Z) = \min \left(l_{\text{total}} - 1, \lfloor \left(\frac{h_Z}{2h_X} + \frac{w_Z}{2w_X} \right) \cdot l_{\text{total}} \rfloor \right), \quad (6)$$

where h_Z, w_Z, h_X, w_X denote images’ heights and widths. Then the scale embedding vector of level $l(Z)$ is retrieved and added back into the original feature. The scale embedding set is randomly initialized and learned during training, and stays fixed during inference.

3.5. Implementation Details

For a fair comparison, we apply the same pre-processing to query images and the feature extractor as in FamNet [29].

Data Pre-processing. We resize the query image while keeping its aspect ratio so that the length of its sides is limited within [384, 1584]. Exemplars are resized to 128×128 before fed into the feature extractor. No data augmentation is applied. During training, the size of all query images within a mini-batch is kept the same by zero-padding.

Network Architecture. The feature backbone consists of the first 4 blocks of ResNet-50 [12], which outputs the feature maps of 1024 channels. For each query image, the number of channels are reduced to 256 using 1×1 convolution. For each exemplar, the feature maps are first processed with global average pooling and then linearly mapped to obtain a 256D feature vector. The counter consists of a few convolution and bilinear upsampling layers to regress a density map of the same size as the query image. When computing channel attention weight α in BMNet+, we apply

Methods	Val MAE	Val MSE	Test MAE	Test MSE
GMN [21]	29.66	89.81	26.52	124.57
FamNet [29]	24.32	70.94	22.56	101.54
FamNet+ [29]	23.75	69.07	22.08	99.54
CFOCNet* [40]	21.19	61.41	22.10	112.71
BMNet (Ours)	19.06	67.95	16.71	103.31
BMNet+ (Ours)	15.74	58.53	14.62	91.83

Table 1. **Comparison with state of the art on the FSC147 dataset.** Best performance is in boldface.

a Linear(128)-ReLU-Linear(256)-Tanh structure, where the number in the bracket denotes the output dimension. Refer to supplementary material for more details.

Training Details. Our model is trained end-to-end. The backbone is initialized via SwAV [6]. Other parameters are randomly initialized. We apply AdamW [20] as the optimizer with a batch size of 8. The model is trained for 300 epochs with a fixed learning rate of $1e-5$. The weight α of similarity loss in Eq. 5 is set to $5e-6$ so that all the loss items are of the same order of magnitude. The total number of scale levels in Eq. 6 is empirically set to 20. We use PyTorch [27] as our experimental platform. Note that the BMNet+ consumes less than 12GB memory on a single GPU during training.

4. Experiments

Here we first showcase the advantage of our models over the state-of-the-art methods. We then validate each component in BMNet+. Next, we analyze the influence of exemplar numbers and discuss how to integrate features before feeding them to the counter. Finally, we show the cross-dataset generality of our method on a car counting dataset.

4.1. Comparison With State of the Arts

The FSC147 Dataset. FSC147 [29] is the first large-scale dataset for class-agnostic counting. It includes 6,135 images from 147 categories varying from animals, kitchen utensils, to vehicles. Given one query image, three instances of the same category are randomly chosen as the exemplars. To validate methods’ generality, the categories in training, validation, and test sets have no overlap. All experiments are done on FSC147 unless otherwise specified.

Comparing Methods. We mainly compare our models with two available CAC methods: GMN (General Matching Network [21]) and FamNet (Few-shot adaptation and matching Network [29]). Since FamNet executes fine-tuning during testing, we denote the fine-tuned version by FamNet+. The other compared methods apply no fine-tuning. Regarding our models, we validate two variants: 1) the baseline BMNet and 2) BMNet+ that implements all

core components, *i.e.*, self-similarity module, dynamic similarity metric, and direct supervision on similarity map. For more comparisons, we also test CFOCNet [40] that applies self-attention similar to our work. We reproduce CFOCNet as its code is unavailable and keep the same exemplar pre-processing and training configuration as in our methods. We denote this by CFOCNet*. Note that the main comparisons are concentrated on the public state-of-the-art FamNet.

Quantitative Results. As shown in Table 1, BMNet exhibits advantage over all of the compared methods with fixed similarity metrics (FamNet, GMN, and CFOCNet). Compared with FamNet, BMNet achieves a relative improvement of 21.63% w.r.t. validation MAE and 25.93% w.r.t. test MAE. Note that BMNet is already a strong baseline over FamNet, which indicates that BMNet is capable of characterizing a novel category without any concerned prior information. One can also observe that BMNet+ reduces the validation MAE by 18.23% and the test MAE by 12.51% compared with BMNet, which validates the effectiveness of our proposed components.

Qualitative Analysis. As shown in Fig. 8, both BMNet and BMNet+ output accurate density maps in whether dense or sparse scenes. Specifically, when counting hot air balloon (the 1st row), FamNet+ and BMNet mistake the tower for counting target, while BMNet+ offers comparatively better discrimination between target and background. In case of strawberries that exhibit large intra-class variation (the 2nd row), FamNet fails while our methods do not. This validates the effectiveness of our bilinear similarity metric (BMNet) and self-similarity in BMNet+. Refer to supplementary materials for more visualizations.

4.2. Ablation Study on BMNet+

Here we justify the effectiveness of each component in BMNet+. We start by testing the supervision on similarity map, because it directly affects the learning of self-similarity and dynamic similarity metric.

Supervision of the Similarity Map. By comparing B1 and B2 in Table 2, we can observe that direct supervision on the similarity map brings a relative improvement of 3.88% and 10.41% on the validation and test set w.r.t. MAE, respectively. This indicates that the similarity loss can help to learn a generic similarity metric.

Self-Similarity With Scale Embedding. Comparing B2 with B4 in Table 2, we can observe that applying self-similarity module and scale embedding improves the validation MAE by 0.97. However, the performance on the test set shows a converse phenomenon (cf. B7 vs. B9). A plausible explanation is that additional parameters in the feature extractor lead to an over-fitting problem. Note that, the comparisons of B3 vs. B4 and B8 vs. B9 show that scale embedding generally improves representations.

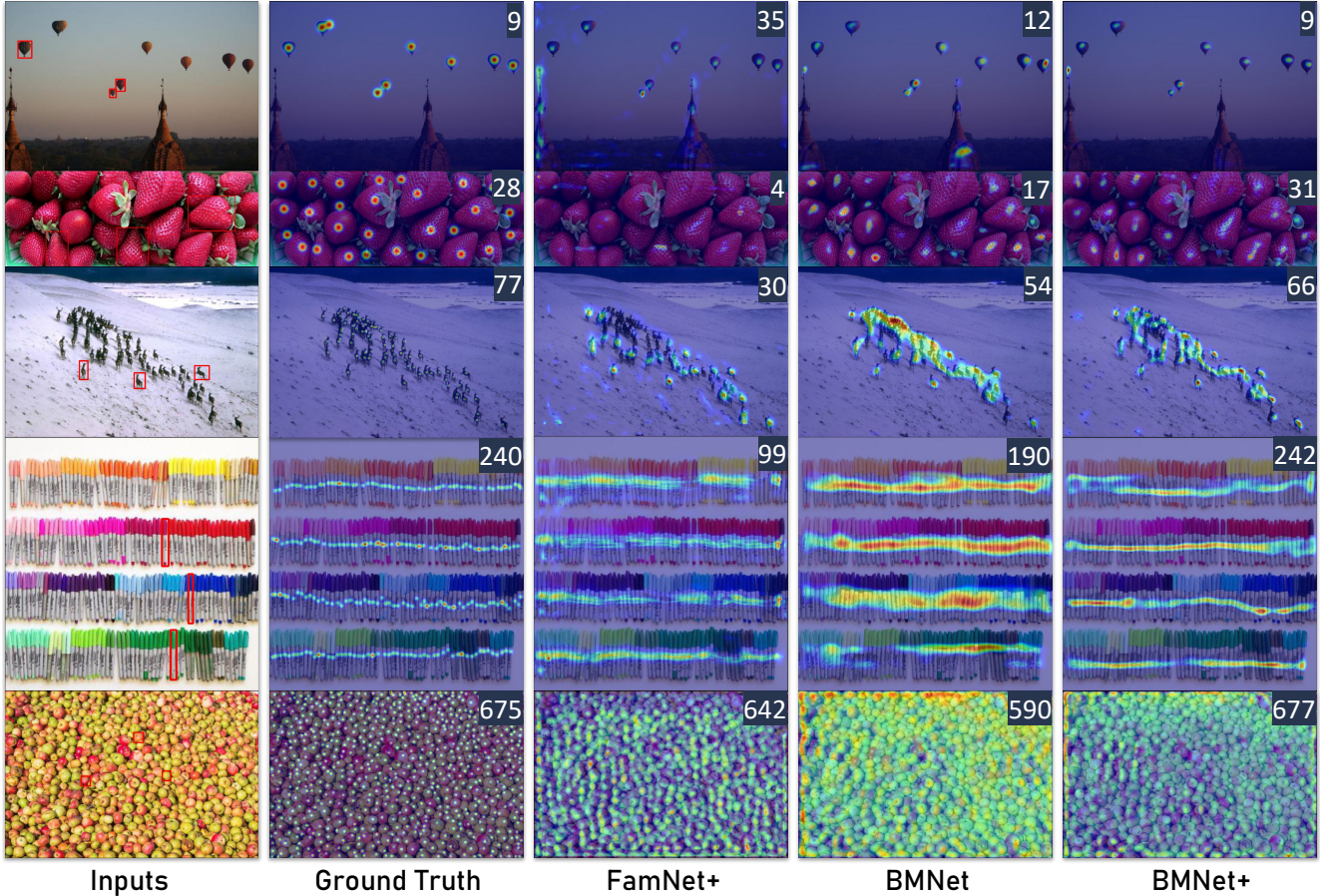


Figure 8. **Qualitative results on the FSC147 dataset.** The samples on the left exhibit significant intra-class variations such as scale, pose, and illumination condition. The red boxes indicate exemplars. Counting values are shown at the top-right corner. Our BMNet and BMNet+ can predict accurate density maps in both dense and sparse scenes. Best viewed by zooming in.

Dynamic Similarity Metric. The inclusion of dynamic similarity metric further brings a relative improvement of 9.28% on validation MAE and 11.29% on test MAE (cf. B4 vs. B5 and B9 vs. B10). In Sec. 3.2, we exemplify that dynamic similarity metric focuses on the exemplar-specific patterns to match similarity. Quantitative results here further demonstrate that the dynamic pattern selection mechanism can improve the naive bilinear similarity metric.

4.3. Number of Exemplars per Task

Here we investigate the impact of the number n of exemplars (randomly chosen) per task. Since the given maximum number of exemplars per query image is 3 in FSC147, we experiment with $n = 1, 2, 3$ and report their results in Table 3. It is foreseeable that more exemplars yield better results as in Table 3. Note that even our method with one single exemplar surpasses the other methods with three exemplars (cf. Table 1). This indicates that our method seeing only one exemplar could still capture information to describe the corresponding category. Besides, CAC meth-

No.	SL	SS	SE	DSM	Val MAE	Val MSE
B1	×	×	×	×	19.06	67.95
B2	✓	×	×	×	18.32	64.01
B3	✓	✓	×	×	17.44	67.07
B4	✓	✓	✓	×	17.35	60.28
B5	✓	✓	✓	✓	15.74	58.53
No.	SL	SS	SE	DSM	Test MAE	Test MSE
B6	×	×	×	×	16.71	103.31
B7	✓	×	×	×	14.97	92.88
B8	✓	✓	×	×	16.53	103.69
B9	✓	✓	✓	×	16.48	96.85
B10	✓	✓	✓	✓	14.62	91.83

Table 2. **Ablation study** on self-similarity (SS), dynamic similarity metric (DSM), similarity loss (SL), and scale embedding (SE).

ods may get more vulnerable to intra-class variations with fewer exemplars, but we find that self-similarity module offers obvious improvement within our method in this sce-

n	Val MAE	Val MSE	Test MAE	Test MSE
1	17.89	61.12	16.89	96.65
2	16.03	58.65	16.16	97.18
3	15.74	58.53	14.62	91.83

Table 3. **Impact of the number of exemplars.**

Combination	Val MAE	Val MSE	Test MAE	Test MSE
S	21.36	69.05	18.76	92.44
$x + z$	19.27	66.75	18.24	84.39
$x + z + S$	18.71	61.88	18.71	88.23
$x + S$ (default)	19.06	67.95	16.53	103.31

Table 4. **Ways to integrate features for the counter.** x and z stand for the features of query and exemplar, respectively, S for similarity map, and “+” for channel-wise concatenation.

nario. Refer to supplementary materials for more details.

4.4. How to Integrate Features for the Counter?

Here we discuss possible ways to integrate the features before feeding them to the counter. Given the exemplar feature z , the query feature x , and the similarity map S , we investigate 4 ways of feature combination as in Table 4, where “+” denotes channel-wise concatenation. According to the results, only using the similarity map to count objects yields the worst performance (the 1st row), while leveraging raw features of exemplars and query images can improve the counting performance (the 3rd and 4th rows). However, excluding the similarity map makes the supervision on the similarity metric impossible (the 2nd row). In addition, concatenating the features of exemplars brings marginal improvements but with increased computation overheads (the 3rd row). Therefore, to leverage the information within similarity map while also maintaining a moderate computational cost, we suggest the combination of similarity map and query features as the default representation. In supplementary materials, we also show that the query features may encode generic semantic information to help correct the mistakes within the similarity map.

4.5. Cross-Dataset Generalization

Following FamNet [29], we test our models’s generality on a car counting dataset CARPK [13]. CARPK contains 1, 448 images of parking lots in a bird view, which differs significantly from the images in FSC147. We exclude the “car” category within FSC147 to ensure that training and test categories have no overlap.

The results are reported in Table 5. We first focus on the models without fine-tuning on the CARPK dataset. It can be observed that our models exhibit strong generality.

Method	fine-tuned	MAE	MSE
FamNet	✓	18.19	33.66
BMNet	✓	8.05	9.70
BMNet+	✓	5.76	7.83
FamNet	×	28.84	44.47
BMNet	×	14.61	24.60
BMNet+	×	10.44	13.77

Table 5. **Generalization performance on the CARPK dataset.** All models are pretrained on the FSC147 dataset. “fine-tuned” denotes whether the pretrained models are further fine-tuned on the CARPK dataset.

Compared with FamNet, BMNet and BMNet+ obtain a relative performance gain of 49.3% and 63.8% on MAE, respectively. Moreover, BMNet and BMNet+ still retain their advantages when compared with FamNet in the fine-tuning scenario, which demonstrates that our designs are orthogonal to fine-tuning. In addition, the improvements of FamNet and BMNet after fine-tuning indicate the benefit of introducing task-specific information.

5. Conclusions and Limitations

In this work, we show that similarity modeling matters for CAC. In particular, we propose a similarity-aware framework for CAC where the feature representation and similarity metric are jointly learned in an end-to-end manner. Then we instantiate our framework with a naive BMNet that learns bilinear similarity. We also show how to extend the BMNet with the idea of exploiting self-similarity among features, learning dynamic similarity metric, and imposing explicit supervision on the similarity map. Both our BMNet and the extended BMNet+ achieve state-of-the-art performance on the large-scale dataset FSC147 and car counting dataset CARPK.

Limitations. Technically, we mainly focus on designing a better similarity metric, while how to obtain better feature representation is not well addressed: 1) the function of self-similarity module is intuitive, and Table 2 shows the self-similarity may hurt the performance on the test set; 2) how to integrate rich representation along with similarity map is also not addressed well in this work. Maybe transformer-based tracking [8] can help. In addition, since our goal is to present a generic framework, some designs in our instantiated models include some heuristics, which could be further studied in detail.

Acknowledgements. This work was funded by the National Natural Science Foundation of China under Grant No. 61876211 and No. 62106080, and the Chinese Fundamental Research Funds for the Central Universities under Grant No. 2021XXJS095.

References

- [1] Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. In *Proc. AAAI Conf. Artificial Intell.*, 2021. [2](#)
- [2] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In *Proc. Eur. Conf. Comput. Vis.*, pages 483–498, 2016. [1](#)
- [3] Mikhail J. Atallah. Faster image template matching in the sum of the absolute value of differences measure. *IEEE Trans. Image Process.*, 10(4):659–663, 2001. [2](#)
- [4] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional siamese networks for object tracking. In *Proc. Eur. Conf. Comput. Vis. Workshop*, pages 850–865, 2016. [2](#)
- [5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. Eur. Conf. Comput. Vis.*, pages 213–229, 2020. [5](#)
- [6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 33, pages 9912–9924, 2020. [6](#)
- [7] A. B. Chan, Zhang-Sheng John Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 1–7, 2008. [2](#)
- [8] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 8122–8131, 2021. [8](#)
- [9] Jian Cheng, Haipeng Xiong, Zhiguo Cao, and Hao Lu. Decoupled two-stage crowd counting and beyond. *IEEE Trans. Image Process.*, 30:2862–2875, 2021. [2](#)
- [10] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):743–761, 2012. [2](#)
- [11] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, volume 2, pages 1735–1742, 2006. [2](#)
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 770–778, 2016. [5](#)
- [13] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H. Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 4165–4173, 2017. [2, 8](#)
- [14] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(8):2011–2023, 2020. [4](#)
- [15] Victor Lempitsky and Andrew Zisserman. Learning To Count Objects in Images. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 23, pages 1324–1332, 2010. [2](#)
- [16] J. P. Lewis. Fast template matching. In *Vision Interface*, 1995. [2](#)
- [17] Liang Liu, Hao Lu, Haipeng Xiong, Ke Xian, Zhiguo Cao, and Chunhua Shen. Counting Objects by Blockwise Classification. *IEEE Trans. Circuits Syst. Video Technol.*, 2019. [2](#)
- [18] Liang Liu, Hao Lu, Hongwei Zou, Haipeng Xiong, Zhiguo Cao, and Chunhua Shen. Weighing counts: Sequential crowd counting by reinforcement learning. In *Proc. Eur. Conf. Comput. Vis.*, pages 164–181, 2020. [2](#)
- [19] Weiyang Liu, Zhen Liu, James M Rehg, and Le Song. Neural similarity learning. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 32, 2019. [3](#)
- [20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. Int. Conf. Learn. Repr.*, 2019. [6](#)
- [21] Erika Lu, Weidi Xie, and Andrew Zisserman. Class-agnostic counting. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Proc. Asi. Conf. Comput. Vis.*, pages 669–684, 2018. [1, 2, 6](#)
- [22] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 3265–3274, 2019. [2](#)
- [23] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Index networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):242–255, 2022. [2](#)
- [24] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 6142–6151, 2019. [2](#)
- [25] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Proc. Eur. Conf. Comput. Vis.*, pages 681–699, 2020. [2, 4](#)
- [26] Daniel Oñoro-Rubio and Roberto J. López-Sastre. Towards perspective-free object counting with deep learning. In *Proc. Eur. Conf. Comput. Vis.*, pages 615–629, 2016. [1](#)
- [27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and Others. PyTorch: An imperative style, high-performance deep learning library. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 8026–8037, 2019. [6](#)
- [28] Hamed Pirsiavash, Deva Ramanan, and Charless Fowlkes. Bilinear classifiers for visual recognition. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 22, 2009. [3](#)
- [29] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 3393–3402, 2021. [1, 2, 5, 6, 8](#)
- [30] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 29, 2016. [2](#)
- [31] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. *CoRR*, abs/2107.12746, 2021. [2](#)
- [32] Qingyu Song, Changan Wang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Jian Wu, and Jiayi Ma. To choose

- or to fuse? scale selection for crowd counting. In *Proc. AAAI Conf. Artificial Intell.*, 2021. [2](#)
- [33] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 6397–6406, 2020. [2](#)
- [34] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. [2](#)
- [35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 30, 2017. [5](#)
- [36] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution Matching for Crowd Counting. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 33, 2020. [2](#)
- [37] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 5265–5274, 2018. [2](#)
- [38] Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 18, 2006. [2](#)
- [39] Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 8361–8370, 2019. [2](#)
- [40] Shuo-Diao Yang, Hung-Ting Su, Winston H. Hsu, and Wen-Chin Chen. Class-agnostic few-shot object counting. In *Proc. Winter Conf. Appl. Comput. Vis.*, pages 869–877, 2021. [1](#), [2](#), [5](#), [6](#)
- [41] Tongtong Yuan, Weihong Deng, Jian Tang, Yinan Tang, and Binghui Chen. Signal-to-noise ratio: A robust distance metric for deep metric learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 4810–4819, 2019. [2](#)
- [42] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-Scene Crowd Counting via Deep Convolutional Neural Networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2015. [1](#), [2](#)
- [43] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks, 2019. [2](#), [5](#)
- [44] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 589–597, 2016. [2](#)