Scale-Prior Deformable Convolution for Exemplar-Guided Class-Agnostic Counting

Wei Lin[1]
wlin38-c@my.cityu.edu.hk
Kunlin Yang[2]
yangkunlin@sensetime.com
Xinzhu Ma[3]
xinzhu.ma@sydney.edu.au
Junyu Gao[4]
gjy3035@gmail.com
Lingbo Liu[5]
liulingbo918@gmail.com
Shinan Liu[2]
liushinan@sensetime.com
Jun Hou[2]
houjun@sensetime.com
Shuai Yi[2]
yishuai@sensetime.com
Antoni B. Chan[1]
abchan@cityu.edu.hk

[1] Department of Computer Science
City University of Hong Kong
Kowloon, Hong Kong SAR

[2] SenseTime Group Limited
Beijing, P.R. China

[3] The University of Sydney
NSW 2006, Australia

[4] School of Artificial Intelligence, Optics and Electronics
Northwestern Polytechnical University
Xi'an, P.R. China

[5] The Hong Kong Polytechnic University
Hunghom, Hong Kong SAR

Abstract

Class-agnostic counting has recently emerged as a more practical counting task, which aims to predict the number and distribution of any exemplar objects, instead of counting specific categories like pedestrians or cars. However, recent methods are developed by designing suitable similarity matching rules between exemplars and query images, but ignoring the robustness of extracted features. To address this issue, we propose a scale-prior deformable convolution by integrating exemplars' information, *e.g.*, scale, into the counting network backbone. As a result, the proposed counting network can extract semantic features of objects similar to the given exemplars and effectively filter irrelevant backgrounds. Besides, we find that traditional L2 and generalized loss are not suitable for class-agnostic counting due to the variety of object scales in different samples. Here we propose a scale-sensitive generalized loss to tackle this problem. It can adjust the cost function formulation according to the given exemplars, making the difference between prediction and ground truth more prominent. Extensive experiments show that our model obtains remarkable improvement and achieves state-of-the-art performance on a public class-agnostic counting benchmark. the source code is available at https://github.com/Elin24/SPDCN-CAC.

1 Introduction

In recent years, remarkable progress has been achieved in counting tasks. However, most methods only work in a category-specific manner, like counting crowd [34] or vehicles [20], and thus they fail to meet the requirements of some real-world applications. For example, there exist demands for counting goods in various categories in supermarkets or warehouses [7]; in agriculture, predicting the crop yield of different fruits/vegetables is required [14, 37]; and some may want to know the number of different trees [3, 21]. However, with traditional counting methods, a separate counting model is needed for each object class, which limits its practical applications.

To tackle the above problem, this paper considers class-agnostic counting, in which counting models predict the number and distribution of objects indicated by a few object exemplars in a set of query images. During training, both images and exemplars are input to the counting model, and then the loss is calculated between the predicted density maps and human-annotated dot maps [23]. Although existing class-agnostic counting methods have achieved good performance, there is still much room for improvement. For example, GMN [16] resizes the given exemplars to a fixed size and then calculates the distance between the exemplar's feature and local regions of the query image to localize the object of interest. One problem in this process is that exemplar features will lose the scale information provided by the exemplar's size. Although BMNet [26] adds a scale embedding to its network to tackle this problem, its function is not intuitive.

To take advantage of scale information, we design a Scale-Prior Deformable Convolution Network (SPDCN) to extract features of objects with specific size. SPDCN embeds the scale information into the deformable convolution, so that its receptive field is adjusted automatically and extracts features corresponding to the scale of the given exemplars. This design significantly boosts the counting performance because objects in the same category typically have similar scale in an image, whereas different object categories may have vastly different scales. With the extracted features, SPDCN then computes the similarity between exemplars and query images to segment out regions containing the counted objects. After that, the generated similarity map and features are sent to a decoder to estimate the density map.

We apply the generalized loss [31] to train SPDCN. However, we find that the vanilla generalized loss is unsuitable for class-agnostic counting because its cost function assumes all objects (people) are the same size, whereas in class-agnostic counting, different object categories have different scales. To tackle this problem, we propose a *scale-sensitive* generalized loss, in which the cost function is adjusted adaptively based on the object scale. Experiments show that the performance is further improved with our adaptive loss function.

To summarize, the key contributions of this paper are:

- To address class-agnostic counting, we propose a scale-prior deformable network to better extract exemplar-related features, followed by a segmentation-then-counting stage to count objects.
- We propose a scale-sensitive generalized loss to make the model training adaptive to objects of different sizes, boosting the performance and generalization of trained models.
- Extensive experiments and visualizations demonstrate these two designs work well, and outstanding performance is obtained when our model is tested on benchmarks.

2 Related Works

Class-Agnostic Counting. Previous counting tasks mainly aim at counting objects in a specific category. The most popular task is crowd counting [17, 19, 29, 31, 32, 34, 35]. Vehicle [11, 20], cell [8, 9] and animal [24] counting also attract researchers' attention, and are applied in various aspects like vehicular management [33], medical research [4], wildlife conservation [1], and so on. However, only a few methods have considered class-agnostic object counting, and relevant datasets are rare. FamNet [23] defines class-agnostic counting as predicting the number of given objects represented by only a few exemplars in the same image and constructs the first dataset called FSC-147 [23]. Its baseline model is designed based on self-similarities matching [25]. One problem is that the scale of exemplars is modeled by the kernel size, which is normally too large to compute, so FamNet freezes the parameters in the extractor to overcome this problem. Another similar work is the generic matching network (GMN) [16], which encodes the semantic feature of exemplars to an embedding, and then uses a matching network to model the relation between the exemplar embedding and the image's feature maps. However, GMN does not consider the scale problem because the size of the embedding vector is fixed. BMNet [26] considers the scale problem and adds scale embedding into its model, but it is not intuitive. Compared with these previous works, our proposed SPDCN embeds scale information into the deformable convolution so that the extracted feature can match the exemplar more accurately, yielding improved performance.

Deformable Convolution. Deformable convolution [2, 38] was proposed for modeling geometric transformations dynamically, and has been applied to video super-resolution [30], font generation [36] and other computer vision tasks. Compared to the previous works, we introduce scale-prior deformable convolution to class-agnostic counting, where the receptive fields of the counting network are adjusted according to the given exemplars.

Generalized Loss. The generalized loss [31] is designed based on the unbalanced optimal transport (UOT) problem. [31] prove that both L2 loss and Bayesian loss [18] are special cases of the generalized loss. In contrast to [31], which uses a fixed cost function assuming all objects are similar sizes, we propose a scale-sensitive generalized loss for class-agnostic counting, where different object categories have different sizes. Experimental results show that class-agnostic counting models perform better with the scale-sensitive generalized loss, compared to the original version.

3 Approach

3.1 Network Architecture

To solve the problem of class-agnostic counting, we design architecture different from previous works [16, 23], which is illustrated in Figure 1. Our architecture contains three key components: (1) scale-prior backbone; (2) counted objects segmentation module (orange part); and (3) class-agnostic density prediction module (blue part).

The scale-prior backbone is modified from an ImageNet pre-trained VGG-19 network [28] due to its strong representation ability to extract counting features from images. We keep the first ten convolutional layers of VGG-19 with three pooling layers. Feature maps output by the backbone are denoted as $F \in \mathbb{R}^{C\times H\times W}$, where C is the number of channels, and $H \times W$ is the spatial size. In the backbone, some specific layers are converted to **scale-prior deformable convolutions**, to utilize the crucial scale information provided by exemplars,

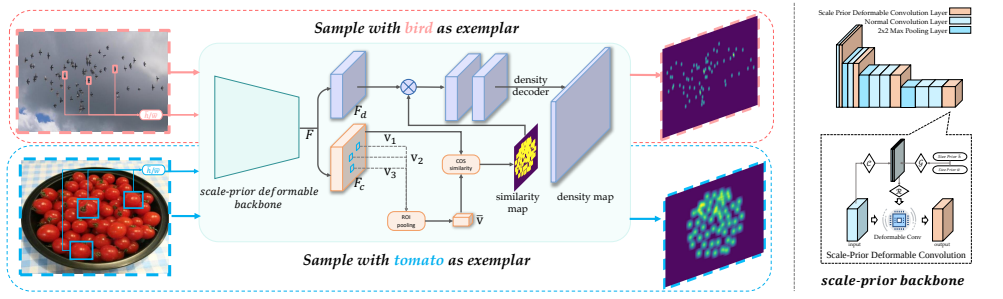

Figure 1: The overall architecture of the proposed SPDCN. Images with different exemplars are inputted into the model. Semantic features are firstly extracted by the scale-prior deformable backbone (details on right side), and then split into two branches, segmentation (orange boxes) and density estimation (blue boxes). The similarity map are obtained through the cosine similarity function. It is used to segment out the foreground of density feature maps, which is then passed to the density decoder to predict the density map.

which will be introduced in the next part. Afterward, the extracted feature map F is encoded as F_c and F_d by two linear functions to represent segmentation and density information, respectively.

In the segmentation branch, an ROIAlign layer [10] is applied to extract semantic vectors $\{v_1, \cdots, v_n\}$ ($n \leq 3$ is the number of labeled exemplars) according to given box coordinates to represent each exemplar. Next, a class-specific representation vector $\bar{v}$ is obtained by averaging these vectors, and the cosine similarity $\tilde{s}_i$ between $\bar{v}$ and each local feature vector f_i of the whole feature map F_c is calculated:

$$\tilde{s}_i = \frac{\bar{v}^\top f_i}{\|\bar{v}\|_2 \cdot \|f_i\|_2}, \quad f_i \in F_c, \quad \bar{v} = \frac{1}{n}\sum_j^n v_j. \tag{1}$$

The predicted similarity map $\tilde{S}$ indicates which regions are foreground or background is constructed by arranging the $\tilde{s}_i$ in spatial order.

The density prediction module uses the similarity results to highlight these regions containing counted objects. In this branch, element-wise multiplication is adopted between F_d and $\tilde{S}$ to generate the class-agnostic density feature, which is then inputted to the final decoder module. Our decoder follows the design of PSCC [35], which uses pixel-shuffling [27] to upsample the feature map, and produces the final density map $\tilde{D}$ with the same size as input images. More details are presented in the supplemental.

It should be noted that our SPDCN is different from previous works [12] that adopt a human head segmentation mask as attention to refine density map results. In our case, we obtain a similarity map under the guidance of exemplar object features, and the similarity map $\tilde{S}$ is used to generate a class-agnostic density feature.

3.2 Scale-Prior Deformable Convolution

Previous exemplar-guided class-agnostic counting methods like GMN [16] usually ignore the scale information. In FamNet [23], exemplars are directly used as convolution kernels so that the scale information is reflected by the kernel size. To take advantage of size information provided by exemplars, here we propose scale-prior deformable convolution to filter

features that are not consistent with given scales and only extract helpful features for class-agnostic counting. Specifically, we use the given exemplars' scales to learn the corresponding scale embeddings, and then adjust the receptive fields of the deformable convolutions according to these embedding vectors.

For a convolution kernel $\mathbf{w}$ with size of $2r+1$, the output value $\mathbf{y}(p)$ while convolving it with input feature map $\mathbf{x}$ at location p is calculated as:

$$\mathbf{y}(p) = \sum_{j \in \mathcal{R}} \mathbf{w}(j) \cdot \mathbf{x}(p+j),\ \mathcal{R} = \{(-r,-r),(-r,-r+1),\cdots,(r,r-1),(r,r)\}. \tag{2}$$

Different from normal convolution, deformable convolution [2] adds learnable offsets Δj to j and (2) becomes:

$$\mathbf{y}(p) = \sum_{j \in \mathcal{R}} \mathbf{w}(j) \cdot \mathbf{x}(p+j+\Delta j). \tag{3}$$

In (3), traditionally Δj is directly learned by mapping $\mathbf{x}$ into an offset tensor [2]. However, exemplars' scales (width and height) are provided in class-agnostic counting. Instead of learning offsets Δj from the input feature map, we believe that the exemplar is a better source to get the offset of the deformable convolution. According to this idea, the offsets of our scale-prior deformable convolution are made up of two parts, the local scale embedding d_c and the global scale embedding d_g.

The local embedding d_c is similar to the vanilla deformable convolution [2], which only utilizes the preceding feature maps to learn offsets locally. In other words, $d_c = \mathcal{C}(\mathbf{x})$ is obtained by applying a non-linear convolutional block $\mathcal{C}$ to the previous image feature maps x. The global scale embedding d_g is obtained from the exemplars' scale, and is the same for the whole feature map. Specifically, d_g is represented by mapping the average width $\bar{w}$ and height $\bar{h}$ of the exemplars to a vector:

$$d_g = \mathcal{G}(\bar{h},\bar{w}),\ \ \bar{h} = \sum_{e_i \in E_I} \frac{h_{e_i}}{|E_I|}, \bar{w} = \sum_{e_i \in E_I} \frac{w_{e_i}}{|E_I|}, \tag{4}$$

where E_I is the exemplar set, h_{e_i} and w_{e_i} is the height and width of the i-th exemplar e_i. Similar to $\mathcal{C}$, $\mathcal{G}$ is a non-linear function transferring a two-dimensional vector to the global scale embedding d_g. Notably, d_g is only related to the size of examples but unrelated to its semantic information. Since d_g is a single vector, it should be expanded along the spatial dimensions to make it the same size as d_c. Afterwards, they are concatenated and another non-linear convolutional block $\mathcal{R}$ is applied to obtain offsets for deformable convolutions. The whole process is displayed in the right part of Figure 1.

Guided by the scale prior, the deformable convolution is able to adjust its receptive fields accordingly, which plays a positive role in extracting an exemplar-related representation. In this work, the last convolution layers in all four stages of VGG-19 [28] are replaced with scale-prior deformable convolutions.

3.3 Scale-Sensitive Generalized Loss

In object-specific counting, the ground truth is generated by convolving the labeled dot map with a Gaussian kernel, and typically the L2 loss is used to measure the quality of the predicted density map. Recently, Generalized loss [31] directly measures the distance between the predicted density map and dot map through an unbalanced optimal transport problem, and shows that L2 loss with Gaussian-blurred ground truth is a special case. The distance

(cost) function in generalized loss reflects the Gaussian kernel size in L2 loss, and both are fixed in object-specific counting. However, in exemplar-guided class-agnostic counting, we can adjust the cost function according to the given exemplars. Inspired by this, we present scale-sensitive generalized loss for improving the training of class-agnostic counting models.

Specifically, the generalized loss is formulated as:

$$\mathcal{L}_{\mathbf{C}} = \min_{\mathbf{P}} \langle \mathbf{C}, \mathbf{P} \rangle - \varepsilon H(\mathbf{P}) + \tau \|\mathbf{P}\mathbf{1}_m - \mathbf{a}\|_2^2 + \tau \|\mathbf{P}^\top \mathbf{1}_n - \mathbf{b}\|_1, \tag{5}$$

where $\mathbf{a}$ is the predicted density map, $\mathbf{b} = \mathbf{1}_m$ plays the role of the target, $H(\cdot)$ is the entropic regularization term, $\mathbf{P}$ is the corresponding unbalanced optimal transport plan, and $\mathbf{C}$ is the cost matrix that defined by the euclidean distance between each pixel in the density map and ground truth. To add the scale information of the counted objects, we define the cost matrix as:

$$\mathbf{C}_{ij} = \|\hat{x}_i - \hat{y}_j\|_2, \quad \begin{bmatrix} \hat{x}_i & \hat{y}_j \end{bmatrix} = \begin{bmatrix} 1/s_h & 0 \\ 0 & 1/s_w \end{bmatrix} \begin{bmatrix} x_i & y_j \end{bmatrix}, \tag{6}$$

in which x_i is the 2-d coordinate of the i-th pixel in the predicted density map, y_i is the 2-d coordinate of the j-th labeled point in the ground truth. s_h and s_w are scale factors that align the two axes. To avoid extreme value, we take a modified sigmoid function $\mathcal{S}(\cdot)$ to limit the range of these two factors:

$$\mathcal{S}(k) = \frac{\alpha}{1 + \exp(-(k-\mu)/\sigma)} + \beta, \tag{7}$$

and then define $s_h = \mathcal{S}(\bar{h})$ and $s_w = \mathcal{S}(\bar{w})$, where $\bar{h}$ and $\bar{w}$ are the average of exemplars' height and width, as in (4). Here μ defines the symcenter of the S-curve; σ is used to smooth the curve; α and β control the range and minimum value of output. These four hyperparameters should be set accordingly before training.

4 Experiments

4.1 Dataset & Evaluation Metrics & Implementation Details

The dataset for class-agnostic counting FSC-147 is introduced by FamNet [23], in which it is proposed for a few-shot counting task. The dataset contains 147 object categories. The whole dataset is split into train, validation, and test sets, and each set does not share any object category with other sets. Thus, the objects in the test set are not seen during the training process, which is also the reason why it is regarded as few-shot counting.

Besides FSC-147, a car counting dataset CARPK [11] is also used to explore whether a model counting general objects can be applied to counting objects in a specific category. There are 989 training and 459 test images, with a total of 89,777 cars. All images are collected with drones from 4 different parking lots.

The counting task is evaluated using Mean Absolute Error (MAE) and root Mean Squared Error (MSE):

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |C_i - \tilde{C}_i| \; ; \; \text{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |C_i - \tilde{C}_i|^2}, \tag{8}$$

where N donates the number of samples in the test dataset, and C_i and $\tilde{C}_i$ are the predicted count number and ground truth of the i-th sample, respectively.

Methods		Validation Set		Test Set	
		MAE	MSE	MAE	MSE
FR FSD [13]	ICCV'19	45.45	112.53	41.64	141.04
FSOD FSD [5]	CVPR'20	36.36	115.00	32.53	140.65
MAML [6]	PRML'17	25.54	79.44	24.90	112.68
GMN [16]	ACCV'18	29.66	89.81	26.52	124.57
FamNet [23]	CVPR'21	23.75	69.07	22.08	99.54
VCN [22]	CVPRW'22	19.38	60.15	18.17	95.60
BMNet [26]	CVPR'22	15.74	58.53	14.62	**91.83**
SPDCN (ours)		15.55	51.00	14.48	100.01
SPDCN† (ours)		**14.59**	**49.97**	**13.51**	96.80

Table 1: Comparison with state of the arts. † indicates training with our scale-sensitive generalized loss.

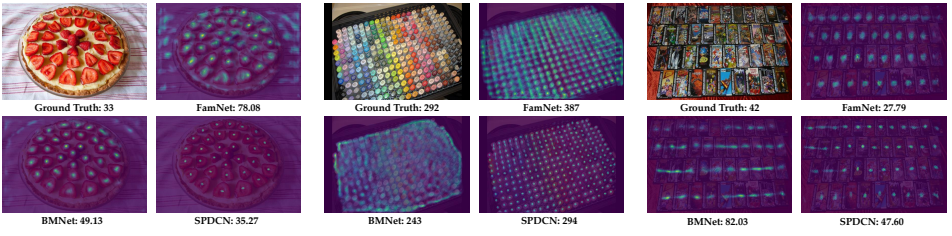

Figure 2: Visualization of density maps predicted by SPDCN and some baselines on FSC-147 dataset. The proposed SPDCN can estimate more sophisticated density maps than other models.

For efficient training, we pad images having lower resolutions with zeros, and resize images having higher resolutions, in order to keep the sizes of all samples the same at 576×384. During training, we set the batch size to 16. AdamW [15] optimizer is adopted with a learning rate of $5e$-5. All models are trained for 50 epochs, and the learning rate decays with a rate of 0.95 after each epoch. $\mathcal{C}$ and $\mathcal{R}$ in the scale-prior deformable convolution are two layers of convolutions with the ReLU active function. $\mathcal{G}$ is constructed by two fully connected layers. Empirically, α, β, μ and σ are set to 256, 128, 64 and 32, respectively.

4.2 Comparison With State-of-the-Arts

We first compare our method with state-of-the-art class-agnostic models on the FSC-147 dataset. As shown in Table 1, few-shot detection based methods like FR [13] and FSOD [5] cannot detect all objects accurately, achieving MAE larger than 30. Compared with them, the meta-learning method MAML [6] obtains MAE of 25.54. Self-similarity-based methods like GMN [16] and FamNet [23] perform better and obtain MAEs of 26.52 and 22.08 respectively. VCN [22] addresses the scarcity of annotated data by using a generator to produce augmented images, and then a counting network predicts the original and augmented image for better generalization. Experiments show that VCN gets a lower MAE of 18.17 than FamNet. BMNet [26] adopts a dynamic similarity metric for better matching between exemplars and query images, which decreases MAEs on the validation and test set to 15.74 and 14.62, respectively.

embedding		validation set		test set	
global	local	MAE	MSE	MAE	MSE
		21.24	71.16	19.41	128.26
	✓	19.99	69.49	17.58	119.69
✓		16.63	53.30	14.54	104.69
✓	✓	**14.59**	**49.97**	**13.51**	**96.80**

Table 2: Results of SPDCN with different deformable convolution types.

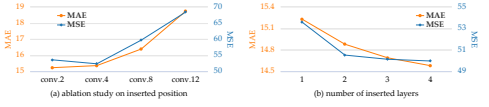

Figure 3: Ablation study on the position and number of scale-prior deformable convolutions.

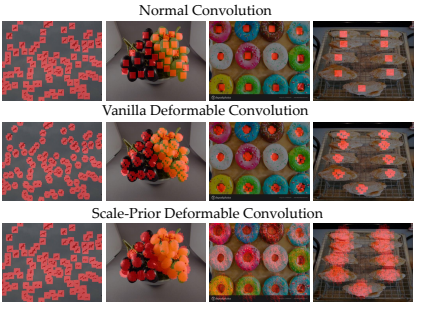

Figure 4: The receptive field comparison among different convolutions.

However, all these models pay little attention to leveraging the scale information in exemplars. In contrast, our SPDCN utilizes the scale information through the proposed scale-prior deformable convolution in a simple matching network, and dramatically improves the counting performance. Training with vanilla generalized loss, SPDCN achieves MAE 14.48, which is is lower than BMNet. If scale-sensitive generalized loss is adopted (SPDCN†), the MAE is further decreased to 13.51, demonstrating our design's effectiveness. Figure 2 presents several examples from the FSC-147 validation set. Although these counted objects vary in scale and quantity, our model obtains good compact predictions compared to other models.

4.3 Ablation Study on Scale-Prior Deformable Convolution

We next present an ablation study on the scale-prior deformable convolution. We compare the proposed method with vanilla deformable and normal convolutions, and then visualize their receptive fields on a few examples. Furthermore, we also explore whether the position and number of scale-prior deformable convolution layers affect the counting performance.

In Table 2, we compare the deformable convolutions with different embeddings. Normal convolutions achieve MAE and MSE of 19.41 and 128.26, respectively. When we replace it with vanilla deformable ones (local embedding) under the same setting shown in Figure 1, the counting errors decrease to 17.58 and 119.69. This illustrates that vanilla deformable convolution helps the model adjust its receptive field adaptively and improve counting performance. However, the estimated scale information is inaccurate, leading to limited improvement. In contrast, the scale-prior version adjusts the receptive fields according to the exact object size information provided by exemplars (global embedding), and the counting errors are reduced to 14.54 and 104.69. Obviously, the receptive field of SPDCN with only global embedding cannot cover all instances due to their variable scale. When combining both, the global embedding gives coarse scale prior and the local embeddings adjust it adaptively, resulting in the best performance (last row in Table 2).

Figure 4 visualizes the receptive fields (red regions) of the output feature maps when models encounter objects with different scales. In the first column, the receptive fields of normal convolution are enough to capture the each instance (bird), so the offsets of both deformable convolutions are close to zero, and the receptive fields (red regions) in these three images are similar. When the object scale increases, normal convolution keeps the same

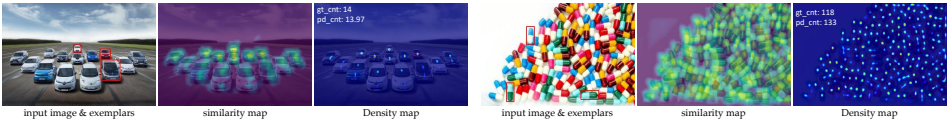

Figure 5: Examples of instances with (left) different scales and (right) different rotations.

receptive field. Vanilla deformable convolution is able to capture a slight change of object scale but cannot work on objects of larger size, which means it is hard to learn scale information of large objects. However, the scale-prior deformable convolution learns the receptive field from both the local features as well as the scale-prior from the given exemplars, which allows the network to capture feature of counted objects with large scale, as illustrated in the right column.

Figure 3(a) studies in which position the scale-prior deformable convolution layer will contribute the largest to counting performance. Putting the scale-prior layer at early layers improves performance more than latter layers. This makes sense because high-level feature maps have smaller resolutions than low-level ones. If the scale of objects is small, adjusting the receptive fields of high-level layers will harm the output features. In contrast, deformable convolution at low-level layers can cover big objects using large offsets (Δj in (3)), and only capture a small region for small objects. Nevertheless, replacing only the final layer with scale-prior deformable convolution still outperforms vanilla deformable convolution (MAE: 18.74 vs. 19.99).

In Figure 3(b), we insert the scale-prior deformable convolutions to the backbone one by one to explore the effect of the number of these layers. The error decreases gradually as more layers are inserted. However, too many deformable layers also requires more computation time. Thus in our model, only one scale-prior deformable convolution is placed in each stage (totaling 4 layers) to maximize the efficiency.

One concern with SPDCN might be that the scale prior is fixed so that it might not handle objects with various scales or rotations. Note that *we only use the scale prior to adjust the receptive field of the network – the matching process and density estimation are based on semantic information instead of scale information.* Furthermore, besides the scale prior, local embeddings are used in the deformable convolution to dynamically adjust the receptive field to each object's appearance. Figure 5 visualizes two example images of objects with various scales and rotation angles. In Figure 5(left), the scales of instances and exemplars change dramatically, and the matching result (similarity map) highlights all instances of interest. In Figure 5(right), instances are rotated at different angles, and our network handles the object-rotation problem by learning rotation-invariant features for matching. Both samples show that our SPDCN is able to work under scale-variance and rotation-variance, thus predict counting results close to ground truth.

4.4 Ablation Study on Scale-Sensitive Generalized Loss

We next conduct an ablation study comparing the proposed scale-sensitive generalized loss with the vanilla one, as well as the corresponding versions of L2 loss, as presented in Table 3. For L2 loss, we use (7) to change the size of the Gaussian kernel, so that a large object also has a large kernel. The results show that the scale-sensitive L2 loss can guide the model to have better generalization no matter the trained model has a common backbone (MAE: 23.67

VGG-19	L2 loss		Generalized loss	
	vanilla	scale-sensetive	vanilla	scale-sensetive
w/o scale-prior	23.67/72.81	22.85/70.90	21.60/71.83	21.23/70.75
w/ scale-prior	15.89/52.24	15.45/50.18	15.55/51.00	**14.59/49.97**

Table 3: Ablation study on scale-sensitive generalized loss (MAE/MSE).

	w/o fine-tuning			w/ fine-tuning		
method	FamNet	BMNet	SPDCN	FamNet	BMNet	SPDCN
MAE	28.84	**17.30**	18.15	18.19	**9.66**	10.07
MSE	44.47	21.89	**21.61**	33.66	14.84	**14.12**

Table 4: Experimental results on CARPK dataset.

$\rightarrow$ 22.85) or scale-prior deformable ones (MAE: 16.15 $\rightarrow$ 15.89). When generalized loss is applied, the performance is also improved, and errors are further decreased (MAE: 15.55 $\rightarrow$ 14.59). These experiments demonstrate that the scale-sensitive loss is more suitable for exemplar-guided class-agnostic counting.

4.5 Test on CARPK

Following FamNet [23], we next present experimental results on CARPK [11]. To adapt the counting everything model to the specific object counting, twelve labeled objects (cars) are randomly selected as exemplars from the training set. As shown in Table 4, SPDCN is compared with FamNet [23] and BMNet [26] (the results of BMNet are reproduced with our code). Without fine-tuning, FamNet achives an MAE of 28.84, which is much worse than BMNet (17.30) and ours SPDCN (18.15). SPDCN achieves the lowest MSE of 21.61. After fine-tuning on the CARPK training set, all models obtain smaller errors than before. FamNet (18.19) is close to SPDCN without fine-tuning. The MAE of BMNet and SPDCN are decreased to 9.66 and 10.07. Our model acheives the smallest MSE of 14.12.

5 Conclusion

In this paper, we explore exemplar-guided class-agnostic counting. To take advantage of scale information provided by exemplars, scale-prior deformable convolution is proposed to adjust the receptive fields according to the given exemplars. Experimental results show that this operation decreases counting errors dramatically and gives a more accurate density distribution. We also propose scale-sensitive generalized loss to adapt the cost function according to exemplars, so that different training samples with different object scales have their own distance function for optimal transport. This new loss further helps our model perform better than previous models on the class-agnostic counting benchmark.

Acknowledgements

This work was supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11212518, CityU 11215820), and by a Strategic Research Grant from City University of Hong Kong (Project No. 7005665).

References

[1] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In *European conference on computer vision*, pages 483–498. Springer, 2016.

[2] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[3] Khelifa Djerriri, Mohamed Ghabi, Moussa Sojiane Karoui, and Reda Adjoudj. Palm trees counting in remote sensing imagery using regression convolutional neural network. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2627–2630. IEEE, 2018.

[4] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1): 67–70, 2019.

[5] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2020.

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[7] Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5236, 2019.

[8] Yue Guo, Jason Stein, Guorong Wu, and Ashok Krishnamurthy. Sau-net: A universal deep network for cell counting. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 299–306, 2019.

[9] Yue Guo, Oleh Krupa, Jason Stein, Guorong Wu, and Ashok Krishnamurthy. Sau-net: A unified network for cell counting in 2d and 3d microscopy images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–1, 2021. doi: 10.1109/TCBB.2021.3089608.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[11] 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.

[12] Shengqin Jiang, Xiaobo Lu, Yinjie Lei, and Lingqiao Liu. Mask-aware networks for crowd counting. *IEEE Transactions on Circuits and Systems for Video Technology*, 30 (9):3119–3129, 2019.

[13] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019.

[14] Bruno T Kitano, Caio CT Mendes, André R Geus, Henrique C Oliveira, and Jefferson R Souza. Corn plant counting using deep learning and uav images. *IEEE Geoscience and Remote Sensing Letters*, 2019.

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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

[17] Ao Luo, Fan Yang, Xin Li, Dong Nie, Zhicheng Jiao, Shangchen Zhou, and Hong Cheng. Hybrid graph neural networks for crowd counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11693–11700, 2020.

[18] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6142–6151, 2019.

[19] Zhiheng Ma, Xing Wei, Xiaopeng Hong, Hui Lin, Yunfeng Qiu, and Yihong Gong. Learning to count via unbalanced optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2319–2327, 2021.

[20] T Nathan Mundhenk, Goran Konjevod, Wesam A Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *European Conference on Computer Vision*, pages 785–800. Springer, 2016.

[21] Lucas Prado Osco, Mauro dos Santos de Arruda, José Marcato Junior, Neemias Buceli da Silva, Ana Paula Marques Ramos, Érika Akemi Saito Moryia, Nilton Nobuhiro Imai, Danillo Roberto Pereira, José Eduardo Creste, Edson Takashi Matsubara, et al. A convolutional neural network approach for counting and geolocating citrus-trees in uav multispectral imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160: 97–106, 2020.

[22] Viresh Ranjan and Minh Hoai. Vicinal counting networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4221–4230, 2022.

[23] 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.

[24] Farah Sarwar, Anthony Griffin, Priyadharsini Periasamy, Kurt Portas, and Jim Law. Detecting and counting sheep with a convolutional neural network. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[25] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[26] 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 (CVPR)*, pages 9529–9538, June 2022.

[27] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[29] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2576–2583, 2021.

[30] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu TDAN. temporally-deformable alignment network for video super-resolution. in 2020 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3357–3366, 2020.

[31] Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1974–1983, 2021.

[32] 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.

[33] Qi Wang, Jia Wan, and Xuelong Li. Robust hierarchical deep learning for vehicular management. *IEEE Transactions on Vehicular Technology*, 68(5):4148–4156, 2019. doi: 10.1109/TVT.2018.2883046.

[34] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2141–2149, 2020.

[35] Qi Wang, Wei Lin, Junyu Gao, and Xuelong Li. Density-aware curriculum learning for crowd counting. *IEEE Transactions on Cybernetics*, 2020.

[36] Yangchen Xie, Xinyuan Chen, Li Sun, and Yue Lu. Dg-font: Deformable generative networks for unsupervised font generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5130–5140, 2021.

[37] Laura Zabawa, Anna Kicherer, Lasse Klingbeil, Reinhard Töpfer, Heiner Kuhlmann, and Ribana Roscher. Counting of grapevine berries in images via semantic segmentation using convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 164:73–83, 2020.

[38] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019.