Sparse Context Transformer for Few-Shot Object Detection

Mingyuan Jiu, Jie Mei, Hichem Sahbi, Xiaoheng Jiang, Mingliang Xu

Sparse Context Transformer for Few-Shot Object Detection*

Mingyuan Jiu*[1,2,3], Jie Mei[1], Hichem Sahbi*[4], Xiaoheng Jiang[1,2,3], and Mingliang Xu*[1,2,3]

[1] School of Computer and Artificial Intelligence, Zhengzhou University, China
[2] Engineering Research Center of Intelligent Swarm Systems, Ministry of Education, Zhengzhou University, Zhengzhou, China
[3] National SuperComputing Center in Zhengzhou, Zhengzhou, China
[4] Sorbonne University, CNRS, LIP6, F-75005, Paris, France
iemyjiu@zzu.edu.cn, hichem.sahbi@lip6.fr, iexumingliang@zzu.edu.cn

Abstract. Few-shot detection is a major task in pattern recognition which seeks to localize objects using models trained with few labeled data. One of the mainstream few-shot methods is transfer learning which consists in pretraining a detection model in a source domain prior to its fine-tuning in a target domain. However, it is challenging for the fine-tuned models to effectively identify new classes in the target domain, particularly when the underlying labeled training data are scarce. In this paper, we devise a novel sparse context transformer (SCT) that effectively leverages object knowledge in the source domain, and automatically learns a sparse context from only few training images in the target domain. As a result, it combines different relevant clues in order to enhance the discrimination power of the learned detectors and reduce class confusion. We evaluated the proposed method on two challenging few-shot object detection benchmarks, and empirical results show that the proposed method obtains competitive performance compared to the related state-of-the-art.

Keywords: Few-shot object detection · Transfer-learning · Transformer · Sparse context

1 Introduction

Although deep learning (DL) models have achieved remarkable performance, these outstanding models are usually label-hungry and their training is time and memory demanding [1,5,10,15,20]. In some scenarios, particularly object detection, labeled data are scarce, and this makes DL-based detection a major challenge [2,3]. Among existing solutions that mitigate scarcity of labeled data, transfer learning is particularly effective and consists in pretraining detection

* This work is supported in part by the National Natural Science Foundation of China (No. U22B2051, 62272422, 62172371, U21B2037), and in part by the National Key Research and Development Program of China under Grant 2021YFB3301500.

models in the source domains — using abundant labeled data — prior to their fine-tuning in the target domains. However, as labeled data are scarce in the target domains, few-shot object detection — based on transfer learning [4,6,7,8,9] — is not yet sufficiently effective in identifying new object classes.

The aforementioned issue is accentuated in few-shot object detection as this task involves both *localization and classification*. Localization focuses on spatial information which is decently obtained from pretrained models in the source domains. Therefore, bounding box regressors (BBOX) trained in the source domain are already reliable for initialization and fine-tuning in the target domain. Hence, detectors fine-tuned with few training samples may effectively locate new object classes. In contrast, classification often requires contextual knowledge of specific categories. In other words, source domain knowledge are insufficient to learn new category distributions in the target domain. Therefore, the underlying models should be completely retrained for new categories. However, scarcity and limited diversity of data in the target domain lowers the accuracy of the new learned category classifiers, and thereby leads to class confusion.

In order to address these issues, we propose in this paper a novel *sparse context transformer* (SCT) that leverages source domain knowledge together with few training images in the target domain. This transformer learns sparse affinity matrices between BBOXs and classification outputs by exploring the most relevant contexts for new object categories in the target domain. Additionally, incorporating SCT during the fine-tuning stage allows the model to focus on the most relevant contextual information and then to refine the decision boundaries between visually similar objects, leading to a more powerful model ability to accurately distinguish between different classes. Through this approach, SCT not only improves recognition performance on new classes but also reduces the likelihood of misclassification during fine-tuning. The proposed transformer consists of two simple yet effective sub-modules: sparse relationship discovery, and aggregation. In the first sub-module, contextual fields are initially designed based on default prior boxes (also called anchor boxes) [11,12] and multi-scale feature maps extracted from a visual encoder. Relationships between each prior box and contextual fields are modeled through a novel sparse attention. In the second sub-module, aggregation further leverages the learned relationships and integrates contextual fields into the relevant prior boxes. Our proposed transformer enhances prior box representations, and mitigates confusion in few-shot object detection and classification. Hence, our contributions include:

- A novel sparse context transformer that effectively explores useful contextual fields from a small number of labeled images. This transformer is embedded into an SSD (plug-and-play style) detector suitable for few-shot object detection.
- A novel attention layer that assists object detection in learning task-relevant knowledge from images by enhancing the underlying task-related features.
- A comprehensive evaluation of our proposed method on the challenging configurations for few-shot detection that shows high performance.

2 Related Work

Recent years have witnessed a significant progress in few-shot object detection, primarily focusing on two core areas: meta-learning and transfer learning. Within the meta-learning framework, researchers have proposed various innovative strategies to address the challenges imposed by few samples. Xiao et al. [13] introduced a meta-learning-based approach that addresses misdetection in the resource-constrained scenarios by integrating query features (ROI generation) with class-related features. Tian et al. [14] presented a model-agnostic meta-learning framework that effectively enhances the cross-domain generalization capabilities of existing meta-learning methods. Han et al. [16] recognized that current few-shot detection models tend to be biased towards base classes while being variance-sensitive to new classes, and proposed a variational feature aggregation method based on meta-learning.

Unlike these meta-learning approaches, which design complex meta-learning models for challenging meta-learning tasks, transfer learning-based methods are often simpler and more efficient. Chen et al. [17] introduced a low shot transfer detector that focuses on foreground objects in the target domain during fine-tuning in order to learn more knowledge on the targeted categories. Khandelwal et al. [18] conjectured that simple fine-tuning may lead to a decrease in the transferability of the models. Hence, they proposed a unified semi-supervised framework that combines weighted multi-modal similarity measures between base and novel classes. With this method, they achieved effective knowledge transfer and adaptation. Unlike these methods, Yang et al. [19] proposed a context-transformer that tackles object confusion in few-shot detection. This transformer relies on a set of contextual fields from different spatial scales and aspect ratios of prior boxes, in order to explore their relationships through dot products. Based on these relationships, the contextual fields are integrated into each prior box and this improves their representation.

Our work is an extension of context transformers that addresses the relatively monotonous contextual fields (constructed in the original version of these transformers) as well as their relationships with prior boxes, which cannot effectively suppress task-independent contextual fields, and further affect the model's ability to recognize novel classes. In this regard, we consider informations from different sources and we model sparse relationships between contextual fields and each prior box to help the model selecting the most effective fields. This also mitigates confusion in few-shot object detection.

3 Proposed Method

In this section, we introduce our novel sparse context transformer. As shown in Fig. 1, our framework relies on an SSD-style detector [12] used as a flexible plug-and-play backbone that delivers rich multi-scale contextual information. The SSD detector consists of K (spatial-scale) heads including bounding boxes regressor (BBOX) and object+background classifiers (OBJ+BG). To generalize few-shot learning in the target domain, we first pretrain the SSD detector

with a large-scale dataset in the source domain. Then, we combine the proposed transformer module with the SSD detector for fine-tuning in the target domain (see again Fig. 1). As shown subsequently, our proposed transformer includes two submodules: one for sparse relationship discovery, and another one for aggregation. These submodules are respectively used to model context/classifier relationships and for context fusion.

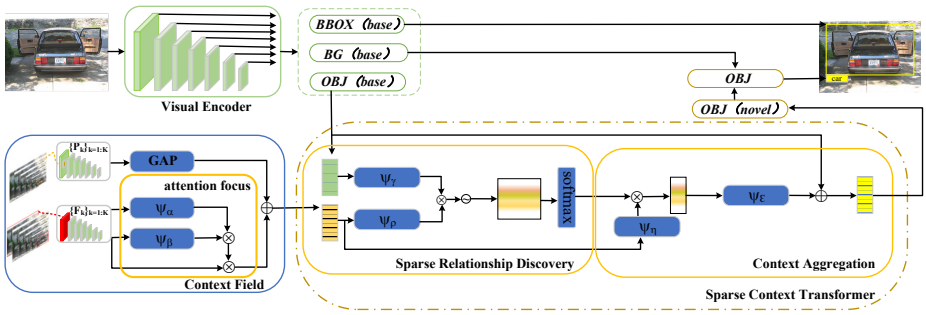


Fig. 1: Few-Shot Detection with Sparse-Context-Transformer. It consists of sparse relationship discovery and context aggregation, which can effectively utilize the context fields of few-shot tasks, improve the context awareness ability of each prior box, and solve the problem of object confusion in few-shot detection. The attention focus module can effectively help us learn task-related contextual fields. "base" refers to the initial pre-training performed on the base classes from the source domain, while "novel" refers to the learning of new classes during the few-shot fine-tuning procedure, where we introduced our sparse context transformer module.

3.1 Sparse Relationship Discovery

Given an image $\mathcal{I}$ fed into an SSD detector, we extract for each prior box in $\mathcal{I}$ a vector of scores $\mathbf{P}_{k,m,h,w} \in \mathbb{R}^{C_s}$; being C_s the number of (source) object categories, $k \in [\![1, K]\!]$ a spatial scale, m an aspect ratio, and (h, w) the prior box coordinates at the k-th scale. In what follows, we reshape the tensor $(\mathbf{P}_{k,m,h,w})_{k,m,h,w}$ as a matrix $\mathbf{P} \in \mathbb{R}^{D_p \times C_s}$; being D_p the total number of prior boxes in $\mathcal{I}$ across all the possible scales, aspect ratios and locations. Scores in $\mathbf{P}$ provide us with a rich semantic representation about object categories [17]; nonetheless, this representation is deprived from contextual relationships between prior boxes. Since SSD usually involves ten thousand prior boxes per image, modeling and training all the relationships between these boxes is clearly intractable, overparameterized and thereby subject to overfitting, particularly in the few-shot scenario. In order to prevent these issues, spatial pooling is first achieved so one may obtain a more compact matrix $\mathbf{Q} \in \mathbb{R}^{D_q \times C_s}$ instead of $\mathbf{P}$, being D_q ($\ll D_p$) the reduced

number of prior boxes after spatial pooling.

Considering that prior boxes capture only object parts (and not their overall extents), multi-scale feature maps extracted by the SSD encoder (denoted as $\{\mathbf{F}_k\}_k$) are also aggregated as $\mathbf{M} = \text{Concat}(\{\mathbf{F}_k\}_k) \in \mathbb{R}^{D_q \times D_f}$; being Concat the concatenation operator and D_f the resulting dimension after the application of this operator. These aggregated features $\mathbf{M}$ enable to provision with complementary visual cues at different scales, and provide us with a more comprehensive contextual information [18]. In the rest of this paper, the *pairs* of pooled prior boxes $\mathbf{Q}$ together with the underlying aggregated features $\mathbf{M}$ are referred to as *contextual fields*.

Attention Focus. In order to learn task-related context from few training data, we design an attention focus layer that enhances the representation of contextual fields and attenuates category confusion. We first define the attention weight matrix $\mathbf{A_M}$ as

$$\mathbf{A_M} = \psi_\alpha(\mathbf{M})^\top \ \psi_\beta(\mathbf{M}), \tag{1}$$

being $\psi_\alpha(.)$ (and $\psi_\beta(.)$) trained fully-connected (FC) layers that increase the expressivity of the attention matrix $\mathbf{A_M} \in \mathbb{R}^{D_f \times C_s}$, and allow obtaining an enhanced representation $\mathbf{M}^* \in \mathbb{R}^{D_q \times C_s}$ as

$$\mathbf{M}^* = \mathbf{M} \ \ \mathbf{A_M}, \tag{2}$$

which also ensures dimension consistency of the learned representation in Eq. (3). By combining the pooled prior boxes in $\mathbf{Q}$ and the underlying multi-scale feature maps $\mathbf{M}^*$, we obtain our contextual field representations that capture both intrinsic (feature) and extrinsic (object-class) information, resulting into

$$\mathbf{C} = \lambda \ \mathbf{M}^* \ + \ \mathbf{Q}, \tag{3}$$

here $\lambda \geq 0$ controls the impact of attention in $\mathbf{M}^*$. Using Eq. (3), we design our sparse attention mechanism in order to explore the affinity relationships between each prior box and contextual fields, and remove spurious ones (i.e., those farther away from the prior boxes) according to the sparse relationship. More precisely, we evaluate the relationship matrix between the contextual fields in $\mathbf{C}$ and prior boxes in $\mathbf{P}$. Here, we adopt a commonly used method for sparse representation–soft thresholding. By shrinking the relationship weights in the relationship matrix $\mathbf{A}$ that fall below a certain threshold to zero, which enables the model to focus on learning the most relevant contextual information:

$$\mathbf{R} = \mathit{SoftShrink}(\mathbf{A}, \zeta) = \text{sign}(\mathbf{A}) \max(|\mathbf{A}| - \zeta, 0) \tag{4}$$

with

$$\mathbf{A} = \text{softmax}\left(\frac{\psi_\gamma(\mathbf{P}) \ \psi_\rho(\mathbf{C})^\top}{\sqrt{C_s}}\right). \tag{5}$$

where softmax is applied row-wise, $\psi_\gamma(.)$, $\psi_\rho(.)$ are again trainable FC layers and sign($\mathbf{A}$) represents the sign of each element in the matrix $\mathbf{A}$, ζ is the threshold, empirically selected to be 0.8. Each row of $\mathbf{R} \in \mathbb{R}^{D_p \times D_q}$ measures the importance

of all contextual fields w.r.t. its underlying prior box. Hence, sparse relationship discovery allows a prior box to identify its important contextual fields and discard those that are not sufficiently important according to various aspect ratios, locations and spatial scales.

3.2 Aggregation

We consider the sparse relationship matrix $\mathbf{R}$ — between prior boxes and contextual fields — as a relational attention in order to derive the representation of each prior box. We also consider a softmax operator on each row i of $\mathbf{R}$ as a gating mechanism that measures how important is each contextual field w.r.t. the i-th prior box. By considering the cross correlations between rows of $\mathbf{R}$ and columns of $\mathbf{C}$, we derive our sparse attention-based representation of prior boxes as

$$\mathbf{W} = \text{softmax}(\mathbf{R})\ \psi_\eta(\mathbf{C}), \tag{6}$$

being $\mathbf{W} \in \mathbb{R}^{D_p \times C_s}$ and ψ_η corresponds again to trainable FC layers. Now, we combine $\mathbf{W}$ with the original matrix of prior boxes $\mathbf{P}$ in order to derive our final context-aware representation $\hat{\mathbf{P}} \in \mathbb{R}^{D_p \times C_s}$

$$\hat{\mathbf{P}} = \mathbf{P} + \psi_\xi(\mathbf{W}). \tag{7}$$

Here ψ_ξ corresponds to other (last) trainable FC layers. Since $\hat{\mathbf{P}}$ is context-aware, it enhances the discrimination power of prior boxes by attenuating confusions between object classes. By plugging the final representation $\hat{\mathbf{P}}$ into a softmax classifier, we obtain our scoring function, on the C_t target classes, as

$$\hat{\mathbf{Y}} = \text{softmax}(\hat{\mathbf{P}}\ \Theta). \tag{8}$$

Note that the representations in $\hat{\mathbf{P}}$ and the underlying parameters $\Theta \in \mathbb{R}^{C_s \times C_t}$ are shared across different aspect ratios and spatial scales, so there is no requirement to design separate classifiers at different scales. This not only reduces computational complexity but also prevents overfitting.

4 Experiments

In this section, we evaluate the proposed framework on two standard benchmarks, namely PASCAL VOC and MS COCO.

4.1 Datasets and Settings

Following [17], we describe the few-shot detection datasets for comparison against the related works. All the quantitative performances correspond to the mean Average Precision (mAP) evaluation metric.

PASCAL VOC 2007/2012 & MS COCO. The train and test sets of PASCAL VOC 2007 and 2012 are split into source (seen) and target (unseen) object

categories. Three different splits are considered for the unseen categories, namely ["bird", "bus", "cow", "motorbike", "sofa"], ["aeroplane", "bottle", "cow", "horse", "sofa"], and ["boat", "cat", "motorbike", "sheep", "sofa"]. As for MS COCO [28], it includes 80 object categories where 20 of them — that overlap with PASCAL VOC — are used as unseen categories. The process of constructing the few shot dataset, and seen/unseen categories in MS COCO is similar to PASCAL VOC.

Implementation Details. We choose a recent SSD detector [12] as a basic architecture built upon 6 heads corresponding to different spatial rescaling factors (taken in $\{1, 3, 5, 10, 19, 38\}$). The contextual fields we designed consist of two parts: in the first one, prior boxes — corresponding to multiple scales and aspect ratios — are max-pooled with different instances of kernel sizes+strides taken in $\{2, 3\}$. For the second one, contextual fields composed of multi-scale features are fused through four spatial scales. In these experiments, we set the hyperparameter λ in Eq. (3) to 0.6, and the embedding functions in the sparse context transformer correspond to the residual FC layer whose input and output have the same number of channels.

We implement our experiments using PyTorch on two Nvidia 3090 GPUs. We pretrain the SSD detectors on the source domain following exactly the original SSD settings in [29], and we fine-tune these SSDs on the target domain using stochastic gradient descent with the following settings: a batch size of 64, a momentum of 0.9, an initial learning rate equal to 4×10^{-3} (decreased by 10 after 3k and 3.5k iterations), a weight decay of 5×10^{-4}, and a total number of training iterations equal to 4k.

4.2 Results on PASCAL VOC

In this section, we show the impact of our models on PASCAL VOC. We first compare our method with the related state-of-the-art, then we carry out ablation study in order to understand the behavior of different components of our proposed sparse context transformer. We also show some qualitative results that highlight the impact of our models. In all these experiments, we take the average performance across 10 random runs.

Comparison. Table 1 shows a comparison of our method against the related state-of-the-art works which mostly report results with multiple random runs. Our proposed sparse context transformer obtains high accuracy on almost all the splits with different shots. Specifically, at extremely low-shot settings (i.e. 1-shot), our method achieves state-of-the-art performance on all the splits. These results demonstrate the ability of our proposed sparse context transformer to effectively combine contextual information of the target domain in order to overcome object class confusion in the few-shot detection scenarios. Furthermore, we also observed that the proposed method — while not totally outperforming DCNet [26], prominently on the Novel set3 — obtains very competitive performances in most shots.

Table 1: Few-Shot detection performance (mAP) on PASCAL VOC dataset. We evaluate 1, 2, 3, 5, 10 shot performance over multiple runs. red and blue indicate the best and second best results, respectively, and '-' indicates no reported results (i.e., not available).

Method/shot	Novel Set 1					Novel Set 2					Novel Set 3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
Shemelkov et al.2017[23]	23.9	-	-	38.8	-	19.2	-	-	32.5	-	21.4	-	-	31.8	-
Meta YOLO[24]	14.8	-	-	33.9	-	15.7	-	-	30.1	-	19.2	-	-	40.6	-
Meta R-CNN [25]	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	39.7	14.3	18.2	27.5	41.2	48.1
DCNet [26]	33.9	37.4	43.7	51.1	59.6	23.2	24.8	30.6	36.7	46.6	32.3	34.9	39.7	42.6	50.7
Cos R-CNN [27]	27.9	33.0	32.1	36.2	33.6	19.4	12.6	14.4	19.1	21.9	16.9	21.6	21.6	27.5	25.5
Baseline	34.2	-	-	44.2	-	26.0	-	-	36.3	-	29.3	-	-	40.8	-
SCT(ours)	37.9	40.8	41.2	45.0	46.8	32.8	33.1	33.8	36.8	37.9	33.7	33.4	35.4	40.8	44.5

Table 2: Ablation studies on the effectiveness of various components in our proposed sparse context transformer. The mAP with IoU threshold 0.5 (AP50) is reported. T stand for target domain categories.

Method	context (w/ multi-scale feature)	Sparse relations	T
Baseline	✗	✗	26.0
Ours	✓	✗	29.4
	✗	✓	30.8
	✓	✓	**32.8**

Ablation. In this ablation study, all the models are trained on the most difficult 1-shot scene set (in the novel set 2 of PASCAL VOC), and then evaluated on the PASCAL VOC 2007 test set. In these results, we again take the average performance across 10 random runs.

Impact of Context & Sparsity. Table 2 (row 2 vs 1) shows the effectiveness of our proposed fusion approach of *contextual fields* from different sources. These results show that at extremely low-shot settings, the fusion of contextual fields from different sources improves the recognition performance of new categories in the target domain. Table 2 (row 3 vs 1) also demonstrates the positive impact of *sparsity*, i.e., constructing sparse relationships between contextual fields and prior boxes. Indeed, at extremely low-shot settings, sparsity improves the accuracy of new object categories in the target domain.

Impact of Attention Focus. Table 3 shows the effectiveness of our designed attention focus layer. Similarly, the comparison results (second row vs third row) indicate that applying an attention focus layer to the contextual fields formed by multi-scale feature maps before performing the GAP layer can effectively enhance the representation of contextual fields and attenuates category confusion. From these results, we observe a clear positive impact of our proposed attention focus layer on target domain (new) classes compared to when no attention is

Table 3: The impact of our proposed attention focus layer. The experiments are conducted on the VOC 2007 test set of the PASCAL VOC dataset with novel split2 and AP50 on 1-shot task. Here T stands for target and GAP for global average pooling layer.

	GAP	Attention Focus	T
Multi-scale Feature Maps	✗	✗	30.3
	✓	✗	29.5
	✓	✓	30.1
	✗	✓	**32.8**

considered. Based on these results, we only adopted the attention focus layer in the final learned model, as shown in Fig. 1.

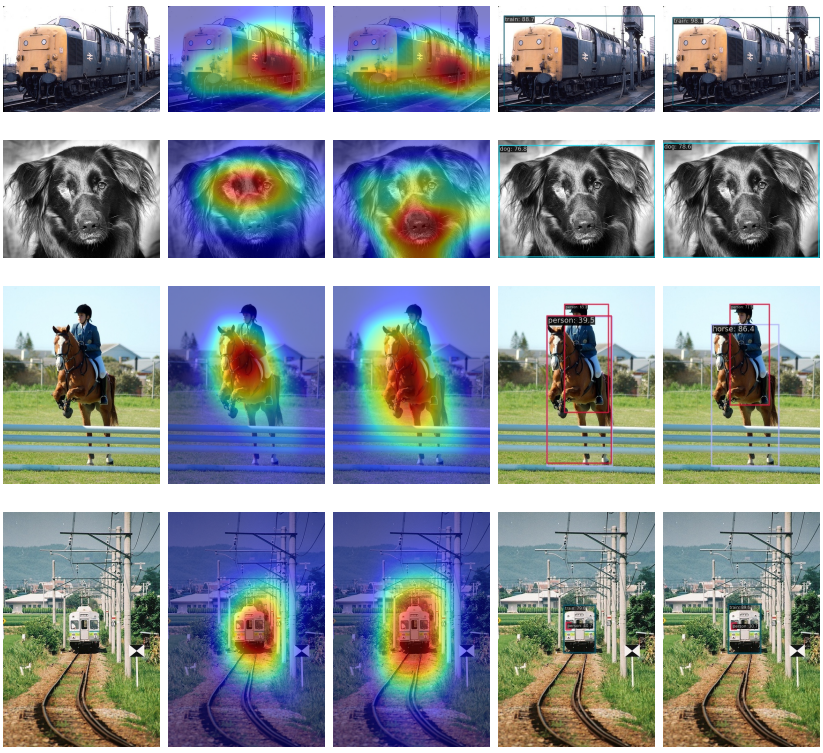
Table 4: Few-Shot detection performance on COCO minival of MS COCO dataset. We report the mean Averaged Precision on the 20 novel classes of COCO. red and blue indicate the best and second best results.

Method/shot	10-shot						30-shot					
	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
LSTD[17]	3.2	8.1	2.1	0.9	2.0	6.5	7.8	10.4	10.4	1.1	5.6	19.6
Meta YOLO [24]	5.6	12.3	4.6	0.9	3.5	10.5	10.1	14.3	14.4	1.5	8.4	28.2
Baseline[19]	7.7	13.5	7.5	2.3	5.4	12.2	11.0	19.5	10.8	2.1	8.3	18.6
SparseCT(ours)	7.9	14.3	7.7	1.9	4.5	12.4	11.2	20.2	11.3	2.2	8.5	19.3

Qualitative Performance. In order to further understand the impact of our sparse context transformer module, we visualize the results with (w/) and without (w/o) SCT. As shown in Fig. 2, the activation maps in the second and third columns show that SCT improves attention to objects in images which eventually leads to better detection performance as shown in the fourth and fifth columns. Furthermore, by comparing visualizations in the third row, we observe that SCT mitigates object confusion in few-shot detection.

4.3 Results on MS COCO

Finally, we provide extra performance on the 10/30-shot setups using the MS COCO benchmark, and we report the average performance using the standard COCO metrics over 10 runs with random shows. Performance on novel classes are shown in Table 4. Compared to the baseline, our proposed method improves performance in the 10/30 shot tasks, and is also competitive compared to the related works. Qualitative results are also shown in Fig. 3; compared to the baseline



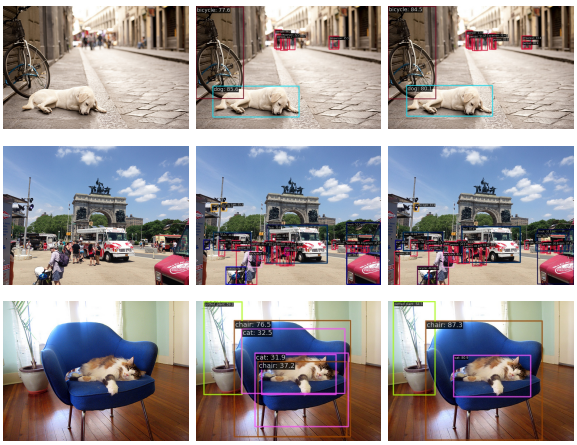
(a) Input (b) W/o SCT (c) W/ SCT (d) W/o SCT (e) W/ SCT

Fig. 2: Visualization of the results with (w/) and without (w/o) Sparse Context Transformer (SCT) on the PASCAL VOC dataset. Different colored bounding boxes represent different categories.

in the related work [19], our proposed method shows a noticeable improvement when handling confusion between detected object categories.

5 Conclusion

In this paper, we propose a novel sparse context transformer that effectively explores useful contextual information for few shot object detection. The strength of the proposed model resides in its ability to learn sparse relevant relationships while discarding irrelevant ones. Extensive experiments conducted on two challenging standard benchmarks (namely PASCAL VOC and MS COCO) show the effectiveness of each component of our sparse context transformer and its outperformance with respect to the related work. As a future work, we will investigate how to further improve the model's generalization ability in limited-sample scenarios and also investigate the extension of this model using other neural architectures, evaluation benchmarks and applications.



(a) Input (b) W/o SCT (c) W/ SCT

Fig. 3: This figure shows comparison results on the MS COCO dataset. Baseline represents the work of Yang et al [19]. Different colored bounding boxes represent different categories.

References

1. M. Jiu and H. Sahbi. Deep representation design from deep kernel networks. Pattern Recognition, 88, 447-457, 2019
2. Köhler M, Eisenbach M, Gross H M. Few-shot object detection: A comprehensive survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023.
3. Jiaxu L, Taiyue C, Xinbo G, et al. A comparative review of recent few-shot object detection algorithms[J]. arXiv preprint arXiv:2111.00201, 2021.
4. Qiao L, Zhao Y, Li Z, et al. Defrcn: Decoupled faster r-cnn for few-shot object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 8681-8690.
5. M. Jiu and H. Sahbi. Context-aware deep kernel networks for image annotation. Neurocomputing, 474, 154-167, 2022.
6. Li Y, Zhu H, Cheng Y, et al. Few-shot object detection via classification refinement and distractor retreatment[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 15395-15403
7. Sun B, Li B, Cai S, et al. Fsce: Few-shot object detection via contrastive proposal encoding[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 7352-7362.
8. Antonelli S, Avola D, Cinque L, et al. Few-shot object detection: A survey[J]. ACM Computing Surveys (CSUR), 2022, 54(11s): 1-37.
9. Tian S, Li L, Li W, et al. A survey on few-shot class-incremental learning[J]. Neural Networks, 2024, 169: 307-324.
10. M. Jiu and H. Sahbi. Nonlinear deep kernel learning for image annotation. IEEE Transactions on Image Processing, 26(4), 1820-1832, 2017.
11. Girshick R. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

12. Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.
13. Tian P, Xie S. An adversarial meta-training framework for cross-domain few-shot learning[J]. IEEE Transactions on Multimedia, 2022, 25: 6881-6891.
14. Xiao Y, Lepetit V, Marlet R. Few-shot object detection and viewpoint estimation for objects in the wild[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(3): 3090-3106.
15. M. Jiu and H. Sahbi. DHCN: Deep hierarchical context networks for image annotation. In ICASSP, 2021.
16. Han J, Ren Y, Ding J, et al. Few-shot object detection via variational feature aggregation[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(1): 755-763.
17. Chen H, Wang Y, Wang G, et al. Lstd: A low-shot transfer detector for object detection[C]. Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).
18. Khandelwal S, Goyal R, Sigal L. Unit: Unified knowledge transfer for any-shot object detection and segmentation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 5951-5961.
19. Yang Z, Wang Y, Chen X, et al. Context-transformer: Tackling object confusion for few-shot detection[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12653-12660.
20. M. Jiu and H. Sahbi. Laplacian deep kernel learning for image annotation. In ICASSP, 2016.
21. Tzeng E, Hoffman J, Darrell T, et al. Simultaneous deep transfer across domains and tasks[C]. Proceedings of the IEEE international conference on computer vision. 2015: 4068-4076.
22. Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.
23. Shmelkov K, Schmid C, Alahari K. Incremental learning of object detectors without catastrophic forgetting[C]. Proceedings of the IEEE international conference on computer vision. 2017: 3400-3409.
24. Kang B, Liu Z, Wang X, et al. Few-shot object detection via feature reweighting[C]. Proceedings of the IEEE/CVF ICCV, 2019: 8420-8429.
25. Yan X, Chen Z, Xu A, et al. Meta r-cnn: Towards general solver for instance-level low-shot learning[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 9577-9586.
26. Hu H, Bai S, Li A, et al. Dense relation distillation with context-aware aggregation for few-shot object detection[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 10185-10194.
27. Data G W P, Howard-Jenkins H, Murray D, et al. Cos r-cnn for online few-shot object detection[J]. arXiv preprint arXiv:2307.13485, 2023.
28. Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.
29. Liu S, Huang D. Receptive field block net for accurate and fast object detection[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 385-400.