Few-Shot Object Detection by Second-order Pooling

Shan Zhang[1], Dawei Luo[2], Lei Wang[3], Piotr Koniusz[4,1(✉)]

[1] Australian National University
[2] Beijing University of Posts and Telecommunications
[3] University of Wollongong
[4] Data61/CSIRO

Abstract. In this paper, we tackle a challenging problem of Few-shot Object Detection rather than recognition. We propose Power Normalizing Second-order Detector consisting of the Encoding Network (EN), the Multi-scale Feature Fusion (MFF), Second-order Pooling (SOP) with Power Normalization (PN), the Hyper Attention Region Proposal Network (HARPN) and Similarity Network (SN). EN takes support image crops and a query image per episode to produce covolutional feature maps across several layers while MFF combines them into multi-scale feature maps. SOP aggregates them per support image while PN detects the presence of visual feature instead of counting its frequency of occurrence. HARPN cross-correlates the PN pooled support features against the query feature map to match regions and produce query region proposals that are then aggregated with SOP/PN. Finally, support and query second-order descriptors are passed to SN.
Our approach performs well because: (i) HARPN leverages SOP/PN for cross-correlation of detected rather than counted support features with query features which improves region proposals, (ii) SOP/PN capture second-order statistics per region proposal and factor out spatial locations, and (iii) PN limits the complexity of the space of functions over which HARPN and SN learn. These properties lead to the state of the art on the PASCAL VOC 2007/12, MS COCO and the FSOD datasets.

1 Introduction

Over the past years, several deep learning object detectors have achieved remarkable performance [1–6]. However, these models usually rely on a large number of fully-annotated bounding boxes for training and they cannot be easily extended to unseen classes not provided during training. Thus, in a practical scenario, the fully-annotated training is insufficient for a given target detection task with novel classes, which limits the applicability of the model.

In contrast, humans learn to recognize new objects even with a little supervision which highlights the superiority of biological vision over artificial CNNs. This inspires us to develop a Few-shot Object Detection (FSOD) network which is trained on just a few of training samples. In recent years, researchers have

Figure 1: The difference between few-shot (left) image- and (right) instance-level recognition. In contrast to classification problems, query images for few-shot object detection contain multiple objects to be localized and recognized.

explored few-shot learning [7–13]. However, such off-the-shelf few-shot classifiers cannot be directly applied to the FSOD problem which requires simultaneous classification (novel classes) and localization of objects. Taking Prototypical Networks [7] and SoSN [14] as examples, it is unclear how to utilize the framework for matching and localization as irrelevant objects within the query image distract the few-shot detector. In short, few-shot classification is an image-level task where the few-shot learner relies on images of a single object to classify. In contrast, few-shot object detector represents an instance classification problem for which a query image includes multiple objects. Thus, the instance-level task needs to predict bounding boxes not just classify objects, as shown in Figure 1.

The idea of few-shot object detection has recently been explored [15–19]. However, different from these approaches, we focus on utilizing robust second-order statistics for FSOD. Furthermore, in contrast to two-stage pipeline [15, 18, 19] where proposals of various sizes produce descriptors of varying sizes, our second-order representation describes regions as SPD matrices which capture multivariate Normal distributions. Such SPD matrices have constant size independent of the spatial dimensions of feature maps. Thus, we disregard the Region of Interest (ROI) pooling which suffers from subsampling and the order of features by their spatial locations (high discriminativity but poor repeatability of exact feature combinations across locations harms similarity learning).

Second-order statistics of data features have advanced the state of art in object recognition [20–22]. Second-order Pooling (SOP) has been extended to CNNs as a trainable layer for few-shot image classification [14, 23–25]. Inspired by these models, we design a Few-shot Object Detection by leveraging SOP. As second-order statistics contain an expected value of co-occurrences of visual features, this often introduces a nuisance variability related to the frequency of certain co-occurrences that vary from object to object of the same class [24]. Thus, SOP requires Power Normalization (PN) which reduces this nuisance variability as demonstrated later in the text. We conjecture that such nuisance variability, relative to limited variations if PN is used, is approximately quadratic w.r.t. the input filter size of Hyper Attention Region Proposal Network (HARPN) as well as the shot number Z. Thus, PN is particularly well-suited for FSOD.

As different convolutional layers capture visual details at different scales of observation *e.g.*, fine-to-coarse or simplistic-to-composite, approaches [6, 26, 27]

have shown that multi-layer feature combinations yield better proposals and detection results than features of a single layer. Particularly, the combination of larger number of fine-to-coarse CNN features is more beneficial compared to using features of neighboring layers which are strongly correlated [6]. Thus, we use coarse-to-fine features to leverage strong semantics from the deep convolutional layers as well as highly localized features from early layers of the network.

Finally, we investigate how to robustly generate bounding boxes via HARPN. In particular, we propose a channel-wise cross-correlation between support features detected via SOP/PN from the support representation (spatial locations are factored out) to highlight matching features across all spatial locations of the query feature map which is fed to RPN for generation of region proposals.

In this paper, we address the challenge of few-shot object detection. At the test time, through a few of support images of novel target object, FSOD strives to detect all objects in the query set that belong to the target object category. To devise the proposed framework termed Power Normalizing Second-order Detector (PNSD), we make four contributions:

i. We make the first attempt to embed SOP and PN into the FSOD.
ii. In our framework, the HARPN remodels the query feature maps via channel-wise cross-correlation to obtain robust ROI proposals whose statistical content is captured by PN-normalized SOP descriptors rather than ROI pooling.
iii. Our SOP descriptors are obtained from coarse-to-fine feature maps which improves the proposal generation and detection.
iv. In our Supplementary Material, we demonstrate statistically the benefit of Power Normalization for HARPN and similarity learning with SN. We show that PN limits the space of functions involved in learning by HARPN and SN which prevents overfitting and benefits FSOD more than classification.

Extensive experimental evaluations in the few-shot setting on three widely used object detection benchmarks, that is PASCAL VOC 2007/12, MS COCO and FSOD, show the effectiveness of our PNSD. The rest of the paper is organized as follows. In Section 2, we summarize the work most related to this paper. Our PNSD approach is described in detail in Section 4 before Preliminaries in Section 3. Experimental results are reported and analyzed in Section 5. Finally, we conclude the paper in Section 6.

2 Related work

In what follows, we describe popular general object detection and few-shot learning algorithms followed by a short discussion on Second-order Pooling.

Object Detection is a classical problem in computer vision. In early years, object detection was usually formulated as a sliding window classification problem using handcrafted features [28–30]. With the rise of deep learning [31], CNN-based detectors have become dominant approaches which can be further divided into two categories: proposal-free detectors and proposal-based detectors. The first line of work follows a one-stage training strategy and does not explicitly

generate proposal boxes [1, 2, 32, 33]. The second line, pioneered by R-CNN, first extracts class-agnostic region proposals of the potential objects from a given image. These boxes are then further rened and classified into different categories by a dedicated module [3–6]. An advantage of this strategy is that it can filter out many negative locations by the Region Proposal Network (RPN) module which facilitates recognition. RPN-based methods usually perform better than proposal-free methods with state-of-the-art results [34] for the detection task. The methods mentioned above, however, work in an intensive supervision manner and are hard to extend to novel classes with only few examples.

Few-shot learning is mainly inspired by the human ability to learn new concepts from a limited number of samples. Recently, many few-shot classification approaches have been developed with the aim to classify images of novel classes given very few labeled examples. These approaches can be divided into metric learning, meta-learning and 'optimization for fast adaptation' approaches. The aim of metric-learning [35, 36, 13] based few-shot classification is to derive a similarity metric that can be directly applied to the inference of unseen classes supported by a set of labeled examples (*ie.*, support set). Koch [35] presents the first principled approach that employs Siamese networks for one-shot image classification. Prototypical Networks [7] learns a model that computes distances between a datapoint and prototype representations of each class. The approach of [7, 37, 38] parametrizes the optimization algorithm to predict the parameters of a few-shot detector via a meta-learner. Ravi and Larochelle [8] propose an LSTM meta-learner that is trained to attain a quick convergence of few-shot learner. Recent methods address subspace-based learning [39], gradient modulation [40] and few-shot action recognition [41]. However, most existing methods focus on image classification but rarely on more practical tasks such as semantic segmentation [42–44], human motion prediction [45] or object detection [15, 19].

Due to numerous bounding-boxes, object detection is more time-consuming than image-level classification. Thus, such work would be practically impactful if the novel classes and object bounding boxes could be predicted by a few-shot learner. Approach [15] transfers knowledge from a larger to a smaller dataset by minimizing the gap between source and target domains but it requires to be fine-tuned for novel categories. To solve this issue, approach [19] proposed a general few-shot object detection network that learns the matching metric between image pairs based on the Faster R-CNN framework, termed Few-shot Object Detection (FSOD). Different from [19] leveraging the first-order representation, we focus on second-order representations to capture co-occurrences of features and we investigate Power Normalization whose goal is to reduce the harmful variability of features. In place of the ROI pooing, SOP/PN are used.

Second-order statistics have been studied in the context of texture recognition [46, 47] through so-called Region Covariance Descriptors (RCD), and further applied to object category recognition [20]. Co-occurrence patterns can also be used in the CNN setting *ie.*, approach [48] extracts feature vectors at two separate locations in feature maps followed by an outer product to form a CNN co-occurrence layer. Approach [24, 22] performs spectral second-order pooling

for fine-grained image classification. SoSN [14, 22] leverages second-order pooling and Power Normalization for end-to-end training of one- and few-shot image classification pipeline (single object per image). In contrast, we develop a few-shot detector that tackle multi-object localisation and classification.

Power Normalization deals with the so-called burstiness of first- and second-order statistics which is 'the property that a given visual element appears more times in an image than a statistically independent model would predict'. Power Normalization [49] suppresses this burstiness by performing likelihood-inspired feature detection rather than feature counting [49, 20, 24]. The specific variant of PN, namely MaxExp feature pooling [49, 22] can be interpreted as a detector of 'at least one particular visual word being present in an image' which is further extended to so-called SigmE pooling [24, 22] for auto-correlation matrices that contain both positive and negative values. Furthermore, Spectral MaxExp pooling [24, 22] performs decorrelation of features which boosts discriminativness. Papers [50, 24] point that many PN functions are closely related, however, they do not analyze why PN is well-suited to few-shot learning. In our Supplementary Material, we analyze PN in the context of few-shot object detection.

3 Preliminaries

Below we review our notations and demonstrate how to calculate second-order statistics and Power Normalization.

Notations. Let $x \in \mathbb{R}^d$ be a d-dimensional feature vector. I_N stands for the index set $\{1, 2, \cdots, N\}$. Moreover, for a matrix $\mathbf{X}$, we denote $\mathbf{X}\mathbf{X}^T = \uparrow \otimes_2 \mathbf{X}$. We also define $\mathbf{1} = [1, ..., 1]^T$ ('all ones' vector). Typically, capitalised boldface symbols such as $\boldsymbol{\Phi}$ denote matrices, lowercase boldface symbols such as $\boldsymbol{\phi}$ denote vectors and regular case such as $\Phi_{i,j}$, ϕ_i, n or Z denote scalars *e.g.*, $\Phi_{i,j}$ is the (i, j)-th coefficient of $\boldsymbol{\phi}$.

Second-order Pooling and Power Normalization.

Let a set $\mathcal{N}$ point to indices of feature vectors stacked as column vectors so that $\boldsymbol{\Phi} = [\boldsymbol{\phi}_n]_{n \in \mathcal{N}}$. To perform SOP, one can simply form an auto-correlation matrix $\mathbf{M} = \frac{1}{N}\boldsymbol{\Phi}\boldsymbol{\Phi}^T$ where $N = |\mathcal{N}|$. As alluded to in the introduction, we desire to perform detection of feature co-occurrences rather than counting. Additionally, we have to deal with the evidence of correlation and anti-correlation in the auto-correlation matrix (positive and negative coefficients). Thus, we employ a so-called SigmE PN function [24] which is designed just for this purpose, and it is defined as:

$$\mathcal{G}_{\text{SigmE}}(\mathbf{M}; \eta) = \frac{2}{1 + e^{-\eta \mathbf{M}/(\text{Tr}(\mathbf{M}) + \lambda)}} - 1, \tag{1}$$

where $1 \leq \eta \approx N$ interpolates between counting and detection, $\lambda \approx 1e^{-6}$ is a regularization constant and the trace $\text{Tr}(\cdot)$ stops diagonal from exceeding one.

To decorrelate features from a support image and match them against query in Hyper Attention RPN introduced below, among other variants, we use a

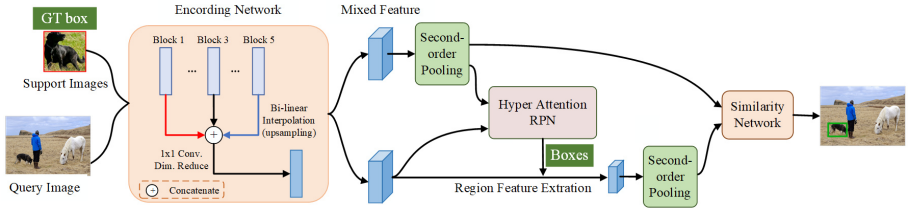

Figure 2: Our PNSD. The query image and support crops are processed by the Encoding Network to form coarse-to-fine features via the Multi-scale Feature Fusion. The Hyper Attention RPN module (Figure 3) detects support features via SOP/PN and cross-correlates them against query regions. Support features per crop and query features per region are passed via SOP/PN to form query-support descriptors passed to the Similarity Network for localization and classification.

variant of Spectral Power Normalization, known as Spectral MaxExp [24, 22]:

$$\widehat{\mathcal{G}}_{\mathrm{MaxExp}}(\mathbf{M};\eta) = \mathbb{I} - (\mathbb{I} - \mathbf{M}/(\mathrm{Tr}(\mathbf{M}) + \lambda))^{\eta}\,. \tag{2}$$

4 Proposed approach

Below we present our Power Normalizing Second-order Detector network followed by a description of its individual components. **Overview.** The algorithm operates on so-called L-way Z-shot episodes which are formed by sampling a query image containing multiple objects, and Z support crops per each of L sampled classes. The training protocol ensures that query classes corresponding to objects in the query image have some matches in the support set. At the test time, given annotated support crops of novel classes, one can localize and classify objects in the query image.

Inspired by a recent FSOD architecture [19], our Power Normalizing Second-order Detector for FSOD, denoted PNSD for short, consists of (i) the Encoding Network (EN), (ii) the Multi-scale Feature Fusion (MFF), (iii) Second-order Pooling (SOP) with Power Normalization (PN), (iv) the Hyper Attention Region Proposal Network (HARPN) and (v) the Similarity Network (SN). Figure 2 shows our architecture for one support image as an example.

The role of EN is to generate image-level convolutional feature vectors (descriptors) whose fine-to-coarse nature is represented by MFF. The task of HARPN is to generate region proposals on the query image. SOP and PN are applied in two manners: (i) as a module of HARPN to improve the region proposals and (ii) as descriptors of the support crops and descriptors of region proposals which results in constant size representations independent of sizes of crops and region proposals. Finally, SN takes such formed query-support pairs and learns localization and similarity with a combination of two objectives.

Encoding Network. We use ResNet-50 with $f : (\mathbb{R}^{W\times H}; \mathbb{R}^{|\mathcal{F}|}) \rightarrow \mathbb{R}^{K\times N}$ realizing EN and MFF, where W and H are the width and height of an input,

$N = N_W \cdot N_H$ and K are the total number of spatial locations and numbers of features (channel-wise) in the feature map after concatenation of outputs of *Block1*, *Block3* and *Block5* (up-sampled to match the large spatial size) of ResNet-50 and reducing their dimensionality channel-wise by 1×1 conv. implemented in MFF. Furthermore, we denote the final support and query maps by $\boldsymbol{\Phi} \in \mathbb{R}^{K \times N}$ and $\boldsymbol{\Phi}^* \in \mathbb{R}^{K \times N^*}$, where $\boldsymbol{\Phi} = f(\boldsymbol{X}; \mathcal{F})$ and $\boldsymbol{\Phi}^* = f(\boldsymbol{X}^*; \mathcal{F})$, the support crop $\boldsymbol{X} \in \mathbb{R}^{W \times H}$ and query image $\boldsymbol{X} \in \mathbb{R}^{W^* \times H^*}$. The parameters $\mathcal{F}$ of EN are shared between support and query passes. EN is shown in Figure 2.

Hyper Attention RPN. The role of the Region Proposal Network (RPN) is to produce candidate regions from the query feature map containing objects. However, traditional RPN generates many candidate regions which become a burden for the detector. In contrast, few-shot detection requires generation of object candidates from the query image that match support regions from a given episode to reduce the number of proposals and improve the quality of recognition. To this end, we introduce the Hyper attention RPN module (Figure 3) which modulates the query feature map to produce proposals relevant to support crops. In contrast to approach [19] which applies average pooling to support features to cross-correlate the feature expectation against the query feature maps, we conjecture that using Power Normalization *e.g.*, SigmE [24] or Spectral MaxExp [22] function, is required to obtain a good matching between support-query pairs. Specifically, the query feature map contains multiple regions which may match objects from the support set. In contrast, support maps each describe a single object. Thus, matching the expectation of features of support set (average pooling) against spatial locations in the query makes such an attention modulator heavily variant w.r.t. the number of activations of a given support feature. These activations depend on the size of support object, pose, repeatable visual stimuli, *etc*. To filter this variability, we propose to use SigmE function which acts as a detector of features rather than counter (average pooling) [24] or Spectral MaxExp which partially decorrelates features (stat. independence) [22]. We propose three variants generating the attention modulator $\boldsymbol{a} \in \mathbb{R}^K$:

$$\text{First-order (FO)+PN:} \quad \boldsymbol{a}_{\text{FO+PN}} = \mathcal{G}_{\text{SigmE}}(\boldsymbol{\Phi} \cdot \mathbf{1}/N; \eta) \tag{3}$$

$$\text{Sec.-order Spec. Diag. Corr. (SOSD)+PN:} \quad \boldsymbol{a}_{\text{SOSD+PN}} = \text{Diag}\left(\widehat{\mathcal{G}}_{\text{MaxExp}}(\mathbf{M}; \eta)\right) \tag{4}$$

$$\text{Second-order Self Corr. (SOSC)+PN:} \quad \boldsymbol{a}_{\text{SOSC+PN}} = \mathcal{G}_{\text{SigmE}}(\mathbf{M} \cdot \mathbf{1}/K; \eta) \tag{5}$$

To summarize, the above operators fulfil the following roles:

i. First-order+PN (FO+PN) detects if on average the majority of features per channel in the support feature matrix $\boldsymbol{\Phi}$ are positive or negative.
ii. Second-order Spectral Diagonal Correlation+PN (SOSD+PN) captures if there is at least one feature detected per channel across support features given the auto-correlation support matrix $\mathbf{M}$ is partially decorrelated.
iii. Second-order Self Correlation+PN (SOSC+PN) detects an evidence of at least one feature per channel across support features and takes into account feature spread to other channels, which is related to compact pooling [51].

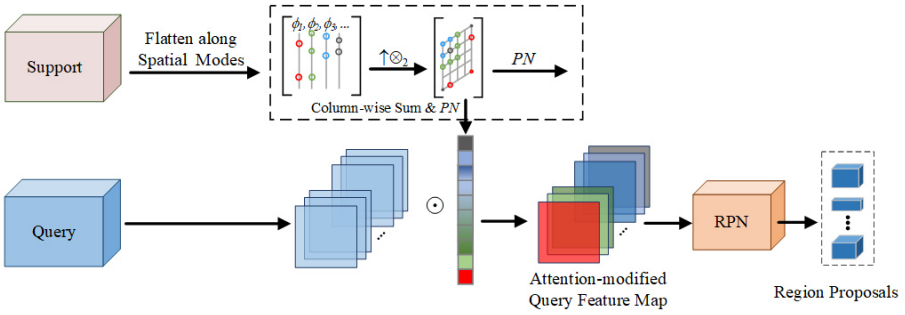

Figure 3: Hyper Attention RPN. The support features are passed by SOP to produce a $K \times K$ matrix to form attention modulator $\boldsymbol{a} \in \mathbb{R}^K$ according to Section 4. Then the attention-modulated query feature map is obtained and fed to RPN for generation of proposals. Fig. 4 is a close-up on the center of HARPN.

In what follows, we evaluate the above three attention modulators by performing cross-correlation with the query feature map by applying $\boldsymbol{\Phi}' = \boldsymbol{a} \odot \boldsymbol{\Phi}^*$, where $\odot$ is the channel-wise multiplication of a chosen attention modulator $\boldsymbol{a}$ with $\boldsymbol{\Phi}^*$ across all spatial locations and $\boldsymbol{\Phi}'$ is the attention-modulated query feature map. Finally, network $h(\boldsymbol{\Phi}'; \mathcal{H})$ produces proposals and $\mathcal{H}$ are the parameters-to-learn of the Hyper Attention RPN. Following approach [3], this module is trained jointly with EN via loss L_{rpn}.

Similarity Network. The similarity network, denoted by $s : (\mathbb{R}^{K\times K}; \mathbb{R}^{|S|}) \rightarrow \mathbb{R}^5$ (sim. score, x_t, y_t, x_b, y_b), is tasked with learning to distinguish similar/dissimilar support-query pairs represented by support-query PN-normalized SOP matrices. Typically, we denote $s(\mathbf{M}, \mathbf{M}^*; \mathcal{S})$, where $\mathbf{M}, \mathbf{M}^* \in \mathbb{R}^{K\times K}$ are support and query matrices, and $\mathcal{S}$ are the parameters-to-learn of the similarity network.

For the L-way Z-shot problem with D proposals from HARPN, we have $L \times Z \times D$ support image regions $\{\mathbf{X}_n\}_{n\in\mathcal{U}}$ from set $\mathcal{U}$ and their corresponding descriptors $\{\boldsymbol{\Phi}_n\}_{n\in\mathcal{U}}$ obtained from EN. We average $\boldsymbol{\Phi}$ of the support crops sampled for a given proposal and belonging to the same category, and we compute $L \times D$ PN-normalized SOP matrices $\mathbf{M}$ each representing one of L classes. We assume one query image $\mathbf{X}^*$ with its query feature maps $\{\boldsymbol{\Phi}^*\}_{n\in\mathcal{W}}$ from set $\mathcal{W}$ of D candidate regions from HARPN. The matrices $\mathbf{M}$ and $\mathbf{M}^*$ belong to one of L classes in the subset $C^\ddagger \equiv \{c_1, ..., c_L\} \subset \mathcal{I}_C \equiv \mathcal{C}$. The L-way Z-shot learning step is dened as alternating between learning (i) feature maps via EN and proposals via HARPN and learning (ii) feature maps via EN, bounding box regression (query) and the query-support similarity by minimizing

$$\sum_{d\in\mathcal{I}_D} \sum_{l\in\mathcal{I}_L} L_{sim}\left(s(\mathbf{M}_{dl}, \mathbf{M}^*_d, \mathcal{S})\right) + L_{box}\left(s(\mathbf{M}_{dl}, \mathbf{M}^*_d, \mathcal{S})\right) \tag{6}$$

with respect to $\mathcal{F}$ and $\mathcal{S}$ parameters of EN and SN. During training, we use the multi-task loss $L = L_{rpn} + L_{sim} + L_{box}$ but we alternate between miniminzing

w.r.t. (i) $\mathcal{F}$ and $\mathcal{H}$ and (ii) $\mathcal{F}$ and $\mathcal{S}$. The loss L_{box} for the bounding-box regression is defined as in [3] and the similarity loss is the binary cross-entropy.

SN Architecture is shown in Fig. 5. Firstly, support/query matrices (64×64×1) are passed to a small CNN to produce $6\times6\times C$ feature maps passed to so-called relation heads. Scores from all heads are averaged for final matching scores. *Global-Relation Head* concatenates support/query maps into a $6\times6\times2C$ map and pools it into a $1\times1\times2C$ map passed to 2 FC layers (+ReLU). *Local-Relation Head* applies $1\times1\times C$ conv. to maps ($6\times6\times2C$). We slide support features as a filter against query features to get local similarity passed to FC. *Patch-Relation Head* uses conv. layers (+ReLU) and pooling (zero padding) on $6\times6\times2C$ map. Two FCs generate matching scores and bounding box predictions, respectively.

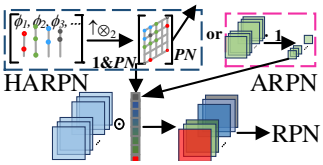

Figure 4: Architecture of HARPN.

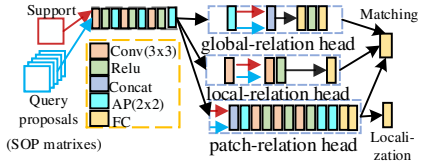

Figure 5: Architecture of SN.

5 Experimental Result

Below we verify the effectiveness of the proposed PNSD approach by comparing it with the state of the art on challenging benchmarks such as PASCAL VOC 2007/12 [52], MS COCO [53] and FSOD [19]. We also perform ablation studies of each component of PNSD on PASCAL VOC 2007 and KITTI [54].

5.1 Datasets

For the PASCAL VOC 2007/12 dataset, we adopt the 15/5 base/novel category split settings as in [16]. As recommended, we use the training and validation sets from PASCAL VOC 2007 and 2012 as training data and the testing set from PASCAL VOC 2007 for evaluation. For the MS COCO dataset, we follow the work of [18] to adopt the 20 categories that overlap with PASCAL VOC as the novel categories for evaluation, and we use the remaining 60 categories of MS COCO as the training categories. For the FSOD dataset, we split its 1000 categories into 800/200 for training and testing, respectively.

5.2 Implementation details

Our model is based on the ResNet-50 pre-trained on ImageNet [55] and MS COCO [53]. We fine-tune the network with a learning rate of 0.002 for the first 56000 iterations and 0.0002 for another 4000 iterations. Unless stated otherwise, all training and testing images are resized such that the shorter side has 600 pixels and the longer side is capped at 1000 pixels to fit into the GPU memory. Each support image in the few-shot object detection is cropped based on ground-truth

Table 1: Comparison with SOTA on the PASCAL VOC 2007 testing set (5-shot protocol). The mAP metric is used.

Method	Novel					
	bird	bus	cow	mbike	sofa	mean
FRCN[3]	31.3	36.9	**54.1**	26.5	36.2	36.9
LSTD[15]	22.8	52.5	31.3	45.6	40.3	38.5
FR[16]	30.0	**62.7**	43.2	**60.6**	39.6	47.2
Meta[18]	52.5	55.9	52.7	54.6	41.6	51.5
PNSD	**53.6**	60.2	51.3	55.6	**42.5**	**52.6**

Method	Base															
	aero	bike	boat	bottle	car	cat	chair	table	dog	horse	person	plant	sheep	train	tv	mean
FRCN[3]	68.4	**75.2**	59.2	**54.8**	74.1	80.8	42.8	56.0	68.9	77/8	75.5	34.7	66.1	71.2	66.2	64.8
LSTD[15]	70.9	71.3	59.8	41.1	**77.1**	81.9	45.1	**67.2**	78.0	78.9	70.7	**41.6**	63.8	**79.7**	66.8	66.3
FR[16]	65.3	73.5	54.7	39.5	75.7	81.1	35.3	62.5	72.8	78.8	68.6	41.5	59.2	76.2	69.2	63.6
Meta[18]	68.1	73.9	59.8	54.2	80.1	82.9	48.8	62.8	80.1	81.4	77.2	37.2	65.7	75.8	70.6	67.9
PNSD	**69.3**	74.8	**61.5**	53.4	80.2	**82.3**	**49.6**	61.8	**80.8**	**82.6**	**77.9**	35.8	**68.6**	78.2	**70.9**	**68.5**

box, bi-linearly interpolated and padded to a square region of 320 × 320 pixels. In the following experiments, we report the commonly used metrics for FSOD such as mAP, AP, AP_{50} and AP_{75}. All codes are implemented in PyTorch.

5.3 Comparison with the state-of-the-art

PASCAL VOC 2007/12. The proposed PNSD is compared with Feature Reweighting (FR) [16], LSTD [15] and FRCN [3] in Table 1. FSOD [19] does not provide results on PASCAL VOC and its authors indicated the code cannot be easily adapted to it. According to Table 1, we achieve the overall best performance for both novel and base classes (5-shot setting). A significant performance gain without any fine-tuning (approximately 5.4% and 2.2% on mAP) is attained against the second best method retrained on novel categories. Table 1 also reveals the generalization fragility of FRCN [3] under the few-shot setting: without adequate training images, it detects poorly objects from novel classes. In contrast, our PNSD the Hyper Attention PRN (HARPN) and the SOP/PN pooling demonstrates superior performance on novel-class object detection.

MS COCO. Table 2 compares our PNSD with FR [16], Meta R-CNN [18] and FSOD [19] on MS COCO minival set for 20 novel categories for a 10-shot detection protocol. As shown, PNSD works best among all the methods in comparison. It outperforms FSOD (the second best) by 4.2%, 1.3% and 1.9% on AP, AP_{50} and AP_{75}, respectively. Although the gains are lesser than those obtained on PASCAL VOC, they are consistent and also significant considering that MS COCO is more challenging in terms of complexity and the dataset size.

FSOD. Below we use the FSOD testing set with 200 novel categories on the 5-shot detection protocol. We compare our PNSD with FSOD [19], LSTD [15] and LSTD (FRCN [3]) where we re-implement BD and TK (modules of LSTD)

based on Faster-RCNN for fair comparison[5]. Table 3 shows that our PNSD produces the highest AP_{50} and AP_{75} values (*ie.*, 29.8% and 22.6%) among the four methods. Note that LSTD has to transfer knowledge from the source to target domain by retraining on novel categories while our PNSD and [19] are directly applied to detect novel categories.

Table 2: Comparison with SOTA on the MS COCO minival set.

Shot	Method	AP	AP_{50}	AP_{75}
10	LSTD[15]	3.2	8.1	2.1
	FR[16]	5.6	12.3	4.6
	Meta[18]	8.7	19.18	6.6
	FSOD [19]	11.1	20.4	10.6
	PNSD	**15.3**	**21.7**	**12.5**

Table 3: Comparison with SOTA on the FSOD testing set.

Shot	Method	AP_{50}	AP_{75}
5	LSTD (FRCN)[15]	23.0	12.9
	LSTD[15]	24.2	13.5
	FSOD [19]	27.5	19.4
	PNSD	**29.8**	**22.6**

5.4 Ablation study

Below we analyze the effectiveness of each component of the proposed PNSD approach. To this end, we compare PNSD with four variations proposed by us in Table 4. We report the mAP with a threshold of 0.5 for evaluation in the real-world application scenarios *e.g.*, urban scene datasets for driving applications (KITTI), and we use the PASCAL VOC setup. The following ablation studies are based on 5/10-shot object detection setting.

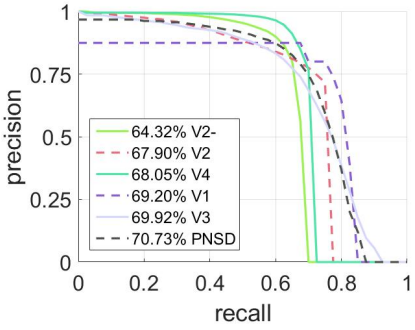

Figure 6: mAP for *car* on four variants of PNSD (KITTI, 10-shot).

Variant	Operation	Validation
V1	HARPN+SOP+PN	MFF
V2	MFF+(FO+PN)+SOP+PN	HARPN
V2-	MFF+RPN+SOP+PN	HARPN
V3	MMF+HARPN+PN	SOP
V4	MFF+HARPN+SOP	PN

Table 4: Four variants of PNSD.

Multi-scale Feature Fusion. To show the advantage of using fine-to-coarse feature representations, we modify variant V1 from Table 4 to use only features from *Block5*. Figure 6 indicates the superiority of MFF with a performance gain of 1.53% (PNSD's 70.73% *vs*. V1's 69.20%) on KITTI dataset. Also, with the PASCAL VOC setup, Table 7 shows that MFF boosts PNSD performance both in novel classes (52.6% *vs*. 52.3% in 5-shot; 61.9% *vs*. 60.1% in 10-shot) and in base classes (68.5% *vs*. 66.8% in 5-shot; 70.3% *vs*. 68.9% in 10-shot).

[5] FSOD is a new dataset released in 2020. Its format differs from datasets such as VOC and COCO. Only few approaches (listed in Table 3) experimented on FSOD.

Hyper Attention RPN. We also analyze the improvement brought by the HARPN which considers an improved feature detection strategy on support representation. As shown in Figure 6, our PNSD outperforms the variant V2 from Table 4 (70.73% *vs*. 67.90%) which leverages the regular attention RPN [19]. For the PASCAL VOC setup, Table 7 shows that our PNSD wins on novel classes (52.6% *vs*. 43.1% in 5-shot; 61.9% *vs*. 55.6% in 10-shot) and base classes (68.5% *vs*. 57.5% in 5-shot; 70.3% *vs*. 66.4% in 10-shot) due to the HAPRN.

Second-order Pooling, Power Normalization and HARPN. Firstly, we note that HARPN has better recall than the regular RPN [19] (96.81% *vs*. 93.57%), as shown in Table 5. Figure 7 evaluates HARPN variants from Section 4 and it shows that SOSD+PN and SOSC+PN outperform standard first-order ARPN [19] by nearly 20% mAP. First-order pooling over support feature maps paired with SigmE function (FO+PN) also shows a large gain. Second-order Self Correlation with SigmE (SOSC+PN) brings visible gain however Second-order Spectral Diagonal Correlation with MaxExp (SOSD+PN) performs the best. We conjecture that SOSC and SOSD take into account the channel- and spectrum-wise correlations between features of the support representation which improves their attentive expressiveness. Finally, PN benefits the attention modulator by detecting features and discarding their counts (nuisance variability). We discuss this theoretical standpoint in the Supplementary Material where we show that using PN reduces the family of functions during learning which reduces the learning complexity.

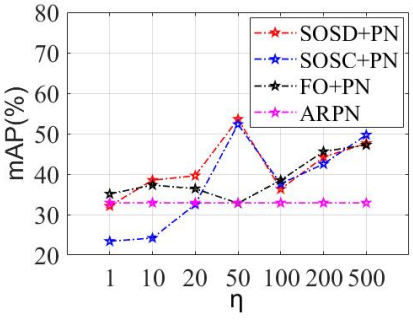

Figure 7: mAP as a function of η for four variants of HARPN (VOC2007 dataset, novel classes, 5-shot).

Method	Proposal Recall
PNSD_{Block5}	88.04
$\text{PNSD}_{Block1,3,5}$	90.43
$\text{PNSD}_{Block1,3,5}$+ARPN	93.57
$\text{PNSD}_{Block1,3,5}$+HARPN	**96.81**

Table 5: Proposal Recall under different variants of PSND. The region proposal number is 100 for evaluation (IoU = 0.5).

Second-order Pooling Regions. As features are pooled over arbitrarily sized ROIs to attain a fixed size representation, typical ROI pooling is highly discriminative but also non-repetitive as features correspond to spatial locations. Thus, we take variant V3 in Table 4 which uses the traditional ROI pooling instead of our SOP, that is the features are bi-linearly interpolated along the spatial modes of feature maps to 7×7 size. Figure 6 shows that PNSD (with SOP/PN) performs better than V3 (70.73% *vs*. 69.92%) with PN. For the PASCAL VOC setup, Table 7 confirms the superiority of our PNSD (with SOP/PN) over ROI pooling with PN in V3. PSND achieves 1.9%/1.6% improvement (52.6% *vs*. 50.7% in

5-shot; 61.9% *vs*. 60.3% in 10-shot) on novel classes. SOP is a good choice for describing ROIs as it captures second-order statistical moments while factoring out spatial modes. In our Supplementary Material we show how to efficiently compute SOP on a large number of region proposals with Integral Histograms.

Power Normalization in SOP descriptors. As indicated previously, PN discards the nuisance variability related to the frequency of certain visual features whose quantity is affected by the scale, pose, and texture areas of objects *etc*. Thus, we compare PNSD with its variant V4 from Table 4 whose SOP descriptors do not use PN. Figure6 shows that our SOP descriptors with PN yield a notable performance gain of 2.68% over V4 on KITTI dataset (*ie*., 70.73% *vs*. 68.05%). For the PASCAL VOC, Table 7 shows that our PNSD significantly outperforms the variant V4 on both base classes (68.5% *vs*. 65.3% in 5-shot ; 70.3% *vs*. 68.7% in 10-shot) and novel classes (52.6% *vs*. 47.5% in 5-shot; 61.9% *vs*. 59.4% in 10-shot). In our Supplementary Material we show that PN on SOP reduces the family of functions during relational learning (SN takes two SOP descriptors to compare them) which reduces the learning complexity.

Aggregating Z-support descriptors. Operator ($\otimes$) first averages over Z-support conv. feat. maps per class followed by the outer-product on the mean support (spatial modes are factored here). Operator ($\otimes$+L) performs the outer-product on Z-support feature vectors per class separately prior to the computation of average. Table 6 shows that operator ($\otimes$+L) outperforms operator ($\otimes$).

Table 6: Ablation studies of SOP descriptors (mAP on VOC2007, testing on novel and base classes, 5-shot).

Method	Novel	Base
SOSD+PN($\otimes$)	52.1	68.0
SOSD+PN($\otimes$+L)	**52.6**	**68.5**
SOSC+PN($\otimes$+L)	51.9	67.8
SOSC+PN($\otimes$)	52.4	68.1

Table 7: Ablation studies on four variants (mAP on VOC2007 testing set for novel classes and base classes).

Shot	Method	Novel	Base
5	V1	52.3	66.8
	V2	43.1	57.5
	V2-	41.9	53.8
	V3	50.7	66.3
	V4	47.5	65.3
	PNSD	**52.6**	**68.5**
10	V1	60.1	68.9
	V2	55.6	66.4
	V2-	54.7	64.3
	V3	60.3	69.0
	V4	59.4	68.7
	PNSD	**61.9**	**70.3**

5.5 Visualization of Detection Results.

To better illustrate the proposed PNSD framework, the detection results of several variants are visualized in Figure 8 to show how they improve upon FSOD. The four columns illustrate PNSD-V1, PNSD-V2, PNSD-V4 and PNSD, respectively. The ground-truth label of object is provided in red, with the green boxes delineating the detected objects.

Comparing the results of PNSD-V1 and PNSD, we can see a performance improvement thanks to its informative features. The differences between PNSD-V2

Figure 8: The visualization of novel-class objects detected by different PNSD variants: PNSD-V1, PNSD-V2, PNSD-V4, and PNSD (from left to right). Some bounding boxes in the first column are missed. The second column shows duplicates and inaccurate localization. The class label is wrongly recognized in the third column. Ground-truth annotated labels are marked with red colour and detection results with green rectangles.

and PNSD demonstrate that HARPN helps RPN reject irrelevant proposals so as to improve detections. In this case, our PNSD detector can potentially avoid duplication. Finally, comparing the results of PNSD-V4 and PNSD, one can clearly see the significant help of Power Normalization because which discards nuisance variability in the same class. SOP with the SigmE function (which performs Power Normalization) benefits the subsequent similarity learning by SN. Finally, it is worth noting that our PNSD framework consistently outperforms better than all four variants constructed for the purpose of ablation studies.

6 Conclusions

In this paper, we propose a Power Normalizing Second-order Detector for Few-Shot Object Detection (PNSD) to address few-shot object detection. Our model extends the R-CNN family through embedding second-order statistics and Power Normalization into HARPN and ROI descriptors. HARPN improve the proposal quality while MFF provide fine-to-coarse features. In order to demonstrate the effectiveness of PNSD, we have conducted extensive quantitative and qualitative experiments on several datasets. Our Supplementary Material demonstrates theoretically why Power Normalization is so valuable for HARPN and similarity learning with SN.

References

1. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society (2017) 6517–6525
2. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. CoRR **abs/1804.02767** (2018)
3. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., eds.: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. (2015) 91–99
4. Girshick, R.B.: Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015, IEEE Computer Society (2015) 1440–1448
5. Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society (2017) 936–944
6. Kong, T., Yao, A., Chen, Y., Sun, F.: Hypernet: Towards accurate region proposal generation and joint object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society (2016) 845–853
7. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R., eds.: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA. (2017) 4077–4087
8. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net (2017)
9. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T.P.: Meta-learning with memory-augmented neural networks. In Balcan, M., Weinberger, K.Q., eds.: Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016. Volume 48 of JMLR Workshop and Conference Proceedings., JMLR.org (2016) 1842–1850
10. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R., eds.: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. (2016) 3630–3638
11. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In Precup, D., Teh, Y.W., eds.: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Volume 70 of Proceedings of Machine Learning Research., PMLR (2017) 1126–1135
12. Cai, Q., Pan, Y., Yao, T., Yan, C., Mei, T.: Memory matching networks for one-shot image recognition. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, IEEE Computer Society (2018) 4080–4088

13. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H.S., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, IEEE Computer Society (2018) 1199–1208
14. Zhang, H., Koniusz, P.: Power normalizing second-order similarity network for few-shot learning. In: IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019, IEEE (2019) 1185–1193
15. Chen, H., Wang, Y., Wang, G., Qiao, Y.: LSTD: A low-shot transfer detector for object detection. In McIlraith, S.A., Weinberger, K.Q., eds.: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press (2018) 2836–2843
16. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE (2019) 8419–8428
17. Karlinsky, L., Shtok, J., Harary, S., Schwartz, E., Aides, A., Feris, R.S., Giryes, R., Bronstein, A.M.: Repmet: Representative-based metric learning for classification and few-shot object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE (2019) 5197–5206
18. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta R-CNN: towards general solver for instance-level low-shot learning. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE (2019) 9576–9585
19. Fan, Q., Zhuo, W., Tai, Y.: Few-shot object detection with attention-rpn and multi-relation detector. CoRR **abs/1908.01998** (2019)
20. Koniusz, P., Yan, F., Gosselin, P., Mikolajczyk, K.: Higher-order occurrence pooling for bags-of-words: Visual concept detection. IEEE Trans. Pattern Anal. Mach. Intell. **39** (2017) 313–326
21. Koniusz, P., Wang, L., Cherian, A.: Tensor representations for action recognition. In: IEEE Trans. Pattern Anal. Mach. Intell. (2020)
22. Koniusz, P., Zhang, H.: Power normalizations in fine-grained image, few-shot image and graph classification. In: IEEE Trans. Pattern Anal. Mach. Intell. (2020)
23. Koniusz, P., Tas, Y., Porikli, F.: Domain adaptation by mixture of alignments of second-or higher-order scatter tensors. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society (2017) 7139–7148
24. Koniusz, P., Zhang, H., Porikli, F.: A deeper look at power normalizations. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, IEEE Computer Society (2018) 5774–5783
25. Zhang, H., Zhang, J., Koniusz, P.: Few-shot learning via saliency-guided hallucination of samples. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE (2019) 2770–2779
26. Liu, L., Shen, C., van den Hengel, A.: Cross-convolutional-layer pooling for image recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39** (2017) 2305–2313
27. Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. CoRR **abs/1608.06993** (2016)

28. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA, IEEE Computer Society (2005) 886–893
29. Forsyth, D.A.: Object detection with discriminatively trained part-based models. IEEE Computer **47** (2014) 6–7
30. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), with CD-ROM, 8-14 December 2001, Kauai, HI, USA, IEEE Computer Society (2001) 511–518
31. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In Bartlett, P.L., Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K.Q., eds.: Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States. (2012) 1106–1114
32. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society (2017) 2999–3007
33. Liu, S., Huang, D., Wang, Y.: Receptive field block net for accurate and fast object detection. In Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., eds.: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI. Volume 11215 of Lecture Notes in Computer Science., Springer (2018) 404–419
34. Singh, B., Najibi, M., Davis, L.S.: SNIPER: efficient multi-scale training. In Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., eds.: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada. (2018) 9333–9343
35. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop. Volume 2., Lille (2015)
36. Shyam, P., Gupta, S., Dukkipati, A.: Attentive recurrent comparators. In Precup, D., Teh, Y.W., eds.: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Volume 70 of Proceedings of Machine Learning Research., PMLR (2017) 3173–3181
37. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In Precup, D., Teh, Y.W., eds.: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Volume 70 of Proceedings of Machine Learning Research., PMLR (2017) 1126–1135
38. Gidaris, S., Komodakis, N.: Dynamic few-shot visual learning without forgetting. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, IEEE Computer Society (2018) 4367–4375
39. Simon, C., Koniusz, P., Nock, R., Harandi, M.: Adaptive subspaces for few-shot learning. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2020. (2020)
40. Simon, C., Koniusz, P., Nock, R., Harandi, M.: On modulating the gradient for meta-learning. In: European Conference on Computer Vision. (2020)

41. Zhang, H., Zhang, L., Qi, X., Li, H., Torr, P.H.S., Koniusz, P.: Few-shot action recognition with permutation-invariant attention. In: European Conference on Computer Vision. (2020)
42. Dong, N., Xing, E.P.: Few-shot semantic segmentation with prototype learning. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018, BMVA Press (2018) 79
43. Michaelis, C., Bethge, M., Ecker, A.S.: One-shot segmentation in clutter. In Dy, J.G., Krause, A., eds.: Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018. Volume 80 of Proceedings of Machine Learning Research., PMLR (2018) 3546–3555
44. Hu, T., Yang, P., Zhang, C., Yu, G., Mu, Y., Snoek, C.G.M.: Attention-based multi-context guiding for few-shot semantic segmentation. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press (2019) 8441–8448
45. Gui, L., Wang, Y., Ramanan, D., Moura, J.M.F.: Few-shot human motion prediction via meta-learning. In Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., eds.: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII. Volume 11212 of Lecture Notes in Computer Science., Springer (2018) 441–459
46. y Terán, A.R.M., Gouiffès, M., Lacassagne, L.: Enhanced local binary covariance matrices (ELBCM) for texture analysis and object tracking. In Eisert, P., Gagalowicz, A., eds.: 6th International Conference on Computer Vision / Computer Graphics Collaboration Techniques and Applications, MIRAGE '13, Berlin, Germany - June 06 - 07, 2013, ACM (2013) 10:1–10:8
47. Tuzel, O., Porikli, F., Meer, P.: Region covariance: A fast descriptor for detection and classification. In Leonardis, A., Bischof, H., Pinz, A., eds.: Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part II. Volume 3952 of Lecture Notes in Computer Science., Springer (2006) 589–600
48. Shih, Y., Yeh, Y., Lin, Y., Weng, M., Lu, Y., Chuang, Y.: Deep co-occurrence feature learning for visual object recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society (2017) 7302–7311
49. Koniusz, P., Yan, F., Gosselin, P.H., Mikolajczyk, K.: Higher-order occurrence pooling on mid-and low-level features: Visual concept detection. (2013)
50. Koniusz, P., Yan, F., Mikolajczyk, K.: Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection. Computer vision and image understanding **117** (2013) 479–492
51. Zhu, H., Koniusz, P.: Generalized factorized bilinear graph convolutional networks for text classification. In: ArXiV. (2020)
52. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88** (2010) 303–338
53. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T., eds.: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings,

Part V. Volume 8693 of Lecture Notes in Computer Science., Springer (2014) 740–755
54. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. I. J. Robotics Res. **32** (2013) 1231–1237
55. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, IEEE Computer Society (2009) 248–255