

# Few-Shot Object Detection: A Comprehensive Survey

Mona Köhler<sup>1</sup>, Markus Eisenbach<sup>1</sup>, and Horst-Michael Gross, *Member, IEEE*

**Abstract**—Humans are able to learn to recognize new objects even from a few examples. In contrast, training deep-learning-based object detectors requires huge amounts of annotated data. To avoid the need to acquire and annotate these huge amounts of data, few-shot object detection (FSOD) aims to learn from few object instances of new categories in the target domain. In this survey, we provide an overview of the state of the art in FSOD. We categorize approaches according to their training scheme and architectural layout. For each type of approach, we describe the general realization as well as concepts to improve the performance on novel categories. Whenever appropriate, we give short takeaways regarding these concepts in order to highlight the best ideas. Eventually, we introduce commonly used datasets and their evaluation protocols and analyze the reported benchmark results. As a result, we emphasize common challenges in evaluation and identify the most promising current trends in this emerging field of FSOD.

**Index Terms**—Few-shot learning, meta learning, object detection, survey, transfer learning.

## I. INTRODUCTION

IN THE last decade, object detection has tremendously improved through deep learning [1], [2]. However, deep-learning-based approaches typically require vast amounts of training data. Therefore, it is difficult to apply them to real-world scenarios involving novel objects that are not present in common object detection datasets. Annotating large amounts of images for object detection is costly and tiresome. In some cases—such as medical applications [3] or the detection of rare species [4]—it is even impossible to acquire plenty of images. Moreover, in contrast to typical deep-learning-based approaches, humans are able to learn new concepts with little data even at an early age [5], [6], [7]. When children are shown new objects, they are able to recognize these objects even if they have seen them only once to a few times.

Therefore, a promising research area in this direction is few-shot object detection (FSOD). FSOD aims at detecting novel objects with only few annotated instances after pre-training in the first phase on abundant publicly available data, as shown in Fig. 1. Consequently, it alleviates the burden of annotating large amounts of data in the target domain.

Manuscript received 26 September 2022; revised 17 January 2023; accepted 24 March 2023. Date of publication 17 April 2023; date of current version 4 September 2024. This work was supported by Carl-Zeiss-Stiftung as part of the Engineering for Smart Manufacturing (E4SM) Project and by the Open Access Publication Fund of Ilmenau University of Technology. (Corresponding author: Mona Köhler.)

The authors are with the Neuroinformatics and Cognitive Robotics Laboratory, Ilmenau University of Technology, 98684 Ilmenau, Germany (e-mail: mona.koehler@tu-ilmenau.de).

Digital Object Identifier 10.1109/TNNLS.2023.3265051

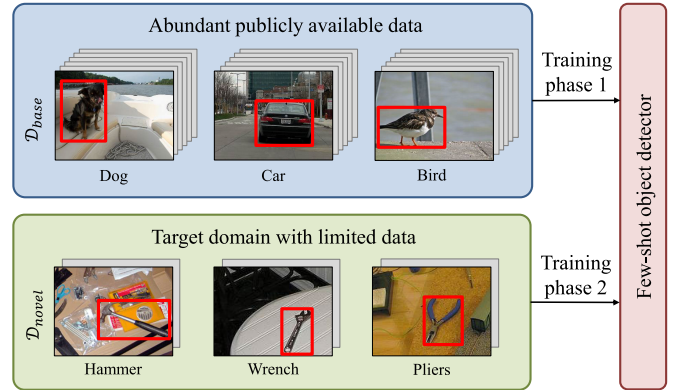


Fig. 1. General idea: by first training on a base dataset with abundant annotated bounding boxes, it is possible to apply few-shot object detectors to settings with only few annotated object instances, such as mechanical tools.

In this survey, we aim to provide an overview of state-of-the-art FSOD approaches for new researchers in this emerging research field. First, we define the problem of FSOD. Afterward, we categorize current approaches and highlight similarities as well as differences. Subsequently, we introduce commonly used datasets and provide benchmark results. Finally, we emphasize common challenges in evaluation and identify promising research directions to guide future research.

## II. PROBLEM DEFINITION

FSOD aims at detecting novel objects with only few annotated instances. Formally, the training dataset  $\mathcal{D} = \mathcal{D}_{base} \cup \mathcal{D}_{novel}$  is separated into two datasets  $\mathcal{D}_{base}$  and  $\mathcal{D}_{novel}$  containing nonoverlapping sets of base categories  $\mathcal{C}_{base}$  and novel categories  $\mathcal{C}_{novel}$ , with  $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \emptyset$ . Each tuple  $(I_i, \hat{y}_{o_1}, \dots, \hat{y}_{o_M}) \in \mathcal{D}$  consists of an image  $I_i = \{o_1, \dots, o_M\}$  containing  $M$  objects  $o_1, \dots, o_M$  and their corresponding labels  $\hat{y}_{o_i} = \{c_{o_i}, b_{o_i}\}$ , including the category  $c_{o_i}$  and the bounding box  $b_{o_i} = \{x_{o_i}, y_{o_i}, w_{o_i}, h_{o_i}\}$  with coordinates  $(x_{o_i}, y_{o_i})$ , width  $w_{o_i}$ , and height  $h_{o_i}$ . For the base categories  $\mathcal{C}_{base}$ , abundant training data are available in the base dataset  $\mathcal{D}_{base}$ . In contrast, the novel dataset  $\mathcal{D}_{novel}$  contains only few annotated object instances for each novel category in  $\mathcal{C}_{novel}$ . For the task of  $K$ -shot object detection, there are exactly  $K$  annotated object instances available for each category in  $\mathcal{C}_{novel}$ . Therefore, the number of annotated novel object instances  $|\{o_j \in I_i \forall I_i \in \mathcal{D}_{novel}\}| = K \cdot |\mathcal{C}_{novel}|$  is relatively small. Note that the number of annotated object instances does not necessarily correspond to the number of images, as one image may contain multiple instances. The most difficult case for FSOD is one-shot object detection, where  $K = 1$ .  $N$ -way object detection denotes a detector that is designed to detect object instances from  $N$  novel categories, where  $N \leq |\mathcal{C}_{novel}|$ . FSOD is therefore often referred to as  $N$ -way  $K$ -shot detection.

Training an object detector only on  $\mathcal{D}_{\text{novel}}$  quickly leads to overfitting and poor generalization due to limited training data [8], [9]. However, training on the highly imbalanced combined data  $\mathcal{D} = \mathcal{D}_{\text{novel}} \cup \mathcal{D}_{\text{base}}$  generally results in a detector that is heavily biased toward the base categories and, therefore, unable to correctly detect instances from novel categories [9]. Therefore, current research focuses on novel approaches for FSOD. Typically, the initial detector model  $\mathcal{M}_{\text{init}}$  equipped with a backbone pretrained on classification data is first trained on  $\mathcal{D}_{\text{base}}$ , resulting in the base model  $\mathcal{M}_{\text{base}}$ . Most approaches then train  $\mathcal{M}_{\text{base}}$  on data  $\mathcal{D}_{\text{finetune}} \subseteq \mathcal{D}$ , including novel categories  $\mathcal{C}_{\text{novel}}$ , resulting in the final model  $\mathcal{M}_{\text{final}}$

$$\mathcal{M}_{\text{init}} \xrightarrow{\mathcal{D}_{\text{base}}} \mathcal{M}_{\text{base}} \xrightarrow{\mathcal{D}_{\text{finetune}}} \mathcal{M}_{\text{final}}. \quad (1)$$

### III. RELATED WORK ON TRAINING WITH LIMITED DATA

There are some related research areas that also focus on training with limited data. In the following, we will briefly discuss differences and similarities with FSOD.

#### A. Related Concepts for Learning With Limited Data

1) *Few-Shot Learning and Classification*: Before being applied to detection, few-shot learning was first explored for classification tasks [10], [11], [12], [13], [14], [15]. As objects with only few training instances do not need to be localized, classification is clearly easier. Yet, many ideas can be adopted for FSOD.

2) *Semisupervised Learning* is related to few-shot learning in that only a few labeled instances of the target categories are available. However, in contrast to few-shot learning, large amounts of additional unlabeled data are often available that help to learn appropriate representations [16], [17], [18], [19].

Thus, when additional unlabeled data are available, methods from semisupervised learning should be considered to improve the learned representations in few-shot learning approaches.

3) *Incremental Learning*: Typical deep-learning approaches suffer from catastrophic forgetting when the model is trained on new data. In contrast, incremental learning approaches [20], [21], [22] aim to retain the performance on old categories when new categories are added incrementally. Some FSOD approaches also incorporate incremental learning techniques.

#### B. Object Detection

1) *Generic Object Detection* is the joint task of localizing and classifying object instances of categories the detector was trained on. Regions of interest (RoIs) are localized by coordinates of bounding boxes and classified into a predefined set of categories. All other object categories, which are not part of the training categories, are regarded as background, and the detector is trained to suppress detections of those other categories. While achieving impressive results, these approaches require loads of annotated object instances per category and typically fail when applied to the few-shot regime. For researchers new in this field, we refer to comprehensive surveys [1], [2] on this topic.

2) *Cross-Domain Object Detection* [23], [24], [25] is the task of first training a detector on abundant labeled data and then adapting this detector to a different domain with limited data; a typical example is synthetic-to-real. However, unlike FSOD, the categories stay the same across different domains.

3) *Zero-Shot Object Detection* can be defined similar to FSOD. However, as an extreme case, the number of annotated object instances is zero ( $K = 0$ ). Zero-shot detectors often incorporate semantic word embeddings [26], [27], [28], i.e., semantically similar categories lead to similar features in the embedding space. This works for detecting everyday objects, which can be easily labeled, but might be problematic when providing a specific label is difficult or when very similar objects need to be distinguished.

4) *Weakly Supervised Object Detection* relaxes the required annotations such that the training data contain only image-level labels, i.e., whether a specific object category is present or absent somewhere in the image [29], [30], [31]. These annotations are much easier to obtain and can often be acquired by keyword search. The challenge for weakly supervised object detectors is detecting all object instances without having any localization information during training. Although alleviating the annotation burden, weakly supervised object detectors still require large amounts of images, which might be hard to obtain for detecting rare objects.

#### C. Learning Techniques for FSOD

In addition to the related research areas described above, in the following, we will address learning techniques that are widely adopted in FSOD.

1) *Transfer Learning* refers to the reuse of network weights pretrained on a baseline dataset to improve generalization capabilities on a new domain with limited data. As in few-shot learning and detection, this usually involves novel categories from the target domain. However, unlike few-shot learning, the number of object instances for novel categories is not necessarily small. Therefore, techniques for learning from few data need to be incorporated in transfer learning approaches for FSOD.

2) *Metric Learning* aims for learning an embedding in which inputs with similar content are encoded in features that have a small distance to each other in terms of the metric, while encoded features from dissimilar inputs are supposed to be far apart [32]. To learn features with low inner-class distances and high inter-class  $\ell^2$  distances, triplet loss [33] or its extensions (see overview in [34]) are often used. Since this learned feature embedding typically generalizes well, the model can also be applied to encode instances of novel categories, which were unknown during training, and make metric-based decisions without the need for retraining. In the context of few-shot classification, this means that during inference, the model extracts feature embeddings of the few annotated examples of  $\mathcal{D}_{\text{novel}}$  as well as of corresponding test images. The test image is then assigned to the category of the closest feature embedding of an annotated example. However, for few-shot detection, concepts for localizing instances in the images need to be integrated.

3) *Meta Learning* approaches learn how to learn in order to generalize for new tasks or new data [13]. For few-shot learning, this means that these approaches learn how to learn to categorize the given inputs even though the categories are not fixed during training. These approaches need to learn how the required knowledge about the category is learned most efficiently so that this category knowledge can also be learned for novel categories with few training examples.

#### D. Related Surveys

Although other surveys on FSOD are available [35], [36], [37], [38], [39], [40], they do not cover as many publications related to FSOD as we do. Works [36], [38], [39] are broader surveys, also addressing self-supervised, weakly supervised, and/or zero-shot learning and do not focus as much on FSOD. Works [35], [36] only cover earlier work on FSOD and hence are somewhat outdated since at least some of the currently best performing approaches on common benchmarks are missing. As work [40] is not available in the English language, it is only accessible to a limited group of researchers. Overall, our survey is most related to [37], as it also elaborates several core concepts and groups approaches according to these concepts. However, with the visual taxonomy in Figs. 4, 7, and 9, we enable the reader to faster grasp which approaches follow similar concepts and what concepts seem to complement each other well. We also provide better guidance on benchmark results by highlighting differences in evaluation protocols and grouping approaches with comparable evaluations. Furthermore, we provide a much more comprehensive survey by covering nearly twice as many FSOD papers as [37] did.

#### IV. CATEGORIZATION OF FSOD APPROACHES

Approaches for FSOD incorporate novel ideas in order to be able to detect objects with only few training examples. In general, the abundant training examples for base categories  $\mathcal{C}_{\text{base}}$  are used to leverage knowledge for the novel categories  $\mathcal{C}_{\text{novel}}$  with limited labeled data.

We categorize approaches for FSOD into meta learning and transfer learning approaches, as shown in Fig. 2. We further divide meta learning approaches into single- and dual-branch architectures. Dual-branch architectures are constituted by a query and a support network branch, i.e., the network processes two inputs (a query and a support image) separately. Single-branch approaches in general resemble the architecture of generic detectors but reduce the number of learnable parameters when training on novel categories or utilize metric learning. Yet, also several dual-branch architectures and some transfer learning approaches incorporate ideas from metric learning. Therefore, to avoid ambiguous categorization, we do not use metric learning as a separate category, as done in early work on FSOD. Instead, we distinguish by training schemes and architectural aspects, which better reflects the different trends in the current state of the art.

FSOD is a rather young but emerging research field as most approaches have been published only within the last three years. Most approaches use transfer learning or dual-branch meta learning.

In the following, we first describe dual-branch meta learning approaches in Section V. We start with the general training scheme for meta learning and follow with the typical realization. In the following, we describe how specific approaches deviate from the general realization. In Section VI, we focus on single-branch meta learning approaches. Although there is no common realization from which others deviate, we still group approaches to their main ideas. In Section VI-D, we cover transfer learning approaches. Similar to dual-branch meta learning approaches, we first describe the general realization and then turn to modifications.

Whenever appropriate, we give short takeaways at the end of the subsections in order to highlight key insights. Some takeaways also contain citations to link the concept to

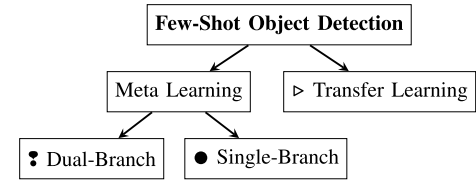


Fig. 2. Categorization of FSOD approaches.

specific well-performing methods with regard to benchmarks in Section IX.

Moreover, we summarize the best performing approaches for each training scheme at the end of the corresponding section. Finally, in Section VIII, we draw a comparison between meta learning and transfer learning approaches before discussing common datasets and benchmark results in Section IX.

#### V. DUAL-BRANCH META LEARNING

A lot of approaches for FSOD utilize meta learning in order to learn how to generalize for novel categories. In this section, we first describe the general training scheme for meta learning in Section V-A. To realize meta learning, dual-branch approaches use a query and a support branch as we outline in Section V-B. After this, we describe how specific approaches deviate from the general realization.

##### A. Training Scheme

For meta learning, the model is trained in multiple stages. First, the model  $\mathcal{M}_{\text{init}}$  is trained only on the base dataset  $\mathcal{D}_{\text{base}}$ , resulting in  $\mathcal{M}_{\text{base}}$ . Typically, an episodic training scheme is applied, where each of the  $E$  episodes mimics the  $N$ -way- $K$ -shot setting. This is called meta training. In each episode  $e$  (also known as few-shot task), the model is trained on  $K$  training examples of  $N$  categories on a random subset  $\mathcal{D}_{\text{meta}}^e \subset \mathcal{D}_{\text{base}}$ ,  $|\mathcal{D}_{\text{meta}}^e| = K \cdot N$ . Therefore, the model needs to learn how to discriminate the presented categories in general depending on the input. Finally, during meta fine-tuning, the model  $\mathcal{M}_{\text{base}}$  is trained on the final task, resulting in  $\mathcal{M}_{\text{final}}$ .

$$\mathcal{M}_{\text{init}} \xrightarrow[\epsilon=1, \dots, E]{\mathcal{D}_{\text{meta}}^e \subset \mathcal{D}_{\text{base}}} \mathcal{M}_{\text{base}} \xrightarrow{\mathcal{D}_{\text{finetune}}} \mathcal{M}_{\text{final}}. \quad (2)$$

If the model is supposed to detect both base and novel categories, it is trained on a balanced set  $\mathcal{D}_{\text{finetune}} \subset \mathcal{D}$  of  $K$  training examples per category, regardless of whether it is a base or a novel category. Otherwise, if we are only interested in the novel categories, the model is trained only on  $\mathcal{D}_{\text{finetune}} = \mathcal{D}_{\text{novel}}$ . Note that some approaches do explicitly not finetune on novel categories but simply apply  $\mathcal{M}_{\text{base}}$  to novel categories, which is called meta testing. During meta testing, the model simply predicts novel objects in the inference mode when presented with  $K$  annotated examples of  $N$  categories.

##### B. General Realization

Dual-branch approaches utilize a two-stream architecture with one query branch  $\mathcal{Q}$  and one support branch  $\mathcal{S}$ , as shown in Fig. 3. The input to the query branch  $\mathcal{Q}$  is an image  $I^{\mathcal{Q}}$  on which the model should detect object instances, whereas the support branch  $\mathcal{S}$  receives the support set  $\mathcal{D}^{\mathcal{S}} = \{(I_i^{\mathcal{S}}, \hat{y}_{o_i})\}_{i=1}^{K \cdot N}$ , with  $K$  support images  $I_i^{\mathcal{S}}$  for each of  $N$  categories and exactly one designated object  $o_j$  and its

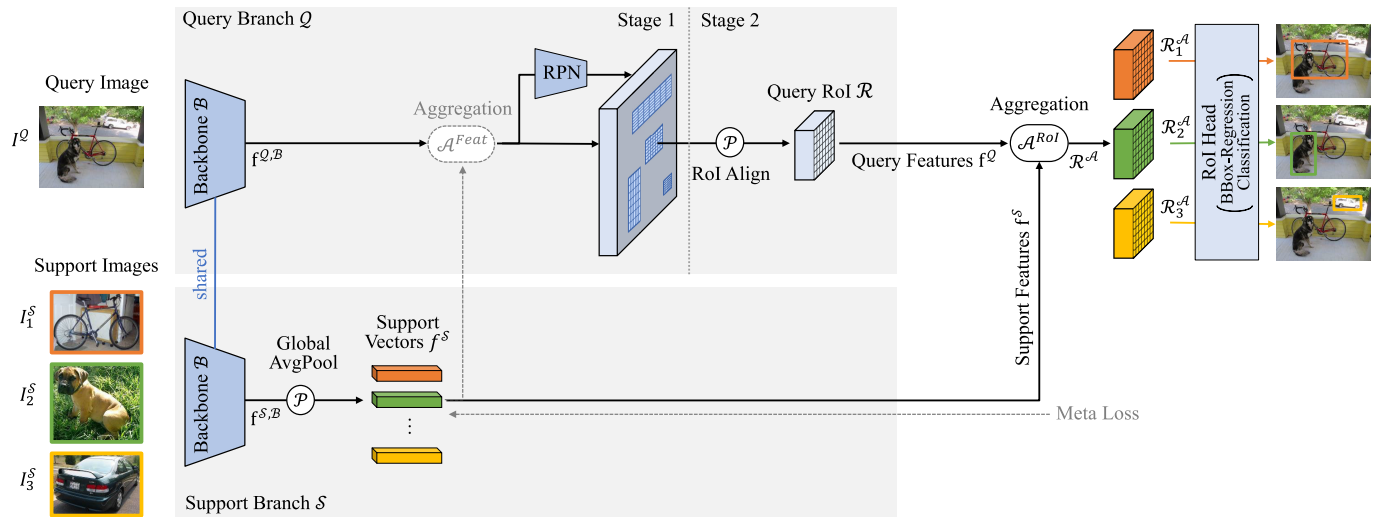


Fig. 3. General architecture for dual-branch meta learning based on Faster R-CNN. Query and support images are fed through a shared backbone. The support features are pooled through global averaging and aggregated with the query features. We show here the one-shot-three-way case without loss of generality.

label  $\hat{y}_{o_j}$  per image. There are three options, how to present the designated object. First, all training examples are already cropped to the designated object by the ground-truth bounding box, as shown in Fig. 3 (bottom). Second, the full-size image and an additional binary mask, indicating the location of the object, are presented. Third, as in [41], the full-size image can be used and the region with features of the designated object is extracted by RoI align [42]. For all three options, we refer to the presented image as support image  $I^S$ . A support image for a specific category  $c$  is denoted as  $I^{S,c}$ .

The support branch  $\mathcal{S}$  is now supposed to extract relevant features  $f^S$  of the support image  $I^S$ . These support features  $f^S$  are then aggregated with the features  $f^Q$  from the query branch  $\mathcal{Q}$ , denoted as  $\mathcal{A}(f^Q, f^S)$ , in order to guide the detector toward detecting object instances of category  $c$  from  $I^{S,c}$  in the query image  $I^Q$ .

Note that the following explanation refers to the most basic and widely used architecture for FSOD with meta learning that is shown in Fig. 3. As shown in Fig. 4, the specific approaches may differ in one or multiple points described here and will be explained in detail in the following.

Many approaches build on top of Faster R-CNN [70] with a ResNet [71] backbone. Often, a Siamese backbone is utilized, i.e., the query branch  $\mathcal{Q}$  and the support branch  $\mathcal{S}$  share their weights. The backbone features  $f^{Q,B}$  of the query branch  $\mathcal{Q}$  are further processed by a region proposal network (RPN) and an RoI align, resulting in the query RoIs  $\mathcal{R}$ . In the support branch  $\mathcal{S}$ , the support features from the backbone  $f^{S,B}$  are pooled through global averaging, resulting in representative support vectors  $f^S$  for each category. In the case of  $K > 1$ , for each category  $c$ , the mean of its support vectors is calculated, resulting in one support vector  $f^{S,c}$  per category. These support vectors encode category-specific information, which are then used to guide the RoI head in recognizing objects of these categories. Therefore, query RoIs  $\mathcal{R}$  and support vectors  $f^S$  are aggregated—shown as  $\mathcal{A}^{\text{RoI}}$  in Fig. 3—in the most simple case by channelwise multiplication  $\mathcal{A}_{\text{mult}}$  as in Equation (3).

After aggregation, for each of the  $N$  categories, there are separate RoIs  $\mathcal{R}^{A,c}$ . Their features are specialized for recognizing objects of the respective category  $c$ . These category-specific RoIs  $\mathcal{R}^{A,c}$  are then fed into a shared RoI

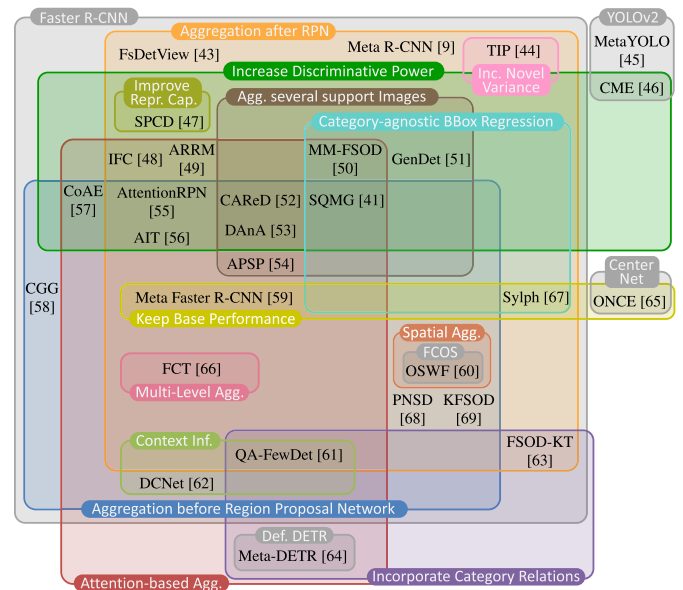


Fig. 4. Categorization of dual-branch meta learning approaches. Best viewed in color.

head for bounding box regression and binary classification. Since the aggregated RoIs  $\mathcal{R}^{A,c}$  already contain category-specific information, the multicategory classification can be replaced by a binary classification that only outputs the information whether the RoI  $\mathcal{R}^{A,c}$  contains an object of the specific category  $c$  or not. To enforce only one category for each RoI  $\mathcal{R}$ , a softmax layer can be applied afterward.

Note that the RoI heads for all categories share the same weights. Therefore, the RoI head must generalize across categories. With this mechanism, it is theoretically possible to detect objects of novel categories without fine-tuning on novel categories, but simply meta testing. This makes meta-learning approaches, especially useful for real-world applications, as no further training is required.

During inference, the support features  $f^S$  of the few images of  $\mathcal{D}_{\text{novel}}$  can be computed once for all  $N$  categories such that the support branch  $\mathcal{S}$  is no longer required.

### C. Variants for Aggregation

The particular dual-branch meta learning approaches differ most in the way the aggregation between query  $f^Q$  and support features  $f^S$  is implemented.

1) *Aggregation Before the RPN*: Typically, the features of the query RoI  $\mathcal{R}$  are aggregated with the support vectors  $f^S$ . However, this requires the RPN to output at least one RoI for each relevant object. Otherwise, even the best aggregation method cannot help in recognizing the desired object. However, the RPN is trained only on base categories. If the novel categories  $\mathcal{C}_{\text{novel}}$  differ a lot from the base categories  $\mathcal{C}_{\text{base}}$ , the RPN might fail to output suitable RoIs for recognizing objects of  $\mathcal{C}_{\text{novel}}$ . Therefore, Fan et al. [55] designed a so-called AttentionRPN, which effectively aggregates query and support features before the RPN. We denote this by  $\mathcal{A}^{\text{Feat}}$  in Fig. 3. Specifically, the support features  $f^S$  are first average pooled and then aggregated with the query features  $f^Q$  by a depthwise cross correlation. Afterward, the RPN is applied onto the enhanced features, resulting in region proposals that are more related to the presented category  $c$  of the support image  $I^{S,c}$ , thus improving the recall.

Zhang et al. [68] (PNSD) built upon this method but replaced average pooling with second-order pooling and power normalization [72]. These second-order representations rather function as a detector of features to capture co-occurrences than a counter as in average pooling. This helps to alleviate the harmful variability of features that stem from varying appearances of objects such as color, viewpoint, and texture.

As second-order pooling is limited to linear correlations, in their follow-up work, Zhang et al. [69] (KFSOD) utilized kernelized covariance matrices [73] and reproducing kernel Hilbert space kernels [74] that capture nonlinear patterns. These kernels can factor out spatial order while keeping rich statistics about each region. Due to this shift invariance, similar objects that vary in physical location, orientation, or viewpoint can be more easily matched.

Furthermore, many others also adopt the idea of AttentionRPN [55], as shown in Fig. 4. Yet, some use a different aggregation operation, which we will discuss in the following.

*Takeaway*: When using Faster R-CNN as a detector, an aggregation before the RPN leads to better region proposals and thus fewer missed detections.

2) *Aggregation Operation*: In the most simple case, the support vectors  $f^S$  and the query features  $f^Q$  are multiplied channelwise

$$\mathcal{A}_{\text{mult}}(f^Q, f^S) = f^Q \odot f^S \quad (3)$$

where  $\odot$  denotes the Hadamard product.

Moreover, different aggregation operations are explored in the state of the art. In AttentionRPN [55] and GenDet [51], support features are convolved/correlated with the query features. Li et al. [60] (OSWF) used cosine similarity between each element of  $f^Q$  and  $f^S$ , which resembles  $\mathcal{A}_{\text{mult}}$  in Equation (3), but with an additional scaling factor.

Michaelis et al. [58] (CGG) and [75] (OSIS) calculated the  $\ell^1$ -distance at each position and concatenated the resulting similarity features to the query features. Xiao and Marlet [43] (FsDetView) used a more complex aggregation operation by combining channelwise multiplication as in  $\mathcal{A}_{\text{mult}}$  with subtraction and query features themselves similar to [58]

$$\mathcal{A}(f^Q, f^S) = [f^Q \odot f^S, f^Q - f^S, f^Q] \quad (4)$$

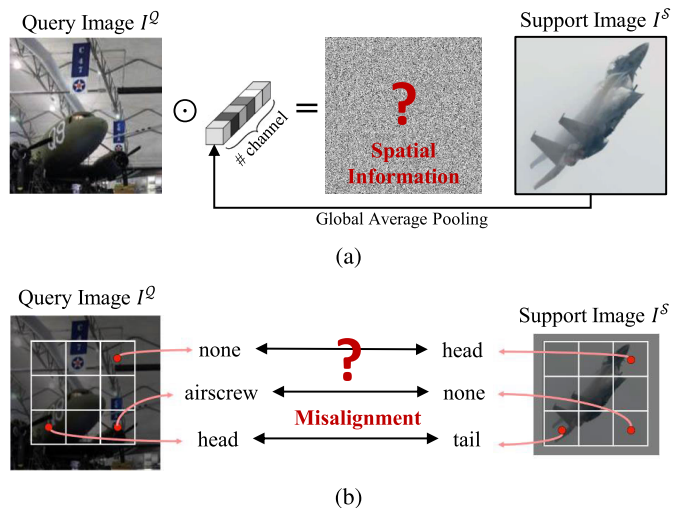


Fig. 5. Common aggregation problems. Images based on [53]. (a) Loss of spatial information due to global average pooling. (b) Spatial misalignment due to convolution-based aggregation.

where  $[\cdot, \cdot]$  denotes channelwise concatenation.

Meta Faster R-CNN [59] builds upon this aggregation

$$\mathcal{A} = [\Phi_{\text{Mult}}(f^Q \odot f^S), \Phi_{\text{Sub}}(f^Q - f^S), \Phi_{\text{Cat}}[f^Q, f^S]] \quad (5)$$

where  $\Phi_{\text{Mult}}$ ,  $\Phi_{\text{Sub}}$ , and  $\Phi_{\text{Cat}}$  each denote a small convolutional network with three conv and rectified linear unit (ReLU) layers.

Zhang et al. [41] (SQMG) decided to enhance the query features  $f^Q$  by support features  $f^S$  with dynamic convolution [76].  $f^S$  is fed into a kernel generator to generate the weights of the convolution. Afterward, the generated weights are convolved with  $f^Q$ .

*Takeaway*: The simple channelwise multiplication of support features  $f^S$  and query features  $f^Q$  cannot fully exploit the information they contain.

3) *Keep Spatial Information for Aggregation*: As opposed to aggregating support features via average pooling, others (see Fig. 4) propose to utilize spatial information. For the object detection task, objects are located by bounding boxes. However, not every part of that bounding box is occupied by the object and, therefore, does not contain relevant information about the respective category. With average pooling however, these irrelevant features are aggregated into the support vector. Moreover, with global average pooling, spatial information is completely lost, as shown in Fig. 5(a).

Therefore, Li et al. [60] (OSWF) first pooled support features to the same spatial dimension as the query RoI  $\mathcal{R}$ . Afterward, these pooled features are concatenated to the query RoI  $\mathcal{R}$ . Finally,  $1 \times 1$  convolutions are used to compare structure-aware local features.

However, Chen et al. [53] argued that a convolution of query features  $f^Q$  and support features  $f^S$  is less suitable since the objects in query images  $I^Q$  and support images  $I^S$  are generally not aligned in the same way, as shown in Fig. 5(b). Therefore, they design an attention-based aggregation as described in the following.

*Takeaway*: In order to incorporate the valuable spatial information of a support image  $I^S$ , its features should not be simply averaged for aggregation.

4) *Attention-Based Aggregation*: Lately, attention mechanisms could significantly improve performance on many vision

tasks [77]. Thus, it is not surprising that the aggregation of support and query features also benefits from incorporating attention mechanisms. These attention mechanisms range from traditional, over nonlocal [78] to multihead attention as in transformers [79]. We will discuss all of them in the following.

Chen et al. [53] (DAnA) aimed to incorporate the spatial correlations between query image  $I^Q$  and support image  $I^S$  but also considered that these images are generally not aligned (see Fig. 5(b)). Therefore, dual-awareness attention first highlights relevant semantic features of the respective category on the support features  $f^S$  and suppresses background information. Afterward, the spatial correlations are incorporated with an attention-based aggregation. This spatial misalignment is also addressed in Meta Faster R-CNN [59]. Using two attention modules, the support and RoI features are first spatially aligned, and then, the foreground regions are highlighted.

Wang et al. [48] (IFC) first used a self-attention module on top of average- and max-pooled query features to separately mine local semantic and detailed texture information. Afterward, with a new feature aggregation mechanism based on a learnable soft-threshold operator [80], redundant information can be shrunk while enhancing feature sensitivity and stability for both novel and base categories.

Huang et al. [49] (ARRM) aimed to achieve a better interaction of support and query features by designing an attention-based affinity relation reasoning module consisting of several convolutions and matrix multiplications of different features. With an additional global-average-pooling branch, also the global semantic context of the support features is integrated. Using this attention-based module for aggregation, misclassifications can be reduced.

Hsieh et al. [57] (CoAE) proposed a coattention method in order to make the query features  $f^Q$  attend to the support features  $f^S$  and vice versa. Therefore, two mutual nonlocal operations [78] are utilized, which receive inputs from both  $f^Q$  and  $f^S$ . This helps the RPN to compute region proposals that are able to better locate objects of the category  $c$  from the support image  $I^{S,c}$ . Moreover, Hsieh et al. [57] proposed a subsequent squeeze-and-coexcitation method—extending the squeeze-and-excitation of SENet [81]—in order to highlight correlated feature channels to detect relevant proposals and eventually the target objects. A similar coattention is utilized by Hu et al. [62] (DCNet).

With AIT, Chen et al. [56] pushed the idea of CoAE [57] a little further. Instead of using a single nonlocal block, multihead coattention is utilized for aggregating query and support features before the RPN. Let  $\mathbf{V}$ ,  $\mathbf{K}$  and  $\mathbf{Q}$  be the value, key, and query of a transformer-based attention [79]. Similar to the coattention in CoAE, query features stem from another branch

$$\mathbf{F}^Q = \text{attn}(\mathbf{V}^Q, \mathbf{K}^Q, \mathbf{Q}^S), \quad \mathbf{F}^S = \text{attn}(\mathbf{V}^S, \mathbf{K}^S, \mathbf{Q}^Q) \quad (6)$$

where superscripts  $Q$  and  $S$  denote whether features are from the query or support branch. The resulting features  $\mathbf{F}^Q$  encode related visual characteristics of both the query image  $I^Q$  and the support image  $I^S$ , which helps the RPN to predict RoIs related to  $I^S$ . According to Chen et al. [56], this improves the accuracy compared to the nonlocal attention block [78] in CoAE [57]. After the RPN, AIT [56] uses a transformer-based encoder–decoder architecture for transforming the RoIs  $\mathcal{R}$  to

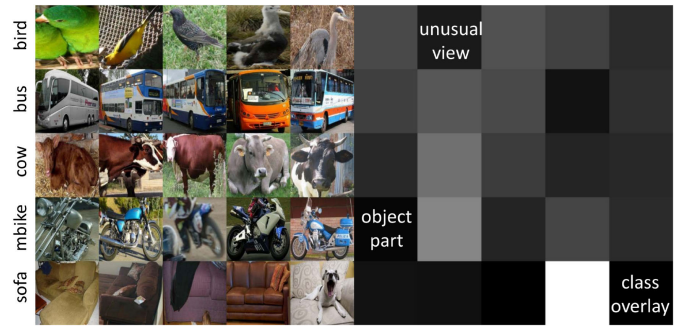


Fig. 6. Different amount of information for several support images of the same category. Image from GenDet [51].

emphasize visual features corresponding to the given support image  $I^S$ .

A similar aggregation to  $\mathbf{F}^S$  in Equation (6) is also used in Meta-DETR [64] and APSP [54]. However, both Meta-DETR and APSP first enhance query or support features as we will describe in the following.

*Takeaway:* The spatial information of a support image  $I^S$ , as well as its relation to a query image  $I^Q$ , is best incorporated through transformer-based attention mechanisms [78], [79].

5) *Multilevel Aggregation:* So far, support and query features were only aggregated after feature extraction in the backbone. However, Han et al. [66] (FCT) argued that multilevel feature interactions between the query and support branch could better align the features. Therefore, they come up with a novel fully cross-transformer based on the improved Pyramid Vision Transformer PVTv2 [82]. The FCT model consists of three interaction stages between query and support in the backbone and one additional interaction stage in the detection head. Finally, a pairwise matching similar to the one from AttentionRPN [55] outputs the final detections.

*Takeaway:* Aggregation of low-, mid-, and high-level features can boost the performance.

6) *Aggregation of Several Support Images:* In the general approach, to fuse all support images of category  $c$ , the mean of their features is calculated

$$\{I_i^{S,c}\}_{i=1}^K : f^{S,c} = \frac{1}{K} \sum_{i=1}^K f_i^{S,c}. \quad (7)$$

However, not all support images provide the same amount of information for the respective category, as shown in Fig. 6. Unusual object views, object parts, or even occlusion by objects of other categories impair the discriminative power if support features  $f_i^{S,c}$  are simply averaged.

Therefore, a weighted average is proposed in GenDet [51]. The weight  $w_i$  for each support image  $I_i^{S,c}$  is computed by the similarity between the single-shot and the mean detector and learned during training

$$\{I_i^{S,c}\}_{i=1}^K : f^{S,c} = \frac{1}{K} \sum_{i=1}^K w_i \cdot f_i^{S,c}. \quad (8)$$

Quan et al. [52] (CAREd) followed a similar approach. However, the weight  $w_i$  is determined by the softmax over the correlation between the support features  $f_i^{S,c}$  and all other support features  $\{f_j^{S,c}\}_{j=1}^K$  of the same category  $c$ . Due to the softmax, the weighting factors already sum up to 1 and the factor  $(1/k)$  is omitted.

DAnA [53], SQMG [41], as well as APSP [54] incorporate the similarity of query  $f^Q$  and different support features  $f^S$ . In DAnA [53], support features  $f_i^{S,c}$  of  $K$  different images  $\{I_i^{S,c}\}_{i=1}^K$  are first aggregated independently with the query features  $f^Q$  based on the correlation between query and support. As the importance of each support images  $I_i^{S,c}$  is already incorporated, the resulting  $K$  aggregated features can be simply averaged

$$\left\{I_i^{S,c}\right\}_{i=1}^K : \mathcal{A}(f^Q, f^{S,c}) = \frac{1}{K} \sum_{i=1}^K \mathcal{A}(f^Q, f_i^{S,c}). \quad (9)$$

In SQMG [41], the support features  $f_i^{S,c}$  of multiple support images  $I_i^{S,c}$  are weighted according to their similarity with the query features  $f^Q$  using an attention mechanism. First, the similarity is computed with a relation network [83]. Afterward, the weighting values  $w_i$  for support features  $f_i^{S,c}$  are computed with a softmax on the similarity score. The final support features are achieved by a weighted sum as in Equation (8).

Lee et al. [54] (APSP) first used a multihead attention to refine each individual support vector  $f_i^{S,c}$  by incorporating all other support vectors of the same category  $c$ . Afterward, instead of computing one single support vector, all  $K$  support vectors  $\{f_i^{S,c}\}_{i=1}^K$  are utilized in a second multihead attention for aggregation with the query features. Thus, not all variances of different support images need to be incorporated in a single support vector and therefore lead to more robust features.

*Takeaway:* As not all support images provide the same amount of information, their individual relevance should be incorporated, as shown in Fig. 6.

#### D. Incorporate Relations Between Categories

Han et al. [61] (QA-FewDet) highlighted the problem that many dual-branch meta-learning approaches work as a kind of single-category detector without modeling multicategory relations. However, especially for novel categories resembling base categories, these relations can help in correctly classifying objects (e.g., a motorbike is more similar to a bicycle than to an airplane).

Therefore, in contrast to using visual features only, Kim et al. [63] (FSOD-KT) additionally incorporated linguistic features. Before aggregation, the support vectors  $f^S$  are fed through a knowledge transfer module, which exploits semantic correlations between different categories. This knowledge transfer module is implemented by a graph convolutional network [84]. The input to this graph convolutional network is a graph where each node represents one category, and the values on the edges represent the similarities between linguistic category names. However, this is only applicable if all categories have predefined and distinct category names and might be hard to transfer to, e.g., medical imaging.

Han et al. [61] (QA-FewDet) also utilized graph convolutions but did not rely on the linguistic category names. In contrast, they build a heterogeneous graph, which enhances support vectors  $f^S$  with multicategory relations in order to better model their relations and incorporate features from similar categories. Moreover, their heterogeneous graph also aligns support and query features. Since the support features  $f^{S,c}$  of one category  $c$  are only extracted from few support images, there might be a huge discrepancy to query RoIs  $\mathcal{R}^c$

that actually belong to the same category  $c$ . Therefore, the heterogeneous graph also contains pairwise edges between RoIs, in order to mutually adapt features of  $f^{S,c}$  and  $\mathcal{R}^c$  and reduce their discrepancy.

Although not using graph convolutions, Zhang et al. [64] (Meta-DETR) also incorporated relations between different categories by transforming their support features. The authors introduce a correlation aggregation module, which is able to simultaneously aggregate multiple support categories in order to capture their interclass correlation. This helps in reducing misclassification and enhances generalization to novel categories. First, the query features  $f^Q$  are matched with multiple support features  $f^S$  simultaneously by utilizing attention modules [79]. Afterward, task encodings help to differentiate these support categories.

*Takeaway:* Incorporating the relations between different categories helps in better representing and classifying the data-sparse novel categories  $\mathcal{C}_{\text{novel}}$ .

#### E. Increase Discriminative Power

After aggregation, for each RoI  $\mathcal{R}$ , there exist  $N$  category-specific RoIs  $\mathcal{R}^{A,c}$ , which are classified independently. If the support features  $f^S$  for different categories are too similar, this independent classification might lead to ambiguities. Therefore, some approaches use an additional meta loss to enforce the support features  $f^S$  to be as diverse as possible. Most often (e.g., in [9], [43], [44], [61], and [63]), the support features  $f^S$  are classified, and a simple cross-entropy loss is applied. This encourages the support vectors to fall in the category the respective object belongs to. More advanced approaches utilize techniques from metric learning to increase the discriminative power, as described in the following.

GenDet [51] and Meta-DETR [64] use a loss based on cosine similarity for more discriminative support vectors. First, the support vectors  $f^S$  are normalized. Afterward, for each pair of support vectors  $(f^{S,c_i}, f^{S,c_j})$ , the cosine similarity is computed, which results in a similarity matrix  $A \in \mathbb{R}^{N \times N}$ , where  $N$  is the number of different categories. With an  $\ell^1$  loss, the similarity matrix  $A$  is constrained to be close to the identity matrix  $I_N \in \mathbb{R}^{N \times N}$ . Intuitively speaking, this results in minimizing the similarity between different support vectors and maximizing the discriminative ability of each support vector, i.e., a high margin between different support vectors.

Wang et al. [48] (IFC), Kobayashi [47] (SPCD), and Huang et al. [49] (ARRM) also used a cosine loss, but in ARRM, an additional margin is added to further increase discrimination and reduce misclassification.

MM-FSOD [50] uses the Pearson distance for aggregating  $f^S$  and  $f^Q$ . Compared to cosine similarity, the Pearson distance first normalizes each dimension with the mean of all dimensions, resulting in a smaller inner class variance. Therefore, there is no need for designing a special distance loss function, and the simple cross-entropy loss can be utilized.

Li et al. [46] (CME) proposed an adversarial training procedure for min-max-margin: Next to a loss for increasing the margin, the features of novel categories are disturbed to reduce the discriminative power of their support vectors and, thus, decrease the margin. To be precise, the most discriminative pixels are erased in an adversarial manner by backpropagating the gradient to the input support image. With this approach, CME [46] is capable of accurately detecting more objects with fewer false positives.

For the meta learning approach, the detector is supposed to detect objects in a query image  $I^Q$  that are of the same category  $c$  as the object in the support image  $I^{S,c}$ . Due to this problem definition, meta learning approaches tend to focus on separating foreground from background instead of distinguishing different categories, as noted by Zhang et al. [41] (SQMG). This often leads to false positives, i.e., predicted bounding boxes, even though the query image  $I^Q$  does not contain any instance of the regarded category  $c$ . However, it is equally important that the detector can distinguish different categories and identify which object categories are not present in the query image.

Therefore, in AttentionRPN [55], a multirelation detector as well as a two-way contrastive training strategy is proposed. The multirelation detector incorporates global, local, and patch-based relations between support features  $f^S$  and query RoIs  $\mathcal{R}$  in order to measure their similarity. The outputs of all three matching modules are summed to give the final matching score. Many others [52], [66], [68], [69] adopt or build upon this multirelation detector. The additionally proposed two-way contrastive training strategy is implemented as follows. In addition to a positive support image  $I^{S,c}$ , a negative support image  $I^{S,n}$  is used from an object category  $n \in \mathcal{C} \setminus \{c\}$  that is not present in the query image  $I^Q$ . This two-way-contrastive training strategy is adapted by DAnA [53] and, similarly, by CAReD [52]. Zhang et al. [41] (SQMG) extended the contrastive loss with an adaptive margin [85] in order to separate the different categories by a proper distance. The adaptive margin incorporates semantic similarity of the categories by word embeddings [86].

A second problem highlighted by Zhang et al. [41] (SQMG) is the extreme imbalance of many background proposals versus few foreground proposals, which impedes distance metric learning. To combat the foreground–background imbalance, the authors use a focal loss [87], which downweights the easy background proposals and focuses on the hard negatives.

CoAE [57] uses an additional margin-based loss to improve the ranking of the RoIs in the RPN. Those RoIs with high similarity to the object in the support image  $I^S$  should be at the top of the ranking since only the top 128 RoIs will be further processed. Therefore, the authors designed a margin-based metric to predict the similarities for all RoIs. Chen et al. [56] (AIT) adopted this margin-based ranking loss.

In typical episodic training, only  $N$  categories are presented in each episode. According to Liu et al. [51] (GenDet), this could lead to a low discriminative ability of the extracted features, as only the sampled categories are distinguished. Thus, their approach GenDet [51] utilizes an additional reference detector during training, where all base categories  $\mathcal{C}_{\text{base}}$  need to be distinguished. The index of a specific base category stays the same over all episodes. Via an additional loss, both detectors are constrained to output similar results. This guides the backbone to extract more discriminative features.

*Takeaway:* In order to increase the discriminative power and differentiate between several categories, ideas from metric learning such as similarity metrics as well as contrastive training should be employed.

### F. Improve Representation Capability

Kobayashi [47] (SPCD) emphasized that during base training, all other nonbase categories are treated as negative. This leads to insufficient expressive power to identify novel categories. Therefore, they introduce an additional self-supervised

module. With selective search [88], rectangular regions different to those from base categories are extracted, and the network is taught to detect the same regions before and after applying strong data augmentation in a self-supervised manner.

### G. Proposal-Free Detectors

Most approaches build on top of the two-stage detector Faster R-CNN [70]. However, these approaches need to deal with possibly inaccurate region proposals and the decision of whether to aggregate support features  $f^S$  and query features  $f^Q$  before or after the RPN or both. When utilizing proposal-free detectors,  $f^S$  and  $f^Q$  can simply be aggregated after feature extraction and before classification and bounding box regression.

Some approaches utilize simple one-stage detectors, such as YOLOv2 [89] in MetaYOLO [45] and CME [46] or RetinaNet [87] in DAnA [53]. Others build on top of anchor-free detectors, such as CenterNet [90] in ONCE [65] or FCOS [91] in Li et al. [60] (OSWF) and GenDet [51]. The transformer-based detector Deformable DETR [92] is utilized in Meta-DETR [64]. Meta-DETR aggregates support features  $f^S$  and query features  $f^Q$  after the shared backbone. Subsequently, a category-agnostic transformer architecture predicts the objects.

*Takeaway:* While most approaches build on top of Faster R-CNN, proposal-free detectors are easier to implement. In particular, transformer-based architectures, such as Meta-DETR [64], already surpass other approaches.

### H. Keep the Performance on Base Categories

In order to better detect base categories and prevent catastrophic forgetting, Han et al. [59] (Meta Faster R-CNN) used an additional branch following the original Faster R-CNN [70] architecture. As Meta Faster R-CNN already aggregates query features  $f^Q$  and support features  $f^S$  before the RPN, only the weights for the backbone are shared between those two branches. After meta training on the base categories  $\mathcal{C}_{\text{base}}$ , the weights of the backbone are fixed and the RPN and RoI head for the base category branch are trained. Finally, the other branch is adapted or simply applied to novel categories with meta fine-tuning or meta testing, respectively (see Section V-A for terminology definitions). As the first branch stays fixed, the performance for base categories  $\mathcal{C}_{\text{base}}$  will not drop due to meta fine-tuning.

For the incremental learning approaches ONCE [65] and Sylph [67], the weights for the already learned categories also stay fixed. Instead of a softmax-based classifier, Sylph uses several independent binary sigmoid-based classifiers (one for each category) such that the categories do not influence each other. For each novel category  $c$ , a hypernetwork on top of the support branch  $\mathcal{S}$  generates the weights for its classifier. Thus, no meta fine-tuning is required.

### I. Increase the Variance of Novel Categories

TIP [44] expands the few training examples for novel categories with data augmentation techniques such as Gaussian noise or cutout. However, naively adding data augmentation impairs detection performance. Therefore, Li and Li [44] (TIP) used an additional transformed guidance consistency loss, implemented by  $\ell^2$  norm, which constrains support vectors  $f_i^S$

and  $f_j^S$  generated by original image  $I_i^S$  and transformed image  $I_j^S = \phi(I_i^S)$  to be close to each other. This results in more similar and representative support vectors even for different support images, thus improving the detection performance of novel categories. Moreover, during training, the query branch  $\mathcal{Q}$  also receives transformed as well as original images. The features of the transformed query image  $I^{\mathcal{Q}}$  are fed into the RPN to predict RoIs. These RoIs are then cropped from the features of the original nontransformed query image via RoI Align [42]. This forces the detector to predict consistent RoIs independent of the transformation used for the query image.

### J. Incorporate Context Information

Typically, by applying RoI pool or RoI align, region proposals are pooled to a specific squared size of, e.g.,  $7 \times 7$ . However, this might lead to information loss during training, which could be remedied with abundant training data. With only a few training examples available, this information loss could result in misleading detections. Therefore, DCNet [62] uses three different resolutions and performs parallel pooling. Similar to the pyramid pooling module in the PSPNet [93] for semantic segmentation, this helps to extract context information, where larger resolutions help to focus on local details, while smaller resolutions help to capture holistic information. In contrast to the pyramid pooling module, the branches are fused with attention-based summation.

Han et al. [61] (QA-FewDet) found that query RoIs  $\mathcal{R}$  might be noisy and may not contain complete objects. Therefore, they built a heterogeneous graph that uses graph convolutional layers [84]. Pairwise edges between proposal nodes incorporate both local and global contexts of different RoIs in order to improve classification and bounding box regression.

### K. Category-Agnostic Bounding Box Regression

Even though parameters for binary classification and bounding box regression are shared for all categories, most approaches compute them for each category-specific RoI independently. In contrast, GenDet [51], MM-FSOD [50], SQMG [41], and Sylph [67] share the bounding box computation among different categories. This follows the intuition that even though different categories vary in their visual appearances, regression of bounding box values has common traits. Moreover, it saves computation overhead.

### Summary of Best Performing Dual-Branch Meta Learning Approaches

In the following, we summarize selected dual-branch meta learning approaches that perform best on FSOD benchmark datasets (see Section IX), in order to highlight their key concepts.

*Meta-DETR* [64] is the first approach building on top of the transformer-based detector DETR. Without depending on accurate region proposals, Meta-DETR circumvents the challenge to adapt these for novel categories. Moreover, in its attention-based aggregation module, the correlation between different categories is incorporated, which reduces misclassification. With an additional loss based on cosine similarity, the learned features are more discriminative and, thus, enhance generalization.

*FCT* [66] also uses a transformer, but instead of DETR, the ResNet backbone of Faster R-CNN is simply replaced by the improved Pyramid Vision Transformer PVTv2. However, support and query features are aggregated at multiple levels to

better align the features. Moreover, a multirelation detector computes similarities between support and query features to output the final detections.

*IFC* [48] does not build on top of transformers but utilizes an interactive self-attention module to capture the discriminating features from scarce novel categories. Moreover, a novel feature aggregation mechanism is introduced, which aims at shrinking redundant information while enhancing feature sensitivity and stability for both novel and base categories. Finally, an orthogonal cosine loss enhances foreground distinguishability.

One of the few approaches not requiring fine-tuning is *SQMG* [41]. In SQMG, both support and query features are enhanced through mutual guidance. First, this helps to generate more category-aware region proposals. Second, the individual relevance of multiple support images is also incorporated. Moreover, SQMG focuses on correct classification with different training techniques. To alleviate the confusion of similar categories, a two-way contrastive training strategy with an adaptive margin is employed. To combat the imbalance between many background proposals versus few foreground proposals, an additional focal loss is incorporated. Finally, the bounding box regression is shared among different categories in order to focus on classification.

### Conclusion on Dual-Branch Meta Learning Approaches

Dual-branch meta learning approaches are very common in FSOD. They enable fast adaption for novel categories or can even be applied to novel categories without fine-tuning but with a simple forward pass, i.e., meta testing. This is especially useful for real-world applications. However, they require a complex episodic training scheme, as described in Section V-A. Nevertheless, by utilizing attention-based aggregations and incorporating metric learning techniques, dual-branch meta learning approaches can achieve state-of-the-art results, as we will discuss in Section IX.

## VI. SINGLE-BRANCH META LEARNING

Single-branch architectures for FSOD follow another approach. Since there are no query and support branches, the general architecture resembles the architecture for generic object detectors such as Faster R-CNN [70]. However, there is no single approach from which others deviate. Still, all approaches use episodic training as described in Section V-A, which is typical for meta learning. In Fig. 7, we display our categorization of single-branch meta learning approaches, which we further describe in the following.

### A. Metric Learning

Similar to dual-branch meta learning, metric learning plays a key role in single-branch approaches.

One of the first approaches for FSOD—RepMet [94]—defines the FSOD task as a distance metric learning problem. For localization, RepMet simply uses the RoIs  $\mathcal{R}$  from Faster R-CNN [70]. Embedded feature vectors  $f^+$  of these RoIs are compared to multiple learned representatives for each category, in order to determine the category for an RoI. To learn an adequate feature embedding, an additional embedding loss is used, which enforces a minimum margin between the distance of the embedded vector to the closest representative of the correct category and the distance to the closest representative of the wrong category.

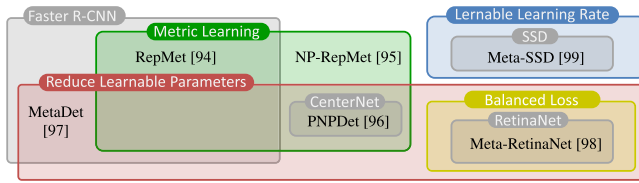


Fig. 7. Categorization of single-branch meta learning approaches. Best viewed in color.

RepMet uses the positive region proposals of a category but discards its negative proposals. However, for learning the embedding space, negative—especially hard negative—proposals are essential. Therefore, NP-RepMet [95] also learns negative embedded feature vectors  $f^-$  and negative representative vectors per category. The embedding spaces for the representatives are learned by utilizing a triplet loss [33].

PNPDet [96] uses cosine similarity for distance metric learning of the objects’ categories to allow for better generalization to novel categories. Cosine similarity computes the similarity of the input image’s features with learned prototypes of each category.

*Takeaway:* Metric learning helps in creating more discriminative features for better distinguishing between different categories.

### B. Reduce Learnable Parameters

Since few training examples of novel categories might not be sufficient to train a deep neural network, some approaches reduce the number of learnable parameters for few-shot fine-tuning.

After training MetaDet [97] on the base dataset, category-agnostic weights (i.e., backbone and RPN of Faster R-CNN) are frozen, and an episodic training scheme is applied to learn how to predict category-specific weights first for the base categories  $C_{\text{base}}$  and then for the novel categories  $C_{\text{novel}}$ . For inference, the meta model can be detached, and the detector looks like the standard Faster R-CNN.

Li et al. [98] (MetaRetinaNet) reduced the number of learnable parameters by freezing all backbone layers after training on  $D_{\text{base}}$  and instead learn coefficient vectors  $v$  initialized to ones. These learnable coefficient vectors  $v$  are multiplied with the convolution weights  $w$ , resulting in a modified convolution operation:  $f_{\text{out}} = f_{\text{in}} \otimes (w \odot v) \oplus b$ .

Zhang et al. [96] (PNPDet) froze the whole network after training on  $D_{\text{novel}}$ . For few-shot fine-tuning, a second small subnetwork is introduced for learning to classify the novel categories  $C_{\text{novel}}$ . This disentangling of novel and base categories prevents a decreasing performance on base categories.

*Takeaway:* When training on data scarce  $D_{\text{novel}}$ , the number of learnable parameters should be reduced.

### C. Learnable Learning Rate

Fu et al. [99] designed their Meta-single shot detector (SSD) such that the model’s parameters can adjust fast—with just one parameter update—to the novel categories  $C_{\text{novel}}$ . All parameters from the original SSD detector [100] get an additional learnable learning rate. During meta learning, these learning rates are learned individually by a meta learner from the distribution of the current task, resulting in neither overfitting nor underfitting.

### D. Balanced Loss Function

Li et al. [98] highlighted that for meta training, in each episode, different training examples of different categories

are sampled, and they achieve different performances. This performance imbalance hinders stability and makes it difficult to adapt the model to novel categories. Thus, in their MetaRetinaNet, a balancing loss is introduced, which constrains the detector to achieve similar performance across episodes.

### Conclusion on Single-Branch Meta Learning Approaches

Single-branch meta learning approaches are much less explored in FSOD. Thus, more advanced dual-branch approaches or transfer learning approaches are able to surpass the approaches presented here.

## VII. TRANSFER LEARNING

Meta learning approaches depend on complex episodic training. In contrast, transfer learning approaches utilize a fairly simple two-phase approach on a single-branch architecture, most often a Faster R-CNN [70], as first proposed by Wang et al. [101] (TFA) and shown in Fig. 8.

In the first phase, the detector is trained on the base categories  $C_{\text{base}}$ . Afterward, all detector weights are frozen except for RoI head, which is responsible for bounding box regression and classification. In the second phase, transfer learning is performed, by fine-tuning the last layers on the base categories  $C_{\text{base}}$  and novel categories  $C_{\text{novel}}$ . For fine-tuning, the training set is composed of balanced subsets of base category data  $D_{\text{base}}$  and novel category data  $D_{\text{novel}}$  with  $K$  shots for each of the base and novel categories. The only modification to Faster R-CNN is the use of cosine similarity for classification, which is crucial to compensate for differences in feature norms of base categories  $C_{\text{base}}$  and novel categories  $C_{\text{novel}}$ , as analyses in [102] have shown. Wang et al. [101] showed that this simple approach is sufficient to adequately learn the novel categories  $C_{\text{novel}}$  and outperform earlier meta learning approaches that are more complex.

Building upon this simple approach, many modifications have been proposed. Fig. 9 shows all transfer learning approaches categorized by the architecture employed and by their modifications. In the following, we describe all the proposed modifications, grouped by the categories shown.

### A. Modifications of the RPN

For the very few-shot setting, where the number of instances  $K$  for novel categories  $C_{\text{novel}}$  is very low, the RPN was identified as a key source for errors [107]. For example, if the detector must learn to detect a category from a single example, the detector can model the categories’ variation only by proposing multiple RoIs that match the object’s ground truth, which is similar to random cropping augmentation, as shown in [102] (see Section VII-C). If the RPN misses even one of these RoIs, the performance on this novel category may drop noticeably. Therefore, Zhang et al. [107] (CoRPN) modified the RPN, by replacing the single binary foreground classifier in the RPN with  $M$  binary classifiers. The goal is that at least one classifier identifies the relevant RoI as foreground. Vu et al. [109] (FORD + BL) added an atrous spatial pyramid pooling (ASPP) [127] context module before the RPN to increase its receptive field. This helps in identifying relevant RoI as foreground.

As in TFA [101], in the second training phase, the weights of the RPN are frozen in many transfer learning approaches. Fan et al. [106] (Retentive R-CNN) observed that the RPN suppresses RoIs of novel categories  $C_{\text{novel}}$  after it was trained

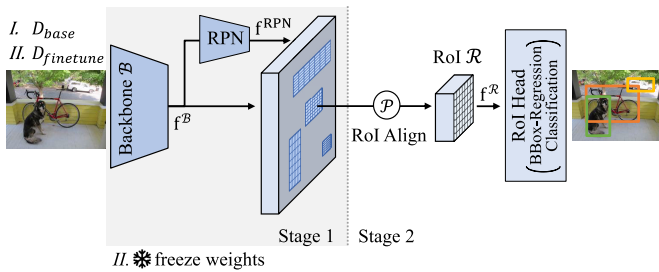


Fig. 8. Realization with transfer learning.

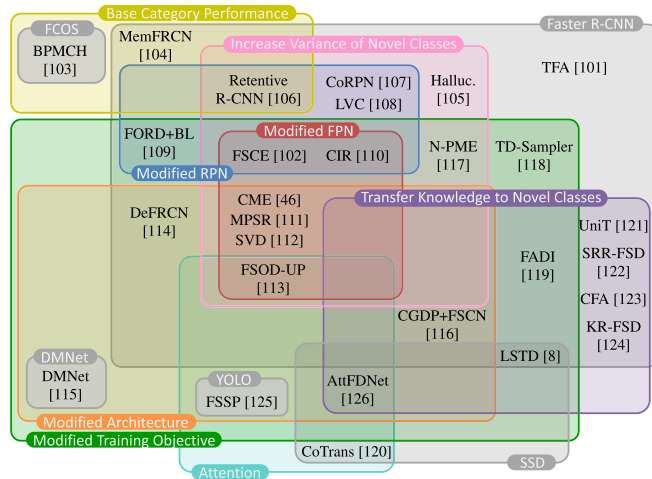


Fig. 9. Transfer learning approaches categorized by detector architecture and types of modifications. Best viewed in color.

only on  $\mathcal{D}_{\text{base}}$  in the first phase. They found that unfreezing the weights of the RPN's final layer that classifies whether objects are foreground or background is sufficient to improve the RPN in the second phase. The same conclusion was drawn by Sun et al. [102] (FSCE), Kaul et al. [108] (LVC), and Wang et al. [110] (CIR), and as a result, all RPN weights were unfrozen. In addition, FSCE and CIR doubled the number of proposals that pass nonmaximum suppression (NMS) to get more proposals for novel categories. FSCE compensates for this by sampling only half the number of proposals in the RoI head used for loss computation, as they observed that in the second training phase, the discarded half contains only backgrounds.

*Takeaway:* To reduce the number of missed detections, the RPN weights should be adapted during fine-tuning on  $\mathcal{D}_{\text{novel}}$  and the number of proposals passing NMS can be increased.

### B. Modifications of the Feature Pyramid Network

Next to unfreezing the RPN, Sun et al. [102] (FSCE) showed that also fine-tuning the feature pyramid network (FPN) in the second phase improves the performance compared to freezing its weights. They assume that the concepts from the base categories cannot be transferred to novel categories without any fine-tuning.

Wu et al. [111] (MPSR) observed that the scales of the FPN do not compensate for the sparsity of the scales of the few samples of novel categories. Therefore, in a refinement branch, specific data augmentation is applied to solve this issue (see Section VII-C). Wang et al. [110] (CIR) designed a context module to enlarge the receptive field of the FPN, which also addresses the problem of varying scales and, in particular, improves the detection of small objects.

*Takeaway:* Also, FPN weights should be adapted during fine-tuning [102].

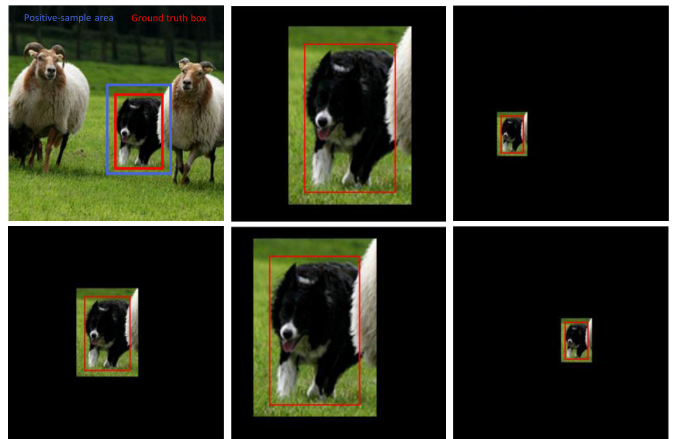


Fig. 10. Augmentation for novel categories regarding scale and translation in FSSP [125].

### C. Increase the Variance of Novel Categories

If training instances for novel categories  $\mathcal{C}_{\text{novel}}$  are limited, also the variance of the data regarding these categories is limited. Therefore, some approaches try to increase the variance of the data for novel categories.

In the refinement branch of MPSR [111], each object is cropped by a square window and resized to various scales. This increases the variance regarding object sizes. This augmentation is also employed in FSOD-UP [113] and CME [46]. A similar approach is taken by Xu et al. [125] (FSSP), where in an auxiliary branch, the objects are augmented regarding scale and translation, as shown in Fig. 10.

Zhang and Wang [105] (Halluc.) introduced a hallucinator network that learns to generate additional training examples for novel categories  $\mathcal{C}_{\text{novel}}$ . To achieve this, the features in the RoI head  $f^{\mathcal{R}}$  of novel category samples are augmented by leveraging the shared within-class feature variation from base categories  $\mathcal{C}_{\text{base}}$ .

Kaul et al. [108] (LVC) showed in their experiments that data augmentation, i.e., color jittering, random cropping, mosaicing, and dropout for the extracted features for each ROI, significantly improves the performance. Sun et al. [102] (FSCE) described the similarity between the augmentation with several random image crops and multiple RoI proposals from the RPN. Thus, increasing the number of proposed RoIs per novel category instance as described in Section VII-A is also increasing the variance of novel categories as it resembles random cropping augmentation. According to [105], increasing the variance of novel categories primarily benefits the extreme few-shot scenario with very few samples  $K$  per novel category.

If additional unlabeled data containing novel categories are available, techniques from semisupervised learning can be applied to increase the number of samples for novel categories. Liu et al. [117] (N-PME) pseudo-labeled the base dataset  $\mathcal{D}_{\text{base}}$  after fine-tuning in order to find additional samples of  $\mathcal{C}_{\text{novel}}$  in  $\mathcal{D}_{\text{base}}$ . The novel samples are then used for an additional fine-tuning phase with more shots by including the samples pseudo-labeled as one of the categories in  $\mathcal{C}_{\text{novel}}$ . Since the bounding boxes for the additional samples are rather imprecise, they are omitted for the regression loss. Kaul et al. [108] (LVC) further improved this attempt by first verifying that the searched novel samples indeed belong to  $\mathcal{C}_{\text{novel}}$  and then correct the inaccurate bounding boxes. For verification, they apply a vision transformer (ViT) [128], which was trained in a self-supervised manner by self-distillation

with no labels (DINO) [129], to get features that can be used in a  $k$ -nearest neighbor classifier to compare to the  $K$  shots of the novel categories. If novel samples can be verified, they are included in  $\mathcal{D}_{\text{novel}}$ ; otherwise, these regions are ignored during the following additional fine-tuning on the extended data. Bounding boxes for verified samples are corrected in the fashion of Cascade R-CNN [130]. Using the high-quality extended data, the detector can be finetuned end-to-end without the need for freezing any components of the detector. While this seems to improve the performance significantly, it should be noted that for the FSOD benchmark datasets Microsoft COCO and PASCAL VOC, searching for novel objects in images of  $\mathcal{D}_{\text{base}}$  is sufficient, but for real-world few-shot applications, such as medical applications or the detection of rare species, additional (unlabeled) data are needed, which could prove problematic.

*Takeaway:* Increasing the variance of training examples from novel categories—e.g., by data augmentation or pseudo-labeling additional data—improves detection accuracy, especially when the number of training examples  $K$  is very low.

#### D. Transfer Knowledge Between Base and Novel Categories

In LSTD [8], using a soft assignment of similar base categories, weights of components for novel categories are initialized by base category weights to transfer base knowledge. Chen et al. [126] (AttFDNet) initialized the parameters of the novel object detector using parameters from the base object detector and an imprinting initialization method [131], [132]. Also, Li et al. [116] (CGDP + FSCN) used imprinting for initialization [131].

By learning and leveraging visual and semantic lingual similarities between the novel and base categories, in the second training phase, Khandelwal et al. [121] (UniT) transferred weights for bounding box regression and classification from base categories to novel categories. Zhu et al. [122] (SRR-FSD) represented each category concept by a semantic word embedding learned from a large corpus of text. The image representations of objects are projected into this embedding space to learn  $\mathcal{C}_{\text{novel}}$  from both the visual information and the semantic relation. Cao et al. [119] (FADI) also incorporated the categories' semantic meaning: After training on  $\mathcal{D}_{\text{base}}$ , they measure the semantic similarity of base and novel categories via WordNet [133]. The authors argue, that in the second fine-tuning phase, associating novel categories to multiple base categories leads to scattered intraclass structures for the novel categories. Thus, each novel category is associated to exactly one base category with the highest similarity. Afterward, each novel category is assigned a pseudo-label of the associated base category. Then, the whole network is frozen—except for the second fully connected layer in RoI Head—and the network is trained such that it learns to align the feature distribution of the novel category to the associated base category. This leads to low intraclass variation of the novel category but inevitably to confusion between  $\mathcal{C}_{\text{base}}$  and  $\mathcal{C}_{\text{novel}}$ . Thus, in a subsequent discrimination step, the classification branches for  $\mathcal{C}_{\text{base}}$  and  $\mathcal{C}_{\text{novel}}$  are disentangled to learn a good discrimination. In [124] (KR-FSOD), a semantic knowledge graph based on word embeddings is used to describe a scene and relations between objects. This helps to improve knowledge propagation between novel and related categories.

*Takeaway:* For initializing the weights of components for each novel category, knowledge from the semantically most similar base category should be transferred [119].

#### E. Keep the Performance on Base Categories

Many approaches suffer from catastrophic forgetting when trained on  $\mathcal{C}_{\text{novel}}$ . Although the model can be trained on  $\mathcal{C}_{\text{base}}$  as well in the fine-tuning phase, the performance still drops compared to before fine-tuning. Therefore, Fan et al. [106] (Retentive R-CNN) proposed to duplicate the RPN and the classification heads for RoI proposal and classification of  $\mathcal{C}_{\text{base}}$  and  $\mathcal{C}_{\text{novel}}$ , respectively. During fine-tuning of the  $\mathcal{C}_{\text{novel}}$  head, a cosine classifier is used to balance the variations in feature norms of  $\mathcal{C}_{\text{base}}$  and  $\mathcal{C}_{\text{novel}}$ . The frozen RPN and RoI head for  $\mathcal{C}_{\text{base}}$  shall keep the performance on the base categories. Feng et al. [103] (BPMCH) combat catastrophic forgetting for base categories  $\mathcal{C}_{\text{base}}$  during the fine-tuning phase mainly by fixing the backbone  $\mathcal{B}_{\text{base}}$  and the classification head for these categories and used an additional backbone  $\mathcal{B}_{\text{novel}}$  as a feature extractor for novel categories  $\mathcal{C}_{\text{novel}}$ .

In MemFRCN [104], additional to the softmax-based classifier in the RoI head, representative feature vectors  $f^{\mathcal{R},c_i}$  for each category  $c_i$  are learned and stored to remember the base categories  $\mathcal{C}_{\text{base}}$  after the RoI head is modified during the fine-tuning phase. During inference, extracted features  $f^{\mathcal{R}}$  can be compared to these category representatives by cosine similarity. This is similar to support vectors in dual-branch meta learning.

Guirguis et al. [123] (CFA) built on the continual learning approaches GEM [134] and A-GEM [135], which observed that catastrophic forgetting occurs when the angle between loss gradient vectors of previous tasks and the gradient update of the current task is obtuse. Therefore, CFA stores  $K$  shots of the base categories in episodic memory, analogous to A-GEM, in order to be able to compute gradients on  $\mathcal{D}_{\text{base}}$ . During the fine-tuning phase, the episodic memory is static, meaning that no further samples are added. The fine-tuning is then conducted as follows. The base category gradient  $g_{\text{base}}$  is calculated on a mini-batch drawn from the episodic memory and the novel category gradient  $g_{\text{novel}}$  is calculated on a mini-batch from  $\mathcal{D}_{\text{novel}}$ . If the angle between  $g_{\text{base}}$  and  $g_{\text{novel}}$  is acute,  $g_{\text{novel}}$  is backpropagated as it is. Otherwise, a new gradient update rule is derived, which averages the base gradients  $g_{\text{base}}$  and novel gradients  $g_{\text{novel}}$ . It also adaptively reweights them in case the novel gradients  $g_{\text{novel}}$  point toward a direction that could lead to forgetting.

*Takeaway:* To prevent catastrophic forgetting and keep the performance on base categories, the angle between gradients of novel and base categories must be considered [123].

#### F. Modify the Training Objective

A modified loss, which updates the training objective, can guide the detector toward focusing on foreground regions or specific aspects, may improve the consistency in multiple branches, and may also help to improve the inner class and interclass variance of features for object classification. Furthermore, restricting the gradient flow in the detector or slightly modifying the training scheme can improve the training of the different components of the detector.

1) *Additional Loss Terms:* Chen et al. [8] (LSTD) used additional background-depression and transfer-knowledge regularization terms in the loss function to help the detector focus on target objects and incorporate source-domain knowledge. Li et al. [116] (CGDP + FSCN) identified unlabeled instances of novel categories  $\mathcal{C}_{\text{novel}}$  in the base dataset  $\mathcal{D}_{\text{base}}$  as problematic. They introduce an additional semisupervised loss term to also utilize these unlabeled instances.

Chen et al. [126] (AttFDNet) proposed two loss terms to maximize the cosine similarity between instances of the same category and to tackle the problem of unlabeled instances in the dataset. Cao et al. [119] (FADI) introduced an additional set-specialized margin loss to enlarge interclass separability. In contrast to previous margin losses such as ArcFace [136], they use scaling factors for different margins, where the scaling factor for  $\mathcal{C}_{\text{novel}}$  is higher than for  $\mathcal{D}_{\text{base}}$ , as novel categories are much more challenging. Liu et al. [117] (N-PME) used a margin loss to exploit error-prone pseudo-labels by evaluating the uncertainty scores of both correct and incorrect pseudo-labels for novel categories on additional data.

2) *Loss for Auxiliary Branches*: Similar to the meta-learning approach TIP [44], Wu et al. [113] (FSOD-UP) used a consistency loss to force features of two branches to be similar. They apply the KL-Divergence loss between these features. The context module of CIR [110] is trained in a supervised manner by an auxiliary classification branch that predicts a binary foreground-background segmentation map. The two branches of MPSR [111] are loosely coupled via shared weights and contributions of both branches to the loss function. CME [46] builds on top of MPSR but introduces an additional adversarial training as we described in Section V-E. Also, Xu et al. [125] (FSSP) introduced an auxiliary branch. It includes a full replication of the detection network used for data augmentation. A modified classification loss combines the decisions in the original branch and this auxiliary branch that processes only one object with most of the background removed.

Sun et al. [102] (FSCE) introduced a new branch in the RoI head. In addition to the standard RoI head, they apply a single fully connected layer as contrastive branch to be able to measure similarity scores between learned object proposal representations. On the contrastive branch, they use a contrastive proposal encoding loss for training that enables increasing the cosine similarity of representation from the same category and reduce the similarity of proposals from different categories. Lu et al. [115] (DMNet) followed a similar approach. They use an auxiliary classification branch in which they compare extracted features to representatives for each category by the Euclidean distance. The feature embedding and the category representatives are learned by triplet-loss-based metric learning.

3) *Modified Gradient Flow*: Qiao et al. [114] (DeFRCN) additionally want to update the backbone in both training phases, but they identified contradictions in training as problematic. The goals of RPN and ROI head are contrary since the RPN tries to learn class-agnostic region proposals, whereas the ROI head tries to distinguish categories. Their extensive experiments showed that it is key to stop the gradient flow from the RPN to the backbone and scale the gradient from the ROI head to the backbone. During training on  $\mathcal{D}_{\text{base}}$  in the first phase, they scale the gradient from the ROI head by 0.75 so that the backbone learns a little less than the rest of the detector. During training on  $\mathcal{D}_{\text{base}} \cup \mathcal{D}_{\text{novel}}$  in the second phase, it proved necessary to scale the gradient by 0.01, which is in the direction of freezing the backbone. Stopping the gradient from the RPN and scaling the gradient from the ROI head significantly boosts the performance, especially in the second phase. The authors observed that this gradient scaling also benefits Faster R-CNN as a generic object detector when trained with sufficient data as well.

Guirguis et al. [123] (CFA) derived a new gradient update rule that takes the angle between gradients for samples of

$\mathcal{D}_{\text{base}}$  and samples of  $\mathcal{D}_{\text{novel}}$  into account in order to combat catastrophic forgetting for base categories  $\mathcal{C}_{\text{base}}$  during the fine-tuning phase, as already described in Section VII-E. While this gradient update rule primarily intends to preserve the performance on base categories  $\mathcal{C}_{\text{base}}$ , it also has a positive impact on the performance regarding novel categories  $\mathcal{C}_{\text{novel}}$ .

4) *Modified Training Scheme*: Inspired by infants beginning to learn from a single observation, in [109] (FORD + BL), it is shown that instead of fine-tuning with  $K$  shots immediately, the performance can be improved by first fine-tuning with a single shot per category and only then fine-tuning with all  $K$  shots. Wu et al. [118] (TD-Sampler) introduced a batch sampling strategy for the fine-tuning phase that enables to use all samples of  $\mathcal{D}_{\text{base}}$  instead of  $K$  shots per base category  $\mathcal{C}_{\text{base}}$  and to use more samples of novel categories  $\mathcal{C}_{\text{novel}}$  per training batch. This is achieved by selecting batches that contain a large number of novel category samples and are unlikely to significantly change the detector activation pattern, judged by an estimated training difficulty (TD).

*Takeaway*: The loss should be modified regarding optimized gradient flow and interclass separability. In an auxiliary branch, a contrastive loss can help to improve the discriminative power of features, like in two-branch meta learning.

### G. Use Attention

Attention blocks help to enhance features. In this sense, Wu et al. [113] (FSOD-UP) used soft attention between learned prototypes (see Section VII-H) and RPN outputs to enhance features in an extra branch. Yang et al. [120] (CoTrans) used the affinity between an anchor box and its contextual field as a relational attention to integrate contexts into the representation of the anchor box. Xu et al. [125] (FSSP) first processed the image by a self-attention module and then processed the attention-enriched input by a one-stage detector. Therefore, the detector can focus on important parts of the input image. Chen et al. [126] (AttFDNet) combined top-down and bottom-up attention. Top-down attention is learned in supervised fashion in a simplified nonlocal block and a squeeze-and-excitation block. Bottom-up attention is computed by a saliency prediction model (boolean map based saliency (BMS) [137] or saliency attentive model (SAM) [138]).

### H. Modify Architecture

1) *Architectures Based on Faster R-CNN*: The majority of transfer learning approaches are based on the Faster R-CNN detector, as shown in Fig. 8. Benchmark results confirm the superiority of this two-stage detector for transfer learning approaches. Only few approaches deviate from this architecture.

Li et al. [116] (CGDP + FSCN) observed that the performance degradation for novel categories in Faster R-CNN is mainly caused by false positive classifications, i.e., by category confusion. Therefore, they refine the classification in an additional discriminability enhancement branch, which is trained with misclassified false positive samples. It directly processes the cropped image of the object to be classified. Then, the classification result is fused with the one of the original Faster R-CNN branch. Qiao et al. [114] (DeFRCN) also observed many low classification scores for novel categories. Similar to Li et al., they conclude that contrary requirements of translation invariant features for classification and translation

covariant features for localization are problematic. To tackle this issue, they propose a prototypical calibration block, which performs score refinement to eliminate high-scored false positive classifications.

Wu et al. [111] (MPSR) used an auxiliary refinement branch for data augmentation during training that is excluded during inference. SVD [112] builds upon MPSR [111]. With a singular value decomposition (SVD), they decompose the backbone features  $f^B$  into eigenvectors with their relevance quantified by the corresponding singular values. The eigenvectors corresponding to the largest singular values are incorporated for localization since they are able to suppress certain variations. In contrast, the eigenvectors corresponding to the smaller singular values are incorporated for category discrimination since they encode category-related information. This discrimination space is further refined by utilizing dictionary learning [139] to facilitate classification. Wu et al. [113] (FSOD-UP) adapted the few-shot learning idea of prototypes [11], [140], [141] that reflect category information. In contrast to category-specific prototypes in dual-branch meta learning, they learn universal prototypes based on all categories in an extra branch that processes backbone features. These universal prototypes are invariant under different visual changes and, thus, enhance the original features from the backbone. After processing original and enhanced features in the RPN, this processing in an auxiliary branch is repeated for RPN features to compute the input for the ROI head.

2) *Incorporating One-Stage Detectors*: One of the earliest FSOD approaches, LSDT [8], combines bounding box regression following the SSD [100] approach and Faster R-CNN [70] concepts for object classification.

Yang et al. [120] (CoTrans) used SSD [100] as a one-stage detector. They argue that the multiscale spatial receptive fields in this architecture provide rich contexts, which are important for knowledge transfer. Chen et al. [126] (AttFDNet) also used the SSD detector, but added two attention branches, to help the detector to focus on the important parts of the image and six prediction heads that predict bounding boxes and categories for objects at different scales.

Lu et al. [115] (DMNet) proposed a one-stage detector that follows the design principles of SSD and YOLO, but uses two decoupled branches for localization and classification. It is argued that this decoupling facilitates adaptation with only a few examples.

Xu et al. [125] (FSSP) show, how a fast one-stage detector, namely YOLOv3 [142], can be made competitive with the slower two-stage detector Faster R-CNN in the vanilla setup as described in [101]. This is possible only by putting in a lot of effort, namely incorporating a self-attention module, using an additional auxiliary branch that contains a full replication of the detection network, augmenting the input data of the auxiliary branch, and applying an additional loss. However, due to these modifications, the fast one-stage detector of Xu et al. [125] is especially performing better than TFA [101] for the extremely low-shot scenario.

*Takeaway*: In detectors based on Faster R-CNN, a score refinement can help to reduce false positive classifications [114]. Single-stage detectors can profit from an auxiliary branch in order to enable data augmentation [125].

### Summary of Best Performing Transfer Learning Approaches

In the following, we summarize selected transfer learning approaches with distinct concepts regarding the categories

described above that perform best on FSOD benchmark datasets (see Section IX).

In *DeFRCN* [114], it was found that the class-agnostic localization task in the RPN and the class-distinguishing task of the ROI head are contrary. Therefore, it is key to stop the gradient flow from the RPN to the backbone and scale the gradient from the ROI head to the backbone. Then, it is possible to train all components of the detector, including the backbone, in both training phases, which significantly boosts the performance, especially in the fine-tuning phase. In addition, a prototypical calibration block performs score refinement in the ROI head to eliminate high-scored false positive classifications, which are a result of contrary requirements of translation invariant features for classification and translation covariant features for localization.

*CFA* [123] can be applied on top of *DeFRCN* to tackle catastrophic forgetting of base categories during fine-tuning, which occurs when angles between base category gradients and novel category gradients are obtuse. Therefore, *CFA* stores  $K$  shots of the base categories in episodic memory for computing gradients on  $\mathcal{D}_{\text{base}}$  during fine-tuning. If the angle between base and novel category gradients is obtuse, both gradients are averaged and adaptively reweighted; otherwise, the novel category gradient can be backpropagated without the risk of catastrophic forgetting. This gradient update rule also benefits the performance regarding novel categories.

Before fine-tuning, *FADI* [119] associates each novel category to exactly one base category by measuring their semantic similarity via WordNet. Then, the network is trained to align the feature distribution of the novel category to the associated base category. This leads to low intra-class variation of the novel category but inevitably to confusion between  $\mathcal{C}_{\text{base}}$  and  $\mathcal{C}_{\text{novel}}$ . Thus, in a subsequent discrimination step, the classification branches for  $\mathcal{C}_{\text{base}}$  and  $\mathcal{C}_{\text{novel}}$  are disentangled to learn good discrimination. In addition, a set-specialized margin loss is employed to enlarge interclass separability.

If additional unlabeled data containing novel categories are available, techniques from semisupervised learning can be applied. On FSOD benchmark datasets  $\mathcal{D}_{\text{base}}$  can be used for this purpose. *LVC* [108] pseudo-labels these data with the detector that was finetuned on  $\mathcal{D}_{\text{novel}}$  in the second training phase. First, the searched novel samples are verified to belong to  $\mathcal{C}_{\text{novel}}$  by using features of a ViT to compare with the shots of  $\mathcal{D}_{\text{novel}}$  in the nearest neighbor fashion. Then, the inaccurate bounding boxes of verified samples are corrected similar to Cascade R-CNN. Using the high-quality extended data for novel categories, all components of the detector can be trained in an additional fine-tuning phase.

### Conclusion on Transfer Learning Approaches

Transfer learning approaches have a much simpler training pipeline, as they do not require complex episodic training as in meta learning. By incorporating specific techniques to be able to finetune as much components of the detector as possible—e.g., modifying the training objective or transferring knowledge between base and novel categories—transfer learning approaches are able to reach the state-of-the-art performance.

## VIII. COMPARISON BETWEEN META LEARNING AND TRANSFER LEARNING

After elaborating on different approaches for meta learning as well as transfer learning, we now want to draw a comparison. Since single-branch meta learning is less explored in recent works and also falls behind in terms of performance,

TABLE I  
COMPARISON BETWEEN DUAL-BRANCH META LEARNING AND  
TRANSFER LEARNING

• Dual-Branch Meta Learning	▷ Transfer Learning
MAIN FOCUS	
How to aggregate information of the support and query branch	How to freeze less components of the detector without performance decline
TRAINING SCHEME	
Complex episodic training scheme	Simple training protocol
ADAPTATION TO $C_{novel}$	
Fast adaptation to $C_{novel}$ (some approaches can even be applied to $C_{novel}$ without finetuning, e.g., SQMG [41])	Requires finetuning on $C_{novel}$
COMBAT TOO FEW PROPOSALS FOR $C_{novel}$	
Easily handled by aggregation before RPN (e.g., AttentionRPN [55]) or proposal-free detectors (e.g., Meta-DETR [64])	Modifications of non-maximum suppression in RPN necessary (harder, e.g., FSCE [102])
REDUCE CLASSIFICATION ERRORS	
Main issue of FSOD is classification of novel categories, not localization To better differentiate between several categories, ideas from metric learning such as similarity metrics, as well as contrastive training can be employed	
GUIDED GRADIENT FLOW	
Harder to achieve in dual branch architecture, not explored yet	Simple to achieve and huge performance gain (see DeFRCN [114])
DATA AUGMENTATION	
Little explored so far, but great potential due to support branch that enables simple application of augmentation	Necessity to increase variance of novel samples to perform well, often needs auxiliary branch for application of complex augmentation
USE OF ATTENTION	
A lot, mainly for aggregation (e.g., APSP [54]) and lately in backbone (e.g., FCT [66])	Only few attempts with mediocre success, out of fashion in most recent attempts, except for attention-based backbone
RELATIONS BETWEEN CATEGORIES	
Can be incorporated to improve the performance to some degree	
SELF- AND SEMI-SUPERVISED LEARNING	
Only one attempt so far (SPCD [47]), improvements in performance observable	Can be applied if additional (unlabeled) data is available, will improve performance a lot (e.g., LVC [108])
KEEP PERFORMANCE ON $C_{base}$	
Simpler due to support features, but less explored in recent approaches	Harder to achieve, often only by duplication of detector components or modified gradient update, but successful in recent approaches (e.g., CFA [123])
VERY LOW SHOT PERFORMANCE	
No approach stands out, performance is relatively poor (• Meta-DETR [64] and ▷ DeFRCN [114] appear most promising)	
ONE-STAGE DETECTORS	
Can be applied, but performance is worse compared to Faster R-CNN, Most approaches build on the latter	

we discard it in the comparison. In Table I, we compare dual-branch meta learning and transfer learning according to several important aspects. Both seem promising for future work and either could benefit by also incorporating ideas from the other training scheme.

## IX. DATASETS, EVALUATION PROTOCOLS, AND BENCHMARK RESULTS

Evaluation of few-shot object detectors requires tailored datasets that distinguish between base and novel categories. Therefore, most approaches use specific splits of the common object detection datasets PASCAL VOC [143] and Microsoft COCO [144]. Only rarely, other datasets, such as FSOD, ImageNet-LOC, or LVIS, are applied. Generally, few-shot object detectors are evaluated in the  $K$ -shot- $N$ -way manner, i.e.,  $\mathcal{D}_{novel}$  consists of  $K$  labeled examples for  $N$  novel categories.

### A. PASCAL VOC Dataset

The PASCAL VOC dataset [143] contains annotations for 20 categories. Commonly, the combination of VOC07 + 12 trainval sets is used for training, and VOC07 test set is used for testing. For evaluating few-shot object detectors, most often three category splits are used, each with 15 base categories and five novel categories ( $N = 5$ ).

- 1) *Set 1*:  $C_{novel} = \{\text{bird, bus, cow, motorbike, sofa}\}$ .
- 2) *Set 2*:  $C_{novel} = \{\text{aeroplane, bottle, cow, horse, sofa}\}$ .
- 3) *Set 3*:  $C_{novel} = \{\text{boat, cat, motorbike, sheep, sofa}\}$ .

The number of shots  $K$  for novel categories is set to 1, 2, 3, 5, and 10. As an evaluation metric, the mean average precision at an intersection over union (IoU) threshold of 0.5 is used (AP<sub>50</sub>). Unfortunately, the specific  $K$ -shot object instances are not fixed, which leads to varying instances used in different approaches. As stated by Wang et al. [101], this high variance in training samples makes it difficult to compare approaches against each other, as approach-based performance differences may be insignificant compared to differences based on different instances. Therefore, Wang et al. [101] proposed a revised evaluation protocol, where results are averaged over 30 runs with different random samples of training shots. Moreover, they also report the performance on base categories since ignoring the performance for base categories might hide a potential performance drop and is, therefore, not suitable for evaluating the overall performance of a model. Currently, approaches focusing on this topic also report results for generalized FSOD (G-FSOD) performance, which refers to the mean over novel and base categories.

In Table II, we list benchmark results of the described approaches for the PASCAL VOC dataset. We split the table according to whether the results are given for a single run or averaged over multiple runs as proposed by Wang et al. [101]. Approaches that give results for both evaluation protocols are marked with “←” to act as anchors for comparison. The performance gap between the two evaluation protocols is not negligible and shows that the averaged results are more reliable. Furthermore, we also denote whether fine-tuning on  $\mathcal{D}_{novel}$  is required or if the results are achieved by simply meta testing. Each approach is characterized by a small symbol at the front. In general, both transfer learning and dual-branch meta learning approaches can achieve similar results. The main characteristics of the best performing approaches are summarized at the end of Section V for dual-branch meta learning and Section VI-D for transfer learning.

Although it is very commonly used, according to Michaelis et al. [58], the PASCAL VOC dataset is too easy. With a dual-branch meta learning approach and uninformative all-black support images, they are still able to locate the novel objects and reach a mAP<sub>50</sub> of 33.2. However, we want to highlight that in general, objects do not simply need to be located but also classified, i.e., detectors need to also determine which category is present in the image.

### B. Microsoft COCO Dataset

In comparison to PASCAL VOC, the Microsoft COCO dataset [144] is more challenging and contains annotations for 80 categories, including the 20 VOC categories. For FSOD, most often, the 20 VOC categories are used as novel categories, leaving the remaining 60 categories as base categories. Typically, the number of shots  $K$  is set to 10 and 30. However,

TABLE II

AP<sub>50</sub> PUBLISHED RESULTS ON THE PASCAL VOC BENCHMARK FOR ALL THREE SETS AND DIFFERENT NUMBER OF SHOTS  $K$ . WE SORT THE APPROACHES BY THE MEAN OVER ALL NOVEL SETS AND SHOTS. —: NO RESULT REPORTED IN PAPER. \*: RESULTS ONLY REPORTED FOR DIFFERENT SHOTS OR SETS, AND THEREFORE, THESE RESULTS ARE NOT INCLUDED HERE. \*: DEVIATING EVALUATION PROTOCOL PREVENTING FAIR COMPARISON AS DESCRIBED IN SECTION IX-C. †: DUAL-BRANCH META LEARNING. ●: SINGLE-BRANCH META LEARNING. ▷: TRANSFER LEARNING

Approach	Publication	Detector	Novel		Novel Set 1					Novel Set 2					Novel Set 3					Base Set 1	
			mean	K=1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	3	10	
<b>Results over a single run:</b>																					
no finetuning	● RepMet [95]	CVPR 2019	Faster R-CNN R-101	30.8	26.1	32.9	34.4	38.6	41.3	17.2	22.1	23.4	28.3	35.8	27.5	31.1	31.5	34.4	37.2	—	—
	† Meta Faster-RCNN [59]	AAAI 2022	Faster R-CNN R-101	36.9	40.2	30.5	33.3	42.3	46.9	26.8	32.0	39.0	37.7	37.4	34.0	32.5	34.4	42.7	44.3	—	—
	● NP-RepMet [95]	NeurIPS 2020	Faster R-CNN R-101	42.6	37.8	40.3	41.7	47.3	49.4	41.6	43.0	43.4	47.4	49.1	33.3	38.0	39.8	41.5	44.8	66.6	68.3
	† SQMG [41]	CVPR 2021	Faster R-CNN R-50	47.7	46.8	49.2	50.2	52.0	52.4	39.4	43.1	43.6	44.1	45.7	44.1	49.8	50.5	52.3	52.8	—	—
	† SQMG [41]	CVPR 2021	Faster R-CNN R-101	49.8	48.6	51.1	52.0	53.7	54.3	41.6	45.4	45.8	46.3	48.0	46.1	51.7	52.6	54.1	55.0	—	—
	▷ AttFDNet (BU+TD) [126]	arXiv 2020	SSD VGG-16	26.9	29.6	34.9	35.1	—	—	16.0	20.7	22.1	—	—	22.6	29.1	32.0	—	—	*	—
	● PNPDet [96]	WACV 2021	CenterNet	27.6	18.2	—	27.3	—	41.0	16.6	—	26.5	—	36.4	18.9	—	27.2	—	36.2	75.5	75.5
	† MetaYOLO [45]	ICCV 2019	YOLOv2	28.4	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9	64.8	69.7
	▷ CME (MetaYOLO) [46]	CVPR 2021	YOLOv2	31.1	17.8	26.1	31.5	44.8	47.5	12.7	17.4	27.1	37.7	40.0	15.7	27.4	30.7	44.9	48.8	—	—
	† Meta R-CNN [9]	ICCV 2019	Faster R-CNN R-101	31.1	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1	64.8	67.9
▷ MemFRCN (DeFRCN) [104]	TFECCS 2022	Faster R-CNN R-101	35.2	36.4	37.4	40.6	45.5	46.6	18.0	26.8	32.1	36.3	42.4	30.3	32.3	37.3	37.8	38.5	—	—	
† APSP (AttentionRPN) [54]	WACV 2022	Faster R-CNN R-101	38.0	31.1	36.1	39.2	50.7	59.4	22.9	29.4	32.1	35.4	32.7	24.3	28.6	35.0	50.0	53.6	—	—	
† FSD-KT [63]	SMC 2020	Faster R-CNN R-101	38.8	27.8	41.4	46.2	55.2	56.8	19.8	27.9	38.7	38.9	41.5	29.5	30.6	38.6	43.8	45.7	69.6	68.1	
▷ TFA w/cos [101] —	ICML 2020	Faster R-CNN R-101	39.9	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8	79.1	78.4	
▷ Halluc. (TFA w/cos) [105]	CVPR 2021	Faster R-CNN R-101	40.6	45.1	44.0	44.7	55.0	55.9	23.2	27.5	35.1	34.9	39.0	30.5	35.1	41.4	49.0	49.3	—	—	
▷ Retentive R-CNN [106]	CVPR 2021	Faster R-CNN R-101	41.1	42.4	45.8	45.9	53.7	56.1	21.7	27.8	35.2	37.0	40.3	30.2	37.6	43.0	49.7	50.1	*	*	
▷ CoRPN w/cos [107]	arXiv 2020	Faster R-CNN R-101	42.2	44.4	38.5	46.4	54.1	55.7	25.7	29.5	37.3	36.2	41.3	35.8	41.8	44.6	51.6	49.6	—	—	
▷ SRR-FSD [122]	CVPR 2021	Faster R-CNN R-101	43.0	46.3	51.1	52.6	56.2	57.3	31.0	29.9	34.7	37.3	41.7	39.2	40.5	39.7	42.2	45.2	—	—	
▷ Halluc. (CoRPN w/cos) [105]	CVPR 2021	Faster R-CNN R-101	43.2	47.0	44.9	46.5	54.7	54.7	26.3	31.8	37.4	37.4	41.2	40.4	42.1	43.3	51.4	49.6	—	—	
▷ CGDP+PFCN [116]	CVPR 2021	Faster R-CNN R-50	43.8	40.7	45.1	46.5	57.4	62.4	27.3	31.4	40.8	42.7	46.3	31.2	36.4	43.7	50.1	55.6	—	—	
▷ CME (MPSR) [46]	CVPR 2021	Faster R-CNN R-101	44.4	41.5	47.5	50.4	58.2	60.9	27.2	30.2	41.4	42.5	46.8	34.3	39.6	45.1	48.3	51.5	—	—	
▷ TD-Sampler [118]	ICCCBDA 2022	Faster R-CNN R-101	44.8	37.1	47.8	50.5	56.2	63.1	26.3	35.0	42.9	46.8	52.0	25.5	36.8	43.5	50.9	58.2	—	—	
▷ MPSR [111]	ECCV 2020	Faster R-CNN R-101	44.8	41.7	—	51.4	55.2	61.8	24.4	—	39.2	39.9	47.8	35.6	—	42.3	48.0	49.7	67.8	71.8	
▷ SRR-FSD [122]	CVPR 2021	Faster R-CNN R-101	44.8	47.8	50.5	51.3	55.2	56.8	32.5	35.3	39.1	40.8	43.8	40.1	41.5	44.3	46.9	46.4	78.2	78.2	
▷ SVD (MPSR) [112]	NeurIPS 2021	Faster R-CNN R-101	44.9	41.5	47.4	51.5	57.7	61.2	29.4	29.6	39.8	41.2	51.5	36.0	39.9	43.4	50.4	51.3	69.4	—	
▷ FSD-UP [113]	ICCV 2021	Faster R-CNN R-101	45.0	43.8	47.8	50.3	55.4	61.7	31.2	30.5	41.2	42.2	48.3	35.5	39.7	43.9	50.6	53.5	*	*	
▷ FSSP [125]	Access 2021	YOLOv3-SPP	45.1	41.6	—	49.1	54.2	56.5	30.5	—	39.5	41.4	45.1	36.7	—	45.3	49.4	51.3	73.5	74.2	
† PNSD * [68]	ACCV 2020	Faster R-CNN R-50	45.4	40.9	—	50.4	56.5	59.8	30.2	—	41.8	46.4	48.3	34.8	—	40.6	46.9	48.6	—	—	
▷ FORD+BL [109]	IMAVIS 2022	Faster R-CNN R-101	45.4	46.3	54.2	49.9	56.3	61.8	19.0	30.8	38.4	39.3	47.3	36.4	46.5	45.4	53.2	55.8	78.7	79.6	
▷ DMNet [115]	TCyb. 2022	DMNet R-101	46.1	39.0	48.9	50.7	58.6	62.5	31.2	32.4	40.3	47.6	52.0	41.7	41.8	42.7	50.3	52.1	—	—	
▷ FSCE [102] —	CVPR 2021	Faster R-CNN R-101	46.6	44.2	43.8	51.4	61.9	63.4	27.3	29.5	43.5	44.2	50.2	37.2	41.9	47.5	54.6	58.5	74.1	—	
▷ SVD (FSCE) [112]	NeurIPS 2021	Faster R-CNN R-101	46.7	46.1	43.5	48.9	60.0	61.7	25.6	29.9	44.8	47.5	48.2	39.5	45.4	48.9	53.9	56.9	74.8	—	
▷ FADI [119]	NeurIPS 2021	Faster R-CNN R-101	49.2	50.3	54.8	54.2	59.3	63.2	30.6	35.0	40.3	42.8	48.0	45.7	49.7	49.1	55.0	59.6	78.9	—	
† KFSOD * [69]	CVPR 2022	Faster R-CNN R-50	49.2	44.6	—	54.4	60.9	65.8	37.8	—	43.1	48.1	50.4	34.8	—	44.1	52.7	53.9	*	*	
● Meta-RetinaNet [98]	BMVC 2020	RetinaNet ResNet18	49.9	38.3	51.8	59.3	65.3	71.5	28.4	36.8	42.4	45.5	50.9	35.9	48.1	53.2	58.0	63.6	—	—	
† Meta-DETR [64] —	TPAMI 2022	Def. DETR R-101	50.2	40.6	51.4	58.0	59.2	63.6	37.0	36.6	43.7	49.1	54.6	41.6	45.9	52.7	58.9	60.6	—	—	
† Meta Faster-RCNN [59]	AAAI 2022	Faster R-CNN R-101	50.5	43.0	54.6	60.6	66.1	65.4	27.7	35.5	46.1	47.8	51.4	40.6	46.4	53.4	59.9	58.6	—	—	
† FCT [66] —	CVPR 2022	Faster R-CNN PVTv2-B2-Li	50.9	49.9	57.1	57.9	63.2	67.1	27.6	34.5	43.7	49.2	51.2	39.5	54.7	52.3	57.0	58.7	—	—	
▷ LVC [108]	CVPR 2022	Faster R-CNN R-101	52.3	54.5	53.2	58.8	63.2	65.7	32.8	29.2	50.7	49.8	50.6	48.4	52.7	55.0	59.6	59.6	—	—	
▷ CFA-DeFRCN [123]	CVPR 2022	Faster R-CNN R-101	57.3	58.2	63.3	65.8	68.9	67.1	37.1	45.5	51.3	55.2	53.8	54.7	57.8	56.9	60.0	63.3	—	—	
▷ UniT * [121]	CVPR 2021	Faster R-CNN R-101	67.8	75.7	75.8	75.9	76.1	76.7	57.2	57.4	57.9	58.2	63.0	67.6	68.1	68.2	68.6	70.0	77.8	77.7	
<b>Results averaged over multiple random runs:</b>																					
no ft.	† QA-FewDet [61]	ICCV 2021	Faster R-CNN R-101	37.0	41.0	33.2	35.3	47.5	52.0	23.5	29.4	37.9	35.9	37.1	33.2	29.4	37.6	39.8	41.5	—	—
	† SPCD [47]	ICIAIP 2022	Faster R-CNN R-50	39.0	46.0	37.0	45.3	51.4	55.0	29.6	26.7	37.0	30.2	34.6	41.0	32.4	30.7	44.0	43.4	—	—
	† MM-FSOD [50]	arXiv 2020	Faster R-CNN R-34	47.6	50.0	—	55.9	57.9	60.9	37.3	—	45.7	46.5	48.2	35.6	—	43.3	44.1	45.4	—	—
with finetuning	● MetaDet [97]	ICCV 2019	Faster R-CNN VGG-16	31.0	18.9	20.6	30.2	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1	—	—
	▷ TFA w/cos [101] —	ICML 2020	Faster R-CNN R-101	34.7	25.3	36.4	42.1	47.9	52.8	18.3	27.5	30.9	34.1	39.5	17.9	27.2	34.3	40.8	45.6	77.3	77.5
	† FsDetView [43]	ECCV 2020	Faster R-CNN R-101	36.7	24.2	35.3	42.2	49.1	57.4	21.6	24.6	31.9	37.0	45.7	21.2	30.0	37.2	43.8	49.6	—	—
	† APSP (FsDetView) [54]	WACV 2022	Faster R-CNN R-101	38.3	24.3	36.5	44.9	52.0	59.2	20.5	27.5	33.1	40.9	47.1	22.4	33.0	37.8	43.9	51.5	—	—
	† TIP [44]	CVPR 2021	Faster R-CNN R-101	38.5	27.7	36.5	43.3	50.2	59.6	22.7	30.1	33.8	40.9	46.9	21.7	30.6	38.1	44.5	50.9	—	—
	† DCNet [62]	CVPR 2021	Faster R-CNN R-101	39.2	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	—	—
	† ARRM [49]	El. Imag. 2022	Faster R-CNN R-101	39.9	37.0	39.8	44.4	48.2	55.4	24.8	25.4	34.3	38.7	45.9	33.3	35.9	41.1	43.8	50.9	*	*
	▷ FSCE [102] —	CVPR 2021	Faster R-CNN R-101	41.2	32.9	44.0	46.8	52.9	59.7	23.7	30.6	38.4	43.0	48.5	22.6	33.4	39.5	47.3	54.0	74.1	—
	† CAReD [52]	Displays 2022	Faster R-CNN R-50	41.8	36.5	45.2	47														

TABLE III

BENCHMARK RESULTS FOR THE MICROSOFT COCO DATASET SORTED BY NOVEL AP<sub>50:95</sub> FOR TEN-SHOT. —: NO RESULT REPORTED IN PAPER.

\*: DEVIATING EVALUATION PROTOCOL PREVENTING FAIR COMPARISON AS DESCRIBED IN SECTION IX-C. †: DUAL-BRANCH META LEARNING.

●: SINGLE-BRANCH META LEARNING. ▷: TRANSFER LEARNING

Approach	Publication	Detector	K=1				K=10				K=30				
			Novel AP			Base AP	Novel AP			Base AP	Novel AP			Base AP	
			50-95	50	75	50-95	50-95	50	75	50-95	50-95	50	75	50-95	
<b>Results over a single run:</b>															
no finetuning	† ONCE [65]	CVPR 2020	CentreNet R-50	—	—	—	—	5.1	—	—	22.9	—	—	—	—
	† Meta Faster-RCNN [59]	AAAI 2022	Faster R-CNN R-101	5.0	10.2	4.6	—	9.7	18.5	9.0	—	10.7	19.6	10.6	—
	† AttentionRPN [55]	CVPR 2020	Faster R-CNN R-50	—	—	—	—	11.1	—	—	—	—	—	—	—
	† DAnA * [53]	TMM 2021	Faster R-CNN R-50	11.9	25.6	10.4	27.8	—	—	—	—	—	—	—	—
	† SQMG [41]	CVPR 2021	Faster R-CNN R-101	—	—	—	—	13.9	29.5	11.7	—	—	—	—	—
with finetuning	● PNPDet [96]	WACV 2021	CenterNet DLA-34 Def.	—	—	—	—	5.5	—	—	25.8	—	—	—	—
	† MetaYOLO [45]	ICCV 2019	YOLOv2	—	—	—	—	5.6	12.3	4.6	—	9.1	19.0	7.6	—
	▷ BMCH [103]	PRL 2022	FCOS R-50	2.4	—	—	29.4	7.0	—	—	35.6	—	—	—	—
	† Meta R-CNN [9]	ICCV 2019	Faster R-CNN R-101	—	—	—	—	8.7	19.1	6.6	—	12.4	25.3	10.8	—
	† Meta-RetinaNet [98]	BMVC 2020	RetinaNet R-18	—	—	—	—	9.7	19.9	7.7	—	13.1	26.7	11.2	—
	▷ MPSR [111]	ECCV 2020	Faster R-CNN R-101	—	—	—	—	9.8	17.9	9.7	—	14.1	25.4	14.2	—
	▷ FSSP [125]	Access 2021	YOLOv3-SPP	—	—	—	—	9.9	20.4	9.6	—	14.2	25.0	13.9	—
	▷ TFA w/cos [101] ←	ICML 2020	Faster R-CNN R-101	1.9	3.8	1.7	—	10.0	—	9.3	—	13.7	—	13.4	—
	▷ Halluc. (TFA w/cos) [105]	CVPR 2021	Faster R-CNN R-101	3.8	6.5	4.3	31.5	—	—	—	—	—	—	—	—
	▷ DMNet [115]	TCyb. 2022	DMNet R-101	—	—	—	—	10.0	17.4	10.4	—	17.1	29.7	17.7	—
	▷ KR-FSD [124]	Electr. 2022	Faster R-CNN R-101	—	—	—	—	10.2	21.5	8.7	—	14.1	28.6	13.2	—
	▷ Retentive R-CNN [106] ←	CVPR 2021	Faster R-CNN R-101	—	—	—	—	10.5	—	—	39.2	13.8	—	—	39.3
	▷ CoRPN w/ cos [107]	arXiv 2020	Faster R-CNN R-101	4.1	7.2	4.4	34.1	10.6	19.9	10.1	34.6	13.9	25.1	13.9	35.8
	▷ Halluc. (CoRPN w/cos) [105]	CVPR 2021	Faster R-CNN R-101	4.4	7.5	4.9	32.3	—	—	—	—	—	—	—	—
	▷ N-PME [117]	ICASSP 2022	Faster R-CNN R-101	—	—	—	—	10.6	21.1	9.4	—	14.1	26.5	13.6	—
	▷ FSOD-UP [113]	ICCV 2021	Faster R-CNN R-101	—	—	—	—	11.0	—	10.7	—	15.6	—	15.7	—
	▷ SVD (MPSR) [112]	NeurIPS 2021	Faster R-CNN R-101	—	—	—	—	11.0	—	10.6	—	16.2	—	15.9	—
	▷ FORD+BL [109]	IMAVIS 2022	Faster R-CNN R-101	3.6	7.1	3.5	—	11.2	22.5	10.2	—	14.8	28.9	13.9	—
	▷ SRR-FSD [122]	CVPR 2021	Faster R-CNN R-101	—	—	—	—	11.3	23.0	9.8	—	14.7	29.2	13.5	—
	▷ CGDP+FSCN [116]	CVPR 2021	Faster R-CNN R-50	—	—	—	—	11.3	20.3	—	—	15.1	29.4	—	—
	▷ FSCE [102] ←	CVPR 2021	Faster R-CNN R-101	—	—	—	—	11.9	—	10.5	—	16.4	—	16.2	—
	▷ SVD (FSCE) [112]	NeurIPS 2021	Faster R-CNN R-101	—	—	—	—	12.0	—	10.4	—	16.0	—	15.3	—
	▷ FADI [119]	NeurIPS 2021	Faster R-CNN R-101	5.7	10.4	6.0	—	12.2	22.7	11.9	—	16.1	29.1	15.8	—
	† Meta Faster-RCNN [59]	AAAI 2022	Faster R-CNN R-101	5.1	10.7	4.3	—	12.7	25.7	10.8	—	15.9	31.9	14.7	—
	† AtFDNet (BU+TD) [126]	arXiv 2020	SSD VGG-16	—	—	—	—	12.9	19.5	13.9	—	16.3	24.6	17.3	—
	† APSP (AttentionRPN) [54]	WACV 2022	Faster R-CNN R-50	—	—	—	—	13.0	24.7	12.1	—	15.3	29.3	14.5	—
	▷ TD-Sampler [118]	ICCCBDA 2022	Faster R-CNN R-101	—	—	—	—	14.8	—	13.6	—	19.9	—	19.2	—
	† ▷ CME (MetaYOLO) [46]	CVPR 2021	YOLOv2	—	—	—	—	15.1	24.6	16.4	—	16.9	28.0	17.8	—
	† PNSD * [68]	ACCV 2020	Faster R-CNN R-50	—	—	—	—	15.3	21.7	12.5	—	—	—	—	—
	† FCT [66] ←	CVPR 2022	Faster R-CNN PVTv2-B2-Li	5.6	—	—	—	17.1	—	—	—	21.4	—	—	—
	▷ LVC [108]	CVPR 2022	Faster R-CNN R-101	—	—	—	—	17.8	30.9	17.8	31.9	24.5	41.1	25.0	33.0
	† KFSOD * [69]	CVPR 2022	Faster R-CNN R-50	—	—	—	—	18.5	26.3	18.7	—	—	—	—	—
	† DAnA * [53]	TMM 2021	Faster R-CNN R-50	—	—	—	—	18.6	—	17.2	—	21.6	—	20.3	—
	▷ LVC [108]	CVPR 2022	Faster R-CNN Swin-S	—	—	—	—	19.0	34.1	19.0	28.7	26.8	45.8	27.5	34.8
	▷ CFA-DeFRCN [123]	CVPR 2022	Faster R-CNN R-101	—	—	—	—	19.1	—	—	35.5	23.0	—	—	35.0
▷ UniT * [121]	CVPR 2021	Faster R-CNN R-50	—	—	—	—	21.7	40.8	20.6	—	23.1	43.0	21.6	—	
<b>Results averaged over multiple random runs:</b>															
no ft.	† SPCD [47]	ICIAP 2022	Faster R-CNN R-50	—	—	—	—	7.8	—	—	—	9.5	—	—	—
	† MM-FSOD [50]	arXiv 2020	Faster R-CNN R-34	—	—	—	—	8.2	19.2	8.0	—	—	—	—	—
	† QA-FewDet [61]	ICCV 2021	Faster R-CNN R-101	5.1	10.5	4.5	—	10.2	20.4	9.0	—	11.5	23.4	10.3	—
with finetuning	● MetaDet [97]	ICCV 2019	Faster R-CNN VGG-16	—	—	—	—	7.1	14.6	6.1	—	11.3	21.7	8.1	—
	▷ TFA w/cos [101] ←	ICML 2020	Faster R-CNN R-101	1.9	3.8	1.7	31.9	9.1	—	—	32.4	12.1	—	12.0	34.2
	▷ Retentive R-CNN [106] ←	CVPR 2021	Faster R-CNN R-101	—	—	—	—	9.5	—	—	39.2	12.4	—	—	39.3
	† GenDet [51]	TNNLS 2021	Faster R-CNN R-101	—	—	—	—	9.9	18.8	9.6	—	14.3	27.5	13.8	—
	▷ FSCE [102] ←	CVPR 2021	Faster R-CNN R-101	—	—	—	—	11.1	—	9.8	—	15.3	—	14.2	—
	† SPCD [47]	ICIAP 2022	Faster R-CNN R-50	—	—	—	—	11.5	—	—	—	16.4	—	—	—
	† QA-FewDet [61]	ICCV 2021	Faster R-CNN R-101	4.9	10.3	4.4	—	11.6	23.9	9.8	—	16.5	31.9	15.5	—
	† FsDetView [43]	ECCV 2020	Faster R-CNN R-50	—	—	—	—	12.5	27.3	9.8	—	14.7	30.6	12.2	—
	† DCNet [62]	CVPR 2021	Faster R-CNN R-101	—	—	—	—	12.8	23.4	11.2	—	18.6	32.6	17.5	—
	† ARRMR [49]	El. Imag. 2022	Faster R-CNN R-101	—	—	—	—	12.9	20.3	13.8	—	—	—	—	—
	† APSP (FsDetView) [54]	WACV 2022	Faster R-CNN R-50	—	—	—	—	13.4	30.6	9.1	—	17.1	35.2	14.7	—
	▷ MemFRCN (DeFRCN) [104]	TFECCS 2022	Faster R-CNN R-101	5.2	—	—	—	14.0	—	—	—	17.5	—	—	—
	† FCT [66] ←	CVPR 2022	Faster R-CNN PVTv2-B2-Li	5.1	—	—	—	15.3	—	—	—	20.2	—	—	—
	† CAREd [52]	Displays 2022	Faster R-CNN R-50	—	—	—	—	15.5	25.1	14.9	—	18.4	30.1	17.7	—
	† TIP [44]	CVPR 2021	Faster R-CNN R-101	—	—	—	—	16.3	33.2	14.1	—	18.3	35.9	16.9	—
	† IFC [48]	APIN 2022	Faster R-CNN R-50	—	—	—	—	16.7	29.2	15.5	—	17.9	33.4	16.5	—
	▷ DeFRCN (base+novel) [114]	ICCV 2021	Faster R-CNN R-101	4.8	—	—	—	16.8	—	—	34.0	21.2	—	—	34.8
	▷ DeFRCN (just novel) [114]	ICCV 2021	Faster R-CNN R-101	9.3	—	—	—	18.5	—	—	—	22.6	—	—	—
† Meta-DETR [64]	TPAMI 2022	Deformable DETR R-101	7.5	12.5	7.7	—	19.0	30.5	19.7	—	22.2	35.0	22.8	—	

CGG [58]) assume each query image  $I^Q$  to have at least one instance of the object category  $c$  from the support image  $I^{S,c}$ . Implicitly, this removes the classification task and only requires localization. The same applies to the one-way training and evaluation setting of DAnA [53], where the detector only

needs to estimate whether the current query image  $I^Q$  contains the object and where it is located, but the difficulty of correct classification is eliminated. In order to report comparable results, we therefore strongly recommend to always evaluate with the  $N$ -way setting.

In contrast, PNSD [68] and KFSOD [69] simplify the generalization to novel categories by already utilizing a ResNet pretrained on COCO such that the novel categories are not really novel anymore.

#### D. Problems of Common Evaluation Protocols

1) *High Variance for Different Samples:* As pointed out by Wang et al. [101], the use of different samples for novel categories can lead to a high variance in performance and, therefore, makes comparison difficult. Hence, we highly recommend to always report the average of the results over multiple random runs.

2) *ImageNet Pretraining and Choice of Novel Categories:* Most approaches use an ImageNet-pretrained backbone. While this is common for generic object detection, it has a negative side effect for FSOD: The novel categories are not truly novel anymore, as the model has probably already seen images of this category. However, omitting ImageNet pretraining altogether results in worse performance even for the base categories. To alleviate this problem, there are two options.

First, the ImageNet categories, which correspond to the novel categories, can be excluded from ImageNet-pretraining as done in CoAE [57], SRR-FSD [122], and AIT [56]. CoAE [57] and AIT [56] even remove all COCO-related categories from ImageNet, which results in 275 categories being removed. However, as argued by Zhu et al. [122] (SRR-FSD), removing all COCO-related categories is not realistic, as these categories are very common in the natural world and removing 275 categories may affect the representational power learned through pretraining. Therefore, Zhu et al. [122] only removed categories corresponding to novel categories for PASCAL VOC, resulting in 50 categories being removed on average. Yet, this requires additional pretraining for every different set of novel categories.

The second option for preventing foreseeing novel categories is using a dataset with novel categories that do not occur in ImageNet. Such a dataset would also be more realistic. Using categories such as cats as novel categories is absurd as there are loads of annotated data. Therefore, a more realistic approach would be to select novel categories that are indeed rare. For example, the LVIS [145] dataset provides a natural long tail with more and less frequent categories. A maximum of ten training instances are available for rare categories. Therefore, they can be used as novel categories. However, Huang et al. [39] pointed out that some rare categories in the training set do not appear in the validation set at all, which hinders the performance evaluation and requires further refinement of balanced splits and evaluation sets.

## X. CURRENT TRENDS

### A. Improvement of Techniques

Currently, dual-branch meta learning approaches improve much by using attention for aggregating features of both branches [48], [54], [64]. By aggregating before the RPN [55] or using a proposal-free transformer as detector [64], the issue of missing proposals for novel categories is effectively solved. Transfer learning approaches currently improve much by guiding the gradient flow to be able to train as many components of the detector as possible [114]. In both types of approaches, a current trend is the use of metric learning concepts by modifying the loss function to enable better category separation [48]. More trends are highlighted as part of the comparison in Section VIII.

### B. Extension to Related Research Areas

Besides these trends toward improving FSOD techniques, the extension of FSOD concepts to further research areas, such as a weakly supervised setting [146], self-supervised learning [39], or to few-shot instance segmentation [147], [148], is also a current trend.

### C. Open Challenges

Many approaches focus on either improving meta learning or transfer learning but often neglect that concepts between both types of approaches are exchangeable as pointed out in Table I, which leaves the potential for improvements in future work. Since the mainly used FSOD benchmarks PASCAL VOC and Microsoft COCO do not contain realistic novel categories that represent rare objects, we would like to encourage future research to additionally evaluate on more realistic datasets such as LVIS or FSOD, as already done in [50], [55], [101], and [149]. In addition, in [150], a framework for creating a customized FSOD dataset is provided. When employing FSOD in a realistic setting, including really rare categories, likely, a domain shift will occur. Therefore, concepts from cross-domain detection [23], [24], [25] should be further explored in future work.

## XI. CONCLUSION

In this survey, we provided a comprehensive overview of the state of the art for FSOD. We categorized the approaches according to their training scheme and architectural layout into single- and dual-branch meta learning and transfer learning. Meta learning approaches use episodic training to improve the subsequent learning with few object instances per novel category. Dual-branch meta learning approaches utilize a separate support branch receiving the image of a designated object, to learn how to represent the objects' category and where to attend in the query image. Transfer learning approaches use a more simplified training scheme, by simply fine-tuning on the novel categories.

After introducing the main concepts, we elaborated on how specific approaches differ from the general realization and gave short takeaways in order to highlight key insights for well performing methods. Based on an analysis of benchmark results on the most widely used datasets PASCAL VOC and Microsoft COCO, we identified current trends in the best performing dual-branch meta learning and transfer learning approaches. It remains an open question, which of these two concepts will prevail.

## REFERENCES

- [1] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.
- [2] L. Liu et al., "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, Feb. 2020.
- [3] A. Katzmann, O. Taubmann, S. Ahmad, A. Mühlberg, M. Sühling, and H.-M. Groß, "Explaining clinical decision support systems in medical imaging using cycle-consistent activation maximization," *Neurocomputing*, vol. 458, pp. 141–156, Oct. 2021.
- [4] L. Mannocci et al., "Leveraging social media and deep learning to detect rare megafauna in video surveys," *Conservation Biol.*, vol. 36, no. 1, Feb. 2022, Art. no. e13798.
- [5] L. B. Smith, S. S. Jones, B. Landau, L. Gershkoff-Stowe, and L. Samuelson, "Object name learning provides on-the-job training for attention," *Psychol. Sci.*, vol. 13, no. 1, pp. 13–19, Jan. 2002.

- [6] L. K. Samuelson and L. B. Smith, "They call it like they see it: Spontaneous naming and attention to shape," *Develop. Sci.*, vol. 8, no. 2, pp. 182–198, Mar. 2005.
- [7] L. A. Schmidt, "Meaning and compositionality as statistical induction of categories and constraints," Ph.D. dissertation, MIT, Cambridge, MA, USA, 2009.
- [8] H. Chen et al., "LSTD: A low-shot transfer detector for object detection," in *Proc. AAAI*, 2018, pp. 1–8.
- [9] X. Yan et al., "Meta R-CNN: Towards general solver for instance-level low-shot learning," in *Proc. ICCV*, 2019, pp. 9577–9586.
- [10] O. Vinyals et al., "Matching networks for one shot learning," in *Proc. NeurIPS*, 2016, pp. 3630–3638.
- [11] J. Snell et al., "Prototypical networks for few-shot learning," in *Proc. NeurIPS*, 2017, pp. 4077–4087.
- [12] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, May 2021.
- [13] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022.
- [14] H.-G. Jung and S.-W. Lee, "Few-shot learning with geometric constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4660–4672, Nov. 2020.
- [15] N. Lai, M. Kan, C. Han, X. Song, and S. Shan, "Learning to learn adaptive classifier–predictor for few-shot learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3458–3470, Aug. 2021.
- [16] O. Chapelle et al., "Semi-supervised learning," *IEEE Trans. Neural Netw.*, vol. 20, no. 3, p. 542, Oct. 2009.
- [17] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, Feb. 2020.
- [18] P. Tang, C. Ramaiah, Y. Wang, R. Xu, and C. Xiong, "Proposal learning for semi-supervised object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2291–2301.
- [19] Z. Song, X. Yang, Z. Xu, and I. King, "Graph-based semi-supervised learning: A comprehensive review," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 18, 2022, doi: [10.1109/TNNLS.2022.3155478](https://doi.org/10.1109/TNNLS.2022.3155478).
- [20] A. Gepperth and B. Hammer, "Incremental learning algorithms and applications," in *Proc. ESANN*, 2016, pp. 1–10.
- [21] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-end incremental learning," in *Proc. ECCV*, 2018, pp. 233–248.
- [22] Y. Wu et al., "Large scale incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 374–382.
- [23] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, "Exploring object relation in mean teacher for cross-domain detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11457–11466.
- [24] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, "Cross-domain detection via graph-induced prototype alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12355–12364.
- [25] C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, "Harmonizing transferability and discriminability for adapting object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8869–8878.
- [26] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," in *Proc. ECCV*, 2018, pp. 384–400.
- [27] S. Rahman, S. Khan, and F. Porikli, "Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts," in *Proc. ACCV*, 2018, pp. 547–563.
- [28] S. Rahman, S. Khan, and N. Barnes, "Polarity loss: Improving visual-semantic alignment for zero-shot detection," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 30, 2022, doi: [10.1109/TNNLS.2022.3184821](https://doi.org/10.1109/TNNLS.2022.3184821).
- [29] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5866–5885, Sep. 2022.
- [30] F. Shao et al., "Deep learning for weakly-supervised object detection and object localization: A survey," 2021, *arXiv:2105.12694*.
- [31] Y. Shen, R. Ji, C. Wang, X. Li, and X. Li, "Weakly supervised object detection via object-specific pixel gradient," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 5960–5970, Dec. 2018.
- [32] M. Kaya and H. C. S. Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, no. 9, p. 1066, Aug. 2019.
- [33] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, no. 2, pp. 1–10, Feb. 2009.
- [34] D. Agaian, M. Eisenbach, J. Wagner, R. Seichter, and H.-M. Gross., "Revisiting loss functions for person re-identification," in *Proc. ICANN*. Cham, Switzerland: Springer, 2021, pp. 30–42.
- [35] S. Antonelli et al., "Few-shot object detection: A survey," *ACM Comput. Surv.*, vol. 54, no. 11, pp. 1–37, 2022.
- [36] L. Jiaxu et al., "A comparative review of recent few-shot object detection algorithms," 2021, *arXiv:2111.00201*.
- [37] T. Liu, L. Zhang, Y. Wang, J. Guan, Y. Fu, and S. Zhou, "An empirical study and comparison of recent few-shot object detection algorithms," 2022, *arXiv:2203.14205*.
- [38] Q. Huang, H. Zhang, M. Xue, J. Song, and M. Song, "A survey of deep learning for low-shot object detection," 2021, *arXiv:2112.02814*.
- [39] G. Huang, I. Laradji, D. Vazquez, S. Lacoste-Julien, and P. Rodriguez, "A survey of self-supervised and few-shot object detection," 2021, *arXiv:2110.14711*.
- [40] C. Liu et al., "A survey of few-shot object detection," *J. Frontiers Comput. Sci. Technol.*, vol. 2022, pp. 1–15, Feb. 2022.
- [41] L. Zhang, S. Zhou, J. Guan, and J. Zhang, "Accurate few-shot object detection with support-query mutual guidance and hybrid loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14424–14432.
- [42] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2961–2969.
- [43] Y. Xiao and R. Marlet, "Few-shot object detection and viewpoint estimation for objects in the wild," in *Proc. ECCV*, 2020, pp. 192–210.
- [44] A. Li and Z. Li, "Transformation invariant few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3094–3102.
- [45] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, "Few-shot object detection via feature reweighting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8420–8429.
- [46] B. Li, B. Yang, C. Liu, F. Liu, R. Ji, and Q. Ye, "Beyond max-margin: Class margin equilibrium for few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7363–7372.
- [47] D. Kobayashi, "Self-supervised prototype conditional few-shot object detection," in *Proc. ICIAP*, 2022, pp. 681–692.
- [48] M. Wang, H. Ning, and H. Liu, "Object detection based on few-shot learning via instance-level feature correlation and aggregation," *Int. J. Speech Technol.*, vol. 53, no. 1, pp. 351–368, Jan. 2023.
- [49] L. Huang, Z. He, and X. Feng, "Few-shot object detection with affinity relation reasoning," *J. Electron. Imag.*, vol. 31, no. 3, pp. 1–10, May 2022.
- [50] Y. Li, W. Feng, S. Lyu, Q. Zhao, and X. Li, "MM-FSOD: Meta and metric integrated few-shot object detection," 2020, *arXiv:2012.15159*.
- [51] L. Liu et al., "GenDet: Meta learning to generate detectors from few shots," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3448–3460, Aug. 2022.
- [52] J. Quan, B. Ge, and L. Chen, "Cross attention redistribution with contrastive learning for few shot object detection," *Displays*, vol. 72, Apr. 2022, Art. no. 102162.
- [53] T.-I. Chen et al., "Dual-awareness attention for few-shot object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 291–301, 2023.
- [54] H. Lee, M. Lee, and N. Kwak, "Few-shot object detection by attending to per-sample-prototype," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2445–2454.
- [55] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-RPN and multi-relation detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4013–4022.
- [56] D.-J. Chen, H.-Y. Hsieh, and T.-L. Liu, "Adaptive image transformer for one-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12247–12256.
- [57] T.-I. Hsieh et al., "One-shot object detection with co-attention and co-excitation," in *Proc. NeurIPS*, 2019, pp. 2725–2734.
- [58] C. Michaelis, M. Bethge, and A. S. Ecker, "Closing the generalization gap in one-shot object detection," 2020, *arXiv:2011.04267*.
- [59] G. Han, S. Huang, J. Ma, Y. He, and S. F. Chang, "Meta faster R-CNN: Towards accurate few-shot object detection with attentive feature alignment," in *Proc. AAAI*, 2022, pp. 780–789.
- [60] X. Li, L. Zhang, Y. Pun Chen, Y.-W. Tai, and C.-K. Tang, "One-shot object detection without fine-tuning," 2020, *arXiv:2005.03819*.
- [61] G. Han et al., "Query adaptive few-shot object detection with heterogeneous graph convolutional networks," in *Proc. ICCV*, 2021, pp. 3263–3272.

- [62] H. Hu, S. Bai, A. Li, J. Cui, and L. Wang, "Dense relation distillation with context-aware aggregation for few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10185–10194.
- [63] G. Kim, H. G. Jung, and S. W. Lee, "Few-shot object detection via knowledge transfer," in *Proc. SMC*, 2020, pp. 3564–3569.
- [64] G. Zhang, Z. Luo, K. Cui, S. Lu, and E. P. Xing, "Meta-DETR: Image-level few-shot detection with inter-class correlation exploitation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 2, 2022, doi: 10.1109/TPAMI.2022.3195735.
- [65] J.-M. Perez-Rua, X. Zhu, T. M. Hospedales, and T. Xiang, "Incremental few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13846–13855.
- [66] G. Han, J. Ma, S. Huang, L. Chen, and S.-F. Chang, "Few-shot object detection with fully cross-transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5321–5330.
- [67] L. Yin, J. M. Perez-Rua, and K. J. Liang, "Sylph: A hypernetwork framework for incremental few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9035–9045.
- [68] S. Zhang, D. Luo, L. Wang, and P. Koniusz, "Few-shot object detection by second-order pooling," in *Proc. ACCV*, 2020, pp. 1–19.
- [69] S. Zhang, L. Wang, N. Murray, and P. Koniusz, "Kernelized few-shot object detection with efficient integral aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19207–19216.
- [70] S. Ren et al., "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NeurIPS*, 2015, pp. 91–99.
- [71] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [72] H. Zhang and P. Koniusz, "Power normalizing second-order similarity network for few-shot learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1185–1193.
- [73] J. Zhang, L. Wang, L. Zhou, and W. Li, "Beyond covariance: SICE and kernel based visual feature representation," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 300–320, Feb. 2021.
- [74] A. J. Smola and R. Kondor, "Kernels and regularization on graphs," in *Learning Theory and Kernel Machines*. Cham, Switzerland: Springer, 2003.
- [75] C. Michaelis, I. Ustyuzhaninov, M. Bethge, and A. S. Ecker, "One-shot instance segmentation," 2018, *arXiv:1811.11507*.
- [76] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11030–11039.
- [77] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.
- [78] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
- [79] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.
- [80] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, "Deep residual shrinkage networks for fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4681–4690, Jul. 2020.
- [81] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
- [82] W. Wang et al., "PVT V2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, Sep. 2022.
- [83] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.
- [84] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. ICLR*, 2017.
- [85] A. Li, W. Huang, X. Lan, J. Feng, Z. Li, and L. Wang, "Boosting few-shot learning with adaptive margin loss," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12576–12584.
- [86] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
- [87] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
- [88] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.
- [89] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
- [90] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*.
- [91] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.
- [92] X. Zhu et al., "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. ICLR*, 2021.
- [93] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
- [94] L. Karlinsky et al., "RepMet: Representative-based metric learning for classification and few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5197–5206.
- [95] Y. Yang, F. Wei, M. Shi, and G. Li, "Restoring negative information in few-shot object detection," in *Proc. NeurIPS*, 2020, pp. 3521–3532.
- [96] G. Zhang, K. Cui, R. Wu, S. Lu, and Y. Tian, "PNPDet: Efficient few-shot detection without forgetting via plug-and-play sub-networks," in *Proc. WACV*, 2021, pp. 3823–3832.
- [97] Y.-X. Wang, D. Ramanan, and M. Hebert, "Meta-learning to detect rare objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9925–9934.
- [98] S. Li et al., "Meta-RetinaNet for few-shot object detection," in *Proc. BMVC*, 2020.
- [99] K. Fu et al., "Meta-SSD: Towards fast adaptation for few-shot object detection with meta-learning," *IEEE Access*, vol. 7, pp. 77597–77606, 2019.
- [100] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. ECCV*, 2016, pp. 21–37.
- [101] X. Wang et al., "Frustratingly simple few-shot object detection," in *Proc. ICML*, 2020, pp. 9919–9928.
- [102] B. Sun, B. Li, S. Cai, Y. Yuan, and C. Zhang, "FSCE: Few-shot object detection via contrastive proposal encoding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7352–7362.
- [103] H. Feng, L. Zhang, X. Yang, and Z. Liu, "Incremental few-shot object detection via knowledge transfer," *Pattern Recognit. Lett.*, vol. 156, pp. 67–73, Apr. 2022.
- [104] T. Lu, S. Jia, and H. Zhang, "MemFRCN: Few shot object detection with memorable faster-RCNN," *IEICE Trans. Fundamentals Electron., Commun. Comput. Sci.*, vol. 105, no. 12, pp. 1626–1630, 2022.
- [105] W. Zhang and Y.-X. Wang, "Hallucination improves few-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13008–13017.
- [106] Z. Fan, Y. Ma, Z. Li, and J. Sun, "Generalized few-shot object detection without forgetting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4527–4536.
- [107] W. Zhang, Y.-X. Wang, and D. A. Forsyth, "Cooperating RPN's improve few-shot object detection," 2020, *arXiv:2011.10142*.
- [108] P. Kaul, W. Xie, and A. Zisserman, "Label, verify, correct: A simple few shot object detection method," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14237–14247.
- [109] A.-K.-N. Vu et al., "Few-shot object detection via baby learning," *Image Vis. Comput.*, vol. 120, Apr. 2022, Art. no. 104398.
- [110] Y. Wang, C. Xu, C. Liu, and Z. Li, "Context information refinement for few-shot object detection in remote sensing images," *Remote Sens.*, vol. 14, no. 14, p. 3255, Jul. 2022.
- [111] J. Wu, S. Liu, Di Huang, and Y. Wang, "Multi-scale positive sample refinement for few-shot object detection," in *Proc. ECCV*, 2020, pp. 456–472.
- [112] A. Wu et al., "Generalized and discriminative few-shot object detection via SVD-dictionary enhancement," in *Proc. NeurIPS*, 2021, pp. 6353–6364.
- [113] A. Wu, Y. Han, L. Zhu, and Y. Yang, "Universal-prototype enhancing for few-shot object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1–4.
- [114] L. Qiao, Y. Zhao, Z. Li, X. Qiu, J. Wu, and C. Zhang, "DeFRCN: Decoupled faster R-CNN for few-shot object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8681–8690.
- [115] Y. Lu, X. Chen, Z. Wu, and J. Yu, "Decoupled metric network for single-stage few-shot object detection," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 514–525, Jan. 2023.

- [116] Y. Li et al., “Few-shot object detection via classification refinement and distractor retreatment,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15395–15403.
- [117] W. Liu, C. Wang, S. Yu, C. Tao, J. Wang, and J. Wu, “Novel instance mining with pseudo-margin evaluation for few-shot object detection,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2250–2254.
- [118] C. Wu, B. Wang, S. Liu, X. Liu, and P. Wu, “TD-Sampler: Learning a training difficulty based sampling strategy for few-shot object detection,” in *Proc. 7th Int. Conf. Cloud Comput. Big Data Anal. (ICCCBDA)*, Apr. 2022, pp. 275–279.
- [119] Y. Cao et al., “Few-shot object detection via association and discrimination,” in *Proc. NeurIPS*, 2021, pp. 16570–16581.
- [120] Z. Yang, Y. Wang, X. Chen, J. Liu, and Y. Qiao, “Context-transformer: Tackling object confusion for few-shot detection,” in *Proc. AAAI*, 2020, pp. 12653–12660.
- [121] S. Khandelwal, R. Goyal, and L. Sigal, “UniT: Unified knowledge transfer for any-shot object detection and segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5951–5961.
- [122] C. Zhu, F. Chen, U. Ahmed, Z. Shen, and M. Savvides, “Semantic relation reasoning for shot-stable few-shot object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8782–8791.
- [123] K. Guirguis, A. Hendawy, G. Eskandar, M. Abdelsamad, M. Kayser, and J. Beyerer, “CFA: Constraint-based finetuning approach for generalized few-shot object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 4039–4049.
- [124] J. Wang and D. Chen, “Few-shot object detection method based on knowledge reasoning,” *Electronics*, vol. 11, no. 9, p. 1327, Apr. 2022.
- [125] H. Xu, X. Wang, F. Shao, B. Duan, and P. Zhang, “Few-shot object detection via sample processing,” *IEEE Access*, vol. 9, pp. 29207–29221, 2021.
- [126] X. Chen, M. Jiang, and Q. Zhao, “Leveraging bottom-up and top-down attention for few-shot object detection,” 2020, *arXiv:2007.12104*.
- [127] L.-C. Chen et al., “Encoder–decoder with Atrous Separable convolution for semantic image segmentation,” in *Proc. ECCV*, 2018, pp. 1–4.
- [128] A. Dosovitskiy et al., “An image is worth 16×16 words: Transformers for image recognition at scale,” in *Proc. ICLR*, 2021.
- [129] M. Caron et al., “Emerging properties in self-supervised vision transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.
- [130] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
- [131] H. Qi, M. Brown, and D. G. Lowe, “Low-shot learning with imprinted weights,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5822–5830.
- [132] X. Chen, Y. Wang, J. Liu, and Y. Qiao, “DID: Disentangling-imprinting-distilling for continuous low-shot detection,” *IEEE Trans. Image Process.*, vol. 29, pp. 7765–7778, 2020.
- [133] G. A. Miller, “WordNet: A lexical database for English,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.
- [134] D. Lopez-Paz and M. A. Ranzato, “Gradient episodic memory for continual learning,” in *Proc. NeurIPS*, 2017, pp. 6467–6476.
- [135] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, “Efficient lifelong learning with A-GEM,” in *Proc. ICLR*, 2019.
- [136] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.
- [137] J. Zhang and S. Sclaroff, “Saliency detection: A Boolean map approach,” in *Proc. ICCV*, 2013, pp. 153–160.
- [138] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, “Predicting human eye fixations via an LSTM-based saliency attentive model,” *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, Oct. 2018.
- [139] H. Zhang, J. Xue, and K. Dana, “Deep TEN: Texture encoding network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 708–717.
- [140] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, “PANet: Few-shot image semantic segmentation with prototype alignment,” in *Proc. ICCV*, 2019, pp. 9197–9206.
- [141] J. Liu, L. Song, and Y. Qin, “Prototype rectification for few-shot learning,” in *Proc. ECCV*, 2020, pp. 741–756.
- [142] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, *arXiv:1804.02767*.
- [143] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VOC) challenge,” *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
- [144] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
- [145] A. Gupta, P. Dollar, and R. Girshick, “LVIS: A dataset for large vocabulary instance segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5356–5364.
- [146] L. Karlinsky et al., “StarNet: Towards weakly supervised few-shot object detection,” in *Proc. AAAI*, 2021, pp. 1743–1753.
- [147] Z. Fan et al., “FGN: Fully guided network for few-shot instance segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9172–9181.
- [148] K. Nguyen and S. Todorovic, “FAPIS: A few-shot anchor-free part-based instance segmenter,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11099–11108.
- [149] X. Hu, Y. Jiang, K. Tang, J. Chen, C. Miao, and H. Zhang, “Learning to segment the tail,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14045–14054.
- [150] W. Bailer, “Making few-shot object detection simpler and less frustrating,” in *Proc. MMM*, 2022, pp. 445–451.



**Mona Köhler** received the bachelor’s degree in media technology and the master’s degree in computer engineering from the Ilmenau University of Technology, Ilmenau, Germany, in 2018 and 2020, respectively.

Since 2020, she has been a Research Assistant at the Neuroinformatics and Cognitive Robotics Laboratory, Ilmenau University of Technology. Her research interests include deep-learning-based object detection, semantic and instance segmentation, and learning with limited data.



**Markus Eisenbach** received the Diploma (M.Sc.) degree in computer science and the Ph.D. degree in robotics from the Ilmenau University of Technology, Ilmenau, Germany, in 2009 and 2019, respectively.

After his studies, he joined the Neuroinformatics and Cognitive Robotics Laboratory, Ilmenau University of Technology, as a Research Assistant. His research interests are robotics, deep and machine learning, and computer vision in the context of human–robot interaction.



**Horst-Michael Gross** (Member, IEEE) received the Ph.D. degree in neuroinformatics from the Ilmenau University of Technology, Ilmenau, Germany, in 1989.

Since 1993, he has been a Full Professor of computer science and the Head of the Chair of Neuroinformatics and Cognitive Robotics, Ilmenau University of Technology. His long-standing research interests include mobile robotics, human–robot interaction, rehabilitation robotics, machine learning, and computer vision with a particular focus on everyday usability of assistive robots in real-world domestic and public application scenarios.