Semantic-aligned Fusion Transformer for One-shot Object Detection

Yizhou Zhao*[1] Xun Guo[2] Yan Lu[2]
[1]Carnegie Mellon University [2]Microsoft Research Asia
yizhouz@andrew.cmu.edu {xunguo, yanlu}@microsoft.com

Abstract

One-shot object detection aims at detecting novel objects according to merely one given instance. With extreme data scarcity, current approaches explore various feature fusions to obtain directly transferable meta-knowledge. Yet, their performances are often unsatisfactory. In this paper, we attribute this to inappropriate correlation methods that misalign query-support semantics by overlooking spatial structures and scale variances. Upon analysis, we leverage the attention mechanism and propose a simple but effective architecture named Semantic-aligned Fusion Transformer (SaFT) to resolve these issues. Specifically, we equip SaFT with a vertical fusion module (VFM) for cross-scale semantic enhancement and a horizontal fusion module (HFM) for cross-sample feature fusion. Together, they broaden the vision for each feature point from the support to a whole augmented feature pyramid from the query, facilitating semantic-aligned associations. Extensive experiments on multiple benchmarks demonstrate the superiority of our framework. Without fine-tuning on novel classes, it brings significant performance gains to one-stage baselines, lifting state-of-the-art results to a higher level.

1. Introduction

Recent years have witnessed the flourish of large-scale perception systems like [3, 23]. Yet it has a long way to go towards real human-like intelligence. Being one of the underlying problems, few-shot learning received more and more interest from language [1, 17, 46, 58] to vision [15, 24, 37, 47, 49, 52, 54] related tasks. This scenario aims at learning a well-generalized model with scarcely labeled data, which challenges conventional learning paradigms.

To bridge the aforementioned gap in few-shot object detection (FSD), existing literature suggests drawing support from transfer-learning [7, 16, 48, 54, 55, 65] or meta-learning [15, 24, 27, 28, 56, 57, 61]. Although the former is simple to conduct through pretraining on massive base classes and fine-tuning on scant novel ones, it suffers from the two-stage redundant procedures. The network should always utilize new-coming few-shot data to optimize parameters before it can well recognize these novel classes, thereby limiting its application. In contrast, the latter trend considers meta-knowledge extraction from sampled meta-tasks. This line of frameworks is expected to adapt directly to similarly organized tasks even without online fine-tuning, though it usually helps in performance. At present, this offline meta-learning paradigm is preferred by one-shot object detection (OSD) specific pipelines, with an out-of-the-box availability.

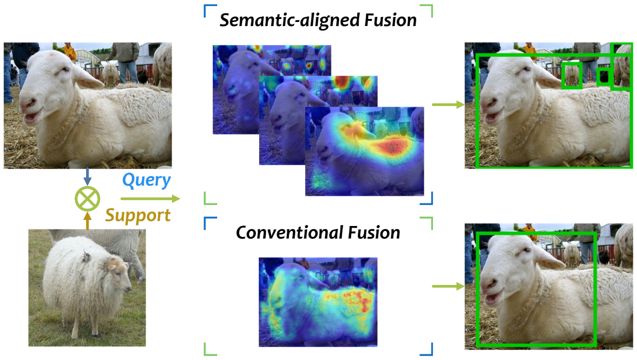

Figure 1. **Comparison of semantic-aligned fusion and conventional fusion.** Heatmaps and detection results of these two are based on our SaFT and a baseline with original cross-sample attention accordingly. Comparing the two schemes, semantic-aligned fusion activates more concentrated heatmaps on various feature levels and produces better OSD results.

In such a setting, the model should be well constructed to learn the relatedness between a given scene, i.e., the query, and an example patch, i.e., the support. To facilitate this, a series of works [15, 21, 24, 28, 38, 40, 57, 61] investigate cross-sample feature fusion, which augments query features with support representations via sample or ROI level correlation. However, neglecting semantic mismatches in space and scale limits their performances in one-shot scenarios.

Concretely, the traditional paradigm suggests generating a prototype [15, 24] or a kernel [61] from the support to associate with query features. With most spatial information compressed, the long-range structure-dependent relation in between remains hardly mined. Although pooled prototypes in Fig. 2(b) and learned kernels in Fig. 2(c) are effec-

*The work was done when the author was with MSRA as an intern.

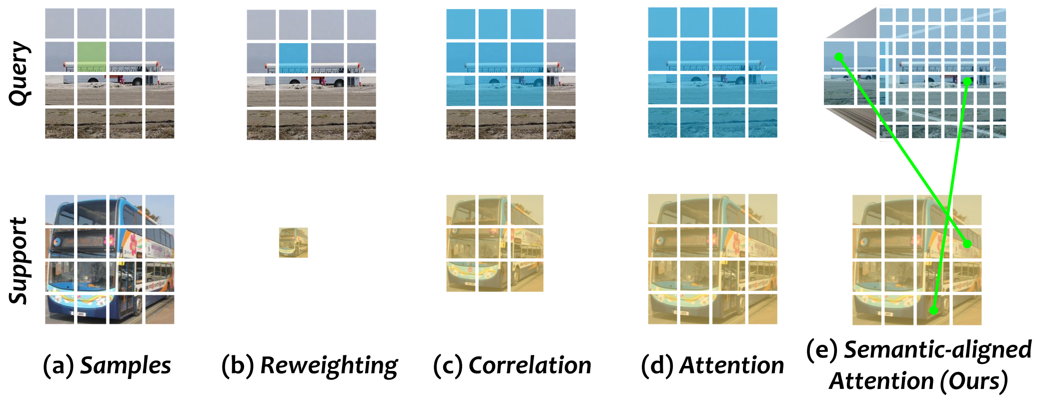

Figure 2. **Visualization of different fusion approaches.** We present previous fusion schemes in (b), (c), (d), and our proposed semantic-aligned attention as a sort of semantic-aligned fusion in (e). Images are split into patches for illustration, with each patch representing the receptive field of a feature point. The only green patch indicates the query where a response is expected, blue patches are values that contribute to one feature point of the fusion result, and yellow ones are keys that interact with these values. Green lines in (e) suggest two pairs of ideal matches, where different granularities of query features are used. Support samples are scaled in reweighting and correlation schemes to visualize their compression in spatial information.

tive in distinguishing one category from another, they contain fewer positioning priors and thus hinder their localization ability. Furthermore, these schemes match global support representations with local query contexts, regardless of semantic misalignment. An emerging trend [6, 22] seeks help from the attention mechanism for adaptive feature fusion. While easing problems discussed to some extent, they commonly focus on feature pairs on a single scale as shown in Fig. 2(d), leaving the multi-scale detection task for later anchor-based detector heads. Therefore, it makes no sense when targets are scattered on different scales. For instance, in Fig. 2(e), ideal matches for the bus wheel and rear windows lie in two distinct levels of query features, making any single-scale attempt sub-optimal. A simple multi-scale implementation cannot resolve it either, since it fuses the query and the support one scale at a time. Without cross-scale long-range interactions, this rigid manner is likely to fail in cases with semantic missing, such as occlusions or query-support inconsistencies in shape and size.

To encourage more appropriate and sufficient feature interactions in OSD, we propose to adaptively fuse each feature point from the support with each from the query feature pyramid. Thus the original attention mechanism is extended to semantic-aligned attention as illustrated in Fig. 2(e). Features from each side are first deconstructed into semantic units, i.e., feature points. Then these units interacts one another in a global manner, not only between query-support sample pairs (horizontally) but also among different scales (vertically). Since objects and parts of objects might exist in different scales and locations, the association process weighted collocates multiple semantic units to make a proper match. In this way, semantic-aligned attention enriches the semantic space that each feature point can utilize, thereby promoting better alignments between the query and the support.

Our Semantic-aligned Fusion Transformer (SaFT) implements this fusion scheme, with Fig. 3 demonstrating its overall structure. It follows a one-stage proposal-free design and can be easily extended to two-stage pipelines through cascading proposal-based heads. Compared with allied frameworks that employ reweighting or correlation, SaFT alternatively contains a vertical fusion module (VFM) and a horizontal fusion module (HFM). The former is placed after the feature extractor to together form a Siamese backbone followed by the latter. VFM prepares scale-attended features via vertical attention (VA) in Fig. 5, and HFM utilizes them from query and support with horizontal attention (HA) in Fig. 4. Note that a single level of support feature interacts with multiple from the other side for a comprehensive view. Thanks to the cross-scale and cross-sample relatedness modeled by the attention mechanism, SaFT achieves remarkable performance gains in both PASCAL-VOC and MS-COCO datasets.

We conclude our contributions as three-fold.

1. To the best of our knowledge, our Semantic-aligned Fusion Transformer is the first to carry out the offline one-shot object detection task with proposal-free one-stage detectors, producing better performance than state-of-the-art two-stage models.
2. We discuss the problems of query-support feature fusion and propose a unified attention mechanism to tackle semantic misalignment in space and scale. Our implementation of this can be used as a general fusion neck.
3. Through qualitative and quantitative experiments, we prove that our novel semantic-aligned fusion is superior to conventional association methods by involving cross-scale long-range relations and collecting more comprehensive meta-knowledge.

2. Related Work

2.1. General Object Detection

Given a plain image, general object detection aims to localize and classify the objects concerned. Modern detectors can be roughly divided into two categories, namely two-stage proposal-based methods and one-stage proposal-free ones. Two-stage pipelines [4, 8, 18, 19, 31, 44] generate a set of class-agnostic region proposals in the first stage and refine as well as classify them into final results in the second. In contrast, one-stage approaches use a class-aware locator to omit the second stage, mostly based on densely placed anchor boxes [32, 36, 43] or anchor points [12, 26, 50, 62]. Different from these, another line of work soaring recently leads a new trend of heuristic-free design. By introducing the attention mechanism, the DETR series [5, 9, 66] have achieved better performance while being fully end-to-end. Our model builds on one-stage detector FCOS [50] for simplicity, while is rather plug-and-play as a fusion neck.

2.2. One/Few-shot Object Detection

With sufficient data of base classes while limited samples of novel ones, few-shot scenarios bring more challenges to object detection. Recent work leads two mainstreams in addressing this problem, using transfer-learning or meta-learning techniques. Transfer-learning-based methods [7, 16, 42, 48, 54, 55, 60, 63, 65] follow a two-stage training schema, which is pretraining and fine-tuning, to transfer knowledge from base classes to novel classes. By comparison, the latter trend [15, 22, 24, 27–29, 56, 57, 61] recasts the problem in a meta-learning form, encouraging efficient knowledge adaptation through meta-tasks sampling and region-based metric-learning. OSD is an extreme case of FSD with merely one label available for each class to detect. Less data needs more generalization, spawning a series of offline models [6, 21] further explore similarity metrics and abandon the fine-tuning phase. While with different task settings, these methods share a common regional similarity comparison strategy with most FSD networks adopting metric-learning [25, 27, 48, 56, 57]. In other words, they highly rely on region proposals, which can be unpredictable in low-shot scenarios. Different from above, our approach learns the metric in a proposal-free manner, facilitating higher efficiency and flexibility.

2.3. Multi-scale Feature Fusion

Unlike humans that are born with a continuously zooming field of vision, modern convolutional feature extractors usually down-sample images in a discrete way. To mitigate this, multi-scale feature fusion techniques are developed in detection networks, bringing remarkable performance boosts. Three paths in the feature pyramid are exploited, i.e., top-down [31], bottom-up [35] and within-scale [5, 53]. Recent work further enriches multi-level information interaction through dense as well as various aggregations [30] and the attention mechanism [59, 64]. Although cross-sample feature fusion is widely investigated in one/few-shot problems [6, 10, 21, 34, 38, 41, 49, 52, 56], its cross-scale counterpart is relatively scarce. Hence, we consider aggregating these two dimensions and propose a unified attention mechanism for feature fusion between samples and among scales. Compared with its counterparts, this design experimentally helps in semantic alignments.

3. Method

3.1. Problem Definition

As in previous literature [6, 21], the one-shot object detection task constitutes of two sets of instances $\mathcal{D} = \mathcal{D}_{base} \cup \mathcal{D}_{novel}$, where $\mathcal{D}_{base}$ denotes a large base set with numerous available annotations and $\mathcal{D}_{novel}$ stands for a small novel set including only one instance per category. Note that the base classes $\mathcal{C}_{base}$ in $\mathcal{D}_{base}$ and novel classes $\mathcal{C}_{novel}$ in $\mathcal{D}_{novel}$ are mutually exclusive, i.e., $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \varnothing$.

We consider this problem in a meta-learning fashion akin to [15, 24, 28] whilst omitting the fine-tuning phase to constrain the setting to fully offline like [6, 21]. Given a query image Q and a support patch S, the task is to find all instances of the same category as S with their bounding boxes in Q. The base set $\mathcal{D}_{base}$ is provided in training to generate both queries Q_{base} and supports S_{base} while the novel set $\mathcal{D}_{novel}$ is utilized in testing for supports S_{novel} only.

3.2. Framework

We propose a concise framework, termed Semantic-aligned Fusion Transformer (SaFT), to settle our motivation. The overall architecture is sketched in Fig. 3. It adopts a Siamese backbone for aligned query-support feature extraction, a shared vertical fusion module (VFM) to enrich per-sample semantic hierarchically, and a subsequent horizontal fusion module (HFM) to aggregate information from both samples for later classification and regression.

3.3. Feature Fusion via Dense Attention

Initially introduced in natural language processing [51] and then borrowed to vision tasks [5, 11, 53], attention mechanism is known by its inductive bias in modeling long-range information. More specifically in location-aware tasks like detection, positional encodings [2, 5, 39] are adopted upon multi-head attention (MHA) to promote a permutation-variant architecture with

$$\text{PMA}(Q, K, V) = \text{MHA}(Q + \text{P}(Q), K + \text{P}(K), V) \quad (1)$$

where PMA abbreviates position-encoded multi-head attention and P represents positional encoding.

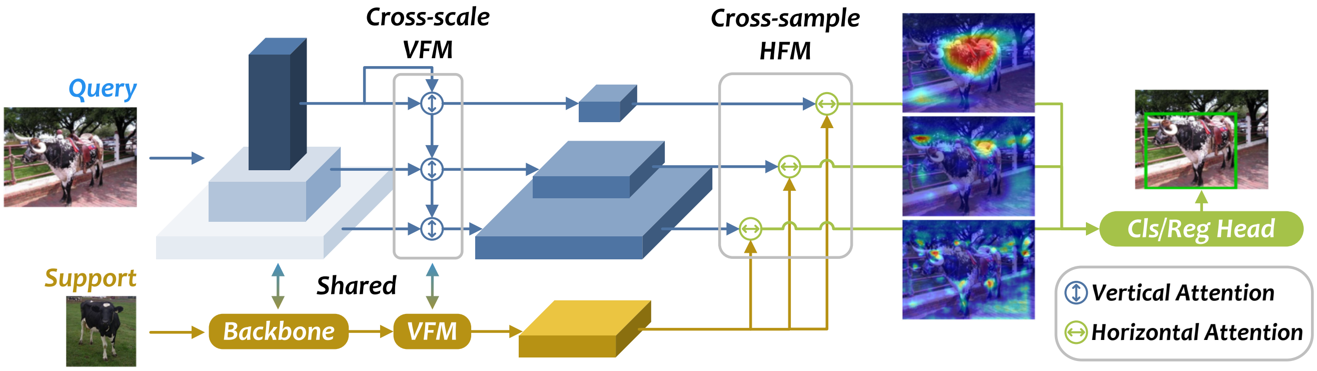

Figure 3. **The architecture of Semantic-aligned Fusion Transformer for One-shot Object Detection.** Darker color indicates features from deeper layers in backbone, the same in Fig. 5. VFM and HFM are vertical fusion module and horizontal fusion module separately.

Based on Eq. (1), we express our dense attention (DA) as follows

$$\mathrm{DA}(F^Q, F^K) = \mathrm{LN}(F^Q + \mathrm{PMA}(F^Q, F^K, F^K)) \quad (2)$$

where $F^Q \in \mathbb{R}^{h_Q w_Q \times d_Q}$, $F^K \in \mathbb{R}^{h_K w_K \times d_K}$, and LN denotes layer normalization. Akin to the decoder in [5] which captures a dense relationship between encoded features and object queries to decode, DA is expected to model a point-to-point correlation between F^Q and F^K, thus named as it is. We further extends DA in form of self-attention (SA)

$$\mathrm{SA}(F^Q) = \mathrm{DA}(F^Q, F^Q) \quad (3)$$

and cross-attention (CA)

$$\mathrm{CA}(F^Q, F^K) = \mathrm{DA}(F^Q, F^K) = F^{K \to Q} \quad (4)$$

$$\mathrm{CAF}(F^Q, F^K) = \mathrm{LN}(F^{K \to Q} + \mathrm{FFN}(F^{K \to Q})) \quad (5)$$

where CAF indicates CA with a consecutive feed-forward network (FFN) and an add-and-norm.

Upon these, we propose two sorts of attention blocks, horizontal attention and vertical attention. The elemental procedure of both of them is consistent, with an SA before a CA. This design helps with adaptiveness since SA selectively expresses information from the query side, and CA weighted balances the two sides.

3.4. Cross-sample Horizontal Attention

DA-based cross-sample horizontal attention (HA) is designed for fusion between features from Q and S samples. For comparison, we first concisely review conventional convolution-based approaches in FSD/OSD tasks and then introduce our methods.

Beginning with a pair of features F^Q and F^S extracted from the query and the support, traditional pair-wise operations either extract a prototype or learn a kernel of S as

$$\phi(F^S) = z^S \quad (6)$$

and then obtain a class-specific enhanced feature with channel-wise multiplication as Fig. 2(b) or convolution as Fig. 2(c)

$$\tilde{F}^Q = F^Q \odot z^S \quad (7)$$

where $\tilde{F}^Q$ denotes the enhanced query feature.

This schema highlights class-related information from support samples while discarding most of the spatial semantics. Furthermore, since the class-related z^S represents the whole support patch while its target is a local area from the query, this global-to-local correlation process may lead to a misalignment in space and scale.

In contrast, HA interacts every location pairs of F^Q and F^S as Fig. 2(d) shows. A single block of HA presented in Fig. 4 constitutes from a pair of SA and CAF

$$\bar{F}^Q_{i+1} = \mathrm{SA}(\tilde{F}^Q_i) \quad (8)$$

$$\bar{F}^S_{i+1} = \mathrm{SA}(\tilde{F}^S_i) \quad (9)$$

$$\tilde{F}^Q_{i+1} = \mathrm{CAF}(\bar{F}^Q_{i+1}, \bar{F}^S_{i+1}) \quad (10)$$

$$\tilde{F}^S_{i+1} = \mathrm{CAF}(\bar{F}^S_{i+1}, \bar{F}^Q_{i+1}) \quad (11)$$

where overlines and tildes stand for self-attended features and cross-attended features accordingly. For more sufficient feature interactions, these operations conduct iteratively with initial $\tilde{F}^Q_0 = F^Q$, $\tilde{F}^S_0 = F^S$ and $i = 0, \ldots, N-1$. Ending attention layers aggregate features from both sides

$$\tilde{F}^Q = \mathrm{HA}(F^Q, F^S) = \mathrm{CAF}(\mathrm{SA}(\tilde{F}^Q_N), \mathrm{SA}(\tilde{F}^S_N)) \quad (12)$$

We term these chained blocks of HA as a horizontal fusion module (HFM).

Intuitively, HFM conducts global-to-global similarity matching and expression. One by one it associates each feature point from the query and one from the support, regardless of their positions. This schema aligns features from two sides in a deformable and reorganizable manner, thereby making them more comparable.

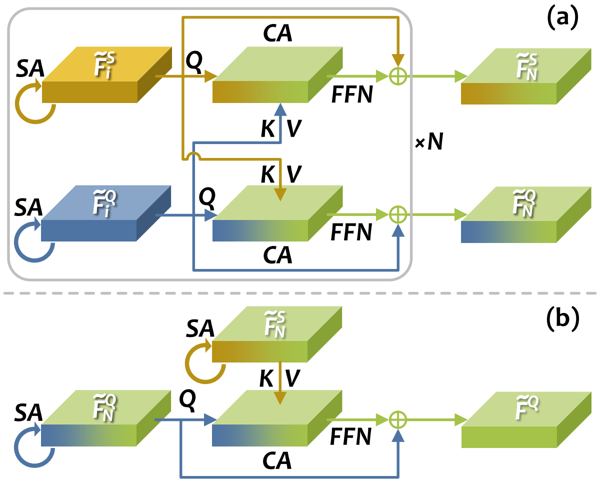

Figure 4. **The horizontal attention (HA) block.** HA includes two sequential processes. (a) Iterative two-way fusion process with pairs of SA and CAF. (b) Finalizing one-way aggregation process with SAs and a CAF.

3.5. Cross-scale Vertical Attention

Besides mutual interactions between Q and S on a single scale, we enhance multi-scale semantics of each sample with cross-scale vertical attention (VA). The whole procedure is shown in Fig. 3 with a close-up illustration in Fig. 5. To show its ability in semantic alignment, we begin with reviewing feature pyramid networks (FPN).

Widely adopted in object detection, FPN is an efficient plug-in to tackle scale variances. Its building blocks can be written as

$$\tilde{F}_j = \text{Conv}_{3\times3}(\text{Conv}_{1\times1}(F_j) + \text{Up}_{2\times}(\tilde{F}_{j+1})) \tag{13}$$

where F_j and $\tilde{F}_j$ are the level-j feature extracted by backbone and the corresponding result after fusion between neighboring scales, separately. We notice that FPN gathers semantics from higher levels to replenish lower levels locally. Notwithstanding its enriched contexts, this in-place scheme falls short in capturing cross-scale long-range information which can be semantically complementary in OSD, e.g., people in a long line are of the same category while having distinct appearance features.

To this end, we introduce VA. Given a feature pyramid $\{F_j|j = 3, \dots, M\}$ having strides $\{2^j|j = 3, \dots, M\}$ extracted by backbone, VA starts from the self-enhancement of top level M as

$$\tilde{F}_M = \text{Conv}_{3\times3}(\text{SA}(\text{Conv}_{1\times1}(F_M))) \tag{14}$$

In a top-down hierarchy, VA adaptively inquires related information globally from upper levels

$$\overline{F}_j = \text{SA}(\text{Conv}_{1\times1}(F_j)) \tag{15}$$

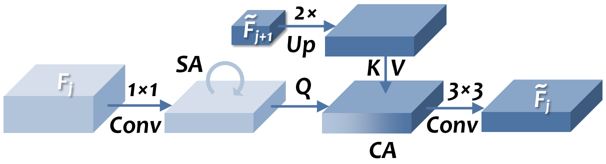

Figure 5. **The vertical attention (VA) block.** VA inserts an SA and CA pair between lateral and output convolutions of FPN.

$$\begin{aligned}\tilde{F}_j &= \text{VA}(F_j, \tilde{F}_{j+1}) \\ &= \text{Conv}_{3\times3}(\text{CA}(\overline{F}_j, \text{Up}_{2\times}(\tilde{F}_{j+1})))\end{aligned} \tag{16}$$

where $\tilde{F}_j$ is the enhanced feature from level j, which is up-sampled before fusion with the next level for alignment. This pyramidal process, named vertical fusion module (VFM), seeks to aggregate multi-scale global semantics.

Compared with FPN in Eq. (13), VFM inserts attention layers between lateral and output convolutions. Rather than linearly combining features in the same position, VFM promotes more flexible cross-scale feature interactions and better matches between query-support representations. The query collects and enriches potential target semantics spread in different positions and scales, whereas the support highlights the main target consistent across scales and dims irrelevant backgrounds. Moreover, VFM cooperates with HFM to expand the attentive field of a support feature point from a single layer to multiple scales as in Fig. 2(e). With this point-to-pyramid connection, richer semantics and cross-scale long-range correlations are available in matching, thus aiding query-support alignments.

4. Experiments

4.1. Experimental Setting

Benchmarks. We follow the previous work [6, 21] to train and evaluate our model on PASCAL-VOC [13, 14] and MS-COCO [33] with the same data splits. For VOC, a split of 20 classes partitions the whole dataset into 16 base classes and 4 novel classes. As for COCO, we divide the whole dataset of 80 classes into four groups, each having 20 classes. By turns, three groups are selected from the four as base classes for training and the remained 20 classes for evaluation. The original setting randomly samples query-support image pairs, which generates a different support patch each time a query image is given, in both training and testing. Differently, we keep the former but replace the latter with fixed seeds as in [24, 54, 56]. This strategy generates one random patch for each support class from COCO17 val and thus restricts the model to see merely one shot in testing rather than the whole set. Compared with the former evaluation, which is at risk of paralleling query and support with identical images, our proposed setting is closer to practical scenarios and de facto one-shot object detection.

Method / Set	Base																	Novel				
	Plant	Sofa	TV	Car	Bottle	Boat	Chair	Person	Bus	Train	Horse	Bike	Dog	Bird	Mbike	Table	Avg.	Cow	Sheep	Cat	Aero	Avg.
FSCE* [48]	**55.0**	69.6	**81.9**	83.9	**71.9**	65.9	45.2	45.9	83.6	85.4	86.4	**85.1**	79.2	79.5	**83.9**	**73.1**	73.5	72.3	67.0	53.9	**48.0**	60.3
DeFRCN* [42]	51.7	**74.4**	78.3	**87.1**	70.8	**67.6**	**52.4**	61.6	**85.0**	85.8	**87.6**	83.1	82.0	**83.8**	82.8	64.0	**74.9**	70.7	59.1	58.8	43.0	57.9
CoAE [21]	30.0	54.9	64.1	66.7	40.1	54.1	14.7	60.9	77.5	78.3	77.9	73.2	80.5	70.8	72.4	46.2	60.1	83.9	67.1	75.6	46.2	68.2
AIT [6]	47.7	62.7	71.9	76.1	51.8	63.5	31.5	**70.3**	84.0	**87.2**	81.2	80.8	**84.5**	72.2	78.7	62.8	69.2	**86.6**	**74.3**	**83.7**	47.7	**73.1**
SaFT (Ours)	**59.7**	**81.3**	**82.4**	**86.9**	**73.0**	**72.0**	**62.3**	**83.7**	**85.9**	**88.1**	**86.7**	**87.7**	**87.7**	**83.5**	**86.1**	**75.1**	**80.1**	**88.1**	**77.0**	**84.3**	**48.5**	**74.5**

Table 1. Experimental results on the VOC 2007 test set in terms of AP50 (%). We evaluate performances of our SaFT over multiple random runs. **RED**/**BLUE** indicate SOTA/the second best, the same below. The superscript * indicates results reproduced under the OSD setting.

Method / Split	Base					Novel				
	1	2	3	4	Avg.	1	2	3	4	Avg.
CoAE [21]	42.2	40.2	39.9	41.3	40.9	23.4	23.6	20.5	20.4	22.0
AIT [6]	**50.1**	**47.2**	**45.8**	**46.9**	**47.5**	**26.0**	**26.4**	**22.3**	**22.6**	**24.3**
SaFT (Ours)	**49.2**	**47.2**	**47.9**	**49.0**	**48.3**	**27.8**	**27.6**	**21.0**	**23.0**	**24.9**

Table 2. Experimental results on the COCO 2017 val set in terms of AP50 (%). Our results are averaged over multiple runs.

Implementation Details. Our approach employs FCOS [50] as our base detector with ResNet-101 [20] pre-trained on ImageNet [45] as backbone. VFM outputs $\{\tilde{F}_4^Q, \tilde{F}_5^Q, \tilde{F}_6^Q\}$ with strides $\{16, 32, 64\}$ from the query while only the intermediate $\tilde{F}_5^S$ from the support for semantic-aligned fusion. HFM iterates $N = 6$ two-way HA blocks. To optimize our network, we use SGD with a mini-batch size of 8, momentum of 0.9 and weight decay of $1e-4$ on both PASCAL-VOC and MS-COCO datasets without online fine-tuning.

4.2. Comparison Results

PASCAL-VOC. We provide performance comparison with current state-of-the-art on VOC in Tab. 1. The first two rows show results we reproduce with FSD methods under the OSD setting. Our SaFT consistently outperforms existing approaches, which demonstrates its effectiveness. We achieve around 5.2% and 1.4% improvements over the best method in base and novel classes respectively. To be specific, we observe a huge upswing in some categories, e.g., Chair with 9.9% and Person with 13.4%. One possible reason might be that objects in these classes have larger diversity in shape and size. Our method aligns query-support semantic units more effectively, thereby favoring these cases. It is also noteworthy that, among all the listed methods, our model is the only one adopting a one-stage framework.

MS-COCO. Similarly, we report evaluation results of COCO on four different splits in Tab. 2. In spite of the challenge of COCO, our model achieves 24.9% novel AP50, better than all existing methods. We further notice that performances on base classes and those on novel classes are not necessarily positively correlated. For example, although SaFT produces relatively low results on the first split of base classes, it brings a 1.8% increase on novel classes over the current SOTA, demonstrating the strong generalization of our approach.

Cross-scale	Cross-sample	Base	Novel
w/o	Reweighting [24]	61.8	53.7
w/o	Correlation [61]	64.1	54.2
w/o	HFM	74.6	65.8
FPN [31]	Reweighting	72.3	62.9
FPN	Correlation	76.6	61.6
FPN	HFM	79.6	69.2
VFM	Reweighting	72.8	64.2
VFM	Correlation	77.7	64.3
VFM	HFM	79.5	71.7

Table 3. Ablation study for different modules of SaFT on VOC. Cross-scale and cross-sample are feature fusion techniques that interact among scales and between samples accordingly.

4.3. Ablation Study

We investigate into the effectiveness of various components of our proposed SaFT. Presented in Tabs. 3 and 4, all relative ablations are conducted on the VOC07 test set with half the batch size as our main experiment and less iterative HA blocks ($N = 4$). Single-scale implementations in rows 1-3 of Tab. 3 utilize Res-4 feature by default, while the rest multi-scale ones use level-4,5,6 features as SaFT.

Impact of different modules. In Tab. 3, reweighting [24] and correlation [61] are borrowed for substitution with HFM as our baseline cross-sample fusion operations. We adopt a 5 × 5 kernel for correlation and simply pool it to 1 × 1 to produce a reweighting prototype. As for cross-scale fusion, we implement pipelines without VFM in rows 1-3 as our baseline and add FPN in rows 4-6 for comparison. Upon these, we go through three phases to finish our exploration of SaFT. (1) Employ HFM for cross-sample fusion. Comparing the results in rows 1-3, we find that HFM improves by 10.5% $\sim$ 12.8% and 11.6% $\sim$ 12.1% respectively on base and novel classes. Similar conclusions can be drawn with respect to rows 4-6, showing the superiority of HFM in modeling cross-sample relatedness. With the help of the attention mechanism, HFM deconstructs support features so that they can be matched with the query deformably rather than always as a whole. In this way, semantics from both sides are more aligned. (2) Use FPN for cross-scale fusion. Before looking into VFM, we first extract multi-scale features with FPN, thereby mitigating the problem

Cross-scale	Cross-sample	Query			Support			Base	Novel
		4	5	6	4	5	6		
FPN [31]	Corresponding-scale	✓	✓	✓	✓	✓	✓	77.4	68.7
FPN	One-to-all-scale	✓	✓	✓		✓		79.6	69.2
VFM	Corresponding-scale	✓	✓	✓	✓	✓	✓	77.5	69.9
VFM	One-to-all-scale	✓	✓	✓		✓		79.5	71.7

Table 4. Ablation study for query-support fusion scales on VOC. One-to-all-scale means associating a single level of support features with all available query features, whereas corresponding-scale is limited to corresponding levels. All experiments use HFM for cross-sample fusion while with different corresponding rules.

of scale variation. Besides the 3.4% ∼ 9.2% performance boost on novel classes with FPN, we notice an interesting phenomenon. Cross-sample fusion approaches with larger receptive fields in the query gain more in single-scale performances, while FPNs attached upon them improve relatively less. This comes not only from the higher baselines they are on, but also from the aggregation of fine-to-coarse features that potentially increase the receptive field of lower levels. (3) Adopt VFM for semantic-aligned fusion. Comparison between rows 6 and 9 presents a 2.5% promotion by replacing FPN with VFM, as a result of the more adaptive VA in VFM that collects and supplements semantics globally rather than locally. Baseline results also grow with 1.3% ∼ 2.7% from rows 4-5 to rows 7-8. Finally, VFM and HFM collaborates in semantic-aligned attention and obtains a 17.5% ∼ 18.0% soar, proving their effectiveness.

Impact of query-support fusion scales. We examine the way to utilize different levels of features for fusion in Tab. 4. To begin with, we propose a simple corresponding-scale strategy that fuses each level of support feature with the same level of query feature. Its results are shown in the first and the third row. While with additional levels of support features participating in query-support fusion, its results turn out relatively lower. This counter-intuitive result might be due to a lack of generality in multi-level support features. Corresponding-scale fusion forces single-level fusion and detection, potentially causing multi-level connections to degrade by over-fitting to each level. In fact, a common down-sample ratio for each corresponding level of query-support feature means always searching targets of the same size. This not only undermines semantic alignments but also confuses multi-scale detection. Differently, since one-to-all-scale fusion matches the support with different sizes of targets on multiple scales, it learns more generalized meta-knowledge. This leads to 0.5% ∼ 1.8% gains in novel classes after shifting from the corresponding-scale strategy. Also, a comparison between rows 2 and 3 shows that VFM with corresponding-scale query-support fusion outperforms FPN with a one-to-all-scale one. We attribute this to the cross-scale long-range correlation modeled by VFM, which mitigates cross-scale semantic misalignment.

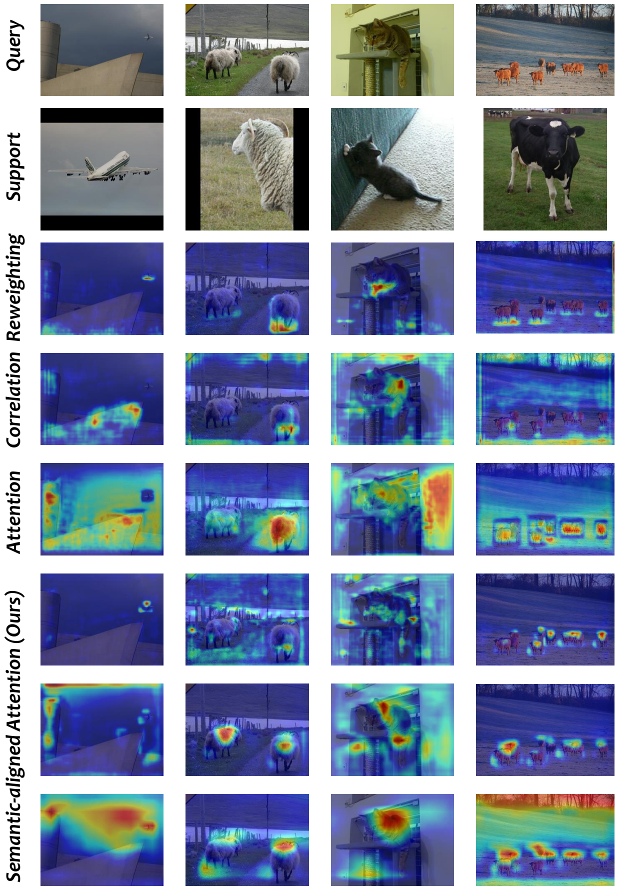

Figure 6. **Feature map responses of different fusion approaches.** Vertically, the four columns are four different classes in the VOC07 test set. Horizontally, the top two rows are query and support samples for fusion, while the rest are feature map responses based on four fusion paradigms. Rows 3, 4, 5 show visualizations for three single-scale implementations. The last three lines demonstrate results of feature levels 4, 5, and 6 for our SaFT with VFM and HFM. These four configurations are respectively corresponding to rows 1-3 and 9 in Tab. 3.

4.4. Qualitative Analysis

Feature map responses of different fusion approaches. To explore the behaviors of four fusion paradigms introduced in Fig. 2, we visualize their feature map responses in Fig. 6. These heatmaps are produced by averaging fusion features in all channels. Comparing row 3 with row 4, we can see that convolution-based methods tend to focus on objects that conform to their kernel sizes. Reweighting with a 1 × 1 kernel recognizes the tiny airplane in the first column while correlation with a 5 × 5 one hardly not. In larger objects, such as the cat in the third column, the heatmap of correlation activates more intensely than that of reweighting. Moreover, most related objects are partially focused due to the query-support mismatch. In contrast, attention-based HFM distributes its attention to more complete object regions. Yet, it activates some unrelated areas as well, especially in cases like the first and third columns, where the query shares a similar background with the support. We attribute the former to scale issues and the latter to the adaptive instinct of attention mechanisms, sometimes misleading. Generally, these conventional techniques are restricted in tackling appearance and scale variations, which leads to their activation biases compared with semantic-aligned attention. Illustrated in the three bottom rows, objects of distinct scales are activated differently in multi-scale heatmaps. As targets grow in size, fusion levels of interest shift from low to high. Clearly shown in the figure, the small-scale airplane, middle-scale sheep, and large-scale cat are highlighted in level 4, 5, and 6 heatmaps, accordingly. As a result, the attention mechanism works in tandem with the semantic-aligned fashion to address feature misalignment as well as scale discrepancies.

Detection results of different fusion approaches. We demonstrate detection results in Fig. 7 for a more intuitive comparison. To be concrete, we explore the figure column by column. The first column clearly shows the differences between convolution-based techniques and attention-based ones. Matching the support patch in local areas of the query, reweighting and correlation are only capable of capturing bounding boxes that cover part of the object. In contrast, attention and semantic-aligned attention produce more accurate results, as they obtain a more global view of the query. Additionally, we can find more false positives in the results of traditional schemes. This is because their misalignment in scale wrongly focuses them on irrelevances. Similar phenomena can be seen in the last column, where a human is recognized as a cow in rows 3-5. From top to bottom, networks successively give higher certainties to the cow and lower to the human. This also proves that space and scale alignments help in query-support associations. In the second column, we examine the ability of different approaches in resolving multi-scale issues. It is worth noticing that rows 3-5 show a downward trend in handling scale variances. Not difficult to understand, the discrepancy between reweighting and correlation is due to the inductive bias of their kernel sizes. With a smaller kernel size, reweighting is better at detecting smaller objects. In terms of attention, it fails to resolve scale variations solely dependent on single-scale matching. Thus, HFM in conjunction with VFM forming semantic-aligned attention presents better predictions. Next, we find an interesting anomaly in the third column. Despite the success of others, reweighting treats all humans as cats while ignoring the real cat. We attribute this to potential semantic conflicts in the support. As in this example, there is a human baby and a cat in the support patch. Then the compression of spatial structure is likely to distract detectors from the never-before-seen cat to the more familiar human that lies in base classes. By contrast, preserving structural information with larger convolution kernels or matching deformably with attention mechanisms remedies this.

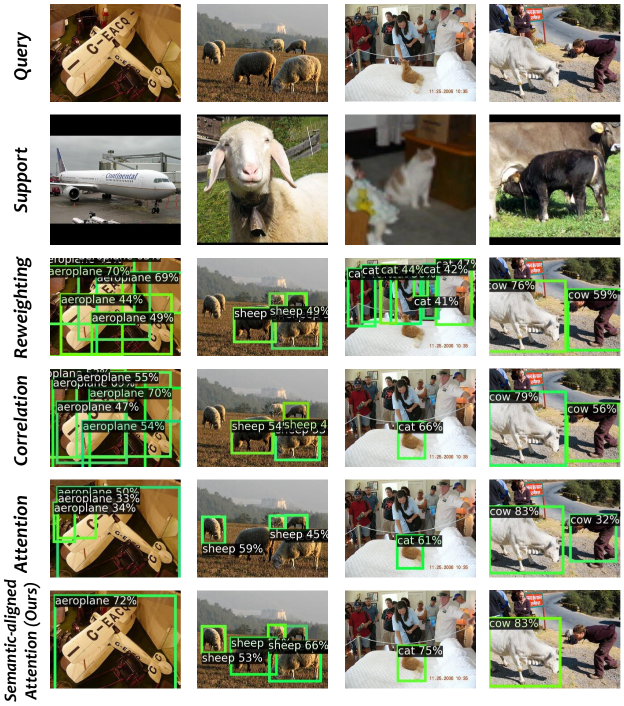

Figure 7. **Detection results of different fusion approaches.** Four columns show original samples and detection results for each novel class in the VOC07 test set. Results in rows 3-6 are of configurations as rows 1-3 (previous schemes) and 9 (our novel scheme) in Tab. 3.

5. Limitation Discussion

As an OSD-specific pipeline, our approach cannot be easily extended to more-shot circumstances. The Siamese design limits its inputs to a paired form, necessitating the construction of a dedicated feature extractor and aggregator for multiple support instances. Also, we notice that our model needs a small learning rate and a long schedule to converge. Hence, we adopt a learning rate of 0.001 instead of the common 0.02. Larger learning rates may result in instability in training. Future work can improve in these aspects.

6. Conclusion

In this paper, we look into the one-shot object detection task from the perspective of space and scale. The proposed Semantic-aligned Fusion Transformer substantially alleviates the underlying query-support misalignment in traditional feature fusion techniques employed by existing schemes. Despite its intuitiveness, our model achieves state-of-the-art performances on various benchmarks.

References

[1] Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. Few-shot text classification with distributional signatures. In *ICLR*, 2020. 1

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *ICCV*, pages 3286–3295, 2019. 3

[3] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, volume 33, pages 1877–1901, 2020. 1

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018. 3

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 3, 4

[6] Ding-Jie Chen, He-Yen Hsieh, and Tyng-Luh Liu. Adaptive image transformer for one-shot object detection. In *CVPR*, pages 12247–12256, 2021. 2, 3, 5, 6

[7] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. Lstd: A low-shot transfer detector for object detection. In *AAAI*, volume 32, 2018. 1, 3

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 3

[9] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *CVPR*, pages 1601–1610, 2021. 3

[10] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. *NeurIPS*, 33:21981–21993, 2020. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[12] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, pages 6569–6578, 2019. 3

[13] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 5

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 5

[15] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. In *CVPR*, pages 4013–4022, 2020. 1, 3

[16] Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. Generalized few-shot object detection without forgetting. In *CVPR*, pages 4527–4536, 2021. 1, 3

[17] Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. Induction networks for few-shot text classification. In *EMNLP/IJCNLP*, 2019. 1

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *PAMI*, 37(9):1904–1916, 2015. 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, June 2016. 6

[21] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. *NeurIPS*, 32, 2019. 1, 3, 5, 6

[22] Hanzhe Hu, Shuai Bai, Aoxue Li, Jinshi Cui, and Liwei Wang. Dense relation distillation with context-aware aggregation for few-shot object detection. In *CVPR*, pages 10185–10194, 2021. 2, 3

[23] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. 1

[24] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *ICCV*, pages 8420–8429, 2019. 1, 3, 5, 6

[25] Leonid Karlinsky, Joseph Shtok, Sivan Harary, Eli Schwartz, Amit Aides, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Repmet: Representative-based metric learning for classification and few-shot object detection. In *CVPR*, pages 5197–5206, 2019. 3

[26] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, pages 734–750, 2018. 3

[27] Aoxue Li and Zhenguo Li. Transformation invariant few-shot object detection. In *CVPR*, pages 3094–3102, 2021. 1, 3

[28] Bohao Li, Boyu Yang, Chang Liu, Feng Liu, Rongrong Ji, and Qixiang Ye. Beyond max-margin: Class margin equilibrium for few-shot object detection. In *CVPR*, pages 7363–7372, 2021. 1, 3

[29] Yiting Li, Haiyue Zhu, Yu Cheng, Wenxin Wang, Chek Sing Teo, Cheng Xiang, Prahlad Vadakkepat, and Tong Heng Lee. Few-shot object detection via classification refinement and distractor retreatment. In *CVPR*, pages 15395–15403, 2021. 3

[30] Di Lin, Dingguo Shen, Siting Shen, Yuanfeng Ji, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Zigzagnet: Fusing top-down and bottom-up context for object segmentation. In *CVPR*, pages 7490–7499, 2019. 3

[31] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 3, 6, 7

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 3

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 5

[34] Weidong Lin, Yuyan Deng, Yang Gao, Ning Wang, Jinghao Zhou, Lingqiao Liu, Lei Zhang, and Peng Wang. Cat: Cross-attention transformer for one-shot object detection. *arXiv preprint arXiv:2104.14984*, 2021. 3

[35] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, pages 8759–8768, 2018. 3

[36] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37, 2016. 3

[37] Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and Tatsuya Harada. Goal-oriented gaze estimation for zero-shot learning. In *CVPR*, pages 3794–3803, 2021. 1

[38] Anton Osokin, Denis Sumin, and Vasily Lomakin. Os2d: One-stage one-shot object detection by matching anchor features. In *ECCV*, pages 635–652, 2020. 1, 3

[39] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, pages 4055–4064, 2018. 3

[40] Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. Incremental few-shot object detection. In *CVPR*, pages 13846–13855, 2020. 1

[41] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *CVPR*, pages 475–484, 2021. 3

[42] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *ICCV*, pages 8681–8690, 2021. 3, 6

[43] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 3

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28:91–99, 2015. 3

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 6

[46] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *EACL*, 2021. 1

[47] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, volume 30, 2017. 1

[48] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *CVPR*, pages 7352–7362, 2021. 1, 3, 6

[49] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, pages 1199–1208, 2018. 1, 3

[50] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019. 3, 6

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 3

[52] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *NeurIPS*, 29, 2016. 1, 3

[53] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 3

[54] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *ICML*, July 2020. 1, 3, 5

[55] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *ECCV*, pages 456–472, 2020. 1, 3

[56] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In *ECCV*, pages 192–210, 2020. 1, 3, 5

[57] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *ICCV*, pages 9577–9586, 2019. 1, 3

[58] Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. Diverse few-shot text classification with multiple metrics. In *NAACL*, pages 1206–1215, June 2018. 1

[59] Dong Zhang, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, and Qianru Sun. Feature pyramid transformer. In *ECCV*, pages 323–339, 2020. 3

[60] Gongjie Zhang, Kaiwen Cui, Rongliang Wu, Shijian Lu, and Yonghong Tian. Pnpdet: Efficient few-shot detection without forgetting via plug-and-play sub-networks. In *WACV*, pages 3823–3832, 2021. 3

[61] Lu Zhang, Shuigeng Zhou, Jihong Guan, and Ji Zhang. Accurate few-shot object detection with support-query mutual guidance and hybrid loss. In *CVPR*, pages 14424–14432, 2021. 1, 3, 6

[62] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *CVPR*, pages 9759–9768, 2020. 3

[63] Weilin Zhang and Yu-Xiong Wang. Hallucination improves few-shot object detection. In *CVPR*, pages 13008–13017, 2021. 3

[64] Gangming Zhao, Weifeng Ge, and Yizhou Yu. Graphfpn: Graph feature pyramid network for object detection. In *ICCV*, pages 2763–2772, 2021. 3

[65] Chenchen Zhu, Fangyi Chen, Uzair Ahmed, Zhiqiang Shen, and Marios Savvides. Semantic relation reasoning for shot-

stable few-shot object detection. In *CVPR*, pages 8782–8791, 2021. 1, 3

[66] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 3