Abstract—While recent progress has significantly boosted few-shot classification (FSC) performance, few-shot object detection (FSOD) remains challenging for modern learning systems. Existing FSOD systems follow FSC approaches, ignoring critical issues such as spatial variability and uncertain representations, and consequently result in low performance. Observing this, we propose a novel Dual-Awareness Attention (DAna) mechanism that enables networks to adaptively interpret the given support images. DAna transforms support images into query-position-aware (QPA) features, guiding detection networks precisely by assigning customized support information to each local region of the query. In addition, the proposed DAna component is flexible and adaptable to multiple existing object detection frameworks. By adopting DAna, conventional object detection networks, Faster R-CNN and RetinaNet, which are not designed explicitly for few-shot learning, reach state-of-the-art performance in FSOD tasks. In comparison with previous methods, our model significantly increases the performance by 47% (+6.9 AP), showing remarkable ability under various evaluation settings.

Index Terms—Deep learning, object detection, visual attention, few-shot object detection.

I. INTRODUCTION

Few-shot object detection (FSOD) is a recently emerging and rapidly growing research topic, which has great potential in many real-world applications. Unlike conventional object detectors, few-shot object detectors can be adapted to novel domains with only few annotated data, saving the costly data re-collection and re-training processes whenever the downstream task changes.

However, though considerable effort in recent years has been devoted, existing FSOD methods [1], [2], [3], [4], [5] suffer extremely low performance in comparison with traditional object detectors [6], [7]. In addition, it seems previous methods suffer performance drop not only on the novel domain but also on the base (training) domain [3], [4]. Moreover, by carrying out experiments, we discovered that previous FSOD models based on Faster R-CNN [6] are incapable of reaching the performance that Faster R-CNN can achieve when being evaluated on base classes (see Tab. [1]). Though the models have been trained on a lot of annotated data, they still have difficulty in recognizing the objects they have seen. We therefore assume the modifications associated with few-shot learning applied in prior attempts somehow undermine the ability of detection networks, resulting in limited performance.

To recast traditional object detectors into few-shot object detectors, prior works tend to leverage the methods that have been proved effective at few-shot classification (FSC). The techniques such as building category prototypes [9], [10], ranking similarity between inputs [11], [3] and feature map concatenation [12], [13] are all widely adopted. However, unlike FSC aiming to classify images, FSOD is a much more complicated task requiring to identify and locate objects precisely. In this work, we summarize three potential concerns that could restrict the performance of FSOD (see Fig. [1]). (a) First of all, prior methods [10], [1], [2], [13], [4], [19] performed global pooling on support features to ensure computational efficiency. However, the lack of spatial information would cause difficulty in measuring object-wise correlations. (b) Secondly, convolutional neural networks (CNN) are physically inefficient at modeling varying spatial distributions [15], so the methods using convolution-based attention [16], [13] would suffer the same restriction. (c) Furthermore, previous works [10], [1], [2], [13], [3, 4] took mean features across multiple support images as class-specific representations, which heavily relies on an implicit assumption that the mean features will still be representative in the embedding space.

To verify the aforementioned concerns are worth studying, we conduct a pilot study with the hypothesis: If the spatial information, variability and feature uncertainty are all trivial in FSOD, the choice of support images will not severely affect the performance. In the experiment, the four well-trained FSOD models are tested on the same query set 100 times but the given support images will be different each time. As shown in Fig. [2] even though we fix the query set, previous methods can easily be influenced by the choice of supports, resulting in a huge range of performance (up to 4.4 AP). On the other hand, our method, which solves the concerns summarized in Fig. [1] achieves the highest and firmest performance.

In this work, we present a novel Dual-Awareness Attention (DAna) mechanism comprised of Background Attenuation (BA) and Cross-Image Spatial Attention (CISA) modules. Inspired by wave interference, we propose a BA module where each feature vector $(1 \times 1 \times C)$ of a high-level feature map $(H \times W \times C)$ is viewed as a wave along the channel dimension. By extracting the most representative feature vector from a feature map and adding it back to the feature map, those local features having different wave patterns from the extracted feature will be blurred and therefore can be easily recognized as noise by the detection network. Those foreground features, on the other hand, can maintain the wave patterns and be preserved after the addition. Thus, the BA module not only attenuates irrelevant features but preserves the target informa-

Dual-Awareness Attention for Few-Shot Object Detection

Tung-I Chen¹, Yueh-Cheng Liu¹, Hung-Ting Su¹, Yu-Cheng Chang¹, Yu-Hsiang Lin¹, Jia-Fong Yeh¹, Wen-Chin Chen¹, Winston H. Hsu¹,²
¹ National Taiwan University, ² Mobile Drive Technology
We demonstrate the impact of each proposed component. We also offer a comprehensive ablation study to further evaluate the generalization ability of each model. In Table III, we show our method significantly outperforms the strongest baseline [17] by 6.9 AP under the 30-shot setting. Furthermore, we are the first to test FSOD based evaluations. In Table I, we show our method significantly increases (see Table II).

By better utilizing support images, our method achieves the most significant improvement as the number of support images increases. Also, CISA provides a more efficient and effective way to summarize information among multiple support images. Those QPA vectors conditioned on the same query region would represent relevant information, and therefore taking the mean of corresponding QPA vectors, the model can easily determine whether the regions should be the parts of the target object. By measuring correlations between the query regions and their corresponding QPA vectors, the model can easily determine whether the regions should be the parts of the target object. Also, CISA provides a more effective and effective way to summarize information among multiple support images. Those QPA vectors conditioned on the same query region would represent relevant information, and therefore taking the mean feature across them is more effective than previous manners. By better utilizing support images, our method achieves the most significant improvement as the number of support images increases (see Table II).

In this paper, we evaluate models across various settings, including the multi-shot, multi-way, cross-domain, and episode-based evaluations. In Table I, we show our method significantly outperforms the strongest baseline [17] by 6.9 AP under the 30-shot setting. Furthermore, we are the first to test FSOD models on novel domains without fine-tuning (Table I and Table III) to further evaluate the generalization ability of each method. We also offer a comprehensive ablation study to demonstrate the impact of each proposed component.

Our main contributions can be summarized as follows:

- We point out the critical issues in previous FSOD methods that lead to the limited performance and the lack of robustness.
- We present novel dual-awareness attention to precisely capture object-wise correlations.
- We conduct comprehensive experiments to fairly assess each approach. Extensive experiments manifest the effectiveness and robustness of the proposed methodology.

II. RELATED WORKS

A. Few-Shot Classification

Few-shot classification (FSC) has multiple branches, including the optimization-based and metric-based approaches. The optimization-based methods [18], [19], [20], [21], [22] aim to learn a good initialization parameter set that can swiftly be adapted to new tasks within few gradient steps. The metric-based methods [11], [23], [9], [12], [24], [25], on the other hand, compute the distance between learned representations in the embedding space. The concept of prototypical representation [9] is widely adopted in FSC, which takes the mean feature over different support images as a class-specific embedding. Such a strategy is intuitive yet the data scarcity in few-shot scenarios might lead to biased prototypes and consequently hinder the performance [24]. To enhance the reliability of class representations, Tian et al. [25] leveraged pre-trained embedding and showed that using good representations is more effective than applying sophisticated meta-learning algorithms. Although the attempts have successfully boosted the performance of FSC, the progress of few-shot object detection (FSOD) is still in a very early stage. In this work, we explore novel attention mechanisms and significantly enhance the performance of FSOD.

B. Attention Mechanism

The attention modules were first developed in natural language processing (NLP) to facilitate machine translation [26], [27], [28]. Recently, inspired by the great success of Transformer [29], researchers start to explore the self-attention
deviations of AP of the four methods are images would seriously influence FSOD results. The standard
test whether the noise and spatial variability of support
previous attempts exploited transfer learning [48]
are usually dominant in performance, able performance in many applications. Two-stage detec-
tors [42], [43], [6], [44] are usually dominant in performance,
C. Few-Shot Object Detection
Deep-learning-based object detectors have shown remark-
able performance in many applications. Two-stage detec-
tors [42], [43], [6], [44] are usually dominant in performance, and one-stage detectors [7], [45], [46], [47] are superior in run-
time efficiency. Most modern object detectors are category-
specific, which means they are incapable of recognizing
objects of unseen categories. To explore generalized object
d Detectors, previous attempts exploited transfer learning [48]

and distance metric learning [10] to achieve quick adaption to
novel domains. As recasting object detection problem into the
few-shot learning paradigm, the idea of current methods could
be somewhat similar to multiple-instance learning (MIL) [49],
[50], [51]. For FSOD, we can also perceive a query image as
a bag, aiming to identify the positive image patches in it by
capturing the contexts relevant to given support images [49].
Recently, there is an important line of works encoding
support images into global vectors, measuring the similarity
between feature vectors and the RoI proposals inferred by
detection networks [11], [2]. Following the spirit, [13] perceived
the task as a guided process where support features are used to
guide the object detection networks. Inspired by [16], Fan et
al. [3] measured the correlations by regarding support images
as kernels and performing a convolution-based operation over
queries. Current works have a tendency to take mean features
as class representations and measure cross-image correlations
by either feature concatenation or element-wise product [2],
[10], [13], [3]. We argue that these techniques will engender
serious issues (illustrated in Fig. 1) in FSOD, degenerating
the performance even on the seen (training) categories.

III. METHODOLOGY

A. Problem Definition
Let $s$ be a support image and $S = \{s_i\}_{i=1}^M$ be a support
set that represents a specific category. An individual FSOD
problem can be formulated as $T = \{(S^1, ..., S^N), I\}$, where $I$
is a query image comprised of multiple instances and back-
grounds. Given $T$, the model should detect all the objects in
$I$ belonging to the target categories $\{S^1, ..., S^N\}$. The object
categories in a training dataset are divided into two disjoint
parts: base classes $C_{\text{base}}$ and novel classes $C_{\text{novel}}$. To train a
FSOD model, a meta-training set $H_{\text{train}} = \{T_i\}_{i=1}^K$ should be
constructed, where all the bounding box annotations belong to
$C_{\text{base}}$. Similarly, a meta-testing set $H_{\text{test}}$ is constructed where
all the target objects belong to $C_{\text{novel}}$. It is allowed to use a fine-
tuning set $H_{\text{finetune}}$ to fine-tune the model before evaluating on
$H_{\text{test}}$. However, in an $N$-way K-shot setting, only K box
annotations of each novel category can be used to fine-tune
the model [10], [11].

To summarize, the primary goal of FSOD is to leverage
rich source-domain knowledge in $H_{\text{train}}$ to learn a model
that can swiftly generalize to target domains where only few
annotated data are available. Instead of only considering the
features of $I$, the model $f(I|S)$ is trained to recognize objects
conditioned on the given support information.

B. Dual-Awareness Attention

1) Overview: FSOD relies on limited support information
to detect novel objects, and therefore we consider two im-
portant aspects: (1) The quality of support features, and (2)
how to better construct correlations between support and query
images. In this work, we propose an attention mechanism
comprised of two novel modules to undermine the influence
of noise and precisely measure object-wise correlations.

C. Few-Shot Object Detection

Deep-learning-based object detectors have shown remark-
able performance in many applications. Two-stage detec-
tors [42], [43], [6], [44] are usually dominant in performance, and one-stage detectors [7], [45], [46], [47] are superior in run-
time efficiency. Most modern object detectors are category-
specific, which means they are incapable of recognizing
objects of unseen categories. To explore generalized object
detectors, previous attempts exploited transfer learning [48]

...
Fig. 3: Illustration of our proposed methods. In the BA block, the extracted feature is superimposed on the feature map to undermine irrelevant background features while maintain distinguishing information. In the CISA block, support feature maps are adaptively transformed into query-position-aware (QPA) vectors, which are customized support information for each query region.

2) Background Attenuation Block: It is infeasible to always ensure high-quality support images in real-world scenarios. Those noise in support images will inhibit models from reaching robustness (see Fig 2). We propose a novel mechanism, Background Attenuation (BA), to undermine the irrelevant support information. The detailed structure of BA block is illustrated in Fig. 3 where a support image s and a query image t are encoded into a support feature map \( Y \in \mathbb{R}^{C \times H_s \times W_s} \) and a query feature map \( X \in \mathbb{R}^{C \times H_Q \times W_Q} \) by a shared CNN backbone. In BA block, the feature map \( Y \) will be reshaped and transformed by a linear learnable matrix \( W_e \in \mathbb{R}^{C \times 1} \). The process can be formulated as

\[
A_{BA}(y_i) = \sigma(W_e y_i) = \frac{\exp(W_e y_i)}{\sum_{j \in \Omega} \exp(W_e y_j)} \tag{1}
\]

where \( y_i \in \mathbb{R}^{1 \times C} \) denotes the feature vector at \( i \)-th pixel of \( Y \); \( \Omega \) is the set of all pixel indices and \( \sigma \) is the softmax function applied along the spatial dimension. Thus, the learned aggregation of \( Y \) can be obtained by

\[
G = \sum_{i \in \Omega} A_{BA}(y_i) \cdot y_i. \tag{2}
\]

Intuitively, \( G \in \mathbb{R}^{1 \times C} \) should represent the most important feature of \( Y \) according to the resulting attention maps. However, we empirically discover that using the learned attention maps to filter out noise would harm the performance (see Tab. VI). By visualization, we observe that only few narrow regions of support image contribute to the aggregated feature \( G \). Such a naive attention process leads to a considerable loss of support information and therefore deteriorate the ability of models.

Consequently, we propose a much softer attention strategy to remove noise. In physics, interference is a term indicating two signals superpose to form a resultant signal of greater or lower amplitudes. Inspired by that, we view each feature vector of \( Y \) as a \( C \)-dimensional signal and superimpose the extracted feature \( G \) on those signals

\[
Z = Y + \alpha \cdot \text{LeakyReLU}(G) \tag{3}
\]

where \( \alpha \) is a constant hyper-parameter, and the nonlinear function is used to rectify the results. As illustrated in Fig. 2 if the original signals (e.g., regions belonging to the dog) are aligned with \( G \) along the channel dimension, the features can be enhanced or maintained. On the contrary, those signals associated with unrelated regions (e.g., the Frisbee) will be blurred after addition due to the significant difference with \( G \). Thus, the BA block can benefit detection networks by providing more discriminative support features. It is worth mentioning the idea of BA is different from the channel-wise attention [33, 32] which re-weights feature maps along the channel dimension. In addition, unlike [33] constructing heavy linear transformation matrices to re-weight semantic dependencies, BA block is more efficient since \( W_e \in \mathbb{R}^{C \times 1} \) is the only learnable weights in it.

3) Cross-Image Spatial Attention Block: Since even the intra-class objects would have obvious deviation in appearance, the model should learn to focus on those most representative parts of the objects to determine the similarity among them. The core idea of Cross-Image Spatial Attention (CISA) is to adaptively transform each support feature map into query-position-aware (QPA) support vectors that represent specific information of a support image. The first step of CISA is
Fig. 5: The model architectures of DAnA-FasterRCNN and DAnA-RetinaNet. Our proposed components are flexible and can be easily combined with existing object detection frameworks which are not originally designed for FSOD.

similar to QKV attention [29], transforming X and Z into the query and key embeddings \( Q = W_q X, K = W_k Z \) by learned weight matrices \( W_q, W_k \in \mathbb{R}^{C \times C'} \) respectively. The similarity scores between the query and support can then be measured by

\[
\delta(X, Z) = \sigma((Q - \mu_Q)^\top(K - \mu_K))
\]

(4)

where \( \mu_Q \) and \( \mu_K \) are the averaged embedding values over all pixels; the softmax function \( \sigma \) is performed over the spatial dimension. Furthermore, we add a simplified self-attention term [33] in CISA because we assume the attention should be based on not only query-support correlations but also the support image itself. Thus, the CISA attention function is formulated as

\[
A_{\text{CISA}}(X, Z) = \delta(X, Z) + \beta \cdot W_i Z
\]

(5)

where \( \beta \) is a constant coefficient; \( W_i \in \mathbb{R}^{C 	imes 1} \). Note that the output of \( A_{\text{CISA}}(X, Z) \) has the shape of \( H_0W_Q \times H_3W_S \), which indicates the fact that we have obtained multiple support attention maps \( (H_3W_S) \) conditioned on each spatial location of the query feature map \( (H_0W_Q) \). The QPA vectors can therefore be obtained by

\[
p_i = \sum_{j \in \Omega} A_{\text{CISA}}(X, Z)_{ij} z_j
\]

(6)

where \( \Omega \) denotes all pixel indices of the support feature map. The equation shows that all the vectors of \( Z \) are adaptively weighted and aggregated into a vector \( p_i \in \mathbb{R}^{1 	imes C} \) according to each query position \( i \) (resolve Fig. 1 (a)). Additionally, if there are \( K \)-shot images available in a support set \( \{s\}^K \), we can perform average pooling across the resulting QPA vectors

\[
p_i = \frac{1}{K} \sum_{k=1}^{K} p_{i,k}
\]

(7)

without the concern of uncertainty, because those QPA vectors \( p_{i,k} \) conditioned on the same entry \( i \) have been refined and will carry relevant features (resolve Fig. 1 (c)). Furthermore, it breaks the physical restriction of CNN since all the query features are aligned with their customized QPA vectors (resolve Fig. 1 (b)).

C. Model Architecture

To apply the DAnA component to existing object detection frameworks, we can simply combine the output \( P \in \mathbb{R}^{C \times H_0 \times W_Q} \) with \( X \) into \( PX \in \mathbb{R}^{2C \times H_0 \times W_Q} \) and send it to the modules such as the region proposal network (RPN) to propose regions having high responses with the given supports (see Fig. 3). In our experiments, we choose Faster R-CNN [6] and RetinaNet [7] as the backbones to verify the effectiveness of the DAnA mechanism. For DAnA-FasterRCNN, one CISA block is employed before RPN and the other is applied in the second stage taking cropped RoI features as inputs. For DAnA-RetinaNet, we apply both BA and CISA modules to each level of the feature pyramids.

In addition, we follow the modification applied in [3], replacing the multi-class classification output with the binary one. Since the pipeline of FSOD can be regarded as a matching process based on given supports, we suggest the binary output could better fit the problem scenarios where an instance will be either positively or negatively labeled. The rest details of the proposed models remain the same as the original Faster R-CNN and RetinaNet.

IV. EXPERIMENTS

A. Implementation Details

By default, we employ the pretrained ResNet50 as the feature extractor, and the batch size is set to 32 for all the experiments. All the models including baselines were implemented in Pytorch [52] on a workstation with 4 NVIDIA Tesla V100. The shorter side of query images is resized to 600 pixels, while the longer side is cropped at 1000. Each support image will be zero-padded and then resized to a square image of 320 \times 320. The embedding dimension \( C \) in the weight matrices \( W_q \) and \( W_k \) is a quarter of the number of original feature channels \( C \). The constant coefficients \( \alpha \) and \( \beta \) of the proposed modules are 0.5 and 0.1 individually. For those

| Method                          | 10shot | 30shot |
|---------------------------------|--------|--------|
|                                | AP     | AP_{50} | AP     | AP_{50} |
| TFA w/cos                      | 10.0   | 9.3     | 13.7   | 13.4    |
| Feature Reweighting            | 5.6    | 4.6     | 9.1    | 7.6     |
| Meta R-CNN                     | 8.7    | 6.6     | 12.4   | 10.8    |
| Attention RPN                  | 11.1   | 10.6    | -      | -       |
| IFSoD                          | 5.1    | -       | -      | -       |
| MPSR                           | 9.8    | 9.7     | 14.1   | 14.2    |
| Viewpoint Estimation           | 12.5   | 9.8     | 14.7   | 12.2    |
| DAnA-FasterRCNN (Ours)         | 18.6   | 17.2    | 21.6   | 20.3    |

TABLE 1: The performance on novel categories of COCO. All the models are trained on base categories and then fine-tuned on a small set of novel samples (e.g., 10 shots of each category). After fine-tuning, models will be evaluated on the novel classes. “-” : no reported results.
models based on Faster R-CNN [6], we adopt SGD with an initial learning rate of 0.001, which decays into 0.0001 after 12 epochs; the four anchor scales are [60^2, 120^2, 240^2, 480^2] and the three aspect ratios are [0.5, 1.0, 2.0]; the momentum and weight decay coefficients are set to 0.9 and 0.0005. For DAnA-RetinaNet, we construct a pyramid with levels P3 through P7 (P has resolution 2^i lower than the input) and adopt Adam with a learning rate of 0.00001. We adopt the two-way contrastive training strategy, which is first proposed in [3], to train the baselines and our models.

### Table II: The 1-way, zero-shot evaluation on COCO. All the few-shot models are trained on base categories and then tested on novel domains without fine-tuning. Note that the performance of Faster R-CNN [6] on base classes can serve as the upper bound because all the methods in this table leverage [6] as their backbone detector. In comparison with baselines, the relative mAP improvement of our method is up to 49% on novel categories, and the performance gap with the traditional object detector has been reduced. The model size and inference speed are reported as well. †: re-implemented results.

| Method                  | AP | AP50 | AP75 | AP | AP50 | AP75 | # parameters | FPS     |
|-------------------------|----|------|------|----|------|------|--------------|---------|
| **Novel Categories**    |    |      |      |    |      |      |              |         |
| Faster R-CNN†           | N/A | N/A  | N/A  | N/A | N/A  | N/A  | 4.76 x 10^1  | 31      |
| Meta R-CNN†             | 1shot | 1shot | 1shot | 1shot | 1shot | 1shot | 50.4 | 52.5 | 52.5 | 27.3 | 28.4 | 28.2 | 4.76 x 10^1  | 28      |
| FGNet† [13]             | 8.7 | 11.1 | 11.2 | 19.9 | 25.3 | 25.9 | 27.3 | 28.6 | 28.5 | 50.4 | 52.5 | 52.5 | 27.3 | 28.4 | 28.2 | 4.76 x 10^1  | 28      |
| Attention RPN† [13]     | 8.7 | 10.1 | 10.6 | 19.8 | 23.0 | 24.4 | 20.6 | 22.4 | 23.0 | 37.2 | 40.8 | 42.0 | 20.5 | 22.2 | 22.4 | 1.03 x 10^1  | 21      |
| DAnA-FasterRCNN         | 11.9 | 14.0 | 14.4 | 25.6 | 28.9 | 30.4 | 27.8 | 29.4 | 32.0 | 46.3 | 50.6 | 54.1 | 27.7 | 30.3 | 32.9 | 1.42 x 10^1  | 24      |

| **Base Categories**     |    |      |      |    |      |      |              |         |
| AP                      | 32.0 | 31.0 | 29.5 | 24.4 | 22.8 | 20.8 |            |         |
| AP50                    | 24.4 | 22.8 | 20.8 | 20.5 | 22.2 | 22.4 |            |         |
| AP75                    | 19.8 | 23.0 | 24.4 | 20.5 | 22.2 | 22.4 |            |         |

### Table III: The multi-way, zero-shot evaluation on COCO. Under the N-way setting, we introduce a support set comprised of N categories to the model and therefore the model should detect the objects belonging to any of these N categories.

### Fig. 6: Visualization of the AP on each base (training) category of COCO. A well-developed few-shot object detector should be comparable to the traditional object detector on the training categories if it has not been fine-tuned. This figure demonstrates our model indeed achieves competitive performance on training categories and even surpasses the traditional one on particular classes.

We define the 20 COCO categories intersecting with PASCAL VOC [54] as the novel classes, while the rest 60 categories covered by COCO but not VOC to be the base classes.

2) **FSOD Training:** Following the paradigm of meta-learning, the training data are organized into episodes. Each episode contains one or more support sets, a query image and corresponding ground truth annotations. Unlike the queries in FSC, a query image in FSOD might contain multiple objects of different categories. Therefore, in each training step we will choose one of the object categories as the target class. A support set of the target class will be offered and the model should learn to detect the target objects conditioned on the given support information. Note that all the bounding box annotations belonging to the novel classes have been removed in order the prevent models from peeking at the novel task.

3) **FSOD Evaluation:** Generally, N-way K-shot in the few-shot paradigm indicates that we can use K images of each novel category to adapt the models before testing. However,
we observe previous works adopted different evaluation settings and fine-tuning protocols, leading to the ambiguity in evaluation. For instance, since a query in FSOD might consist of multiple objects of different categories, Kang et al. [1] defined that there should be only $K$ box annotations of each category in the fine-tuning set, while Chen et al. [48] adopted the protocol of collecting $K$ query images of each category to fine-tune their models. To thoroughly evaluate the proposed method, in Tab. I we follow the widely-used protocol in [1] and compare our results with those works adopting the same setting. Furthermore, we carry out additional experiments under our proposed zero-shot protocol (Tab. II III IV V), where we do NOT leverage any annotated data of the novel classes to fine-tune models. We expect a well-developed few-shot object detector can deal with novel objects as long as few support images are given, reducing the costly data re-collection and annotations processes.

### C. Generic FSOD Protocol

The performance of each method under the general FSOD protocol is reported in Tab. I. In this table, we comply with the evaluation protocol widely used in previous works and compare our results with the numbers reported in previous papers. As we have discussed in Sec. IV-B1, the models will not be exposed to the annotations of the 20 novel classes during training. After training, there will be $K = 10, 30$ bounding box annotations per novel class can be used to fine-tune the models. Under such a challenging setting, Tab. I shows our method significantly outperforms the other baselines. By equipping with DAnA, Faster R-CNN conspicuously outperforms the strongest baseline [17] by 6.1 and 6.9 AP under the 10-shot and 30-shot settings respectively.

### D. Zero-shot FSOD Protocol

In this section, we evaluate the models under a more challenging setting without fine-tuning. Note that the word “zero-shot” here is not referred to the zero-shot learning paradigm [55] but referred to the fact that we do not take any novel image to fine-tune the models. In order to evaluate each method under an unified protocol, we re-implement three previous methods: Meta R-CNN [2], FGN [13] and Attention RPN [3]. These baselines and our model are all based on Faster R-CNN, and therefore we can fairly analyze the impact brought by different attention mechanisms. Note that we have slight modifications on these methods: 1) For the model architectures, we replace the multi-class classification output with the binary one as explained in III-C. 2) For training, we apply the two-way contrastive strategy [3] to each method since we empirically observe it can improve the performance. In addition, as we discussed in Sec. IV-B3 all the methods are evaluated without fine-tuning under such a protocol.

Tab. II shows the results predicted with different numbers of support images. With 5-shot support images given at inference, DAnA-FasterRCNN outperforms FGN and Attention RPN by 3.5/6.4/4.0 and 3.8/6.0/4.7 mAP respectively on $AP/AR_{50}/AR_{5}$ metrics. Moreover, the proposed model achieves the most significant improvement as the shot increases, supporting our argument that DAnA can better retrieve the information carried within support images by leveraging QPA features.

Intuitively, without being fine-tuned on the novel domain, a well-developed FSOD model should perform well on base classes. In order to evaluate to what extent the ability of detection networks has been undermined by the modifications, we also provide the performance of Faster R-CNN [6], which is the backbone of these few-shot object detectors. The Faster R-CNN is trained on the 60 base classes and has not been exposed to the novel categories. The performance of it can be regarded as the upper bound of performance on base classes. By precisely capturing object-wise correlations, our model has a much smaller gap with the upper bound comparing with baselines. Additionally, we report the model sizes and inference speed in Tab. II as well. The proposed DAnA component only causes little additional cost in comparison to prior approaches.

In Tab. III we further evaluate each method under the multi-way evaluation setting. Under the multi-way setting, models must detect more than one object category conditioned on the given support set. Suppose there are $M$ categories in a query image, we will randomly sample $N$ classes from them to form an $N$-way support set. Sometimes the number of object classes in a query image will be less than $N$, and we just select other

---

**TABLE IV:** The results of PASCAL2COCO evaluation. All the models are trained on PASCAL VOC 2007 and tested on COCO 2014. In this setting, the base domain denotes the 20 classes shared between PASCAL VOC and COCO, while the novel domain denotes the other 60 classes in the COCO dataset.

| Method                  | Novel Categories | Base Categories |
|-------------------------|------------------|-----------------|
|                         | AP               | AP             |
|                         | # Way            | 1way 3shot 5shot | 1way 3shot 5shot |
| # Given Supports        | 1shot 3shot 5shot | 1shot 3shot 5shot |
| Meta R-CNN† [2]         | 7.8 9.0 9.3      | 33.7 34.8 35.1  |
| FGN† [13]               | 6.7 7.4 7.7      | 30.5 31.2 32.1  |
| Attention RPN† [3]      | 6.8 7.4 7.7      | 23.7 26.6 27.2  |
| DAnA-RetinaNet          | 9.1 10.3 11.5    | 36.5 36.9 37.2  |
| DAnA-FasterRCNN         | 10.0 11.6 11.9   | 36.5 38.1 38.6  |

**TABLE V:** Following [10], we construct class-balanced episodes from COCO for evaluation. In addition, we report the performance of two different detection networks adopting the proposed DAnA component. Even though the performance of DAnA-RetinaNet is less remarkable than DAnA-FasterRCNN, it still significantly outperforms the baselines which are based on Faster R-CNN, showing the effectiveness of our proposed attention mechanism.

| Method                  | Novel Categories | Base Categories |
|-------------------------|------------------|-----------------|
|                         | AP               | AP             |
|                         | # Way            | 1way 3shot 5shot | 1way 3shot 5shot |
| # Given Supports        | 1shot 3shot 5shot | 1shot 3shot 5shot |
| Meta R-CNN† [2]         | 14.7 17.0 17.4   | 13.1 15.0 15.2  |
| FGN† [13]               | 14.9 16.7 17.9   | 13.8 15.5 16.8  |
| Attention RPN† [3]      | 15.0 17.1 18.1   | 13.7 15.4 16.2  |
| DAnA-RetinaNet          | 16.6 18.8 19.5   | 15.5 17.7 18.4  |
| DAnA-FasterRCNN         | 17.9 21.3 21.6   | 17.0 20.2 20.4  |
presenting the advantages of DAnA, which is verified by the
To conclude, we suggest RPN plays a deep role in fully
object locations. Furthermore, in FSOD, those foregrounds
networks must tackle a nearly exhaustive list of potential
does not help DAnA-RetinaNet reach better results. The
On the other hand, it seems the use of feature pyramids
performs previous methods using two-stage networks, which
RetinaNet is based on the one-stage object detector, it still out-
work (RPN) have higher performance than the one-stage [1],
detectors [8], [2], [3], [5], [17] with region proposal net-
has never been trained to recognize these objects. However,
instances in Fig. 7 belong to the novel classes, so the model
category of the given support set. Notably, all the target
instances in Fig. 7. In FSOD, the prediction is dependent on the object
for example, given the head (colored in red) or the feet of
semantic correspondence between the query and support. Take
recognizing and locating the instances. The last row of Fig. 7
presents a failure case where a motorcycle is given yet the
transforming ordinary features into QPA vectors indeed
brings improvement on the performance.

TABLE VI: The results of ablations. (a,b) shows the effect-
iveness of using QPA features. (b,f) and (d,f) show the ef-
ectiveness of the proposed BA mechanism. (e,f) shows RPN is a critical component to fully represent the advantages of DAnA.

$$N - M$$ classes to form an $$N$$-way support set.

Fan et al. [13] evaluated models under the COCO2VOC setting, where models are trained on the 60 COCO categories disjoint with VOC, and then tested on PASCAL VOC [54]. In this work, we consider a much more challenging setting termed PASCAL2COCO, where models are trained on PASCAL VOC 2007 and tested on the COCO benchmark. The 60 categories disjoint with VOC will serve as the novel categories, which is opposite to the setting in [13]. In Tab. IV, even the models are applied to a domain that is more complicated than the training domain, the models equipped with DAnA still demonstrate remarkable ability.

In Tab. IV we apply the same zero-shot protocol but follow the tradition in FSC to construct episodes. Instead of directly testing by COCO validation split, we follow the episode-based evaluation protocol defined in RepMet [10] to prepare 500 random evaluation episodes in advance. In each episode, take $$N$$-way $$K$$-shot evaluation for example, there will be $$K$$ support images and 10 query images for each of the $$N$$ categories. Consequently, there will be $$K \times N$$ support images and $$10 \times N$$ query images in each episode.

In Tab. IV and Tab. V we have compared DAnA-FasterRCNN with another proposed framework, DAnA-RetinaNet, to evaluate the performance as adopting different detection networks. Generally, the two-stage few-shot object detectors [3], [2], [4], [5], [17] with region proposal network (RPN) have higher performance than the one-stage [11], [4] methods (see Tab. I). Nonetheless, though our DAnA-RetinaNet is based on the one-stage object detector, it still outperforms previous methods using two-stage networks, which suffices to verify the effectiveness and dexterity of our DAnA. On the other hand, it seems the use of feature pyramids does not help DAnA-RetinaNet reach better results. The result could be caused by the lack of RPN. Without RPN, networks must tackle a nearly exhaustive list of potential object locations. Furthermore, in FSOD, those foregrounds irrelevant to the presented support set should be classified as backgrounds as well. Therefore, the huge amount of anchors and the extremely imbalanced foreground-background ratio would seriously undermine the ability of DAnA-RetinaNet. To conclude, we suggest RPN plays a deep role in fully presenting the advantages of DAnA, which is verified by the aforementioned experiments and the ablation study.

E. Ablations

In Tab. VI we show the results of the ablation study to analyze the impact brought by each component.

a) Whether CISA is effective at utilizing support images: In Sec. III-B3 we argue that the issue of feature uncertainty can be addressed by performing average pooling across query-position-aware (QPA) vectors. In the ablations, we investigate the improvement brought by our CISA component. For setting (a), we directly take the mean feature over multiple images as the class representations. The ablation (a, b) suffices to show that transforming ordinary features into QPA vectors indeed brings improvement on the performance.

b) Why not just denoise by soft attention: To attenuate noise in support images, it is intuitive to directly learn a soft attention mask and reweights the importance of each pixel. In Tab. VI “Mask” means we leverage CNN to learn a soft attention mask and perform element-wise product between the mask and the support feature map. As it can be observed, the result suggests that using the BA module is more effective than using general soft attention by preserving much more information.

c) How to combine query features and QPA vectors: Since the resulting QPA features will have the same size as query feature maps, they can be combined by either concatenation or element-wise product. According to (b, c), we conclude that concatenation is a better strategy for DAnA.

d) The importance of RPN: In the experiments, we have compared the two proposed detection networks, DAnA-FasterRCNN (with RPN) and DAnA-RetinaNet (without RPN). Based on the results in Tab. IV and V we conclude that RPN is a critical component in our problem. Nonetheless, there could be other factors influencing the results, such as the use of feature pyramids and Focal Loss [7]. Therefore, in the ablations (e), we re-train a DAnA-RetinaNet but remove the feature pyramids from it. However, we still adopt the Focal Loss in (e) because either RPN or Focal Loss is necessary to prevent the model from being severely deteriorated by numerous negative anchors. The result (e, f) shows that, without the influence of feature pyramids, the model including RPN has significantly better performance. To conclude, we suggest RPN is an indispensable component to fully represent the effectiveness of DAnA.

F. Visualizations

The examples of few-shot object detection are demonstrated in Fig. 7 In FSOD, the prediction is dependent on the object category of the given support set. Notably, all the target instances in Fig. 7 belong to the novel classes, so the model has never been trained to recognize these objects. However, given few support images, the proposed model is capable of recognizing and locating the instances. The last row of Fig. 7 presents a failure case where a motorcycle is given yet the bicycle in the query is detected.

We also visualize the cross-image spatial attention (CISA) in Fig. 8. The CISA module is capable of capturing the semantic correspondence between the query and support. Take (e) for example, given the head (colored in red) or the feet of
Query and support  Prediction  Ground truth  Query and support  Prediction  Ground truth

Fig. 7: At training, the model has never seen the novel categories, including bird, car, dog, person, motorcycle, bicycle, etc. However, given different support examples, our model is capable of recognizing unseen target objects in the query image. The last row represents the failure case where the model confuses a bicycle with a motorcycle.

Fig. 8: For each query image, we consider two different regions (colored in red and blue) and visualize their corresponding support attention maps. The CISA module is capable of capturing the semantic correspondence between the query and support images (e.g., the head or legs of an animal). (f) represents the case where neither of the support categories exists in the query image.

a human (colored in blue), the attention maps will highlight the head or the feet areas of the support images. The result (f) represents the case that neither of the support categories exists in the query image, which shows that if there is no corresponding contextual information, CISA has a tendency to highlight the most distinguishing components that can best describe an object.

V. CONCLUSION

In this work, we observed the problem of spatial misalignment and feature uncertainty in the challenging few-shot object detection (FSOD) task. We propose a novel and effective Dual-Awareness Attention (DAnA) mechanism to tackle the problem. Our method is adaptable to both one-stage and two-stage object detection networks. DAnA remarkably boosts the FSOD performance on the COCO benchmark and reaches state-of-the-art results. We are excited to point out a new direction to solve FSOD tasks. We encourage future works to extend our method to other challenging tasks such as few-shot instance segmentation and co-salient object detection.

ACKNOWLEDGMENT

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 110-2634-F-002-026. We benefit from NVIDIA DGX-1 AI Supercomputer and are grateful to the National Center for High-performance Computing.

REFERENCES

[1] B. Kang, Z. Liu, X. Wang, F. Yu, J. Feng, and T. Darrell, “Few-shot object detection via feature reweighting,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8420–8429.
P. Dollár, B. Babenko, S. Belongie, P. Perona, and Z. Tu, “Multiple component learning for object detection,” in *European conference on computer vision*. Springer, 2008, pp. 211–224.

B. Babenko, M.-H. Yang, and S. Belongie, “Robust object tracking with online multiple instance learning,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 8, pp. 1619–1632, 2010.

A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.

M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

R. Socher, M. Ganjo, H. Sridhar, O. Bastani, C. D. Manning, and A. Y. Ng, “Zero-shot learning through cross-modal transfer,” *arXiv preprint arXiv:1301.3666*, 2013.

**Tung-I Chen** received the B.E. degree in the Department of Biomedical Engineering, National Cheng Kung University (NCKU) in 2019, and the M.S. degree in the Department of Computer Science & Information Engineering, National Taiwan University (NTU) in 2021. His research interests include computer vision, object detection and machine learning theory.

**Yueh-Cheng Liu** received M.S. degree from computer science department in National Taiwan University (NTU) in 2020. He is currently a research assistant in vision science lab in EE department at National Tsing Hua University. His research interests include computer vision, robotic learning, and 3D scene understanding.

**Hung-Ting Su** is currently pursuing the Ph.D. degree with the Graduated Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan. His current research interests include multimodal comprehension and unsupervised learning.

**Yu-Cheng Chang** received his M.S. degree from the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, in 2020. His research interests include machine learning and medical image processing.

**Yu-Hsiang Lin** received the M.S. degree in statistics from National Yang Ming Chiao Tung University, Hsinchu, Taiwan, in 2019. He is currently a research assistant at National Tsing Hua University, Taiwan. His research interests include object detection, natural language processing, sentiment analysis, and deep learning.

**Jia-Fong Yeh** received the B.S. degree and the M.S. degree in Computer Science and Information Engineering from National Taiwan Normal University (NTNU), Taiwan, in 2017 and 2019, respectively. He is currently pursuing his Ph.d. degree in Computer Science and Information Engineering at National Taiwan University (NTU), Taiwan. His research interests include machine learning (ML), evolutionary algorithms (EAs), and computer games (CG). Recently, He is devoted to the study on few-shot learning. Jia-Fong is a student member of IEEE.

**Winston H. Hsu** (S’03–M’07–SM’12) received the Ph.D. degree in electrical engineering from Columbia University, New York, NY, USA. He is keen to realizing advanced researches towards business deliverables via academia-industry collaborations and co-founding startups. Since 2007, he has been a Professor with the Graduate Institute of Networking and Multimedia and the Department of Computer Science and Information Engineering, National Taiwan University. His research interests include large-scale image/video retrieval/mining, visual recognition, and machine intelligence. Dr. Hsu served as the Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA and on the Editorial Board for the IEEE MultiMedia Magazine.

**Wen-Chin Chen** received a BS in mathematics from National Taiwan University and a PhD in computer science from Brown University in 1976 and 1984. He has been the Professor of Computer science and Information Engineering department of National Taiwan University since 1987. His research interests are in the areas of design and analysis of algorithms, multimedia systems and quantumn algorithms.