Adaptive Image Transformer for One-Shot Object Detection

Ding-Jie Chen[1], He-Yen Hsieh[1], and Tyng-Luh Liu[1,2]
[1]Institute of Information Science, Academia Sinica, Taiwan
[2]Taiwan AI Labs

Abstract

*One-shot object detection tackles a challenging task that aims at identifying within a target image all object instances of the same class, implied by a query image patch. The main difficulty lies in the situation that the class label of the query patch and its respective examples are not available in the training data. Our main idea leverages the concept of language translation to boost metric-learning-based detection methods. Specifically, we emulate the language translation process to adaptively translate the feature of each object proposal to better correlate the given query feature for discriminating the class-similarity among the proposal-query pairs. To this end, we propose the Adaptive Image Transformer (AIT) module that deploys an attention-based encoder-decoder architecture to simultaneously explore intra-coder and inter-coder (*i.e.*, each proposal-query pair) attention. The adaptive nature of our design turns out to be flexible and effective in addressing the one-shot learning scenario. With the informative attention cues, the proposed model excels in predicting the class-similarity between the target image proposals and the query image patch. Though conceptually simple, our model significantly outperforms a state-of-the-art technique, improving the unseen-class object classification from 63.8 mAP and 22.0 AP50 to 72.2 mAP and 24.3 AP50 on the PASCAL-VOC and MS-COCO benchmark datasets, respectively.*

1. Introduction

Object detection is considered one of the core techniques in computer vision. Learning such a system [24, 27, 28, 35] reliably often requires a large amount of labeled training data over a wide range of object categories . These days a state-of-the-art object detector is expected to perform well in localizing those objects in a scene, whose class labels have been *seen* in training, but it can still be easily confused by those of *unseen* classes. To alleviate the predicament, the task of One-Shot object Detection (OSD) [2, 12, 26] is introduced to extend the system to detect objects of an arbitrary unseen class, which is implicitly hinted by a given inference query often in the form of an image patch.

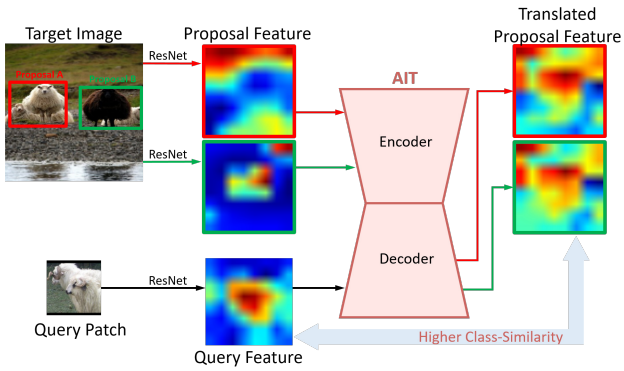

Figure 1: Adaptive Image Transformer (AIT): The proposed AIT module can adaptively represent each region proposal so that the similarity to the query patch can be properly evaluated. Specifically, AIT *translates* the feature of a region proposal to match the query feature. The adaptive mechanism can improve the similarity measurement and boost the performance of our metric-learning based one-shot object detector. Here the illustrated visual features are selected from the 508th channel of ResNet-50 stage-4.

Understandably, the OSD problem leads to a challenging task as on-line model fine-tuning is not performed and less feasible under the one-shot setting. In addition, even of the same object class, the query and the corresponding objects in a target image could differ substantially in size, shape, color, texture and appearance. To account for such unforeseen variations, we aim to develop a neural network approach that explores the multi-head attention mechanism to adaptively represent each potential region (*i.e.*, a region proposal in our formulation) so that its relatedness to the query patch can be properly evaluated. Like in most of the attention-based architectures, the proposed network design may appear to be engineered solely for better performance. It

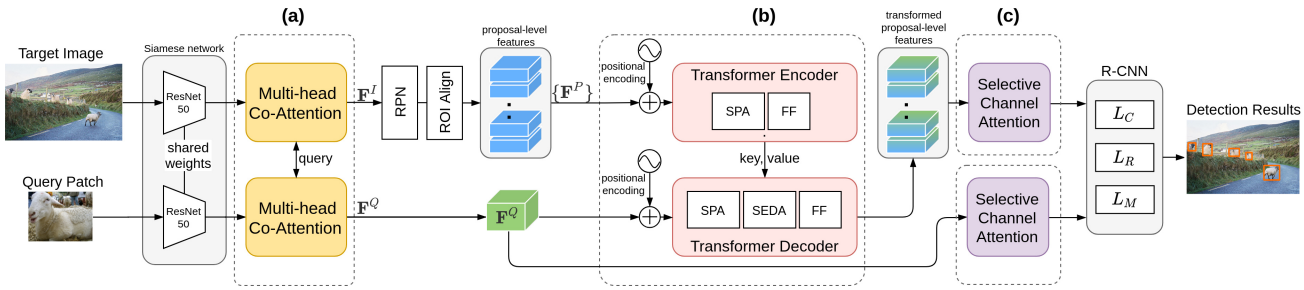

Figure 2: Overview of our one-shot object detection model. (a) Multi-head Co-Attention (MCA). (b) Adaptive Image Transformer (AIT). (c) Selective Channel Attention (SCA). In AIT, SPA: Selective Parallel Attention; FF: Feed Forward; SEDA: Selective Encoder-Decoder Attention, which replaces the original encoder-decoder attention with the intra-coder SPA.

indeed goes beyond this perspective and intends to establish a unified principle for solving the OSD problem based on the proposed *adaptive image transformer* (AIT), as shown in Figure 1. To better convey our motivations, we describe the reasoning behind three key components of our method and their advantages over existing relevant SOTA techniques.

Our proposed OSD method is a two-stage technique that relies on the region proposal network (RPN) to generate candidate regions. Despite being class-agnostic, the original RPN is trained without access to any examples from the unseen object classes, and thus could degrade the detection performance in inference due to excluding some legitimate region proposals for a given one-shot query. In dealing with this issue, we develop the first key component of our method that uses *multi-head co-attention* (MCA) to correlate the target image and query patch through various embeddings. The attention mechanism jointly considers the target image and the query by exploring different aspects of visual characteristics and spawns a corresponding feature map that encode such relatedness, upon which the RPN could generate more relevant region proposals to the query. Compared with the non-local scheme in Hseish *et al*. [12], the proposed MCA mechanism resolves the one-shot issue more effectively as supported by the improved detection accuracy.

The AIT module, illustrated in Figures 1 and 2(b), constitutes the second key component of our method. It is designed to explore how each proposal-query pair shares common semantic attributes over the deep visual features. Specifically, we employ a feature translation scheme, inspired by the attention-based paradigm [31] which shows the advantages of tackling the task of language translation by leveraging the intra-coder and inter-coder attention. In our formulation, AIT would adaptively transform the feature map of each proposal to match the query feature via employing the learned attention mechanisms. That is, given a query patch, the aspects of visual characteristics, *e.g*., shape, texture, and color, to be emphasized could vary among the region proposals. Note that AIT is more general than the co-excitation module in [12] where channel re-weighting is applied to the whole feature map rather than adapted to each proposal. In Figure 1, we see that the two region proposals are adaptively translated to match the query patch by attending on different aspects of visual features, namely, color and texture.

The third key component in our formulation concerns the use of *selective channel attention* (SCA) to improve the effectiveness of optimizing with the ranking loss. Although the AIT module can transform the proposal feature to match the query feature, the similarity could differ significantly over the respective channel dimension. Thus, it would be beneficial to enhance the importance of those channels of high similarity before evaluating a proposal-query pair. The consideration prompts our design of the SCA module, which is implemented with the SK-Net [21] to make each neuron based on multi-scale input information to adjust its receptive field size. Figure 2(c) shows that SCA is separately applied to each proposal feature and the query feature before computing the losses. We characterize the main contributions of our method to one-shot object detection as follows:

- We introduce the Adaptive Image Transformer (AIT) to effectively address the OSD task. AIT can adaptively translate the feature of each object proposal to better correlate the given query feature. The region-aware attention mechanism is more general than the whole-image co-excitation scheme in [12].
- We develop a Multi-head Co-Attention (MCA) module to explore feature relatedness and then use the information to refine both the feature maps of the target image and the query patch. As a result, the quality of the RPN region proposals can be significantly enhanced.
- We propose a novel Selective Parallel Attention (SPA) mechanism to improve the performance of intra-coder or inter-coder multi-head attention, which is the corner stone of Transformer-based techniques.
- Our method for one-shot object detection achieves state-of-the-art experimental results over existing techniques on two standard benchmark datasets.

2. Related work

This section overviews concisely about recent efforts relevant to the tasks of object detection, few-shot object detection, and attention mechanism.

2.1. Object detection

An object detector aims to localize objects of certain target classes in a given image and labels each object instance a corresponding class. The recent object detectors can be categorized into one-stage methods [17, 18, 20, 24, 27, 36] and two-stage methods [3, 9, 10, 28]. One-stage detectors simultaneously reason the locations and class labels of objects. In contrast, two-stage detectors first generate region proposals for locating the potential objects and then infer each proposal's class label. Note that the one-stage detectors [20, 24] are also called *single-shot* detectors. However, these methods aim to identify and localize seen-class objects. The goal is fundamentally different from our *one-shot* object detector. Our method follows Faster R-CNN [28] to design a two-stage object detector. The region proposal network of Faster R-CNN is employed to localize the potential target regions with respect to the query patch.

2.2. Few-shot object detection

The term of *few-shot* borrows the setting from the task of metric-learning-based few-shot classification [22, 29, 32, 34]. This sort of classification task aims to learn a metric for reasoning the unseen classes supported by labeled examples. Generally speaking, the N-way K-shot setting means K labeled samples available per class to handle the N-class classification. With the few-shot setting, the few-shot object detection aims to use a few supported samples for localizing and recognizing the objects. The previous works employ metric learning, transfer learning, meta learning, or contrastive training to address few-shot object detection. The metric learning based methods [12, 15, 25] include a metric classifier in their detectors. The transfer-learning based method [5] uses a regularization to alleviate the over-fitting while training on a handful of unseen-class labeled images. The meta-learning based method [14] trains a few-shot meta-model to re-weight the image features extracted from a detection model. The contrastive-training based method [8] exploits the attention-RPN and multi-relation detector for estimating the similarity between the support images and the target image regions. To detect unseen-class objects specified by one single example, Osokin *et al*. [26] densely match and align the target-image-feature and the query-image-feature to recognize the specified objects. Akin to the image classification [16] and detection [12] based on the metric learning under a one-shot setting, our model learns a similarity metric from the image pairs. Though each image pair provides only one supported sample, the learned metric is able to decide whether the classes of the two images are the same.

2.3. Attention mechanism

The attention mechanism has been shown the advantages of natural language processing [1, 31] and vision-related tasks [6, 13, 30] for capturing some specific properties while training features. There are various ways to implement the attention mechanism, and here we focus on *Transformers*, which was introduced by Vaswani *et al*. [31] as a building block for language translation. Like the non-local block [33], the transformer designs its attention mechanism by scanning each sentence's element and updating it with respect to the entire sentence's aggregated information. The AIT module borrows the merits from *Transformers* to leverage the intra-coder and inter-coder attention of each target-query pair. Precisely, the AIT models the intra-coder attention for target proposals and query patch and models the inter-coder attention of the target-query pair. Our idea is to transform the feature map of each target proposal to match the query feature via employing the learned attention mechanisms from AIT. As a result, our metric classifier for discriminating class-similarity is excel to lean the relation of each target-query pair through the better correlate features and hence shows a superior learning performance. Note that the *Transformers* is an auto-regressive model for generating the output token one by one; in contrast, our AIT directly generates the output at once instead and prevents this sort of iterative process.

3. Method

In this section, we first define the problem of one-shot object detection and then elaborate on how the proposed key components of our method facilitate tackling the OSD task.

Problem formulation Consider the OSD task over a set of object class labels $\mathcal{C} = \mathcal{S} \cup \mathcal{U}$, where $\mathcal{S}$ and $\mathcal{U}$ denote the sets of seen-class and unseen-class labels, respectively. The seen-class $\mathcal{S}$ includes those class labels of objects available during training, and $\mathcal{U}$ comprises the remaining unseen class labels, from which a one-shot inference query could assume. That is, the two sets $\mathcal{S}$ and $\mathcal{U}$ are mutually exclusive. In a valid implementation, one needs to exclude all the training images containing at least one unseen-class object to adhere to the OSD problem formulation.

Overview To mimic the one-shot inference, the proposed system is trained by providing a query patch Q from some object class in $\mathcal{S}$ and a target image I comprising at least one object of the underling class, the training goal is to learn to generate an adaptive feature map for each region proposal so that high similarity values can be obtained from those regions in I matching the ground truths.

Figure 2 sketches the overall architecture of the proposed one-shot object detection method. Our model is based on the two-stage detector Faster R-CNN [28], which employs visual feature extractor ResNet-50 [11] as the backbone, to

carry out the OSD task. The first stage begins with generating region proposals for the potential locations containing an object of the given query class. In addition to the classification and the regression losses, the following second stage scores each region proposal via the learned similarity metric for estimating the semantic (class) similarity per proposal with respect to the query image patch.

In the first stage, a Siamese network of ResNet-50 is adopted to respectively extract the visual representations of the target image I and the query patch Q. Then, multi-head co-attention is applied to correlate I and Q via multiple embeddings, where the proposed MCA module enables the RPN to generate region proposals more relevant to the query in various visual aspects. To sum up, the target image I is decomposed into a set of N region proposals, denoted as $\mathcal{P} = \{P_1, P_2, \ldots, P_N\}$, where each proposal P_i suggests potential presence of a query-class object and retrieves its visual feature via ROI-Align [28].

In the second stage, the proposed AIT module adaptively transforms each region proposal's visual representation into the feature space of the query patch. The region-aware feature transformation enables our system to uncover regions with the same object class to the query, but their resembling to the query could base on different visual aspects. However, the similarity between an AIT-transformed proposal feature and the query feature differs significantly in the channel dimension. We design the SCA module to weigh the importance of each channel respectively for both feature maps, as shown in Figure 2(c). In the experiment, our ablation study supports that SCA complements the *margin-based ranking loss* [12] and improves the overall OSD performance.

3.1. Multi-head co-attention

The quality of object proposals is pivotal to a two-stage object detector. In dealing with the scenario of one-shot object detection, this aspect of concern is even more crucial as the RPN is trained with the ground-truth bounding boxes only over the seen classes in $\mathcal{S}$. Without re-designing the learning strategy about the RPN, the resulting object proposals may fail to include some regions that correspond to the one-shot query patch of an unseen class. Different from CoAE [12], which employs the non-local block [33] as the attention mechanism to overcome the issue, our method considers a more powerful scheme termed as *multi-head co-attention* (MCA) to retain the effectiveness of RPN for one-shot object detection. Compared with the non-local proposals in CoAE, the proposed MCA module could explore co-attention from various aspects of visual feature representations and encode such relatedness in the resulting feature maps, upon which the RPN operates. We show in the experimental results that the resulting object proposals are of better quality and more effective than those by CoAE in locating the potential regions of interest to the one-shot query.

Attention unit Consider that a self-attention function [31, 33] is operated as a re-weighting of `value` $\mathbf{v}$ concerning the compatibility of its `key` $\mathbf{k}$ and a `query` $\mathbf{q}$, where the value, key, and query are derived from the *same* input feature vector but with different embeddings. Following the Transformer [31], we define a basic attention function f by

$$f(\mathbf{v}, \mathbf{k}, \mathbf{q}) = \mathrm{softmax}\left(\frac{\mathbf{q}\mathbf{k}^\intercal}{\sqrt{d_\mathbf{k}}}\right)\mathbf{v}\,, \tag{1}$$

where $\mathbf{v}$ is an embedded feature vectors of dimension $d_\mathbf{v}$, and $\mathbf{k}, \mathbf{q}$ are embedded feature vectors of dimension $d_\mathbf{k}$.

MCA As the multi-head attention [31] is defined to jointly attend to the information collected *in parallel* from various representation spaces, we express the attention function by

$$f^p(\mathbf{v}, \mathbf{k}, \mathbf{q}) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)\mathbf{W}^o\,, \tag{2}$$

$$\mathrm{head}_i = f(\mathbf{W}_i^\mathbf{v}\mathbf{v}, \mathbf{W}_i^\mathbf{k}\mathbf{k}, \mathbf{W}_i^\mathbf{q}\mathbf{q})\,, \tag{3}$$

where the weight matrices are $\mathbf{W}_i^\mathbf{v} \in \mathbb{R}^{d_\mathbf{m} \times d_\mathbf{v}}$, $\mathbf{W}_i^\mathbf{k} \in \mathbb{R}^{d_\mathbf{m} \times d_\mathbf{k}}$, $\mathbf{W}_i^\mathbf{q} \in \mathbb{R}^{d_\mathbf{m} \times d_\mathbf{k}}$, and $\mathbf{W}^o \in \mathbb{R}^{hd_\mathbf{v} \times d_\mathbf{m}}$. In this work, we use the default number of attention heads, *i.e.*, $h = 8$, and $d_\mathbf{v} = d_\mathbf{k} = d_\mathbf{m}/h = 64$.

Given a target image I and a query patch Q, we can leverage (2) to establish the multi-head co-attention (MCA) between I and Q:

$$\mathbf{F}^I = f^p(\mathbf{v}^I, \mathbf{k}^I, \mathbf{q}^Q)\,, \tag{4}$$

$$\mathbf{F}^Q = f^p(\mathbf{v}^Q, \mathbf{k}^Q, \mathbf{q}^I)\,, \tag{5}$$

where the superscript denotes the source of the feature. From (4) and (5), we see that unlike the self-attention mechanism, the MCA feature $\mathbf{F}^I$ of target image I is obtained by considering feature embeddings from two different sources, I and Q, and the same applies to the other MCA feature $\mathbf{F}^Q$ of the query patch Q. As mentioned before, the RPN employs $\mathbf{F}^I$ to generate region proposals $\mathcal{P}$ and extracts the proposal-level features $\{\mathbf{F}^P\}$ via ROI-Align. Observe that, owing to the attended target feature $\mathbf{F}^I$ involving the weighted features between I and Q, it is expected that RPN could generate proposals more relevant to the query Q and hence more suitable for the one-shot object detection task. Figure 3(a) shows the data flow of the proposed MCA module.

3.2. Adaptive image transformer

At the core of our method is the proposed adaptive image transformer which enables our method to correlate the query patch to each RPN region proposal, rather than the whole target image. To facilitate the use of a transformer-like module for the OSD task, we have proposed a number of effective modifications leading to the AIT, which is applied to transform the feature of each region proposal to account for the one-shot query feature. Figure 2(b) sketches the architecture of the proposed AIT module.

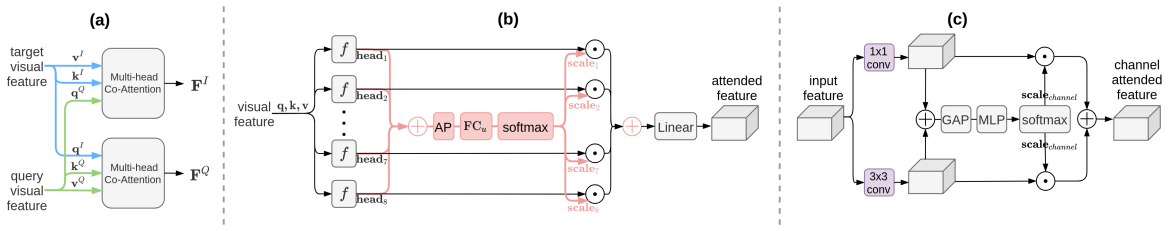

Figure 3: Main attention mechanisms in our method. (a) Multi-head Co-Attention (MCA). (b) Selective Parallel Attention (SPA). (c) Selective Channel Attention (SCA). The highlighted non-gray parts indicate the modifications in our model compared to the original ones mentioned in the context. Please refer to supplementary material for architecture details.

Transformer The *Transformer* framework [31] is established upon an encoder-decoder architecture. Both the encoder and decoder receive the encoded inputs via *positional encoding*, and generate the embedded outputs through a feed-forward network. In particular, the positional encoding adds relative positions in the form of sinusoidal vectors to the features, while the feed-forward network can be considered as two convolutions with kernel size 1. Regarding the main components, the Transformer encoder includes a multi-head attention module for exploring the intra-coder attention over all the input target word-embedding vectors. On the other hand, except the masked strategy to preserve the auto-regressive property, the Transformer decoder also comprises a similar multi-head attention module for exploring the intra-coder attention over all the input query word-embedding vectors. Further, the Transformer decoder has an encoder-decoder multi-head attention module which leverages the key and value from the output of the encoder to probe the inter-coder attention. Note that all the above-mentioned multi-head attention modules can essentially be implemented with the operations in (2) and (3). For further details about the Transformer architecture, we refer the readers to the inspiring work [31].

From our brief description about Transformer, it can be comprehended that the multi-head attention plays an important role in the mechanism. However, the integration of the individual attention from each head is carried out via a simple concatenation, followed by a linear embedding as in (2). In our experiment, we observe that it can be improved by a more flexible attention fusion scheme, which we term it as *selective parallel attention* (SPA) and describe in what follows. We remark that in the AIT module, the SPA scheme is always used to integrate the multi-head attention.

Selective parallel attention The original Transformer employs concatenation to directly combine all heads and resumes the proper dimension via a linear embedding as in (2). We instead propose the SPA module to more effectively fuse the information from the h heads, where the underlying idea is motivated by the SE-Net [13] and SK-Net [21]. We denote the selective parallel attention as f^s and design the SPA to make each neuron weightily select its representation from the multiple heads. (See Figure 3(b).) We have

$$f^s(\mathbf{v},\mathbf{k},\mathbf{q}) = \sum\nolimits_i (\text{scale}_i \times \text{head}_i)\mathbf{W}^s\,, \tag{6}$$

$$\text{scale}_i = \text{softmax}([\mathbf{s}_i]_{i=1}^h)\,, \tag{7}$$

$$[\mathbf{s}_i]_{i=1}^h = \text{FC}_u(\text{AP}(\sum\nolimits_i \text{head}_i))\,, \tag{8}$$

$$\text{head}_i = f(\mathbf{W}_i^{\mathbf{v}}\mathbf{v}, \mathbf{W}_i^{\mathbf{k}}\mathbf{k}, \mathbf{W}_i^{\mathbf{q}}\mathbf{q})\,, \tag{9}$$

where the weight matrices $\mathbf{W}_i^{\mathbf{v}}$, $\mathbf{W}_i^{\mathbf{k}}$, $\mathbf{W}_i^{\mathbf{q}}$ are defined as before, and $\mathbf{W}^s \in \mathbb{R}^{d_{\mathbf{v}} \times d_{\mathbf{m}}}$. The $\sum$ denotes the element-wise summation of the results from all heads. The AP denotes the average pooling of all pixel-level vectors. The FC_u is a fully-connected layer that expands the dimension from $d_{\mathbf{v}}$ to $h \times d_{\mathbf{v}}$ so that a corresponding scale vector $\mathbf{s}_i$ can be obtained for each head. Note that the softmax in (8) operates over the head dimension h.

In addition to the way that the multi-head attention is fused, AIT also adopts a different data feeding strategy from Transformer. In particular, all the pixel-level features of $\mathbf{F}^Q$ are simultaneously fed to the decoder, which enables the AIT module to skip the auto-regressive process and achieve a more effective implementation of feature translation.

We now justify the design of the proposed AIT module. In the adaptive image transformer, each pixel-level visual feature of $\mathbf{F}^P$ or $\mathbf{F}^Q$ plays the role akin to the word embedding vector in the Transformer. Our formulation respectively inputs all pixel-level features from $\mathbf{F}^P$ and $\mathbf{F}^Q$ into the encoder and decoder such that AIT functions as projecting each proposal feature $\mathbf{F}^P$ into the feature space of query feature $\mathbf{F}^Q$. As a result of applying the AIT, we are able to learn the pixel-level intra-coder attention of $\mathbf{F}^P$ with the AIT encoder, and of $\mathbf{F}^Q$ with the AIT decoder. More importantly, the correlation between between $\mathbf{F}^P$ and $\mathbf{F}^Q$ can be explored by the inter-coder multi-head attention, which for brevity we denote it as SEDA in Figure 2(b) for selective encoder-decoder attention. The architecture of SEDA is the same as SPA, yet SEDA leverages the key and value from the output of the AIT encoder. The experimental results show the feature translation from $\mathbf{F}^P$ to $\mathbf{F}^Q$ via our AIT is beneficial for addressing metric-learning based one-shot object detection.

3.3. Selective channel attention and ranking loss

Following [12], we consider a margin-based loss function to rank the proposal-query pairs. As the resemblance between an AIT-transformed proposal feature and the query feature could vary significantly over the channel dimension. We plug in two independent modules of selective channel attention, as shown in Figure 2(c), to properly re-weight the importance of each feature channel so that the metric learning can be carried out more effectively.

Selective channel attention Figure 3(c) illustrates the proposed SCA module. We implement the proposed SCA based on the SK-Net [21]. It re-weights each feature channel by leveraging different receptive filed sizes on multiple scales of the input feature maps. In this work, we empirically employ the SK-Net of the kernel sizes of 1×1 and 3×3 to adjust the proposal-level features and the query-patch feature. The main operations of the SK-Net are analogous to (7) and (8); we refer the reader to [21] for more details.

Margin-based ranking loss We adopt the margin-based ranking loss introduced in [12] to score the proposal-query pairs. Specifically, we employ a two-layer MLP network as metric learning to generate a two-class softmax prediction, one for same-category and the other for different-category. In the ranking loss, a concatenated feature vector $\mathbf{x} = [\mathrm{GAP}(\mathbf{F}^P); \mathrm{GAP}(\mathbf{F}^Q)]$ combines the spatially global average pooling (GAP) vectors of the proposal feature and query feature. The label y of $\mathbf{x}$ is 1 if the underlying proposal and the ground truth have their IoU larger than 0.5, and otherwise, is set to 0. The margin-based ranking loss function $\mathcal{L}_{\mathrm{M}}$ is defined by

$$\begin{aligned}\mathcal{L}_{\mathrm{M}}(\{\mathbf{x}_i\}) = &\sum_{i=1}^{N} y_i \times \max\{m^+ - s_i, 0\} \\ &+ (1 - y_i) \times \max\{s_i - m^-, 0\} + \Delta_i \,, \quad (10)\end{aligned}$$

$$\begin{aligned}\Delta_i = &\sum_{j=i+1}^{N} [y_i = y_j] \times \max\{|s_i - s_j| - m^-, 0\} \\ &+ [y_i \neq y_j] \times \max\{m^+ - |s_i - s_j|, 0\} \,, \quad (11)\end{aligned}$$

where s is the same-category probability by the MLP, $[\cdot]$ is the Iverson bracket. The margin m^+ is the expected lower bound of the same-category probability, and m^- is the expected upper bound of the different-category probability.

Finally, the complete loss for training our one-shot object detection model can be expressed by

$$\mathcal{L} = \mathcal{L}_{\mathrm{C}} + \mathcal{L}_{\mathrm{R}} + \lambda \mathcal{L}_{\mathrm{M}} \,, \quad (12)$$

where $\mathcal{L}_{\mathrm{C}}$ and $\mathcal{L}_{\mathrm{R}}$ are the cross entropy and regression loss of Faster R-CNN, respectively. We set $m^+ = 0.7$, $m^- = 0.3$, and $\lambda = 3$ as [12].

4. Experiments

In this section, we evaluate our model on two datasets comparing to previous methods [4, 12, 19, 25, 37]. We first illustrate the two datasets, implementation details, and the process to generate the experiments' target-query image pairs. Then, we compare our model against other state-of-the-art methods to demonstrate its advantages in addressing the one-shot object detection task. Finally, we provide an ablation study of the architecture and visualizes some examples for realizing the effectiveness of each model component.

Datasets We follow the previous work [12, 25] to train and evaluate our model on datasets PASCAL-VOC [7] and MS-COCO [23]. For using the PASCAL-VOC dataset, we train our model on the set composed of 'PASCAL-VOC 2007 train&val' and 'PASCAL-VOC 2012 train&val,' and then test on the 'PASCAL-VOC 2007 test' set. We follow [37] to partition each set of the 20 object classes for fitting the one-shot objection detection scenario. There are 16 seen classes (PASCAL-VOC-train-16) for model training, and the remained four unseen classes (PASCAL-VOC-test-4) are for testing. In the MS-COCO dataset, we use the image set 'train 2017' for training our model, and then evaluate with the image set 'val 2017.' As the work [25], we partition each set of 80 classes into four groups. During the model training, we employ three groups of 60 seen classes (MS-COCO-train-60). As a result, the remained group of 20 unseen classes (MS-COCO-test-20) is used for testing.

Implementation details We use the SGD optimizer to train our model with momentum of 0.9 for ten epochs and weight decay of $1e-4$. We train our model with a batch size of 32 on four V100 GPUs in parallel. The learning rate starts at 0.01 and gradually decays by a ratio of 0.1 for every four epochs. There are two pre-trained weights available for initialing the backbone network, *i.e.* ResNet-50. The first pre-trained weight is from the original ImageNet, which contains 1,284,168 images of 1,000 classes. The second weight is from the reduced ImageNet, which includes 933,052 images of 725 classes. In the reduced ImageNet, the MS-COCO related classes are removed [12] while training the backbone network; hence it is guaranteed to exclude the PASCAL-VOC related classes from the ImageNet for ensuring the trained model not foresee the unseen-class objects.

Target-query pairs We follow [12] to generate target-query image pairs. The target images are first chosen from the datasets. In the training phase, we randomly select a query seen-class image patch for a given target image containing the same seen-class object. In the testing phase, for each class in a target image, the query image patches of the same class are shuffled with a random seed of target image ID, then the first five query image patches are selected with averaging their Average Precision (AP) scores. The performance is evaluated by averaging all AP scores.

Method	Seen Class																	Unseen Class				
	plant	sofa	tv	car	bottle	boat	chair	person	bus	train	horse	bike	dog	bird	mbike	table	mAP	cow	sheep	cat	aero	mAP
SiamFC	3.2	22.8	5.0	16.7	0.5	8.1	1.2	4.2	22.2	22.6	35.4	14.2	25.8	11.7	19.7	27.8	15.1	6.8	2.28	31.6	12.4	13.3
SiamRPN	1.9	15.7	4.5	12.8	1.0	1.1	6.1	8.7	7.9	6.9	17.4	17.8	20.5	7.2	18.5	5.1	9.6	15.9	15.7	21.7	3.5	14.2
CompNet	28.4	41.5	65.0	66.4	37.1	49.8	16.2	31.7	69.7	73.1	75.6	71.6	61.4	52.3	63.4	39.8	52.7	75.3	60.0	47.9	25.3	52.1
CoAE (1000)	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
Ours (1000)	**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**
CoAE (725)	24.9	50.1	58.8	64.3	32.9	48.9	14.2	53.2	71.5	74.7	74.0	66.3	75.7	61.5	68.5	42.7	55.1	78.0	61.9	72.0	43.5	63.8
Ours (725)	**46.4**	**60.5**	**68.0**	**73.6**	**49.0**	**65.1**	**26.6**	**68.2**	**82.6**	**85.4**	**82.9**	**77.1**	**82.7**	**71.8**	**75.1**	**60.0**	**67.2**	**85.5**	**72.8**	**80.4**	**50.2**	**72.2**

Table 1: Performance comparison on the PASCAL-VOC dataset in terms of mAP (%). '(725)' means the model is pre-trained on a reduced ImageNet dataset for preventing from foreseeing the unseen-class objects. Note that SiamFC, SiamRPN, and CompNet use all classes in their ImageNet pre-trained backbones.

Method	split-1	split-2	split-3	split-4	Average	Method	split-1	split-2	split-3	split-4	Average
SiamMask (seen)	38.9	37.1	37.8	36.6	37.6	SiamMask (unseen)	15.3	17.6	17.4	17.0	16.8
CoAE (seen)	42.2	40.2	39.9	41.3	40.9	CoAE (unseen)	23.4	23.6	20.5	20.4	22.0
Ours (seen)	**50.1**	**47.2**	**45.8**	**46.9**	**47.5**	Ours (unseen)	**26.0**	**26.4**	**22.3**	**22.6**	**24.3**

Table 2: Performance comparison on the MS-COCO val 2017 dataset in terms of AP50 score (%).

4.1. Comparison with state-of-the-art methods

Some previous methods related to our method are used for evaluating on two datasets. In the PASCAL-VOC dataset, the compared methods are SiamFC [4], SaimRPN [19], CompNet [37], and CoAE [12]. SiamFC and SaimRPN aim to tackle the tracking task. CompNet is based on Faster R-CNN to equip with the metric-based classifiers in RPN and R-CNN. CoAE aims to address the one-shot object detection task with the proposed co-attention, co-excitation, and margin-based ranking loss. In the MS-COCO dataset, the compared methods are SiamMask [25] and CoAE. SiamMask is based on Mask R-CNN [10] with a metric-based classifier within R-CNN.

For the PASCAL-VOC dataset, Table 1 shows that our method 'Ours (725)' achieves seen-class mAP of 67.2% and unseen-class mAP of 72.2% by pre-training on the reduced ImageNet. Our performance improvements are 12.1% mAP of the seen-class and 8.4% mAP of the unseen-class compared with 'CoAE (725).' The results show that our model outperforms all the other methods among all classes in large margins, which demonstrates that our transformer-based OSD model evidently benefits the one-shot object detection task. Furthermore, our performance could be further boosted once the ImageNet pre-trained backbone is trained over 1,000 classes, *i.e.*, 'Ours (1000).'

For the MS-COCO dataset, Table 2 shows that our model 'Ours' achieves seen-class average-AP50 of 47.5% and unseen-class average-AP50 of 24.3%. That is, our results are better than that of the state-of-the-art CoAE by 6.6% and 2.3%, respectively. The results on the MS-COCO also show that our model outperforms all other methods among all splits. Therefore, the comparisons on the two datasets show the strong generalization property of our OSD model.

4.2. Ablation analysis

The experiment in this part compares different configurations of the proposed model for assessing each component's effectiveness. Table 3 presents the performance of our model under various configurations on the PASCAL-VOC dataset with the mAP metric. In Table 3, the configuration in row 1, namely the CoAE, in which the co-excitation is akin to SE-Net [13] for re-weighting per feature channel. The comparison of row 1 and row 2 shows significant performance-boosting while employing the proposed AIT module. Besides, while integrating the Multi-head Co-Attention (MCA) and the Selective Channel Attention (SCA), we can further improve the detection performance as shown in row 5. The comparison of row 3 and row 4 shows that the selective parallel attention module is no need for the masked strategy as the Transformer. The ablation study shows the effectiveness of each component in our model. This performance-boosting by our model obviously demonstrates the proposal-feature translation concerning the query-feature is helpful to address the one-shot object detection task.

	Configuration			Seen Class	Unseen Class				
	Figure 2(a)	Figure 2(b)	Figure 2(c)	mAP	cow	sheep	cat	aero	mAP
1	Co-attention	Co-excitation	none	55.1	78.0	61.9	72.0	43.5	63.8
2	Co-attention	AIT	none	65.5	82.6	72.2	**80.6**	48.1	70.9
3	Co-attention	AIT	SCA	66.6	**85.5**	72.2	80.1	49.8	71.3
4	Co-attention	AIT-MSPA	SCA	65.5	84.1	69.1	77.0	50.0	70.0
5	MCA	AIT	SCA	**67.2**	**85.5**	**72.8**	80.4	**50.2**	**72.2**

Table 3: Ablation study for various configurations of the proposed OSD model on the PASCAL-VOC dataset in terms of mAP (%). The configuration comprises the three-part settings that corresponding to the (a), (b), and (c) in Figure 2. MCA: Multi-head Co-Attention; AIT-MSPA: Adaptive Image Transformer with Masked Selective Parallel Attention; SCA: Selective Channel Attention.

4.3. Visualization

To analyze the property of multi-head co-attention, Figure 4 visualizes the proposal distribution as a heatmap. Each pixel counts how many proposals cover it. Normalizing the pixel count hence generates the probability map as the heatmap. The results show that the multi-head co-attention enables the RPN to generate proposals focusing on the query-class object. Figure 5 visualizes the transformed deep visual feature maps in specific channels to realize the advantage of our adaptive image transformer. Each feature map means the specific feature channel selected from the ResNet-50 stage-4, and our AIT module translates it concerning the query feature and hence generates the translated feature. The results show that the correct proposal (red box in the target image) has a better translated feature quality, *i.e.*, more similar to the query feature. As a result, our AIT module helps learn metrics for ranking target proposals. Figure 6 shows the usage of our one-shot object detection model. Our model is able to detect the correct query-class object even with the problematic query-patch covering the partial object regions.

5. Conclusion

We have demonstrated the feasibility to boost the metric-learning based one-shot object detection through image feature translation. The proposed AIT module adaptively translates the feature of each region proposal to better correlate the given query feature for discriminating the class-similarity among the proposal-query pairs. As a result, our model makes a metric-learning-based object detector easier to learn such a class-similarity metric and consequently enhances its effectiveness for addressing the one-shot object detection task. Besides benefiting from the learned intra-coder and inter-coder attention within our AIT module, we present the multi-head co-attention, selective parallel attention, and selective channel attention to gain more OSD performance improvements. Combining all these attention mechanisms manifests our model's advantage in achieving state-of-the-art OSD performance compared with the existing methods.

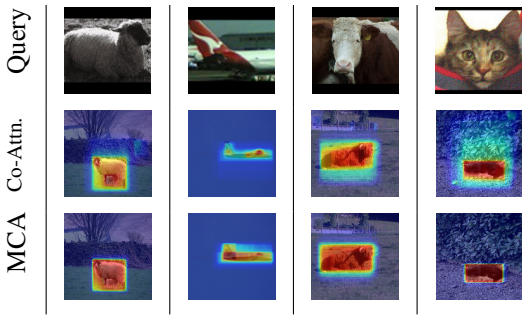

Figure 4: Visualization of the RPN's proposal quality. The first row shows the query images. The second row shows the RPN's proposal quality by using co-attention (Co-Attn.) in CoAE. The third row shows the RPN attracts more proposals on the query-class object with the proposed multi-head co-attention in our one-shot object detection model.

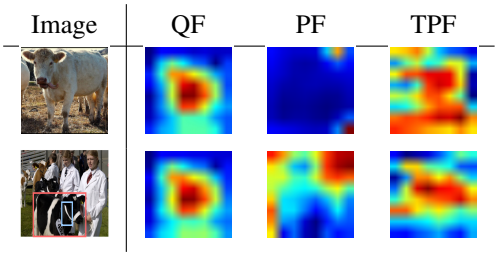

Figure 5: The left-most column from top to bottom shows the query patch and target image; the right three columns of the top row show the 254th feature channel corresponding to the low-quality proposal (cyan box in the target image); the right three columns of the bottom row show the 286th feature channel corresponding to the high-quality proposal (red box in the target image). The right three columns show the specific feature channels from the query feature (QF), proposal feature (PF), and translated proposal feature (TPF).

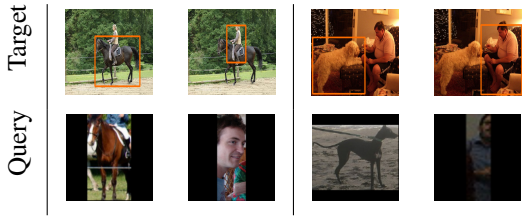

Figure 6: Our one-shot object detection model is able to make a target image result in the different detected regions with respect to the different query image patches.

Acknowledgement. This work was supported in part by the MOST, Taiwan under Grant 110-2634-F-001-009. We are also grateful to the *National Center for High-performance Computing* for providing computational resources and facilities.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[2] Sujoy Kumar Biswas and Peyman Milanfar. One shot detection with laplacian object and fast matrix cosine similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(3):546–562, 2016.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018.

[4] Miaobin Cen and Cheolkon Jung. Fully convolutional siamese fusion networks for object tracking. In *ICIP*, pages 3718–3722, 2018.

[5] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. LSTD: A low-shot transfer detector for object detection. In *AAAI*, pages 2836–2843, 2018.

[6] Misha Denil, Loris Bazzani, Hugo Larochelle, and Nando de Freitas. Learning where to attend with deep architectures for image tracking. *Neural Computation*, 24(8):2151–2184, 2012.

[7] Mark Everingham, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[8] 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 4012–4021. IEEE, 2020.

[9] Ross B. Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):386–397, 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[12] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. In *NeurIPS*, pages 2721–2730, 2019.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[14] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *ICCV*, pages 8419–8428, 2019.

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

[16] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML 2015 Deep Learning Workshop*, 2015.

[17] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Trans. Image Process.*, 29:7389–7398, 2020.

[18] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. *Int. J. Comput. Vis.*, 128(3):642–656, 2020.

[19] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, 2018.

[20] Shuai Li, Lingxiao Yang, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Dynamic anchor feature selection for single-shot object detection. In *ICCV*, pages 6608–6617, 2019.

[21] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *CVPR*, pages 510–519, 2019.

[22] Xiang Li, Lin Zhang, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. One-shot object detection without fine-tuning. abs/2005.03819, 2020.

[23] Tsung-Yi Lin, Michael Maire, Serge J. 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.

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

[25] Claudio Michaelis, Ivan Ustyuzhaninov, Matthias Bethge, and Alexander S. Ecker. One-shot instance segmentation. volume abs/1811.11507, 2018.

[26] Anton Osokin, Denis Sumin, and Vasily Lomakin. OS2D: one-stage one-shot object detection by matching anchor features. In *ECCV*, 2020.

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

[28] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

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

[30] Yichuan Tang, Nitish Srivastava, and Ruslan Salakhutdinov. Learning generative models with visual attention. In *NIPS*, pages 1808–1816, 2014.

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

[32] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NIPS*, pages 3630–3638, 2016.

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

[34] 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.

[35] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization for object detection. In *CVPR*, pages 10183–10192, 2020.

[36] 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 9756–9765, 2020.

[37] Tengfei Zhang, Yue Zhang, Xian Sun, Hao Sun, Menglong Yan, Xue Yang, and Kun Fu. Comparison network for one-shot conditional object detection. *CoRR*, abs/1904.02317, 2019.