Aggregating Bilateral Attention for Few-Shot Instance Localization

He-Yen Hsieh, Ding-Jie Chen, Cheng-Wei Chang, and Tyng-Luh Liu
Institute of Information Science, Academia Sinica, Taiwan
heyen@iis.sinica.edu.tw, djchen.tw@gmail.com, johnnyccw.tw@gmail.com, liutyng@iis.sinica.edu.tw

Abstract

Attention filtering under various learning scenarios has proven advantageous in enhancing the performance of many neural network architectures. The mainstream attention mechanism is established upon the non-local block, also known as an essential component of the prominent Transformer networks, to catch long-range correlations. However, such unilateral attention is often hampered by sparse and obscure responses, revealing insufficient dependencies across images/patches, and high computational cost, especially for those employing the multi-head design. To overcome these issues, we introduce a novel mechanism of aggregating bilateral attention *(ABA) and validate its usefulness in tackling the task of few-shot instance localization, reflecting the underlying query-support dependency. Specifically, our method facilitates uncovering informative features via assessing:* i) *an* embedding norm *for exploring the semantically-related cues;* ii) context awareness *for correlating the query data and support regions. ABA is then carried out by integrating the affinity relations derived from the two measurements to serve as a lightweight but effective query-support attention mechanism with high localization recall. We evaluate ABA on two localization tasks, namely, few-shot action localization and one-shot object detection. Extensive experiments demonstrate that the proposed ABA achieves superior performances over existing methods.*

1. Introduction

Non-local block (NLB) [39] is one self-attention mechanism, also known as an essential unit, i.e., the scaled dot-product attention unit within the multi-head attention in the prominent Transformer [37]. Both NLB and Transformer show the success of feature enhancement in addressing the various tasks of natural language processing and computer vision. For example, such an attention mechanism's footprints exist in several vision tasks of action localization [4, 13, 23, 44], object detection [6, 19], image denoising [27], 3D imaging [33], and semantic segmentation [51].

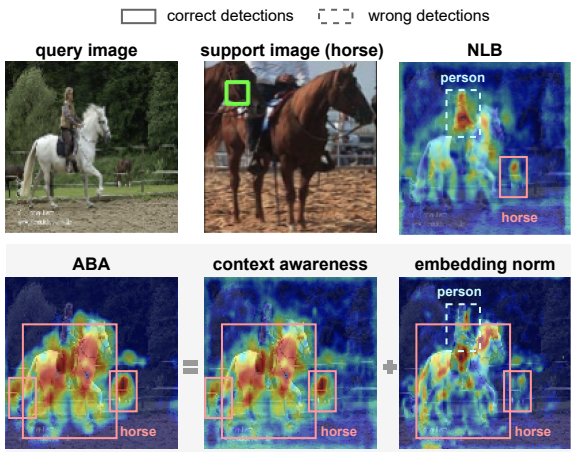

Figure 1: **The effect of Aggregating Bilateral Attention**. The top-left two images are the input pair, and the remaining four show the attention heatmaps stacking on top of the query image. Each attention heatmap depicts the pixel-level feature similarity against the green-box region within the support image. The proposed ABA can increase the localization recall to benefit the few-shot instance localization.

This paper is motivated by observing that most of the existing few-shot instance localization methods, which directly employ the typical non-local block, suffer the issues of *sparse attention* and *high computational cost*. The issue of sparse attention is observed in calculating the query-support dependency via the typical NLB could result in high accuracy yet low recall of regions of interest. The top-right image in Figure 1 shows that one detected brown horse of focused attention and the missed white horse of suppressed attention. Such a situation is caused by the affinity calculation step within NLB, including the dot product operation and the subsequent $\mathrm{softmax}$ operation. Precisely, one high-similarity pixel-pair of, *e.g.*, query brown horse against support brown horse, could rapidly accumulate its similarity over feature channels via the dot product operation. Further, the subsequent $\mathrm{softmax}$ operation could speed up such a

high-similarity pixel-pair to stand out from the other dissimilar pixel pairs, including query white horse against support brown horse and any other query non-horse region against support non-horse region. However, while tackling a few-shot instance localization task, we expect to retrieve any *potential* pixel pairs across the query image and the support image for further recognition. Unfortunately, such sparse attention derived from the dot product and $\mathrm{softmax}$ could reduce these potential pairs' recall and hence degrade the model performance. The second issue is observed from another self-attention mechanism, *i.e*., Transformer, which integrates multi-headed multi-layer dot-product calculations concerning multiple representation sub-spaces. Though this manipulation could relieve the first issue; however, the additional computational cost becomes the other issue.

This work discusses two few-shot instance localization tasks: few-shot action localization [4, 13, 23, 44, 45] and one-shot object detection [6, 19]. The former task aims to localize the unseen-class action within an untrimmed video sequence, where the unseen-class action is hinted at by one support set of trimmed videos possessing that unseen-class action. The latter task aims to localize within an image the unseen-class object, which is hinted at by one support image possessing that unseen-class instance. In general, the query-support dependency is essential in addressing the few-shot instance localization tasks. The mentioned attention mechanisms of the non-local block [39] and Transformer [37] are usually used in practice, for example, NLB-based methods [19, 23, 44] and Transformer-based methods [6, 45], to construct such a dependency for making the query aware of the unseen class of interest indicated by the support set.

The proposed *Aggregating Bilateral Attention* aims to be a low computational cost query-support attention mechanism, which can increase the localization recall to benefit the few-shot instance localization tasks. Our core idea is to increase the number of informative pairwise affinities across the query and the support, where the informative affinities trigger the high-recall region proposals generated from a localization model. To this end, our ABA discovers these informative pairwise affinities by integrating the affinities deriving from the proposed *embedding norm* and *context awareness*. The embedding norm employs the p-norm to measure query-support feature distance per data pair. The measurement reduces the similarity differences between query-support pairs and hence increases the chance of discovering the data pairs with minor similarities. The context awareness employs the support context information captured by global average pooling. The captured support context guides the query-support affinity calculation to make aware of the class of the salient instance depicted in the support. As a result, integrating the affinities from the two measurements mentioned above concerns not only discovering the data pairs of various similarities but also focusing on the support context information. Besides, in order to achieve low computational cost, ABA embeds data with a higher dimension reduction ratio and discards the design of multi-headed or multi-layer. Moreover, since the affinities derived from the two measurements may not be equally important, our affinity matrices fusion step employs convolution for learning the fusion. Experiments show that our design benefits the few-shot instance localization tasks. We identify the main contributions of ABA as follows:

- We introduce novel query-support affinity aggregation accounting for the learned embedding norm and the context-aware similarity to enhance the query feature concerning the support. The proposed ABA is lightweight and easy to be incorporated into existing networks to yield informative query-support affinities.
- We conduct extensive experiments to demonstrate that ABA improves existing models to achieve state-of-the-art performances on both few-shot action localization and one-shot object detection.

2. Related work

Attention mechanism. The attention mechanism has demonstrated its advantages in addressing linguistic-related [1, 2, 37] and vision-related [6, 10, 21, 35, 39] tasks. Concerning the pairwise correlation on input data, an attention mechanism is able to capture long-range dependencies for enhancing the data representations. For example, Deng *et al*. [9] propose accumulating three attentions of query, image, and objects for distilling information while excluding the noise. Wu *et al*. [40] employ a non-local block to associate long-term features. Xu *et al*. [43] modulate the relationship between visual and linguistic representation to better produce image captions via an attention mechanism. Recently, several other attention mechanisms have been proposed, such as channel attention mechanism [14], disentangled non-local neural networks [46], non-local blocks ensemble method [50], and additional terms for the auxiliary position [21, 22]. Rather than directly employing the existing attention mechanisms, we present a novel aggregating bilateral attention that distills the informative ingredients to form robust attention via two affinity measurements.

Few-shot action localization. Concerning the type of the support set, we briefly categorize the few-shot action localization tasks as *sentence-supported* [8, 15, 18, 47], *video-supported* [13, 44], and *image-supported* [49] tasks. The sentence-supported few-shot action localization attempts to localize within an untrimmed query video the video segments matched with the support sentences. Chen *et al*. [8] propose to incorporate local and global video features via a

non-local block. Zhang *et al*. [47] use an iterative graph adjustment network to associate the proposal encoding with temporal structural reasoning. The video-supported few-shot action localization aims to retrieve the video segments comprising the action instance hinted by a few trimmed support videos. Feng *et al*. [13] propose a cross-gated bilinear matching approach to align the support video into the untrimmed query video semantically. Yang *et al*. [44] employ a non-local block to correlate the representations of the query video's action proposals with the support videos. With the Transformer mechanism, Yang *et al*. [45] design a few-shot Transformer with a dedicated encoder-decoder structure to localize action instances more accurately. Unlike the video-supported few-shot action localization, the image-supported few-shot action localization retrieves the video segments with stricter conditions, namely, being supported with the image patches sacrificing temporal information. Zhang *et al*. [49] apply an attention-based mechanism to correlate the objects described within the support images with the query video. In this paper, we plug in ABA with FSCAL [44] to gain more improvements than the original FSCAL, as demonstrated in the experiments.

Few-shot object detection. The few-shot object detection tasks can be grouped concerning the learning strategies, such as transfer-learning [7], metric-learning [25, 29, 31, 34, 38, 41], meta-learning [20, 24], or contrastive-learning [12]. The transfer-learning-based approach [7] aims to mitigate the over-fitting issue over the few unseen images by using a regularization approach. The goal of the metric-learning-based few-shot object detection methods [25, 29, 31, 34, 38, 41] is to construct a learn-able metric classifier for reasoning the unseen classes hinted by a few labeled examples. The meta-learning-based method [24] leverages a trained few-shot meta-model to refine the image representations derived from a detector. Hu *et al*. [20] introduce a variant of the non-local block by replacing a single embedding layer with multiple convolutional layers to enhance co-existing features among the query image and the support images. The contrastive-training-based approach [12] associates the query image and the support images through an attention-based RPN and multi-relation detector. While using merely one support image as a more strict constraint, Hsieh *et al*. [19] suggest employing a non-local block to mutually associate the features from support-image and query-image. Osokin *et al*. [32] semantically align features of query-image and support-image. Chen *et al*. [6] propose the Transformer structure to correlate the query image proposals and the support image patch. This paper plugs in our ABA with CoAE [19] and AIT [6] to improve these two one-shot object detection methods.

3. Preliminaries

Problem definition. The problem of few-shot instance (query-support) localization follows the standard M-way K-shot protocol to discriminate within the query all the instances of the M classes hinted by the K-labeled data per class. Assume that the localization task concerns data from totally $\mathcal{N} = \bar{\mathcal{U}} \cup \mathcal{U}$ classes, where the mutually exclusive $\bar{\mathcal{U}}$ and $\mathcal{U}$ ($\bar{\mathcal{U}} \cap \mathcal{U} = \emptyset$) denote the seen classes and the unseen ones in training, respectively. Note that in inference the M classes in a support set could be from either $\bar{\mathcal{U}}$ or $\mathcal{U}$.

In the few-shot instance localization task, a feature point $\mathbf{x}$ encodes the proposal/frame-level feature from a video input or the pixel-level feature from an image input. For the video input, we follow [44] to adopt the C3D backbone [36] for encoding each query video as a tensor in $\mathbb{R}^{t\times d}$ and each support video as $S \in \mathbb{R}^{t\times d}$, where t denotes the temporal dimension and d denotes the feature dimension. Following [44], we then use the proposal subnet, R-C3D [42], to retrieve the proposal-level representation of the encoded query video as $Q \in \mathbb{R}^{p\times d}$, where p denotes the number of proposals. Hence, we denote one proposal-level feature point from Q as $\mathbf{x}^Q \in \mathbb{R}^{1\times d}$ and one support feature point from S as $\mathbf{x}^S \in \mathbb{R}^{1\times d}$. For the image input, we follow CoAE [19] to employ the ResNet backbone [16] for encoding the query image as $Q \in \mathbb{R}^{hw\times d}$ and each support image as $S \in \mathbb{R}^{hw\times d}$, where h and w denote the height and width of one feature channel, respectively. Hence, we denote one pixel-level feature point from Q as $\mathbf{x}^Q \in \mathbb{R}^{1\times d}$ and one pixel-level feature point from S as $\mathbf{x}^S \in \mathbb{R}^{1\times d}$.

The non-local block [39] is designed to capture the self-attended long-range dependencies by densely correlating every data pair from one input. To address the few-shot instance localization task, various attention-related models are proposed to calculate the query-support attention, such as FSCAL [44] for the few-shot action localization task, CoAE [19] and AIT [6] for the one-shot object detection task. It is worth mentioning that Transformer [37] has shown that stacking multiple non-local blocks for simultaneous encoding intra-attention and inter-attention can boost the performance of several attention-based models.

Non-local block. The non-local block is one sort of self-attention modeled as the non-local mean [3], which formulates the correlation between each element with all the other elements. Hence, each element can correlate itself with all data elements as one sort of long-range dependency. Formally, taking an image I, we denote $\mathbf{x}_i \in I$ as the image's pixel-level feature at position i, and the attended feature $\mathbf{y}_i$ output from a typical non-local block is defined as

$$\mathbf{y}_i = \frac{1}{Z(\mathbf{x})} \sum_{j\in\Omega} \omega(\mathbf{x}_i, \mathbf{x}_j)\theta(\mathbf{x}_j)\,, \tag{1}$$

where Ω denotes all the pixels of I, the function ω calculates the embedded feature similarity of the pixel pair $\mathbf{x}_i$ and $\mathbf{x}_j$. The function θ forms a *value* embedding, and $Z(\mathbf{x})$ denotes the normalization factor. The function ω calculates the pairwise similarity by employing the dot product as

$$\omega(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\intercal \rho(\mathbf{x}_j) = \sum\nolimits_e \phi(\mathbf{x}_i)_e \times \rho(\mathbf{x}_j)_e \,, \quad (2)$$

where $\sum_e$ denotes summation over channels, ϕ and ρ respectively denote the *query* and *key* embedding functions, $\times$ denotes the multiplication operation, and the symbol e denotes the index of the feature dimension after embedding. While inserting the non-local block into a neural network, the attended feature $\mathbf{y}_i$ is usually accompanied by additional linear transformation ψ and a residual connection as the enhanced feature, $\mathbf{x}_i + \psi(\mathbf{y}_i)$, as shown in Figure 2 (a).

Mutual enhancement block. One existing query-support attention modeling is to interchange the query-support features via two non-local blocks [6, 19, 44]. Figure 2(b) shows an example of the mutual enhancement block in FSCAL [44] for reference, and another way to employ a Transformer module for carrying out such query-support attention can refer to [6, 37, 45] for more details. Briefly, the mutual enhancement block takes query and support with the following measurements:

$$\begin{aligned} \mathbf{y}_i^Q &= \frac{1}{Z(\mathbf{x})} \sum_{j \in S} \omega(\mathbf{x}_i^Q, \mathbf{x}_j^S) \theta(\mathbf{x}_j^S) \,, \\ \mathbf{y}_j^S &= \frac{1}{Z(\mathbf{x})} \sum_{i \in Q} \omega(\mathbf{x}_j^S, \mathbf{x}_i^Q) \theta(\mathbf{x}_i^Q) \,, \end{aligned} \quad (3)$$

where i and j respectively denote one pixel-level feature from Q and S, function ω measures the pairwise similarity in eq. (2). Hereafter, each output of eq. (3), *i.e.*, $\mathbf{y}$, linearly embeds itself, *i.e.*, $\psi(\mathbf{y})$, for adding back with $\mathbf{y}$.

4. Aggregating bilateral attention

Unlike the previous attention-based methods [6, 19, 37, 39, 44] to capture long-range dependencies using pairwise dot-product measurement, our ABA, as shown in Figure 2(c), distills the feature attention derived from embedding norm and context awareness. Specifically, the ABA integrates two kinds of query-support pairwise affinities via two measurements: *i)* one affinity derived from the distance-based *embedding norm* shows the ability to retrieve more semantically-related data; *ii)* the other affinity derived from *context awareness* shows the ability to focus on the salient support instance via the captured support context information. Moreover, we provide an effective *affinity matrices fusion* mechanism to integrate the two affinities. Experiments show that the ABA helps increase the localization recall for the few-shot instance localization tasks.

4.1. Embedding norm ω^N

This measurement aims at retrieving more semantically-related data neighboring within an embedding space. The embedding norm calculates per query-support pair's distance within a learned embedding space, which treats the locally neighboring data as the potential candidates comprising similar (semantically-related) classes. We carry out the embedding norm by p-norm, defined as

$$\|\mathbf{x}_i^Q, \mathbf{x}_j^S\|_p = \Big(\sum\nolimits_e (\phi(\mathbf{x}_i^Q)_e - \rho(\mathbf{x}_j^S)_e)^p \Big)^{\frac{1}{p}} \,, \quad (4)$$

where the functions ϕ and ρ represent the linear embeddings, and the symbol e denotes the index of the feature dimension after embedding. To convert the distance into similarity, we employ an operation as

$$\omega^N(\mathbf{x}_i^Q, \mathbf{x}_j^S) = \frac{1}{1 + \|\mathbf{x}_i^Q, \mathbf{x}_j^S\|_p} \,. \quad (5)$$

Without loss of generality, 2-norm is the default setting.

Discussion. The proposed embedding norm employs the p-norm to measure query-support feature distance per data pair. Compared with the typical dot-product, such a distance-based measurement replaces the *per channel pairwise* multiplication in eq. (2) with subtraction in eq. (4). As a result, the replacement slows down the similarity accumulation over channels per data pair, especially the high-similarity one. In this way, the embedding norm reduces the difference between a high-similarity data pair and other data pairs. It increases the chance of discovering the medium-similarity data pair, *e.g.*, the query white horse against support brown horse, or even the low-similarity data pair, *e.g.*, the query person against support brown horse, as shown in Figure 1. On the other hand, reducing the differences between query-support pairs means a higher chance of discovering the potential instances, namely, getting rid of merely focusing on the sparse high-similarity instances as the typical NLB does in the scenario of few-shot localization.

4.2. Context awareness ω^C

The context awareness aims to retrieve the instance class concerning the salient instance in the support. This affinity measurement forces the query data to consider the support context by employing the global average pooling and the dot-product. Precisely, we measure pairwise similarity as

$$\omega^C(\mathbf{x}_i^Q, \mathbf{x}_j^S) = [\phi(\mathbf{x}_i^Q) \odot \sigma(\mathsf{GAP}(\rho(S)))]^\intercal \rho(\mathbf{x}_j^S) \,, \quad (6)$$

where σ denotes the sigmoid function, GAP means the global average pooling, $\odot$ denotes the channel-wise scaling, and ϕ and ρ denote the linear embedding functions. In eq. (6), $\sigma(\mathsf{GAP}(\rho(S)))$ is a channel scaling vector derived from the support S to re-weight each channel of $\phi(\mathbf{x}_i^Q)$.

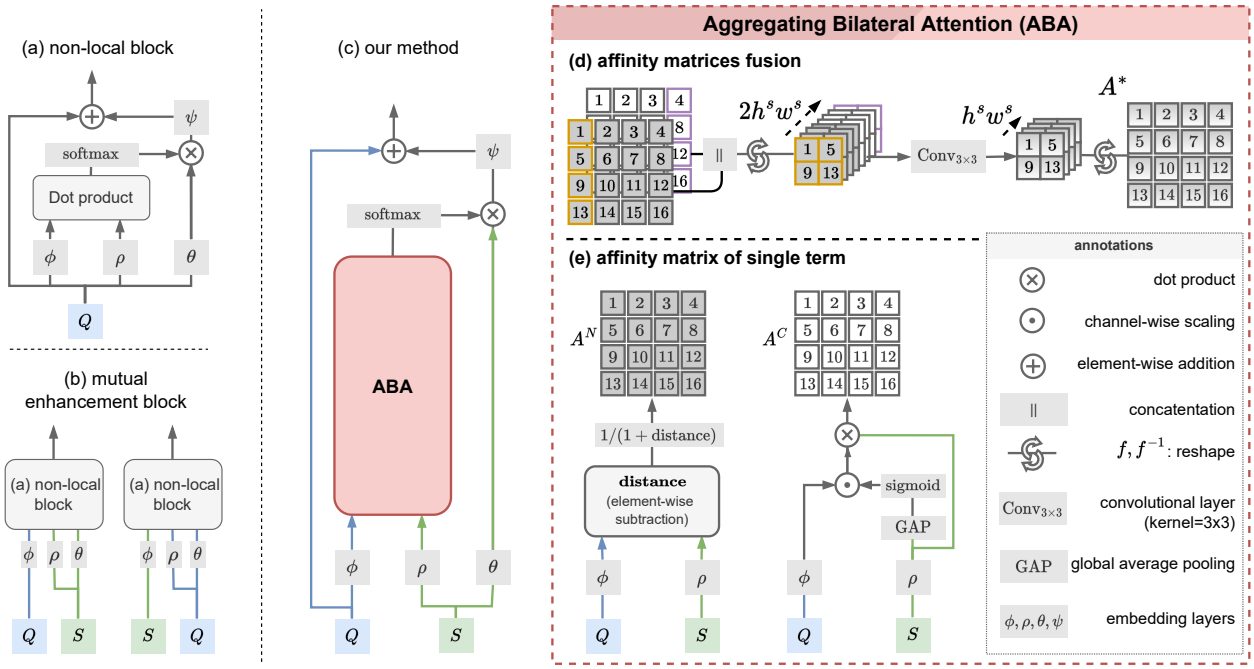

Figure 2: **Comparison of various attention mechanisms.** (a) *non-local block* [39] for capturing the self-attended long-range dependencies. (b) *mutual enhancement block* [44] employs two non-local blocks to capture query-support attention. (c) *Aggregating Bilateral Attention* fuses two query-support affinities to enhance the query data. (d) The affinity matrices fusion integrates two affinity matrices concerning the spatial data neighborhood. (e) The two query-support affinity measurements generate the affinity matrices A^N and A^C. The symbols Q and S denote the inputs from the query and support, respectively.

Discussion. The context awareness measures query-support data pairs' affinities concerning the support context captured by global average pooling. In each query-support data pair, the corresponding affinity is formulated as one query datum against the entire support context describing the salient instance within the support. Such a design shows the ability to enforce the query focus on the salient support class, whether the class is seen or unseen. For example, the successfully detected three horses shown in Figure 1.

4.3. Bilateral affinity fusion

We carry out the affinity matrices fusion concerning the spatial data neighborhood. Suppose the query and the support are of the spatial resolution $h^q \times w^q$ and $h^s \times w^s$, our fusion mechanism aims to integrate the two affinity matrices $A^N \in \mathbb{R}^{h^q w^q \times h^s w^s}$ and $A^C \in \mathbb{R}^{h^q w^q \times h^s w^s}$, resulting from ω^N and ω^C, into the final affinity matrix $A^* \in \mathbb{R}^{h^q w^q \times h^s w^s}$. To this end, our fusion mechanism is defined as

$$A^* = f^{-1}\left(\mathrm{Conv}\left(f(A^N) \| f(A^C)\right)\right), \tag{7}$$

where the function $f : \mathbb{R}^{h^q w^q \times h^s w^s} \rightarrow \mathbb{R}^{h^q \times w^q \times h^s w^s}$ reshapes a 2D affinity matrix in a 3D representation of $height \times width \times channel$, $\|$ denotes the concatenation over the channel dimension, the convolution function $\mathrm{Conv} : \mathbb{R}^{h^q \times w^q \times 2h^s w^s} \rightarrow \mathbb{R}^{h^q \times w^q \times h^s w^s}$ employs the number of $h^s w^s$ 2D kernels on the concatenated affinities, and the function $f^{-1} : \mathbb{R}^{h^q \times w^q \times h^s w^s} \rightarrow \mathbb{R}^{h^q w^q \times h^s w^s}$ reshapes a 3D affinity matrix back to a 2D representation. In eq. (7), we can integrate two affinity matrices concerning each query pixel's *spatial neighborhood* defined in a 2D kernel of the convolution function Conv.

Discussion. Two affinity matrices generated from ω^N and ω^C are used to discover the semantically-related data yet aware of the instance class in support. Since the affinities derived from the two measurements may not be equally important, our fusion step concerns spatial data neighborhood to employ convolution for learning the fusion.

Complete mechanism. The complete aggregating bilateral attention mechanism enhances each feature of the query Q as

$$\mathbf{z}_i^Q = \mathbf{x}_i^Q + \psi\left(\mathrm{softmax}(A^*_{i,j})\theta(\mathbf{x}_j^S)\right), \tag{8}$$

where i and j respectively denote each feature point from Q and S, $A^*_{i,j}$ denotes the fused pairwise similarity between $\mathbf{x}_i^Q$ and $\mathbf{x}_j^S$, softmax normalizes the affinity values along the dimension j, and ψ means a linear embedding function.

5. Experiments

5.1. Datasets and metrics

Few-shot action localization. We follow the previous methods [13, 44] to reconstruct the ActivityNet-1.3 video dataset for evaluating the few-shot action localization models. The ActivityNet-1.3 dataset includes 14,950 annotated videos of 200 action classes. Video sequences containing multiple instances are decomposed into independent videos of one instance according to methods [13, 44]. Those video sequences with more than 768 frames are neglected, and the rest video sequences are randomly selected concerning action classes to obtain the training, validation, and testing splits in the ratio of 80%, 10%, and 10%, respectively. Experiments are conducted on the setting of one single instance within the target untrimmed video. We then evaluate a few-shot action localization model equipped with our ABA mechanism with the mean Average Precision (mAP) metric as [17]. Since most evaluations under some specific overlap thresholds are unclear in [44], we report the mAP at the overlap of 0.5, *i.e.*, mAP@0.5 ($\%$), on this task.

One-shot object detection. We evaluate models on two standard datasets, PASCAL-VOC [11] and MSCOCO [30], using the same settings as [5, 19, 28, 31]. In the PASCAL-VOC dataset, we employ the 'PASCAL-VOC 2007 train&val' and 'PASCAL-VOC 2012 train&val' for training and use the 'PASCAL-VOC 2007 test' for testing. Following the same setting as methods [19, 48], we consider the object classes to organize the PASCAL-VOC dataset. For such a dataset of 20 object classes, we form the training and testing split with ratios of 80% and 20%, respectively. For the MS-COCO dataset, the model training employs the 'train 2017' split, and the model testing uses the 'val 2017' split. We obtain four groups to test the model's generality by separating the 80 object classes under criterion [48]. Three groups of 60 object classes serve as the training split comprising seen classes, while the rest group of 20 object classes is provided as the unseen testing split. We adopt the CoAE [19] protocol to prepare the target-query image pairs; readers are referred to [19] for data preparation details. In evaluating phase, we report the averaged Average Precision (AP) scores for the first five sampled query image patches to ensure consistent statistics. Notice that in the MS-COCO dataset, we follow previous methods to evaluate with the metric AP50, *i.e.*, AP with a fixed IoU threshold at 50%.

5.2. Implementation details

Few-shot action localization. We select the few-shot common action localization 'FSCAL' [44][1] for directly replacing their non-local block with the proposed aggregating bilateral attention mechanism. Since FSCAL is a non-local block based approach, we denoted it as FSCAL (NLB). The resulting model after the attention mechanism replacement, *i.e.*, 'FSCAL (NLB-ABA),' is optimized with Adam optimizer initiated by a learning rate of $1e$-5 with one NVIDIA GTX 1080Ti GPU. We train 'FSCAL (NLB-ABA)' with a batch size of 1 and then decay the learning rate to $1e$-6 after 25K iterations over 40K iterations. For a fair comparison, we use the same video encoder of the C3D backbone [36] pre-trained on Sports-1M [26] in which the unseen-class actions in the ActivityNet-1.3 dataset are excluded. To retrieve a high-quality action proposal set, we employ the R-C3D [42] to obtain various action proposals followed by filtering out foreground-irrelative proposals with a confidence score lower than 0.3. We keep 128 and 300 proposals after the non-maximum suppression for the training and testing, respectively. The input features are altered with a dimension reduction ratio of 8 in functions ρ, ϕ, and θ.

	Measurements		Affinity Matrices Fusion	One-Shot Object Detection				
	ω^N	ω^C	Configurations	cow	sheep	cat	aero	mAP
1	✓	-	n/a	84.7	68.7	79.1	51.7	71.1
2	-	✓	n/a	**85.6**	70.8	79.6	49.4	71.3
3	✓	✓	$A^N \odot A^C$	83.4	68.7	78.4	48.5	69.7
4	✓	✓	$\sigma(A^N \odot A^C)$	83.6	70.0	80.2	46.9	70.2
5	✓	✓	$\sigma(A^N) \odot \sigma(A^C)$	82.7	69.4	79.5	50.0	70.4
6	✓	✓	$\mathrm{Conv}_{1\times1}(A^N \| A^C)$	83.5	69.5	76.5	46.7	69.1
7	✓	✓	$f^{-1}\left(\mathrm{Conv}_{1\times1}\left(f(A^N) \| f(A^C)\right)\right)$	84.7	70.9	80.8	53.5	72.5
8	✓	✓	$f^{-1}\left(\mathrm{Conv}_{5\times5}\left(f(A^N) \| f(A^C)\right)\right)$	83.9	70.2	**81.8**	51.4	71.8
9	✓	✓	$f^{-1}\left(\mathrm{Conv}_{3\times3}\left(f(A^N) \| f(A^C)\right)\right)$	84.0	**73.4**	**81.8**	**53.6**	**73.2**

Table 1: **Ablation Study** on one-shot object detection task using PASCAL-VOC dataset in terms of AP score (%). The three configuration sets of the affinity matrices fusion step from top to bottom concern *single affinity*, *naive affinity fusion*, and *our affinity fusion*. The kernel size within a convolution function Conv eq. (7) is denoted as its subscript.

One-shot object detection. We select the one-shot object detection models of 'CoAE (NLB)' [19][2] and 'AIT (TF)' [6][3] for directly replacing their non-local block and Transformer mechanism with the proposed ABA. The resulting models, *i.e.*, 'CoAE (NLB-ABA)' and 'AIT (TF-ABA),' are optimized via an SGD optimizer with a momentum of 0.9 for ten epochs using four V100 GPUs in parallel. We train 'CoAE (NLB-ABA)' and 'AIT (TF-ABA)' with a batch size of 32 and scale the learning rate, initiated at 0.01, by 0.1 degradedly for every four epochs. For a fair comparison without foreseeing the unseen-class objects during the training, we exclude the PASCAL-VOC and MS-COCO related classes in the dataset to get the reduced ImageNet. The reduced ImageNet has 933,052 images of 725 classes to train the initial weight of the backbone. The input features have a dimension reduction ratio of 2 in functions ρ, ϕ, and θ.

[1] https://github.com/PengWan-Yang/commonLocalization

[2] https://github.com/timy90022/One-Shot-Object-Detection

[3] https://github.com/WOMMOW/AIT

5.3. Ablation study

The experiment in this part compares different configurations of the proposed aggregating bilateral attention for assessing each component's effectiveness. To analyze the proposed aggregating bilateral attention mechanism, we take the one-shot object detection model of CoAE [19] for equipment with our ABA mechanism in this experiment. Table 1 summarizes the results of various configurations, including *single affinity*, *naive affinity fusion*, and *our affinity fusion*.

Single affinity. Row 1 and row 2 in Table 1 show the results of replacing the *dot-product* within the non-local block in the CoAE model, as shown in Figure 2 (a), with the proposed affinity A^N or A^C, deriving from eq. (5) or eq. (6), respectively. Compared with the baseline model, *i.e.*, CoAE using a typical NLB in Table 3, both affinities are able to improve more 1.8% mAP scores over the unseen classes.

Naive affinity fusion. Row 3 to row 6 in Table 1 show the various naive affinity fusion strategies, such as the element-wise product ($\odot$), convolution ($\mathrm{Conv}_{1\times1}$), and sigmoid activation (σ). It shows that naively integrating the affinities *element-wise* does not work for affinity fusion.

Our affinity fusion. In Table 1, row 7 to row 9 considers the various kernel size of the convolution function (Conv) in eq. (7). The results demonstrate that our affinity fusion within the spatial domain can successfully integrate the two affinities derived from the *embedding norm* and *context awareness*. All the fusion of different kernels contributes positively. According to the ablation study, we adopt a kernel size of 3×3 in the following experiments. The possible reason for such a successful fusion is the *spatial data neighborhood* represented on the spatial domain, enabling the fusion mechanism to concern each query pixel's affinity with its neighbors' affinities within the neighborhood defined by the convolution $\mathrm{Conv}_{3\times3}$ in eq. (7). As a result, the robust affinity correlating query and support could be retrieved.

5.4. State-of-the-art comparison

The experiments here compare different methods tackling the few-shot instance localization tasks to assess the efficacy of our ABA mechanism. Furthermore, we also compare the model efficiency with three metrics and visualize some examples to realize our model better.

Few-shot action localization. Table 2 shows the comparison results with the state-of-the-art methods on the few-shot action localization task with the ActivityNet-1.3 dataset. The state-of-the-art methods in this comparison experiment include Buch's model [4], Hu's model [23], Feng's model [13], FSCAL (NLB) [44], and Yang's model [45].

Methods	1-shot	2-shot	3-shot	4-shot	5-shot
Buch's model [4]	-	-	-	-	39.7
Hu's model [23]	41.0	-	-	-	45.4
Feng's model [13]	43.5	-	-	-	-
FSCAL (NLB) [44]	53.1	53.8	54.9	55.4	56.5
Yang's model [45]	57.5	-	-	-	60.6
FSCAL (NLB-ABA)	56.9	57.1	57.8	57.9	58.0
FSCAL (TF-1L-1H)	56.2	56.6	56.8	56.2	57.2
FSCAL (TF-6L-8H)	57.1	58.3	58.8	59.1	59.7
FSCAL (TF-1L-1H-ABA)	**60.7**	**60.9**	**61.5**	**61.6**	**61.2**

Table 2: **State-of-the-Art Comparison** on few-shot action localization task using ActivityNet-1.3 dataset in terms of mAP@0.5 score (%). '-': not available.

We build the FSCAL (NLB-ABA) model after replacing the within FSCAL (NLB) non-local block with our aggregating bilateral attention mechanism. The results in Table 2 show the consistent positive performance gains of FSCAL (NLB-ABA) in 3.8%, 3.3%, 2.9%, 2.5%, and 1.5% mAP@0.5(%) over five different shot settings compared to FSCAL (NLB). While equipping the FSCAL with a typical six-layer-eight-head Transformer, *i.e.*, 'FSCAL (TF-6L-8H)', it can boost the FSCAL (NLB) by 3.2% mAP@0.5 within the 5-shot scenario. However, by replacing the Transformer's inter-attention module with our ABA, *i.e.*, 'FSCAL (TF-1L-1H-ABA)', we can boost the model again even with a simplified Transformer merely leveraging one layer and one head.

One-shot object detection. Table 3 shows the comparison results with the state-of-the-art one-shot object detection methods on the PASCAL-VOC dataset. The methods in this experiment include SiamFC [5], SaimRPN [28], CompNet [48], CoAE (NLB), and AIT (TF), where the CoAE (NLB) is a non-local block based method and AIT (TF) is a Transformer based one. The result in Table 3 shows that 'CoAE (NLB-ABA)' gains clear improvement against CoAE (NLB) for most object classes while replacing the NLB with our ABA, demonstrating the ABA's contribution. Surprisingly, 'CoAE (NLB-ABA)' shows better results than the transformer-based model, 'AIT (TF).' Replacing the Transformer's inter-attention module with our ABA, *i.e.*, 'AIT (TF-ABA),' can obtain the best performance.

Model efficiency. We employ the open-source tool[4] to evaluate the model efficiency. The evaluation metrics include the trainable parameters (Params), FLOPs, and computational latency per video (Latency); the lower value per metric means better model efficiency. Each query and support video is encoded with a C3D backbone using the input resolution of $3 \times 768 \times 112 \times 112$ and $3 \times 64 \times 112 \times 112$, respectively. The experiment is conducted upon FSCAL-

[4] https://github.com/Lyken17/pytorch-OpCounter

Methods	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 (NLB)	43.9	59.7	72.0	74.5	53.6	64.6	21.7	68.8	**85.2**	86.3	82.1	80.7	**85.1**	75.0	77.4	61.1	68.2	84.3	69.8	83.2	49.9	71.8
AIT (TF)	**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 (NLB-ABA)	46.7	**66.6**	**72.7**	73.3	**53.7**	**65.3**	25.5	**70.9**	84.4	86.1	**84.2**	78.1	84.9	**76.3**	76.8	60.4	69.1	86.3	72.1	**84.6**	**55.7**	**74.6**
CoAE (NLB)	39.3	53.3	72.9	70.8	49.7	60.7	16.6	65.1	82.3	85.4	79.0	75.8	79.1	**71.8**	74.0	56.8	64.5	85.0	69.1	78.7	44.4	69.3
CoAE (NLB-ABA)	48.5	65.5	73.8	76.5	51.2	60.5	26.6	67.4	**85.1**	**86.6**	81.1	77.8	82.5	71.3	73.3	59.1	67.9	84.0	73.4	**81.8**	**53.6**	73.2
AIT (TF)	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
AIT (TF-ABA)	**51.3**	**68.6**	**76.9**	**78.0**	**55.3**	64.5	**34.5**	**69.9**	85.0	83.7	**83.8**	**81.9**	**83.9**	71.1	**77.3**	**66.1**	**70.7**	84.6	**74.5**	**81.8**	53.1	**73.5**

Table 3: **State-of-the-Art Comparison** on one-shot object detection task using PASCAL-VOC dataset in terms of AP score (%). The top method set is pre-trained on the 1000-class ImageNet dataset, yet the bottom method set is pre-trained on the reduced 725-class ImageNet dataset to avoid foreseeing the unseen classes in testing. CoAE is a re-implementation version.

Methods	Params (M)	FLOPs (G)	Latency (s)
FSCAL (NLB) [44]	5.254	1.614	**0.712**
FSCAL (TF-1L-1H)	22.964	2.976	0.755
FSCAL (TF-6L-8H)	220.355	18.166	0.826
FSCAL (NLB-ABA)	**0.990**	**1.286**	0.716

Table 4: **Model Efficiency.** Comparison of model efficiency with the few-shot action localization method, FSCAL, equipping with typical non-local block, Transformer, and our aggregating bilateral attention mechanism.

based models, where FSCAL employs the proposal subnet of R-C3D for the proposal-level representation.

Table 4 shows the results among four FSCAL configurations. The baseline model, *i.e.*, 'FSCAL (NLB),' is the original FSCAL model using NLB to capture long-range dependency. Besides, we build 'FSCAL (TF-1L-1H)' and 'FSCAL (TF-6L-8H)' by replacing the FSCAL's NLB with the Transformer using one-layer-one-head and six-layer-eight-head, respectively. More layers and heads bring higher model complexities. The bottom row in table 4 shows the efficiency of the FSCAL model while equipped with our *ABA* mechanism. Notice that the metrics of *Params* and *FLOPs* are estimated on merely the attention mechanism, yet the *Latency* metric is estimated on the entire few-shot action localization model. Since the attention block only occupies a small part of the entire model, our attention mechanism does not cause a great model latency reduction.

Visualization. Figure 3 visualizes the attention heat maps in our aggregating bilateral attention mechanism and the non-local block. The top two rows show that our ABA mechanism retrieves better attention results than the non-local block. In these two rows, the non-local block using dot-product struggles to recall the regions hinted by the support image, *i.e.*, bicycle or car. In contrast, our ABA mechanism shows a more remarkable ability to recall these regions of interest owing to the *embedding norm* better correlates semantically-related object classes, such as person/bicycle and traffic-sign/car. Our *context awareness* is designed to retrieve the instance class concerning the salient instance in the support, hence making more attention to the bicycle and car. The bottom row shows that non-local block can focus on the correct regions of the person, yet our ABA completes the regions of the person and partial semantically-related regions of cars. Please refer to supplementary material for more experimental results.

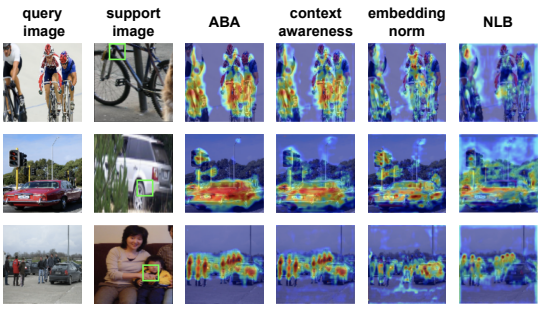

Figure 3: **Attention Heat Map Visualization.** Columns from left to right show the inputs and attention heat maps.

6. Conclusions

This paper points out the issues of sparse attention and high computational cost existing in the previous models. To alleviate these issues, we propose aggregating bilateral attention concerning the embedding norm and context awareness. The former affinity discovers the semantically-related data with minor similarities, and the latter affinity correlates the query data and support regions by focusing on the salient instance depicted in the support. Such an attention mechanism is lightweight and easily integrated into existing models. The experimental results demonstrate the effectiveness and efficiency of our ABA mechanism, which helps existing few-shot instance localization models by raising their localization recall to achieve superior performances.

Acknowledgements. This work was supported in part by the MOST grants 110-2634-F-007-027, 110-2221-E-001-017 and 111-2221-E-001-015 of Taiwan. We are grateful to 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] Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc V. Le. Massive exploration of neural machine translation architectures. In *arXiv*, 2017.

[3] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A non-local algorithm for image denoising. In *CVPR*, pages 60–65, 2005.

[4] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. SST: single-stream temporal action proposals. In *CVPR*, pages 6373–6382, 2017.

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

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

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

[8] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, pages 162–171, 2018.

[9] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *CVPR*, pages 7746–7755, 2018.

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

[11] Mark Everingham, Luc 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.

[12] 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, 2020.

[13] Yang Feng, Lin Ma, Wei Liu, Tong Zhang, and Jiebo Luo. Video re-localization. In *ECCV*, pages 55–70, 2018.

[14] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, pages 3146–3154, 2019.

[15] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: temporal activity localization via language query. In *ICCV*, pages 5277–5285, 2017.

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

[17] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[18] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. In *ICCV*, pages 5804–5813, 2017.

[19] 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.

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

[21] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, pages 3588–3597, 2018.

[22] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *ICCV*, pages 3463–3472, 2019.

[23] Tao Hu, Pascal Mettes, Jia-Hong Huang, and Cees Snoek. SILCO: show a few images, localize the common object. In *ICCV*, pages 5066–5075, 2019.

[24] 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.

[25] 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.

[26] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Fei-Fei Li. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.

[27] Stamatios Lefkimmiatis. Non-local color image denoising with convolutional neural networks. In *CVPR*, pages 5882–5891, 2017.

[28] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, pages 8971–8980, 2018.

[29] Xiang Li, Lin Zhang, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. One-shot object detection without fine-tuning. In *arXiv*, 2020.

[30] 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.

[31] Claudio Michaelis, Ivan Ustyuzhaninov, Matthias Bethge, and Alexander S. Ecker. One-shot instance segmentation. In *arXiv*, 2018.

[32] Anton Osokin, Denis Sumin, and Vasily Lomakin. OS2D: one-stage one-shot object detection by matching anchor features. In *ECCV*, pages 635–652, 2020.

[33] Jiayong Peng, Zhiwei Xiong, Xin Huang, Zheng-Ping Li, Dong Liu, and Feihu Xu. Photon-efficient 3d imaging with A non-local neural network. In *ECCV*, pages 225–241, 2020.

[34] 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.

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

[36] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

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

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

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

[40] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross B. Girshick. Long-term feature banks for detailed video understanding. In *CVPR*, pages 284–293, 2019.

[41] 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.

[42] Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: region convolutional 3d network for temporal activity detection. In *ICCV*, pages 5794–5803, 2017.

[43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

[44] Pengwan Yang, Vincent Tao Hu, Pascal Mettes, and Cees G. M. Snoek. Localizing the common action among a few videos. In *ECCV*, pages 505–521, 2020.

[45] Pengwan Yang, Pascal Mettes, and Cees G. M. Snoek. Few-shot transformation of common actions into time and space. In *IEEE*, pages 16031–16040. Computer Vision Foundation / IEEE, 2021.

[46] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *ECCV*, pages 191–207, 2020.

[47] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S. Davis. MAN: moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, pages 1247–1257, 2019.

[48] Tengfei Zhang, Yue Zhang, Xian Sun, Hao Sun, Menglong Yan, Xue Yang, and Kun Fu. Comparison network for one-shot conditional object detection. In *arXiv*, 2019.

[49] Zhu Zhang, Zhou Zhao, Zhijie Lin, Jingkuan Song, and Deng Cai. Localizing unseen activities in video via image query. In *IJCAI*, pages 4390–4396, 2019.

[50] Lei Zhu, Qi She, Duo Li, Yanye Lu, Xuejing Kang, Jie Hu, and Changhu Wang. Unifying nonlocal blocks for neural networks. In *ICCV*, 2021.

[51] Zhen Zhu, Mengdu Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *ICCV*, pages 593–602, 2019.