Received 11 August 2022, accepted 27 August 2022, date of publication 1 September 2022, date of current version 12 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3203399

RESEARCH ARTICLE

DAFA: Diversity-Aware Feature Aggregation for Attention-Based Video Object Detection

SI-DONG ROH, (Graduate Student Member, IEEE), AND KI-SEOK CHUNG, (Member, IEEE)
Department of Electronic Engineering, Hanyang University, Seoul 04763, South Korea

Corresponding author: Ki-Seok Chung (kchung@hanyang.ac.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by the Korean Government [Ministry of Science and ICT (MSIT)] through the Development of Intelligent Edge Computing Semiconductor for Lightweight Manufacturing Inspection Equipment under Grant 2021-0-00131.

ABSTRACT We present a framework for attention-based video object detection using a simple yet effective external memory management algorithm. An attention mechanism has been adopted in video object detection task to enrich the features of key frames using adjacent frames. Although several recent studies utilized frame-level first-in-first-out (FIFO) memory to collect global video information, such a memory structure suffers from collection inefficiency, which results in low attention performance and high computational cost. To address this issue, we developed a novel scheme called diversity-aware feature aggregation (DAFA). Whereas other methods do not store sufficient feature information without expanding memory capacity, DAFA efficiently collects diverse features while avoiding redundancy using a simple Euclidean distance-based metric. Experimental results on the ImageNet VID dataset demonstrate that our lightweight model with global attention achieves 83.5 mAP on the ResNet-101 backbone, which exceeds the accuracy levels of most existing methods with a minimum runtime. Our method with global and local attention stages obtains 84.5 and 85.9 mAP on ResNet-101 and ResNeXt-101, respectively, thus achieving state-of-the-art performance without requiring additional post-processing methods.

INDEX TERMS Attention mechanism, diversity-aware, neural networks, spatio-temporal, video object detection.

I. INTRODUCTION

Recent advances in deep convolutional neural networks [1], [2], [3], as well as the successful development of object detection networks [4], [5], [6], [7], have driven significant progress in image object detection. However, single-image-based object detectors fail to achieve sufficiently high accuracy when detecting objects in videos, mainly due to severe deterioration effects such as motion blur, partial occlusion, camera defocus, and pose variation. To remedy this problem, video object detection methods commonly utilize spatiotemporal information to enhance the current frame features. In particular, attention-based methods [8], [9], [10], [11] model relationships between object features using attention mechanisms, which are derived from so-called *multi-head attention* [12]. Because the construction of an informative key set is important for attention, these mechanisms collect diverse features from images within a video. Whereas local attention methods [11] utilize features from adjacent frames to construct a key set, global attention methods [8], [10] form a richer key set by collecting features from randomly sampled frames. Most attention-based methods save the extracted features into their external memory structure and reuse them to keep their key set informative. For example, as shown in Fig. 1(a), methods that employ a FIFO-type memory structure [8], [10], [11] sample features from reference frames and collect them in the same order they were sampled.

However, existing external memory structures suffer from *collection inefficiency* due to object-level redundancy, which occurs because most images in a video include objects that do not exhibit significant changes in appearance over time. Methods that employ FIFO-type memory do not

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

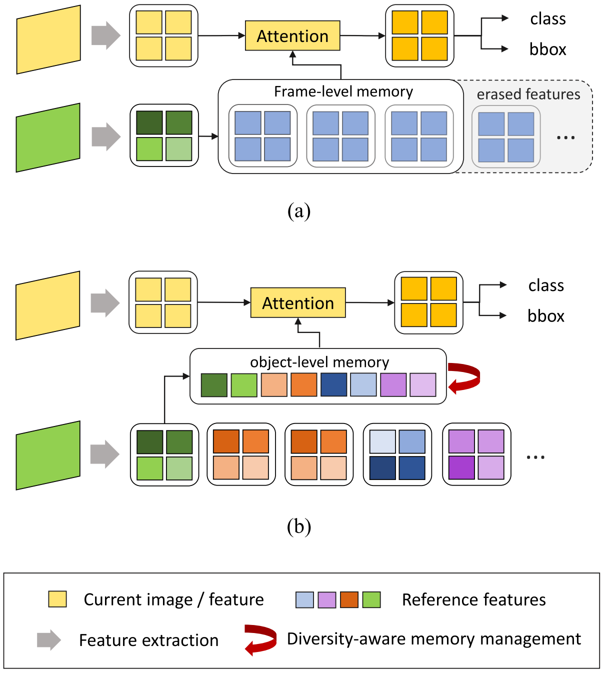

FIGURE 1. Comparison of memory management methods. (a) First-in first-out (FIFO) memory: The oldest frame features in memory are abandoned without considering redundancy. (b) Diversity-aware memory: Our method selects features to be saved in an object-wise manner based on diversity. Without redundancy, all feature information can be collected in compact memory. Best viewed in color.

appropriately account for object-level redundancy, which limits their performance for the following two reasons:

A. DIFFICULTY OF COLLECTING DIVERSE FEATURES

In the FIFO-type memory structure, certain features that are relatively rare yet informative may inevitably be erased in the process of an update because the FIFO approach only considers the sampled time of each frame, not feature information itself. Conversely, these features are less likely to be resampled because they appear less frequently.

B. KEY SET IMBALANCE

Because FIFO-type memory does not handle feature redundancy, semantically similar features can take up the majority of memory space. Although existing methods often employ attention mechanisms with imbalanced memory as the key set, these mechanisms are vulnerable to key set imbalance because the attention score proportionately increases with the number of redundant features, which leads to performance degradation. Therefore, even when informative features are collected in memory, they have limited influence when the attention score is biased toward major redundant features.

To address these issues, we propose a novel framework called diversity-aware feature aggregation (DAFA), which enhances the current feature using a nonredundant and balanced reference feature set. This feature set is managed by a novel memory management scheme called diversity-aware memory management (DAMM), which manages the contents of external memory based on the diversity of features. DAMM attempts to collect diverse features by iteratively selecting a feature that is least similar to the samples. Similarity is estimated using Euclidean distance, which is a simple and natural indicator. Because our method traverses diverse points from the sampled features regardless of their frequency of occurrence, it generates an informative and efficient feature set that improves attention performance. As shown in Fig. 1(b), DAMM collects various object-level features (denoted by color) from a video with a minimal amount of information redundancy, whereas the FIFO-style external memory approach illustrated in Fig. 1(a) does not. Our main contributions are therefore summarized as follows:

- We propose a novel memory management scheme called DAMM for object-level fine-grained key set construction. Using diversity as a quantitative indicator, DAMM ensures efficiency and diversity when collecting global information.
- We propose a video object detection framework called DAFA, which fully aggregates both local and global information into the current feature. Experiments on the ImageNet VID [13] and YouTube Objects dataset [14] demonstrated that DAFA achieves state-of-the-art performance.
- Our experimental results confirm that fine-grained diversity-aware key set construction achieves both high performance and low computational cost in video object detection tasks.

II. BACKGROUND AND RELATED WORK

A. SINGLE IMAGE OBJECT DETECTION

Advances in deep convolutional neural networks (CNNs) have made it possible to achieve excellent performance in image-classification tasks. For instance, ResNet [1] and ResNeXt [3] show excellent performance in classification tasks and they are commonly used as the pretrained backbone for object detection. The single-image object detection task, commonly defined as the problem of inferring multiple object classes and locations in a single image, can be addressed in two ways: a two-stage approach or a one-stage approach. Two-stage object detection typically comprises a region proposal stage and a detection stage. The region proposal network (RPN) infers object proposals in an image, and the cropped features of each object region are fed into the detection network for classification and box refinement. In contrast, single-stage object detection generally omits the region proposal network and detects objects directly from the feature grid. Although both approaches have advantages and disadvantages, modern video object detection networks generally employ the two-stage method owing to the ease of utilizing object-level features.

TABLE 1. Comparison of DAFA with previous attention-based methods.

Methods	OGEMN [9]	SELSA [10]	RDN [11]	MEGA [8]	MAMBA [15]	**DAFA (Ours)**
Attention mechanism?	✓	✓	✓	✓	✓	✓
Use local information?			✓	✓	✓	✓
Use global information?	✓	✓		✓	✓	✓
Object-level memory update?	✓				✓	✓
Memory management based on indirect info?	✓				✓	✓
Memory management based on direct info?						✓

B. VIDEO OBJECT DETECTION

Video object detection (VOD) is more challenging than single-image object detection because image deterioration frequently occurs as a result of object or camera motion. To alleviate this problem, early video object detectors predicted pixel- and object-level movements, and utilized them to aggregate features between the current frame and its adjacent frames. DFF [16], FGFA [17], and THP [18] employ optical flow estimation to aggregate information on spatially adjacent pixels. MANet [19] predicts both pixel- and object-level movements to conduct object-level aggregation. D&T [20] includes both a detection model and a tracking model to complement each other. STSN [21] uses deformable convolution to learn pixel-level feature movements. RNN-based methods [22], [23] model the relationship between extracted features and updated frames to construct long-term memory. Attention-based methods [8], [9], [10], [11], [15] exploit the attention mechanism [12], [24] to train feature dependency. SELSA [10] measures the relationship between proposals and the full sequence by aggregating global features from randomly sampled images. RDN [11] constructs an informative key set by selecting the features of local frames based on objectness scores. MEGA [8] uses both local and global features to improve the measure of dependency and exploits a long-range memory queue to increase the lifetime of the precedent local features. OGEMN [9] and MAMBA [15] presents their own memory management scheme which writes or erase features based on their defined criteria.

C. ATTENTION MECHANISM

Attention mechanism [12], [25] has been proposed as a way to capture the relationship between sequential inputs in natural language processing tasks. The most widely used form of attention is scaled dot-product attention [12]. The formal definition of scaled dot-product attention is as follows.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

It calculates the combined attention weights of a key set K for a query set Q, and scale the weights by dimension d_k. Then the scaled weights are used to calculate a weighted sum. Usually, query is the information we want to know, and key is the candidate information to be combined. value set V is generally same as the key. Recently, the scope of attention mechanism is not limited to sequential inputs and it has been widely used in various vision tasks [8], [24], [26], [27].

D. SAMPLING LOCAL AND GLOBAL INFORMATION IN VIDEO TASKS

We investigated several video-related tasks which deal with local and global information. In the video object detection (VOD) task, recent methods (MEGA [8], SELSA [10], RDN [11] and FGFA [17]) sample features in sequential (local information) or random (global informaion) manner and save them in FIFO. Methods in video object segmentation (VOS) task (such as STCN [26]) mainly focus on local information than the global information, since detailed pixel-by-pixel classification is required. Thus, FIFO is mainly used in VOS task. Methods in the action recognition task (such as TSN [28]) use random sampling to obtain global information, which is similar to the sampling method of SELSA and MEGA in VOD. However, they just use multiple samples and no further management scheme exists. Considering the above investigations, unfortunately, even relatively recent papers still use a simple FIFO or do not have memory management schemes and the limitation of FIFO is not discussed. Due to the structural characteristics of FIFO, memory features that could potentially help current features are inevitably deleted over time. Thus We can expect deterministic information loss in FIFO leads to low performance in its task. Our paper addresses and improves this limitation by presenting a new external memory management method.

E. EXTERNAL MEMORY APPROACH FOR VOD

The neural Turing machine [29], which is an neural networks model designed with an external memory network, demonstrated that external memory outperforms long short-term memory in several memorization tasks. Subsequently, OGEMN [9] first employed an external memory approach to VOD. Unlike previous methods that utilize FIFO memory, the objectness score and class label is a novel criterion for storing and excluding each feature. Furthermore, whereas other approaches [8], [10], [11], [17] save and erase features frame-wise (all object features in a frame are saved or erased at the same time), the minimal save and

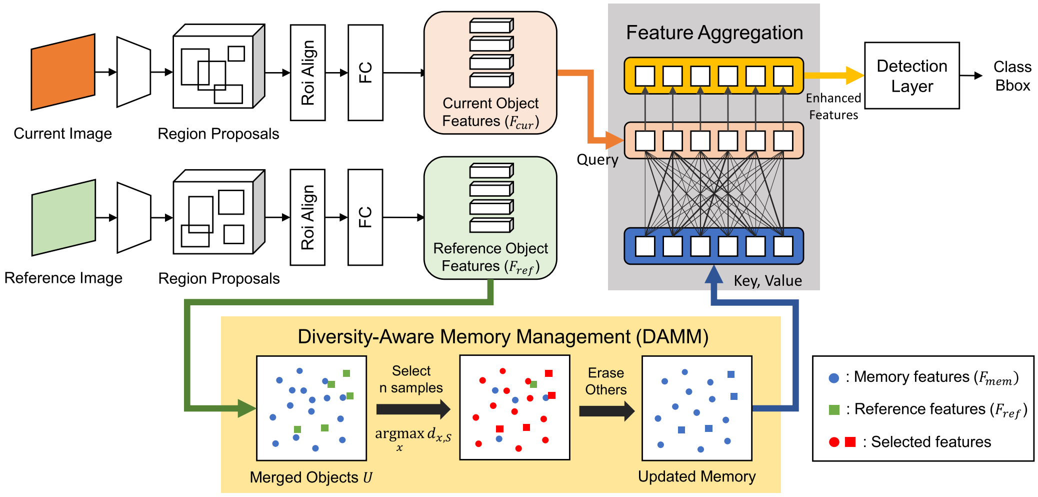

FIGURE 2. Simplified view of the DAFA framework. At each frame time, both the current image and the randomly-sampled reference image are fed into the feature extraction layers. After the region proposal by RPN, extracted features are ROI-aligned by the region of interests (ROIs), and object-level features (F_{cur} and F_{ref}) are generated. The external memory is then updated using the object features (F_{mem} and F_{ref}) as input by Diversity-aware Memory Management (DAMM) algorithm. The updated memory features(F_{mem}) are utilized as a key set of the feature aggregation module to enhance the current feature queries (F_{cur}). Using enhanced current features, the detection layer outputs detection results. Best viewed in color.

erase unit of OGEMN is pixel or object-level feature, which is a more fine-grained approach. MAMBA [15] utilizes a random sampling strategy to read, write, and erase procedures to alleviate memory redundancy and reduce computation costs. Our approach also adopts external memory to manage long-term information. Unlike previously designed external memory methods, our method does not rely solely on indirect information (objectness score, attention score) or randomness, instead using direct information (comparing extracted vectors itself) to manage object features. A comparison of our method with existing approaches is presented in Table 1.

III. OUR APPROACH

A. DIVERSITY-AWARE FEATURE AGGREGATION NETWORK

In this paper, we propose a novel framework called DAFA, which enhances current features using a carefully managed key set. The key set is managed in a fine-grained object-level manner using a novel memory management scheme called DAMM.

The DAFA framework operates as follows: To enhance the current frame features (F_{cur}), a set of features (F_{ref}) are extracted from randomly selected images in the same video sequence as F_{cur}. Subsequently, a memory feature set (F_{mem}) is updated using F_{ref} as the candidate feature. The memory update involves both the collection and deletion processes of object-level features using the DAMM method. After F_{mem} is updated, the feature aggregation module is applied to enhance F_{cur} by using F_{mem} as a key set. Finally, the enhanced features are fed into a detection network to generate the detection results. An overview of DAFA is presented in Fig. 2, and the inference procedure is described in Algorithm 1.

Algorithm 1 Inference Algorithm of DAFA

```
# $I_t$: image at time $t$
# Feat: backbone and region proposal network
# DAMM: diversity-aware memory management
# FA: feature aggregation module
# Detect: detection head for object proposals
Input: $V$ (video sequence)
Output: $B$ (object boxes)
for $i \leftarrow 1$ to $n$ do
    # extract features from reference images
    $I_{ref}$ = random_sample($V$)
    $F_{ref}$ = Feat($I_{ref}$)
    # memory update with merged feature set
    $F_{mem}$ = DAMM($F_{mem} {}^\frown F_{ref}$)
end for
for $I_t$ in $V$ do
    # extract features from current frame image
    $F_{cur}$ = Feat($I_t$)
    # enhance current feature with feature aggregation
    $\hat{F}_{cur}$ = FA($F_{cur}$, $F_{mem}$)
    # detect on enhanced current feature
    $B_t$ = Detect($\hat{F}_{cur}$)
end for
```

B. DIVERSITY-AWARE MEMORY MANAGEMENT

We introduce a novel memory management method called DAMM, which stores the maximum amount of information with a minimum memory cost. Unlike previous methods, our reading and writing protocols were unified into a single operation. DAMM updates the contents of external memory at each frame time to enhance global attention performance.

Assume that F_{mem} is stored in external memory, and F_{ref} is newly sampled. For feature selection, F_{mem} and F_{ref} are first merged into U. We can then define the memory management problem by selecting the least redundant n samples from the candidates in U. However, finding the best set using an exhaustive search is an NP-hard problem. Therefore, we propose a greedy algorithm that sequentially selects the optimal nonredundant feature from U.

We refer to the Euclidean distance between two features x, y as $r_{x,y}$. Although we cannot define an exact threshold for feature redundancy, we can determine that feature redundancy is inversely proportional to $r_{x,y}$. To sequentially select a feature, we then define the diversity between a feature and a set of selected features. A diversity metric between a feature $x \in U$ and feature set S can be calculated as

$$d_{x,S} = \min_{y \in S} r_{x,y}. \tag{2}$$

At each selection iteration, DAMM selects a feature x with largest diversity, which can be defined as follows.

$$\arg\max_{x \in U} d_{x,S} \tag{3}$$

Using this objective, DAMM selects a feature from U at each selection iteration and sequentially appends it to the set S. After n selection iterations, the completely selected set S is updated to the new F_{mem}. Our algorithm is similar to the farthest point sampling (FPS) algorithm, which is widely employed in the point cloud domain [30], [31]. See Algorithm 2 for additional details.

C. FEATURE AGGREGATION MODULE

After F_{mem} is attained by the DAMM algorithm, we update F_{cur} using a feature aggregation module. F_{cur} and F_{mem} are used as query set $Q = \{q_i\}$ and key-set $K = \{k_j\}$, respectively. We used a feature aggregation mechanism similar to those found in [8], [9], [10], and [11], which were inspired by the multihead attention in [12]. However, we did not apply spatiotemporal positional encoding in the global attention stages because the time difference between q_i and k_j can be much longer than that incurred by previous methods. Our feature aggregation module jointly attends to information from diverse perspectives using multiple heads. First, we multiply q_i and k_j by the linear transformation matrices W_Q and W_K, respectively, to project them into the same vector space. Each embedding is then split into M smaller features channel-wise to conduct M multihead attention. We can formulate each item of projected queries (q_i^m) and keys (k_j^m) as

$$W_Q \cdot q_i = \text{concat}[\{q_i^m\}_m^M], \tag{4}$$

$$W_K \cdot k_j = \text{concat}[\{k_j^m\}_m^M], \tag{5}$$

where m denotes the mth attention head. The attention score and weight between a query and key are computed by

$$s(q_i^m, k_j^m) = \frac{(k_j^m)^T (q_i^m)}{\sqrt{d}}, \tag{6}$$

$$w_{ij}^m = \frac{\exp(s(q_i^m, k_j^m))}{\sum_j \exp(s(q_i^m, k_j^m))}, \tag{7}$$

where d is the channel dimension of the split queries and keys. We can then calculate the aggregation for the ith query as

$$\text{agg}(q_i, K) = W_V \cdot \text{concat}[\{\sum_j (w_{ij}^m k_j)\}_{m=1}^M]. \tag{8}$$

where *concat* denotes a channel-wise concatenation and W_K denotes a linear transformation matrix. Finally, the augmented feature f is obtained as follows:

$$f(q_i, K) = q_i + \text{agg}(q_i, K). \tag{9}$$

In the case of multiple aggregation stages, a nonlinear transformation composed of a fully-connected layer and a ReLU activation function is inserted between each pair of stages.

Algorithm 2 Diversity-Aware Memory Management

```
# $m$: number of features in $U$
# $n$: number of features in $S$
# $D$: $m * m$ Euclidean distance matrix
# ds: diversity between features in $U$ and $S$
# sel: indices of selected features
Input: $U$ (candidate features)
Output: $S$ (selected features)
# compute distance matrix
$D$ = compute_euclidean($U$, $U$)
# initialize index array
sel = zeros($n$)
# initially select a feature
ds = $D$[0,:]
for $i \leftarrow 1$ to $n$ do
    # select a farthest one from selected features
    idx = argmax(ds)
    # save selected feature index
    sel[$i$] = idx
    # update diversity metric for updated $S$
    ds = pointwise_min(ds, $D$[idx,:])
end for
$S$ = $U$.index_select(sel)
```

D. ANALYSIS OF MEMORY MANAGEMENT POLICY

A video generally includes information regarding various visual changes caused by rare poses, motion blurs, and occlusions. However, most other information is redundant because cameras capture similar scenes with small temporal gaps. In the context of information entropy, the redundancy removal function of DAMM is effective because it increases entropy and makes memory more informative. In contrast, most previous methods exhibit low entropy as a result of redundant information. In extreme cases where most frames visually overlap, because memory is updated by the FIFO approach, object features are deleted without accounting for priority. Consequently, relevant information that occurs early in the video is deleted, and most of the memory space is occupied by duplicate features. To avoid this, our method rejects low-diversity features, thereby maintaining the diversity of memory even under inputs with highly redundant features. Therefore, our method preserves overall feature information with a smaller memory capacity than existing methods.

Also, DAMM's diversity-based feature selection is robust against deteriorated objects without any algorithmic complementation because DAMM takes objects suggested by RPN as input. RPN returns top-k objects in order of objectness score per image, and thus object candidates of DAMM are limited to features with high objectness scores.

IV. EXPERIMENTS

A. DATASETS AND EVALUATION

We used the same settings as those in FGFA [17] to train our model. Because ImageNet VID does not have sufficient samples, we combined it with the ImageNet DET dataset. Of the 200 categories comprising the ImageNet DET dataset, we used 30 categories that overlap with the ImageNet VID dataset. We evaluated the proposed method using the ImageNet VID validation set, which consists of 3862 videos for training and 555 videos for validation. To compare the object detection performance of our method with that of existing methods, we report mean average precision (mAP) with threshold of IoU > 0.5.

B. BACKBONE AND DETECTION NETWORK

We employed ResNet-101 and ResNeXt-101 as the main backbone architectures. To increase the resolution of extracted features, we changed the stride of the first convolution block in the conv5 stage from 2 to 1. We implemented a two-stage Faster-RCNN network, which has an RPN between the backbone and detection networks. The RPN head predicts objectness scores and box coordinates for 12 anchors (three aspect ratios of {1:2, 1:1, 2:1} and four scales of $\{64^2, 128^2, 256^2, 512^2\}$) for each feature pixel. After the proposals were sorted by objectness scores, we used the top 300 proposals from the current frame and 75 from the reference frames for training and testing, respectively. For each proposal, a 7×7 ROI-align and 1024-D fully-connected layer were used to extract object-level features.

C. IMPLEMENTATION DETAILS OF DAFA

There are two variants of our method: DAFA_G and DAFA_F. Whereas DAFA_G only exploits global attention, DAFA_F initially adds local attention stages to ensure more effective information collection. In the local attention stage, we applied a FIFO-type memory similar to that in RDN [11] or MEGA [8]. The size of the temporal window (T_l) was set to 25. Based on the objectness scores from each frame in the temporal window, 75 object features were selected as base features, and 15 were selected as advanced features. The local stage consists of three phases that gradually enhance the current features. In the first phase, the current frame features and advanced features are enhanced by the base features using a feature aggregation module. In the second phase, the current and advanced features are enhanced by the advanced features. In the final phase, the current features are enhanced by the advanced features. The *Long Range Memory* of MEGA, which was introduced to extend the visible frames of local memory, is not used, as that range is covered by global attention. Features from randomly shuffled images are managed in global memory using DAMM. The global attention stage further enhances the current features with features in global memory. The enhanced current features are then fed into the fully-connected layer to predict the classes and bounding boxes.

D. TRAINING DETAILS

We initialized the backbone using the ImageNet pretrained weights. In both training and testing stages, the input images were resized to a shorter side of 600 pixels. We trained the entire model on four RTX 3090 GPUs with the minibatch size set to one per graphics card for a total batch size of 4. A single minibatch consisted of one current frame, four global reference frames, and two local reference frames. The global reference frames were randomly sampled from a video with the same sequence as the current frame. Because each reference frame had 75 object features, a total of 300 object features were obtained. To mimic the managed global memory in the testing phase, we selected 50% of the global reference features using DAMM. Local reference frames were randomly sampled from adjacent frames within the temporal window. We trained our model using the SGD optimizer with a learning rate of 0.001, momentum of 0.9, and weight decay of 0.0001. We trained the entire architecture over 120K iterations. We set the learning rate to 0.001 for the first 80K iterations and decayed it to 0.0001 for the last 40K iterations. The RPN and detection losses were only calculated for the current frame.

E. MAIN RESULTS

1) ACCURACY COMPARISON

Table 2 compares the accuracy results of the proposed method with those of existing state-of-the-art methods. To ensure fairness, all methods were compared without applying any postprocessing-based re-scoring techniques such as Seq-NMS [32] or BLR [11]. All performance results are based on mAP@50 (mean average precision with IoU > 0.5) unless otherwise stated. Existing methods used for comparison include motion-based methods, such as FGFA, MANET, THP, and STSN, and attention-based methods, such as OGEMN, SELSA, RDN, MEGA, and MAMBA. Under the ResNet-101 backbone settings, our DAFA_F exhibited an almost negligible performance difference (84.5%) compared to MAMBA (84.6%), as summarized in Table 2. Our method with only the global attention module (DAFA_G) achieved 83.5% mAP. Note that DAFA_G achieved higher accuracy than MEGA, which utilizes both local and global attention.

TABLE 2. Accuracy comparison of the video object detection methods for the ResNet-101 (R101) and ResNeXt-101 (X101) backbones on the ImageNetVID validation set.

Methods	Backbone	local	global	mAP(%)
FGFA [17]	R101	✓		76.3
MANet [19]	R101	✓		78.1
THP [18]	R101+DCN	✓		78.6
STSN [21]	R101+DCN	✓		78.9
OGEMN [9]	R101+DCN		✓	80.0
SELSA [10]	R101		✓	80.3
RDN [11]	R101	✓		81.8
MEGA [8]	R101	✓	✓	82.9
MAMBA [15]	R101		✓	84.6
DAFA_G	R101		✓	83.5
DAFA_F	R101	✓	✓	84.5
SELSA [10]	X101		✓	83.1
RDN [11]	X101	✓		83.2
MEGA [8]	X101	✓	✓	84.1
MAMBA [15]	X101		✓	85.4
DAFA_F	X101	✓	✓	85.9

TABLE 3. Comparison of the attention-based methods with the ResNet-101 backbone and the FasterR-CNN detector. The runtime of all compared methods were measured on an RTX3090 GPU.

Methods	stages (local/global)	ref. feats (local/global)	mAP(%)	time(ms)
SELSA [10]	0 / 2	0 / 12600	80.3	123.7
RDN [11]	3 / 0	6105 / 0	81.8	93.1
MEGA [8]	3 / 1	5250 / 750	82.9	121.8
DAFA_G	0 / 2	0 / 900	83.5	54.9
DAFA_F	3 / 1	2625 / 750	84.5	108.1

Our key component, DAMM, computes diversity based on object features, thus preforming better with a more robust feature extractor. With a stronger ResNeXt-101 backbone, our full-featured DAFA_F achieved an 85.9% mAP, which is the highest accuracy among all competitors.

2) SPEED-ACCURACY TRADE-OFF

Table 3 presents a runtime comparison between our method and existing attention-based methods. We also show the number of iteration stages and reference features for a fair comparison. The number of reference frames is not displayed because our method uses fine-grained object-level management. For a quantitative comparison, the number of reference features for each iteration stage was summed. For example, the total number of local reference features of DAFA_F was calculated as $25(T_l)*(75(\text{stage1})+15(\text{stage2})+15(\text{stage3})) = 2625$. We set the number of proposals of RPN in SELSA to 300 and that in other methods to 75 to maintain consistent settings with those used in the original studies. SELSA and RDN are representative examples of methods that employ global and local attention, respectively. SELSA models a global relationship by randomly sampling images throughout a video, whereas RDN models the local relationship around the current feature. MEGA models both global and local relationships with a FIFO-type memory and applies *Long Range Memory* to extend the range of visible frames. As summarized in Table 3, DAFA_F outperformed MEGA by 1.6% while using the same local and global attention stages. Although the two algorithms feature the same number of local and global attention stages, MEGA uses *Long Range Memory* to improve coverage of the local range. Consequently, the number of local reference features is higher than that of DAFA_F, which incurs a higher computational cost. This confirms that DAMM creates a robust feature set even with smaller object features. Fig. 3 displays a comparison of video object detection results for MEGA and DAFA_F. It is apparent our detection results are robust even without the use of *Long Range Memory*. DAFA_G, which uses two-staged global attention, also produced 0.6% higher performance than MEGA while achieving a runtime of 54.9 ms, the lowest among all compared methods.

TABLE 4. Comparision of vanilla attention module [12] and Feature Aggregation module. The module comparison is conducted on DAFA_G model.

Attention Module	Transformer [12]	Feature Aggregation
mAP(%)	83.1	83.5
runtime(ms)	52.7	54.9

TABLE 5. Net effects on the accuracy by adopting various combinations of DAMM and the attention stages.

Model	(a)	(b)	(c)	(d)	(e)
global attention		✓	✓	✓	✓
local attention			✓		✓
DAMM				✓	✓
mAP(%)	75.4	81.8	82.5	83.5	84.5

F. ABLATION STUDY

We conducted an extensive set of experiments to determine the net effects of the key components of DAFA.

1) NET EFFECTS ON ATTENTION MODULE

Table 4 presents differences in accuracy and inference speed. It is apparent that the feature aggregation module produces 0.4% higher performance than the vanilla transformer at a slightly higher runtime (2.2 ms), which verifies that our feature aggregation module enhances overall performance with negligible cost.

2) NET EFFECTS ON ACCURACY ENHANCEMENT

Table 5 presents the net effect on accuracy under various combinations of attention stages and DAMM. Model (a) is the baseline model, which is a single-frame-based Faster-RCNN. Model (b) features two global attention stages with FIFO-type memory. Model (c) includes three local attention stages and one global attention stage. Model (d) has a global attention

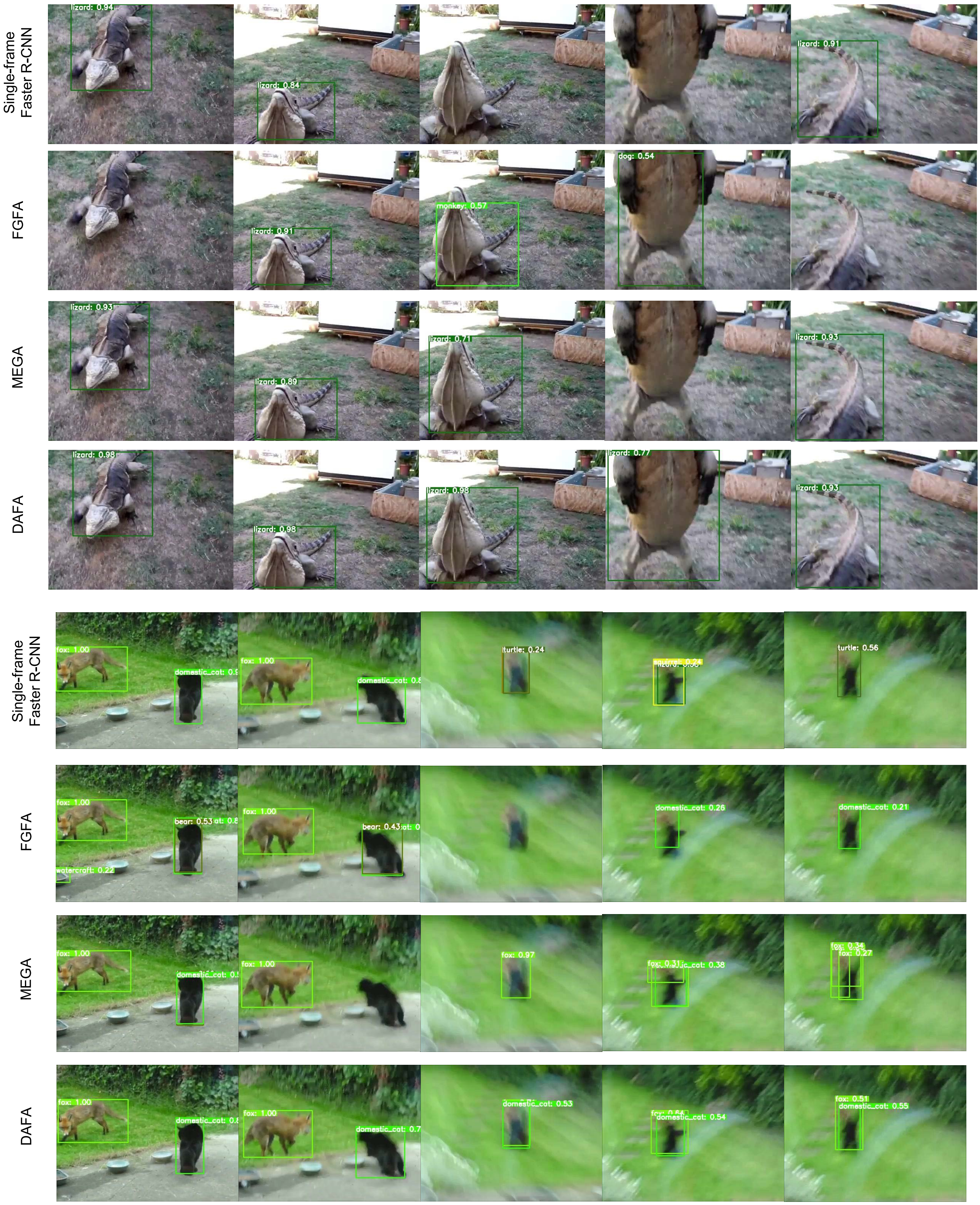

FIGURE 3. Visualized detection comparison between FasterRCNN [31], FGFA [17], MEGA [8], and DAFA_F on two videos in ImageNet VID dataset. The first to fourth rows show the rare pose case, and the fifth to eighth rows show the motion blur case.

model with DAMM, which corresponds to DAFA_G. Model (e) is our full-featured DAFA_F model that employs both local and global attention. By default, we set the global memory size to 750 and the temporal window size to 25. Note that Model (d) achieved 1.7% higher performance than Model-(b), which implies that DAMM yields a significant

TABLE 6. Effect of varying global attention stages of DAFA_G.

N_{stages_g}	1	2	3
mAP(%)	82.8	83.5	83.2
runtime(ms)	51.8	54.9	58.0

TABLE 7. Effect of varying the number of reference frames during training.

N_{ref}	1	2	3	4*	5
FIFO	81.0	81.6	81.7	81.7	81.8
Manage	81.7	82.4	82.7	82.8	82.9

improvement by creating an informative object feature set. Similarly, Model (e) achieved a 2.0% higher performance than Model (c). These results show that even when object features are enhanced from a wide range of local features, our global attention model with well-managed memory further improves performance by collecting diverse features from overall video frames.

3) THE NUMBER OF ATTENTION STAGES

We conducted comparison experiments on global attention stages N_{stages_g}. We experimented with our models by disabling local attention stages for a fair comparison. Each attention stage consists of our feature aggregation module with same number of heads ($M = 16$). Table 6 shows that performance grows until $N_{stages\_g} = 2$ and degrades after. Because $N_{stages\_g} = 2$ performs best with reasonable time consumption, we selected it as default for DAFA_G model.

4) THE NUMBER OF REFERENCE FRAMES DURING TRAINING

In this ablation study, we aimed to determine the effect of the number of global reference frames (N_{ref}) during the training phase. Intuitively, it is expected that an increase in N_{ref} will improve the robustness of the global attention module by using more diverse reference features in the training phase. Note that N_{ref} does not affect the inference speed. FIFO in Table 7 denotes a model that uses frame-level FIFO-type global memory, and Manage denotes a model that employs object-level management, as in DAMM. For a fair comparison, the number of global attention stages was fixed at 1, and the local attention module was eliminated. For both FIFO and Manage, mAP performance steadily increased with the increase in N_{ref}. Therefore, the aggregation of a richer set of features in the training phase positively affects performance. We note that Manage performed better than FIFO by up to 1.1%. This result strongly implies that DAMM increases the range of actually visible features in the test phase regardless of N_{ref} by collecting diverse features. For an appropriate trade-off between performance and training time, we set N_{ref} to 4, as in the other experiments in this study.

TABLE 8. Effect of varying global memory size.

N_{mem_g}	300	600	900*	1500	3000
mAP(%)	83.3	83.4	83.5	83.5	83.4
runtime(ms)	53.8	54.2	54.9	56.5	58.3

5) MEMORY SIZE IN TESTING

We conducted an experiment by varying the global memory capacity in the test phase (N_{mem_g}). In this experiment, we used the DAFA_G model to exclude the effect of local attention. Because the total amount of information within a video is limited, no further improvement is expected when the memory capacity is sufficient to store all the information from the video. Table 8 presents the results of this experiment. As expected, performance increased with memory size until N_{mem_g} reached 1500, after which point performance slightly decreased. This decrease in performance may be caused by an increase in the number of false positive samples, such as backgrounds, in the key set. From this experiment, we found that a global memory size of 750 achieves the optimal trade-off between performance and runtime.

G. EVALUATION ON YOUTUBE OBJECTS DATASET

We further evaluated DAFA on the YouTube-Objects (YTO) dataset [14] to test our model's generalizability. The YTO dataset contains 150 videos with a total of 720,000 frames and 10 categories, which correspond to a subset of ImageNet VID tasks. Each video contains several shots of consecutive frames. Because only a few frames were labeled for each shot, 6,087 frames were annotated with 6,975 bounding boxes. The annotated frames were allocated into training and test sets, with the latter consisting of 1781 annotated frames. To evaluate DAFA's performance on the test set, we reused the model trained on the ImageNet DET and VID sets. No additional fine-tuning was applied. Localization accuracy was measured using CorLoc [36], an object localization metric calculated by dividing the number of correctly localized images by the number of ground truth images. Results are presented in Table 9. Data from existing methods [33], [34], [35] were retrieved from corresponding studies, and we reproduced Faster-RCNN and MEGA [8] with the ResNet-101 backbone for fair comparison. [33], [34], [35] boost performance with strong post-processing, even though it uses a weak feature extractor (HoG or GoogleNet). Note that T-CNN shows comparable performance with Faster-RCNN and MEGA. Our method outperformed all existing methods by large margins without additional post-processing. DAFA_G outperforms MEGA by 0.6% and DAFA_F outperforms MEGA by 1.2%, which are similar to the main results of Table 2.

H. QUALITATIVE ANALYSIS

1) FAILURE CASE ANALYSIS

We show some failure cases of DAFA in Fig 4. The first row is an example of missing objects. This occurs when RPN

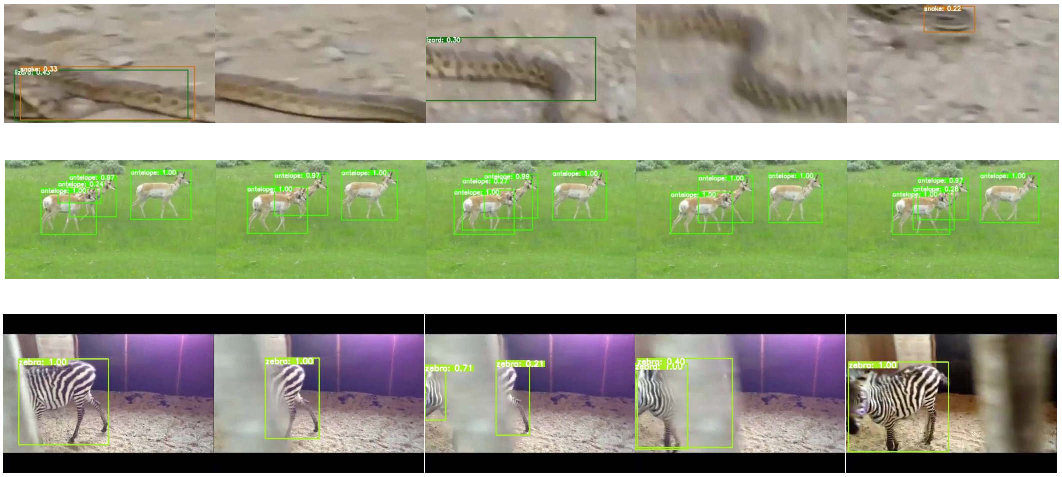

FIGURE 4. Failure cases of DAFA on ImageNet VID validation dataset. Each row lists five frames in a video that contains false negative (snake, motion blur) or false positive (antelope and zebra, occlusion) results.

TABLE 9. Localization performance evaluation on YouTube-Objects Dataset. CorLoc is used for evaluation metric.

Method	airplane	bird	watercraft	car	cat	cattle	dog	horse	motorcycle	train	Avg.
Kwak [33]	56.5	66.4	58.0	76.8	39.9	69.3	50.4	56.3	53.0	31.0	55.7
TCN [34]	94.1	69.7	88.2	79.3	76.6	18.6	89.6	89.0	87.3	75.3	76.8
T-CNN [35]	91.8	98.7	85.4	95.0	92.2	100	95.7	93.4	93.9	84.2	93.0
FasterRCNN	97.8	100	94.9	96.9	76.4	87.3	75.1	78.8	82.6	85.4	87.5
MEGA [8]	98.9	100	94.4	98.0	89.1	100	91.3	88.3	83.6	87.3	93.1
DAFA_G	98.3	100	97.0	98.5	86.7	99.2	96.0	91.1	83.1	86.7	93.7
DAFA_F	99.4	100	96.1	98.8	89.1	100	93.1	97.2	80.8	88.6	94.3

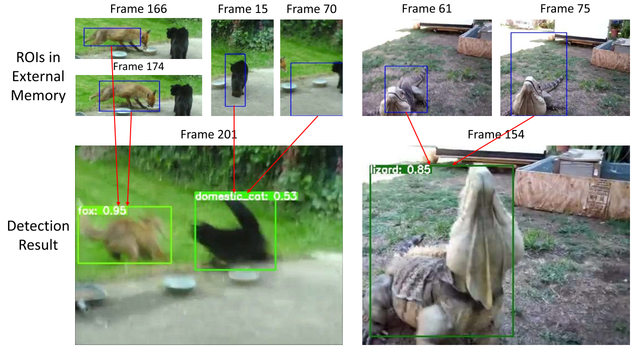

FIGURE 5. Visualization of feature aggregation and external memory.

misses objects in significantly deteriorated images caused by motion blur. DAFA can not aggregate information when no candidate object exists in the current image, as in the second column image. Also, If a video is highly deteriorated, DAFA can collect and aggregate inaccurate information, leading to misclassifying objects (misclassifying a snake as a lizard in the first and third column in the first row). Images in the second and third rows show false positive cases in object occlusion situations. Because DAFA's feature aggregation module cannot distinguish redundant ROIs (like multiple ROIs for a single object in two videos), they are not suppressed and can be generated as false-positive results. DAFA's failure cases are usually caused by low performance of RPN. Thus, combining the pixel-level attention before RPN, such as [9], [15], can be a possible solution to this problem.

2) VISUALIZATION OF FEATURE AGGREGATION AND EXTERNAL MEMORY

Fig. 5 shows how external memory and feature aggregation work for two video examples. The blue boxes in the figure in the upper row show the ROIs of memory features in the external memory, which has the top-2 highest attention scores for each current feature. Feature aggregation enhances the current features using memory features. The box in the lower row of Fig. 5 shows the detection results of enhanced current features. We observed that visually similar features, although very far from the current features, tend to have high attention scores, which shows that the proposed DAMM and feature aggregation module work as intended.

V. CONCLUSION

In this paper, we present an effective attention-based video object detection framework, DAFA. DAFA accounts for diversity when collecting global information to perform video object detection tasks. One of the key contributions of DAFA is a novel memory management scheme called DAMM. DAMM efficiently collects diverse features and alleviates the imbalance of sampled features to construct an efficient and robust key set. Experimental results show that DAFA_G and DAFA_F achieve state-of-the-art performance on the challenging ImageNet VID and YouTube Objects dataset in terms of speed and accuracy.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[3] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2015, pp. 91–99.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[6] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[7] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[8] Y. Chen, Y. Cao, H. Hu, and L. Wang, "Memory enhanced global-local aggregation for video object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10337–10346.

[9] H. Deng, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, and H. Guan, "Object guided external memory network for video object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6678–6687.

[10] H. Wu, Y. Chen, N. Wang, and Z.-X. Zhang, "Sequence level semantics aggregation for video object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9217–9225.

[11] J. Deng, Y. Pan, T. Yao, W. Zhou, H. Li, and T. Mei, "Relation distillation networks for video object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7023–7032.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[13] O. Russakovsky, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[14] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3282–3289.

[15] G. Sun, Y. Hua, G. Hu, and N. Robertson, "Mamba: Multi-level aggregation via memory bank for video object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2620–2627.

[16] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2349–2358.

[17] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 408–417.

[18] X. Zhu, J. Dai, L. Yuan, and Y. Wei, "Towards high performance video object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7210–7218.

[19] S. Wang, Y. Zhou, J. Yan, and Z. Deng, "Fully motion-aware network for video object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 542–557.

[20] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3038–3046.

[21] G. Bertasius, L. Torresani, and J. Shi, "Object detection in video with spatiotemporal sampling networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 331–346.

[22] M. Liu, M. Zhu, M. White, Y. Li, and D. Kalenichenko, "Looking fast and slow: Memory-guided mobile video object detection," 2019, *arXiv:1903.10172*.

[23] M. Zhu and M. Liu, "Mobile video object detection with temporally-aware feature maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5686–5695.

[24] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3588–3597.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[26] H. K. Cheng, Y.-W. Tai, and C.-K. Tang, "Rethinking space-time networks with improved memory coverage for efficient video object segmentation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 11781–11794.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[28] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.

[29] A. Graves, G. Wayne, and I. Danihelka, "Neural Turing machines," 2014, *arXiv:1410.5401*.

[30] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," 2017, *arXiv:1706.02413*.

[31] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10529–10538.

[32] W. Han, P. Khorrami, T. L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. S. Huang, "Seq-NMS for video object detection," 2016, *arXiv:1602.08465*.

[33] S. Kwak, M. Cho, I. Laptev, J. Ponce, and C. Schmid, "Unsupervised object discovery and tracking in video collections," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3173–3181.

[34] K. Kang, W. Ouyang, H. Li, and X. Wang, "Object detection from video tubelets with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 817–825.

[35] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, and W. Ouyang, "T-CNN: Tubelets with convolutional neural networks for object detection from videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2896–2907, Oct. 2018.

[36] T. Deselaers, B. Alexe, and V. Ferrari, "Localizing objects while learning their appearance," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 452–466.

SI-DONG ROH (Graduate Student Member, IEEE) received the B.S. degree in electronic engineering from Hanyang University, Seoul, South Korea, in 2015, where he is currently pursuing the Ph.D. degree in electronic engineering. His research interests include video-object detection and image-based defect inspection.

KI-SEOK CHUNG (Member, IEEE) received the B.S. degree in computer engineering from Seoul National University, Seoul, South Korea, in 1989, and the Ph.D. degree in computer science from the University of Illinois at Urbana–Champaign, in 1998. He was a Senior Research and Development Engineer at Synopsys Inc., Mountain View, CA, USA, from 1998 to 2000, and a Staff Engineer at Intel Corporation, Santa Clara, CA, USA, from 2000 to 2001. He also worked as an Assistant Professor at Hongik University, Seoul, from 2001 to 2004. Since 2004, he has been a Professor at Hanyang University, Seoul. His research interests include low-power embedded system design, multi-core architecture, image processing, reconfigurable processor and DSP design, SoC-platform-based verification, and system software for MPSoCs.