Dual Semantic Fusion Network for Video Object Detection

Lijian Lin*†
Fujian Key Laboratory of Sensing and
Computing for Smart City, School of
Informatics, Xiamen University, Xiamen, China.
ljlin@stu.xmu.edu.cn

Haosheng Chen†
Fujian Key Laboratory of Sensing and
Computing for Smart City, School of
Informatics, Xiamen University, Xiamen, China.
haoshengchen@stu.xmu.edu.cn

Honglun Zhang
Applied Research Center (ARC), Tencent PCG
honlanzhang@tencent.com

Jun Liang
Fujian Key Laboratory of Sensing and
Computing for Smart City, School of
Informatics, Xiamen University, Xiamen, China.
Junliang@stu.xmu.edu.cn

Yu Li, Ying Shan
Applied Research Center (ARC), Tencent PCG
{ianyli,yingsshan}@tencent.com

Hanzi Wang‡
Fujian Key Laboratory of Sensing and
Computing for Smart City, School of
Informatics, Xiamen University, Xiamen, China.
hanzi.wang@xmu.edu.cn

ABSTRACT

Video object detection is a tough task due to the deteriorated quality of video sequences captured under complex environments. Currently, this area is dominated by a series of feature enhancement based methods, which distill beneficial semantic information from multiple frames and generate enhanced features through fusing the distilled information. However, the distillation and fusion operations are usually performed at either frame level or instance level with external guidance using additional information, such as optical flow and feature memory. In this work, we propose a dual semantic fusion network (abbreviated as DSFNet) to fully exploit both frame-level and instance-level semantics in a unified fusion framework without external guidance. Moreover, we introduce a geometric similarity measure into the fusion process to alleviate the influence of information distortion caused by noise. As a result, the proposed DSFNet can generate more robust features through the multi-granularity fusion and avoid being affected by the instability of external guidance. To evaluate the proposed DSFNet, we conduct extensive experiments on the ImageNet VID dataset. Notably, the proposed dual semantic fusion network achieves, to the best of our knowledge, the best performance of 84.1% mAP among the current state-of-the-art video object detectors with ResNet-101 and 85.4% mAP with ResNeXt-101 without using any post-processing steps.

CCS CONCEPTS

• **Computing methodologies → Object detection**.

---

*This work is done while interning at Applied Research Center (ARC), Tencent PCG.
†Equal contribution.
‡Corresponding author.

KEYWORDS

Video Object Detection, Semantic Fusion, Information Distillation, Geometric Similarity

1 INTRODUCTION

Video object detection aims to detect objects of interest on consecutive video frames, which is a vital task in the multimedia area. Despite the great success achieved by still image detection works, video object detection is still challenging due to the deteriorated video quality caused by motion blur, video defocus, pose variation and occlusion (see Figure 1 for some examples). However, along with these challenges, videos inherently contain much richer context information in the spatio-temporal domain compared with individual images, which gives a clear direction of exploiting the rich information in video sequences to improve the performance of video object detection.

Based on the success of single-frame detectors, cutting-edge video object detection works (such as [9, 19, 53, 59]) tend to consider more than one frame as support frames to leverage the context information of video sequences. Specifically, these works distill useful information from the support frames and then fuse the distilled information into the deteriorated frames to generate enhanced features for robust detection. The distillation and fusion operations are usually performed on either frame-level or instance-level features with external guidance, such as optical flow [58, 59] or global/local feature memory [7]. The external guidance is used to measure the similarities among pixels or instances, which are employed to guide the following fusion process. Since the external guidance is usually implemented by using additional deep neural networks, the performance of the guidance based object detection methods cannot be assured by themselves and they may suffer from the instability of the external guidance. In particular, false positive estimations introduced by the external guidance are fatal for the guidance based methods and they are more likely to cause the failure of detection.

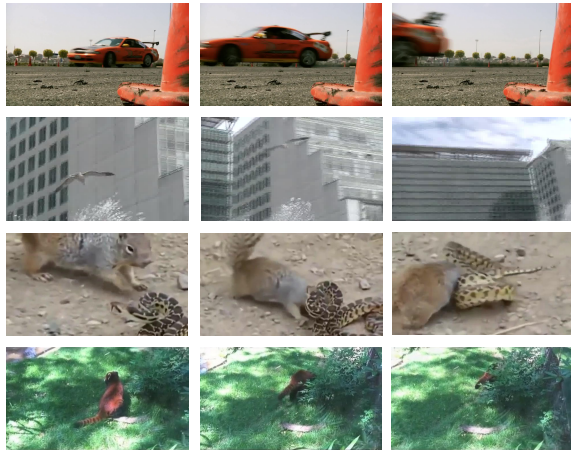arXiv:2009.07498v1 [cs.CV] 16 Sep 2020

Figure 1: Some examples of deteriorated video sequences. From the top row to the bottom row, the corresponding challenges are motion blur, video defocus, pose variation and occlusion, respectively.

Considering that there are frame-level and instance-level semantic information that can be extracted by the dominated two-stage detection framework before and after the region proposal network (RPN), there is a natural desire to perform the distillation and fusion operations at both levels. Consequently, in this paper, we present a Dual Semantic Fusion Network (called DSFNet) to exploit both frame-level and instance-level semantic information in video sequences. Moreover, we propose a geometric similarity measure to measure the geometric similarities among object instances. Then, the geometric similarities are used to cooperate with its corresponding appearance information to mitigate the information distortion problem caused by noise during the dual semantic fusion procedure. Compared with the current one-stage feature enhancement methods, the proposed DSFNet can generate more robust features through the multi-granularity fusion.

Different from the existing guidance based methods, we argue that the proposed dual semantic fusion network can be learned through a unified framework in a fully end-to-end manner, even if there is no external guidance for the fusion. The reason is that the frame-level fusion in DSFNet can provide rich but object-agnostic information, which consists of relatively low-level semantics. On the contrary, the instance-level fusion in DSFNet can distill object-specific but limited information, which includes relatively high-level semantics, as the complementary cue. Through the combination of the frame-level and instance-level semantic fusions, the distilled information from one level becomes the internal guidance of the other level in DSFNet. In addition, since we do not use any external guidance in our network, the proposed DSFNet is self-contained and it does not need to rely on the precision and reliability of the corresponding external guidance.

In summary, we make the following contributions in this paper:

- We present a dual semantic fusion network, which performs a multi-granularity semantic fusion at both frame level and instance level in a unified framework and then generates enhanced features for video object detection.
- We introduce a geometric similarity measure into the proposed dual semantic fusion network along with the widely used appearance similarity measure to alleviate the information distortion caused by noise during the fusion process.
- We explain the video object detection process from a novel information theory perspective and then give a detailed analysis to show the effectiveness of the proposed dual semantic fusion network.

We evaluate the proposed DSFNet on the large scale ImageNet VID dataset [41]. The experimental results demonstrate the superiority of our DSFNet over several state-of-the-art methods. Especially, DSFNet outperforms its baseline detector by a large margin of 9.4% mAP and achieves, to the best of our knowledge, the highest mAP of 84.1%/85.4% with ResNet-101/ResNeXt-101 when it is compared with the published state-of-the-art video object detection methods without using additional post-processing steps.

2 RELATED WORK

In this section, we briefly review several representative still image object detectors and video object detectors.

Still Image Object Detectors. Object detection in still images is one of the fundamental tasks in the multimedia and computer vision communities with a variety of applications (*e.g.*, [4, 5, 17, 18, 29, 36, 37, 44, 57]). State-of-the-art still image object detectors can be roughly classified into two types: two-stage detectors (such as [3, 10, 13, 16, 27, 28, 38, 46, 47]), and one-stage detectors (*e.g.*, [2, 15, 23, 25, 30, 39, 56]).

As one of the most representative two-stage detectors, Faster R-CNN [40] proposes to generate region proposals by using CNNs, and then classifies and refines the generated proposals for object detection. FPN [30] improves Faster R-CNN by designing a feature pyramid network for detecting objects at different scales. [22] proposes to model the similarities among instances to capture the contextual information in a whole image, which yields promising performance on the object detection task.

In contrast, the one-stage detectors directly make predictions with a single detection network. For example, SSD [34] and RetinaNet [31] place some pre-designed anchor boxes densely over feature maps, and directly classify and refine each anchor box. CenterNet [12] and FCOS [48] propose to detect objects in images without designing a set of anchor boxes, and they achieve better performance than most of the anchor based one-stage detectors.

Different from detecting objects in still images, a video sequence contains much richer spatio-temporal information for detection. The rich information can be leveraged to solve the challenging situations (such as occlusion, motion blur, rare poses) when detecting objects in videos. Therefore, in this paper, we propose to exploit the spatio-temporal information in video frames to improve the video object detection performance. Similar to most of the state-of-the-art video object detectors, our proposed DSFNet is also built upon the effective Faster R-CNN framework.

Video Object Detectors. For the task of video object detection, one of the main challenges is how to improve single-frame detection performance by exploring temporal information of videos. One
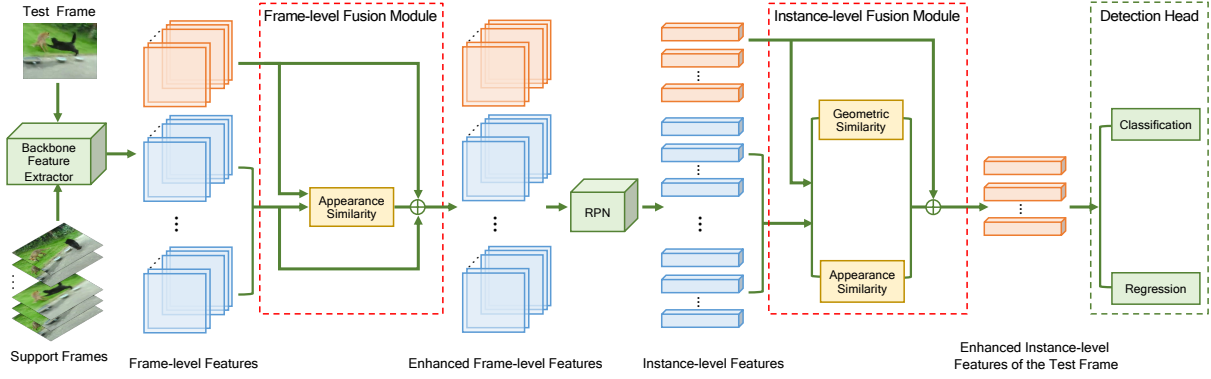
Figure 2: The pipeline of the proposed DSFNet. Given a test frame and several support frames, we first extract their frame-level features. The features marked in orange/blue are the features of the test/support frames. Then, these features are enhanced by fusing them according to their appearance similarities (Eq. (2)). We apply RPN on the enhanced frame-level features to get instance-level features. After that, the instance-level features are enhanced based on their appearance and geometric similarities (Eq. (5)). Finally, the enhanced instance-level features of the test frame are fed into the detection head for final detection.

common solution is to apply post-processing techniques to the predicted bounding boxes obtained by still image detectors [20, 24, 26, 45]. For example, SeqNMS [20] conducts a sequence-level NMS on the detected bounding boxes, and it uses high-scoring detection results to boost the scores of weaker detection results. T-CNN [24] adopts optical flow to propagate the predicted detection results to neighboring frames and re-scores the detection results by incorporating an additional object tracker. These post-processing methods can improve the performance of still image object detectors when they are applied to videos. However, the performance of these methods highly relies on their associated image object detectors, and it is difficult for them to correct the errors produced by the associated image object detectors. In contrast, our DSFNet focuses on using the temporal information at feature level rather than at the final bounding box level, and it can be trained in an end-to-end manner without using any post-processing steps.

Recently, state-of-the-art video object detectors tend to fuse features extracted from multiple frames in a video sequence to boost the detection performance. There are mainly two ways to fuse features: frame-level fusion [19, 32, 33, 35, 54, 58] and instance-level fusion [9, 42, 53]. For the frame-level fusion based methods, they propose to fuse the features extracted from multiple frames at frame level. For example, FGFA [59] proposes to warp the features from adjacent frames and fuse them to the reference frame according to the optical flow generated by [11]. THP [58] also uses optical flow to propagate the extracted features from keyframes to non-keyframes. STSN [1] adopts deformable convolution to align and aggregate features between frames without using optical flow information.

For the instance-level fusion based methods, they fuse the features extracted from object instances for detection. For example, SELSA [53] introduces a semantics aggregation module to fuse the features extracted from object instances and produce enhanced features for detection. RDN [9] designs relation distillation networks to measure the relation among object instances and then aggregates them to augment the features of instances for detection. The work in [42] modifies the non-local block in [51] to learn the appearance

similarities between the target proposals and the proposals generated from multiple support frames to enrich the target proposal features, which boosts its detection performance. Different from the above-mentioned methods, we propose a dual semantic fusion network, which fuses features on both frame level and instance level. By doing this, the features from both levels are fused to generate enhanced features by utilizing the spatial-temporal information in videos, leading to a better video object detection performance.

3 METHOD

In this section, we first provide an overview of the proposed DSFNet. Then, we introduce the proposed dual semantic fusion network, which generates enhanced features to perform robust video object detection. Finally, we analyze the proposed DSFNet from an information theory viewpoint.

3.1 Overview

The overall video object detection pipeline of the proposed DSFNet is illustrated in Figure 2. Given a test frame of a video sequence, we first sample a set of support frames from the video sequence and extract the frame-level features of these frames by using the backbone feature extractor. Then, we apply the proposed frame-level semantic fusion module on these features to obtain the corresponding enhanced frame-level features. These enhanced features are fed into the Region Proposal Network (RPN) to generate a set of object instances. We further enhance the features of these instances by using the proposed instance-level semantic fusion module. Finally, we feed the enhanced instance-level features into the detection head for object classification and bounding box regression.

3.2 Dual Semantic Fusion

Considering that there are usually deteriorated video frames occurring in the task of video object detection, the main challenge of accurately detecting objects in videos lies in how to leverage the rich information in videos. In this subsection, we describe the proposed dual semantic fusion network, which consists of a frame-level

fusion module and an instance-level fusion module. Both fusion modules can enhance the features extracted from individual frames by fusing the rich information in videos.

Frame-level Semantic Fusion. Given a test frame in a video sequence, we first sample $n - 1$ support frames from the rest of the video sequence. With the n frames, we extract a series of frame-level features $\mathcal{F} = \{F_1, F_2, ..., F_n\}$, where $F_i \in \mathcal{F}$ indicates the frame-level feature extracted from the i-th input frame. Since each of the frame-level features in $\mathcal{F}$ has d channels, we split the frame-level features in $\mathcal{F}$ into $n * d$ separated channel-wise features $\mathcal{F}^c = \{F_1^c, F_2^c, ..., F_{n*d}^c\}$. During the frame-level semantic fusion, inspired by the non-local network in [51], we calculate a similarity matrix S^F of $\mathcal{F}^c$ to represent the appearance similarities among the features in $\mathcal{F}^c$. Then, for the i-th feature F_i^c in $\mathcal{F}^c$, we fuse all the features in $\mathcal{F}^c$ into F_i^c based on S^F to generate the corresponding i-th enhanced feature F_i^e. Here, we denote the generated enhanced features as $\mathcal{F}^e = \{F_1^e, F_2^e, ..., F_{n*d}^e\}$. Specifically, the i-th enhanced feature $F_i^e \in \mathcal{F}^e$ is calculated by the following equation:

$$F_i^e = F_i^c + \sum_{j=1}^{n*d} S_{i,j}^F \cdot \theta(F_j^c), i = 1, 2, 3, ..., n * d \qquad (1)$$

where $\theta(\cdot)$ denotes a general transformation function parameterized by fully connected layers. $S_{i,j}^F \in S^F$ means the appearance similarity between F_i^c and F_j^c, which is calculated as follows:

$$S_{i,j}^F = \frac{exp(a_{i,j})}{\sum_{u=1}^{n*d} exp(a_{i,u})} \qquad (2)$$

where $a_{i,j}$ is the cross product between F_i^c and F_j^c, and it is formulated as follows:

$$a_{i,j} = < \phi(F_i^c), \varphi(F_j^c) > \qquad (3)$$

$\phi(\cdot)$ and $\varphi(\cdot)$ are two general transformation functions, which are similar to $\theta(\cdot)$. After the fusion, the information contained in the i-th feature $F_i^c \in \mathcal{F}^c$ is propagated to the other features in $\mathcal{F}^c$. As a result, each of the enhanced features $\mathcal{F}^e$ can distill rich information from the frame-level features of the other frames.

Instance-level Semantic Fusion. For the instance-level semantic fusion, the enhanced features $\mathcal{F}^e$ generated by the frame-level semantic fusion module are fed into RPN to generate a set of object instances with the associated bounding boxes $\mathcal{B} = \{B_1, B_2, ..., B_m\}$. Here m is the number of the generated instances. Each bounding box in $\mathcal{B}$ contains the spatial location and the scale information of an instance. Then, a RoI layer is applied on the bounding boxes in $\mathcal{B}$ and the enhanced frame-level features in $\mathcal{F}^e$ to generate the corresponding RoI features $Q = \{Q_1, Q_2, ..., Q_m\}$ of the instances. After the instance-level semantic fusion, the final enhanced instance-level features $Q^e = \{Q_1^e, Q_2^e, ..., Q_m^e\}$ are generated by fusing all the RoI features in Q, which is written as follows:

$$Q_k^e = Q_k + \sum_{l=1}^{m} S_{k,l}^I \cdot \gamma(Q_l), k = 1, 2, 3, ..., m \qquad (4)$$

where $\gamma(\cdot)$ is a general transformation function and $S_{k,l}^I$ indicates the instance-level similarity between Q_k and Q_l.

Since geometric information plays an important role in representing an object as well as the appearance information, for the instance-level fusion, we propose to measure the similarities among

the instances not only based on the appearance information contained in Q, but also based on the geometric information contained in $\mathcal{B}$, which is

$$S_{k,l}^I = \frac{exp(z_{k,l} + r_{k,l})}{\sum_{v=1}^{m} exp(z_{k,v} + r_{k,v})} \qquad (5)$$

where $z_{k,l}$ is the appearance similarity between Q_k and Q_l. $r_{k,l}$ indicates the geometric similarity between the k-th and l-th bounding boxes B_k and B_l in $\mathcal{B}$. Specifically, $z_{k,l}$ is formulated as:

$$z_{k,l} = < \xi(Q_k), \zeta(Q_l) > \qquad (6)$$

where $\xi(\cdot)$ and $\zeta(\cdot)$ are two general transformation functions parameterized by fully connected layers. Since different objects may have similar spatial locations in different frames, the scale information (i.e., the width w and the height h) contained in $\mathcal{B}$ is more reliable in measuring the geometric similarity than the spatial information. Therefore, we propose to measure the geometric similarity $r_{k,l}$ between B_k and B_l, as follows:

$$r_{k,l} = \psi(\varrho(log(\frac{w_k}{w_l}), log(\frac{h_k}{h_l}), log(|\frac{w_k}{h_k} - \frac{w_l}{h_l}|))) \qquad (7)$$

$\psi(\cdot)$ indicates a general transformation function, which plays a similar role as $\xi(\cdot)$ and $\zeta(\cdot)$. $\varrho(\cdot)$ is the embedding function used in [22], which embeds the primitive low-dimensional geometric similarity $r_{k,l}$ into a high-dimensional representation for the proposed deep detection network. By exploiting the geometric information and the appearance information, our DSFNet can alleviate the information distortion problem caused by noise during the fusion process. Finally, the enhanced features in Q^e that correspond to the test frame are fed to the detection head for the final object detection.

3.3 An Information Theory Viewpoint

As described in the pioneering study [43], the learning process of a deep neural network based object detector can be mathematically analyzed from the perspective of information bottleneck (IB) theory [49], as shown in the state-of-the-art still image object detection work [52]. Similarly, from the IB perspective, learning a deep video object detection network can be considered as a Markov process with a concise Markov chain:

$$V \rightarrow F \rightarrow O \qquad (8)$$

where V is the input variable (i.e., the input video sequence tensor), F means the intermediate variable, which is related to the frame-level features extracted by the backbone network, and O stands for the output variable, which consists of the final predicted object labels and locations. Then, the goal of learning the whole detection network is to minimize the mutual information between the input video tensor V and the frame-level features F, and to maximize the mutual information between F and O, which is formulated as:

$$\min_{\omega_b, \omega_d} \{I(V; F) - \beta I(F; O)\} \qquad (9)$$

where ω_b and ω_d are the learnable parameters of the backbone and the detection head, respectively. β is a Lagrange multiplier. $I(X, Y)$ is the mutual information between X and Y, which is defined as:

$$I(X; Y) = \mathbb{E}_{x \sim p(x)} \mathbb{E}_{y \sim p(y|x)} \log \frac{p(y|x)}{p(y)} \qquad (10)$$

where x and y are respectively a specific instance in X and the corresponding instance in Y. $p(\cdot)$ is the prior distribution, and $\mathbb{E}$ refers to the expectation function. In Eq. (9), $I(V;F)$ is minimized to distill the useful information for the video object detection task from the input variable V, while $I(F;O)$ is maximized to preserve more distilled information for the final detection.

According to the data processing inequality concept in information theory, there is no post-processing that can increase the information contained in the input variable V. Specifically, the information contained in the input variable V can not be increased through the given Markov chain, which can be formulated as:

$$I(V;F) \geqslant I(V;O) \qquad (11)$$

The equality of Eq. (11) can be achieved if and only if F and O contain the same information about V, which is impractical due to the high compression in Eq. (9). As a consequence, the information contained in V is gradually decreased during the learning process.

Since most of the state-of-the-art video object detectors and the proposed DSFNet have two stages (*i.e.*, RPN based object proposal generation and RCNN based final prediction) between F and O, the Markov chain in Eq. (8) can be rewritten as:

$$V \rightarrow F \rightarrow P \rightarrow O \qquad (12)$$

where P represents the instance-level features generated by the RoI layer. As a result, following the afore-mentioned data processing inequality rule, we can rewrite Eq. (11) as:

$$I(V;F) \geqslant I(V;P) \geqslant I(V;O) \qquad (13)$$

According to the IB principle in Eq. (9), $I(F;O)$ should be maximized to preserve more distilled information. Different from still image object detection, for video object detection, the information contained in the support frames of a video sequence can boost the performance of detection on the test frame of the video sequence. Since the information contained in F is gradually decreased in Eq. (13), and the final prediction is solely based on the information contained in the test frame, the information contained in the support frames should be distilled and fused into the enhanced features of the test frame before and after the RoI layer.

In the proposed DSFNet, the first frame-level fusion is employed before the RoI layer to distill frame-level information from support frames. The distilled frame-level information is then fused into the frame-level features of the test frame. Then, the second fusion is performed at the instance level after the RoI layer. The instance-level fusion, which is a high-level semantic fusion, is based on the object-level similarity. By doing the dual semantic fusion, the information distilled at both the frame level and the instance level is fused into the RoI features of the test video frame for the final object detection. As a result, compared with the current single fusion based methods, the proposed DSFNet can preserve more beneficial information contained in F to perform more accurate video object detection.

4 EXPERIMENTS

In this section, we first introduce the dataset and evaluation protocols for the video object detection task. Then, we present the implementation details of the proposed DSFNet. We also carry out several ablation studies on the ImageNet VID validation set [41]

to verify the effectiveness of the proposed dual semantic fusion network. Finally, we compare DSFNet with several other state-of-the-art video object detection methods.

4.1 Dataset and Evaluation Protocols

We conduct the experiments on the ImageNet VID dataset [41], which is a large scale benchmark for video object detection. The ImageNet VID dataset contains 4,417 video snippets for training and validation. There are 3,862 video snippets in the training set. And the validation set consists of 555 video snippets. The frames in these snippets are fully annotated over 30 object categories with bounding boxes. Following the widely adopted protocols in video object detection [1, 9, 24, 53, 54, 59], we evaluate the proposed DSFNet on the ImageNet VID validation set and use the mAP@IoU=0.5 scores as the evaluation metric. Moreover, as in [42, 53, 59], all the objects in the ImageNet VID validation set are categorized into three groups (*i.e.*, the slow, medium and fast motion groups), according to their motion speed. We evaluate DSFNet on the objects with different motion groups for better analysis.

Although there are more than a million frames in the ImageNet VID training set, some of these frames are redundant. Thus, the appearance diversity of the objects in the ImageNet VID training set is limited, which makes the training process less effective. Therefore, as in the previous works [1, 9, 24, 53, 54, 59], we train the proposed DSFNet on the intersection of the ImageNet VID and DET datasets [41]. The ImageNet DET dataset is a still image detection dataset.

4.2 Implementation Details

Next, we will discuss the implementation details of the proposed DSFNet from four aspects, including feature extractor, detection network, dual semantic fusion, and training/inference details.

Feature Extractor. We use ResNet-101 [21] or ResNeXt-101-32×4d [55] as our backbone feature extractor. Following the work in [9, 50, 53, 59], we modify the convolutional stride of the last block of the last stage (i.e, *conv*5 for both ResNet-101 and ResNeXt-101) from 2 to 1. As a result, the total stride of *conv*5 is changed from 32 to 16, which increases the resolution of the extracted feature maps. In addition, we set the dilation rate to 2 in those convolutional layers in *conv*5, where their kernel size is larger than 1, to retain the receptive field of the backbone feature extractor.

Detection Network. We adopt Faster R-CNN [40] as our baseline detection network. RPN is applied to the output of *conv*4. To reduce redundancy, a non-maximum suppression (NMS) with a IoU threshold of 0.7 is adopted and 300 candidate boxes are generated in each frame during both training and inference phases. A RoI pooling layer is applied to the output of *conv*5 with the generated candidate boxes to extract a series of RoI-pooled features for these boxes. Then, these RoI-pooled features are fed into the detection head for object classification and bounding box regression.

Dual Semantic Fusion. We apply the frame-level fusion module to the output of *conv*4 to generate enhanced frame-level features. These enhanced features are then fed into RPN to get a series of object instances. For the instance-level fusion module, we insert it after the RoI pooling layer twice to generate enhanced instance-level features for final detection.
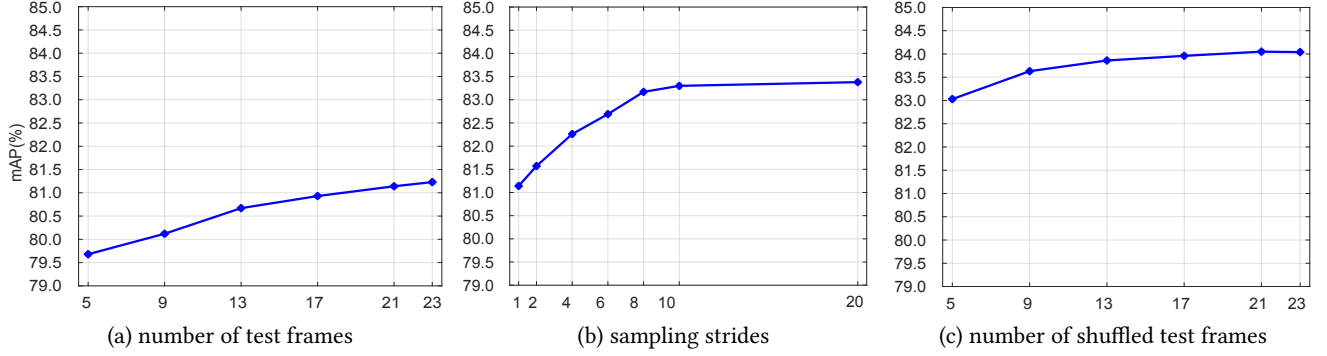
Figure 3: The test performance on the ImageNet validation set obtained by the proposed DSFNet with (a) different numbers of test frames, (b) different sampling strides, and (c) different numbers of shuffled test frames.

Training and Inference Details. For both training and inference, the input images are resized to have a shorter side of 600 pixels. During training, the backbone feature extraction network (*i.e.*, ResNet-101 or ResNeXt-101) is initialized with the weights that were pre-trained on the ImageNet classification dataset [8]. The whole network is trained on 8 GPUs using SGD with cross-entropy loss. The total batch size is 8 with each GPU holding one sample. During training, a sample contains 3 frames: One is the current frame for training and the other two are the support frames that provide temporal information. For the ImageNet VID dataset, the two support frames are randomly sampled in the current video sequence. And for the ImageNet DET dataset, the three frames from the dataset (*i.e.*, the still image dataset) are identical. We train the proposed network for a total of 247k iterations. The initial learning rate is set to 2.5×10^{-4} and it is respectively dropped by a factor of 10 at the 109k and 219k iterations. In addition, we adopt the same data augmentation strategy as in [53]. During inference, we sample n frames in a video for the proposed DSFNet. The influence of the parameter n will be discussed in the next subsection.

4.3 Frame Sampling Strategies

Frame sampling is an essential part of feature enhancement based video object detection methods. This has been reported by the previous works (*e.g.*, [53, 59]). Therefore, it is worth investigating the effectiveness of DSFNet under different frame sampling strategies.

Here, we evaluate the performance of the proposed DSFNet using a fixed interval sampling strategy with different numbers of test frames and various sampling strides. Moreover, we also use a stochastic sampling strategy to evaluate DSFNet. Let n be the number of the test frames. The n test frames consist of the evaluated frame and the $n - 1$ sampled support frames.

Firstly, we evaluate the proposed DSFNet with different numbers of test frames using a fixed interval sampling strategy. During the evaluation, the $n - 1$ adjacent frames are sampled as the support frames with a fixed sampling stride of 1. By increasing the number of the test frames from 5 to 23, the performance of DSFNet is improved from 79.7% to 81.2% mAP (+1.5%), as shown in Figure 3(a). From the figure, it is clear that the performance of DSFNet can be improved with the increasing number of the test frames. However, more test

frames require more computing resources. Therefore, more choices of the sampling stride should be considered before the value of n (*i.e.*, the number of test frames) is determined.

Then, we examine the influence of different sampling strides on the performance of DSFNet. Let s be the sampling stride. The $n - 1$ support frames are uniformly sampled at every s frames. We fix the n value to 21 but use various sampling strides on DSFNet. The experimental results are reported in Figure 3(b). As we can see, the performance of DSFNet can be improved with the increasing of the sampling stride. In particular, DSFNet achieves the highest mAP of 83.4% with the largest stride of 20, which is large enough to traverse most of the test video sequences in the ImageNet VID set. Actually, the current fixed interval sampling strategy can be considered as a special case of the stochastic sampling strategy, which randomly samples the support frames from the whole test video sequence.

Finally, we replace the fixed interval sampling strategy with the stochastic sampling strategy to further improve the performance of our DSFNet. The test frames are shuffled at the beginning. After that, we adjust the number of the shuffled test frames from 5 to 23 to evaluate the performance of DSFNet. The obtained results are given in Figure 3(c), from which we can see that by leveraging the rich context information in the temporal domain, DSFNet achieves the highest mAP of 84.1% when the number of the shuffled test frames is set to 21. Moreover, the performance of DSFNet is saturated when the number of the shuffled test frames is more than 21, as illustrated in Figure 3(c). Consequently, we choose the stochastic sampling strategy to sample the test frames and fix the value of n to 21 in the proposed DSFNet for all the following experiments, which is a good trade-off between effectiveness and efficiency.

4.4 Ablation Study

We perform several ablation studies on the ImageNet VID validation set to evaluate the effectiveness of the proposed DSFNet. Table 1 reports the quantitative results obtained by four variants of DSFNet, which are respectively: (a) the baseline, (b) the baseline with the proposed frame-level fusion module, (c) the baseline with the proposed instance-level fusion module, and (d) the proposed DSFNet. All the results in Table 1 are based on the ResNet-101 backbone. In
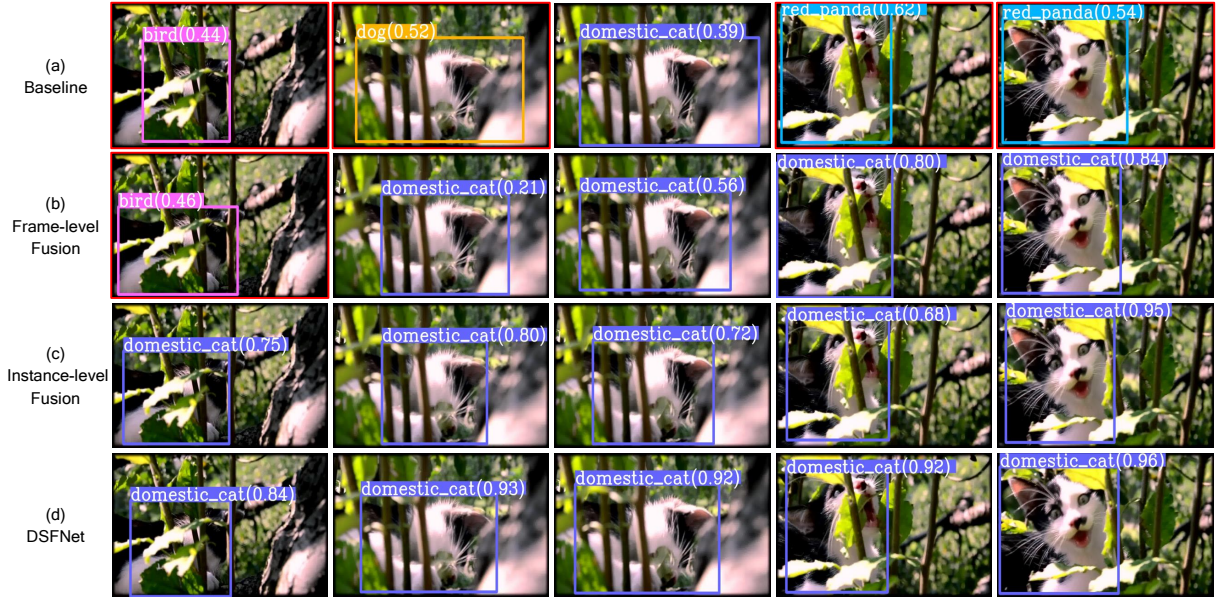
Figure 4: Qualitative results obtained by four variants of our DSFNet. The results include object labels and the corresponding confidence scores in brackets. The four variants of DSFNet are listed at the left of this figure. The detection results are marked in different colors according to their predicted labels. Those frames with false positive results are highlighted by red rectangles.

Table 1: Ablation study on the ImageNet VID validation set. The results are obtained by four variants of DSFNet. The best results are highlighted by bold.

Variants	(a) Baseline	(b) Frame-Level Fusion	(c) Instance-Level Fusion	(d) DSFNet
mAP (%)	74.7	77.0	83.3	$\textbf{84.1}_{\uparrow 9.4}$
mAP (%) (slow)	83.3	85.7	89.6	$\textbf{90.0}_{\uparrow 6.7}$
mAP (%) (medium)	72.3	74.8	81.4	$\textbf{82.6}_{\uparrow 10.3}$
mAP (%) (fast)	52.3	54.0	66.2	$\textbf{67.0}_{\uparrow 14.7}$

particular, Table 1(a) provides the results obtained by the baseline detector (*i.e.*, Faster R-CNN).

Frame-level Fusion. From the results in Table 1(b), we can see that introducing the proposed frame-level fusion module into the baseline detector leads to +2.3% gain in terms of mAP. This is because that the proposed frame-level fusion module is capable of producing enhanced features by fusing the frame-level information. As a result, the frame-level fusion module can effectively propagate the beneficial semantic information across frames, by which it boosts the performance of the baseline detector.

Instance-level Fusion. Table 1(c) shows the results obtained by applying the proposed instance-level feature fusion module to the baseline detector. Compared with the baseline detector, a significant +8.6% gain on mAP is achieved. This performance gain can be attributed to the improvement of fusing the rich semantic context information across instances and leveraging the proposed

geometric similarity measure. The rich instance-level information makes the detector robust against object appearance variations (such as motion blur, occlusion, and deformation) in videos.

The overall performance of leveraging the above two fusion modules is presented in Table 1(d). Jointly applying both of the proposed frame-level and instance-level feature fusion modules to the baseline detector leads to a considerable gain of +9.4% mAP on the test dataset. Moreover, as shown in Table 1, DSFNet can significantly improve the detection performance of the baseline detector on all the three types of motion groups in [59]. Specifically, DSFNet achieves +6.7%, +10.3%, and +14.7% mAP gains for the object detection on the slow, medium, and fast motion groups, respectively. The most significant improvement of +14.7% is achieved by DSFNet on the fast motion group. This is because that DSFNet can effectively enhance the deteriorated features of fast moving objects by fusing the features among frames and instances. Thus, the enhanced features contain beneficial semantic information from other high-quality frames and instances, which makes DSFNet more robust in dealing with the fast moving objects. Overall, the results in Table 1 show the effectiveness of combining both the frame-level and instance-level fusions in the proposed DSFNet within a unified framework for detecting objects in videos.

Qualitative Detection Results. Besides the quantitative results, we also provide some qualitative detection results obtained by these variants in Figure 4. The video sequence in Figure 4 is very challenging due to the deteriorated appearance of the cat caused by serious occlusions and significant pose variations. As shown in Figure 4(a), the baseline detector tends to classify the detected objects into incorrect categories. The frame-level fusion module utilizes the features from more than one frame and achieves much better results than the baseline detector using the features from a single

Table 2: Comparison with state-of-the-art competitors on the ImageNet VID validation set. * indicates the methods with post-processing steps. The best results are highlighted by bold.

Methods	Backbone	mAP (%)
FGFA [59]	ResNet-101	76.3
MANet [50]	ResNet-101	78.1
THP [58]	ResNet-101	78.6
STSN [1]	ResNet-101	78.9
LRTR [42]	ResNet-101	81.0
RDN [9]	ResNet-101	81.8
SELSA [53]	ResNet-101	82.7
FGFA* [59]	ResNet-101	78.4
ST-Lattice* [6]	ResNet-101	79.6
D&T* [14]	ResNet-101	79.8
MANet* [50]	ResNet-101	80.3
STSN* [1]	ResNet-101	80.4
STMN* [54]	ResNet-101	80.5
RDN* [9]	ResNet-101	83.8
DSFNet (ours)	ResNet-101	**84.1**
RDN [9]	ResNeXt-101	83.2
LRTR [42]	ResNeXt-101	84.1
SELSA [53]	ResNext-101	84.3
FGFA* [59]	Inception-ResNet	80.1
D&T* [14]	Inception-v4	82.1
RDN* [9]	ResNeXt-101	84.7
DSFNet (ours)	ResNeXt-101	**85.4**

frame. However, it fails in some hard cases (see the most left frame in Figure 4(b)). Meanwhile, the detector with only the proposed instance-level fusion module can correctly detect those objects affected by occlusions and pose variations with relatively low confidence scores and less accurate bounding boxes, as shown in Figure 4(c). Finally, by leveraging both the frame-level and instance-level fusion modules, the proposed DSFNet yields the best performance among those variants, which shows the effectiveness of DSFNet.

4.5 Comparison with State-of-the-art Methods

We compare the proposed DSFNet with several state-of-the-art video object detection methods, including MANet [50], FGFA [59], THP [58], ST-Lattice [6], D&T [14], STSN [1], STMN [54], RDN [9], SELSA [53], and LRTR [42]. Table 2 summarizes the results obtained by the proposed DSFNet and the other state-of-the-art methods on the ImageNet VID validation set. As shown in Table 2, the proposed DSFNet with ResNet-101 obtains 84.1% mAP, outperforming all the other competing video object detectors.

Among these detectors, FGFA and THP propose to improve per-frame features by fusing the features across frames with external guidance using optical flow information estimated by [11]. Thus, these two detectors may suffer from the instability of their guidance. In contrast, the proposed DSFNet aims to enhance the features for video object detection in a unified framework without using any external guidance, which yields much better performance than

FGFA (+7.8% mAP) and THP (+5.5% mAP). In addition, STSN and STMN only use the aggregated frame-level features to perform robust video object detection. Compared with them, DSFNet achieves better results by fusing the frame-level and instance-level features, outperforming these two methods by +5.2% and +3.6% mAP, respectively. Meanwhile, SELSA is a newly proposed video object detection method that utilizes the appearance similarities among instances to perform instance-level semantic fusion. Compared with SELSA, DSFNet adopts both appearance similarity and geometric similarity in the instance-level semantic fusion module to mitigate the information distortion problem. As a result, DSFNet achieves the highest mAP of 84.1%, which outperforms SELSA by +1.4% mAP. RDN also aggregates the instance-level features across frames to generate the enhanced instance-level features for robust detection, and it achieves a satisfying performance of 81.8% mAP. Moreover, RDN employs additional post-processing techniques to boost its performance from 81.8% to 83.8% mAP. Nevertheless, the performance of RDN is still inferior to that of the proposed DSFNet, which does not use any post-processing techniques.

Moreover, by changing the backbone feature extractor from ResNet-101 to a stronger backbone feature extractor ResNeXt-101, our DSFNet achieves a better performance of 85.4% mAP without using any post-processing steps. This result still outperforms the reported results from the current state-of-the-art video object detection methods that use stronger backbone networks, as shown in Table 2. The +1.3% performance gain on mAP achieved by the ResNeXt version of DSFNet can be ascribed to the more powerful features extracted by the stronger backbone network. As a result, the fused features in the ResNeXt version of DSFNet contain more beneficial semantic information, which makes it more robust in handling the aforementioned challenges in video object detection.

5 CONCLUSION

In this paper, we present a novel dual semantic fusion network (named DSFNet) for video object detection. In DSFNet, both frame-level and instance-level semantics contained in input videos are distilled and fused to generate enhanced features for robust video object detection. Different from the existing one-stage feature enhancement methods that perform the feature fusion at either frame level or instance level with external guidance, DSFNet combines both frame-level and instance-level feature fusions, which can be learned in a unified fusion framework without any external guidance. In addition, we also introduce a new geometric similarity measure to mitigate the information distortion caused by noise during the fusion process. Extensive experiments on the large scale ImageNet VID dataset demonstrate the effectiveness and superiority of the proposed DSFNet. In particular, compared with several other cutting-edge methods, DSFNet has achieved the best performance of 84.1% mAP with ResNet-101 and 85.4% mAP with ResNeXt-101 without using any post-processing steps. Moreover, the proposed two-stage semantic fusion scheme in DSFNet is generic for video object detection, which can inspire more future works.

REFERENCES

[1] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. 2018. Object detection in video with spatio temporal sampling networks. In *Proc. of ECCV*. 331–346.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).

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

[4] Scott Carter, Laurent Denoue, and Daniel Avrahami. 2019. Documenting physical objects with live video and object detection. In *Proc. of ACM MM*. 1032–1034.

[5] Wenbin Che, Xiaopeng Fan, Ruiqin Xiong, and Debin Zhao. 2018. Paragraph generation network with visual relationship detection. In *Proc. of ACM MM*. 1435–1443.

[6] Kai Chen, Jiaqi Wang, Shuo Yang, Xingcheng Zhang, Yuanjun Xiong, Chen Change Loy, and Dahua Lin. 2018. Optimizing video object detection via a scale-time lattice. In *Proc. of CVPR*. 7814–7823.

[7] Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. 2019. Object guided external memory network for video object detection. In *Proc. of ICCV*. 6678–6687.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. of CVPR*. 248–255.

[9] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. 2019. Relation distillation networks for video object detection. In *Proc. of ICCV*. 7023–7032.

[10] Xuanyi Dong, Deyu Meng, Fan Ma, and Yi Yang. 2017. A dual-network progressive approach to weakly supervised object detection. In *Proc. of ACM MM*. 279–287.

[11] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. 2015. Flownet: Learning optical flow with convolutional networks. In *Proc. of CVPR*. 2758–2766.

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

[13] Christian Eggert, Dan Zecha, Stephan Brehm, and Rainer Lienhart. 2017. Improving small object proposals for company logo detection. In *Proc. of ACM MM*. 167–174.

[14] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2017. Detect to track and track to detect. In *Proc. of ICCV*. 3038–3046.

[15] Zhihang Fu, Zhongming Jin, Guo-Jun Qi, Chen Shen, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. 2018. Previewer for multi-scale object detector. In *Proc. of ACM MM*. 265–273.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. of CVPR*. 580–587.

[17] Yundi Guo, Beiji Zou, Ju Ren, Qingqing Liu, Deyu Zhang, and Yaoxue Zhang. 2019. Distributed and efficient object detection via interactions among devices, edge, and cloud. *IEEE TMM* 21, 11 (2019), 2903–2915.

[18] Chaojun Han, Fumin Shen, Li Liu, Yang Yang, and Heng Tao Shen. 2018. Visual spatial attention network for relationship detection. In *Proc. of ACM MM*. 510–518.

[19] Guangxing Han, Xuan Zhang, and Chongrong Li. 2018. Semi-supervised DFF: Decoupling detection and feature flow for video object detectors. In *Proc. of ACM MM*. 1811–1819.

[20] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. 2016. Seq-NMS for video object detection. *arXiv preprint arXiv:1602.08465* (2016).

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

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

[23] Zhong Ji, Qiankun Kong, Haoran Wang, and Yanwei Pang. 2019. Small and dense commodity object detection with multi-scale receptive field attention. In *Proc. of ACM MM*. 1349–1357.

[24] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, et al. 2017. T-CNN: Tubelets with convolutional neural networks for object detection from videos. *IEEE TCSVT* 28, 10 (2017), 2896–2907.

[25] Hei Law and Jia Deng. 2018. Cornernet: Detecting objects as paired keypoints. In *Proc. of ECCV*. 734–750.

[26] Byungjae Lee, Enkhbayar Erdenee, Songguo Jin, Mi Young Nam, Young Giu Jung, and Phill Kyu Rhee. 2016. Multi-class multi-object tracking using changing point detection. In *Proc. of ECCV*. 68–83.

[27] Jianan Li, Xiaodan Liang, ShengMei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. 2017. Scale-aware fast R-CNN for pedestrian detection. *IEEE TMM* 20, 4 (2017), 985–996.

[28] Jianan Li, Yunchao Wei, Xiaodan Liang, Jian Dong, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. 2016. Attentive contexts for object detection. *IEEE TMM* 19, 5 (2016), 944–954.

[29] Tianwei Lin, Xu Zhao, and Zheng Shou. 2017. Single shot temporal action detection. In *Proc. of ACM MM*. 988–996.

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

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proc. of ICCV*. 2999–3007.

[32] Mason Liu and Menglong Zhu. 2018. Mobile video object detection with temporally-aware feature maps. In *Proc. of CVPR*. 5686–5695.

[33] Mason Liu, Menglong Zhu, Marie White, Yinxiao Li, and Dmitry Kalenichenko. 2019. Looking fast and slow: Memory-guided mobile video object detection. *arXiv preprint arXiv:1903.10172* (2019).

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

[35] Yongyi Lu, Cewu Lu, and Chi-Keung Tang. 2017. Online video object detection using association LSTM. In *Proc. of ICCV*. 2344–2352.

[36] Jacinto C Nascimento and Jorge S Marques. 2006. Performance evaluation of object detection algorithms for video surveillance. *IEEE TMM* 8, 4 (2006), 761–774.

[37] Kaoru Ota, Minh Son Dao, Vasileios Mezaris, and Francesco GB De Natale. 2017. Deep learning for mobile multimedia: A survey. *ACM TOMM* 13, 3s (2017), 1–22.

[38] Heqian Qiu, Hongliang Li, Qingbo Wu, Fanman Meng, Linfeng Xu, King N Ngan, and Hengcan Shi. 2020. Hierarchical context features embedding for object Detection. *IEEE TMM* (2020).

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

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

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

[42] Mykhailo Shvets, Wei Liu, and Alexander C Berg. 2019. Leveraging long-range temporal relationships between proposals for video object detection. In *Proc. of ICCV*. 9756–9764.

[43] Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810* (2017).

[44] Xu Sun, Tongwei Ren, Yuan Zi, and Gangshan Wu. 2019. Video visual relation detection via multi-modal feature fusion. In *Proc. of ACM MM*. 2657–2661.

[45] Peng Tang, Chunyu Wang, Xinggang Wang, Wenyu Liu, Wenjun Zeng, and Jingdong Wang. 2020. Object detection in videos by high quality object linking. *IEEE TPAMI* 42, 5 (2020), 1272–1278.

[46] Yuxing Tang, Xiaofang Wang, Emmanuel Dellandréa, and Liming Chen. 2016. Weakly supervised learning of deformable part-based models for object detection via region proposals. *IEEE TMM* 19, 2 (2016), 393–407.

[47] Qingyi Tao, Hao Yang, and Jianfei Cai. 2018. Exploiting web images for weakly supervised object detection. *IEEE TMM* 21, 5 (2018), 1135–1146.

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

[49] Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057* (2000).

[50] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. 2018. Fully motion-aware network for video object detection. In *Proc. of ECCV*. 542–557.

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

[52] Ziwei Wang, Ziyi Wu, Jiwen Lu, and Jie Zhou. 2020. BiDet: An efficient binarized object detector. In *Proc. of CVPR*.

[53] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. 2019. Sequence level semantics aggregation for video object detection. In *Proc. of ICCV*. 9217–9225.

[54] Fanyi Xiao and Yong Jae Lee. 2018. Video object detection with an aligned spatial-temporal memory. In *Proc. of ECCV*. 485–501.

[55] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proc. of CVPR*. 1492–1500.

[56] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. 2016. UnitBox: An advanced object detection network. In *Proc. of ACM MM*. 516–520.

[57] Hao Zhou, Chongyang Zhang, and Chuanping Hu. 2019. Visual relationship detection with relative location mining. In *Proc. of ACM MM*. 30–38.

[58] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. 2018. Towards high performance video object detection. In *Proc. of CVPR*. 7210–7218.

[59] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Flow-guided feature aggregation for video object detection. In *Proc. of ICCV*. 408–417.