Exploring Spatial-temporal Instance Relationships In an Intermediate Domain For Image-to-video Object Detection

Zihan Wen, Jin Chen, and Xinxiao Wu*

Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing, China
{wenzihan,chen_jin,wuxinxiao}@bit.edu.cn

Abstract. Image-to-video object detection leverages annotated images to help detect objects in unannotated videos, so as to break the heavy dependency on the expensive annotation of large-scale video frames. This task is extremely challenging due to the serious domain discrepancy between images and video frames caused by appearance variance and motion blur. Previous methods perform both image-level and instance-level alignments to reduce the domain discrepancy, but the existing false instance alignments may limit their performance in real scenarios. We propose a novel spatial-temporal graph to model the contextual relationships between instances to alleviate the false alignments. Through message propagation over the graph, the visual information from the spatial and temporal neighboring object proposals are adaptively aggregated to enhance the current instance representation. Moreover, to adapt the source-biased decision boundary to the target data, we generate an intermediate domain between images and frames. It is worth mentioning that our method can be easily applied as a plug-and-play component to other image-to-video object detection models based on the instance alignment. Experiments on several datasets demonstrate the effectiveness of our method. Code will be available at: https://github.com/wenzihan/STMP.

Keywords: Deep learning · Object detection · Domain adaptation.

1 Introduction

Tremendous progress has been achieved on object detection in videos [10, 9, 23, 19, 3] thanks to the great success of deep neural networks. However, training deep object detectors requires annotating large-scale video frames, which is often time-consuming and labor-intensive. On the other hand, images are much easier and cheaper to be annotated, and there are also many existing labeled image datasets that can be readily utilized.

Therefore, image-to-video object detection has been proposed to leverage annotated images for detecting objects in unannotated videos, as to break the heavy

* Corresponding author

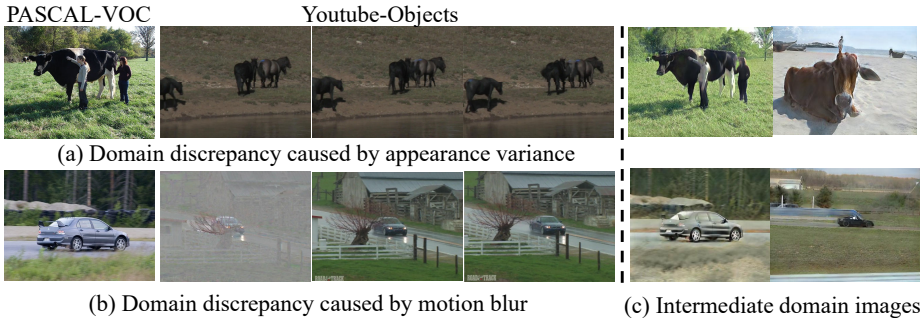


Fig. 1. Illustration of the domain discrepancy between images from the PASCAL-VOC dataset and video frames from the Youtube-Objects dataset. Although both datasets have the samples of "horse" and "car", the domain discrepancy caused by appearance variance (a) and motion blur (b) still make it challenging to apply an object detector learned from source images to target video frames. In order to reduce the domain discrepancy between images and video frames, we propose to generate an intermediate domain and some intermediate domain images are shown in (c).

dependency on the expensive annotation of frames. However, directly applying the detector trained on source images may significantly hurt its performance on detection in target videos, since there exists serious domain discrepancy between images and frames caused by appearance variance and motion blur, as shown in Fig 1. To address this problem, several prior works [4, 17, 22, 1] mainly focus on doing image-level and instance-level alignments of source images and target video frames to minimize the domain discrepancy. Although these methods have achieved promising results, the bounding box deviation, occlusion and out of focus may lead to the false alignments between cross-domain instances, which degrades their performance and limits their applications in real scenarios.

Taking into account the spatial-temporal context of instances as a potentially valuable information source for alleviating the false instance alignments, in this paper, we propose a novel spatial-temporal message propagation method to model the contextual relationships between instances for image-to-video object detection. By propagating message over the graph, the visual information of neighboring object proposals both within the same frame and from the adjacent frames are adaptively aggregated to enhance the current object instance representation. This feature aggregation strategy can mitigate the influence of bounding box deviation, occlusion and out of focus in images or frames. For example, spatially neighboring proposal features that are aggregated according to the intersection over union are helpful for relieving the deviation of bounding boxes, and temporally adjacent proposal features that are aggregated under the guidance of optical flow are beneficial to the alleviation of effects from the occlusion and motion blur. To be more specific, we first build an undirected graph where the nodes are represented by the region proposals and the edges are represented by the spatial-temporal instance relationships between the nodes.

Then we introduce a single-layer graph convolution network with normal propagation rule [12] for message propagation over the graph and update the node representation by aggregating the propagated features from the spatial-temporal neighboring nodes.

Moreover, to adapt the source-biased decision boundary to the target domain, we generate an intermediate domain between the source images and target video frames via generative adversarial learning. It serves as a bridge for connecting the source and target domains, and the cross-domain alignments on both the image level and the instance level are performed on this domain, which further boosts the positive transfer of the object detector between different domains.

The main contributions of this work are three-fold:

- We propose a novel spatial-temporal graph to model the contextual relationships between object instances for alleviating the false instance alignments in image-to-video object detection. It can be easily and readily applied as a plug-and-play component for other detection models based on the instance alignment.
- We propose an intermediate domain between the source images and the target video frames to reduce the domain discrepancy, thereby facilitating the cross-domain alignment.
- Extensive experiments on several datasets demonstrate that our method achieves better performance than existing methods, validating the effectiveness of modeling the instance relationships via a spatial-temporal graph.

2 Related work

Video object detection is vulnerable to motion blur, illumination variation, occlusion and scale changes. To address this problem, many methods have been proposed to utilize temporal context for detection, roughly falling into two categories: box-level propagation and feature-level propagation. The box-level propagation methods [10, 9] explore bounding box relations and apply temporal post-processing to suppress false positives and recover false negatives. T-CNN [10] incorporates temporal and contextual information from tubelets obtained in videos which dramatically improves the baseline performance of existing still image detection frameworks. Seq-NMS [9] uses high-scoring object detections from nearby frames to boost scores of weaker detections within the same clip. The feature-level propagation methods [23, 19, 3] use the temporal coherence on features to solve the problem. FGFA [23] improves the per-frame features by the aggregation of nearby features along the motion paths, and thus improves the video detection accuracy. Based on FGFA, MANet [19] jointly calibrates the features of objects on both pixel-level and instance-level in a unified framework. MEGA [3] augments proposal features of the key frame by effectively aggregating global and local information. All these methods of video object detection heavily depend on the manually per-frame annotations of bounding box coordinates and categories that is usually time-consuming and labor-expensive.

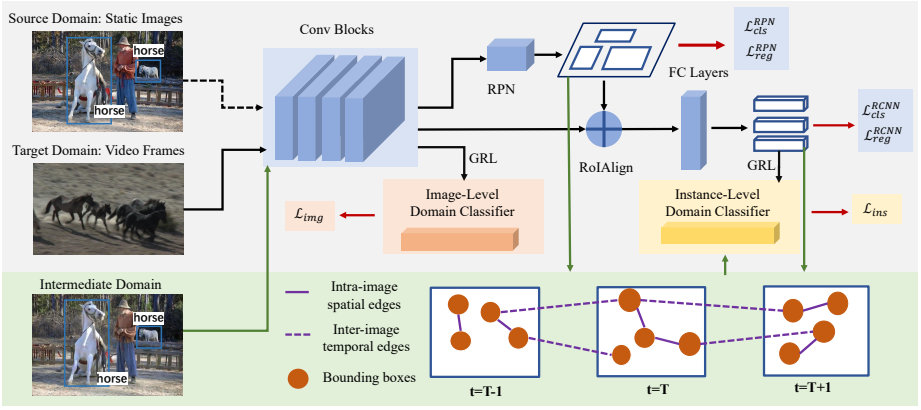


Fig. 2. Overview of our framework. The top of the framework is a basic image-to-video detector based on [4]. The bottom of the framework is our proposed spatial-temporal message propagation (STMP) and intermediate domain, which serves as a plug-and-play component of basic image-to-video detector for positive alignments on both image and instance levels. The source images linked with a dotted line are utilized to generate the intermediate domain and not used for the training of the image-to-video detector.

To break the dependency on the large-scale annotated video frames, image-to-video object detection is proposed to leverage existing annotated images to help object detection in unannotated videos. Several image-image object detection methods have been proposed in recent years. DA-Faster [4] is a prominent and effective approach for domain adaptive object detection, which introduces two domain classifiers on both image and instance levels to alleviate the performance drop caused by domain shift. SW-Faster [17] is an improvement of DA-Faster, which proposes a novel approach based on strong local alignment and weak global alignment. SCDA [22] improves DA-Faster by replacing the plain image-level alignment model with a region-level alignment model. HTCN [1] extends the ability of previous adversarial-based adaptive detection methods by harmonizing the potential contradiction between transferability and discriminability. These cross-domain object detection methods can also be used to reduce the domain gap between images and video frames in image-to-video object detection.

In the aforementioned methods, there often exists false alignments between instances across the domains due to appearance variations and motion blur. In this paper, we attempt to handle the false instance alignments by taking full advantage of spatial and temporal information within and across frames in videos to enhance primal instance representations, thus boosting the positive instance alignments.

3 Our Method

3.1 Overview

In this paper, we propose a novel spatial-temporal graph to address the false instance alignments in image-to-video object detection and an intermediate domain to reduce the domain discrepancy. Specifically, we first generate an intermediate domain by learning a transformation between source images and target video frames to reduce the domain discrepancy at the image level. Based on the intermediate domain and the target video domain, we then construct a spatial-temporal graph to model the instance relationships, and incorporates it into the domain adaptive Faster R-CNN [4, 17] to relieve the false instance alignments. The overview of our method is illustrated in Fig. 2.

For the image-to-video object detection task, we have an annotated source image domain of N_s images, denoted as $\mathcal{D}_s = \{\boldsymbol{x}_i^s, y_i^s\}_{i=1}^{N_s}$, and an unannotated target video domain that consists of N_t video frames, denoted as $\mathcal{D}_t = \{\boldsymbol{x}_i^t\}_{i=1}^{N_t}$. The region proposals are generated via a region proposal network. Let $B_i^s = \{b_{i,j}^s|_{j=1}^{N_s^b}\}$ represent a set of region proposals in $\boldsymbol{x}_i^s$, where N_s^b is the number of region proposals in $\boldsymbol{x}_i^s$. Let $B_i^t = \{b_{i,j}^t|_{j=1}^{N_t^b}\}$ represent a set of region proposals in $\boldsymbol{x}_i^t$, where N_t^b is the number of region proposals in $\boldsymbol{x}_i^t$.

3.2 Intermediate Domain

Due to the large discrepancy between the source and target domains, the performance of detector degrades substantially. In this paper, we propose to generate an intermediate domain $\mathcal{D}_f$ to bridge the source and target domains by learning to translate the source images into the target video frames. A typical and powerful image-to-image translation network, CycleGAN [21], is employed to learn a transformation between the source image domain $\mathcal{D}_s$ and the target video domain $\mathcal{D}_t$. Since ground truth labels are only accessed for source domain, we merely consider the transformation from source images to target frames after training CycleGAN and then translate the source domain $\mathcal{D}_s$ into an intermediate domain $\mathcal{D}_f = \{\boldsymbol{x}_i^f, y_i^s\}_{i=1}^{N_s}$. $\mathcal{D}_s$ and $\mathcal{D}_f$ are similar in image content, but diverges in visual appearances, while $\mathcal{D}_f$ and $\mathcal{D}_t$ differ in image content, but have similar distributions on the pixel-level. Therefore, our intermediate domain constitutes an intermediate feature space distributed neutrally between the source and target domains. As shown in Fig 1, we give some visualization results of the generated intermediate-domain images.

3.3 Domain Adaptive Faster R-CNN

Based on the generated intermediate domain $\mathcal{D}_f$ and the target domain $\mathcal{D}_t$, We utilize DA-Faster [4] to enable a basic image-to-video object detection model, which consists of an object detector (Faster R-CNN [16]) and an adaptation module.

Faster R-CNN is a two-stage detector mainly consisting of three main components: a deep convolutional neural network (*i.e.* "Conv Blocks" in Fig. 2) to extract features, a region proposal network (*i.e.*"RPN" in Fig. 2) to generate region proposals, and a full connected network (*i.e.*"FC Layers" in Fig. 2) to focus on bounding box detection and regression. The loss function of Faster R-CNN is summarized as

$$\mathcal{L}_{det} = \mathcal{L}_{cls}^{RPN} + \mathcal{L}_{reg}^{RPN} + \mathcal{L}_{cls}^{RCNN} + \mathcal{L}_{reg}^{RCNN}. \tag{1}$$

The adaptation module aims at aligning the distribution between the intermediate domain and the target domain at both image and instance levels, which consists of an image-level domain classifier D_{img} and an instance-level domain classifier D_{ins}. Specifically, let d_i denote the domain label of the i-th training image either in the intermediate domain or the target domain, with $d_i = 0$ for the intermediate domain and $d_i = 1$ for the target domain. By denoting the output of D_{img} located at (u, v) as $p_i^{(u,v)}$, the image-level adaptation loss can be written as

$$\mathcal{L}_{img} = \sum_{i,u,v} \Big[d_i \log p_i^{(u,v)} + (1 - d_i) \log \Big(1 - p_i^{(u,v)} \Big) \Big]. \tag{2}$$

Let $p_{i,j}$ denote the output of the instance-level domain classifier D_{ins} for the j-th region proposal in the i-th image. The instance-level adaptation loss can be written as

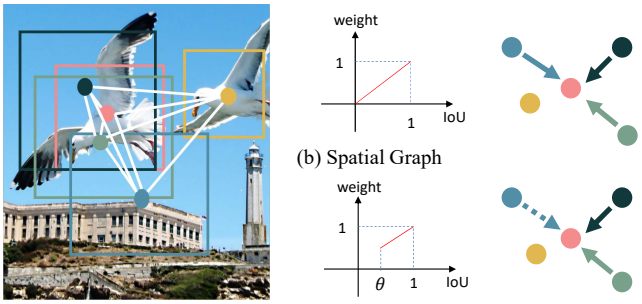
$$\mathcal{L}_{ins} = \sum_{i,j} [d_i \log p_{i,j} + (1 - d_i) \log (1 - p_{i,j})]. \tag{3}$$

To align the domain distributions, the parameters of image-level and instance-level domain classifiers should be optimized to minimize the above corresponding domain classification loss, while the base network should be optimized to maximize the training loss. For the implementation, the gradient is reversed by Gradient Reverse Layer (GRL) [7] to conduct the adversarial training between (D_{img},D_{ins}) and the base network of Faster R-CNN.

3.4 Spatial-Temporal Instance Relationships Construction

To avoid the false instance alignments during the learning of domain adaptive Faster R-CNN, we propose a spatial-temporal graph to model the contextual relationships between instances on both the spatial and temporal dimensions. As shown in Fig. 2, an intra-image spatial sparse graph and an inter-image temporal sparse graph are successively constructed for spatial-temporal message propagation. After message propagation, we obtain more accurate instance representations, which are fed into the instance-level domain classifier D_{ins} for positive instance alignment. Note that the inter-image temporal sparse graph is an extension of the intra-image spatial sparse on the temporal dimension.

Intra-image Spatial Sparse Graph Construction. We construct an intra-image spatial sparse graph for spatial message propagation. Specifically, for an



(a) Fully Connected Graph (b) Spatial Graph (c) Spatial Sparse Graph

Fig. 3. Different choices of constructing a graph to encode spatial relationships: (a) Fully connected graph: implicitly learning a fully connected graph between region proposals. The learned graph is redundant and ignores spatial information between region proposals. (b) Spatial graph: using Intersection Over Union (IoU) between region proposals to learn a spatial graph. However, there still exist redundant edges, which are useless to get a better instance representation, and even damage the primal features. (c) Spatial sparse graph: a spatial sparse graph is learned via adding constraints on the sparsity of a spatial graph.

image $\boldsymbol{x}_i^f$ in the intermediate domain $\mathcal{D}_f$ or a frame $\boldsymbol{x}_i^t$ in the target domain $\mathcal{D}_t$, we structure its region proposals as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ is the set of nodes corresponding to the proposal set B_i^f of $\boldsymbol{x}_i^f$ or B_i^t of $\boldsymbol{x}_i^t$ and each node is corresponding to one proposal in the proposal set. $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$ is the set of edges and represents the relationships between proposals within an image or a video frame. Intuitively, two spatially neighboring proposals are likely to represent the same object and should be highly correlated. Intersection Over Union (IoU) is a broadly used metric that measures the spatial correlation of two proposals by their bounding box coordinates. Hence, we employ IoU to construct a spatial-aware adjacency matrix $\mathbf{A}$, formulated as

$$A_{j,k} = \text{IoU}\left(b_{i,j}, b_{i,k}\right) = \frac{b_{i,j} \cap b_{i,k}}{b_{i,j} \cup b_{i,k}}, \tag{4}$$

where $A_{j,k}$ is the element in the j-th row and the k-th column of $\mathbf{A}$. $b_{i,j}$ and $b_{i,k}$ denote the j-th and k-th region proposals in $\boldsymbol{x}_i^f$ or $\boldsymbol{x}_i^t$, respectively. Although some irrelevant edges have been removed by using spatial information, the subsequent graph still over-propagates information and leads to confused proposal features. To solve this problem, we impose constraints on graph sparsity by only retaining the edges between region proposal $b_{i,j}$ and $b_{i,k}$ only if $\text{IoU}\left(b_{i,j}, b_{i,k}\right)$ is greater than a threshold θ_{spt}. In other words, for a node $b_{i,j}$, we select some relevant nodes as its neighborhoods, formulated as

$$\text{Neighbour}(\text{Node } b_{i,j}) = \{b_{i,k} | \text{IoU}(b_{i,j}, b_{i,k}) > \theta_{spt}\}. \tag{5}$$

As shown in Fig. 3, compared to a fully connected graph between region proposals, our spatial sparse graph picks up relevant neighborhoods for each node,

which greatly reduces the noise of irrelevant nodes and leads to low computation cost.

Inter-image Temporal Sparse Graph Construction. An inter-image temporal sparse graph is constructed to model the instance relationships on the temporal dimension, where the optical flow between consecutive video frames is utilized to improve the coordinates of region proposals from neighboring frames. Given two adjacent video frames $\boldsymbol{x}_{i-1}^t$ and $\boldsymbol{x}_i^t$ with their region proposals B_{i-1}^t and B_i^t, the temporal region proposal propagation is formulated by

$$B_{i-1\to i}^t = \text{flowbox}(B_{i-1}^t, \boldsymbol{\mathcal{F}}_{i-1\to i}), \tag{6}$$

$$\overline{B}_i^t = [B_{i-1\to i}^t,\ B_i^t], \tag{7}$$

where $\boldsymbol{\mathcal{F}}_{i\to i+1}$ is the optical flow map between $\boldsymbol{x}_{i-1}^t$ and $\boldsymbol{x}_i^t$, and flowbox() is a propagation function to generate pseudo proposals for $\boldsymbol{x}_i^t$ by adding the mean flow vectors to the region proposals in $\boldsymbol{x}_{i-1}^t$. After propagation, we construct an inter-image temporal sparse graph for the current frame $\boldsymbol{x}_i^t$ with region proposals $\boldsymbol{x}_i^t$. Different from the intra-image sparse spatial graph introduced above, the set of proposals $\mathcal{V}$ contains not only the proposals from its self frame, but also the pseudo proposals from adjacent frames. Accordingly, $\mathcal{E}$ denotes the set of edges between these proposals. We constantly use the IoU metric to form an adjacency matrix $\mathbf{A}$ and put constraints on sparsity by a threshold θ_{tmp} .

Spatial-temporal Message Propagation. By utilizing the intra-image spatial sparse graph and inter-image temporal sparse graph constructed above, we conduct spatial-temporal message propagation (STMP) to achieve accurate instance representation. With the assistance of adjacency matrix $\mathbf{A} \in \mathbb{R}^{N^b \times N^b}$ ($N^b = 256$ for spatial graph, while $N^b = 512$ for temporal graph), region proposals either from the intermediate or target domain of proposal features $\mathbf{F} \in \mathbb{R}^{N^b \times d}$ (d is the dimension of proposal feature) are aggregated by

$$\widetilde{\mathbf{F}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{F}, \tag{8}$$

where $\mathbf{D} = \sum_k A_{j,k}$ is the diagonal degree matrix of $\mathbf{A}$. After aggregating the intra-image and inter-image adjacent proposals, proposal feature $\widetilde{\mathbf{F}} \in \mathbb{R}^{N^b \times d}$ is discriminative enough and expresses more accurate instance-level information especially for low image quantity cases. Compared to the conventional graph convolution [12], we leave the trainable weight matrix $\mathbf{W}$ out. After STMP, we use $\widetilde{p}_{i,j}$ to denote the output of the instance-level domain classifier D_{ins} for the j-th region proposal in the i-th image in the intermediate domain or target domain. The instance-level alignment loss (*i.e.*, Eq. (3)) is rewritten as

$$\mathcal{L}_{ins}^{STMP} = \sum_{i,j} [d_i \log \widetilde{p}_{i,j} + (1 - d_i) \log (1 - \widetilde{p}_{i,j})]. \tag{9}$$

Objective Function. The overall objective function for jointly learning the domain adaptive Faster R-CNN and spatial-temporal message propagation is formulated by

$$\mathcal{L}_{DA-STMP} = \mathcal{L}_{det} + \lambda(\mathcal{L}_{img} + \mathcal{L}_{ins}^{STMP}), \tag{10}$$

where λ is a balance hyper-parameter to control the relative importance of detection and adaptation, and is set to 0.1 in our experiments.

4 Experiment

4.1 Datasets

To evaluate the effectiveness of our method, we conduct experiments on three public datasets, including PASCAL-VOC (VOC) [6], COCO [14], and Youtube-Objects (YTO) [15]. With these three datasets, we construct two image to video transfer tasks: VOC→YTO and COCO→YTO.

VOC→YTO. The VOC has 20 categories and consists of about $5,000$ training images with bounding box annotations. The YTO is a sparsely annotated video frame dataset for video object detection, which has about $4,300$ frames for training and $1,800$ frames for test with 10 categories. There are 10 common categories between VOC and YTO, including "aeroplane", "bird", "boat", "car", "cat", "cow", "dog", "horse", "bike", and "train". The images of the common 10 categories on the VOC dataset are used as the source image domain. We use unannotated sparse training frames associated with their adjacent six frames as the target video domain.

COCO→YTO. The COCO is a large-scale real-world image dataset with 80 object categories. We randomly select $4,000$ images of the common 10 categories between COCO and YTO from the training set as the source domain. For the target domain, we use the same setting as the VOC→YTO task.

4.2 Experiment Settings

Baselines and Comparison Methods. In our experiment, DA-Faster [4] and SW-Faster [17] are adopted as our baseline detectors. The modified overall objectives of DA-Faster and SW-Faster are both equipped with Eq. (9). We also compare our method with Faster R-CNN [16] that directly adapts the model trained on images to videos and several other image-to-video object detection methods [13, 2, 20].

Implementation Details. The instance-level domain classifier D_{ins} for "SW-Faster" is constructed by three full-connected layers ($4096 \rightarrow 100 \rightarrow 100 \rightarrow 2$) and the first two layers are activated by the ReLU [8] function. For the domain classifiers in "SW-Faster" and "DA-Faster", we follow the settings in original papers [4, 17]. The learning ratio of each domain classifier to the backbone network of Faster R-CNN is set as $1:1$, *i.e.*, setting the parameter of GRL layer as 1. We adopt the VGG-16 [18] pretrained on ImageNet [5] as the backbone of Faster R-CNN, and finetune the overall network with a learning rate of 1×10^{-3} for 50k iterations and then reduce the learning rate to 1×10^{-4} for another 30k iterations. Each batch consists of one image from the intermediate domain and one video frame from the target domain. We employ RoIAlign (*i.e.*"RoIAlign" in Fig. 2) for RoI feature extraction. As for the training of CycleGAN, we set

the batch size to 1, and adopt the Adam optimizer [11] with a momentum of 0.5 and an initial learning rate of 0.0002. For evaluation, both the Average Precision (AP) of each category and the mean Average Precision (mAP) of all categories are computed with an IoU threshold 0.5 for both two transfer tasks.

Table 1. Experimental results (%) on the VOC→YTO task.

Methods	aero	bird	boat	car	cat	cow	dog	horse	bike	train	mAP
Faster R-CNN [16]	75.0	90.4	37.3	71.4	58.0	52.8	49.1	42.2	62.8	39.2	57.8
CycleGAN [13]	78.5	**97.2**	31.5	72.3	66.3	59.7	45.9	43.6	66.3	49.4	61.1
SW-ICR-CCR [20]	79.3	95.2	36.9	**75.6**	58.7	61.4	38.7	45.4	66.6	42.6	60.0
SIR [2]	78.9	95.3	31.1	66.1	61.3	56.3	48.1	42.7	64.6	34.4	57.9
DA-Faster [4]	77.1	96.8	31.8	72.5	60.3	59.4	39.6	43.0	63.7	40.3	58.4
DA-Faster-inter-STMP	81.2	95.4	**42.6**	71.4	**66.5**	**72.0**	48.2	49.3	67.1	47.1	64.1
SW-Faster [17]	81.2	95.2	29.4	74.4	56.4	55.3	41.2	44.3	66.5	40.9	58.5
SW-Faster-inter-STMP	**84.4**	95.8	39.2	75.1	64.9	61.7	**53.8**	**50.6**	**68.4**	**52.9**	**64.7**

Table 2. Experimental results (%) on the COCO→YTO task.

Methods	aero	bird	boat	car	cat	cow	dog	horse	bike	train	mAP
Faster R-CNN [16]	57.9	91.4	29.6	68.9	51.9	51.4	**64.2**	55.2	63.2	52.3	58.6
CycleGAN [13]	75.5	87.8	37.6	69.4	52.3	62.8	61.0	57.8	64.4	58.0	62.7
SW-ICR-CCR [20]	69.8	90.6	34.6	**72.4**	54.9	61.1	58.0	58.3	**66.4**	47.5	61.4
SIR [2]	65.6	91.6	25.8	67.6	47.6	60.7	54.8	54.3	60.9	50.1	57.9
DA-Faster [4]	**84.9**	**93.3**	32.2	71.6	61.7	**66.9**	47.8	47.9	64.8	43.6	61.5
DA-Faster-inter-STMP	78.3	92.7	**41.8**	69.7	60.1	64.5	54.7	55.4	63.8	63.3	**64.4**
SW-Faster [17]	71.7	90.7	29.9	71.6	53.0	60.8	59.3	**58.8**	60.7	56.8	61.3
SW-Faster-inter-STMP	81.8	91.3	30.1	70.2	**64.5**	60.5	58.2	55.6	64.3	**67.4**	**64.4**

4.3 Results

VOC→YTO. Table 1 shows the comparison results on YTO. First, our method outperforms all the compared methods on mAP, clearly demonstrating the effectiveness of our proposed method. Second, our proposed intermediate domain and spatial-temporal message propagation consistently boosts the performance of "DA-Faster" and "SW-Faster" detectors with gains of 5.7% and 6.2% on mAP, respectively. Third, it is noteworthy that for some difficult categories, "DA-Faster" and "SW-Faster" perform worse than "Faster R-CNN", probably due to that there exist false alignments across domains and the performance of domain adaptive detectors is limited. As for the false alignment categories in "SW-Faster" such as "boat", "cat" and "dog", "SW-Faster-inter-STMP" improves them by 9.8%, 8.5% and 12.6%, respectively. The performance drops a

Table 3. Ablation studies (%) on VOC→YTO and COCO→YTO tasks.

Methods	mAP	
	VOC→YTO	COCO→YTO
DA-Faster [4]	58.4	61.5
DA-Faster-inter	61.9	61.9
DA-Faster-inter-SMP	63.4	62.8
DA-Faster-inter-STMP	64.1	64.4
SW-Faster [17]	58.5	61.3
SW-Faster-inter	60.6	62.2
SW-Faster-inter-SMP	64.1	63.5
SW-Faster-inter-STMP	64.7	64.4

lot on the false alignment categories such as "boat" and "dog" in "DA-Faster", and "DA-Faster-inter-STMP" can greatly improve these difficult categories by 10.8% and 8.6%, respectively.

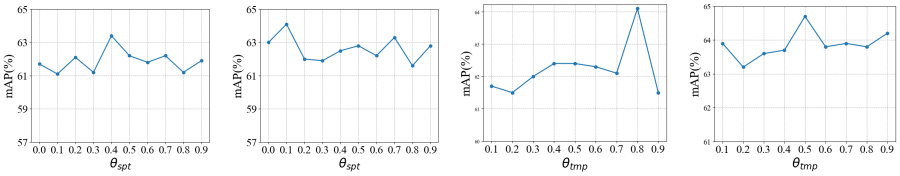
COCO→YTO. As shown in Table 2, our method outperforms all the compared methods on mAP. Moreover, we observe that our proposed framework improves "DA-Faster" and "SW-Faster" by 2.9% and 3.1%, respectively. Similar to the observation on the VOC→YTO task, we significantly improve the detection result of some false alignment categories such as "bike" by 3.6% for "SW-Faster" and "dog" by 6.9%, "horse" by 7.5%, "train" by 19.7% for "DA-Faster". In addition to these difficult categories, we further promote positive alignments in other simple categories, which validates the effectiveness of our method.

4.4 Ablation Study

To evaluate the effectiveness of each component, we conduct ablation studies on both the VOC→YTO and COCO→YTO tasks. The results are shown in Table 3, where "inter", "SMP", and "STMP" denote the intermediate domain, spatial message propagation, and spatial-temporal message propagation, respectively.

Effect of the intermediate domain: To evaluate the effect of the intermediate domain, we compare "DA-Faster-inter" with "DA-Faster" and "SW-Faster-inter" with "SW-Faster". For the VOC→YTO task, we observe that "DA-Faster-inter" and "SW-Faster-inter" achieve 3.5% and 2.1% improvements over "DA-Faster" and "SW-Faster", respectively. Similar improvements can be found for the COCO→YTO task, clearly demonstrating the effectiveness of the intermediate domain on promoting positive alignments between the source and target domains.

Effect of the spatial message propagation: To evaluate the effectiveness of the spatial message propagation, we compare "DA-Faster-inter-SMP" with "DA-Faster-inter" and "SW-Faster-inter-SMP" with "SW-Faster-inter". Their difference is whether handling false instance alignments by propagating spatial message within an image or video frame. From the results, "DA-Faster-inter-SMP" achieves better results compared to 'DA-Faster-inter" by spatial-temporal message propagation on two transfer tasks. Similar improvements are achieved



(a) (b) (c) (d)

Fig. 4. Analysis on spatial graph sparsity threshold θ_{spt} and temporal graph sparsity threshold θ_{tmp} on the VOC→YTO task. (a) DA-Faster-inter-SMP, (b) SW-Faster-inter-SMP, (c) DA-Faster-inter-STMP, (d) SW-Faster-inter-STMP.

with "SW-Faster" as the base detector. These improved results strongly validate the effectiveness of spatial message propagation in relieving the false instance alignments.

Effect of the temporal message propagation: To evaluate the effectiveness of the temporal message propagation, we compare "DA-Faster-inter-STMP" with "DA-Faster-inter-SMP" and "SW-Faster-inter-STMP" with "SW-Faster-inter-SMP". From the results shown in Table 3, "SW-Faster-inter-STMP" works better than "SW-Faster-inter-SMP" for VOC→YTO and COCO→YTO, respectively. Also, "DA-Faster-inter-STMP" outperforms "DA-Faster-inter-SMP". It validates that temporal message propagation can contribute to better instance representations for improving instance-level alignment.

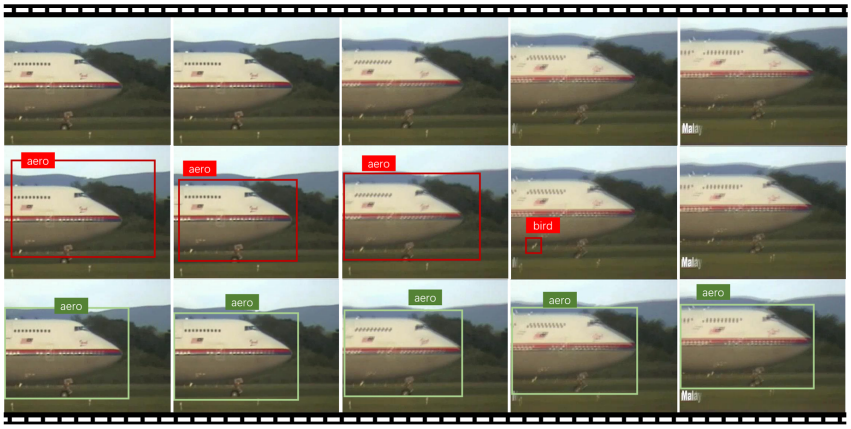
4.5 Parameter Analysis

To analyze the influence of the spatial graph sparsity threshold θ_{spt} and the temporal graph sparsity threshold θ_{tmp} on spatial-temporal message propagation, we conduct experiments using "DA-Faster" and "SW-Faster" as the baseline detectors for the VOC→YTO task. We select θ_{spt} and θ_{tmp} in the range of {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}, and show the mAP-threshold curve in Fig. 4, where the horizontal axis represents the value of θ_{spt} or θ_{tmp} and the vertical axis represents the mAP.

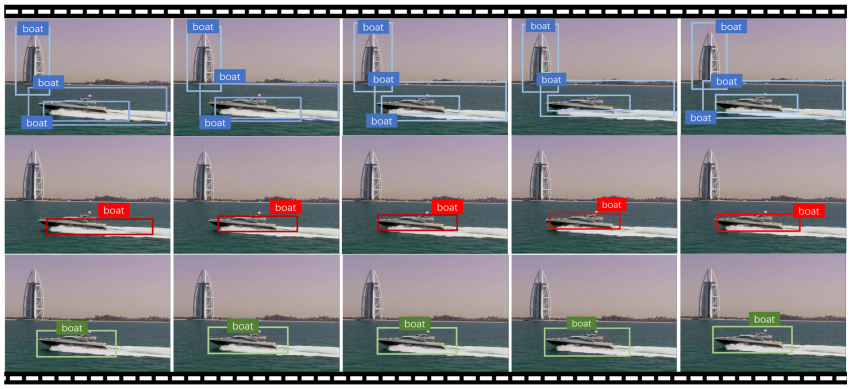
We conduct experiments on "DA-Faster-inter-SMP" and "SW-Faster-inter-SMP" to analyze the performance of θ_{spt}. From the results in Fig. 4 (a) and (b), we can find that small and large spatial graph sparsity thresholds θ_{spt} both lead to the decreasing of mAP. This is probably because that the smaller the spatial graph sparsity θ_{spt} , the more noise will be introduced and the larger the spatial

Table 4. Setting of spatial-temporal graph sparsity thresholds for both VOC→YTO and COCO →YTO tasks.

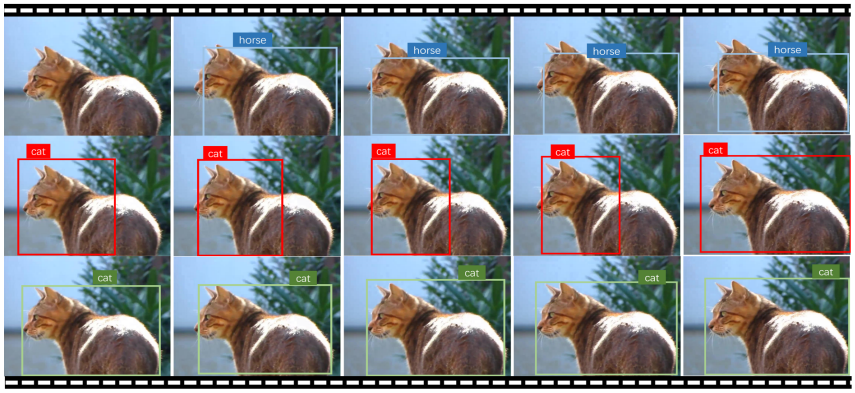
Base detector	θ_{spt}	θ_{tmp}
DA-Faster	0.4	0.8
SW-Faster	0.1	0.5



(a) Missing detections



(b) Redundant detections



(c) Wrong detections

Fig. 5. Detection examples on the COCO→YTO task. Bounding boxes in blue, red, and green denote the detection results of "SW-Faster", "SW-Faster-inter-STMP" and ground truth, respectively. (**Best viewed in color.**)

graph sparsity θ_{spt} , the less message will be propagated between instances to form better instance representations. Based on the experimental results, we set $\theta_{spt} = 0.4$ using the base detector of "DA-Faster" and $\theta_{spt} = 0.1$ using the base detector of "SW-Faster" to balance message propagation and noise filtering. To analyze the performance of θ_{tmp}, we use the θ_{spt} selected before to conduct experiments on "DA-Faster-inter-STMP" and "SW-Faster-inter-STMP". From the results in Fig. 4 (c) and (d), we select $\theta_{tmp} = 0.8$ and $\theta_{tmp} = 0.4$ as the sweet spot between message propagation and noise filtering for "DA-Faster" and "SW-Faster", respectively. For the COCO→YTO task, we use the same θ_{spt} and θ_{tmp}. Table 4 gives a detailed summary of the graph sparsity thresholds using two base detectors.

4.6 Qualitative Analysis

Fig.5 shows some detection examples of the COCO→YTO task by "SW-Faster" and "SW-Faster-inter-STMP (Ours)". There are three examples of five consecutive frames from YTO. As shown in Fig.5 (a), the base object detector fails to detect the blurred "aeroplane", while our method could partially recover the false negatives. It probably benefits from our proposed intermediate domain module that reduces the domain discrepancy. As shown in Fig.5 (b) and (c), the base object detector misclassifies the background tall building into a "boat", and misclassifies the foreground "cat" as a "horse". However, our method perfectly solves the redundant and wrong detections. This is probably because our proposed spatial-temporal message propagation module can successfully relieve the false instance alignments. In general, our method achieves more accurate detection results under the domain shift with poor image quality.

5 Conclusion

In this paper, we have presented a novel spatial-temporal graph to exploit spatial-temporal contextual relationships between object instances for alleviating the false instance alignments in image-to-video object detection. The generated intermediate domain can bridge the source image domain and the target video domain. With this intermediate domain and the target video domain, an intra-image spatial sparse graph and an inter-image temporal sparse graph have been constructed to enable the spatial-temporal message propagation, which can enrich the instance representation according to the guidance of spatial-temporal contextual. Extensive experiments on several datasets have demonstrated the effectiveness of our method.

Acknowledgments This work was supported in part by the Natural Science Foundation of China(NSFC) under Grant No 62072041.

References

1. Chen, C., Zheng, Z., Ding, X., Huang, Y., Dou, Q.: Harmonizing transferability and discriminability for adapting object detectors. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8869–8878 (2020)
2. Chen, J., Wu, X., Duan, L., Chen, L.: Sequential instance refinement for cross-domain object detection in images. IEEE Transactions on Image Processing (TIP) **30**, 3970–3984 (2021)
3. Chen, Y., Cao, Y., Hu, H., Wang, L.: Memory enhanced global-local aggregation for video object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10337–10346 (2020)
4. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 3339–3348 (2018)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition (CVPR). pp. 248–255. Ieee (2009)
6. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision (IJCV) **88**(2), 303–338 (2010)
7. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning (ICML). pp. 1180–1189. PMLR (2015)
8. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics (AISTATS). pp. 315–323. JMLR Workshop and Conference Proceedings (2011)
9. Han, W., Khorrami, P., Paine, T.L., Ramachandran, P., Babaeizadeh, M., Shi, H., Li, J., Yan, S., Huang, T.S.: Seq-nms for video object detection. arXiv preprint arXiv:1602.08465 (2016)
10. Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Zhang, C., Wang, Z., Wang, R., Wang, X., et al.: T-cnn: Tubelets with convolutional neural networks for object detection from videos. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) **28**(10), 2896–2907 (2017)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
13. Lahiri, A., Ragireddy, S.C., Biswas, P., Mitra, P.: Unsupervised adversarial visual level domain adaptation for learning video object detectors from images. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1807–1815. IEEE (2019)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision (ECCV). pp. 740–755. Springer (2014)
15. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3282–3289. IEEE (2012)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems (NeurIPS) **28**, 91–99 (2015)
17. Saito, K., Ushiku, Y., Harada, T., Saenko, K.: Strong-weak distribution alignment for adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6956–6965 (2019)

18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
19. Wang, S., Zhou, Y., Yan, J., Deng, Z.: Fully motion-aware network for video object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 542–557 (2018)
20. Xu, C.D., Zhao, X.R., Jin, X., Wei, X.S.: Exploring categorical regularization for domain adaptive object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11724–11733 (2020)
21. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp. 2223–2232 (2017)
22. Zhu, X., Pang, J., Yang, C., Shi, J., Lin, D.: Adapting object detectors via selective cross-domain alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 687–696 (2019)
23. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 408–417 (2017)