Mining Inter-Video Proposal Relations for Video Object Detection

Mingfei Han[1,2⋆], Yali Wang[1⋆], Xiaojun Chang[2], and Yu Qiao[1]

[1] Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China
{yl.wang, yu.qiao}@siat.ac.cn
[2] Faculty of Information Technology, Monash University, Melbourne, Australia
{hmf282, cxj273}@gmail.com

Abstract. Recent studies have shown that, context aggregating information from proposals in different frames can clearly enhance the performance of video object detection. However, these approaches mainly exploit the intra-proposal relation within single video, while ignoring the intra-proposal relation among different videos, which can provide important discriminative cues for recognizing confusing objects. To address the limitation, we propose a novel Inter-Video Proposal Relation module. Based on a concise multi-level triplet selection scheme, this module can learn effective object representations via modeling relations of hard proposals among different videos. Moreover, we design a Hierarchical Video Relation Network (HVR-Net), by integrating intra-video and inter-video proposal relations in a hierarchical fashion. This design can progressively exploit both intra and inter contexts to boost video object detection. We examine our method on the large-scale video object detection benchmark, i.e., ImageNet VID, where HVR-Net achieves the SOTA results. Codes and models are available at https://github.com/youthHan/HVRNet.

Keywords: Video Object Detection, Inter-Video Proposal Relation, Multi-Level Triplet Selection, Hierachical Video Relation Network

1 Introduction

Video object detection has emerged as a new challenge in computer vision [1, 5, 27, 28, 35, 41]. The traditional image object detectors [12, 22, 24, 25] often fail in this task, due to the fact that objects in videos often contain motion blur, sudden occlusion, rare pose, etc. Recent studies [5, 28] have shown that, modeling the relation of object proposal from different frames can effective aggregate spatio-temporal context and yield better representation for detection task. These approaches, however, only utilize the proposal-relations within the same video,

⋆ Equal contribution.

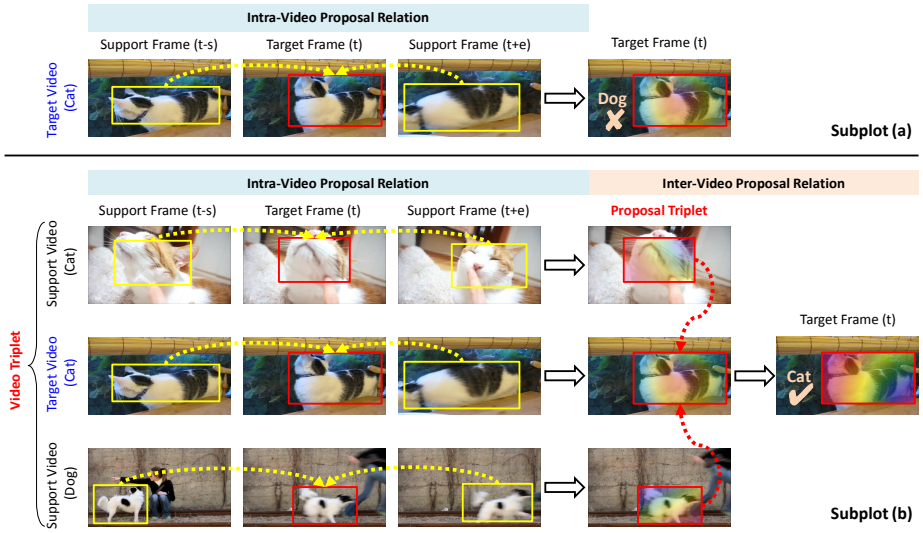

Fig. 1. Motivation. Subplot (a): The intra-video proposal relation often captures what *Cat* looks like and how it moves in this video. But it has little clue about object variations among different videos. As a result, the detector mistakenly recognizes *Cat* as *Dog* at the target frame t, even though it leverages spatio-temporal contexts from other support frames $t-s$ and $t+e$. Subplot (b): To tackle such problem, we design a novel inter-video proposal relation module, which can adaptively mine hard object proposals (i.e., proposal triplet) from highly-confused videos (i.e., video triplet), and effectively learn and correct their relation to reduce object confusion among videos.

and still face the difficulty to distinguish the confusing objects which have similar appearance and/or motion characteristics from different videos.

As shown in Fig. 1(a), the detector mistakenly recognizes *Cat* as *Dog* at the target frame t, even though it leverages spatio-temporal contexts from other support frames $t-s$ and $t+e$ to enhance the proposal representation of current frame. The main reason is that, such intra-video relation only describes what this *Cat* looks like and how it moves in this video. It has little clue about object relations or variations among different videos, e.g., in Fig. 1(b), *Cat* in the target video looks similar to *Dog* in the support video, while it looks different from *Cat* in the support video. In this case, the detector would get confused to distinguish *Cat* from *Dog*, if it only focuses on each individual video but without understanding object relations among different videos.

To address such difficulty, we design a novel Inter-Video Proposal Relation method, which can effectively leverage inter-video proposal relation to learn discriminative representations for video object detection. Specifically, we first introduce a multi-level triplet selection scheme to select hard training proposals among confused videos. Since these proposal triplets are the key factors to avoid confusion, we can exploit the relation on each of them to construct better object features. Moreover, we propose a powerful Hierarchical Video Relation

Network (HVR-Net), by integrating intra-video and inter-video proposal relation modules in a unified framework. In this case, it can progressively utilize both intra-video and inter-video contextual dependencies to boost video object detection. We investigate extensive experiments on the large-scale video object detection benchmark, i.e., ImageNet VID. Our HVR-Net shows its superiority with **83.8** mAP by ResNet101 and **85.4** mAP by ResNeXt101 32x4d.

2 Related Works

Object Detection in Still Images Object detection in still images [3, 8, 9, 14, 22, 24, 26] has recently achieved remarkable successes, with the fast development of deep neural networks [13, 18, 29, 30, 36] and large-scale well-annotated datasets [21, 27]. The existing approaches can be mainly categorized into two frameworks, i.e., two-stage frameworks (such as RCNN [9], Fast-RCNN [8], Faster-RCNN [26]) and one-stage frameworks (such as YOLO [24], SSD [22], RetinaNet [20]). Two-stage detectors often achieve a better detection accuracy, while one-stage detectors often maintain computation efficiency. Recently, anchor-free detectors also show the impressive performance [6, 19, 38, 39], inspired by key point detection. However, these image-based detectors often fail in video object detection, since they often ignore the challenging spatio-temporal characteristics in videos such as motion blur, object occlusion, etc.

Object Detection in Videos To improve object detection in still images, the previous works of video object detection often leverage temporal information, by box-level association and feature aggregation. Box-level association mainly designs the post-processing steps on image object detector, which can effectively produce tubelet of objects in videos. Such approaches [7, 11, 16] have been widely-used to boost performance. On the other hand, feature aggregation mainly takes nearby frames as video context to enhance feature representation of current frame [1, 5, 28, 35, 41]. In particular, recent studies have shown that, learning proposal relations among different frames can alleviate difficulty in detecting objects in videos [5, 28], via long-term dependency modeling [31, 33]. However, these approaches mainly focus on intra-video relations of object proposals, while neglecting object relations among different videos. As a result, they often fail to detect objects which contain highly-confused appearance or motion characteristics in videos. Alternatively, we design a novel HVR-Net, which can progressively integrate intra-video and inter-video proposal relations, based on a concise multi-level triplet selection scheme. This allows to effectively alleviate object confusion among videos to boost detection performance.

3 Our HVR-Net

Overview. In this section, we systematically introduce our Hierachical Video Relation Network (HVR-Net), which can boost video object detection by lever-

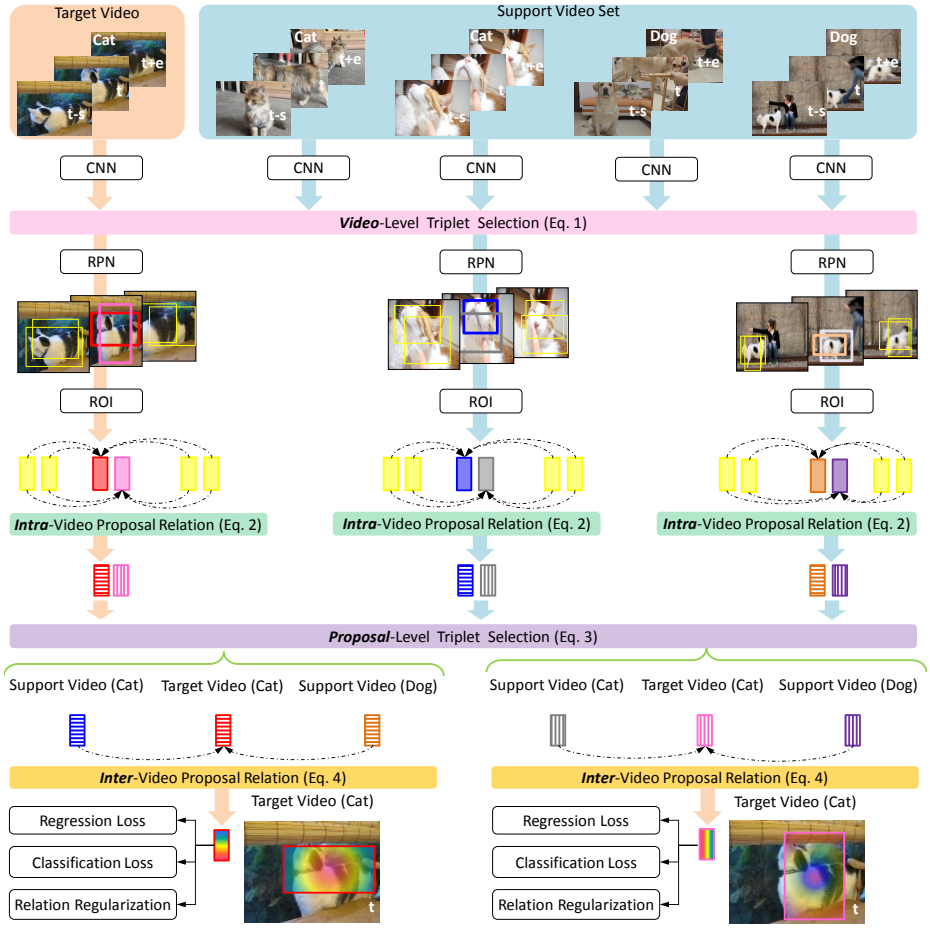

Fig. 2. Our HVR-Net Framework. It can effectively boost video object detection, by integrating intra-video and inter-video proposal relation progressively within a multi-level triplet selection scheme. More explanations can be found in Section 3.

aging both intra-video and inter-video contexts within a multi-level triplet selection scheme. The whole framework is shown in Fig. 2. **First**, we design a video-level triplet selection module. For a target video, it can flexibly select two confused videos from a set of support videos, i.e., the most dissimilar video in the same category, the most similar video in the different categories, according to their CNN features. As a result, we obtain a triplet of confused videos in each training batch, which can guide our HVR-Net to model object confusion among videos. **Second**, we introduce an intra-video proposal relation module. For each video in the triplet, we feed its sampled frames (e.g., $t-s$, t and $t+e$) into RPN and ROI layers of Faster RCNN. This produces feature vectors of object proposals for each frame. Subsequently, we aggregate proposals from sup-

port frames (e.g., $t-s$, $t+e$) to enhance proposals in the target frame t. As a result, each proposal feature in the target frame t integrates long-term dependencies in the corresponding video, which can address intra-video problem such as motion blur, occlusion, etc. **Third**, we develop a proposal-level triplet selection module. Note that, the intra-video-enhanced proposals mainly contain object semantics in each individual video, while ignoring object variations among videos. To model such variations, we select hard proposal triplets from the video triplet, according to the intra-video-enhanced features. **Finally**, we design an inter-video proposal relation module. For each proposal triplet, it can aggregate proposals from support videos to enhance proposals in the target video. In this case, each proposal feature further leverages inter-video dependencies to tackle object confusion among videos.

3.1 Video-Level Triplet Selection

To effectively alleviate inter-video confusions, we start finding a triplet of hard videos for training. Specifically, we randomly sample K object categories from training set, and randomly sample N videos per category. Hence, there are $K \times N$ videos in a batch. Then, we randomly select one video as *target video*, and use other $(K \times N - 1)$ videos as a set of *support videos*. For each video, we randomly sample one frame as *target frame* t, and sample other $T-1$ frames as *support frames*, e.g., frame $t-s$ and frame $t+e$ in Fig. 2.

For each video, we feed its T frames individually into the CNN backbone of Faster RCNN for feature extraction. As a result, the feature tensor of this video is with size of $H \times W \times C \times T$, where $H \times W$ and C are respectively the spatial size and the number of feature channels. Then, we perform global average pooling along spatial and temporal dimensions of this tensor, which produces a C-dimension video representation. According to cosine similarity between video representations, we find the video triplet

$$\mathcal{V}^{triplet} = \{\mathcal{V}^{target}, \mathcal{V}^{+}, \mathcal{V}^{-}\}, \tag{1}$$

where $\mathcal{V}^{+}$ is the most dissimilar support video in the class which $\mathcal{V}^{target}$ belongs to, and $\mathcal{V}^{-}$ is the most similar support video in the other classes.

3.2 Intra-Video Proposal Relation

After finding $\mathcal{V}^{triplet}$, we generate object proposals for each video in this triplet. Specifically, we feed the sampled T frames of each video into RPN and ROI layers of Faster RCNN, which produces M proposal features per frame.

Recent studies have shown that, spatio-temporal proposal aggregation among different frames [5, 28] can boost video object detection. Hence, we next introduce intra-video proposal relation module, which builds up proposal dependencies between target frame and support frames in each video. Specifically, we adapt a concise non-local-style relation module for $\mathcal{V}^{v}$ in the video triplet

$(v \in \{target, +, -\})$,

$$\boldsymbol{\alpha}_{t,m}^{v} = \mathbf{x}_{t,m}^{v} + \sum_{i \in \Omega} \sum_{j} g(\mathbf{x}_{t,m}^{v}, \mathbf{x}_{i,j}^{v}) \times \mathbf{x}_{i,j}^{v}, \tag{2}$$

where $\mathbf{x}_{t,m}^{v}$ is the m-th proposal feature in the target frame t, $\mathbf{x}_{i,j}^{v}$ is the j-th proposal feature in the support frame i, and i belongs to the set of support frames Ω (e.g., $\Omega = \{t-s, t+e\}$). As shown in Eq. (2), we first compare similarity between $\mathbf{x}_{t,m}^{v}$ and $\mathbf{x}_{i,j}^{v}$, by a kernel function $g(\cdot,\cdot)$, e.g., Embedded Gaussian in [33]. Then, we aggregate $\mathbf{x}_{t,m}^{v}$ by weighted sum over all the proposal features of support frames. As a result, $\boldsymbol{\alpha}_{t,m}^{v}$ is an enhance version of $\mathbf{x}_{t,m}^{v}$, which contains video-level object semantics to tackle motion blur, object occlusion, etc.

3.3 Proposal-Level Triplet Selection

After intra-video relation module, $\boldsymbol{\alpha}_{t,m}^{v}$ in Eq. (2) can integrate spatio-temporal object contexts from video $\mathcal{V}^{v}$ itself. However, it contains little clue to describe object relation among confused videos. To discover such inter-video relation, we propose to further select hard proposal triplets, from intra-video-enhanced proposals in the video triplet $\mathcal{V}^{triplet}$. Specifically, we compare cosine similarity between these proposals, according to their features in Eq. (2). For a proposal $\mathcal{P}_{t,m}^{target}$ in the target video, we obtain its corresponding proposal triplet,

$$\mathcal{P}^{triplet} = \{\mathcal{P}_{t,m}^{target}, \mathcal{P}^{+}, \mathcal{P}^{-}\}, \tag{3}$$

where $\mathcal{P}^{+}$ is the most dissimilar proposal in the same category, $\mathcal{P}^{-}$ is the most similar proposal in the other categories.

3.4 Inter-Video Proposal Relation

After finding all the proposal triplets, we model relation on each of them in order to describe object variation among videos. To achieve this goal, we use a concise non-local-style relation module for each proposal triplet,

$$\boldsymbol{\beta}_{t,m}^{target} = \boldsymbol{\alpha}_{t,m}^{target} + f(\boldsymbol{\alpha}_{t,m}^{target}, \boldsymbol{\alpha}^{+}) \times \boldsymbol{\alpha}^{+} + f(\boldsymbol{\alpha}_{t,m}^{target}, \boldsymbol{\alpha}^{-}) \times \boldsymbol{\alpha}^{-}, \tag{4}$$

where $f(\cdot,\cdot)$ is a kernel function (e.g., Embedded Gaussian) for similarity comparison, $\boldsymbol{\alpha}^{+}$ is the intra-video-enhanced feature of hard positive proposal $\mathcal{P}^{+}$, and $\boldsymbol{\alpha}^{-}$ is the intra-video-enhanced feature of hard negative proposal $\mathcal{P}^{-}$. By Eq. (4), we further aggregate the proposal $\mathcal{P}_{t,m}^{target}$ in the target video, with inter-video object relationships. Finally, to effectively reduce object confusions when performing detection, we introduce the follow loss for a target video,

$$\mathcal{L} = \mathcal{L}_{detection} + \gamma \mathcal{L}_{relation}, \tag{5}$$

where $\mathcal{L}_{detection} = \mathcal{L}_{regression} + \mathcal{L}_{classification}$ is the tradition detection loss (i.e., bbox regression and object classification) on the final proposal features $\boldsymbol{\beta}_{t,m}^{target}$

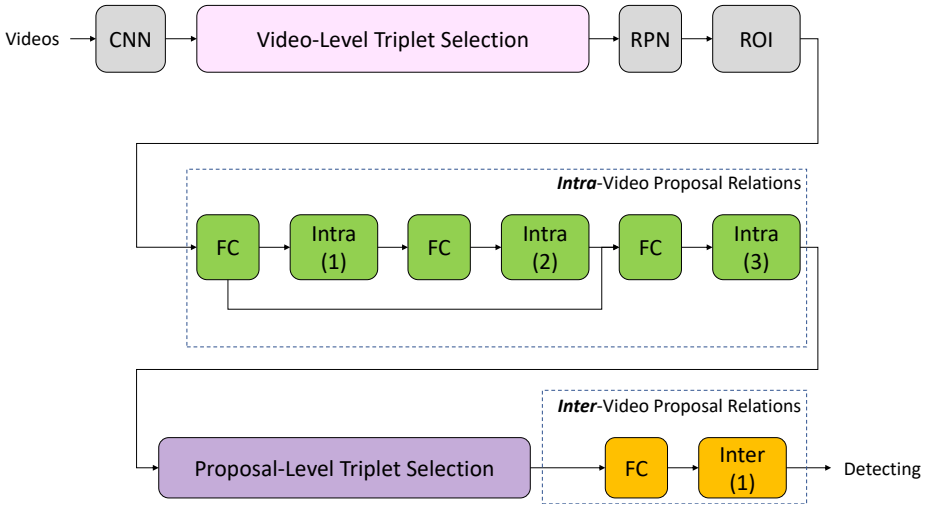

Fig. 3. HVR-Net Architecture. We flexibly adapt the widely-used Faster RCNN architecture as our HVR-Net. More implementation details can be found in Section 4.1.

in the target frame. γ is a weight coefficient. $\mathcal{L}_{relation}$ is a concise triplet-style metric loss to regularize Eq. (4),

$$\mathcal{L}_{relation} = \max(d(\boldsymbol{\alpha}_{t,m}^{target}, \boldsymbol{\alpha}^{-}) - d(\boldsymbol{\alpha}_{t,m}^{target}, \boldsymbol{\alpha}^{+}) + \lambda, 0). \tag{6}$$

Via this loss, we emphasize a discriminative relation constraint when computing $\boldsymbol{\beta}_{t,m}^{target}$ in Eq. (4), i.e., $d(\boldsymbol{\alpha}_{t,m}^{target}, \boldsymbol{\alpha}^{+}) > d(\boldsymbol{\alpha}_{t,m}^{target}, \boldsymbol{\alpha}^{-}) + \lambda$, where d is Euclidean distance. In this case, $\boldsymbol{\beta}_{t,m}^{target}$ becomes more discriminative to alleviate inter-video object confusion, by increasing relation between $\mathcal{P}_{t,m}^{target}$ and $\mathcal{P}^{+}$ as well as decreasing relation between $\mathcal{P}_{t,m}^{target}$ and $\mathcal{P}^{-}$.

4 Experiments

We mainly evaluate our HVR-Net on the large-scale ImageNet VID dataset [18]. It consists of 3862 training videos (1,122,397 frames) and 555 validation videos (176,126 frames), with bbox annotations across 30 object categories. Moreover, we train our model on intersection of ImageNet VID and DET dataset [5, 28], and report mean Average Precision (mAP) on validation set of VID.

4.1 Implementation Details

Architecture We flexibly adapt Faster RCNN to our HVR-Net with the following details. The architecture is shown in Fig. 3. We use ResNet-101 [13] as backbone for ablation studies, and also report the results on ResNeXt-101-32x4d [36] for SOTA comparison. We extract the feature of each sampled frame

Table 1. Effectiveness of our HVR-Net.

Methods	Intra-Video	Inter-Video	mAP(%)
Baseline: Faster-RCNN	-	-	73.2
Our HVR-Net	✓	-	$80.6_{\uparrow \mathbf{7.4}}$
Our HVR-Net	✓	✓	$\mathbf{83.2}_{\uparrow \mathbf{10.0}}$

Table 2. Multi-Level Triplet Selection of Our HVR-Net.

Multi-Level Triplet Selection	mAP(%)
Simple	81.0
Our	**83.2**

after the *conv*4 stage, in order to select video triplet in a training batch. Region Proposal Network (RPN) is used to generate proposals from each frame of the selected video triplet, by using the feature maps after the *conv*4 stage. We introduce three intra modules in Fig. 3. Before each of them, we add 1024-dim fully-connected (FC) layer. Additionally, we use a skip connection between intra(1) and intra(3), to increase learning flexibility. In this case, the intra(3) module can use both initial and transformed proposals of support frames to enhance proposals in the target frame.

We introduce one inter module which is added upon a 1024-dim FC layer. Additionally, for both intra and inter modules, the kernel function is set as Embedded Gaussian in [33], where each embedding transformation in this kernel is a 1024-dim FC layer.

Training Details We implement our HVR-Net on Pytorch, by 8 GPUs of 1080Ti. In each training batch, we randomly sample $K = 3$ object categories from training set, and randomly sample $N = 3$ videos per category. Hence, there are 9 videos in a batch. Then, we randomly select one video as target video, and use other 8 videos as the support video set. For each video, we randomly sample 3 frames, where the middle frame is used as target frame. More inplementation details could be found in supplementary materials.

4.2 Ablation Studies

Effectiveness of HVR-Net We first compare our HVR-Net with the baseline architecture, i.e., Faster RCNN. As shown in Table 1, our HVR-Net significantly outperforms Faster RCNN, indicating its superiority in video object detection. More importantly, HVR-Net with both intra-video and inter-video is better than that with intra-video only (83.2 vs. 80.6). It demonstrates that, learning proposal interactions inside each single video is not sufficient to describe category differences among videos. When adding inter-video proposal relation module, our

Table 3. Supervision in Our HVR-Net.

Detection Loss	Relation Regularization	mAP(%)
✓	-	80.0
✓	✓	**83.2**

Table 4. Number of Intra and Inter Modules in Our HVR-Net.

No. of Intra	No. of Inter	mAP(%)
2	1	81.8
3	1	**83.2**
3	2	82.1

Table 5. Number of Testing Frames in Our HVR-Net.

Testing Frames	5	11	17	21	31
mAP(%)	80.5	81.6	82.0	82.9	**83.2**

HVR-Net can flexibly select hard proposals from confused videos, and effectively build up relations among these proposals to distinguish object confusions.

Multi-Level Triplet Selection Our HVR-Net is built upon a multi-level triplet selection scheme, including video-level and proposal-level proposal selection. To demonstrate the effectiveness, we replace these two selection modules with a simple approach, i.e., selecting random videos and using all proposals in each video. In Table 2, when using the straightforward selection, the performance of HVR-Net is getting worse. The main reason is that, blindly selected videos and proposals do not guide our HVR-Net to focus on object confusion in videos. Alternatively, when we add our video and proposal triplet selection, HVR-Net can effectively leverage hard proposals of confused videos to learn and correct inter-video object relations, in order to boost video object detection.

Supervision in HVR-Net As mentioned in Section 3.4, we introduce a relation regularization in Eq. (6), in order to emphasize the correct relation constraint on inter-video relation module in Eq. (4). We investigate it in Table 3. As expected, this regularization can boost HVR-Net by a large margin, by enhancing similarity between proposals in the same category, and reducing similarity between proposals in the different categories.

Number of Intra and Inter Relation Modules We investigate the performance of our HVR-Net, with different number of intra-video and inter-video proposal modules. When changing the number of intra modules (or inter modules), we fix the number of inter modules (or intra modules). The results are

Table 6. Comparison with the state-of-the-art methods on ImageNet VID (mAP).

Methods	Backbone	Post-processing	Base detector	mAP(%)
D&T[7]	ResNet101	-	R-FCN	75.8
MANet[32]	ResNet101	-	R-FCN	78.1
LWDN[15]	ResNet101	-	R-FCN	76.3
RDN[5]	ResNet101	-	Faster-RCNN	81.8
LongRange[28]	ResNet101	-	FPN	81.0
Deng [4]	ResNet101	-	R-FCN	79.3
PSLA [10]	ResNet101+DCN	-	R-FCN	80.0
THP [40]	ResNet101+DCN	-	R-FCN	78.6
STSN[1]	ResNet101+DCN	-	R-FCN	78.9
Ours	ResNet101	-	Faster-RCNN	**83.2**
TCNN [17]	DeepID+Craft[37, 23]	Tublet Linking	RCNN	73.8
STMN [35]	ResNet101	Seq-NMS	R-FCN	80.5
FGFA[41]	Align. Incep.-ResNet	Seq-NMS	R-FCN	80.1
D&T($\tau = 10$)[7]	ResNet101	Viterbi	R-FCN	78.6
D&T($\tau = 1$)[7]	ResNet101	Viterbi	R-FCN	79.8
MANet[32]	ResNet101	Seq-NMS	R-FCN	80.3
ST-Lattice[2]	ResNet101	Tublet-Rescore	R-FCN	79.6
SELSA[34]	ResNet101	Seq-NMS	Faster-RCNN	82.5
Deng [4]	ResNet101	Seq-NMS	R-FCN	80.8
PSLA [10]	ResNet101+DCN	Seq-NMS	R-FCN	81.4
STSN+[1]	ResNet101+DCN	Seq-NMS	R-FCN	80.4
Ours	ResNet101	Seq-NMS	Faster-RCNN	**83.8**
D&T[7]	ResNeXt101	Viterbi	Faster-RCNN	81.6
D&T[7]	Inception-v4	Viterbi	R-FCN	82.1
LongRange[28]	ResNeXt101-32×8d	-	FPN	83.1
RDN[5]	ResNeXt101-64×4d	-	Faster-RCNN	83.2
RDN[5]	ResNeXt101-64×4d	Seq-NMS	Faster-RCNN	84.5
SELSA[34]	ResNeXt101-32×4d	-	Faster-RCNN	84.3
SELSA[34]	ResNeXt101-32×4d	Seq-NMS	Faster-RCNN	83.7
Ours	ResNeXt101-32×4d	-	Faster-RCNN	**84.8**
Ours	ResNeXt101-32×4d	Seq-NMS	Faster-RCNN	**85.5**

shown in Table 4. As expected, when increasing the number of both modules, the performance of HVR-Net is getting better and tends to become flat. Hence, in our experiment, we set the number of intra modules as three, and set the number of inter module as one.

Number of Testing Frames We investigate the performance of HVR-Net, w.r.t., the number of sampled frames in a testing video. As expected, when increasing the number of testing frames, the performance of HVR-Net is getting better and tends to become stable. Hence, we choose the number of testing frames as 31 in our experiment. Besides, we test HVR-Net by unloading inter-video

Table 7. Comparison with state-of-the-art methods in mAP.

Methods	Fast (mAP)	Medium (mAP)	Slow (mAP)
FGFA [41]	57.6	75.8	83.5
MANet [32]	56.7	76.8	86.9
Deng[4]	61.1	78.7	86.2
LongRange[28]	64.2	79.5	86.7
Ours	**66.6**	**82.3**	**88.7**

proposal relation module in the testing phase, which achieves the comparable mAP.

4.3 SOTA Comparison

We compare our HVR-Net with a number of recent state-of-the-art approaches on ImageNet VID validation set. As shown in Table 6 and Table 7, HVR-Net achieves the best performance among various settings and object categories.

In Table 6, we first make comparison without any video-level post-processing techniques. Under the same backbone, We significantly outperform the well-known approaches such as FGFA [41] and MANet [32], which uses expensive optical flow as guidance of feature aggregation. More importantly, our HVR-Net outperform the recent approaches [5, 28] that mainly leverage proposal relations among different frames for spatio-temporal context aggregation. This further confirms the effectiveness of learning inter-video proposal relation. Second, we equip HVR-Net with the widely-used post-processing approach Seq-NMS. Once again, we outperform other state-of-the-art approaches under the same backbone. It shows that, our HVR-Net is compatible and complementary with post-processing of video object detection, which can further boost performance.

Additionally, we follow FGFA [41] to evaluate detection performance on the categories of slow, medium, and fast objects, where these three categories are divided by their average IoU scores between objects across nearby frames, i.e., Slow (score>0.9), Medium (score$\in$[0.7,0.9]), Fast (Others). As shown in Table 7, our HVR-Net boost the detection performance on all these three categories, showing the importance of inter-video proposal relation for confusion reduction.

4.4 Visualization

Detection Visualization We show the detection result of HVR-Net in Fig. 4. Specifically, we compare two settings, i.e., baseline with only intra-video proposal relation module, and HVR-Net with both intra-video and inter-video proposal relation modules. As expected, when only using intra-video relation aggregation, baseline fails to recognize the object in the video, e.g., a female lion in Subplot (a) is mistakenly recognized as a horse with confidence larger than 0.9. The main reason is that, intra-video relation mainly focuses on what the object looks like

Fig. 4. Detection Visualization. For each video, the first row shows the baseline with only intra-video proposal relation module. The second row shows HVR-Net with both intra-video and inter-video proposal relation modules. Clearly, our inter-video can effectively guide HVR-Net to tackle object confusion in videos. For example, a female lion in Subplot (a) looks quite similar to a horse, due to its color and its motion in this video. As a result, the baseline mistakenly recognizes it as a horse, when only using intra-video relation aggregation. By introducing inter-video proposal relation, our HVR-Net successfully distinguish such object confusion in videos. Other subplots also exhibit the similar result, i.e., it is necessary and important to learn inter-video proposal relations to boost video object detection.

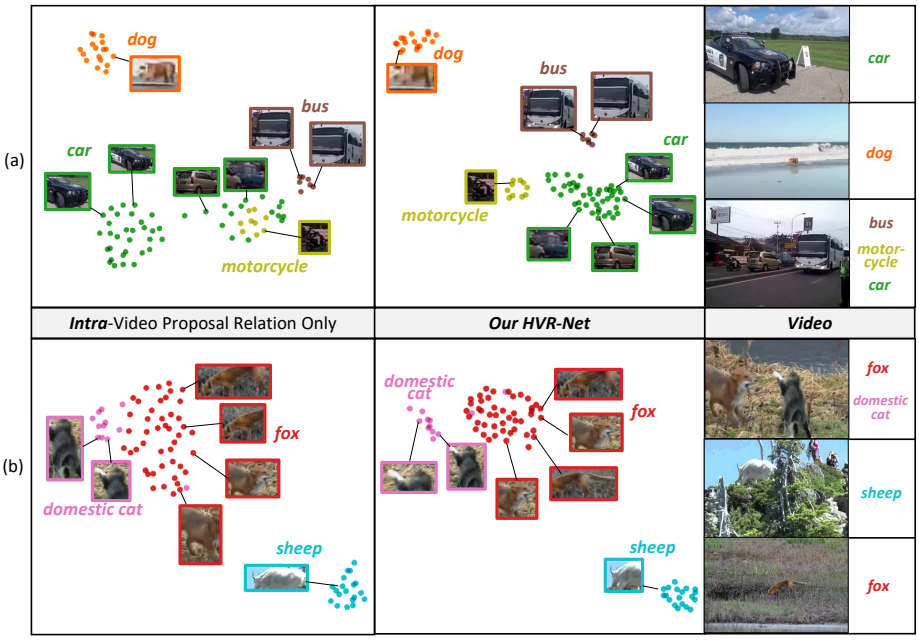

Fig. 5. Proposal Feature Visualization of Video triplet by t-SNE. With intra-video relation only, proposals of confusing objects mistakenly stay together as a cluster (i.e. domestic cats and foxes in (b), cars and motobikes in (a)). Our HVR-Net can learn the discriminative cues and clarify those proposals of confusing objects. For each video triplet, three target frames and their proposals are shown.

and how it moves in this video. For the video in Subplot (a), the appearance and motion of this lion are quite similar to a horse, leading to high confusion. Alternatively, when we introducing inter-video proposal relation module, HVR-Net successfully distinguish such object confusion in videos. Hence, it is necessary and important to learn inter-video proposal relations for video object detection.

Video and Proposal Feature Visualization in HVR-Net We visualize the proposal features of target frames in video triplets with t-SNE in Fig. 5. As expected, with inter-video proposal relation integrated, the proposal features of confusing objects can be clarified, while baseline, with intra-video proposal relations only, mistakenly clusters the proposals not belong to same category, e.g., in Fig. 5 (b), proposals of domestic cat mistakenly stay with proposals of fox together as a cluster, while our HVR-Net can learn a compact cluster (e.g., proposals of fox) and assign proposals of domestic cat correctly. The reason is that the object confusion is clarified with inter-video proposal relation integrated, leading to enlarged difference of confused proposals in feature embedding.

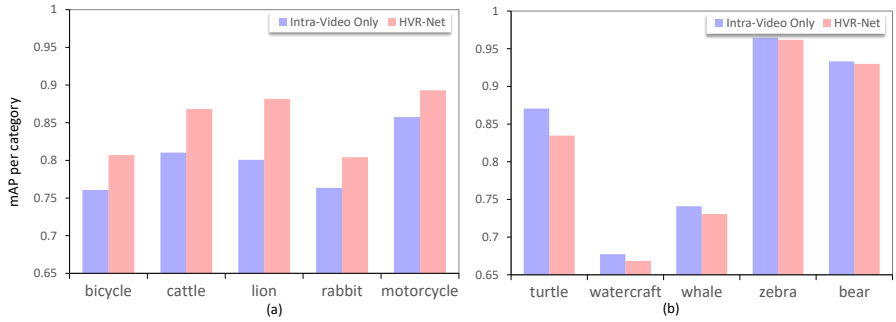

Fig. 6. Comparison of mAP per Category. Top-5 improved most categories and top-5 declined most categories are shown in subplot (a) and (b) separately. For each category, mAP is shown for baseline with only intra-video proposal relation module and our HVR-Net.

Performance Analysis on Object Categories We show the accuracy (mAP) comparison of 10 categories with our HVR-Net and baseline with intra-video proposal relation only. Top-5 improved most categories and top-5 declined most categories are shown in Fig. 6. The proposed inter-video proposal relation module boosts performance a large margin in cattle, rabbit, lion and other mammal categories. The reason is that objects in those categories usually share similar motion and appearance characteristics. With the inter-video proposal relation integrated, the object confusion is clarified, as illustrated in Fig. 4.

5 Conclusion

In this work, we propose to learn inter-video object relations for video object detection. Based on a flexible multi-level triplet selection scheme, we develop a Hierachical Video Relation Network (HVR-Net), which can effectively leverage intra-video and inter-video relation in a unified manner, in order to progressively tackle object confusions in videos. We perform extensive experiments on the large-scale video object detection benchmark, i.e., ImageNet VID. The results show that our HVR-Net is effective and important for video object detection.

Acknowledgement

This work is partially supported by Science and Technology Service Network Initiative of Chinese Academy of Sciences (KFJ-STS-QYZX-092), Guangdong Special Support Program (2016TX03X276), Shenzhen Basic Research Program (CXB201104220032A), National Natural Science Foundation of China (61876176, U1713208), the Joint Lab of CAS-HK. This work is also partially supported by Australian Research Council Discovery Early Career Award (DE190100626). The corresponding author is Yu Qiao.

References

1. Bertasius, G., Torresani, L., Shi, J.: Object detection in video with spatiotemporal sampling networks. In: ECCV. pp. 331–346 (2018)
2. Chen, K., Wang, J., Yang, S., Zhang, X., Xiong, Y., Change Loy, C., Lin, D.: Optimizing video object detection via a scale-time lattice. In: CVPR. pp. 7814–7823 (2018)
3. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. In: Advances in neural information processing systems. pp. 379–387 (2016)
4. Deng, H., Hua, Y., Song, T., Zhang, Z., Xue, Z., Ma, R., Robertson, N., Guan, H.: Object guided external memory network for video object detection. In: ICCV. pp. 6678–6687 (2019)
5. Deng, J., Pan, Y., Yao, T., Zhou, W., Li, H., Mei, T.: Relation distillation networks for video object detection. In: ICCV (2019)
6. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: ICCV. pp. 6569–6578 (2019)
7. Feichtenhofer, C., Pinz, A., Zisserman, A.: Detect to track and track to detect. In: ICCV. pp. 3038–3046 (2017)
8. Girshick, R.: Fast r-cnn. In: ICCV. pp. 1440–1448 (2015)
9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. pp. 580–587 (2014)
10. Guo, C., Fan, B., Gu, J., Zhang, Q., Xiang, S., Prinet, V., Pan, C.: Progressive sparse local attention for video object detection. In: ICCV. pp. 3909–3918 (2019)
11. 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)
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: CVPR. pp. 2961–2969 (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
14. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: CVPR. pp. 3588–3597 (2018)
15. Jiang, Z., Gao, P., Guo, C., Zhang, Q., Xiang, S., Pan, C.: Video object detection with locally-weighted deformable neighbors. In: AAAI. vol. 33, pp. 8529–8536 (2019)
16. Kang, K., Li, H., Xiao, T., Ouyang, W., Yan, J., Liu, X., Wang, X.: Object detection in videos with tubelet proposal networks. In: CVPR. pp. 727–735 (2017)
17. 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. TCSVT **28**(10), 2896–2907 (2017)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
19. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: ECCV. pp. 734–750 (2018)
20. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)
21. 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: ECCV. pp. 740–755. Springer (2014)

22. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV. pp. 21–37. Springer (2016)
23. Ouyang, W., Luo, P., Zeng, X., Qiu, S., Tian, Y., Li, H., Yang, S., Wang, Z., Xiong, Y., Qian, C., et al.: Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection. arXiv preprint arXiv:1409.3505 (2014)
24. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. pp. 779–788 (2016)
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: IEEE-TPAMI (2017)
27. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
28. Shvets, M., Liu, W., Berg, A.C.: Leveraging long-range temporal relationships between proposals for video object detection. In: ICCV (2019)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
30. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR. pp. 1–9 (2015)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NIPS. pp. 5998–6008 (2017)
32. Wang, S., Zhou, Y., Yan, J., Deng, Z.: Fully motion-aware network for video object detection. In: ECCV. pp. 542–557 (2018)
33. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR. pp. 7794–7803 (2018)
34. Wu, H., Chen, Y., Wang, N., Zhang, Z.: Sequence level semantics aggregation for video object detection. In: ICCV (2019)
35. Xiao, F., Jae Lee, Y.: Video object detection with an aligned spatial-temporal memory. In: ECCV. pp. 485–501 (2018)
36. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR. pp. 1492–1500 (2017)
37. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Craft objects from images. In: CVPR. pp. 6043–6051 (2016)
38. Zhou, X., Zhuo, J., Krahenbuhl, P.: Bottom-up object detection by grouping extreme and center points. In: CVPR. pp. 850–859 (2019)
39. Zhu, C., He, Y., Savvides, M.: Feature selective anchor-free module for single-shot object detection. In: CVPR. pp. 840–849 (2019)
40. Zhu, X., Dai, J., Yuan, L., Wei, Y.: Towards high performance video object detection. In: CVPR. pp. 7210–7218 (2018)
41. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. In: ICCV. pp. 408–417 (2017)