Relation Distillation Networks for Video Object Detection*

Jiajun Deng†, Yingwei Pan‡, Ting Yao‡, Wengang Zhou†, Houqiang Li†, and Tao Mei‡
† CAS Key Laboratory of GIPAS, University of Science and Technology of China, Hefei, China
‡JD AI Research, Beijing, China
{djiajun1206, panyw.ustc, tingyao.ustc}@gmail.com, {zhwg, lihq}@ustc.edu.cn, tmei@jd.com

Abstract

It has been well recognized that modeling object-to-object relations would be helpful for object detection. Nevertheless, the problem is not trivial especially when exploring the interactions between objects to boost video object detectors. The difficulty originates from the aspect that reliable object relations in a video should depend on not only the objects in the present frame but also all the supportive objects extracted over a long range span of the video. In this paper, we introduce a new design to capture the interactions across the objects in spatio-temporal context. Specifically, we present Relation Distillation Networks (RDN) — a new architecture that novelly aggregates and propagates object relation to augment object features for detection. Technically, object proposals are first generated via Region Proposal Networks (RPN). RDN then, on one hand, models object relation via multi-stage reasoning, and on the other, progressively distills relation through refining supportive object proposals with high objectness scores in a cascaded manner. The learnt relation verifies the efficacy on both improving object detection in each frame and box linking across frames. Extensive experiments are conducted on ImageNet VID dataset, and superior results are reported when comparing to state-of-the-art methods. More remarkably, our RDN achieves 81.8% and 83.2% mAP with ResNet-101 and ResNeXt-101, respectively. When further equipped with linking and rescoring, we obtain to-date the best reported mAP of 83.8% and 84.7%.

1. Introduction

The advances in Convolutional Neural Networks (CNN) have successfully pushed the limits and improved the state-of-the-art technologies of image and video understanding [16, 18, 19, 22, 24, 25, 35, 34, 37, 42, 43, 44]. As one of the most fundamental tasks, object detection in still images has attracted a surge of research interests and the recent methods [3, 5, 10, 14, 39] mostly proceed along the region-based detection paradigm which is derived from the work of R-CNN [11]. In a further step to localize and recognize objects in videos, video object detection explores spatio-temporal coherence to boost detectors generally through two directions of box-level association [8, 13, 20, 21] and feature aggregation [46, 49, 53, 54]. The former delves into the association across bounding boxes from consecutive frames to generate tubelets. The latter improves per-frame features by aggregation of nearby features. Regardless of these different recipes for enhancing video object detection, a common issue not fully studied is the exploitation of object relation, which is well believed to be helpful for detection.

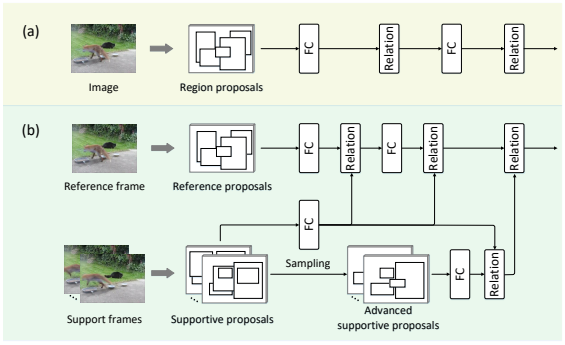

Figure 1. Modeling object relations by employing (a) stacked relation within an image and (b) distillation in an cascade manner across video frames.

Object relations characterize the interactions or geometric positions between objects. In the literature, there has been strong evidences on the use of object relation to support various vision tasks, e.g., recognition [48], object detection [17], cross-domain detection [2], and image captioning [52]. One representative work that employs object relation is [17] for object detection in images. The basic idea is to measure relation features of one object as the weighted sum of appearance features from other objects in the image and the weights reflect object dependency in terms of both appearance and geometry information. A stacked relation module as shown in Figure 1(a) aggregates relation features and augments the object features in a multi-step fashion. The method verifies the merit on modeling object

*This work was performed at JD AI Research.

relation to eventually enhance image object detection. Nevertheless, the extension of mining object relation in an image to in a video is very challenging due to the complex spatio-temporal context. Both the objects in the reference frame and all the supportive objects extracted from nearby frames should be taken into account. This distinction leads to a huge rise in computational cost and memory demand if directly capitalizing on the measure of object relation in [17], not to mention that the increase of supportive object proposals results in more invalid proposals, which may affect the overall stability of relation learning. To alleviate these issues, we propose a new multi-stage module as illustrated in Figure 1(b). Our unique design is to progressively schedule relation distillation. We select object proposals with high objectness scores from all support frames and only augment the features of these proposals with object relation to further distill the relation with respect to proposals in reference frame. Such cascaded means, on one hand, could reduce computation and filter out invalid proposals, and on the other, refine object relation better.

By consolidating the idea of modeling object relation in spatio-temporal context, we novelly present Relation Distillation Networks (RDN) for boosting video object detection. Specifically, Region Proposal Network (RPN) is exploited to produce object proposals from the reference frame and all the support frames. The object proposals extracted from support frames are packed into supportive pool. The goal of our RDN is to augment the feature of each object proposal in the reference frame by aggregating its relation features over the proposals in the supportive pool. RDN employs multi-stage reasoning structure, which includes basic stage and advanced stage. In the basic stage, RDN capitalizes on all the proposals in the supportive pool to measure relation features measured on both appearance and geometry information. The interactions are explored holistically across all the supportive proposals in this stage irrespective of the validity of proposals. Instead, RDN in the advanced stage nicely selects supportive proposals with high objectness scores and first endows the features of these proposals with relation against all the supportive proposals. Such aggregated features then in turn strengthen the relation distillation with respect to proposals in the reference frame. The upgraded feature of each proposal with object relation is finally exploited for proposal classification and regression. Moreover, the learnt relation also benefit the post-processing of box linking. Note that our RDN is applicable in any region-based vision tasks.

2. Related Work

Object Detection. The recent advances in deep convolutional neural networks [16, 22, 43, 44] and well-annotated datasets [28, 40] have inspired the remarkable improvements of image object detection [5, 10, 11, 14, 15, 23, 26, 27, 30, 38, 39, 41]. There are generally two directions for object detection. One is proposal-based two-stage detectors (e.g., R-CNN [11], Fast R-CNN [10], and Faster R-CNN [39]), and the other is proposal-free one-stage detectors (e.g., SSD [30], YOLO [38], and RetinaNet [27]). Recently, motivated by the success of attention model in NLP field [9, 45], [17, 47] extend attention mechanisms to support computer vision tasks by exploiting the attention/relations among regions/CNN features. In particular, [17] presents an object relation module that models the relations of region proposals through the interaction between their appearance features and coordinates information. [47] plugs non-local operation into conventional CNN to enable the relational interactions within CNN feature maps, aiming to capture contextual information and eventually boost both object detection and video classification tasks.

The Relation Distillation Networks in our work is also a type of relation modeling among objects. Unlike [17] that is developed for object detection in images, ours goes beyond the mining of object relation within one image and aims to explore the object interactions across multiple frames in the complex spatio-temporal context of video object detection. Moreover, a progressive schedule of relation distillation is devised to refine object relations and meanwhile reduce the computational cost on measuring object relations between reference frame and all nearby support frames.

Video Object Detection. Generalizing still image detectors to video domain is not trivial due to the spatial and temporal complex variations existed in videos, not to mention that the object appearances in some frames may be deteriorated by motion blur or occlusion. One common solution to amend this problem is feature aggregation [1, 29, 49, 53, 54, 55] that enhances per-frame features by aggregating the features of nearby frames. Specifically, FGFA [54] utilizes the optical flow from FlowNet [7] to guide the pixel-level motion compensation on feature maps of adjacent frames for feature aggregation. [49] devises a spatio-temporal memory module to perform frame-by-frame spatial alignment for aggregation. Another direction of video object detection is box-level association [8, 13, 20, 21, 46] which associates bounding boxes from consecutive frames to generate tubelets via independent processes of linking/tracking. Seq-NMS [13] builds temporal graph according to jaccard overlap between bounding boxes of consecutive frames and searches the optimal path with highest confidence as tubelets. D&T [8] integrates a tracking formulation into R-FCN [5] to simultaneously perform object detection and across-frame track regression. [46] further extends FGFA [54] by calibrating the object features on box level to boost video object detection.

Despite both feature-level and box-level methods have generally enhanced video object detection with higher quantitative scores, the object relations are not fully ex-

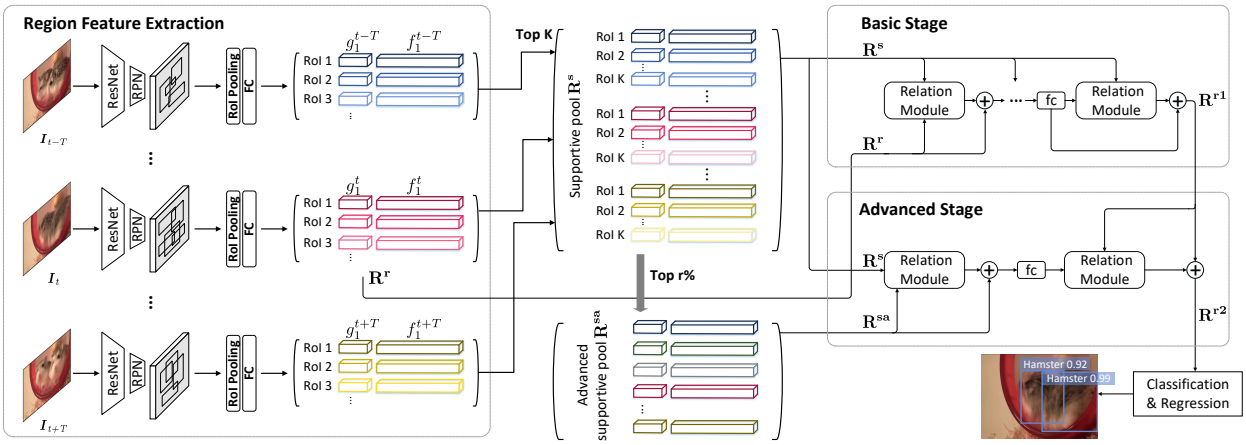

Figure 2. An overview of Relation Distillation Networks (RDN) for video object detection. Given the input reference frame I_t and all support frames $\{I_\tau\}_{\tau=t-T}^{t+T}$, Region Proposal Networks (RPN) is first employed to produce object proposals (*i.e.*, Region of Interests (RoI)) from reference frame and all support frames. We select the top-K object proposals from reference frame as the reference object set $\mathbf{R^r}$ and pack all the top-K object proposals from support frames into the supportive pool $\mathbf{R^s}$. After that, RDN is devised to augment the feature of each reference proposal in $\mathbf{R^r}$ by aggregating its relation features over the supportive proposals in $\mathbf{R^s}$, enabling the modeling of object relations in spatio-temporal context. Specifically, RDN is a multi-stage module which contains *basic stage* and *advanced stage* to support multi-stage reasoning and relation distillation. In basic stage, all supportive proposals in $\mathbf{R^s}$ are leveraged to measure the relation feature of each reference proposal in $\mathbf{R^r}$ via exploring the interactions across all the supportive proposals, outputting a set of refined reference proposals $\mathbf{R^{r1}}$. In the advanced stage, we first select $r\%$ supportive proposals in $\mathbf{R^s}$ with high objectness scores to form the advanced supportive pool $\mathbf{R^{sa}}$, where the feature of each supportive proposal is endowed with relation against all the supportive proposals. Such aggregated features then in turn strengthen the relation distillation with respect to proposals in $\mathbf{R^{r1}}$ from basic stage. Finally, the upgraded features of all reference proposals ($\mathbf{R^{r2}}$) output from advanced stage is exploited for proposal classification and regression.

ploited across frames for object detection in videos. In contrast, we exploit the modeling of object relations in spatio-temporal context to facilitate video object detection. To this end, we design a novel Relation Distillation Networks to aggregate and propagate object relation across frames to augment object features in a cascaded manner for detection.

3. RDN for Video Object Detection

In this paper, we devise Relation Distillation Networks (RDN) to facilitate object detection in videos by capturing the interactions across objects in spatio-temporal context. Specifically, Region Proposal Networks (RPN) is first exploited to obtain the object proposals from the reference frame and all the support frames. RDN then aggregates and propagates object relation over the supportive proposals to augment the feature of each reference object proposal for detection. A multi-stage module is employed in RDN to simultaneously model object relation via multi-stage reasoning and progressively distill relation through refining supportive object proposals with high objectness scores in a cascaded manner. The learnt relation can be further exploited in both classification & regression for detection and the detection box linking in post-processing. An overview of our RDN architecture is depicted in Figure 2.

3.1. Overview

Notation. In the standard task of video object detection, we are given a sequence of adjacent frames $\{I_\tau\}_{\tau=t-T}^{t+T}$ where the central frame I_t is set as the reference frame. The whole sequence of adjacent frames $\{I_\tau\}_{\tau=t-T}^{t+T}$ is taken as support frames and T represents the temporal spanning range of support frames. As such, the goal of video object detection is to detect objects in reference frame I_t by additionally exploiting the spatio-temporal correlations in the support frames. Since the ultimate goal is to model object relation in spatio-temporal context to boost video object detection, RPN is first leveraged to generate object proposals of reference frame and all support frames. The set of selected top-K object proposals from reference frame is denoted as $\mathbf{R^r} = \{R_i^r\}$. All the top-K object proposals from support frames are grouped as the supportive pool, denoted as $\mathbf{R^s} = \{R_i^s\}$. In addition, we further refine the supportive pool $\mathbf{R^s}$ by sampling $r\%$ supportive object proposals with high objectness scores, leading to the advanced supportive pool $\mathbf{R^{sa}} = \{R_i^{sa}\}$. Both of the supportive pool $\mathbf{R^s}$ and advanced supportive pool $\mathbf{R^{sa}}$ will be utilized in our devised Relation Distillation Networks to enable the progressive scheduling of relation distillation.

Problem Formulation. Inspired by the recent success of exploring object relations in various vision tasks (e.g., recognition [48] and object detection [17]), we formulate our video object detection method by modeling interactions between objects in spatio-temporal context to boost video object detectors. Given the set of reference proposals $\mathbf{R^r}$, the supportive pool $\mathbf{R^s}$ and the advanced supportive pool $\mathbf{R^{sa}}$, we are interested to progressively augment the fea-

ture of each reference proposal in $\mathbf{R^r}$ with distilled relations against supportive proposals in $\mathbf{R^s}$ and $\mathbf{R^{sa}}$. To do this, a novel Relation Distillation Networks is built based on the seminal detector Faster R-CNN [39]. A multi-stage reasoning structure consisting of basic and advanced stages is adopted in RDN for progressively scheduling relation distillation in a cascaded manner. Such design of cascaded means not only reduces computation and filters out invalid proposals, but also progressively refines object relations of reference proposals against supportive ones to boost detection. Most specifically, in the basic stage, all supportive proposals in $\mathbf{R^s}$ are utilized to measure relation features of reference proposals in $\mathbf{R^r}$ on both appearance and geometry information. As such, the output set of refined reference proposals $\mathbf{R^{r1}} = \{R_i^{r1}\}$ from basic stage is obtained via a stacked relation module which explores the interactions between reference proposals and all supportive proposals irrespective of the validity of proposals. In the advanced stage, we first enhance the feature of each selected supportive proposal in the advanced supportive pool $\mathbf{R^{sa}}$ with relation against all the supportive proposals in $\mathbf{R^s}$. Such aggregated features of distilled supportive proposals then in turn strengthen the relation distillation with respect to reference proposals in $\mathbf{R^{r1}}$ output from basic stage. Once the upgraded reference proposals $\mathbf{R^{r2}} = \{R_i^{r2}\}$ from advance stage are obtained, we directly exploit them to improve object detection in reference frame. More details about the multi-stage reasoning structure of our RDN will be elaborated in Section 3.3. Moreover, by characterizing the natural interactions between objects across frames, the learnt relations can be further leveraged to guide detection box linking in post-processing, which will be presented in Section 3.4.

3.2. Object Relation Module

We begin by briefly reviewing object relation module [17] for object detection in images. Motivated from Multi-Head Attention in [45], given the input of proposals $\mathbf{R} = \{R_i\}$, object relation module is devised to enhance each proposal R_i by measuring M relation features as the weighted sum of appearance features from other proposals. Note that we represent each object proposal with its geometric feature g_i (*i.e.*, the 4-dimensional coordinates of object proposal) and appearance feature f_i (*i.e.*, the RoI pooled feature of object proposal). Formally, the m-th relation feature of proposal R_i is calculated conditioning on $\mathbf{R}$:

$$f_{rela}^m(R_i, \mathbf{R}) = \sum\nolimits_j \omega_{ij}^m \cdot (W_L^m \cdot f_j), \quad m = 1, \cdots, M, \quad (1)$$

where W_L^m denotes the transformation matrix. ω_{ij} is an element in relation weight matrix ω and represents the pairwise relation between proposals R_i and R_j which is measured based on their appearance and geometric features. By concatenating all the M relation features of each proposal R_i and its appearance feature, we finally obtain the relation-augmented feature output from object relation module:

$$f_{rm}(R_i, \mathbf{R}) = f_i + concat[\{f_{rela}^m(R_i, \mathbf{R})\}_{m=1}^M]. \quad (2)$$

3.3. Relation Distillation Networks

Unlike [17] that explores object relations within an image for object detection, we facilitate the modeling of object relations in video object detection by exploiting the object interactions across multiple frames under the complex spatio-temporal context. One natural way to extend the relation-augmented detector in image to video is to capitalize on the object relation module in [17] to measure the interactions between the objects in reference frame and all supportive objects from nearby frames. Nevertheless, such way will lead to a huge rise in computational cost, not to mention that the increase of supportive proposals results in more invalid proposals and the overall stability of relation learning will be inevitably affected. To alleviate this issue, we devise Relation Distillation Networks to progressively schedule relation distillation for enhancing detection via a multi-stage reasoning structure, which contains basic stage and advanced stage. The spirit behind follows the philosophy that basic stage explores relations holistically across all the supportive proposals with respect to reference proposals, and advanced stage progressively distills relations via refining supportive proposals, which are augmented with relations to further strengthen reference proposals.

Basic Stage. Formally, given the set of reference proposals $\mathbf{R^r}$ and the supportive pool $\mathbf{R^s}$, the basic stage predicts the relation features of each reference proposal as the weighted sum of features from all supportive proposals via a stacked relation module:

$$\mathbf{R^{r1}} = \mathcal{N}_{basic}(\mathbf{R^r}, \mathbf{R^s}), \quad (3)$$

where $\mathcal{N}_{basic}(\cdot)$ represents the function of the stacked relation module in basic stage and $\mathbf{R^{r1}}$ denotes the output enhanced reference proposals from basic stage. Please note that in the complex spatio-temporal context of video object detection, a single relation module is insufficient to model the interactions between objects among multiple frames. Therefore, we iterate the relation reasoning in a stacked manner equipped with N_b object relation modules to better characterize the relations across all the supportive proposals with regard to reference proposals. Specifically, for the k-th object relation module in basic stage, the i-th reference proposal is augmented with the relation features over all proposals in supportive pool $\mathbf{R^s}$:

$$R_i^{r1,k} = \begin{cases} f_{rm}(R_i^r, \mathbf{R^s}), & k = 1, \\ f_{rm}(h(R_i^{r1,k-1}), \mathbf{R^s}), & k > 1, \end{cases} \quad (4)$$

where $h(\cdot)$ denotes the feature transformation function implemented with a fully-connected layer plus ReLU. Each

Algorithm 1 Inference Algorithm of our RDN

Input: video frames $\{I_t\}$, temporal spanning range T.
for $t = 1$ **to** $T + 1$ **do** ▷ initialize proposal feature buffer
  $\mathbf{R}_t = \mathcal{N}_{RoI}(I_t)$ ▷ region proposal and feature extraction
  $\mathbf{R}_t^{\mathbf{s}} = \mathbf{Sample}_{top\text{-}K}(\mathbf{R}_t)$ ▷ sample top-K proposals
end for
for $t = 1$ **to** ∞ **do**
  $\mathbf{R}^{\mathbf{r}} = \mathbf{R}_t$ ▷ reference proposal set
  $\mathbf{R}^{\mathbf{s}} = \mathbf{R}^{\mathbf{s}}_{max(1,t-T)} \cup \cdots \cup \mathbf{R}^{\mathbf{s}}_{t+T}$ ▷ supportive pool
  $\mathbf{R}^{\mathbf{r1}} = \mathcal{N}_{basic}(\mathbf{R}^{\mathbf{r}}, \mathbf{R}^{\mathbf{s}})$ ▷ basic stage
  $\mathbf{R}^{\mathbf{sa}} = \mathbf{Sample}_{top\text{-}r\%}(\mathbf{R}^{\mathbf{s}})$ ▷ sample top-$r\%$ proposals
  $\mathbf{R}^{\mathbf{r2}} = \mathcal{N}_{adv}(\mathbf{R}^{\mathbf{r1}}, \mathbf{R}^{\mathbf{s}}, \mathbf{R}^{\mathbf{sa}})$ ▷ advanced stage
  $\mathbf{D}_t = \mathcal{N}_{det}(\mathbf{R}^{\mathbf{r2}})$ ▷ classification and regression
  $\mathbf{R}_{t+T+1} = \mathcal{N}_{RoI}(I_{t+T+1})$
  $\mathbf{R}^{\mathbf{s}}_{t+T+1} = \mathbf{Sample}_{top\text{-}K}(\mathbf{R}_{t+T+1})$
  update proposal feature buffer
end for
Output: detection results $\{\mathbf{D}_t\}$

relation module takes the transformed features of reference proposals from previous relation module as the reference inputs. We stack N_b relation modules in basic stage and all the enhanced reference proposals from the N_b-th relation module are taken as the output $\mathbf{R^{r1}}$ of basic stage.

Advanced Stage. The relation reasoning in basic stage only explores the modeling of interactions between reference proposal and all the supportive proposals, while leaving the relations among supportive proposals in $\mathbf{R^s}$ unexploited. Furthermore, we present a novel advanced stage to explore the interactions between supportive proposals by enhancing the distilled supportive proposals with relations against all supportive proposals. Next, the enhanced distilled supportive proposals are utilized to further strengthen the reference proposals from basic stage via relation reasoning in between. Such design of progressively distilling supportive proposals in advanced stage not only reduces the computation cost of measuring relations among supportive proposals, but also filters out invalid supportive proposals for relation reasoning and eventually improves detection.

Technically, given the output reference proposals $\mathbf{R^{r1}}$ from basic stage, the supportive pool $\mathbf{R^s}$, and the advanced supportive pool $\mathbf{R^{sa}}$, the advanced stage further strengthens all reference proposals $\mathbf{R^{r1}}$ through progressively scheduling relation distillation:

$$\mathbf{R^{r2}} = \mathcal{N}_{adv}(\mathbf{R^{r1}}, \mathbf{R^s}, \mathbf{R^{sa}}), \tag{5}$$

where $\mathcal{N}_{adv}(\cdot)$ denotes the operation in advanced stage and $\mathbf{R^{r2}}$ represents the output relation-augmented reference proposals from advanced stage. Most specifically, we first refine the distilled supportive proposals in $\mathbf{R^{sa}}$ with relation reasoning against all supportive proposals in $\mathbf{R^s}$:

$$R_i^a = f_{rm}(R_i^{sa}, \mathbf{R^s}), \tag{6}$$

where R_i^a denotes the i-th refined supportive proposal. After that, the refined supportive proposals $\mathbf{R^a} = \{R_i^a\}$ are utilized to further distill the relation with respect to reference proposals $\mathbf{R^{r1}}$ from basic stage:

$$R_i^{r2} = f_{rm}(R_i^{r1}, \mathbf{R^a}), \tag{7}$$

where R_i^{r2} denotes the i-th upgraded reference proposal. Finally, all the upgraded reference proposals $\mathbf{R^{r2}} = \{R_i^{r2}\}$ are exploited for proposal classification and regression.

Training and Inference. At training stage, we adopt the strategy of temporal dropout [54] to randomly select two support frames $I_{t+\tau_1}$ and $I_{t+\tau_2}$ ($\tau_1, \tau_2 \in [-T, T]$) from the adjacent frames $\{I_\tau\}_{\tau=t-T}^{t+T}$. Accordingly, the whole RDN is optimized with both classification and regression losses over the relation-augmented reference proposals $\mathbf{R^{r2}}$ from advanced stage in an end-to-end manner.

During inference, we follow [54] and sequentially process each frame with a sliding proposal feature buffer of adjacent frames $\{I_\tau\}_{\tau=t-T}^{t+T}$. The capacity of this proposal feature buffer is set as the length of adjacent frames (*i.e.*, $2T+1$), except for the beginning and ending T frames. The detailed inference process of RDN is given in Algorithm 1.

3.4. Box Linking with Relations

To further boost video object detection results by re-scoring individual detection boxes among consecutive frames, we adopt the post-processing of linking detection boxes across frames as in [12, 13, 21]. Despite the box-level post-processing methods have generally enhanced video object detection with higher quantitative scores, the object relations between detection boxes are not fully studied for box linking. In contrast, we integrate the learnt object-to-object relations into post-processing of box linking to further propagate the confidence scores among high-related detection boxes and thus improve the detection.

Specifically, we formulate the post-processing of box linking as an optimal path finding problem. Note that since the box linking is independently applied for each class, we omit the notation of class here for simplicity. Given two detection boxes d_i^t and d_j^{t+1} from consecutive frames I_t and I_{t+1}, the linking score between them is calculated as:

$$S(d_i^t, d_j^{t+1}) = \{s_i^t + s_j^{t+1} + iou(d_i^t, d_j^{t+1})\} \cdot e^{\bar{\omega}_{ij}}, \tag{8}$$

where s_i^t and s_j^{t+1} are confidence scores of the two boxes, and $iou(\cdot)$ indicates jaccard overlap. $\bar{\omega}_{ij}$ represents the pairwise relation weight between the two boxes d_i^t and d_j^{t+1}, which is measured as the average of all the M relation weights obtained in the last relation module at basic stage: $\bar{\omega}_{ij} = \frac{1}{M}\sum_{m=1}^{M} \omega_{ij}^m$. Accordingly, for each class, we seek the optimal path as:

$$\bar{P}^* = \arg\max_{\bar{P}} \frac{1}{\mathcal{L}} \sum_{t=1}^{\mathcal{L}-1} S(\mathbf{D}_t, \mathbf{D}_{t+1}), \tag{9}$$

where $\mathbf{D}_t = \{d_i^t\}$ denotes the set of detection boxes in frame I_t and $\mathcal{L}$ is the duration of video. This problem can be solved by Viterbi algorithm [12]. Once the optimal path for linking boxes is obtained, we follow [8] and re-score detection boxes in each tube by adding the average value of the top-50% classification score of boxes in this path.

4. Network Architecture

Backbone. We exploit two kinds of backbones, *i.e.*, ResNet-101 [16] and ResNeXt-101-64×4d [50], for our RDN. Specifically, to enlarge the resolution of feature maps, we modify the stride of first conv block in last stage of convolutional layers from 2 to 1. As such, the effective stride in this stage is changed from 32 to 16 pixels. Besides, all the 3×3 conv layers in this stage are modified by the "hole algorithm" [4, 32] (*i.e.*, "atrous convolution" [31]) to compensate for the receptive fields.

Region Feature Extraction. We utilize RPN [39] on the top of conv4 stage for region feature extraction. In particular, we leverage 12 anchors with 4 scales $\{64^2, 128^2, 256^2, 512^2\}$ and 3 aspect ratios $\{1{:}2, 1{:}1, 2{:}1\}$ for classification and regression. During training and inference, we first pick up 6,000 proposals with highest objectness scores and then adopt Non Maximum Suppression (NMS) with threshold of 0.7 Intersection-over-Union (IoU) to obtain $N = 300$ proposals for each frame. After generating region proposals, we apply RoI pooling followed by a 1,024-d fully-connected layer on the top of conv5 stage to extract RoI feature of each proposal.

Relation Distillation Networks. For each relation module in RDN, the number of relation features is set as $M = 16$. The dimension of each relation feature is 64. As such, by concatenating all the $M = 16$ relation features as in Equation 2, the dimension of relation-augmented feature output from relation module is 1,024. In basic stage, we stack $N_b = 2$ relation modules. In advanced stage, one relation module is first employed to enhance proposals in advanced supportive pool $\mathbf{R^{sa}}$. Next we apply another relation module to strengthen the reference proposals output from basic stage. Finally, we utilize two parallel branches (*i.e.*, classification and regression) to obtain detection boxes based on the refined RoI features from advanced stage.

5. Experiments

5.1. Dataset and Evaluation

We empirically verify the merit of our RDN by conducting experiments on ImageNet object detection from video (VID) dataset [40]. The ImageNet VID dataset is a large-scale benchmark for video object detection task, consisting of 3,862 training videos and 555 validation videos from 30 classes. Given the fact that the ground truth of the official testing set are not publicly available, we follow the widely adopted setting as in [8, 20, 46, 49, 54, 53] to report mean Average Precision (mAP) on validation set.

Table 1. Performance comparison with state-of-the-art end-to-end video object detection models on ImageNet VID validation set.

Methods	Backbone	Base Detector	mAP (%)
FGFA [54]	ResNet-101	R-FCN	76.3
	ResNet-101	Faster R-CNN	77.5
MANet [46]	ResNet-101	R-FCN	78.1
THP [53]	ResNet-101 + DCN [6]	R-FCN	78.6
STSN [1]	ResNet-101 + DCN [6]	R-FCN	78.9
RDN	ResNet-101	Faster R-CNN	**81.8**
	ResNeXt-101-64×4d	Faster R-CNN	**83.2**

Following the common protocols in [8, 46, 49, 54], we utilize both ImageNet VID and ImageNet object detection (DET) dataset to train our RDN. Since the 30 classes in ImageNet VID are a subset of 200 classes in ImageNet DET dataset, we adopt the images from overlapped 30 classes in ImageNet DET for training. Specifically, due to the redundancy among adjacent frames, we sample 15 frames from each video in ImageNet VID for training. For ImageNet DET, we select at most 2,000 images from each class to make the class distribution more balanced.

5.2. Implementation Details

At training and inference stages, the temporal spanning range is set as $T = 18$. We select the top $K = 75$ proposals with highest objectness scores from each support frame and pack them into the supportive pool $\mathbf{R^s}$. We obtain the advanced supportive pool $\mathbf{R^{sa}}$ by sampling 20% supportive proposals with highest objectness scores from $\mathbf{R^s}$.

We implement RDN mainly on Pytorch 1.0 [36]. The input images are first resized so that the shorter side is 600 pixels. The whole architecture is trained on four Tesla V100 GPUs with synchronized SGD (momentum: 0.9, weight decay: 0.0001). There is one mini-batch in each GPU and each mini-batch contains one image/frame. For reference frame, we sample 128 RoIs with a ratio of 1:3 for positive:negatives. We adopt a two-phase strategy for training our RDN. In the first phase, we train the basic stage together with backbone & RPN over the combined training set of ImageNet VID and ImageNet DET for 120k iterations. The learning rate is set as 0.001 in the first 80k iterations and 0.0001 in the next 40k iterations. In the second phase, the whole RDN architecture is trained on the combined training set with another 60k iterations. The learning rate is set as 0.001 in the first 40k iterations and 0.0001 in the last 20k iterations. The whole procedure of training takes about 15 hours in the first phase and 8 hours in the second phase. At inference, we adopt NMS with a threshold of 0.5 IoU to suppress reduplicate detection boxes.

5.3. Performance Comparison

End-to-End models. The performances of different end-to-end video object detection models on ImageNet VID

Table 2. Performance comparison with state-of-the-art video object detection methods plus post-processing on ImageNet VID validation set. BLR: Box Linking with Relations in Section 3.4.

Methods	Backbone	Base Detector	mAP (%)
T-CNN [21]	DeepID [33] + Craft [51]	R-CNN	73.8
FGFA [54] + [13]	ResNet-101	R-FCN	78.4
	Aligned Inception-ResNet	R-FCN	80.1
D&T [8]	ResNet-101	R-FCN	79.8
	ResNet-101	Faster R-CNN	80.2
	Inception-v4	R-FCN	82.0
STMN [49]	ResNet-101	R-FCN	80.5
RDN + [13]	ResNet-101	Faster R-CNN	82.6
	ResNeXt-101-64×4d	Faster R-CNN	83.9
RDN + [12]	ResNet-101	Faster R-CNN	83.4
	ResNeXt-101-64×4d	Faster R-CNN	84.5
RDN + BLR	ResNet-101	Faster R-CNN	**83.8**
	ResNeXt-101-64×4d	Faster R-CNN	**84.7**

validation set are shown in Table 1. Note that for fair comparison, here we only include the state-of-the-art end-to-end techniques which purely learn video object detector by enhancing per-frame feature in an end-to-end fashion without any post-processing. Overall, the results under the same backbone demonstrate that our proposed RDN achieves better performance against state-of-the-art end-to-end models. In particular, the mAP of RDN can achieve 81.8% with ResNet-101, which makes 2.9% absolute improvement over the best competitor STSN. As expected, when equipped with a stronger backbone (ResNeXt-101-64×4d), the mAP of our RDN is further boosted up to 83.2%. By additionally capturing global motion clues to exploit instance-level calibration, MANet exhibits better performance than FGFA that performs pixel-level calibration with the guidance from optical flow. Different from the flow-guided methods (FGFA, MANet, and THP) which estimate the motion across frames for warping the feature map, STSN spatially samples features from adjacent frames for feature aggregation and achieves better performance. Nevertheless, the performance of STSN is still lower than that of our RDN which models object relation in spatio-temporal context. The results highlight the advantage of aggregating and propagating object relation to augment object features for video object detection.

Add Post-Processing. In this section, we compare our RDN with other state-of-the-art methods by further applying post-processing of box linking. Table 2 summarizes the results on ImageNet VID validation set. In general, when equipped with existing post-processing techniques (Seq-NMS and Tube Linking), our RDN exhibits better performances than other state-of-the-art post-processing based approaches. In addition, by leveraging our Box Linking with Relations (BLR) that integrates the learnt object relations into Tube Linking, the performances of RDNs are further boosted up to 83.8% and 84.7% with ResNet-101 and ResNeXt-101-64×4d, respectively. This confirms the effectiveness of propagating confidence scores among detection boxes with high relations via box linking in our BLR.

Table 3. Performance comparisons across different ways on the measure of object relation, *i.e.*, Faster R-CNN on single frame irrespective of relation, stacked relation within frame in [17], RDN with relation only in basic stage (**BASIC**), full version of RDN with advanced stage (**ADV**)). The backbone is ResNet-101.

Methods	BASIC	ADV	mAP (%)
Faster R-CNN			75.4
+ Relation [17]			$78.5_{\uparrow 3.1}$
Faster R-CNN			
+ BASIC	✓		$80.9_{\uparrow 5.5}$
RDN	✓	✓	$\mathbf{81.8}_{\uparrow \mathbf{6.4}}$

Figure 3. Examples of video object detection results by different ways of relation modeling in our RDN.

5.4. Experimental Analysis

Ablation Study. Here we study how each design in our RDN influences the overall performance. Faster R-CNN [39] simply executes object detection on single frame irrespective of object relation. [17] models relation in an image via stacked relation modules. We extend this idea to learn the interactions between objects in a video frame and re-implement [17] in our experiments. The run of Faster R-CNN + BASIC only exploits the basic stage for relation reasoning and RDN further integrates the advanced stage.

Table 3 details the performances across different ways on the measure of object relation. Directly performing Faster R-CNN on single frame achieves 75.4% of mAP. The mining of relation in [17] leads to a boost of 3.1%. The results verify the idea of exploring object relation to improve video object detection, even in case when the relation is measured within each frame. By capturing object interactions across frames in the basic stage, Faster R-CNN + BASIC boosts up the mAP from 75.4% to 80.9%. The improvements indicate learning relation in spatio-temporal context is superior to spatial dimension only. RDN is benefited from the mechanism of cascaded relation distillation in advanced stage and the mAP of RDN finally reaches 81.8%. Figure 3 showcases one example of video object detection results with different ways of relation modeling in our RDN. As illustrated in the figure, the detection results become increasingly robust as more designs of relation modeling are included.

Effect of Temporal Spanning Range T. To explore the effect of temporal spanning range T in our RDN, we show

Table 4. Performance and run time comparisons by using different temporal spanning range T in our RDN.

# T	3	6	9	12	15	18	21	24
mAP (%)	80.3	80.7	80.9	81.3	81.6	**81.8**	81.7	81.7
runtime (ms)	90.1	90.3	91.5	93.0	93.5	94.2	97.3	103.1

Table 5. Performance comparisons by using different number of relation module in basic stage.

# N_b	0	1	2	3	4
mAP (%)	75.4	79.4	**80.9**	80.8	80.4

the performance and run time by varying this number from 3 to 24 within an interval of 3 in Table 4. The best performance is attained when the temporal spanning range is set to $T = 18$. In particular, once the temporal spanning range is larger than 12, the performances are less affected with the change of the temporal spanning range, which eases the selection of the temporal spanning range in our RDN practically. Meanwhile, enlarging the temporal spanning range generally increases run time at inference. Thus, the temporal spanning range is empirically set to 18, which is a good tradeoff between performance and run time.

Effect of Relation Module Number N_b in Basic Stage. Table 5 shows the performances of employing different number of relation module in basic stage. In the extreme case of $N_b = 0$, no relation module is utilized and the model degenerates to Faster R-CNN on single frame. With the use of only one relation module, the mAP is increased from 75.4% to 79.4%. That basically validates the effectiveness of modeling relation for object detection. The mAP is further boosted up to 80.9% with the design of two modules but the performance slightly decreases when stacking more modules. We speculate that this may be the result of unnecessary information repeat from support frames and that double proves the motivation of designing the advanced stage in RDN. In practice, the number N_b is generally set to 2.

Effect of Sampling Number K in Basic Stage and Sampling Ratio $r\%$ in Advanced Stage. We firstly vary K from 25 to 300 in basic stage to explore the relationship between the performance/run time and the sampling number K. As shown in Table 6, the performances are very slightly affected with the change of sampling number K. Specifically, the best performance is attained when the sampling number K is 75. Meanwhile, the run time at inference is gradually increased when enlarging the sampling number. Therefore, we set the sampling number K to 75 practically. Next, to investigate the effect of sampling ratio $r\%$ in advanced stage, we further compare the results of performance and run time by varying the sampling ratio from 10% to 100% in Table 7. The best performance is obtained when the sampling ratio is set as 20%. Meanwhile, the performances are relatively smooth when the sampling ratio varies. That practically eases the selection of of the sampling ratio $r\%$ in advanced stage. In addition, when the sampling ratio increases, the run time is significantly increased. Accordingly, the sampling ratio is empirically set as $r = 20\%$, which seeks a better tradeoff between performance and run time.

Table 6. Performance and run time comparisons by using different sampling number K in basic stage of our RDN.

# K	25	50	75	100	150	200	250	300
mAP (%)	80.2	80.5	80.9	80.7	80.4	80.4	80.2	80.1
runtime (ms)	80.0	81.7	84.9	86.3	94.9	107.2	125.3	152.7

Table 7. Performance and run time comparisons by using different sampling ratio $r\%$ in advanced stage of our RDN.

# r (%)	10	20	30	40	50	60	80	100
mAP (%)	81.3	81.8	81.7	81.6	81.5	81.5	81.3	81.3
runtime (ms)	92.8	94.2	96.9	100.2	104.0	108.9	114.8	125.6

Complementarity of Two Stages. In RDN, basic stage augments reference proposals with relation features of supportive proposals, which enhances reference proposals with first-order relation from supportive proposals on a star-graph. Then advanced stage progressively samples supportive proposals with high objectness scores and first enhances sampled/advanced proposals with relation against all supportive proposals. In this way, the advanced supportive proposals is endowed with first-order relation from supportive proposals on a full-connected graph. Next, advanced stage strengthens reference proposals with advanced supportive proposals. As such, the reference proposals are further endowed with higher-order relation from supportive proposals, which are naturally complementary to basic stage.

6. Conclusions

We have presented Relation Distillation Networks architecture, which models object relation across frames to boost video object detection. Particularly, we study the problem from the viewpoint of employing multi-stage reasoning and scheduling relation distillation progressively. To verify this, we utilize RPN to generate object proposals in the reference and support frames. The supportive pool is comprised of all the proposals extracted from support frames. In the basic stage, RDN measures the relation of each object proposal in reference frame over all the proposals in the supportive pool and augments the features with relation. In the advanced stage, RDN self adjusts the selected supportive proposals with the relation against all the supportive ones firstly and then capitalizes on such selected proposals to distill the relation of each proposal in reference frame. Extensive experiments conducted on ImageNet VID dataset validate our proposal and analysis. More remarkably, we achieve to-date the best reported mAP of 84.7%, after post-processing of linking and rescoring.

Acknowledgements This work was supported in part by NSFC under contract No. 61836011, No. 61822208, and No. 61632019, and Youth Innovation Promotion Association CAS (No. 2018497).

References

[1] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *ECCV*, 2018.

[2] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *CVPR*, 2019.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *ICLR*, 2015.

[5] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, 2016.

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017.

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

[8] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *ICCV*, 2017.

[9] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *ICML*, 2017.

[10] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[12] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *CVPR*, 2015.

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

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. on PAMI*, 2015.

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

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

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[20] Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, and Xiaogang Wang. Object detection in videos with tubelet proposal networks. In *CVPR*, 2017.

[21] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Trans. on CSVT*, 2017.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[23] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018.

[24] Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. Recurrent tubelet proposal and recognition networks for action detection. In *ECCV*, 2018.

[25] Dong Li, Ting Yao, Ling-Yu Duan, Tao Mei, and Yong Rui. Unified spatio-temporal attention networks for action recognition in videos. *IEEE Transactions on Multimedia*, 2018.

[26] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *ICCV*, 2017.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[29] Mason Liu and Menglong Zhu. Mobile video object detection with temporally-aware feature maps. In *CVPR*, 2018.

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[32] Stéphane Mallat. *A wavelet tour of signal processing*. 1999.

[33] Wanli Ouyang, Ping Luo, Xingyu Zeng, Shi Qiu, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Yuanjun Xiong, Chen Qian, et al. Deepid-net: multi-stage and deformable deep convolutional neural networks for object detection. *arXiv:1409.3505*, 2014.

[34] Yingwei Pan, Yehao Li, Ting Yao, Tao Mei, Houqiang Li, and Yong Rui. Learning deep intrinsic video representation by exploring temporal coherence and graph structure. In *IJCAI*, 2016.

[35] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, 2016.

[36] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Workshop on Machine Learning Systems, NIPS*, 2017.

[37] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, 2017.

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

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[41] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Robert Fergus, and Yann Lecun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014.

[42] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[44] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[46] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. Fully motion-aware network for video object detection. In *ECCV*, 2018.

[47] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[48] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018.

[49] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. In *ECCV*, 2018.

[50] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

[51] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z Li. Craft objects from images. In *CVPR*, 2016.

[52] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.

[53] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. Towards high performance video object detection. In *CVPR*, 2018.

[54] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *CVPR*, 2017.

[55] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *CVPR*, 2017.