Interventional Video Relation Detection

Yicong Li

liyicong@u.nus.edu.sg
National University of
Singapore

Xun Yang*

hfutyangxun@gmail.com
National University of
Singapore

Xindi Shang

shangxin@comp.nus.edu.sg
National University of
Singapore

Tat-Seng Chua

dcscts@nus.edu.sg
National University of
Singapore

ABSTRACT

Video Visual Relation Detection (VidVRD) aims to semantically describe the dynamic interactions across visual concepts localized in a video in the form of $\langle \text{subject}, \text{predicate}, \text{object} \rangle$. It can help to mitigate the *semantic gap* between vision and language in video understanding, thus receiving increasing attention in multimedia communities. Existing efforts primarily leverage the multimodal/spatio-temporal feature fusion to augment the representation of object trajectories as well as their interactions and formulate the prediction of predicates as a multi-class classification task. Despite their effectiveness, existing models ignore the severe long-tailed bias in VidVRD datasets. As a result, the models' prediction will be easily biased towards the popular *head* predicates (e.g., *next-to* and *in-front-of*), thus leading to poor generalizability.

To fill the research gap, this paper proposes an Interventional Video Relation Detection (IVRD) approach that aims to improve not only the accuracy but also the robustness of the model prediction. Specifically, to better model the high-level visual predicate, our IVRD consists of two key components: 1) we first learn a set of predicate prototypes, where each prototype vector describes a set of relation references with the same predicate; and 2) we apply a causality-inspired intervention on the model input $\langle \text{subject}, \text{object} \rangle$, which forces the model to fairly incorporate each possible predicate prototype into consideration. We expect the model to focus more on the visual content of the dynamic interaction between subject and object, rather than the spurious correlations between the model input and predicate labels. Extensive experiments on two popular benchmark datasets show the effectiveness of IVRD and also its advantages in reducing the bad long-tailed bias.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

Video relation detection; Video understanding; Causal Intervention

ACM Reference Format:

Yicong Li, Xun Yang, Xindi Shang, and Tat-Seng Chua. 2021. Interventional Video Relation Detection. In *Proceedings of the 29th ACM International*

*Corresponding author

This work is licensed under a Creative Commons Attribution International 4.0 License.

MM '21, October 20–24, 2021, Virtual Event, China.

© 2021 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8651-7/21/10.

<https://doi.org/10.1145/3474085.3475540>

Figure 1: (a) The pipeline of the VidVRD task. Relations are separately detected in each short segment and merged afterwards. Both (b) and (c) denote the long-tailed distributions of predicates in ImageNet-VidVRD (b) and Vidor (c). Only 20 predicates are depicted to avoid cluttering

Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3474085.3475540>

1 INTRODUCTION

Deep learning techniques has boosted unprecedented progress on low-level visual tasks, e.g., the classification, detection, matching [50, 53]. However, it is still challenging for the machine to understand the visual content at a higher semantic level. Recent research efforts have focused on the detection of the relations in the visual (e.g., image and video) domain as a key step to narrowing the *semantic gap* between low-level visual content and high-level semantic description. Visual Relation Detection not only involves detecting visual concepts but also identifying the high-level interactions across visual concepts (e.g., spatial/temporal relations, actions) in the form of relation triplet $\langle \text{subject}, \text{predicate}, \text{object} \rangle$ [33, 44]. It can significantly improve the ability of cross-media reasoning of multimedia systems, thus facilitating diverse downstream multimedia tasks, e.g., QA [45], and retrieval [9, 42, 43, 48].

Although extensive efforts have been devoted to relation detection, most works just focus on image-level detection. They lack the ability of reasoning complex relations that are inherently temporal and/or correlated in nature. For example, it is ambiguous to infer from a static image whether the *dog is faster than or near* the person. Therefore, Video Visual Relation Detection (VidVRD) has emerged

as a new challenging but important task, especially after the publication of the two benchmark datasets, ImageNet-VidVRD [34] and VidOR [33]. VidVRD usually requires to first extract the object trajectories from the given video, and then predict the semantic label (e.g., *predicate*) of the dynamic interactions between the two video objects (e.g., *subject* and *object*). Compared with image-level detection, the task of VidVRD is much more challenging due to its complex temporal-spatial characteristics. Given a pair of localized video objects, there are usually more than one target relations between them, such as static spatial relations (e.g., next to, stand left) and dynamic temporal relations (e.g., move away, move toward), thus leading to a much larger search space and higher complexity. Besides, the user annotations in current VidVRD datasets are usually unbalanced, as shown in Figure 1 (b) and (c). For example, the number of spatial relation predicate *next-to* is significantly larger than that of temporal relation predicate *move-away* in Figure 1 (b), since the spatial relation is much easier to be recognized and then annotated. Overall, VidVRD is a relatively high-level task and still in the early stage of research.

Existing efforts [1, 24, 31, 34, 36, 37, 39] in VidVRD primarily leveraged the multimodal/spatio-temporal feature fusion [24, 31, 34, 39] to augment the representation of object trajectories as well as their dynamic interaction. They then formulated the prediction of the *predicate* label in a relation triplet as a multi-class classification problem. Despite their effectiveness, existing methods did not consider the severe long-tailed bias in VidVRD datasets. In fact, prior works achieved their improvement mainly on the frequently-appearing (e.g., top-20) predicates, leaving the performance of the rest untouched. The long-tailed training data forces the model to learn a shortcut to maximize the overall likelihood of training data and therefore fail to deal with the less dominant (i.e., few-shot) predicates. If a model is trained to predict *next-to* much more frequently than *walk-with*, then the former is more likely to prevail over the latter during testing. From a **causal** perspective, such a long-tailed dataset bias will inevitably lead to a bad *spurious correlation* between the model input $\langle \text{subject}, \text{object} \rangle$ and the predicate label.

To fill the research gap in VidVRD, we develop an Interventional Video Relation Detection (IVRD) approach that aims to improve not only the accuracy but also the robustness of the model prediction on predicate labels. Our IVRD is mainly inspired by the causal intervention techniques [41, 49, 52, 54, 57] in the field of vision and language, which can effectively prevent the model from utilizing the spurious correlation in biased training data. Specifically, in order to better model the high-level visual predicate, our IVRD consists of two key components: 1) we first learn a set of predicate prototypes, where each prototype vector describes a set of relation triplets with the same predicate; and 2) we apply an intervention on the model input $\langle \text{subject}, \text{object} \rangle$, which forces the model to fairly incorporate each possible predicate prototype into consideration for making a more robust prediction. We expect our model to focus more on the visual content of the dynamic subject-object interaction, thus improving the prediction on temporal relations rather than the frequently-appearing spatial relations. We conducted extensive experiments and analysis on the two popular benchmark datasets, ImageNet-VidVRD [34] and VidOR [33]. Extensive evaluation results show that our IVRD can effectively reduce the long-tailed bias

and improve the performance on the tail categories, thus indicating the effectiveness of the intervention strategy in our IVRD. We also observed substantial improvements over state-of-the-art methods.

In summary, our work makes the following main contributions:

- We are the first to address the long-tailed dataset bias in the task of video relation detection.
- We develop an IVRD method for improving the task of VidVRD. It applies a causal intervention on the model input to prevent the model from relying on the bad spurious correlations.
- We conduct extensive experiments on two public VidVRD datasets that demonstrate the effectiveness and advantages of IVRD.

2 RELATED WORK

Visual Relation Detection. Visual relation detection has been intensively studied since [21] first brought up this notion in static images (ImgVRD). Understanding this diversity of relationships in images is central to accurate visual retrieval [2, 3, 5, 12, 59] and to a richer semantic understanding of our visual world. Ever since then, a huge proportion of works were dedicated to better feature extracting module [6, 14, 22, 46]. [55] is the first work that brings the unbalance problem onto the table, and a following work [4] proposes an unbiased metric for comprehensive evaluation. [23] also attacked the tail part of the dataset by pruning those dominant and easy-to-predict relations in the training set.

Compare to ImgVRD, VidVRD is more difficult because the temporal nature of video has enriched objects with more diverse combinations, such as relations involving actions (e.g. *chase*) or position changes (e.g. *run-toward*). This poses a greater challenge for the model to incorporate both the statistical dependency and the visual information provided in videos. It also has a great potential for enhancing the downstream video tasks [11, 16–18, 44, 51]. VidVRD has not received sufficient attention until the release of ImageNet-VidVRD dataset [34] and VidOR dataset [33], the former also proposed a three-stage detection framework that adopted by most following works. Qian et al. [31] built a spatio-temporal graph convolutional networks within adjacent video segments to refine the object and relation features. Sun et al. [37] utilized language context feature along with spatial-temporal feature for predicate prediction. Cao et al. [1] proposed a 3DRN to fuse the spatio-temporal visual characteristics, object label features, and spatial interactive features for learning the relation instances with multi-modal cues. In order to capture relations involving long motions, Liu et al. [24] proposed a sliding-window scheme to predict short-term and long-term relations simultaneously. Our IVRD is the first to introduce the notion of causal intervention in VidVRD, as well as tackling the long-tail issue in existing benchmarks.

Video Object Detection. Video object detection is a fundamental step to tackling the VidVRD, some methods utilized optical flow to either propagate features from key frames to non-key frames [61] or enhance the features of each frame [60]. Other works also employed RNN [20] to process the video sequence, [25] reported the offline LSTM solutions that exploit all frames in the video, whereas the online solution [56] only considered adjacent frames. Tracking-based methods utilize the temporal information in an efficient way.

[10, 27] detected objects on fixed interval frames and tracked objects across frames. [26] further made their improvement by detecting interval frames adaptively. In this paper, we re-generated the object trajectories for ImageNet-VidVRD based on [13], which lies in the last category. It computed the cross-frame affinity score via visual feature similarity and IOU score and then associated bounding-box regions accordingly.

3 THE PROPOSED APPROACH

Given an untrimmed video V with arbitrary temporal lengths, our goal is to detect visual relations from the video in the form of triplet $\langle \text{subject}, \text{predicate}, \text{object} \rangle \in \mathcal{O} \times \mathcal{P} \times \mathcal{O}$. The *subject* (*object*) corresponds to the semantic label of spatio-temporally localized subject (*object*) trajectory $\mathcal{T}_s$ ($\mathcal{T}_o$) in a temporal window. Each trajectory is a sequence of object bounding-boxes detected by the pretrained object detectors, e.g., Faster-RCNN [32]. The *predicate* denotes the label of dynamic interaction between the *subject* and *object*. The $\mathcal{O}$ denotes the set of video object semantic categories and $\mathcal{P}$ denotes the set of *predicate* semantic categories. As defined by Shang et al. [34], the standard video relation detection usually consists of three components: 1) detecting all the possible object trajectory proposals $\{\mathcal{T}_i\}_{i=1}^T$ in a short video segment; 2) predicting the short-term video relations; and 3) associating short-term relation prediction into video-level relation instances. In this work, we mainly focus on relation prediction, which can be flexibly integrated into the relation detection frameworks in [34, 37].

In this section, we describe our proposed method in three steps: 1) object trajectory encoding and prediction in section 3.1; 2) interventional relationship prediction with predicate prototype learning in section 3.2; and 3) learning objective in section 3.3.

3.1 Object Trajectory Encoding and Prediction

3.1.1 Trajectory Proposal Generation. The first step to training a good VidVRD model is to detect object trajectory proposals from the given video V . Following the popular strategies [34, 37] in VidVRD, we first employ the pretrained image-level object detector, such as Faster-RCNN [32], to densely detect visual objects in each video frame; we then can apply robust object tracking techniques [7] or customized bounding-box linking strategies [13, 37] to link the detected object bounding-boxes across the video frames for generating valid trajectory proposals. The non-maximum suppression (NMS) [15] algorithm is applied to reduce the number of overlapping proposals.

Formally, given a video (segment), we generate a set of object trajectory proposals $\{\mathcal{T}_i\}_{i=1}^N$, and each proposal is a set of (sampled) spatio-temporal correlated bounding-boxes $\mathcal{T} = \{B_j\}_{j=1}^M$, where N is the number of objects in a video segment, and M is the number of frames in a video segment. Here, each bounding-box B is represented by a high-dimensional visual Region-of-Interest (RoI) feature $\mathbf{v} \in \mathbb{R}^{d^*}$, extracted by pre-trained region proposal networks, and the box coordinates (x, y, w, h) where (x, y) are the central coordinates of the bounding-box, and w and h denote its width and height, respectively.

3.1.2 Trajectory Proposal Encoding. To effectively model the spatio-temporal characteristic of object trajectories, we apply a

simple and effective method to learn the trajectory representation in this work. Given the set of bounding-box RoI features $\{\mathbf{v}_j\}_{j=1}^M$ of a trajectory proposal, we use the bidirectional Long short-term memory (LSTM) [20] network to model the temporal dependencies between adjacent object bounding boxes. Formally, the LSTM unit, at the j -th time step, takes the bounding box RoI feature $\mathbf{v}_j$, the previous hidden state $\mathbf{h}_{j-1}$, and cell state $\mathbf{c}_{j-1}$ as inputs, and outputs the current hidden state $\mathbf{h}_j$ and cell state $\mathbf{c}_j$:

$$(\mathbf{h}_j, \mathbf{c}_j) = \text{LSTM}(\mathbf{v}_j, \mathbf{h}_{j-1}, \mathbf{c}_{j-1}). \quad (1)$$

The output of the bidirectional LSTM network is $\mathcal{H} = \{\mathbf{h}_1^*, \dots, \mathbf{h}_M^*\}$ where $\mathbf{h}_j^* \in \mathbb{R}^d$ is the channel-wise vector concatenation of forward and backward hidden states at the j -th time step. The final video object trajectory representation $\mathbf{x} \in \mathbb{R}^d$ is obtained by applying average-pooling over $\mathcal{H}$.

Note that we only exploit the visual RoI feature to learn the object trajectory representation. We do not use any form of language features, e.g., *classme* [34, 37], since it will exacerbate the *spurious correlation* between semantic context and predicate labels. A similar issue is also pointed out and addressed in a scene-graph generation work [38] based on a causal inference method.

3.1.3 Video Object Prediction. We predict the semantic label of the object trajectory proposal by a prediction function $f^o(\cdot)$:

$$y^o = \text{Softmax}(f^o(\phi(\mathbf{x}))), \quad (2)$$

where $f^o(\cdot)$ is implemented by a fully-connected layer with learnable weighting parameter $\mathbf{W}_o \in \mathbb{R}^{d \times (|O|+1)}$, $\phi(\cdot)$ denotes the non-linear activation function *Relu*, and y^o is a normalized probabilistic distribution over the $|O|$ classes and a *background* class. The class with the highest score is treated as the object label of the trajectory.

3.2 Interventional Relation Prediction

3.2.1 Baseline Solution. The key of the VidVRD task is to predict the *predicate* label y^p of the dynamic interaction between the localized *subject* trajectory $\mathcal{T}_s$ and *object* trajectory $\mathcal{T}_o$. The first step is to represent the dynamic interaction by an interaction function $g(\mathcal{T}_s, \mathcal{T}_o)$ which should consider not only the visual feature interaction but also the spatial interaction across two trajectories. We implement $g(\cdot)$ in a simple and effective way:

$$g(\mathcal{T}_s, \mathcal{T}_o) = [\mathbf{x}_s, \mathbf{x}_o, \mathbf{r}_{loc}], \quad (3)$$

where $[\cdot]$ denotes the channel-wise vector concatenation, $\mathbf{x}_s \in \mathbb{R}^d$ and $\mathbf{x}_o \in \mathbb{R}^d$ denote the visual features of subject and object trajectories, respectively, and $\mathbf{r}_{loc}$ is the spatial location feature of the paired trajectories. We first compute the spatial location feature of the paired bounding boxes (B^s, B^o) as a vector $[s_x, s_y, s_w, s_h, s_a]$. It is calculated as follows: $s_x = \frac{x^s - x^o}{x^o}$, $s_y = \frac{y^s - y^o}{y^o}$, $s_w = \log \frac{w^s}{w^o}$, $s_h = \log \frac{h^s}{h^o}$, and $s_a = \log \frac{w^s \cdot h^s}{w^o \cdot h^o}$, where (x^s, y^s, w^s, h^s) , and (x^o, y^o, w^o, h^o) are the subject and object bounding-boxes respectively. The trajectory-level spatial location feature $\mathbf{r}_{loc} \in \mathbb{R}^d$ is computed by applying a bidirectional LSTM over the sequence of box-level spatial location feature, followed by an average-pooling. Then, we cast the output of the interaction into the predicate space by a predication function $f^p(\cdot)$. Formally, we can describe our

Figure 2: A brief overview of the *predicate prediction* in our IVRD based on an *intervention mechanism*. $\oplus$ denotes element-wise add operation and "fc" denotes a fully-connected layer. (some operations are omitted for simplicity.)

baseline solution of predicate prediction as

$$P(y^p | \mathcal{T}_s, \mathcal{T}_o) = \sigma(f^p(g(\mathcal{T}_s, \mathcal{T}_o))), \quad (4)$$

where $f^p(\cdot)$ is implemented by a fully-connected layer with weighting parameter $\mathbf{W}_p \in \mathbb{R}^{3d \times |\mathcal{P}|}$, and $\sigma(\cdot)$ is the sigmoid function $\sigma(x) = \frac{1}{1+\exp(-x)}$ that transforms the output of $f^p(\cdot)$ into $(0, 1)$.

Weakness: Most existing methods [24, 31, 37] use Eq. (4) for predicting the predicate in a relation triplet with different implementation of the interaction function $g(\cdot)$. Despite the effectiveness and simplicity of Eq. (4), it is easily misled by the long-tailed dataset bias that leads to a spurious correlation between the paired input $(\mathcal{T}_s, \mathcal{T}_o)$ and the predicate label. The first reason is that current pretrained object detection networks may entangle the semantic context of the detected object with its visual content. Then, due to the imbalanced and incomplete relation annotations in current VidVRD datasets [33, 34], the model would spuriously predict the predicate of $(\mathcal{T}_s, \mathcal{T}_o)$ as a frequently-appearing label, e.g., $\langle \text{person}, \text{next-to}, \text{dog} \rangle$, while ignores the temporal-aware transitive label, e.g., $\langle \text{person}, \text{walk-with}, \text{dog} \rangle$. Although the spatial predicate *next-to* sometimes is also a right answer, it is not the main focus of our VidVRD task. In short, existing methods mostly achieve their performance improvement on the frequently-appearing predicate categories while leaving the performance of the rest untouched. They thus fail to deal with less-dominant but sometimes more meaningful predicates, leading to poor generalization.

3.2.2 Our Solution (IVRD). Our goal is to prevent the model from relying on the spurious correlation caused by the long-tailed dataset bias. We expect the model to treat all predicates fairly during training. Inspired by the recent advances of causal reasoning in vision and language [41, 49, 52, 54, 57], we develop an **Interventional Video Relation Detection (IVRD)** method. As shown in Figure 2, the basic idea of IVRD is to intervene the paired model input $(\mathcal{T}_s, \mathcal{T}_o)$ using the *do-calculus* [28, 29] that forces the model to fairly incorporate each possible predicate into consideration:

$$P(y^p | do(\mathcal{T}_s, \mathcal{T}_o)) = \sum_{z \in \mathcal{Z}} P(y^p | (\mathcal{T}_s, \mathcal{T}_o), z) P(z), \quad (5)$$

where $\mathcal{Z} = \{z_1, \dots, z_{K_z}\}$ ($K_z = |\mathcal{P}|$) denotes a set of visual predicate prototypes [30]. $P(z)$ denotes the prior probability of the predicate prototype z . Note that we force the prior $P(z)$ to be a constant $\frac{1}{K_z}$ to alleviate the long-tailed dataset bias. We can see from Eq. (5) that the decision on the predicate label y^p depends on not only the model input $(\mathcal{T}_s, \mathcal{T}_o)$ but also all the predicate prototypes $\{z_1, \dots, z_{K_z}\}$. It can be seen as a way of removing spurious *observational bias* by forcing the model to virtually see more.

3.2.3 Implementation of IVRD. To implement Eq. (5), we organize all the predicate prototypes into a prototype dictionary $\mathbf{Z} = [z_1, \dots, z_{K_z}] \in \mathbb{R}^{K_z \times d^p}$ where $z \in \mathbb{R}^{d^p}$ represents the predicate prototype z . Then, we first rewrite Eq. (5) as

$$P(y^p | do(\mathcal{T}_s, \mathcal{T}_o)) := \mathbb{E}_z[\sigma(f(g(\mathcal{T}_s, \mathcal{T}_o), z))] \approx \sigma(\mathbb{E}_z[f(g(\mathcal{T}_s, \mathcal{T}_o), z)]), \quad (6)$$

where the approximation is based on the Normalized Weighted Geometric Mean (NWGM) [47] to avoid the expensive sampling of $\mathbb{E}_z$. We further implement $f(\cdot)$ as a linear model:

$$f(g(\mathcal{T}_s, \mathcal{T}_o), z) = \mathbf{W}_p^T g(\mathcal{T}_s, \mathcal{T}_o) + \mathbf{W}_z^T h_{so}(z), \quad (7)$$

where $\mathbf{W}_z \in \mathbb{R}^{d^p \times |\mathcal{P}|}$ is a learnable matrix. $h_{so}(z)$ is a transformation function of z , parameterized by the representation of $(\mathcal{T}_s, \mathcal{T}_o)$: $[x_s, x_o]$. Then, we move $\mathbb{E}_z$ into $f(\cdot)$ and rewrite Eq. (6) as:

$$P(y^p | do(\mathcal{T}_s, \mathcal{T}_o)) \approx \sigma(\mathbf{W}_p^T g(\mathcal{T}_s, \mathcal{T}_o) + \mathbb{E}_z[h_{so}(z)]). \quad (8)$$

The key to compute Eq. (8) is $\mathbb{E}_z[h_{so}(z)]$. We implement $h_{so}(z)$ as the scaled Dot-Product attention [40] to adaptively assign weights on different predicate prototypes in dictionary $\mathbf{Z}$ with the input $[x_s, x_o]$ as *query*. Then we have:

$$\mathbb{E}_z[h_{so}(z)] = \sum_z \left[\text{Softmax}\left(\mathbf{K}^T \mathbf{q} / \sqrt{d^p}\right) \odot \mathbf{Z} \right] P(z), \quad (9)$$

where $\mathbf{q} = \mathbf{W}_1[x_s, x_o]$, $\mathbf{K} = \mathbf{W}_2 \mathbf{Z}^T$ with learnable parameters $\mathbf{W}_1 \in \mathbb{R}^{d^p \times 2d}$ and $\mathbf{W}_2 \in \mathbb{R}^{d^p \times d^p}$. If removing $\mathbb{E}_z[h_{so}(z)]$ from Eq. (8), we can see that Eq. (8) will be the same as the baseline Eq. (4).

Prototype Dictionary Z: The quality of the prototype dictionary $\mathbf{Z}$ is also important to our IVRD. We construct $\mathbf{Z}$ in a simple and effective manner. Let z denote the prototype vector of a predicate $p \in \mathcal{P}$, e.g., *ride*. We first obtain all the relation triplets sharing the predicate p from the training set, $\{\langle s_i, p, o_i \rangle\}_{i=1}^{N_p}$. For each triplet $\langle s_i, p, o_i \rangle$, we represent it as $[\mathbf{v}_{s_i}^p, \mathbf{v}_{o_i}^p]$, where $[\cdot]$ denotes the channel-wise vector concatenation, and $\mathbf{v}_{s_i}^p \in \mathbb{R}^{d^p}$ is the average RoI features of the subject s_i 's bounding-boxes in all the annotated training instances of the triplet $\langle s_i, p, o_i \rangle$. We use the same way to compute the averaged object feature vector $\mathbf{v}_{o_i}^p$. During training, for the prototype of predicate p , we randomly sample K_p ($K_p \leq N_p$) triplets from $\{\langle s_i, p, o_i \rangle\}_{i=1}^{N_p}$ and calculate the prototype vector z as the averaged representation of the selected triplets: $z = \frac{1}{K_p} \sum \mathbf{W}_3[\mathbf{v}_{s_i}^p, \mathbf{v}_{o_i}^p]$ where $\mathbf{W}_3 \in \mathbb{R}^{d^p \times 2d^p}$ denotes a trainable transformation matrix.

So far, we have finished the introduction of our proposed IVRD method. Given the representation of subject trajectory and object trajectory and the predicate prototype dictionary, we can predict the predicate label using Eq. (8). Together with the video object

prediction in Eq. (2), we can finally finish the prediction of a video relation $\langle \text{subject}, \text{predicate}, \text{object} \rangle$.

3.3 Learning Objective

To train our method, we use the cross entropy loss function to calculate the loss of classifying video object trajectories in Eq. (2), $\mathcal{L}_{obj}(\mathcal{T}, \hat{y}^o)$. Due to the fact that a subject-object pair might have multiple predicate labels, we use the binary cross entropy loss function for predicate prediction in Eq. (8), $\mathcal{L}_{pred}(\mathcal{T}_s, \mathcal{T}_o, \mathbf{Z}, \hat{y}^p)$. Both $\mathcal{L}_{obj}$ and $\mathcal{L}_{pred}$ are computed as the average loss over a training batch. $\hat{y}^o$ and $\hat{y}^p$ denote the ground-truths of video object and predicate, respectively. The final loss is given as:

$$\mathcal{L} = \mathcal{L}_{obj} + \lambda \mathcal{L}_{pred}, \quad (10)$$

where λ is a trade-off hyperparameter.

4 EXPERIMENTS

4.1 Datasets, Settings, and Details

4.1.1 Datasets. To validate the effectiveness of our method, we perform comprehensive experiments on two datasets ImageNet-VidVRD [34] and VidOR [33]. (1) ImageNet-VidVRD is the first VidVRD dataset, which consists of 1,000 videos and is split into 800 training videos and 200 test videos. It covers 35 categories of subject/objects and 132 categories of predicates in total. The videos are densely annotated with relation triplets in the form of $\langle \text{subject}, \text{predicate}, \text{object} \rangle$ as well as the corresponding subject/object trajectories. There are 2,961 relation triplets in training set and 1,011 in test set. Particularly, the test set contains 258 zero-shot relation triplets that are unseen in training set. (2) VidOR consists of 7,000 videos for training, 835 videos for validation, and 2,165 videos for testing. We just use the validation set for evaluation, since the test set is used for grand challenge and not released for public use. Its annotations cover 80 categories of subject/objects and 50 categories of predicates. It has 6,258 relation triplets in the training set and 295 zero-shot relation triplets in the validation set.

4.1.2 Task Settings. Following [34], we evaluate our method on two tasks: *relation detection* and *relation tagging*. (1) The objective of relation detection is to generate a set of $\langle \text{subject}, \text{predicate}, \text{object} \rangle$ relation triplets with spatio-temporally localized trajectories in the video. A correct prediction should have the same relation triplet as the ground-truth and both the subject and object trajectories should have a high voluminal Intersection over Union (vIoU) score with those in ground-truths. In our experiments, the overlapping threshold of vIoU is set to 0.5. (2) The tagging task is less strict. In the evaluation stage, it just requires the predicted relation triplets to fall into the corresponding video's ground-truth set without the requirement of accurate object localization. The effectiveness of tagging task can support various video applications, such as video retrieval.

4.1.3 Evaluation protocols and Metrics. Following the standard evaluation protocol in [34], for relation detection task, we adopt mean Average Precision (mAP) and Recall@K (R@K) to evaluate the model performance. The mAP score calculates the average precision on each video and then takes the average over all videos. The Recall@K score measures the fraction of the positive relation

Table 1: Comparison with SOTA methods on ImageNet-VidVRD. Z-mAP denotes the mAP in zero-shot setting.

Method	Relation Detection				Relation Tagging		
	mAP	R@50	R@100	Z-mAP	P@1	P@5	P@10
Shang's [34]	8.58	5.54	6.37	0.40	43.00	28.90	20.80
GSTEG [39]	9.52	7.05	7.67	0.15	51.50	39.50	28.23
MHRA [8]	13.27	6.82	7.39	0.51	41.00	28.70	20.95
VRD-GCN [36]	16.26	8.07	9.33	0.74	57.50	41.00	28.50
PPN-ST [24]	18.38	11.21	13.69	-	60.00	43.10	32.24
MHA [36]	19.03	9.53	10.83	1.18	57.50	41.10	29.45
Baseline	21.24	11.96	14.17	0.60	66.50	47.60	34.88
IVRD	22.97	12.40	14.46	1.47	68.83	49.87	35.57

¹ We report the average performance of multiple runs with different random seeds.

² "-" means that the result is not reported by the paper.

detection in the *top-K* results, and K is set to 50 and 100. As for relation tagging, following [34], we use Precision@K (P@K) to measure the tagging accuracy. K is set to 1, 5, and 10.

4.1.4 Implementation Details. (1) For frame-level pre-processing on ImageNet-VidVRD [34], we adopt the fine-tuned Faster-RCNN [32] on this dataset with Resnet101 [19] backbone to detect object bounding boxes and extract RoI features, and videos are cut into consecutive 30-frames segments with 15-frames overlapping. We use our re-generated trajectories (termed **Traj-O**) based on the linking strategy in [13] for evaluation on ImageNet-VidVRD by default. The standard trajectories (termed **Traj-S**) released by [34] is also used for comparison with Traj-O in Table 4. (2) For frame-level preprocessing on VidOR [33], the pre-trained Faster-RCNN on the MS-COCO dataset is adopted with Resnet101 backbone to detect object bounding boxes and extract RoI features. All the trajectories are adopted from [37]. (3) During training, the batch size is set to 64 for ImageNet-VidVRD and 256 for VidOR. The number of trajectory proposals in each batch is set to 8. We use all the 30 frames in each segment for trajectory encoding on ImageNet-VidVRD by default and randomly select 5 frames in each segment on VidOR for efficiency. All the experiments are conducted on Nvidia-DGX with the Adam optimizer and the learning rate of 0.001. We select the best model with the highest mAP score on a validation set for final evaluation. The hyperparameter λ in Eq. (10) is set to 50.

4.2 Overall Performance Comparison

In this section, we compare our proposed IVRD method with the reported results of state-of-the-art (SOTA) methods on the two public benchmark datasets ImageNet-VidVRD [34] and VidOR [33]. The experimental results are depicted in Table 1 and 2. The **Baseline** method in Table 1 and 2 denotes our baseline solution in Eq. (3.2) without using the intervention strategy. It is implemented in the same framework and settings as **IVRD** for fair comparison.

The SOTA methods on the two datasets include Shang's [34], GSTEG [39], PPN-ST [24], VRD-GCN [31], MHRA [8], MHA [36], MAGUS.Gamma [37], RELAbuilder [58], and the regenerated results of CAI [62] and GSTEG in MAGUS.Gamma.

Table 2: Comparison with SOTA methods on VidOR.

Method	Relation Detection			Relation Tagging	
	mAP	R@50	R@100	P@1	P@5
RELAbuilder [58]	1.47	1.58	1.85	33.05	35.27
OTD+CAI [37, 62]	5.65	6.19	8.16	48.31	38.49
OTD+GSTEG [37, 39]	5.58	6.40	8.43	51.20	37.26
MAGUS.Gamma [37]	6.56	6.89	8.83	51.20	40.73
PPN-ST [24]	6.85	8.21	9.90	48.92	36.78
MHA [36]	6.59	6.35	8.05	50.72	41.56
Baseline	6.08	7.07	9.08	48.60	40.00
IVRD	7.42	7.36	9.41	53.40	42.70

- Shang’s [34] is the first baseline method for the VidVRD task and released the trajectories and relation association method for public evaluation on ImageNet-VidVRD.
- GSTEG, VRD-GCN, and PPN-ST mainly focus on modeling the spatio-temporal dependencies/contexts by different learning techniques, such as Markov conditional random fields and graph convolutional network (GCN). Particularly, PPN-ST extracts refined video features with spatial-temporal GCN.
- MHRA [8] and MHA [36] are mainly designed for improving the short-term relation association, where MHRA generates the hypothesis for each subject-object pair, whereas MHA generates the hypothesis on each possible relation instead.
- MAGUS.Gamma and RELAbuilder are the top 2 competitors in ACM MM 2019 Video Relation Understanding Challenge [35]. In particular, MAGUS.Gamma released its generated object trajectories on VidOR for public evaluation (used by our IVRD).

We have the following observations from Table 1 and 2:

- Our IVRD achieves the best performance on ImageNet-VidVRD in all metrics. IVRD obtains the highest mAP score and the best tagging performance on VidOR. The worse Recall scores in Table 2 are mainly because PPN-ST [24] uses a much stronger object detector for generating trajectories while we just use the released trajectories from MAGUS.Gamma [37]. The overall comparison results clearly indicate the effectiveness of our IVRD.
- IVRD consistently outperforms the baseline in all metrics on two datasets, which clearly demonstrates the effectiveness of our proposed intervention strategy in Eq. (5).
- IVRD achieves more significant relation tagging performance than all other methods on two datasets. This is because our intervention strategy focuses on the predicate label prediction, thus leading to a large improvement on the relation tagging task where accurate object localization is not required
- In Table 1, our IVRD achieves the best **zero-shot** detection result (Z-mAP, %), which reflects the advantages of IVRD in detecting unseen video relations and indicates its good generalizability.

Table 3: Effects of different construction strategies of the prototype dictionary on ImageNet-VidVRD.

Method	Relation Detection				Relation Tagging		
	mAP	R@50	R@100	Z-mAP	P@1	P@5	P@10
Baseline	21.24	11.96	14.17	0.60	66.50	47.60	34.88
IVRD-1	21.67	11.98	14.15	0.75	65.00	47.8	34.56
IVRD-2	22.18	12.36	14.40	0.47	66.25	49.70	35.50
IVRD	22.97	12.40	14.46	1.47	68.83	49.87	35.57

4.3 Study of IVRD

We conduct extensive experiments to investigate the effectiveness of IVRD by answering the following research questions: 1) how does the construction strategies of the predicate prototypes affect the performance of IVRD? (**R1**), 2) how does IVRD perform with different features of Sub-Object (SO) Interaction, different trajectory generation strategies when compared to the baseline? (**R2**), 3) how does the hyperparameters, like the number of sampled frames in segments, affect the performance of IVRD? (**R3**), 4) how does IVRD perform on the *tail* and *head* predicate groups? (**R4**).

4.3.1 Predicate Prototype Dictionary Construction. As shown in Eq. (8) in the section 3.2.3, our IVRD requires a predicate prototype dictionary $\mathbf{Z}$ for intervening the common input $(\mathcal{T}_s, \mathcal{T}_o)$. In this section, we design two variants of IVRD, termed **IVRD-1** and **IVRD-2**, to investigate the effect of the construction strategies: 1) IVRD-1: for constructing the dictionary, we directly compute each predicate prototype vector as the channel-wise concatenation of the averaged subject RoI features and the averaged object RoI features in all the annotated relation instances with the corresponding predicate. 2) IVRD-2: we compute the triplet-level prototype vector in the same way as IVRD in the section 3.2.3, and then compute each predicate prototype vector as the averaged triplet-level representation. The predicate prototype dictionaries of both IVRD-1 and IVRD-2 do not require randomly sample the triplets and are fixed (without the learnable transformation matrix.)

Table 3 shows the performance comparison among IVRD, IVRD-1, IVRD-2, and the baseline. We have the following observations:

- Both IVRD and its variants clearly improve the baseline in all metrics, especially in mAP and P@1, which indicates the prototype dictionary-based intervention strategy is effective.
- IVRD performs much better than IVRD-1 and IVRD-2. It indicates that our learnable prototype dictionary with a dynamic triplet sampling can work effectively in IVRD. In our design, the triplet sampling and the learnable pooling operation enables the loss to be back-propagated to the triplet level, so the model can perceive the discrepancy between subject-object pairs with the same predicate, e.g., $\langle \text{dog}, \text{chase}, \text{car} \rangle$ and $\langle \text{boy}, \text{chase}, \text{ball} \rangle$.

4.3.2 SO-Interaction Representation and Trajectories. As mentioned in Eq. (3), to effectively represent the dynamic subject-object (SO) interaction, we concatenate the subject’s visual embedding, object’s visual embedding, and their relative spatial location feature. We investigate the effect of this design in Table 4 on the two datasets by comparing the performance of IVRD with/without the spatial location feature. We also conduct the comparison on

Table 4: Comparison with the baseline on two datasets using different subject-object interaction features and trajectories. "-V" denotes that only visual RoI feature is used.

Method	ImageNet-VidVRD						VidOR		
	Traj-O			Traj-S					
	mAP	R@50	R@100	mAP	R@50	R@100	mAP	R@50	R@100
Baseline-V	19.45	10.01	11.83	17.59	9.45	11.23	5.12	6.57	8.80
IVRD-V	21.20	10.51	12.89	19.24	9.89	12.08	6.66	6.92	9.11
Baseline	21.24	11.96	14.17	18.48	10.44	12.16	6.07	7.07	9.08
IVRD	22.97	12.40	14.46	19.88	10.69	12.60	7.42	7.36	9.41

ImageNet-VidVRD in two different trajectory settings: 1) Traj-O: our re-generated object trajectories (Used by default) and 2) Traj-S: the standard trajectories released by Shang's [34]. We have the following observations from Table 4:

- Both IVRD and the baseline improve over their counterparts, IVRD-V and Baseline-V, in all metrics and trajectory settings. It indicates the effectiveness of the relative spatial location feature in Eq. (3), since the prediction of most predicates in ImageNet-VidVRD requires a good understanding the relative spatial locations of subject and object, such as *ride*, *follow*, and *walk-with*.
- Our re-generated object trajectories (Traj-O) achieve better relation detection results than the released trajectories (Traj-S) by [34], which demonstrates the advantage of our bounding-box linking technique [13] over the widely used method in [34].
- IVRD outperforms the baseline consistently in all features and trajectory settings, which further demonstrates the robustness of our intervention strategy.

4.3.3 Hyperparameters. We test the performance of IVRD on ImageNet-VidVRD with: 1) different numbers of sampled triplets for learning the prototype dictionary, $K_p \in \{5, 10, 15, \dots, 50\}$, as depicted in Figure 3 (a), and 2) different numbers of sampled frames in a segment for trajectory encoding, $M \in \{3, 5, 7, 10, 15, 20, 30\}$, as depicted in Figure 3 (b). In Figure 3 (b), "Ave" denotes a variant of IVRD which uses the averaged bounding-box RoI features to encode trajectories. "Ave" is shown as a baseline result. We have the following observations:

- We observe in Figure 3 (a) that the mAP score of IVRD goes up stably when we increase K_p from 5 to 20 and tends to fluctuate when $K_p \in \{20, \dots, 50\}$. It is reasonable, since the number of relation triplets of each predicate is unbalanced. Using too larger K_p will lead to over-smoothed prototype representation of the *head* predicates. In this work, we set K_p to 20 as a trade-off, because the average number of relation triplets of each predicate in the training set is 22.8.
- We see in Figure 3 (b) that a larger number of sampled frames in a segment can lead to a better performance of IVRD, since the number of sufficient frames may facilitate the modeling of temporal dependence across frames. But a large number of sampled frames in a segment also incurs a significantly high computational complexity for a larger video dataset. Therefore, we use all the 30 frames for the small dataset, while we just sample 5 frames in a segment, as a trade-off, for

Figure 3: Evaluation of IVRD with (a) different numbers of sampled triplets for learning the dictionary, and (b) different numbers of sampled frames for trajectory encoding.

Figure 4: Performance comparison (P-mAP, %) on 10 predicate groups in ImageNet-VidVRD dataset. P-mAP is a macro-averaging mAP metric over the predicate categories in the same group. The solid line in blue depicts the average training instances in each group.

the large dataset, VidOR. The performance on VidOR could be higher, if we increase the number of sampled frames.

4.3.4 Long-tailed Predicate Distribution. We investigate how our IVRD performs on the long-tailed predicate categories in ImageNet-VidVRD. We first sort the 132 predicate categories in a descending order based on their corresponding training instances and then we divide the sorted 132 predicate categories into 10 predicate groups evenly for easy illustration. As depicted in Figure 4, the top groups mostly correspond to the *head* predicate categories. We have the following observations from Figure 4:

- IVRD performs slightly worse than the baseline on the first group and achieves a bit better performance than the baseline on the top-6 predicate groups except the first group. However, our IVRD can significantly outperform the baseline on the four *tail* groups. The observation verifies the rationale of our aforementioned analysis on the long-tailed dataset bias. By relying on the dataset bias, the baseline performs very well on the popular predicate categories, but it cannot effectively generalize to the few-shot and zero-shot predicate categories. While, IVRD has a stronger ability of predicting the *tail* predicates, which should be attributed to the effectiveness of the intervention strategy in Eq. (5).

4.4 Quantitative Results and Analysis

4.4.1 Visualization of the Attended Prototypes. The key of our IVRD is to compute $\mathbb{E}_z [h_{so}(z)]$. It is implemented by a scaled

Figure 5: Qualitative examples of IVRD and Shang's [34]. The correct relation instances in the top 30 results are shown, and their ranks are marked in front of them with parentheses. (a) Results of relation detection and (b) results of relation tagging.

Dot-Product attention over the prototype dictionary, given the features of the subject-object trajectories as the *query*. Figure 6 shows two interesting real cases where the top-5 attended predicate prototypes are shown. We have the following observations:

- In Figure 6 (a), given the paired trajectories of $\langle \text{person}, \text{car} \rangle$ as input, we surprisingly find that both the top-2 attended prototypes (*toward* and *move-toward*) are the target predicates to be predicted and they semantically overlap with each other. Although the 5-th prototype *chase* is not the target predicate, it is semantically correlated with *move-toward*. When the input is $\langle \text{person}, \text{bicycle} \rangle$, all the top-5 attended prototypes (*ride*, *above*, *sit-above*, *taller*, and *stand-above*) are semantically correlated/similar. It demonstrates the effectiveness of the computation of $\mathbb{E}_z [h_{so}(z)]$. Less popular predicates have more chances to join the model training based on our intervention strategy, and the model is also able to capture the latent relationship between different predicates rather than treating each predicate independently.
- In the case depicted in Figure 6 (b), where the input is $\langle \text{adult}, \text{adult} \rangle$. Three target predicates (*in-front-of*, *hold-hand-of*, and *watch*) fall in the top-5 attended prototypes. In particular, the 4-th attended prototype is *hug* which has a strong temporal causal relationship with *hold-hand-of*. In the training set of VidOR, the action *hold-hand-of* is frequently (30%) followed by the action *hug* in the same video. It further verifies the rationale of our IVRD.

4.4.2 Visualization of Detection and Tagging Results. We illustrate the qualitative results of the relation detection and tagging tasks in Figure 5. Three real cases are shown for each task. In Figure 5 (a), we observe that our IVRD can not only well predict the spatial predicates, e.g., *stop-right*, but also the transitive predicates, e.g. *toward*, *creep-front*, *fight*, and *walk-with*, where the model is required to accurately understand the spatio-temporal interaction across objects. We see a wrong prediction, $\langle \text{dog}, \text{play}, \text{snake} \rangle$, in Figure 5 (a). But such prediction is acceptable and even may be correct, just because the annotators thought the *dog* was fighting with the *snake*. The action *fight* is visually similar to the action *play*. In Figure 5 (b), more accurate tagging results are ranked in top-20,

Figure 6: Visualization of the top-5 attended predicate prototypes, given the features of the subject-object trajectories, on ImageNet-VidVRD (a) and VidOR (b), where the numbers denote the normalized attention weights. Besides, the prototypes in red correspond to the ground-truth predicate labels.

where the temporal predicate (e.g. *feed*) and few-shot predicate (e.g. *faster*) are successfully predicted by IVRD.

5 CONCLUSIONS

This paper presented a causality-inspired video relation detection approach, termed IVRD, that improved the overall accuracy, while strengthen the robustness of model prediction. In particular, the intervention is performed via a predicate prototype dictionary that enables the model to incorporate every predicate prototype fairly for making a more robust decision. Extensive experiments on two popular VidVRD datasets show that our IVRD is effective on few-shot and even zero-shot transitive predicate categories. In future, we intend to investigate how video relation helps downstream task like VideoQA, Video Caption.

6 ACKNOWLEDGMENTS

This research is funded by Sea-NExT Joint Lab.

REFERENCES

- [1] Qianwen Cao and Heyan Huang. 2021. VSRN: Visual-Semantic Relation Network for Video Visual Relation inference. *IEEE TCSVT* (2021).
- [2] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *ACM MM*. 32–41.
- [3] Jing-jing Chen, Chong-Wah Ngo, and Tat-Seng Chua. 2017. Cross-modal recipe retrieval with rich food attributes. In *ACM MM*. 1771–1779.
- [4] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. 2019. Knowledge-embedded routing network for scene graph generation. In *CVPR*. 6163–6171.
- [5] H. Cui, L. Zhu, J. Li, Y. Yang, and L. Nie. 2020. Scalable Deep Hashing for Large-Scale Social Image Retrieval. *IEEE TIP* 29 (2020), 1271–1284.
- [6] Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. In *CVPR*. 3076–3086.
- [7] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. 2014. Accurate scale estimation for robust visual tracking. In *BMVC*. BMVA Press.
- [8] Donglin Di, Xindi Shang, Weinan Zhang, Xun Yang, and Tat-Seng Chua. 2019. Multiple Hypothesis Video Relation Detection. In *BigMM*. IEEE, 287–291.
- [9] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual encoding for video retrieval by text. *IEEE TPAMI* (2021).
- [10] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2017. Detect to Track and Track to Detect. In *ICCV*.
- [11] Zan Gao, Leming Guo, Weili Guan, An-An Liu, Tongwei Ren, and Shengyong Chen. 2021. A Pairwise Attentive Adversarial Spatiotemporal Network for Cross-Domain Few-Shot Action Recognition. *IEEE TIP* 30 (2021), 767–782.
- [12] Zan Gao, Yinming Li, and Shaohua Wan. 2020. Exploring Deep Learning for View-Based 3D Model Retrieval. *ACM TOMM* 16, 1 (2020), 1–21.
- [13] Georgia Gkioxari and Jitendra Malik. 2015. Finding action tubes. In *CVPR*. 759–768.
- [14] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. 2019. Scene graph generation with external knowledge and image reconstruction. In *CVPR*. 1969–1978.
- [15] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. 2016. Seq-nms for video object detection. *arXiv preprint arXiv:1602.08465* (2016).
- [16] Yanbin Hao, Zi-Niu Liu, Hao Zhang, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. 2020. Person-level Action Recognition in Complex Events via TSD-TSM Networks. In *ACM MM*. 4699–4702.
- [17] Yanbin Hao, Chong-Wah Ngo, and Benoit Huet. 2019. Neighbourhood structure preserving cross-modal embedding for video hyperlinking. *IEEE TMM* 22, 1 (2019), 188–200.
- [18] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, Qiang Liu, and Xiaojun Hu. 2020. Compact Bilinear Augmented Query Structured Attention for Sport Highlights Classification. In *ACM MM*. 628–636.
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.
- [20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [21] J. Johnson, R. Krishna, M. Stark, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. 2015. Image retrieval using scene graphs. In *CVPR*. 3668–3678.
- [22] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017. Scene graph generation from objects, phrases and region captions. In *ICCV*. 1261–1270.
- [23] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. 2019. Vrr-vg: Refocusing visually-relevant relationships. In *ICCV*. 10403–10412.
- [24] Chenchen Liu, Yang Jin, Kehan Xu, Guoqiang Gong, and Yadong Mu. 2020. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *CVPR*. 10840–10849.
- [25] Mason Liu, Menglong Zhu, Marie White, Yinxiao Li, and Dmitry Kalenichenko. 2019. Looking Fast and Slow: Memory-Guided Mobile Video Object Detection. *arXiv:1903.10172 [cs.CV]*
- [26] Hao Luo, Wenxuan Xie, Xinggang Wang, and Wenjun Zeng. 2019. Detect or track: Towards cost-effective video object detection/tracking. In *AAAI*, Vol. 33. 8803–8810.
- [27] Huizi Mao, Taeyoung Kong, and William J. Dally. 2019. CaTDet: Cascaded Tracked Detector for Efficient Object Detection from Video. In *SysML*.
- [28] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.
- [29] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic Books.
- [30] Francois Plesse, Alexandru Ginsca, Bertrand Delezoide, and Francoise Preteux. 2020. Modelling relations with prototypes for visual relation detection. *Multimedia Tools and Applications* (2020), 1–22.
- [31] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. 2019. Video relation detection with spatio-temporal graph. In *ACM MM*. 84–93.
- [32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*.
- [33] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. In *ICMR*. 279–287.
- [34] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video visual relation detection. In *ACM MM*. 1300–1308.
- [35] Xindi Shang, Junbin Xiao, Donglin Di, and Tat-Seng Chua. 2019. Relation Understanding in Videos: A Grand Challenge Overview. In *ACM MM*. 2652–2656.
- [36] Zixuan Su, Xindi Shang, Jingjing Chen, Yu-Gang Jiang, Zhiyong Qiu, and Tat-Seng Chua. 2020. Video Relation Detection via Multiple Hypothesis Association. In *ACM MM*. 3127–3135.
- [37] Xu Sun, Tongwei Ren, Yuan Zi, and Gangshan Wu. 2019. Video visual relation detection via multi-modal feature fusion. In *ACM MM*. 2657–2661.
- [38] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *CVPR*. 3716–3725.
- [39] Yao-Hung Hubert Tsai, Santosh Divvala, Louis-Philippe Morency, Ruslan Salakhutdinov, and Ali Farhadi. 2019. Video relationship reasoning using gated spatio-temporal energy graph. In *CVPR*. 10424–10433.
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*. 5998–6008.
- [41] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. 2020. Visual commonsense r-cnn. In *CVPR*. 10760–10770.
- [42] Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie, Zhouchen Lin, and Baoquan Chen. 2019. Neural multimodal cooperative learning toward micro-video understanding. *IEEE TIP* 29 (2019), 1–14.
- [43] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *ACM MM*. 1437–1445.
- [44] Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. 2020. Visual relation grounding in videos. In *ECCV*. Springer, 447–464.
- [45] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NeXT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *CVPR*.
- [46] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *CVPR*. 5410–5419.
- [47] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*. 2048–2057.
- [48] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. 2020. Tree-Augmented Cross-Modal Encoding for Complex-Query Video Retrieval. In *SIGIR*. 1339–1348.
- [49] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded Video Moment Retrieval with Causal Intervention. In *SIGIR*.
- [50] Xun Yang, Xiangnan He, Xiang Wang, Yunshan Ma, Fuli Feng, Meng Wang, and Tat-Seng Chua. 2019. Interpretable fashion matching with rich attributes. In *SIGIR*. 775–784.
- [51] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. 2020. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *ACM MM*. 1939–1947.
- [52] Xu Yang, Hanwang Zhang, and Jianfei Cai. 2020. Deconfounded image captioning: A causal retrospect. In *ECCV*.
- [53] Xun Yang, Peicheng Zhou, and Meng Wang. 2018. Person reidentification via structural deep metric learning. *IEEE TNNLS* 30, 10 (2018), 2987–2998.
- [54] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. 2020. Interventional Few-Shot Learning. In *NeurIPS*.
- [55] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *CVPR*. 5831–5840.
- [56] Chen Zhang and Joohye Kim. 2019. Modeling Long- and Short-Term Temporal Context for Video Object Detection. In *ICIP*. 71–75.
- [57] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. 2020. Causal intervention for weakly-supervised semantic segmentation. In *NeurIPS*.
- [58] Sipeng Zheng, Xiangyu Chen, Shizhe Chen, and Qin Jin. 2019. Relation understanding in videos. In *ACM MM*.
- [59] Lei Zhu, Xu Lu, Zhiyong Cheng, Jingjing Li, and Huaxiang Zhang. 2020. Deep Collaborative Multi-View Hashing for Large-Scale Image Search. *IEEE TIP* 29 (2020), 4643–4655.
- [60] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Flow-guided feature aggregation for video object detection. In *ICCV*. 408–417.
- [61] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Deep Feature Flow for Video Recognition. In *CVPR*.
- [62] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. 2017. Towards context-aware interaction recognition for visual relationship detection. In *ICCV*. 589–598.