Adaptive Image-to-Video Scene Graph Generation via Knowledge Reasoning and Adversarial Learning

Jin Chen, Xiaofeng Ji, Xinxiao Wu*
Beijing Laboratory of Intelligent Information Technology
School of Computer Science, Beijing Institute of Technology, Beijing, China
{chen_jin, jixf, wuxinxiao}@bit.edu.cn

Abstract

Scene graph in a video conveys a wealth of information about objects and their relationships in the scene, thus benefiting many downstream tasks such as video captioning and visual question answering. Existing methods of scene graph generation require large-scale training videos annotated with objects and relationships in each frame to learn a powerful model. However, such comprehensive annotation is time-consuming and labor-intensive. On the other hand, it is much easier and less cost to annotate images with scene graphs, so we investigate leveraging annotated images to facilitate training a scene graph generation model for unannotated videos, namely image-to-video scene graph generation. This task presents two challenges: 1) infer unseen dynamic relationships in videos from static relationships in images due to the absence of motion information in images; 2) adapt objects and static relationships from images to video frames due to the domain shift between them. To address the first challenge, we exploit external commonsense knowledge to infer the unseen dynamic relationship from the temporal evolution of static relationships. We tackle the second challenge by hierarchical adversarial learning to reduce the data distribution discrepancy between images and video frames. Extensive experiment results on two benchmark video datasets demonstrate the effectiveness of our method.

Introduction

The task of generating a scene graph in a video aims to detect objects and their relationships on both spatial and temporal dimensions, which provides a fine-grained representation of the video and underpins numerous downstream visual tasks, such as action recognition (Girdhar et al. 2017),video captioning (Xu et al. 2019; Hao, Zhou, and Li 2020; Cao, Zhao, and Fu 2020), video retrieval (Girdhar et al. 2017) and visual question answering (Liu and Huet 2016). Existing methods require a large number of training videos to be annotated with objects and their relationships in each video frame. However, it is a time-consuming and labor-intensive process to acquire such comprehensive annotation. On the other hand, it is much easier and less cost to annotate scene graphs in images and also there exist several available annotated image datasets such as Visual Genome (Krishna et al. 2017) and Visual Relationship Dataset (Lu et al. 2016).

*Corresponding author: Xinxiao Wu

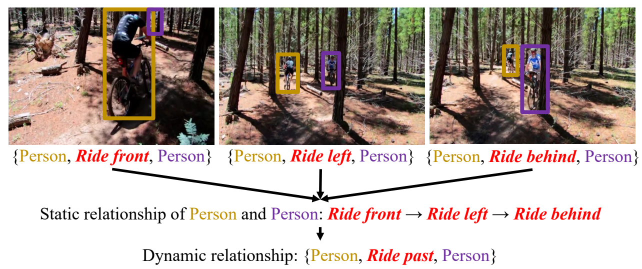


Figure 1: An example of inferring the dynamic relationship from static relationships on the time dimension. The subject and object are denoted in the brown box and the purple box, respectively.

Therefore, we investigate exploiting existing annotated images to facilitate training a video scene graph generation model without video annotations, called *image-to-video scene graph generation*, which breaks the heavy dependency on the large-scale annotated training videos. This new task introduces two challenges. First, since the temporal motion information is absent in images, it is difficult for a scene graph generation model trained using images to capture the dynamic object relationships in videos. Second, the domain shift between images and video frames makes the difficulty to adapt the detection models of objects and static relationships from images to videos.

To address the first challenge, we propose knowledge reasoning to infer unseen dynamic relationships in videos. Our insight is that a dynamic relationship can be inferred from the temporal evolution of related static relationships. As illustrated in Figure 1, the dynamic relationship {person, ride past, person} can be inferred from sequential static relationships: {person, ride front, person} $\rightarrow$ {person, ride left, person} $\rightarrow$ {person, ride behind, person}. We denote such association between static and dynamic relationships as commonsense knowledge that can be exploited from many external text resources such as Action genome (Ji et al. 2020) and Wikipedia (Pataki, Vajna, and Marosi 2012). To be more specific, starting with learning a shared embedding space between visual features and language features of relationships, called visual-language embedding space, we then learn to generate the embedding feature of an unseen dynamic relationship from its associated sequential static relationships in the visual-language embedding space for prediction, with the guidance of the commonsense knowledge.

To tackle the second challenge, we propose hierarchical adversarial learning to reduce the domain shift in both image and instance levels for adapting an object detection model from images to video frames. Specifically, the image-level shift (*e.g, variance of image style, illumination, etc.*) is minimized by aligning the second-order statistics of the image and video frame features via adversarial training between a domain classifier and a feature extractor. The instance-level shift (*e.g, variance of object appearance, size, etc.*) is minimized by aligning the appearance of region proposals extracted from images and video frames in a similar adversarial manner. In this way, we learn the domain-invariant visual features of images and video frames, thus benefiting the prediction of static relationships in video frames.

The contributions of this work are three-fold: (1) We propose a new task, image-to-video scene graph generation, that adapts well the scene graph generation model trained using annotated images to unannotated videos. This task breaks the heavy dependency on large-scale videos annotated with objects and their relationships for training, making it more practical and general in real-world scenarios. (2) We propose a knowledge reasoning method that exploits external commonsense knowledge to infer unseen dynamic relationships from the temporal evolution of static relationships. (3) We propose a hierarchical adversarial learning method to reduce the domain shift between image and video domains for promoting the adaptation of objects and static relationships.

Related Work

Video Scene Graph Generation

Video scene graph generation is more challenging than image scene graph generation since there exist dynamic relationships with complex changes over both space and time dimension. Shang *et al.* (Shang et al. 2017) firstly build a dataset for video visual relationship detection and propose a three-stage scheme including object tracklet proposal generation, relationship prediction and relationship association. Later, several methods focus on learning relationship representation via constructing spatial-temporal graph by conditional random fields (Tsai et al. 2019) or graph convolutional networks (Qian et al. 2019; Liu et al. 2020). In (Su et al. 2020), Su *et al.* propose a multiple hypothesis association method to handle the inaccurate or missing problem in the relationship association.

All existing methods require a large-scale number of annotated videos for training, but it is costly to label objects and relationships in every frame. In contrast, our method does not rely on the annotated videos and uses existing available annotated images for training the video scene graph generation model, which is more suitable for realistic applications.

Image-to-video Adaptation

Image-to-video adaptation has been applied into many visual tasks such as action recognition (Li et al. 2017; Yu et al. 2018; Liu et al. 2019; Yu et al. 2019; Chen et al. 2021b) and object detection (Chanda et al. 2017; RoyChowdhury et al. 2019; Lahiri et al. 2019), which transfers the knowledge from images to videos in order to relieve the reliance on the large-scale training videos. In video action recognition, Chen *et al.* (Chen et al. 2021b) introduce a spatial-temporal causal inference framework, which can help infer how the spatial and temporal domain shifts affect the adaptation via counterfactual causality. In video object detection, Chanda *et al.* (Chanda et al. 2017) transfer the knowledge from labeled images to weakly labeled videos with a two-stream architecture trained on images and video frames.

To the best of our knowledge, we are the first to apply the image-to-video adaptation on video relation detection, *i.e.*, image-to-video scene graph generation. Compared with the aforementioned two tasks, our task involves both cross-domain object detection and cross-domain relationship detection and is more challenging.

Our Method

Overview

In this paper, we propose an image-to-video scene graph generation method that infers unseen dynamic relationships in videos via knowledge reasoning and reduces the domain shift via hierarchical adversarial learning. Our method consists of three modules: a cross-domain object detection module, a static relationship prediction module and a dynamic relationship learning module, as illustrated in Figure 2.

Formulation

In this task, we are given an annotated source image domain and an unannotated target video domain. The source domain is denoted as $\mathcal{D}_s = \{(\boldsymbol{x}_i^s, \mathcal{G}_i^s)|_{i=1}^{N_s}\}$, where $\boldsymbol{x}_i^s$ represents the i-th image, and $\mathcal{G}_i^s$ denotes the scene graph annotation of $\boldsymbol{x}_i^s$. Each scene graph annotation $\mathcal{G}$ is represented as a 3-tuple set $\mathcal{G} = \{B, O, R\}$. $B = \{\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_n\}$ is a region proposal set, where $\boldsymbol{b}_k \in \mathbb{R}^4$ denotes the bounding box of the k-th region proposal. $O = \{o_1, o_2, ..., o_n\}$ is an object set, where $o_k \in \mathcal{C}$ is the class label of $\boldsymbol{b}_k$, and $\mathcal{C}$ is the set of all object classes including background. $R = \{r_{1\to 2}, r_{1\to 3}, ..., r_{n\to n-1}\}$ is a relationship set, where $r_{k\to q}$ is a triplet of a subject $(o_k, \boldsymbol{b}_k) \in O \times B$, an object $(o_q, \boldsymbol{b}_q) \in O \times B$ and a predicate label $y^p_{k\to q} \in \mathcal{P}$, and $\mathcal{P}$ is the set of all predicate classes including non-relationship. The target domain is formulated as $\mathcal{D}_t = \{\boldsymbol{x}_i^t|_{i=1}^{N_t}\}$, where $\boldsymbol{x}_i$ represents the i-th video. Each video consists of multiple video frames, formulated as $\boldsymbol{x}_i^t = \{\boldsymbol{f}_{i,j}|_{j=1}^{N_i}\}$, where $\boldsymbol{f}_{i,j}$ denotes the j-th frame of the i-th video.

Cross-domain Object Detection by Hierarchical Adversarial Learning

There exist two-level domain shifts between images and video frames: 1) the image-level shift caused by the variances of image styles, illumination and the motion blur in

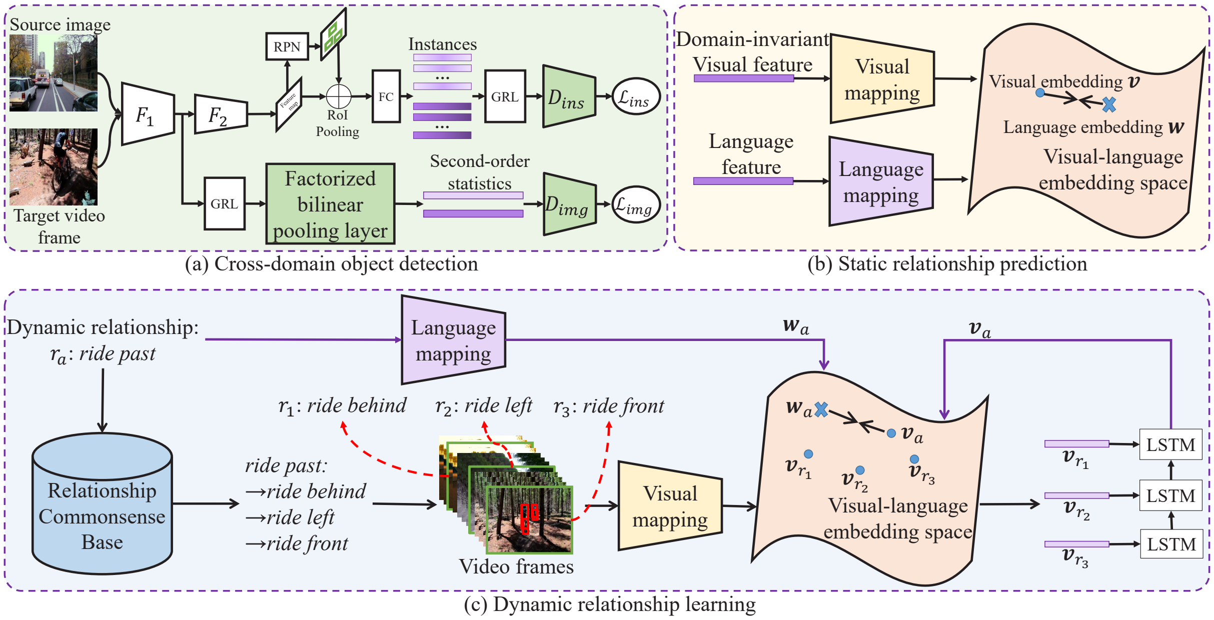


Figure 2: Overview of our method. (a) The cross-domain object detection module learns domain-invariant features between images and video frames by hierarchical adversarial learning to reduce the image-level and instance-level shifts simultaneously, where GRL is a gradient reverse layer (Ganin and Lempitsky 2015) to enable adversarial training. (b) The static relationship prediction module learns to project both the domain-invariant visual feature and the language feature of each relationship into a visual-language embedding space to generate visual embedding and language embedding, respectively. (c) The dynamic relationship learning module learns to generate the visual embedding of dynamic relationship from sequential static relationships by exploiting the external Relationship Commonsense Base.

videos; 2) the instance-level shift caused by the variances of object appearances. To reduce them, we propose hierarchical adversarial learning that incorporates two adversarial learning components into the training of detection model to learn the domain-invariant features of images and video frames.

Image-level Adversarial Learning. We align the second-order statistics of the image and video frame features to reduce the image-level shift since the second-order statistics contain pairwise correlations between features, well reflecting the detailed information in images. Since the low-level feature contains more texture information, an adversarial learning component is constructed on the low-level feature, which consists of a domain classifier and a feature extractor of the object detector.

The feature extractor F consists of F_1 and F_2, and the domain classifier D_{img} is designed to predict the domain labels of the second-order statistics of features extracted from F_1. Given an input image $\boldsymbol{x}$, we represent the convolutional features extracted from F_1 as $F_1(\boldsymbol{x}) \in \mathbb{R}^{C\times W\times H}$, where C is the number of distinct filters (the number of feature maps), W and H are the width and height of each feature map, respectively. A factorized bilinear pooling scheme (Gao et al. 2020) is utilized to compute the second-order statistics of image features and video frame features. Concretely, the convolutional features $F_1(\boldsymbol{x})$ are reshaped into a matrix $\boldsymbol{M} = [\boldsymbol{m}_1, ..., \boldsymbol{m}_N] \in \mathbb{R}^{C\times N}$ where $\boldsymbol{m}_j \in \mathbb{R}^C$ represents the j-th column. A d-dimensional second-order statistic descriptor $\boldsymbol{g} \in \mathbb{R}^d$ of $\boldsymbol{M}$ is computed by

$$\boldsymbol{g} = \sum_j \boldsymbol{A}(\boldsymbol{U}^\top \boldsymbol{m}_j \circ \boldsymbol{V}^\top \boldsymbol{m}_j), \tag{1}$$

where $\boldsymbol{U} \in \mathbb{R}^{C\times L}$ and $\boldsymbol{V} \in \mathbb{R}^{C\times L}$ are learnable parameters, $L = r \times d$, and r is a hyperparameter. $\circ$ denotes the Haramard product. $\boldsymbol{A} \in \mathbb{R}^{d\times L}$ is a fixed binary matrix and in the l-th row of $\boldsymbol{A}$ with $l \in [1, d]$, the elements from column $((l-1)\times r + 1)$ to column $(l \times r)$ are set to "1" and others are set to "0".

For the domain classifier D_{img}, its input is the second-order statistics of image features $\boldsymbol{g}_i^s$ or video frame features $\boldsymbol{g}_{i,j}^t$, and the output of D_{img} is the domain label of the second-order statistics of input features, *i.e.,* 0 for source image and 1 for target video frames. We utilize a least-squares loss (Mao et al. 2017) to train D_{img} for distinguishing the second-order statistics of images from that of video frames, formulated by

$$\mathcal{L}_{img} = \sum_i \left(D_{img}(\boldsymbol{g}_i^s)\right)^2 + \sum_{i,j} \left(1 - D_{img}(\boldsymbol{g}_{i,j}^t)\right)^2. \tag{2}$$

The feature extractor F_1 tries to confuse D_{img} to make the second-order statistics of the two different domains as indistinguishable as possible. Hence, D_{img} and F_1 are optimized via adversarial learning: $\max_{F_1} \min_{D_{img}} \mathcal{L}_{img}$.

Instance-level Adversarial Learning. We employ a patch-based adversarial learning method (Chen et al. 2021a) to reduce the instance-level domain shift, thus further improving the detection performance. Specifically, an instance domain classifier D_{ins} is introduced to predict multiple domain labels for pixels of a region proposal of images or video frames. Let W_2 and H_2 denote the width and height of a region proposal, respectively. The output of D_{ins} is a domain prediction map with the size of $W_2 \times H_2$, and $D_{ins}(\boldsymbol{p})_{(w,h)}$ denotes the domain prediction of the pixel (w, h) of the region proposal $\boldsymbol{p}$. Let P_i^s and $P_{i,j}^t$ denote the region proposal sets of source image $\boldsymbol{x}_i^s$ and target video frame $\boldsymbol{f}_{i,j}^t$, respectively. The loss of D_{ins} is formulated by

$$\begin{aligned}\mathcal{L}_{ins} = &\sum_i \sum_{\boldsymbol{p} \in P_i^s} \sum_{w,h} \left(D_{ins}(\boldsymbol{p})_{(w,h)}\right)^2 \\ &+ \sum_{i,j} \sum_{\boldsymbol{p} \in P_{i,j}^t} \sum_{w,h} \left(1 - D_{ins}(\boldsymbol{p})_{(w,h)}\right)^2.\end{aligned} \tag{3}$$

Similarly, D_{ins} and F are optimized via adversarial learning: $\max_{F} \min_{D_{ins}} \mathcal{L}_{ins}$, to make the region proposals of the two different domains as distinguishable as possible.

Therefore, the complete objective is given by

$$\mathcal{L}_{adpt} = \mathcal{L}_{det} + \mathcal{L}_{img} + \mathcal{L}_{ins}, \tag{4}$$

where $\mathcal{L}_{det}$ denotes the detection losses detailed in (Ren et al. 2015), including a classification loss and a bounding box regression loss.

Predicating Static Relationship by Visual-language Embedding Space

We learn a visual-language embedding space to bridge the visual and language modalities for predicting static relationships. We construct a visual mapping ϕ and a language mapping φ to project the domain-invariant visual features and the language features (*i.e.,* word vectors) of relationships into the visual-language embedding space, respectively, where the distance of the matched visual and language embeddings is minimized and that of the dismatched ones is maximized.

We use images and their corresponding scene graph annotations to learn the visual-language embedding space. Let $\boldsymbol{z}$ and $\boldsymbol{e}$ denote the visual feature and the language feature of a relationship $r_{k\to q} = \{o_k, y_{k\to q}, o_q\}$, respectively, where o_k and o_q represent the subject class label and the object class label, respectively, and $y_{k\to q}$ represents a predicate label between subject o_k and object o_q. The visual feature $\boldsymbol{z}$ is extracted from images, consisting of 1) domain-invariant visual features of subject, object, and predicate, and 2) a spatial feature (Liang et al. 2018) of the relative location of subject and object. All domain-invariant visual features are extracted by RoI pooling from the object detector via the corresponding bounding box, and the bounding box of the predicate is the union bounding box of the subject bounding box $\boldsymbol{b}_k$ and the object bounding box $\boldsymbol{b}_q$. The language feature $\boldsymbol{e}$ is represented by an word vector of the predicate label $y_{k\to q}$, extracted from GloVe (Pennington, Socher, and Manning 2014). We project $\boldsymbol{z}$ and $\boldsymbol{e}$ by the visual mapping ϕ and the language mapping φ, respectively, formulated as

$$\boldsymbol{v} = \phi(\boldsymbol{z}), \boldsymbol{w} = \varphi(\boldsymbol{e}), \tag{5}$$

where $\boldsymbol{v}$ and $\boldsymbol{w}$ represent the visual and language embeddings of the relationship $r_{k\to q}$, respectively.

The visual-language embedding space is learned by minimizing the distance of the matched visual and language embeddings and maximizing that of the dismatched ones, and the loss is given by

$$\begin{aligned}\mathcal{L}_{emb} = &\sum_i \sum_{r_{k\to q} \in R_i^s} \mathbb{I}_{y_{k\to q}=1} \log\left(\frac{1}{1+e^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v}}}\right) \\ &+ \sum_i \sum_{r_{k\to q} \in R_i^s} \mathbb{I}_{y_{k\to q}=0} \log\left(\frac{1}{1+e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{v}}}\right),\end{aligned} \tag{6}$$

where $\mathbb{I}_{y_{k\to q}=0}$ and $\mathbb{I}_{y_{k\to q}=1}$ are indicator functions. When $y_{k\to q} = 1$, $\mathbb{I}_{y_{k\to q}=1} = 1$, which means that $\boldsymbol{v}$ and $\boldsymbol{w}$ are matched and otherwise dismatched. R_i^s is the relationship set of the source image $\boldsymbol{x}_i^s$.

Learning Dynamic Relationship by Knowledge Reasoning

Due to the absence of dynamic relationships in images, it is impossible to optimize the distance between the domain-invariant visual features and language features of dynamic relationships in the visual-language embedding space learned with source images. Fortunately, there exists the association between a dynamic relationship and a sequence of static relationships, and a dynamic relationship can be represented by the temporal evolution of static relationships. Such association can be regarded as commonsense knowledge of the dynamic relationship. In this paper, we propose knowledge reasoning to first generate an associated sequential static relationships for a dynamic relationship with the guidance of commonsense knowledge, and then learn a visual embedding of the dynamic relationship from the generated sequence by minimizing its distance to the language embedding.

Commonsense Knowledge. We generate commonsense knowledge from both the popular action recognition dataset, *i.e.,* Action Genome (Ji et al. 2020), and the most widely used visual relationship detection datasets, *i.e.,* the VidVRD dataset (Shang et al. 2017) and the VidOR dataset (Shang et al. 2019). In Action Genome, each action corresponds to five temporally sequential scene graphs. The action is actually a dynamic relationship, and the static relationships with the same subject and object to this action are chosen from the five scene graphs as a sequence of static relationships associated with the dynamic relationship. For example, for the "awakening in bed" action, the two relationship triplets {person, lying on, bed} and {person, sitting on, bed} are selected from the scene graphs, and formulated as a rule {awakening in: lying on→sitting on}. For the VidVRD dataset (Shang et al. 2017) and the VidOR dataset (Shang et al. 2019), we first count the frequency of static relationships that appear together with the dynamic relationship of the same subject and object and then summarize a rule manually according to the frequency. For example, for the dynamic relationship "past", the three most frequent static relationships are "front", "behind", and "right", and the association between these relationships are formulated as a rule manually, *i.e.,* {past: front→right→behind}. Totally, we obtain 249 rules about 35 dynamic relationships and 76 static relationships to build a Relationship Commonsense Base (RCB).

Learning Visual Embeddings of Dynamic Relationships. We propose knowledge reasoning to learn dynamic relationships. First, we generate sequential static relationships by sampling video frames according to the rules in RCB. Second, the visual embedding of a dynamic relationship is learned by modeling the temporal evolution of the generated sequential static relationships via LSTM. Finally, the distance between the visual embedding and the language embedding of the dynamic relationship is minimized to optimize LSTM.

Specifically, for a dynamic relationship r_a, we retrieve its corresponding rule $\{r_a : r_1, r_2, ..., r_t\}$ from RCB where $\{r_1, r_2, ..., r_t\}$ represents a sequence of static relationships associated with the dynamic relationship r_a. With the retrieved rule, we obtain domain-invariant features of these static relationships by sampling relationship instances of the corresponding labels (*i.e.,* $\{r_1, r_2, ..., r_t\}$) from video frames, and extract their visual embeddings $\{\boldsymbol{v}_{r_1}, \boldsymbol{v}_{r_2}, ..., \boldsymbol{v}_{r_t}\}$ using the visual mapping ϕ learned by the static relationship prediction module. With the visual embeddings of the sequential static relationships, the visual embedding $\boldsymbol{z}_a$ and the language embedding $\boldsymbol{w}_a$ of the dynamic relationship r_a are obtained via LSTM and the language mapping φ, respectively, formulated as

$$\boldsymbol{z}_a = \mathrm{LSTM}(\boldsymbol{v}_{r_1}, \boldsymbol{v}_{r_2}, ..., \boldsymbol{v}_{r_t}), \boldsymbol{w}_a = \varphi(\boldsymbol{e}_{r_a}), \quad (7)$$

where $\boldsymbol{e}_{r_a}$ is the language feature (*i.e.,* word vector) of r_a. Afterwards, the distance between $\boldsymbol{z}_a$ and $\boldsymbol{w}_a$ is minimized to optimize LSTM:

$$\min_{LSTM} \mathcal{L}_{dis} = ||\boldsymbol{w}_a - \boldsymbol{z}_a||_2. \quad (8)$$

Scene Graph Generation in Videos

During testing, given an input video, we first detect objects for each video frame via the cross-domain object detector and then predict static relationships for all the combinations of detected objects by finding the most similar language embedding as the relationship label in the visual-language embedding space. Afterwards, the static relationships between the same subject and the same object on the time dimension form a sequence of static relationships, which are fed into LSTM to generate the visual embedding of a dynamic relationship. And then the class label of the dynamic relationship is determined by finding the most similar language embedding to its visual embedding. Finally, scene graphs are generated using both static and dynamic relationships.

Experiments

Datasets

To evaluate the proposed method, we conduct experiments on two video benchmark datasets, *i.e.*, the VidVRD dataset (Shang et al. 2017) and the VidOR dataset (Shang et al. 2019). With the VidVRD dataset as the target domain, we use the VRD dataset (Lu et al. 2016) as the source image domain. With the VidOR dataset as the target video domain, we use the VG dataset (Zhang et al. 2017) as the source image domain. Therefore, we construct two image-to-video scene graph generation tasks: VRD→VidVRD and VG→VidOR. For the two tasks, we use the objects and their relationships shared by the source and target domains to train and evaluate. The dynamic relationships that only exist in the video domain. For the VRD→VidVRD task, there are 15 object categories and 89 relationship categories (74 static relationship categories and 15 dynamic relationship categories). For the VG→VidOR task, there are 41 object categories and 26 relationship categories (16 static relationship categories and 10 dynamic relationship categories). We adopt the unsupervised domain adaptation protocol, where the training data consists of annotated images from the source domain and unannotated videos from the target domain. The annotations of target videos are only used for evaluation.

Task	VidVRD			VidOR		
	#Img	#Obj	#Rel	#Img	#Obj	#Rel
Video SGG	32160	83865	314340	4970534	16195788	42777103
Ours	1572	4280	6050	22188	74747	14371

Table 1: Numbers of annotations on the VidVRD and VidOR datasets. #Img, #Obj and #Rel denote the numbers of annotated images/video frames, object instances and relationship instances, respectively.

The numbers of annotations of the image-to-video scene graph generation task ("Ours") and the video scene graph generation task ("Video SGG") are shown in Table 1. It is noteworthy that our task requires much fewer annotations, clearly showing it can relieve the heavy dependency on the large-scale annotated videos for training by leveraging existing available images.

Implement Details

Network Architecture. We use Faster R-CNN (Ren et al. 2015) as the object detection model and an MS COCO-pretrained ResNet101 (He et al. 2016) as the backbone of the detection model, following (Xu et al. 2017; Zhang et al. 2019). The shorter side of images and video frames is resized into 600 while preserving its aspect ratio. The dimension of the second-order statistic descriptor is set to 512 and the hyperparameter r in the factorized bilinear pooling is set to 5. The domain classifier D_{img} and the instance domain classifier D_{ins} are designed using five fully-connected layers ($1024 \rightarrow 512 \rightarrow 256 \rightarrow 128 \rightarrow 1$) and three convolution layers ($512 \rightarrow 128 \rightarrow 1$), respectively. The visual mapping ϕ and the language mapping φ consist of three fully-connected layers ($256 \rightarrow 256 \rightarrow 300$) and two fully-connect layers ($1024 \rightarrow 300$), respectively.

Training and Test Details. During training, a three-stage training strategy is employed. First, the object detection model is optimized by the loss function shown in Eq. (4), where gradient reverse layer (Ganin and Lempitsky 2015) is used for hierarchical adversarial training. Second, the

Method	Object Detection mAP	Relationship Detection R@50		R@100		mAP		Relationship Tagging P@1		P@5		P@10	
		sta	dyn	sta	dyn	sta	dyn	sta	dyn	sta	dyn	sta	dyn
VidVRD (Shang et al. 2017)	-	13.35	0.00	14.64	0.00	17.20	0.00	54.03	0.00	34.03	0.00	23.23	0.00
w/o adversarial learning	36.70	4.01	0.00	4.53	0.00	6.73	0.00	30.65	0.00	23.23	5.00	16.45	4.03
w/o image-level adversarial	45.17	7.09	0.00	8.14	0.00	9.89	0.00	48.39	9.09	32.90	4.55	21.96	4.55
w/o instance-level adversarial	41.29	6.16	1.47	7.09	1.47	8.20	0.15	42.74	0.00	32.26	12.88	23.06	12.88
w/o knowledge reasoning	49.40	7.67	0.00	9.36	0.00	12.79	0.00	43.55	0.00	33.23	0.00	23.15	0.00
Ours	49.40	7.67	2.94	9.36	2.94	12.79	0.38	43.55	13.64	33.23	12.58	23.15	12.58
Oracle	-	30.58	4.41	36.74	4.41	36.11	1.36	58.87	18.19	42.58	16.29	30.40	16.29

Table 2: Results on the VidVRD dataset. R@K and P@K are abbreviations of Recall@K and Precision@K, respectively. "sta" and "dyn" denote the static relationship and the dynamic relationship, respectively.

Method	Object Detection mAP	Relationship Detection R@50		R@100		mAP		Relationship Tagging P@1		P@5		P@10	
		sta	dyn	sta	dyn	sta	dyn	sta	dyn	sta	dyn	sta	dyn
w/o adversarial learning	20.42	1.54	0.00	2.22	0.00	1.28	0.00	9.64	0.25	8.83	0.10	8.02	0.06
w/o image-level adversarial	26.52	2.72	0.11	3.69	0.11	2.01	0.03	23.46	7.52	22.44	4.31	17.61	3.76
w/o instance-level adversarial	27.69	2.77	0.04	3.75	0.04	2.30	0.03	25.14	7.77	22.22	4.16	17.17	3.53
w/o knowledge reasoning	28.13	2.84	0.00	3.95	0.00	2.61	0.00	24.02	0.00	21.84	0.00	17.38	0.00
Ours	28.13	2.84	0.21	3.95	0.21	2.61	0.14	24.02	7.77	21.84	5.43	17.38	5.48
Oracle	-	18.55	1.35	25.92	1.35	16.63	0.53	46.93	16.54	38.86	9.77	30.86	8.40

Table 3: Results on the VidOR dataset.

visual-language embedding space is optimized according to Eq. (6). Third, we use greedy association algorithm (Shang et al. 2017) to obtain the visual embeddings of static relationships at the video level by merging detected static relationships at frame level. With the guidance of RCB, we sample the generated static relationships to generate sequential static relationships and train LSTM by Eq. (8). During test, we use non maximum suppression with an IoU threshold of 0.3 to select boxes from object proposals and then take the selected boxes with a confidence score greater than 0.5 as the final detected objects to predicate relationships. Besides, we use language priors of the image domain to further improve predictions following (Zellers et al. 2018).

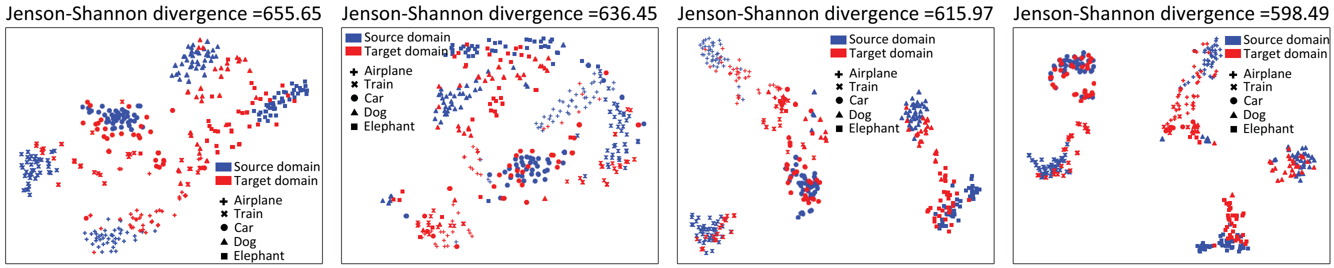
Evaluation Metrics

We utilize three existing evaluation metrics of *object detection*, *relationship detection* and *relationship tagging* to evaluate the performance of the proposed method. Object detection aims to localize objects in each video frame and we adopt mean average precisions (mAP) as the metric of the object detection task. The threshold of mAP is set to 0.5. Relationship detection aims at first detecting objects and then predicting the relationships of detected objects. A detected relationship is considered correct if it has the same relationship triplet in the ground truth and the detected object and subject trajectories have sufficient voluminal intersection over union (vIoU) to those in the ground truth. The threshold of vIoU is set to 0.5, and we adopt mean average precision (mAP) and Recall@K (K equals to 50 and 100) metrics following (Shang et al. 2017). Relationship tagging focuses on only relationship detection in videos. A detected relationship is considered correct if it has the same relationship triplet in the ground truth without taking the object trajectories into account. We adopt Precision@K (K equals to 1, 5, and 10) metrics following (Shang et al. 2017).

Results

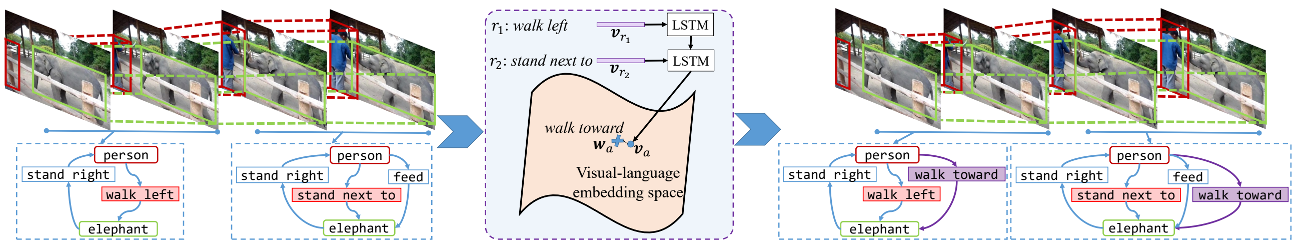
To the best of our knowledge, this is the first work for the new task of image-to-video scene graph generation. So the most related methods to our method are the methods of video scene graph generation that use annotated videos for training. Among these methods, only VidVRD (Shang et al. 2017) releases code on the VidVRD dataset, respectively, so we implement it using our training data on the corresponding dataset for comparison. We also compare our method with several variants (*i.e.,*"w/o adversarial learning","w/o image-level adversarial", "w/o instance-level adversarial", "w/o knowledge reasoning") to demonstrate the effect of each individual component. Since both the relationship detection task and the relationship tagging task are based on the object detection results, we use the ground truth of object detection as the object detection results to evaluate the relationship tasks deeper, denoted as "Oracle".

The comparison results on the VidVRD and VidOR datasets are shown in Table 2 and Table 3, respectively. We have the following observations: 1) in comparison with the VidVRD method, our method performs worse on static relationships due to the unavailable annotations of videos, but achieves better performance on dynamic relationships with 13.64%, 12.58%, and 12.58% gains on P@1, P@5 and P@10, respectively. These promising results show that it is beneficial to exploit external knowledge for inferring dynamic relationships from sequential static relationships; 2) when removing the knowledge reasoning, our method fails to predict dynamic relationships. For example, in the relationship detection task on the VidVRD dataset, "w/o knowledge reasoning" cannot predict any dynamic relationships, while our method achieves 2.94%, 2.94%, and 0.38% on R@50, R@100, and mAP, respectively; 3) the results of "w/o adversarial learning" are far from that of "Ours", clearly demonstrating the existence of domain shift and the



(a) w/o adversarial learning (b) w/o image-level adversarial (c) w/o instance-level adversarial (d) Ours

Figure 3: Object feature visualization on the VRD→VidVRD task. Red and blue colors denote the target video feature and the source image feature, respectively. Different shapes denote different classes as shown in the legend.



(a) w/o knowledge reasoning (b) knowledge reasoning (c) Ours

Figure 4: One example of dynamic relationship prediction from sequential static relationships on the VRD→VidVRD task.

effectiveness of our hierarchical adversarial learning on reducing the domain shift; 4) compared with "w/o image-level adversarial" or "w/o instance-level adversarial", our method achieves better results, showing that both image-level and instance-level adversarial learning benefit improving the performance; 5) "Oracle" achieves better results than supervised video scene graph generation method ("VidVRD"), showing the feasibility of learning a relationship prediction model from existing annotated images with given good object detector and the significance of learning a better cross-domain object detection model.

Feature Visualization

To further analyze the effectiveness of the hierarchical adversarial learning module on reducing the domain shift, we visualize the object features (extracted from RoI pooling of the object detector) of images and video frames learned by "w/o adversarial learning", "w/o image-level adversarial", "w/o instance-level adversarial", and "Ours" using t-SNE (Maaten and Hinton 2008). Due to the large amount of objects, only five object classes are chosen and for each class, 40 instances are randomly sampled from source images and target video frames to show the visualization results of different methods in Figure 3. We also show the Jenson-Shannon divergence of source and target data distributions. The larger the Jenson-Shannon divergence is, the more different the data distributions are. In Figure 3 (a), the data distributions of different domains are quite different, indicating that there is a large domain gap. Compared Figure 3 (d) with others, we can find that the data distribution discrepancy is largely reduced when performing both image-level and instance-level adversarial learning.

Qualitative Evaluation

We illustrate our qualitative results in Figure 4. Our method detects static relationships of "stand right", "walk left", "stand next to" and "feed" well. By the knowledge reasoning module, our method succeeds in inferring the dynamic relationship of "walk toward", guided by the rule {walk toward: walk left→stand next to}. In other words, the LSTM learns the visual embedding of dynamic relationships successfully via transferring commonsense in RCB to the visual-language embedding space, further demonstrating the effectiveness of the knowledge reasoning.

Conclusion

We have presented a new task called image-to-video scene graph generation that leverages annotated images to train a scene graph generation model for videos. This task breaks the heavy dependency on large-scale annotated training videos, making it more approaching to real-world application. To infer dynamic relationships in videos, we have proposed a knowledge reasoning method that can generate visual embedding representations of unseen dynamic relationships for prediction via exploiting commonsense knowledge. To reduce the domain shift between images and videos, we have proposed a hierarchical adversarial learning method that can learn domain-invariant visual features to enable the adaption of objects and static relationships from images to video frames. Extensive experiments on the benchmark dataset have validated the effectiveness of our method.

Acknowledgments

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

References

Cao, D.; Zhao, Q.; and Fu, Y. 2020. Using Spatial Temporal Graph Convolutional Network Dynamic Scene Graph for Video Captioning of Pedestrians Intention. In *Proceedings of the International Conference on Natural Language Processing and Information Retrieval*, 179–183.

Chanda, O.; Teh, E. W.; Rochan, M.; Guo, Z.; and Wang, Y. 2017. Adapting Object Detectors from Images to Weakly Labeled Videos. In *Proceedings of the British Machine Vision Conference (BMVC)*, 56.1–56.12.

Chen, J.; Wu, X.; Duan, L.; and Chen, L. 2021a. Sequential Instance Refinement for Cross-Domain Object Detection in Images. *IEEE Transactions on Image Processing*, 30: 3970–3984.

Chen, J.; Wu, X.; Hu, Y.; and Luo, J. 2021b. Spatial-temporal Causal Inference for Partial Image-to-video Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, online.

Ganin, Y.; and Lempitsky, V. S. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1180–1189.

Gao, Z.; Wu, Y.; Zhang, X.; Dai, J.; Jia, Y.; and Harandi, M. 2020. Revisiting Bilinear Pooling: A Coding Perspective. In *AAAI Conference on Artificial Intelligence (AAAI)*, 3954–3961.

Girdhar, R.; Ramanan, D.; Gupta, A.; Sivic, J.; and Russell, B. 2017. ActionVLAD: Learning spatio-temporal aggregation for action classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3165–3174.

Hao, X.; Zhou, F.; and Li, X. 2020. Scene-Edge GRU for Video Caption. In *Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 1290–1295.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Ji, J.; Krishna, R.; Fei-Fei, L.; and Niebles, J. C. 2020. Action Genome: Actions as Composition of Spatio-temporal Scene Graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10233–10244.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1): 32–73.

Lahiri, A.; Ragireddy, S. C.; Biswas, P.; and Mitra, P. 2019. Unsupervised adversarial visual level domain adaptation for learning video object detectors from images. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1807–1815.

Li, J.; Wong, Y.; Zhao, Q.; and Kankanhalli, M. S. 2017. Attention transfer from web images for video recognition. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 1–9.

Liang, K.; Guo, Y.; Chang, H.; and Chen, X. 2018. Visual relationship detection with deep structural ranking. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 7098–7105.

Liu, C.; Jin, Y.; Xu, K.; Gong, G.; and Mu, Y. 2020. Beyond short-term snippet: Video relation detection with spatio-temporal global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10840–10849.

Liu, X.; and Huet, B. 2016. Event-based cross media question answering. *Multimedia Tools Application (MTA)*, 75(3): 1495–1508.

Liu, Y.; Lu, Z.; Li, J.; Yang, T.; and Yao, C. 2019. Deep image-to-video adaptation and fusion networks for action recognition. *IEEE Transactions on Image Processing*, 29: 3168–3182.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 852–869.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9(Nov): 2579–2605.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2794–2802.

Pataki, M.; Vajna, M.; and Marosi, C. A. 2012. Wikipedia as Text. *ERCIM News*, 2012(89).

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Qian, X.; Zhuang, Y.; Li, Y.; Xiao, S.; Pu, S.; and Xiao, J. 2019. Video relation detection with spatio-temporal graph. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 84–93.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 91–99.

RoyChowdhury, A.; Chakrabarty, P.; Singh, A.; Jin, S.; Jiang, H.; Cao, L.; and Learned-Miller, E. 2019. Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 780–790.

Shang, X.; Di, D.; Xiao, J.; Cao, Y.; Yang, X.; and Chua, T.-S. 2019. Annotating objects and relations in user-generated videos. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR)*, 279–287.

Shang, X.; Ren, T.; Guo, J.; Zhang, H.; and Chua, T.-S. 2017. Video visual relation detection. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 1300–1308.

Su, Z.; Shang, X.; Chen, J.; Jiang, Y.-G.; Qiu, Z.; and Chua, T.-S. 2020. Video Relation Detection via Multiple Hypothesis Association. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 3127–3135.

Tsai, Y.-H. H.; Divvala, S.; Morency, L.-P.; Salakhutdinov, R.; and Farhadi, A. 2019. Video relationship reasoning using gated spatio-temporal energy graph. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10424–10433.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5410–5419.

Xu, N.; Liu, A. A.; Wong, Y.; Zhang, Y.; Nie, W.; Su, Y.; and Kankanhalli, M. 2019. Dual-Stream Recurrent Neural Network for Video Captioning. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 29(8): 2482–2493.

Yu, F.; Wu, X.; Chen, J.; and Duan, L. 2019. Exploiting images for video recognition: Heterogeneous feature augmentation via symmetric adversarial learning. *IEEE Transactions on Image Processing*, 28(11): 5308–5321.

Yu, F.; Wu, X.; Sun, Y.; and Duan, L. 2018. Exploiting images for video recognition with hierarchical generative adversarial networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 1107–1113.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5831–5840.

Zhang, H.; Kyaw, Z.; Chang, S.-F.; and Chua, T.-S. 2017. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5532–5540.

Zhang, J.; Kalantidis, Y.; Rohrbach, M.; Paluri, M.; Elgammal, A.; and Elhoseiny, M. 2019. Large-scale visual relationship understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 9185–9194.