DDS: Decoupled Dynamic Scene-Graph Generation Network

A S M Iftekhar[*‡], Raphael Ruschel[*‡], Satish Kumar[‡], Suya You[§], B. S. Manjunath[‡]

[‡] University of California Santa Barbara

[§] Army Research Laboratory, Intelligent Perception, CISD, Los Angeles, CA

`(iftekhar, raphael251, satishkumar, manj)@ucsb.edu,`
`suya.you.civ@army.mil`

Abstract

Scene-graph generation involves creating a structural representation of the relationships between objects in a scene by predicting subject-object-relation triplets from input data. Existing methods show poor performance in detecting triplets outside of a predefined set, primarily due to their reliance on dependent feature learning. To address this issue we propose DDS– a decoupled dynamic scene-graph generation network– that consists of two independent branches that can disentangle extracted features. The key innovation of the current paper is the decoupling of the features representing the relationships from those of the objects, which enables the detection of novel object-relationship combinations. The DDS model is evaluated on three datasets and outperforms previous methods by a significant margin, especially in detecting previously unseen triplets.

1. Introduction

Dynamic Scene-Graph (DSG) provides a graph structure presenting the relationships among different objects in a scene by predicting relationship triplets composed of $\langle subject, object, relationship \rangle$ at each frame of an input video. This acts as a foundational block for various computer vision tasks [30, 46]. Current DSG generation systems [5, 15, 24] have poor performance when facing a triplet that was not present during training, even though the individual components have been seen. However, in a more realistic deployment scenario, the network will likely encounter triplets that it has not seen before. Therefore, a system should be able to transfer the learned concepts of relationships and objects to compose unseen triplets. This poor performance is mainly attributed to the learning of highly dependent feature representations of relationships and objects. The proposed decoupled dynamic scene-graph (DDS)

*represents equal contribution
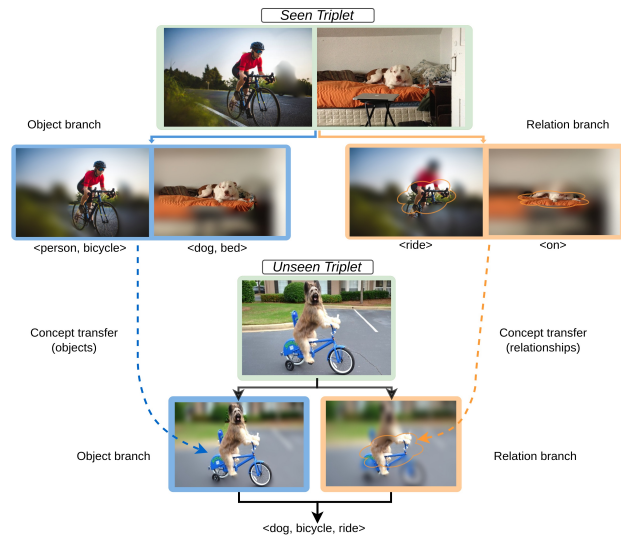Iftekhar is currently with the Responsible AI team, Microsoft

Figure 1. Diagram to show the concept learning and transferring in DDS. By focusing on different spatial regions, DDS learns the concept of relationships (ride, on) and objects (person, bicycle, bed) independently.

addresses this issue.

Fig. 1 shows the core idea of the proposed DDS network. This architecture utilizes two different branches to learn decoupled features for relationships and objects. As shown in the figure, DDS learns the concept of 'ride', 'on', 'person', 'bicycle', and 'bed' from the training examples of a 'person riding a bike' and a 'dog on the bed' independently. The decoupled design makes DDS look into different spatial regions for relationships and objects. These learned concepts are transferred to successfully detect the unseen triplet $\langle dog, bicycle, ride \rangle$.

DDS ensures the learning of discriminative spatio-temporal cues for relationships and objects. Fig. 2 shows the overview of our architecture. It consists of two separate branches: the relation and the object branch. We chose a transformer-based encoder-decoder [2] architecture for these branches with two different sets of queries. More-

over, a novel temporal decoder is added to embed temporal information into the queries. These separate sets of queries focus on learning generalized representations for relationships and objects from differently encoded feature maps in both temporal and spatial domains. This is significantly better than the existing works, where the same object features are used for both object and relationship detection. Also, DDS does not have the dependence on off-the-shelf object detectors like previous works.

Our proposed model is thoroughly evaluated on the Action-Genome [16] dataset for DSG generation, where it achieves significant performance gains compared to the SOTA models. Additionally, we evaluate DDS on the task of static scene-graph (SSG) generation on the HICO-DET [3] dataset and unusual SSG generation on the UnRel [37] dataset, where DDS outperforms all the existing models in both datasets. Finally, the proposed design choices are evaluated in an extensive ablation study.

2. Related Works

DDS is built on the previously developed works in SSG and DSG generation. This section is used to review the literature in the mentioned areas along with additional relevant publications on scene-graph generation under the compositional setting.

2.1. Static Scene-Graph (SSG) Generation

SSG generation is proposed for the task of image retrieval. The initial works rely heavily on two-stage (object detection and then scene-graph generation) structures. Also, many authors utilize prior knowledge [13,31,45] (e.g. semantic knowledge, statistical heuristics) for SSG generation. Despite recent improvements in SSG generation, these methods are heavily constrained by their reliance on the object detection quality as noted in [12].

Modern works in SSG generation focus on utilizing a one-stage Transformer based architecture to deal with the aforementioned issues. These architectures rely on set-based predictions to generate SSG. Among these works, Qpic [42] uses a single encoder-decoder model while CDN [51] extends Qpic by using sequential decoding of objects and relationships. Additionally, MSTR [20] enables the use of multi-scale feature maps to these networks. Another concurrent work, SSRT [12] refines the overall architecture with spatial and semantic support features. Moreover, a recent line of research heavily exploits the usage of very large-scale semantic knowledge engines (e.g. CLIP [39]) [12,26,38]. Apart from the obvious limitation of these works being unable to utilize temporal dependencies, they perform poorly while detecting unseen triplets. With the decoupled multi-branch design, we significantly differ from

these works by using separate sets of queries for relationship and object detection.

2.2. Dynamic Scene-Graph (DSG) Generation

DSG is an extension of the SSG where the scene-graph is created for videos. This process is harder since temporal cues need to be utilized [5, 15, 24]. Current works in this area have two-stage architectures following the initial works on SSG. Among these works, STTran [5] utilizes a temporal decoder based on the self-attention mechanism. DSGAP [24] expands STTran with an anticipatory pre-training paradigm. On the other hand, HORT [15] utilizes a multi-branch design with different types of Transformers. Both STTran and HORT use similar features for relationship and object detection. These features come from the object bounding boxes predicted by off-the-shelf object detectors. However, using similar features for relationship and object detection forces the learning of relationships and objects to be dependent on each other.

2.3. Compositionality in Scene-Graph Generation

Creating new compositions from base known concepts during inference is known as compositional zero-shot learning (CZSL) under the compositional setting [22, 34]. In this paper, we utilize this setting to evaluate our model. Kato et al. [17] introduce CZSL in SSG generation with an embedding-based model. Many following works [9–11] adapt different object-affordance ideas. These works assume there exist common relationships between the subjects and the objects. This work does not have such limited assumptions, and as a result, can generate scene graphs even when the relationships are very unusual (See Table 4).

3. Method

We propose a multi-branch network to address the challenge of dependent feature learning in DSG generation, which decreases the performance of detection of unseen relationship triplets in current models. The network learns distinct feature representations for relationships and objects. Before detailing the architecture, we first formulate the problem.

3.1. Problem Formulation

Given an input video, $\mathbf{V} = \{\mathbf{I}_1, \mathbf{I}_2, .., \mathbf{I}_t, .., \mathbf{I}_T\}$ with T frames, the task in DSG generation is to predict a set of relationship triplets, $\{R_1, R_2, .., R_t, .., R_M\}$ at every frame of the video. Every frame has N_M number of relationship triplets. Each relationship triplet can be presented by $\langle s, o, r_{so} \rangle$. Here, s, o refers to subject, object and are represented by bounding boxes and category labels. r_{so} is the relationship between s and o. In a single frame $\mathbf{I}_t$, s and o can have multiple relations, as shown in the sample input-output pair in Figure 2.
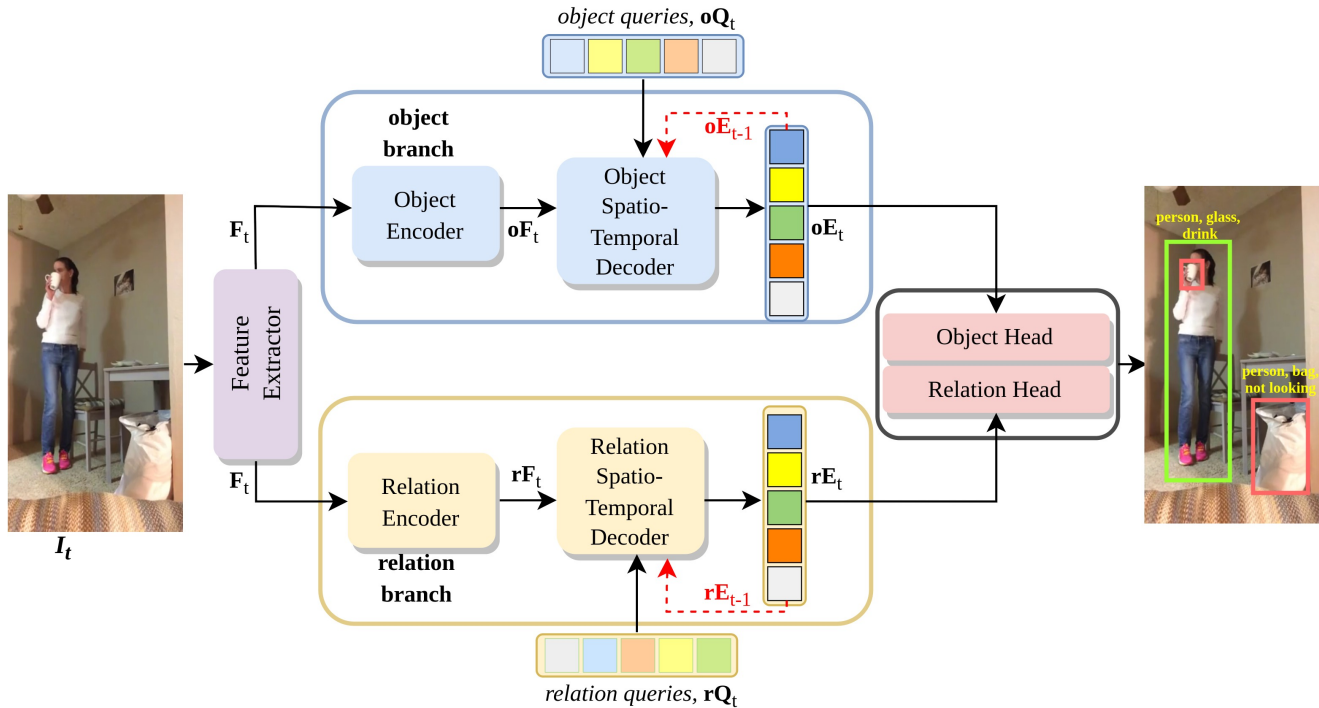
Figure 2. Overview of DDS's architecture. Given an input frame $\mathbf{I}_t$, features are extracted by the backbone and fed to decoupled object and relation branches, each with an encoder and spatio-temporal decoder. The decoders process queries and previous frame embeddings (red arrow) to produce learned embeddings, which are used by the object and relation heads to predict relationship triplets.

In this paper, the main goal is to predict relationship triplets under the compositional setting. In this setting, the networks see all subjects, objects, and relationships during training but not all combinations available, leaving a portion to be exclusive only to the test set so we can evaluate the performance on unseen triplets.

3.2. Technical Overview

The proposed work adopts a one-stage approach for DSG generation compared to the current two-stage methods [5, 15, 24] as the former [4, 19, 42, 54] have shown impressive performance in creating SSG. However, these image-based works present poor generalization capabilities. Therefore, we propose a network that uses a different set of queries with two branches, where each branch follows a Transformer-like encoder-decoder architecture, similar to [21, 23]. Fig. 2 shows a diagram of the model, where a convolutional neural network (CNN) extracts features from the input frame, and those are encoded differently by the object and the relation encoders. Each spatio-temporal decoder takes encoded features from their respective encoder in addition to two types of inputs: queries for the current frame ($\mathbf{oQ}_t$, $\mathbf{rQ}_t$) and the embeddings ($\mathbf{oE}_{t-1}$, $\mathbf{rE}_{t-1}$) propagated from the previous frame. As the encoded features differ for each branch, the queries learn decoupled features for relationships and objects. The decoder outputs are the

learned object and relation spatio-temporal embeddings and are used as input to the relation and the object heads for final predictions.

3.3. Feature Extraction & Encoders

Consider a frame $\mathbf{I}_t \in \mathbb{R}^{N_C \times H \times W}$ at time t of the input video V. A CNN network is used as backbone, $\mathbf{B}$ (e.g. resnet-50 [7]) to extract features that are then reduced using 1×1 convolutions, flattened, and added positional embeddings to get the feature map $\mathbf{F}_t \in \mathbb{R}^{(H'W') \times d'}$ that is used as a common feature for both the relation and the object branches.

Both of the network's branches have an encoder comprising of stacked multi-head self-attention layers [47] with a feed-forward network (FFN). The output of the encoders are two separate feature maps, $\mathbf{rF}_t$, and $\mathbf{oF}_t$, which represent the features from the relationship and object branch, respectively.

3.4. Input Queries

The DDS framework consists of 2 independent branches with each one containing one decoder. Each decoder takes a set of randomly initialized, trainable queries $\mathbf{rQ}_t$ and $\mathbf{oQ}_t$ for the relationship and object branch, respectively, and process them through a dual-stage process. This process, designed to explore spatio-temporal features, begins

with the temporal decoders aggregating information across frames, followed by spatial decoders enriching this information with the current frame context.

3.5. Spatio-Temporal Decoders

After encoding the features using a standard transformer encoder, DDS uses a two-stage decoding process, where each decoder consists of two small components: temporal and spatial decoder, which are shown in Fig. 3 for reference. The proposed multi-branch design ensures discriminative feature learning for the queries of each branch.

Temporal Decoders: These decoders allow queries to leverage temporal dependencies. Each temporal decoder takes two sets as inputs: the current frame's queries and the embeddings from the previous frames. For frame $\mathbf{I}_t$, the current frame's relation and object queries sets are defined as $\mathbf{rQ}_t \in \mathbb{R}^{N_q \times d}$ and $\mathbf{oQ}_t \in \mathbb{R}^{N_q \times d}$. Every query is a d dimensional vector, and every branch has N_q number of queries. Embeddings from the previous frames for the relation and the object branches are presented by $\mathbf{rE}_{t-1} \in \mathbb{R}^{N_q \times d}$ and $\mathbf{oE}_{t-1} \in \mathbb{R}^{N_q \times d}$, and are marked with a red arrow in Fig.3.

The cross-attention in the temporal decoders allows the current frame's queries to select what to learn from the previous frame's embeddings. The outputs of the temporal decoders are the temporally aggregated queries $\mathbf{rA}_t$, $\mathbf{oA}_t$. They are fed to their respective spatial decoders. In the case of the first frame of a video, the temporal decoders directly output $\mathbf{rQ}_t$ and $\mathbf{oQ}_t$ as $\mathbf{rA}_t$ and $\mathbf{oA}_t$ without passing them through the cross-attention and FFN blocks as there is no previous frame.

Spatial Decoders: The spatial decoders architecture is similar to the standard Transformer decoder [2]. These decoders consist of both self-attention and cross-attention layers along with FFN networks. Each decoder takes encoded feature maps ($\mathbf{rF}_t$ or $\mathbf{oF}_t$) along with the aggregated queries of the temporal decoders ($\mathbf{rA}_t$ or $\mathbf{oA}_t$) from their respective branch as inputs. Also, these decoders take learnable positional embeddings, $\mathbf{rPE}$, and $\mathbf{oPE}$, for the relationship and object decoder, respectively. These embeddings for the relation and the object branch are defined as $\mathbf{rPE} \in \mathbb{R}^{N_q \times d}$, $\mathbf{oPE} \in \mathbb{R}^{N_q \times d}$.

$$\mathbf{rE}_t = \text{Relation Decoder}(\mathbf{rN}_t, \mathbf{rPE}, \mathbf{rQ}_t) \quad (1)$$
$$\mathbf{oE}_t = \text{Object Decoder}(\mathbf{oN}_t, \mathbf{oPE}, \mathbf{oQ}_t) \quad (2)$$

The outputs of the decoders are the learned spatio-temporal embeddings. They are used in the object and the relation heads to make the final relationship triplet predictions.

3.6. Object & Relation Heads

The output embeddings from the object spatio-temporal decoder, $\mathbf{oE}_t$ are fed to four different FFNs. For input frame $\mathbf{I}_t$, these FFNs predict subject bounding boxes, $\mathbf{sB}_t \in [0,1]^{N_q \times 4}$, object bounding boxes, $\mathbf{oB}_t \in [0,1]^{N_q \times 4}$, subject prediction vectors, $\mathbf{sP}_t \in [0,1]^{N_q \times O}$, and object prediction vectors, $\mathbf{oP}_t \in [0,1]^{N_q \times N_o}$. Here, N_q is the number of queries, and N_o is the total number of objects. Similarly, $\mathbf{rE}_t$, are fed to two FFNs that produce as output the relation prediction vectors, $\mathbf{rP}_t[0,1]^{N_q \times N_r}$ and relation region bounding boxes $\mathbf{rB}_t \in [0,1]^{N_q \times 4}$. Notice that the relation region bounding box is defined as the union between the subject and object bounding boxes, and is used solely during training.

3.7. Inference

During inference, we take the outputs from the object and relation heads as in 3.6 and compose N_q relationship pairs by leveraging the maximum confidence score in $\mathbf{sP}_t$ and $\mathbf{oP}_t$. The maximum confidence score is used to create $\mathbf{sP}_{tmax} \in [0,1]^{N_q}$ and $\mathbf{oP}_{tmax} \in [0,1]^{N_q}$ and the corresponding index is used to determine the category label for each of the bounding boxes. For every composed relationship pair, the final relation score is calculated as a multiplication of each category's confidence score

3.8. Training

For training DDS, we utilize losses similar to Qpic [42]. This loss calculation implicitly binds the two sets of queries from the relation and the object branch. The loss calculation happens in two stages:

In the first stage, we find the bipartite matching between the predictions $\mathbf{P}_t$ for frame $\mathbf{I}_t$ and the ground truths $\mathbf{G}_t$. One important detail to note here is that there are three kinds of ground truth bounding boxes: subject bounding boxes, object bounding boxes, and relation regions. Ground truth relation regions refer to the union bounding boxes between the subject and the object bounding boxes for the current interaction and are only used during the training phase. The matching cost metric is defined as:

$$\mathbf{C}^{(i,j)} = \eta_b(\mathbf{C}_{sb}^{(i,j)} + \mathbf{C}_{ob}^{(i,j)} + \mathbf{C}_{rb}^{(i,j)}) + \eta_o \mathbf{C}_o^{(i,j)} + \eta_r \mathbf{C}_r^{(i,j)} \quad (3)$$

Where $\mathbf{C}_{sb}^{(i,j)}, \mathbf{C}_{ob}^{(i,j)}, \mathbf{C}_{rb}^{(i,j)}$ are the subject bounding box, the object bounding box and the relation region matching costs, $\mathbf{C}_o^{(i,j)}$ is the object label matching cost, and $\mathbf{C}_r^{(i,j)}$ is the relation label matching cost between i^{th} element from $\mathbf{P}_t$ and j^{th} element from $\mathbf{G}_t$. These costs are calculated following [42]. η_b, η_o, η_r are fixed hyper-parameters. The Hungarian matching algorithm [2] is used to find the optimal matching between the predictions and the ground truths by using these cost metrics. After this matching, every prediction is associated with a ground truth. Next, the following loss is calculated for training the network:

$$\mathcal{L} = \lambda_g \mathcal{L}_{GIOU} + \lambda_l \mathcal{L}_{L1} + \lambda_o \mathcal{L}_{obj} + \lambda_r \mathcal{L}_{rel}, \quad (4)$$
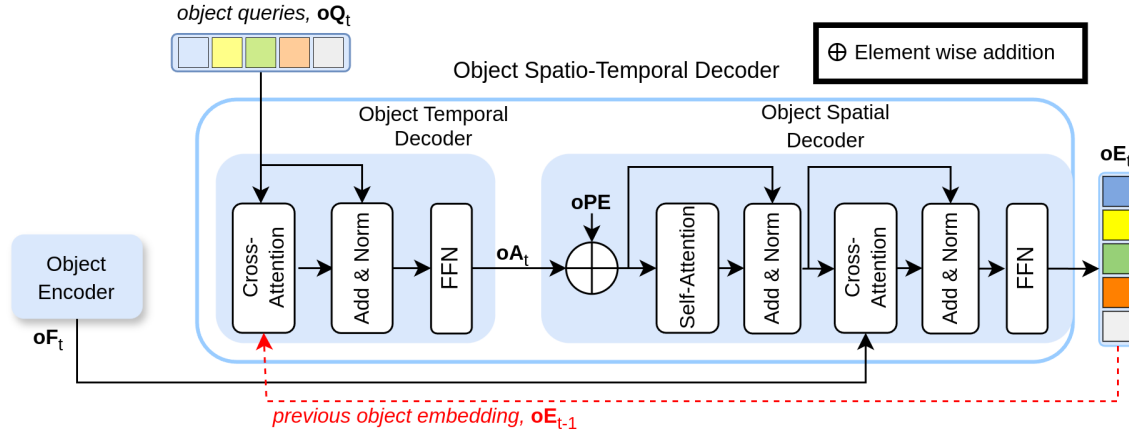
Figure 3. Design of the spatio-temporal decoders using the Object decoder as an example. The relationship decoder uses the same architecture, however, with its corresponding inputs adjusted.

Here, $\mathcal{L}_{GIOU}$ and $\mathcal{L}_{L1}$ are the generalized intersection over union (gIOU) and L1 box regression losses for the predicted subject bounding boxes, object bounding boxes, and relation regions. $\mathcal{L}_{obj}$ is the cross-entropy loss for subject and object label predictions. $\mathcal{L}_{rel}$ is the binary cross-entropy loss for the relationship label predictions. λ_o, λ_g, λ_l, and λ_r are the corresponding hyper-parameters.

Notice that a portion of the datasets [3, 16] fixes the subject as humans. In this case, the subject prediction vectors and subject bounding boxes are not used for loss calculation.

4. Experiments

4.1. Experimental Setup

To evaluate DDS's performance on video data, we use the Action Genome (AG) [16] dataset. Moreover, we show our model's performance in SSG generation datasets: HICO-DET [3] and UnRel [37]. As these datasets only contain images, each sample is treated as a single-frame video, and thus the only difference is that the feedback mechanism will not be used. Next, the used datasets will be presented in more detail:

Action Genome (AG) [16]: This dataset is built on top of the Charades [41] dataset, provides frame-level annotations, and is extensively used in the literature for DSG generation. It has 36 distinct object classes and 25 relationship classes. The object classes are common household items such as doors, windows, and cups, and have a total (train and test set) of $476,229$ bounding boxes. In total, AG provides $1,715,568$ instances of the mentioned classes contained in $135,484$ subject-object pairs. Every subject-object pair can have multiple relations. Also, on AG the subject class is always human.

Originally, the AG dataset provided $7,464$ videos with

$166,785$ frames in the training set and $1,737$ videos with $54,371$ frames in the test set. The original training set contains 530 relationship triplets. All these relationship triplets are present in the test set. We refer to this setting as the fully-supervised setting. As the main interest in this paper is to evaluate DDS's performance in the compositional setting, a new training split of the data is created. This new proposed training set contains $6,784$ videos with $146,517$ frames containing 421 relationship triplets. The original test set is not changed. It contains 499 object-relationship, where 80 of them are not present in our new training set.

HICO-Det [3]: This dataset has 80 objects and 117 relationship classes. In the literature, this dataset is used for evaluating SSG generation performance under compositional setting [8–11]. DDS's performance is reported following the RF (Rare First) protocol provided by [8]. This protocol has $37,328$ images in the training set with 480 relationship triplets. The test set has $9,552$ images with 480 seen relationship triplets and 120 unseen relationship triplets.

UnRel [37]: This dataset provides extremely unusual SSG triplets, for example: $(\langle elephant, bike, riding\rangle)$. It has 4000 training and 1000 test images with 100 objects and 70 relationships. The original train/test split provided by the authors already provides a compositional setting where the test set has 65 unseen relationship triplets. The training set contains 4000 seen relationship triplets.

4.2. Evaluation Metrics

Following existing works [5, 15, 16, 24], we report our performances in AG dataset with Recall@K metric with $K = [20, 50]$. We utilize only the SGDet [16] protocol to report our performances, as PREDCLS and SGPRED are not applicable to one-stage methods. In this protocol, the network needs to detect relationship triplets along with sub-

ject and object bounding boxes. We carry out experiments on two different scenarios: *With constraint*: Each query is only allowed to predict a single triplet, and *No constraint*: Where each *subject-object* pair can generate multiple verb predictions. Given that AG presents a long-tailed distribution over the relationship classes, recall alone might result in biased results if the model is favorable towards the most frequent classes. To provide a more complete analysis, we also present the mean Recall@k (mR@K), where the recall of each predicate is measured and then averaged over the number of classes. Moreover, mAP (mean average precision) was selected to report the performances in UnRel and HICO-Det datasets similar to current works [9, 10, 36, 37]. Here, performances are reported in three categories: unseen (only unseen relationship triplets), seen (only seen relationship triplets), and full (all relationship triplets) [8]. For all datasets, a prediction triplet from DDS is considered correct if subject and object bounding boxes have at least 0.5 Intersection over Union (IoU) with ground truth bounding boxes, and subject, object, and relationship labels match with ground truth labels.

4.3. Implementation details

ResNet-50 [7] is used as the CNN backbone. Both temporal decoders inside the spatio-temporal decoders have a single layer. We follow Qpic's [42] setup for the encoders. We select 6 layers for the object spatial decoder with 3 layers for the relation spatial decoder. All loss coefficients in equation 3 and equation 4 are set as [42]. The number of queries in each branch is 64. Each query is a vector of size 256. The model is trained with AdamW [14] optimizer and uses the training strategy in [40] for efficiently handling video sequences. We initialize the parameters of DDS from DETR [2] trained on COCO [28] object detection dataset. The initial learning rate for the backbone network is 10^{-6} and for the other part of the network is 10^{-5}. All evaluations are done following the codebase provided by [43].

When training in the AG dataset, we drop the learning rate by 10 times at every 40 epochs and utilize a batch size of 128. DDS processes each video frame from a single video sequentially. We utilize scale augmentation like [2]. Input frames are resized with the shortest side being at least 480 and at most 800 pixels, and longest side is at most 800.

In the other datasets [3, 37], the learning rate is dropped by 10 times at every 60 epochs with a batch size of 16. We use a scale augmentation scheme similar to the one used for AG, except that the longest side of the resized image is chosen as 1333. The training schedule is selected based on the convergence of losses. Upon acceptance, we will publicly release our trained models and code.

5. Results & Analysis

5.1. Comparison with the SOTA models

In the AG [16] dataset, we report DDS's performances in Table 2 under the compositional setting. In this setting, there are $\sim 12\%$ less training data with 80 unseen relationship triplets. We retrain STTRan [5] in the mentioned setting for comparison. It is important to note here, that among the three recent DSG generation models [5, 15, 24], only STTran's code is publicly available, therefore limiting the capacity to evaluate other models. DDS outperforms STTran in all recall levels in both unseen and seen relationship triplet detection. Especially, for detecting unseen triplets DDS achieves $4 - 24$ times improvement over the SOTA model. It shows the generalization power of DDS.

Additionally, for a fair comparison with other models, we train DDS in the full training set of AG dataset and report performance in Table 1. With similar training data, DDS achieves SOTA performance on all metrics, especially at R@10, thus requiring fewer predictions to capture the actions in the scene. Among the compared works, HORT [15] has a mutlti-branch Transformer based architecture which is similar to DDS on a high level. However, the primary innovation of DDS lies in its ability to learn object and relationship features separately. Since HORT has not made their code publicly available, a direct comparison of performance in detecting unseen triplets is not feasible. However, considering the significant improvements demonstrated by our method, as shown in Table 1 — an average recall enhancement of 7% in the fully supervised settings — underscoring its efficacy. We anticipate that this performance gap would be even higher in the context of detecting unseen triplets. Similarly, most recent papers as [29], [33], and [25] are not designed with the end goal of predicting unseen triplets. Given that our method outperforms them by a significant margin using the full data of AG, we would expect a significant performance gap in the compositional setting as well, however, a comparison is again limited due to the lack of a publicly available implementation of their approach.

In Table 3 and Table 4, DDS outperforms all existing methods in HICO-Det [3] and UnRel [37] datasets. In summary, DDS outperforms existing works in both seen and unseen SSG generation (full category) in HICO-Det by 5% and in UnRel by 33%. It is worth noting that different from the works in [1, 27, 27, 52], we use the mentioned datasets to assess only our performance under the compositional setting to evaluate DDS' capability of predicting unseen triplets.

5.2. Ablation Studies

We perform ablations for different design choices of our network in this section. Other than the temporal decoder

| | SGDet | | | | | | | | | | | |
| | With constraint | | | | | | No constraint | | | | | |
Method	mR@10	mR@20	mR@50	R@10	R@20	R@50	mR@10	mR@20	mR@50	R@10	R@20	R@50
ReIDN [53]	3.3	3.3	3.3	9.1	9.1	9.1	7.5	18.8	33.7	13.6	23	36.6
TRACE [44]	8.2	8.2	8.2	13.9	14.5	14.5	22.8	31.3	41.8	26.5	35.6	45.3
iSGG [18]	-	19.7	22.9	-	29.2	35.3	-	-	-	-	-	-
STTran [5]	16.6	20.8	22.2	25.2	29.1	37	20.9	29.7	39.2	24.6	36.2	48.8
STTran-TPI [50]	15.6	20.2	21.8	26.2	29.1	34.6	-	-	-	25.7	37.9	50.1
APT [24]	-	-	-	26.3	29.1	38.3	-	-	-	-	-	-
TEMPURA [35]	18.5	22.6	23.7	28.1	33.4	34.9	24.7	33.9	43.7	29.8	38.1	46.4
VsCGG [33]	18.7	-	24.2	27.4	35.8	38.2	24.3	33.1	42.8	29.3	40.2	48.9
TD2-Net (p) [29]	20.4	-	23	26.1	28.7	37.1	27.9	33	46.3	30.5	-	49.3
DSG-DETR [6]	-	-	-	30.3	34.8	36.1	-	-	-	32.1	40.9	48.3
OED [48]	-	-	-	33.5	40.9	**48.9**	-	-	-	35.3	**44**	**51.8**
DDS (ours)	**24.5**	**29.1**	**32.2**	**36.2**	**42.0**	47.3	**32.9**	**41.3**	**48.7**	**37.3**	43.3	51.5

Table 1. DDS's performance comparison in AG test set. Here, like other models, DDS is trained in the full training set of AG dataset. The best results are shown in **bold**. For the other models, all the reported numbers are taken from the original publications.

| | Seen | | Unseen | |
	R@20	R@50	R@20	R@50
STTran [5]*	33.7	36.6	0.3	4.4
DDS (Ours)	**41.8**	**48.8**	**7.4**	**18.2**

Table 2. DDS's performance comparison in AG test set under the compositional setting. Both reported models are trained on the proposed small-size training set under the compositional setting. * means the model was trained using publicly available code. Among recent DSG generation models, only STTran's [5] code is publicly available. The best results are shown in **bold**.

Method	Unseen (mAP)	Seen (mAP)	Full (mAP)
VCL [8]	10.1	24.3	21.4
ATL [9]	9.2	24.7	21.6
FCL [10]	13.2	24.2	22.0
THID [49]	15.5	24.3	23.0
SCL [11]	19.1	30.4	28.1
DDS (Ours)	**21.1**	**31.7**	**29.6**

Table 3. DDS's performance comparison in HICO-Det test set under RF (Rare-First) compositional setting. The best results are shown in **bold**.

Method	Unseen (mAP)	Seen (mAP)	Full (mAP)
VRD [32]	-	-	7.2
WSL [37]	-	-	9.9
DUV [36]	-	-	13.4
DDS (Ours)	**16.3**	**27.4**	**17.9**

Table 4. DDS's performance comparison in UnRel test set. The best results are shown in **bold**. Results not reported are marked with a '-'.

ablation, we report our performances in the HICO-Det [3] dataset. We utilize AG dataset for the temporal decoder ablation. The complete tables with the results of each ablation study are presented in the supplemental material.

Multi-branch Design: We first validate the decoupled multi-branch design and present the results in 5. The single-branch network, which uses a shared set of queries for both object and relationship detection, performs poorly, especially on unseen categories. Adding two spatio-temporal decoders improves performance, and further introducing two separate encoders and relation region prediction yields significant gains. This demonstrates the benefit of decoupled learning.

Type	Relation Region	Separate Encoders	Separate Decoders	Unseen (mAP)	Seen (mAP)
Base network	✗	✗	✗	17.9	29.9
Multi branch	✗	✗	✓	18.7	30.5
	✗	✓	✓	19.7	31.6
	✓	✓	✓	**21.1**	**31.7**

Table 5. Impact of different components on our decoupled multi-branch design.

Relation Region Ground Truths: During training, the relation branch requires ground truth relation regions. We experiment with two approaches: (1) Mixture, where the relation region is the intersection of subject and object boxes if IoU exceeds θ, otherwise their union, and (2) Union box, where the union of subject and object boxes defines the relation region. We noticed that performance is correlated with the value of θ and the best value is the union box, which corresponds to $\theta = 1$, likely due to its inclusion of spatial information for non-contact relationships (e.g., subject looks at object).

Share Queries: We test two query-sharing strategies between the object and relation branches: (1) o to r (object to relation) and (2) r to o (relation to object). DDS performs best without shared queries, especially for unseen triplets, reinforcing the importance of decoupled queries. Sharing
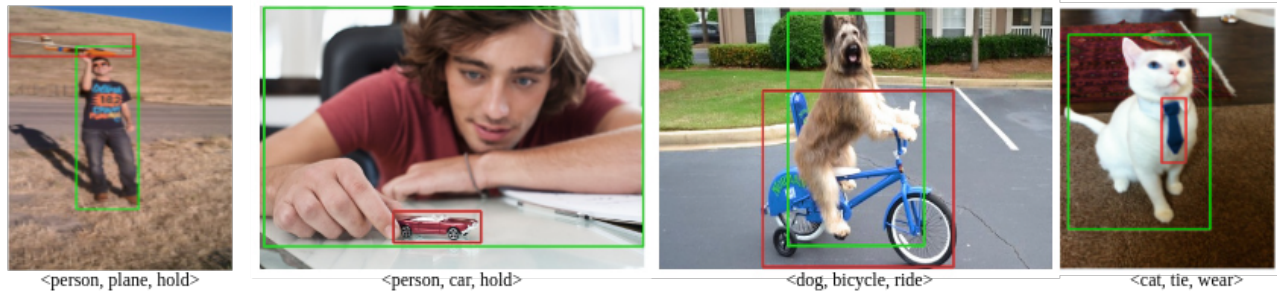
Figure 4. Qualitative results of DDS for predicting unusual relationship triplets in UnRel [37] dataset. The subject bounding box is green and the object bounding box is red.

queries, particularly from relation to object, reduces performance, indicating that object queries generalize better.

Spatial Decoders: We evaluate different numbers of layers for the spatial decoders. Performance suffers with too few or too many layers, with the best results when the object decoder has more layers than the relation decoder, reflecting the need for more capacity when decoding two entities (subject and object).

Temporal Decoders: Temporal decoders play a crucial role in detecting relationship triplets. Without them, performance drops by about 3% recall, confirming that temporal dependencies across frames are key for accurate triplet predictions.

5.3. Qualitative Results

This section compares DDS's performance with our base network (details in section 5.2). This comparison is made using the UnRel [37] dataset. Fig. 4 illustrates the samples where DDS is successful; however, our base network predicted bounding boxes do not match with ground truth. More results can be seen in the supplemental materials.

We compare the attention maps from DDS and the base network to further analyze our improved performances. Fig. 5 presents attention maps for the samples where both DDS and the base network have correct predictions. The attention maps are of the queries that predict the marked subject and object bounding boxes from the last layer of the spatio-temporal decoders. We overlap attention maps from both our spatio-temporal decoders to get the final attention map. As can be seen, although both networks have correct predictions, DDS's attention maps cover the correct spatial region compared to spotty locations produced by the base network.

6. Conclusion

This paper proposes DDS, a multi-branch decoupled network for DSG generation. it is comprised of two encoder-decoder based Transformer branches. This design enables independent learning of objects and relationships, thus im-
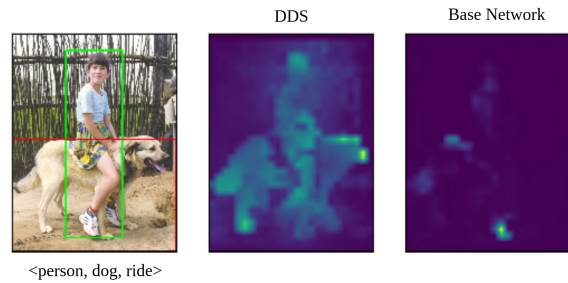

Figure 5. Performance analysis of DDS over the base network. The attention maps are visualized from the last layer of the spatio-temporal decoder.

proving the performance when detecting unseen relationship triplets. The effectiveness of DDS is demonstrated through extensive experiments where it achieves SOTA performance on three benchmark datasets. Moreover, the conducted ablation studies have provided the motivation and significance for different components of DDS. However, while successful compared to the existing works, the quantitative results show room for improvement in detecting unseen relationship triplets. Future research will focus on improving DDS for a better generalized DSG generation.

Acknowledgments

References

[1] Yichao Cao, Qingfei Tang, Feng Yang, Xiu Su, Shan You, Xiaobo Lu, and Chang Xu. Re-mine, learn and reason: Exploring the cross-modal semantic correlations for language-guided hoi detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23492–23503, 2023. 6

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 4, 6

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 2, 5, 6, 7

[4] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021. 3

[5] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16372–16382, 2021. 1, 2, 3, 5, 6, 7

[6] Shengyu Feng, Hesham Mostafa, Marcel Nassar, Somdeb Majumdar, and Subarna Tripathi. Exploiting long-term dependencies for generating dynamic scene graphs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5130–5139, 2023. 7

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 6

[8] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020. 5, 6, 7

[9] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, 2021. 2, 5, 6, 7

[10] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14646–14655, 2021. 2, 5, 6, 7

[11] Zhi Hou, Baosheng Yu, and Dacheng Tao. Discovering human-object interaction concepts via self-compositional learning. In *ECCV*, 2022. 2, 5, 7

[12] ASM Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, and Davide Modolo. What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5353–5363, 2022. 2

[13] ASM Iftekhar, Satish Kumar, R Austin McEver, Suya You, and BS Manjunath. Gtnet: Guided transformer network for detecting human-object interactions. *arXiv preprint arXiv:2108.00596*, 2021. 2

[14] Loshchilov Ilya, Hutter Frank, et al. Decoupled weight decay regularization. *Proceedings of ICLR*, 2019. 6

[15] Jingwei Ji, Rishi Desai, and Juan Carlos Niebles. Detecting human-object relationships in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8106–8116, 2021. 1, 2, 3, 5, 6

[16] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 2, 5, 6

[17] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–251, 2018. 2

[18] Siddhesh Khandelwal and Leonid Sigal. Iterative scene graph generation. *Advances in Neural Information Processing Systems*, 35:24295–24308, 2022. 7

[19] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83, 2021. 3

[20] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19578–19587, 2022. 2

[21] Satish Kumar, Ivan Arevalo, ASM Iftekhar, and BS Manjunath. Methanemapper: Spectral absorption aware hyperspectral transformer for methane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17609–17618, 2023. 3

[22] Satish Kumar, ASM Iftekhar, Ekta Prashnani, and BS Manjunath. Locl: Learning object-attribute composition using localization. *arXiv preprint arXiv:2210.03780*, 2022. 2

[23] Satish Kumar, Bowen Zhang, Chandrakanth Gudavalli, Connor Levenson, Lacey Hughey, Jared A Stabach, Irene Amoke, Gordon Ojwang, Joseph Mukeka, Stephen Mwiu, et al. Wildlifemapper: Aerial image analysis for multi-species detection and identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12594–12604, 2024. 3

[24] Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13874–13883, 2022. 1, 2, 3, 5, 6, 7

[25] Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pre-training. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13864–13873, 2022. 6

[26] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance

interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20123–20132, 2022. 2

[27] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20123–20132, 2022. 6

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[29] Xin Lin, Chong Shi, Yibing Zhan, Zuopeng Yang, Yaqi Wu, and Dacheng Tao. Td2-net: Toward denoising and debiasing for dynamic scene graph generation. *ArXiv*, abs/2401.12479, 2024. 6, 7

[30] Xiaochen Liu, Pradipta Ghosh, Oytun Ulutan, BS Manjunath, Kevin Chan, and Ramesh Govindan. Caesar: cross-camera complex activity recognition. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, pages 232–244, 2019. 1

[31] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4235–4243, 2020. 2

[32] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016. 7

[33] Jiale Lu, Lianggangxu Chen, Youqi Song, Shaohui Lin, Changbo Wang, and Gaoqi He. Prior knowledge-driven dynamic scene graph generation with causal inference. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 4877–4885, New York, NY, USA, 2023. Association for Computing Machinery. 6, 7

[34] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 953–962, 2021. 2

[35] Sayak Nag, Kyle Min, Subarna Tripathi, and Amit K. Roy-Chowdhury. Unbiased scene graph generation in videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22803–22813, 2023. 7

[36] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1981–1990, 2019. 6, 7

[37] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Weakly-supervised learning of visual relations. In *Proceedings of the ieee international conference on computer vision*, pages 5179–5188, 2017. 2, 5, 6, 7, 8

[38] Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong, and Dacheng Tao. Distillation using oracle queries for transformer-based human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19558–19567, 2022. 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[40] Raphael Ruschel, A. S. M. Iftekhar, B. S. Manjunath, and Suya You. Bload: Enhancing neural network training with efficient sequential data handling, 2023. 6

[41] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 5

[42] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021. 2, 3, 4, 6

[43] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019. 6

[44] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13688–13697, 2021. 7

[45] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13617–13626, 2020. 2

[46] Oytun Ulutan, Swati Rallapalli, Mudhakar Srivatsa, Carlos Torres, and BS Manjunath. Actor conditioned attention maps for video action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 527–536, 2020. 1

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[48] Guan Wang, Zhimin Li, Qingchao Chen, and Yang Liu. Oed: Towards one-stage end-to-end dynamic scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27938–27947, 2024. 7

[49] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2022. 7

[50] Shuang Wang, Lianli Gao, Xinyu Lyu, Yuyu Guo, Pengpeng Zeng, and Jingkuan Song. Dynamic scene graph generation via temporal prior inference. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM

'22, page 5793–5801, New York, NY, USA, 2022. Association for Computing Machinery. 7

[51] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, 34:17209–17220, 2021. 2

[52] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13319–13327, 2021. 6

[53] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11535–11543, 2019. 7

[54] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11825–11834, 2021. 3