VirtualHome Action Genome: A Simulated Spatio-Temporal Scene Graph Dataset with Consistent Relationship Labels

Yue Qiu, Yoshiki Nagasaki, Kensho Hara, Hirokatsu Kataoka,
Ryota Suzuki, Kenji Iwata, Yutaka Satoh
National Institute of Advanced Industrial Science and Technology (AIST)
{qiu.yue, yoshiki.nagasaki, kensho.hara, hirokatsu.kataoka,
ryota.suzuki, kenji.iwata, yu.satou}@aist.go.jp

Abstract

Spatio-temporal scene graph generation is an essential task in household activity recognition that aims to identify human-object interactions. Constructing a dataset with per-frame object region and consistent relationship annotations requires extremely high labor costs. Existing datasets sparsely annotate frames sampled from videos, resulting in the lack of dense spatio-temporal correlation in videos. Additionally, existing datasets contain inconsistent relationship annotations, leading to the problem of learning ambiguous temporal associations. Moreover, existing datasets mainly discuss relationships that can be inferred from a single frame, ignoring the significance of temporal associations. To resolve those issues, we created a simulated dataset with per-frame consistent annotations and introduced a range of relationships requiring both spatial and temporal context. Most existing methods explore spatial correlations within single images and do not explicitly consider the dynamic changes across frames. Therefore, we proposed a tracking-based approach that explicitly grasps spatio-temporal human-object interactions while simultaneously localizing humans and objects. Our proposed approach achieved state-of-the-art performance on scene graph generation and outperformed existing methods in scene graph localization by large margins on the proposed dataset. Moreover, the experiments show the efficacy of pre-training on the proposed dataset while adapting to a previous benchmark consisting of real daily videos, indicating the potential of the proposed dataset in real-world scenarios.

1. Introduction

Video recognition plays an essential role in various applications due to the increasing use of videos. Action recognition [1, 2, 3], which generates single labels from videos, has been widely discussed. Recently, a range of tasks that require detailed video semantics have been proposed, such as video question answering [4, 5, 6], video captioning [7, 8], spatio-temporal action localization [9, 10].

Figure 1. Previous dataset Action Genome (top) contains inconsistent relationships across video frames (better viewed in color), leading to the problem of learning ambiguous temporal associations. The VirtualHAG dataset is automatically generated from scripts with consistent per-frame annotations (bottom).

Scene graph [11], which can describe objects and object relationships in images, has proved to be effective in various image recognition downstream tasks, such as visual question answering [12], and image captioning [13, 14]. More recently, Ji *et al*. proposed a counterpart for use in videos by defining a spatio-temporal scene graph generation task [15], aimed at identifying humans, objects, and their relationships from video frames.

Ji *et al*. also proposed a novel dataset Action Genome [15]. That dataset consists of videos collected from an existing daily activity dataset Charades [16] with additional scene graph annotation added for video frames. Separately, Rai *et al*. proposed the Home Action Genome (HOMAGE) dataset [17], which contains multi-view daily activity videos. However, there are two major issues with the abovementioned datasets. First, due to the difficulties in annotating videos with accurate and consistent relationship labels, existing manual-labeled datasets contain a part of inconsistent relationships and labeling errors (Figure 1 (top) and Figure 2), raising the concerns for learning ambiguous relationships. Second, most of the relationships in the two datasets can be predicted from a single image, thus requiring fewer temporal associations of their video frames.

To address those issues, we propose a novel dataset VirtualHome Action Genome (VirtualHAG) (Figure 1, bottom), consisting of 2,588 indoor activity videos based on the household activity simulator VirtualHome [18]. VirtualHAG contains consistent relationships and various relationships requiring spatio-temporal context. Due to the high labor costs in dataset constructing, existing datasets sparsely annotated videos. On the contrary, VirtualHAG provides per-frame annotation and can be extended without requiring manual labeling, thus enabling the evaluation and diagnosis of various abilities required in this task.

The experimental results on VirtualHAG show that previous methods mainly explore the spatial correlations within single frames, limiting their performance in distinguishing temporal changes. Object tracking frameworks explicitly focus on spatio-temporal context contained in videos for continuously localizing objects, which could be highly useful for identifying and localizing human-object relationships. Therefore, we propose a tracking-based framework Scene Graph Tracker (SGTracker) that explicitly explores temporal contexts by tracking object changes and associates context intra- and inter-video frames in an end-to-end manner without using object detectors. SGTracker achieved state-of-the-art results on VirtualHAG dataset and outperformed previous methods on localization by large margins. Additionally, we conducted simulation-to-real (sim2real) study on existing benchmark Action Genome. The experimental results show that pre-training on VirtualHAG helps elevating the model performance on Action Genome dataset.

The contributions of our work are four-fold: (i) We propose a novel spatio-temporal scene graph dataset with consistent annotations and various relationships that require spatio-temporal contexts. (ii) We benchmark existing methods, and the results revealed the deficiencies of temporal associations of previous methods. (iii) We propose a method that explicitly associates spatio-temporal contexts by incorporating a tracking framework which obtains high performance on the proposed dataset. (iv) We conduct a sim2real study on the existing benchmark Action Genome and show the potential of the proposed dataset in real-world human-object interaction recognition.

2. Related Work

2.1. Scene Graphs

Johnson *et al*. first introduced the scene graph [11] - a directed graph structure consisting of nodes (objects) and edges (predicates) for use in image retrieval. Scene graph structure has proven efficacy in various downstream tasks, such as image captioning [13, 14], and image generation [19, 20]. The dataset bias problem in the widely used Visual Genome dataset [21] has been discussed extensively. Zellers *et al*. proposed MotifNet [22] to utilize dataset biases for predicting the most frequent relationships between objects. Meanwhile, Tang *et al*. proposed a counterfactual causality-based method [23] for achieving unbiased scene graph generation, and Zhang *et al*. proposed RelDN [24] to incorporate contrastive learning into this task.

Ji *et al*. proposed the spatio-temporal scene graph generation task and Action Genome dataset [15] consisting of third-person-view daily activity videos. Meanwhile, Rai *et al*. proposed the HOMAGE dataset [17], which consists of indoor human-object interaction videos, and extended the task to include multiple viewpoints and sensory inputs. However, those datasets contain inconsistent relationship annotations across video frames, making part of the relationships ambiguous to distinguish. Moreover, most relationships can be determined from a single image. Accordingly, we propose a simulated dataset in which relationships can be automatically computed and in which temporal changes are essential to relationship predictions.

Ji *et al*. evaluated a series of image-based methods [15], including RelDN [24], for use in spatio-temporal scene graph generation. Cong *et al*. introduced a spatio-temporal transformer-based [25] approach STTran [26]. Teng *et al*. proposed a method TRACE [27] to enhance performance by integrating video-based features and object coordinates in images. However, STTran and TRACE implicitly integrate spatio-temporal features and require the use of object detectors. Li *et al*. [28] proposed a method to integrate temporal features but focused on the pre-training process. In contrast, we present a tracking-based method that explicitly considers spatial and temporal scene graph changes and simultaneously tracks object regions.

2.2. Object Tracking

Object tracking has been acknowledged as a fundamental task in the computer vision field, and the introduction of CNNs has elevated object tracking to applicable levels in various applications such as vehicle [29, 30] and pedestrian

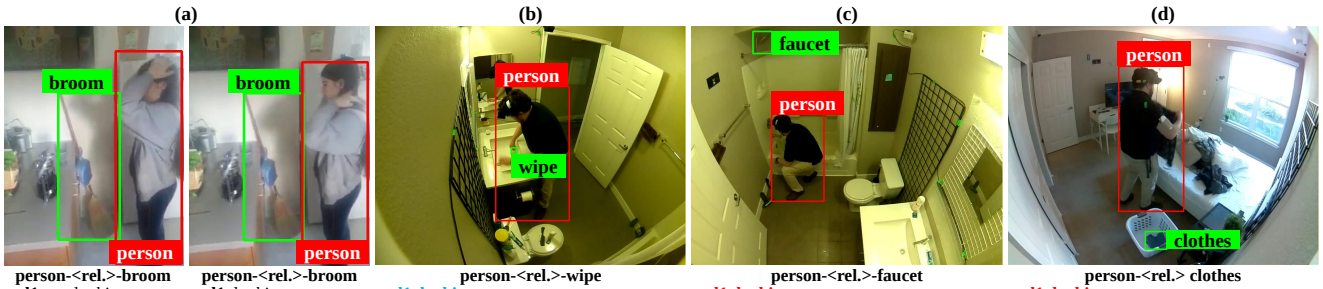

Figure 2. Examples of inconsistent relationship annotations and annotation errors in the Action Genome (a) and the HOMAGE (b), (c), (d) datasets. We highlighted inconsistent relationships in blue, contradictory relationships in cyan, and incorrect relationships in red.

[31, 32] tracking. SiameseFC [33] introduced two share-weighted networks and a similarity function for use in single object tracking. SiameseRPN [34] introduced region proposal networks for better determining object regions, and SiameseRPN++ [35] enhanced SiameseRPN by introducing deep networks. Later, SiamMASK [36] further enabled object mask prediction, while SiamMOT [37] adapted Siamese networks for multi-object tracking.

Recently, transformers have shown promising performance in object tracking [38, 39, 40, 41, 42, 43, 44, 45]. HiFT [40] used transformers to grasp the hierarchical correlations between a template object and search patches, thus providing discriminative representation for the challenging aerial tracking task. TransMOT [41] adopted spatio-temporal graph transformers for associating multiple objects with their trajectories in videos. UTT [45] adopted a unified transformer for single and multiple object tracking.

Since daily human activities often involve complicated changes in object appearance and human-object interaction, tracking-based approaches that explicitly integrate spatial and temporal information in videos have the potential to alleviate the information deficiency in single frames, and are highly suitable for spatio-temporal scene graph generation. Therefore, we adopted a tracking-based framework for improving performance related to this task.

2.3. Household Activity Simulators

Household activity simulators [46, 47, 48, 49, 18] have the potential to provide photo-realistic environments for training and evaluating daily activity recognition tasks. AI2-THOR [46] is a widely used human-object interaction simulator with 120 scenes and 102 interactable objects aimed at facilitating first-person-view applications. The author of ALFRED [47] expanded the AI2-THOR simulator with additional activity language instructions. RoboTHOR [48] further introduced both real scenes and their simulated counterparts to enable sim2real usage. Ultimately, the VirtualHome simulator [18], which consists of human-shaped avatars and allows a variety of human-object interactions, was selected for use in our dataset generation process because of its diversity in objects and interaction types and its high levels of realism.

(a)

attention	spatial	contact	
looking at	in front of	carrying	covered by
not looking at	behind	drinking from	eating
unsure	on the side of	have it on the back	holding
	above	leaning on	lying on
	beneath	not contacting	sitting on
	in	standing on	touching
		twisting	wearing
		wiping	writing on

(b)

spatial (1)	spatial (2)	contact	
close	in front of	touching	sitting on
not close	above	not contacting	standing up
	in	grabbing	switching on
	approaching	putting back	switching off
	not approaching	holding	opening
		drinking from	closing

Figure 3. Relationships defined in Action Genome and HOMAGE datasets (a) and VirtualHAG (b). Relationships requiring temporal context for distinguishing are highlighted with the same colors.

Recently, there have also been several studies on synthetic datasets for human activities [50, 51, 52]. PHASE [50] is a 2D image-based dataset for simulating human social interactions. Watch-And-Help [51] is constructed based on VirtualHome for social perception and human-AI collaboration. BEHAVIOR [52] is also targeted at household activity recognition. Different from our work, BEHAVIOR focuses on activity recognition, rather than human-object interaction, and is a first-person-view dataset.

3. VirtualHAG Dataset

The Action Genome and HOMAGE datasets have two main issues. First, due to the high labor costs and human annotation errors during dataset construction, both datasets contain semantically similar frames with different relationship annotations and examples with incorrect annotations (*e.g.* Figure 2). Second, most relationships in the two datasets, including contact relationships, are descriptive of states (Figure 3 (a)) and can usually be inferred from sin-

Dataset	Videos	Total Frames	Views	Objects	Relationships	Scenes	Person	Seg. Mask	Per-frame Anno.	Expandability	Label Consistency
Action Genome [15]	9,848	496k	1	34	26	-	-	✗	✗	High labor costs required	Unsure
HOMAGE [17]	1,583	383k	2∼5	28	25	2	6	✗	✗	High labor costs required	Unsure
VirtualHAG	2,588	574k	4∼8	50	19	7	4	✓	✓	✓	✓

Table 1. Spatio-temporal scene graph dataset comparison (Seg.: segmentation; Anno.: annotation).

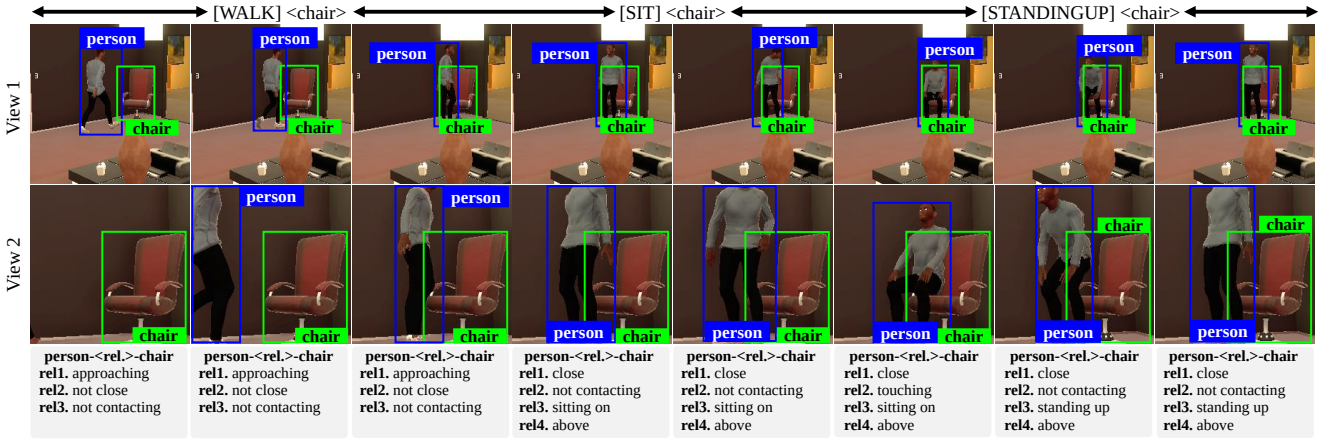

Figure 4. VirtualHAG dataset example sampled from two observation viewpoints (total of six viewpoints for this example).

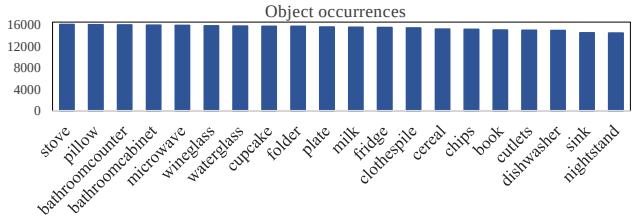

Figure 5. Occurrences of top 20 objects in the VirtualHAG dataset.

gle images. For example, "eating" and "drinking from" can be distinguished from each other from the objects. As a result, both datasets experience difficulties when tasked with diagnosing model abilities in terms of temporal reasoning.

To address these above issues, we propose VirtualHAG dataset, in which consistent relationships are automatically computed, and in which relationships requiring temporal associations and challenging to identify from appearance only are included ((Figure 3 (b)). We show the dataset comparison in Table 1.

3.1. VirtualHome Simulator

We built the VirtualHAG dataset based upon the existing indoor household activities simulator VirtualHome [18], which automatically generates videos from action-describing scripts. For example, based on the following simple script, "[WALK] $\langle sofa \rangle$", the avatar will walk from its original position to a sofa. VirtualHome allows the recording of segmentation masks, avatar poses, action types, object states (*e.g.* open, switch-on), and spatial relationships (*e.g.* "book" is inside the "bookshelf") for each video frame. It contains 308 object categories, executable non-object actions (*e.g.* walk, turn left), and various actions involving human-object interactions (*e.g.* grab, open). With these, users can generate a variety of household videos by designing scripts.

3.2. Objects and Relationships

A spatio-temporal scene graph annotation consists of categories, and the bounding box coordinates of objects and the subject ("human"), and the interaction relationships. We carefully chose 50 daily object categories from VirtualHome for inclusion in VirtualHAG scene graphs.

We constructed our relationship set based on existing dataset setups (Figure 3 (a)) and the executable action types defined in VirtualHome. We deleted attention relationships because of the difficulty in differentiating "looking at" and "not looking at" in simulated environment. To enhance the model's temporal reasoning abilities, we added four relationship pairs "grabbing" and "putting back", "opening" and "closing", "switching on" and "switching off", "standing up" and "sitting on", each of which require the temporal context to differ from each other. We added "close" and "not close" to indicate if the object is within handling distance to the human, and "approaching" and "not approaching" to indicate whether or not the human is walking towards the object.

In VirtualHAG, frame-by-frame human-object relationships are fully computable based on the action types, the position of the human and objects, and the state of objects. The existence of each relationship is computed and recorded during the dataset generation, and all of the rela-

tionships defined in VirtualHAG are shown in Figure 3 (b).

3.3. Video Generation Scripts

To generate videos in the VirtualHome simulator, we manually designed 108 unique scripts (*e.g*. Figure 4, top) by adjusting the object categories, interaction types, and interaction order. Each script consists of a sequence of non-object human actions and human-object interactions defined in VirtualHome.

3.4. Dataset Generation Process

Video Generation. Before executing each script, the script id, avatar id (from four types of avatars), and scene id (from seven scenes) were randomly determined and followed a uniform distribution. We also randomly picked four to eight viewpoints for each script from a group of predefined camera positions where the scene could be well observed. Each script was then executed in the VirtualHome simulator, and multiview videos were recorded. From this step, 2,588 multiview videos were generated.

Dataset Balancing. After video generation, we computed bounding box coordinates of the human and objects and removed all frames with bounding box edges shorter than five pixels. After completing the above step, we obtained 2,345,231 valid frames. Next, we further balanced the dataset to form a uniform distribution of object categories and opposite relationships (*e.g*. "open" and "close") to prevent models from overfitting. The resulting balanced dataset contains 574,635 valid frames. Figure 4 shows a dataset example. Figure 5 shows the distribution of the top-20 objects. For each script, we randomly chose five scenes for use as training and two for use as test data. Unlike previous datasets (Table 1), the VirtualHAG contains per-frame annotation and can be easily extended with minimal labor costs as well as allowing the use of multiple additional sensory information types, such as segmentation masks and depth images. Additional dataset and script examples will be provided in the supplementary material.

4. SGTracker

Given a specific image and a video clip that includes that image, the spatio-temporal scene graph task aims to identify the human and object region, and predict labels of objects and relationships. The spatial context contained inside the image is crucial for detecting the human and object regions and determining their labels. Some interaction types can also be predicted from a single image frame, such as "in front of" and "touching", but some relationships such as "sitting on" and "standing up" or "opening" and "closing", could be ambiguous when viewed in single frames. Moreover, the video context can alleviate information deficiencies in single frames (*e.g*. motion blur and occlusions).

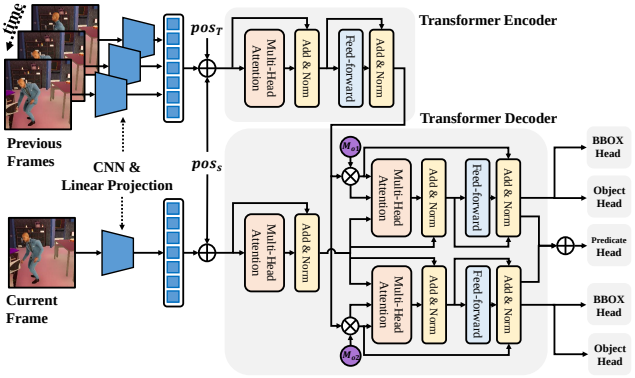

Figure 6. Overview of the proposed approach SGTracker.

Despite the significance of image and video information in this task, a number of methods only consider single image inputs [15, 53]. Two previous state-of-the-art methods, STTran [26] and TRACE [27], consider both spatial and temporal context but they do not explicitly focus on the changed regions in video frames. Moreover, STTran and TRACE are built upon object detectors, which means they cannot achieve end-to-end scene graph generation.

In contrast, we propose a transformer encoder-decoder framework called SGTracker to explicitly incorporate temporal and spatial contexts (Figure 6). More specifically, SGTracker is a tracking-based framework that utilizes temporal contexts to track objects and simultaneously determine the object and predicate labels. The encoder of SGTracker grasps the spatio-temporal contexts of previous video frames while the decoder explores spatial information contained in current frames and then further associates current and previous frames via a cross-attention mechanism. These framework details are discussed in the following.

4.1. Feature Extraction and Encoder

The temporal context is critical for recognizing dynamic changes in human-object interactions, and has the potential to provide clues for tracking objects in the upcoming frames. Here, we introduce a transformer-encoder for obtaining the dynamic changes in previous image frames.

Given the previous frames ($0, ..., T-1$ frames) and the current frame (T frame), we first use CNNs to extract input image features and obtain $V_{prev}(i,j,t) = I_0, I_1, ..., I_{T-1}$ and $I_{curr}(i,j) = I_T \in \mathbb{R}^{H \times W \times D}$, where H and W are the height and width of the features, respectively, and D is the dimension of each spatial location. Next, we adopt a linear projection LP (with weight W_{LP} and bias b_{LP}) to each frame in V_{prev} and the I_{curr} to reduce the channel depth from D to C. To enhance the model's spatial and temporal reasoning abilities, we use two different position embeddings pos_S and pos_T, where pos_S transfers each spatial location (i,j) in $H \times W$ to a C-dimensional embedding,

and pos_T encodes temporal index t to a C-dimensional embedding. The transformer encoder and decoder inputs are shown in the following equations:

$$V = W_{LP}V_{prev} + b_{LP} + pos_S(i,j) + pos_T(t) \quad (1)$$

$$I = W_{LP}I_{curr} + b_{LP} + pos_S(i,j) \quad (2)$$

We then feed V into a standard transformer encoder $\mathrm{Att}(query, key, value)$ and obtain $\hat{V} = \mathrm{Att}(V,V,V) \in \mathbb{R}^{T\times H\times W\times C}$. We omit the notations of add, normalization, and feed-forward operations and show them in Figure 6.

4.2. Decoder

Given the input $\hat{V}$ from the encoder and current image feature I, the decoder aims to combine the spatio-temporal context contained in $\hat{V}$ and I in order to determine the object regions and labels of the objects and predicates. To accomplish this, we first adopt a multi-head attention $\hat{I} = \mathrm{Att}(I,I,I) \in \mathbb{R}^{H\times W\times C}$ on I for spatial reasoning.

To highlight the object regions in previous frames, we adapt the Gaussian-shaped masks used in [54] to compute masks from ground truth bounding boxes. With two successive regions for Objects 1 and 2 in $\hat{V}$ (representing humans and objects), we then compute two masks M_{o1} and M_{o2} for Objects 1 and 2 using the following equation, where c is the ground truth target position, y indicates each spatial location of input features.

$$M(y) = exp(-\frac{\|y-c\|^2}{2\sigma^2}) \quad (3)$$

Next, we obtain masked object features $\hat{V}_{o1} = \hat{V}\cdot M_{o1}$ and $\hat{V}_{o2} = \hat{V}\cdot M_{o2}$ by dot production of $\hat{V}$ with M_{o1} and M_{o2}. Next, we adapt cross-attention $\mathrm{Att}(\hat{I},\hat{V}_{o1},\hat{V}_{o1})$, $\mathrm{Att}(\hat{I},\hat{V}_{o2},\hat{V}_{o2})$ for correlations between M_{o1} and M_{o2} with $\hat{I}$, respectively with shared weight. We use a skip-connection to add up the output of cross-attention with $\hat{V}_{o1}$ and $\hat{V}_{o2}$. The final decoder outputs are shown as follows:

$$\hat{I}_{o1} = \mathrm{Att}(\hat{I},\hat{V}_{o1},\hat{V}_{o1}) + \hat{V}_{o1} \quad (4)$$

$$\hat{I}_{o2} = \mathrm{Att}(\hat{I},\hat{V}_{o2},\hat{V}_{o2}) + \hat{V}_{o2} \quad (5)$$

Finally, we transfer $\hat{I}_{o1}$ and $\hat{I}_{o2}$ using a linear projection and obtain object classification results for Objects 1 and 2 separately. For predicate prediction, we first add up the features $\hat{I}_{o1}$ and $\hat{I}_{o2}$ and then adapt a linear classification for determining the predicate labels. For bounding box regression, we use a process that was inspired by [42] and [39] to transfer each embedding in all of the $H\times W$ regions in $\hat{I}_{o1}$ and $\hat{I}_{o2}$ using a linear projection. Then, we separately predict the objectness and bounding box offset of each $H\times W$ region in $\hat{I}_{o1}$ and $\hat{I}_{o2}$.

Tracking loss L_{bbox}	Attention type	Object accuracy	Predicate accuracy
Without	S	73.8	79.7
	T	73.4	79.6
	ST	73.8	79.9
With	S	74.4	81.1
	T	74.5	81.9
	ST	**75.0**	**82.4**

Table 2. Object and predicate accuracy of different model designs applied to the VirtualHAG dataset. (Attention types: S (Spatio-only), T (Temporal-only), and ST (Spatio-Temporal)).

4.3. Loss Function

As shown in Equation (6), our loss consists of three parts, with each weighted by λ_1, λ_2, and λ_3.

$$L = \lambda_1 L_{obj} + \lambda_2 L_{pred} + \lambda_3 L_{bbox} \quad (6)$$

We adopt cross-entropy loss L_{obj} for object classification and multi-label cross-entropy loss L_{pred} for predicate classification. Similar to [42] and [39], the bounding box regression loss L_{bbox} can be formulated as follows:

$$L_{bbox} = \lambda_4 L_{cls} + \lambda_5 L_{reg} \quad (7)$$

In Equation (7), we adopt the cross-entropy for L_{cls} to evaluate if each region is correctly identified as a "background" or an "object". We use the L_{reg} with the same setup cited in [39], which evaluates the intersection of union (IoU) for each predicted region with ground truth.

5. Experiments

5.1. Experimental Setups

Datasets. We evaluated different methods on the VirtualHAG dataset. In addition, we also conducted sim2real experiments on Action Genome dataset while pre-trained with VirtualHAG dataset.

Evaluation Metrics. Similar to existing methods evaluated on the Action Genome dataset, we also adapted three major metrics: predicate classification (PredCLS), scene graph classification (SGCLS), and scene graph detection (SGDET) [23]. All three metrics evaluate recall@K relationships in predicted relationships compared to ground truth. The ground truth object region and labels are provided in PredCLS and the object region is provided in SGCLS. In SGDET, models are expected to predict object and predicate categories and detect object region (successful detections have over 0.5 IoU with ground truth bounding boxes). Similar to STTran [26], we evaluated metrics both with and without graph constraint setups. The with graph constraint setup means that each human-object pair has at most one predicate for each type of relationship in Figure 3.

Methods. In addition to the proposed method, we also evaluated two previous representative methods RelDN [15]

Methods	Input	Region proposal	Backbone	End-to-end	With graph constraint						Without graph constraint					
					PredCLS			SGCLS			PredCLS			SGCLS		
					R@3	R@5	R@10	R@3	R@5	R@10	R@3	R@5	R@10	R@3	R@5	R@10
RelDN [15]	Single image	FasterRCNN [55]	ResNet101 [56]	✗	61.9	67.2	68.9	47.9	54.6	60.8	62.0	81.2	94.7	48.3	62.6	74.9
STTran [26]	Video (4 frames)	FasterRCNN [55]	Transformer [25]	✗	62.4	68.3	71.2	48.1	55.2	62.6	59.9	82.2	95.0	46.8	62.6	75.0
RelDN [15]	Video (32 frames)	-	3DResNet [2]	✓	68.8	75.0	77.4	43.8	50.7	58.8	68.8	88.0	**98.3**	44.4	55.4	66.4
RelDN [15]	Video (32 frames)	-	TimeSformer [57]	✓	**69.8**	**76.0**	**78.4**	44.8	52.4	61.2	**69.8**	**88.4**	98.0	45.3	56.9	69.5
SGTracker	Video (4 frames)	-	Transformer [25]	✓	65.9	71.3	74.0	**52.0**	**59.0**	**65.3**	65.7	86.1	96.9	**51.9**	**65.5**	**77.2**

Table 3. Results comparison for PredCLS and SGCLS using different methods on the VirtualHAG dataset (R@K: recall@K).

Methods	With graph constraint			Without graph constraint		
	R@3	R@5	R@10	R@3	R@5	R@10
RelDN [15]	25.4	26.9	28.0	25.4	31.2	35.7
STTran [26]	23.6	27.3	28.6	25.7	34.9	42.4
SGTracker	**36.7**	**39.4**	**41.8**	**35.8**	**44.0**	**50.1**

Table 4. SGDET results of different methods on VirtualHAG.

and STTran [26] in the VirtualHAG dataset. RelDN predicts scene graphs based on single images, while STTran first uses FasterRCNN [55] for object detection and then adopts a spatio-temporal transformer structure for combining context among video frames. To enhance RelDN model in temporal reasoning, we introduced two video recognition methods, 3DResNet [2] and TimeSformer [57], into RelDN for feature extraction. During implementation, we clip 32-frame videos at the coordinates of the ground truth bounding boxes of the target frame.

Implementation Details. We trained all models for 40 epochs. During all of the experiments, we set the image resolution to 224×224. In implementing the 3DResNet and TimeSformer-based methods, we set the initial learning rate to 0.001 and 0.0001 for all other methods. We used the 3DResNet (pre-trained on Kinetics-700) and TimeSformer (pre-trained on Kinetics-600) provided by the official implementation, respectively. The input video frame number for 3DResNet and TimeSformer was set to 32, and 4 for STTran and SGTracker. For SGDET evaluations, we trained the FasterRCNN model on VirtualHAG for 20 epochs and adopted the same model into RelDN and STTran. We set $\lambda 1$ to $\lambda 5$ in Equations (6) and (7) to 1. We used ResNet101 [56] pre-trained on ImageNet for image feature extraction for SGTracker, resulting in $7 \times 7 \times 2048$-dimensional features. The heads and layers of transformers were set to 2 with 2048 feed-forward dimensions.

5.2. Experiments on VirtualHAG Dataset

Ablation Study. We first examined the performance on object and predicate predictions of different SGTracker model designs on VirtualHAG in Table 2. The models were adopted with and without the tracking loss L_{bbox}, and with spatio-only, temporal-only, and spatio-temporal attention to combine information from previous frames in the transformer encoder. The experimental results indicate that integrating tracking loss improved model performance for all different attention types and that the model performance can be improved by using spatio-temporal attention when adopting tracking loss. Therefore, we used SGTracker with spatio-temporal attention in the remaining experiments.

Scene Graph Generation Evaluation. Table 3 shows the scores of PredCLS and SGCLS on VirtualHAG. For PredCLS, RelDN with video-level features (3DResNet and TimeSformer) achieved relatively high performance, but singe-image RelDN and STTran exhibited a noticeable performance gap with video-based methods. In the VirtualHAG dataset, a range of relationships require temporal reasoning (*e.g.* "opening" and "closing"), thus making video feature-based methods more useful in predicate predictions. Singe-image RelDN and STTran obtained higher accuracy on SGCLS than video feature-based RelDNs. Unlike the PredCLS evaluation, the SGCLS evaluation also considers object classification. Therefore, the spatial information contained in the target image is beneficial for obtaining high accuracy. Finally, our proposed SGTracker obtained higher PredCLS scores than single-image RelDN and STTran, and achieved the highest SGCLS scores of all methods examined, indicating that the SGTracker could efficiently integrate spatial and temporal information in video frames.

The SGDET evaluation results are shown in Table 4. It is noteworthy that while singe-image RelDN and STTran use FasterRCNN to extract object proposals, SGTracker achieves end-to-end scene graph generation and object tracking. Additionally, SGTracker outperformed both two methods by a large margin, showing its effectiveness in human-object interaction recognition and localization.

Qualitative Results. In Figure 7, we show example results for frames and viewpoints sampled from a video describing an avatar putting a book back on a bookshelf. SGTracker exhibited promising localization and object classification performance, including a small object (book).

However, in frames 277 and 290 (views 4 and 7), the human-book relationships predicted by SGTracker were false. In frame 290, one false relationship was "drinking from", which is nearly impossible when humans and books are paired. The results could possibly be improved by incorporating object labels in predicate prediction. In frame

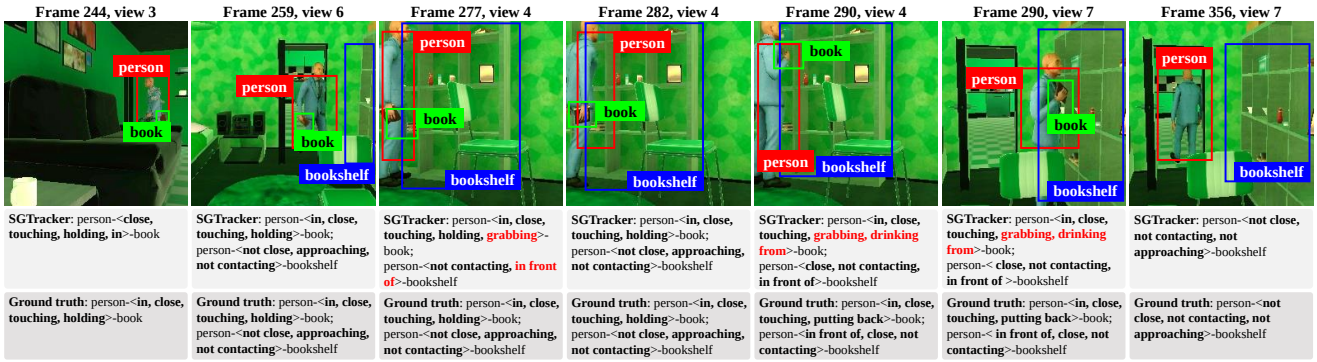

Figure 7. Example result on VirtualHAG. The incorrect predictions are hightlighted in red.

Pre-train (VirtualHAG)	PredCLS		SGCLS	
	R@10	R@20	R@10	R@20
Scratch	67.1	70.9	45.3	46.8
Pre-train	**68.4**	**72.1**	**45.9**	**47.7**

Table 5. Sim2real study on Action Genome dataset.

277, SGTracker predicted incorrect predicates for the bookshelf, while in this example, the predicted predicate "in front of" appears to be correct. When human avatars are walking towards one object, we annotate the relationship as "approaching". When avatars stop in front of the objective object, the relationship "in front of" is annotated. However, the above relationships are ambiguous and difficult to annotate in frame 277, so we consider this to be one of the limitations of our proposed VirtualHAG dataset and will endeavor to introduce more flexible relationship annotations in the future. More experimental results will be provided in the supplementary material.

5.3. Sim2real Study on Action Genome Dataset

To evaluate the dataset efficiency of VirtualHAG when applied to real-world situations, we evaluated the sim2real performance on Action Genome dataset [15] (Figure 1 (top)) of our proposed SGTracker pre-trained on VirtualHAG. Since Action Genome and VirtualHAG have different object and predicate labels, we re-labeled both datasets by combining all labels in these two datasets while keeping the labels of the same object and predicate to be consistent. We pre-trained SGTracker on VirtualHAG dataset for 10 epochs. Then, we trained models on Action Genome for 20 epochs. As shown in Table 5, although VirtualHAG is purely synthetic while Action Genome consists of real-scenario videos, we found that pre-trained on VirtualHAG elevated the model performance. The results indicate the potential of VirtualHAG to be adopted in human-object interaction recognition in real-world scenarios.

5.4. Limitations

Since SGTracker uses a tracking backbone, it requires previous frames with known object regions to detect relationships. Further studies on the online tracking framework could help achieve tracking relationships without ground truth object region. Currently, SGTracker processes each paired object region within an image separately, which means that computing costs increase significantly when multiple objects interact with humans with each frame. Adopting a multi-object tracking framework might be helpful in dealing with situations involving multiple objects.

6. Conclusion

This paper proposed a novel spatio-temporal scene graph dataset and a transformer-based method that simultaneously recognizes and localizes human-object relationships. Existing manually constructed datasets contain inconsistent relationship annotations and mainly consider relationships that can be inferred from a single frame, thereby limiting their evaluation ability in temporal reasoning. To resolve that issue, we proposed a simulated dataset VirtualHAG which contains per-frame consistent annotations and various relationships that require temporal associations. Most existing methods do not explicitly explore the temporal changes between frames, thus limiting their ability to distinguish and localize temporal changes. Accordingly, we proposed a method that explicitly identifies temporal changes by tracking human-object relationships between frames, which outperformed existing methods by large margins in localizing human-object interactions on VirtualHAG.

Acknowledgements

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used.

References

[1] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, pages 568–576, 2014.

[2] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555, 2018.

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *CVPR*, pages 6202–6211, 2019.

[4] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

[5] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*, 2019.

[6] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *CVPR*, pages 6576–6585, 2018.

[7] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055, 2017.

[8] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, pages 8739–8748, 2018.

[9] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, pages 4405–4413, 2017.

[10] Dejun Zhang, Linchao He, Zhigang Tu, Shifu Zhang, Fei Han, and Boxiong Yang. Learning motion representation for real-time spatio-temporal action localization. *Pattern Recognition*, 103:107312, 2020.

[11] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, pages 3668–3678, 2015.

[12] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *CVPR*, pages 1–9, 2017.

[13] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, pages 10685–10694, 2019.

[14] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *ICCV*, pages 10323–10332, 2019.

[15] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, pages 10236–10247, 2020.

[16] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526, 2016.

[17] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *CVPR*, pages 11184–11193, 2021.

[18] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *CVPR*, pages 8494–8502, 2018.

[19] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *CVPR*, pages 1219–1228, 2018.

[20] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *CVPR*, pages 8584–8593, 2019.

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[22] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018.

[23] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, pages 3716–3725, 2020.

[24] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *CVPR*, pages 11535–11543, 2019.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[26] Yuren Cong, Wentong Liao, Hanno Ackermann, Michael Ying Yang, and Bodo Rosenhahn. Spatial-temporal transformer for dynamic scene graph generation. In *ICCV*, pages 16372–16382, 2021.

[27] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *ICCV*, pages 13688–13697, 2021.

[28] Yiming Li, Xiaoshan Yang, and Changsheng Xu. Dynamic scene graph generation via anticipatory pre-training. In *CVPR*, pages 13874–13883, 2022.

[29] Shuai Hua, Manika Kapoor, and David C Anastasiu. Vehicle tracking and speed estimation from traffic videos. In *CVPR*, pages 153–160, 2018.

[30] Minghu Wu, Yeqiang Qian, Chunxiang Wang, and Ming Yang. A multi-camera vehicle tracking system based on city-scale vehicle re-id and spatial-temporal information. In *CVPR*, pages 4077–4086, 2021.

[31] Daniel Stadler and Jurgen Beyerer. Improving multiple pedestrian tracking by track management and occlusion handling. In *CVPR*, pages 10958–10967, 2021.

[32] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljosa Osep, Simone Calderara, Laura Leal-Taixe, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *CVPR*, pages 10849–10859, 2021.

[33] Miaobin Cen and Cheolkon Jung. Fully convolutional siamese fusion networks for object tracking. In *ICIP*, pages 3718–3722, 2018.

[34] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, pages 8971–8980, 2018.

[35] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*, pages 4282–4291, 2019.

[36] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *CVPR*, pages 1328–1338, 2019.

[37] Bing Shuai, Andrew Berneshawi, Xinyu Li, Davide Modolo, and Joseph Tighe. Siammot: Siamese multi-object tracking. In *CVPR*, pages 12372–12382, 2021.

[38] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, pages 8126–8135, 2021.

[39] Bin Yu, Ming Tang, Linyu Zheng, Guibo Zhu, Jinqiao Wang, Hao Feng, Xuetao Feng, and Hanqing Lu. High-performance discriminative tracking with transformers. In *ICCV*, pages 9856–9865, 2021.

[40] Ziang Cao, Changhong Fu, Junjie Ye, Bowen Li, and Yiming Li. Hift: Hierarchical feature transformer for aerial tracking. In *ICCV*, pages 15457–15466, 2021.

[41] Peng Chu, Jiang Wang, Quanzeng You, Haibin Ling, and Zicheng Liu. Transmot: Spatial-temporal graph transformer for multiple object tracking. *arXiv preprint arXiv:2104.00194*, 2021.

[42] Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. Siamcar: Siamese fully convolutional classification and regression for visual tracking. In *CVPR*, pages 6269–6277, 2020.

[43] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: Multi-object tracking with memory. In *CVPR*, pages 8090–8100, 2022.

[44] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *CVPR*, pages 8771–8780, 2022.

[45] Fan Ma, Mike Zheng Shou, Linchao Zhu, Haoqi Fan, Yilei Xu, Yi Yang, and Zhicheng Yan. Unified transformer tracker for object tracking. In *CVPR*, pages 8781–8790, 2022.

[46] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[47] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *CVPR*, pages 10740–10749, 2020.

[48] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *CVPR*, pages 3164–3174, 2020.

[49] Kwanyoung Park, Hyunseok Oh, and Youngki Lee. Veca: A toolkit for building virtual environments to train and test human-like agents. *arXiv preprint arXiv:2105.00762*, 2021.

[50] Aviv Netanyahu, Tianmin Shu, Boris Katz, Andrei Barbu, and Joshua B Tenenbaum. Phase: Physically-grounded abstract social events for machine social perception. In *AAAI*, volume 35, pages 845–853, 2021.

[51] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. *ICLR*, 2021.

[52] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *CoRL*, pages 477–490, 2022.

[53] Yichao Lu, Cheng Chang, Himanshu Rai, Guangwei Yu, and Maksims Volkovs. Multi-view scene graph generation in videos. In *International Challenge on Activity Recognition (ActivityNet) CVPR 2021 Workshop*, 2021.

[54] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *CVPR*, pages 1571–1580, 2021.

[55] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015.

[56] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[57] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021.