Spatial-Temporal Transformer for Dynamic Scene Graph Generation

Yuren Cong , Wentong Liao , Hanno Ackermann , Bodo Rosenhahn , Michael Ying Yang

Presenter: Çağhan Köksal
Tutor: Ege Özsoy

Technische Universität München
Johns Hopkins
Whiting School of Engineering

Motivation

Technische Universität München

Johns Hopkins Whiting School of Engineering

Motivation and Contributions

- **Motivation**
  - Most of the previous works **omit temporal domain**

- **Goal**
  - Generating dynamic scene graphs from videos.
  - Leveraging temporal domain to model dynamic relationships between objects

- **Contributions**
  - They create a **novel framework**, Spatial-Temporal Transformer (STTran).
  - **Multi-label classification** is used in a relationship classification task.
  - **Novel thresholding strategy** to select additional confident relations between objects
  - **Extensive experiments and ablation studies** to show the effectiveness of the model to use temporal information.

Related Works

Image Retrieval using Scene Graphs

- Novel framework for Semantic Image Retrieval

- Scene Graphs as a query
  - Retrieve similar images
  - More precise semantic description

- CRF reason about groundings of scene graphs
- Likelihoods as a ranking
- Novel dataset of 5000 scene graphs
- Scene graphs
  - **Objects**
    - Man
    - Boat
  - **Relationships between objects**
    - Man “standing” on boat
  - **Attributes of objects**
    - Boat is white

(a) Results for the query on a popular image search engine.

(b) Expected results for the query.

Figure 1: Image search using a complex query like “man holding fish and wearing hat on white boat” returns unsatisfactory results in (a). Ideal results (b) include correct *objects* (“man”, “boat”), *attributes* (“boat is white”) and *relationships* (“man on boat”).

Johnson, J., Krishna, R., Stark, M., Li, L. J., Shamma, D., Bernstein, M., & Fei-Fei, L. (2015). Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3668-3678).

Scene Graph Generation by Iterative Message Passing

Motivation

- **Relations between interacting objects**

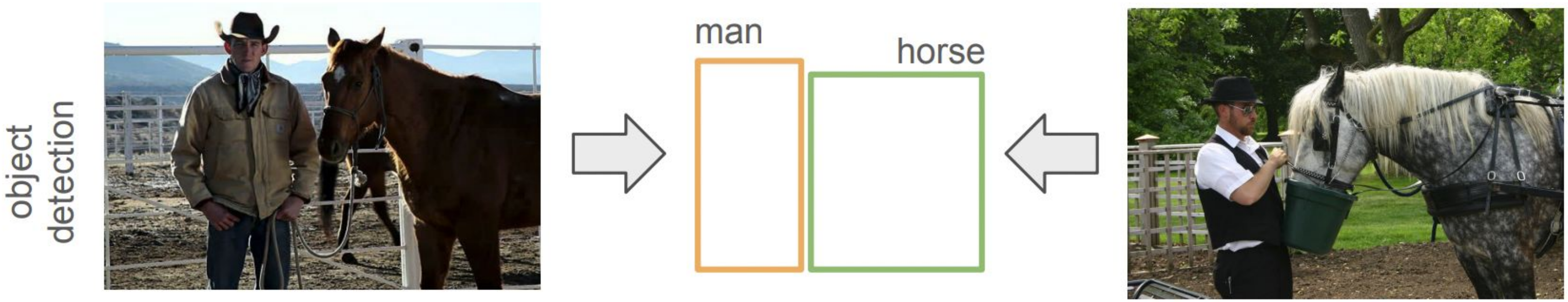


Figure: Two semantically different images have the same representation

- Novel Scene Graph Generation method
- **Learns to improve it's predictions**
    - Iterative message passing algorithm
- Contextual information

Scene Graph Generation by Iterative Message Passing

- Object Proposal Network
- Graph Inference Network
  - **Input :**
    - Features of object regions
  - **Output:**
    - Categories of object
    - Relationship types between object pairs

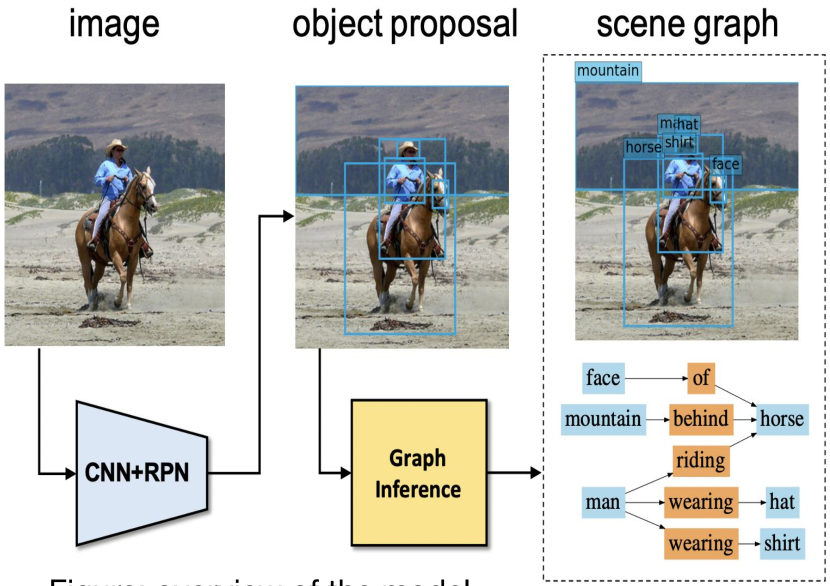


Figure: overview of the model

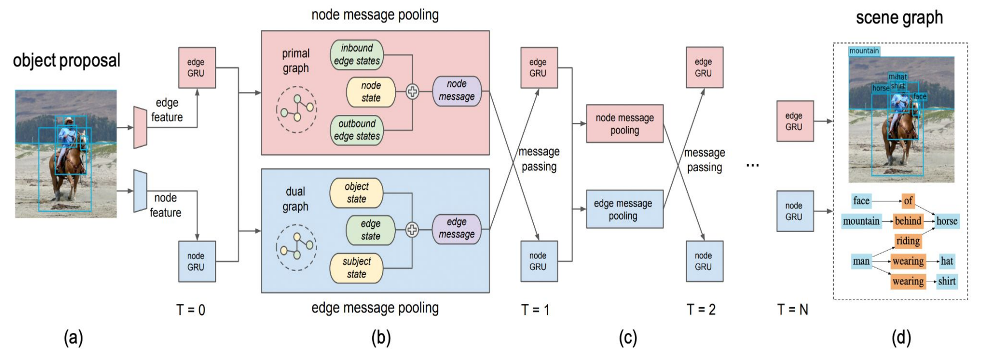


Figure: Detailed Model

Xu, D., Zhu, Y., Choy, C. B., & Fei-Fei, L. (2017). Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5410-5419).

Graph R-CNN for Scene Graph Generation

- Goal: Reduce quadratic complexity of pairwise relationships

- Solution: Relation Proposal Network (RePN)

Contributions

- Contextual representation with Attentional Graph Convolutional Networks
- Relation Proposal Networks
- More realistic evaluation metric

Yang, J., Lu, J., Lee, S., Batra, D., & Parikh, D. (2018). Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 670-685)

Graph R-CNN for Scene Graph Generation

Region Proposal Networks

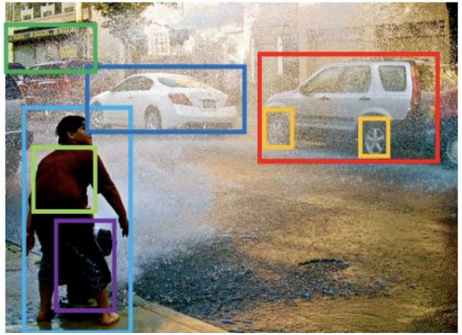
Attentional Graph Convolutional Networks

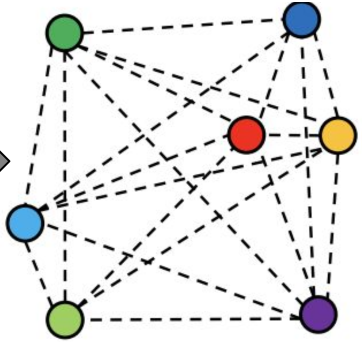
Faster-RCNN detects all possible nodes

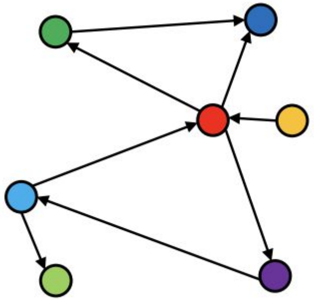
All Possible Relationships

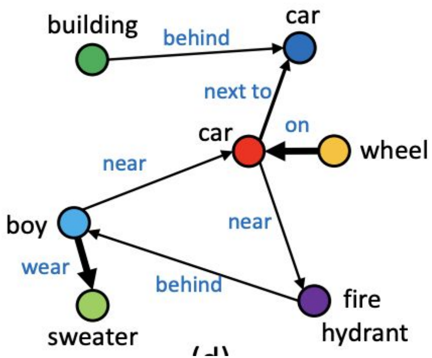
RePN prunes -> sparser graph

(AGCN) -> Predicates and node labels









Yang, J., Lu, J., Lee, S., Batra, D., & Parikh, D. (2018). Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 670-685)

Attention Is All You Need

- Main building block of the modern systems.
- Encoder - Decoder Architecture
- Fully Attentional Networks
- Self Attention
  - Contextualized representation
- Parallelizable architecture
- SOTA results on NLP

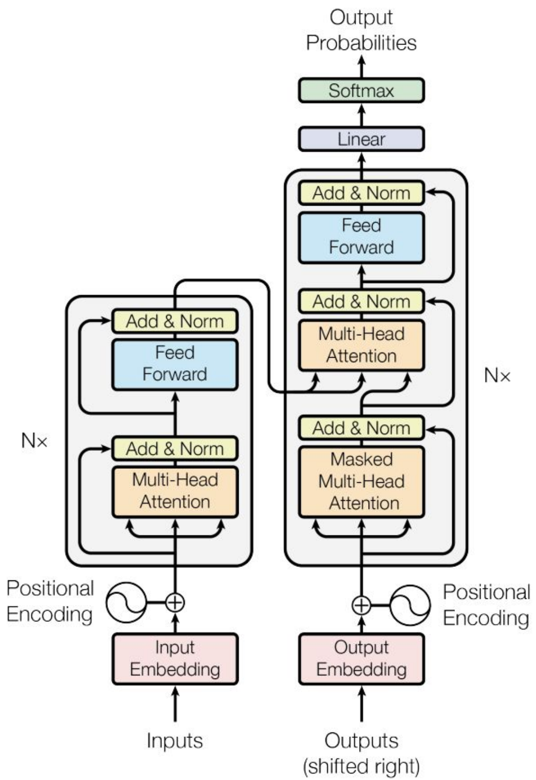


Figure: Transformer Architecture

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

- Transformers limited use in vision
- Sequence of image patches
- Good performance on image classification
- Large amount of training data is needed

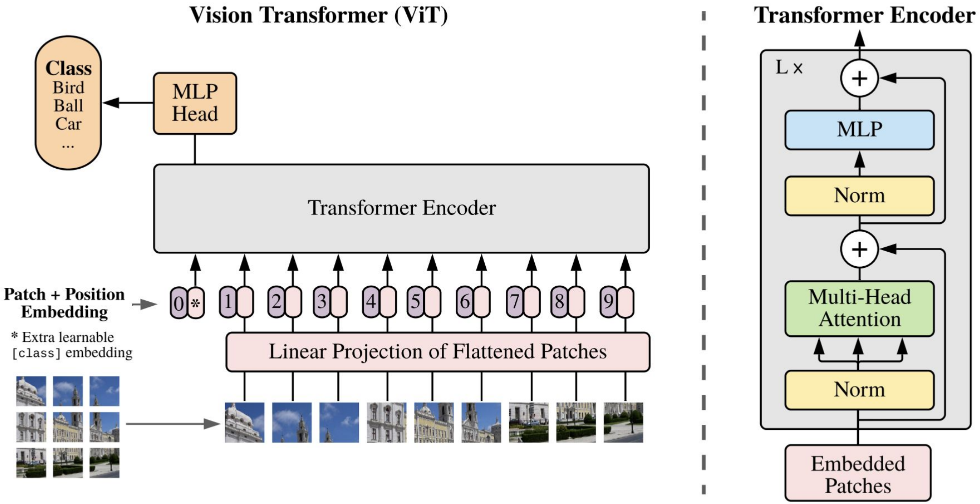


Figure: Vision Transformer Architecture

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

End-to-End Object Detection with Transformers(DETR)

- Object detection as set prediction
- Remove NMS
- Set based loss via bipartite matching
- Learned object queries

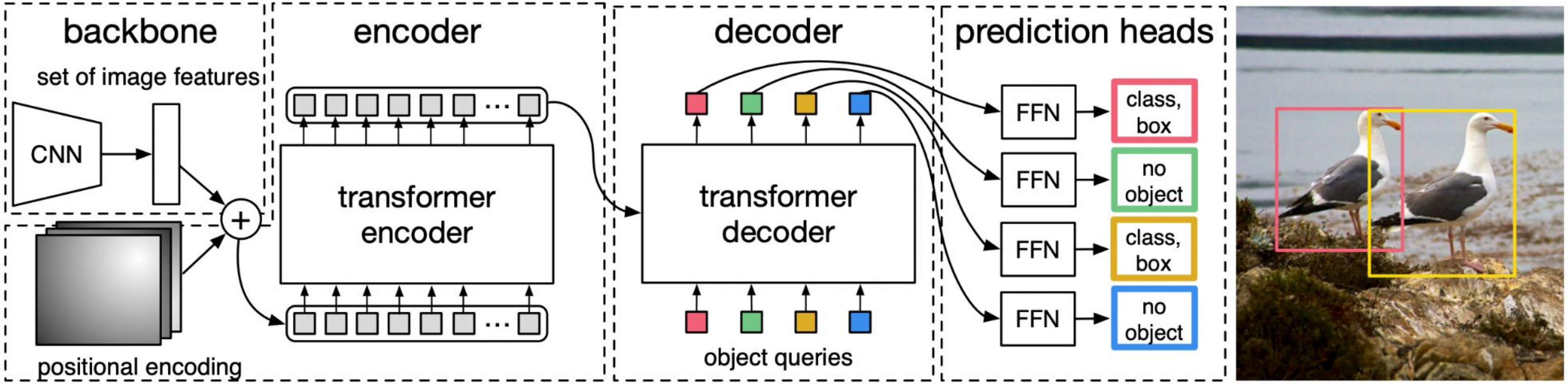


Figure: DETR Architecture

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213-229). Springer, Cham.

Video Action Transformer Network

- Recognize and Localize human actions
- Spatiotemporal context
- Learns to track individual people
- Pick up on semantic context from the actions of others.
- Attention on hands and faces

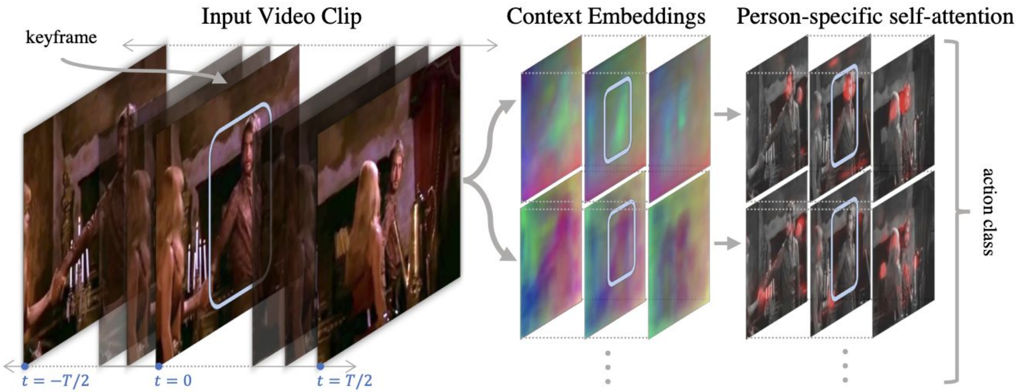


Figure: Overview of Video Action Transformer Network

Girdhar, R., Carreira, J., Doersch, C., & Zisserman, A. (2019). Video action transformer network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 244-253).

End-to-End Video Instance Segmentation with Transformers

- Video instance segmentation (VIS)
  - Classification
  - Segmentation
  - Object Tracking
- Input:
  - Sequence of images
- Output
  - Sequence of masks for each instance

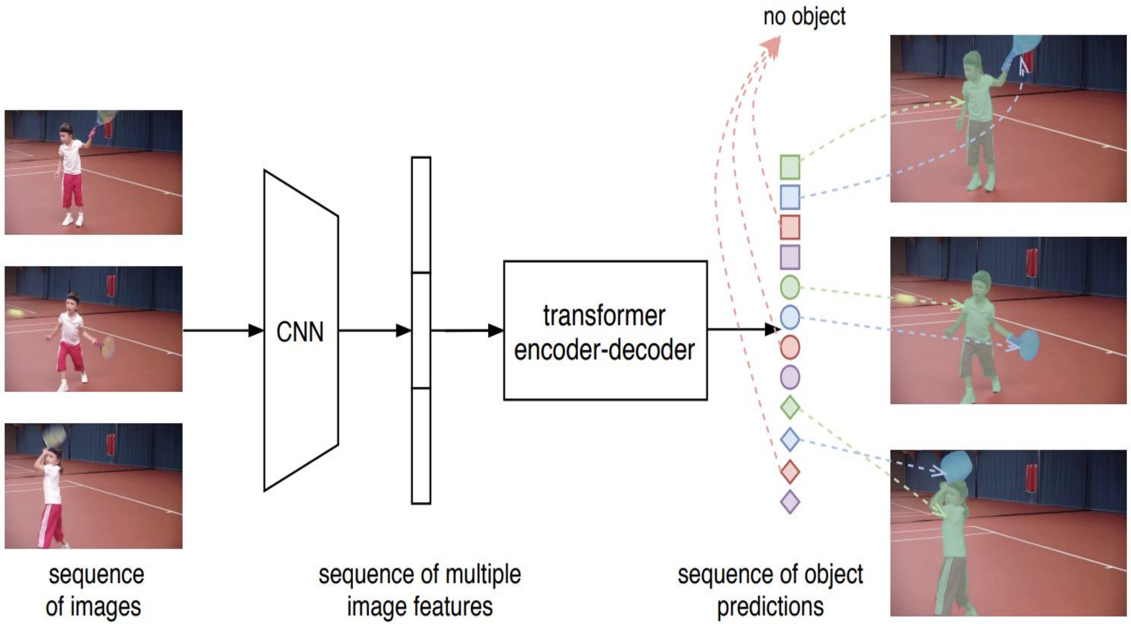


Figure: Architecture Overview

Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., & Xia, H. (2021). End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8741-8750).

Two-Stream Convolutional Networks for Action Recognition in Videos

- Goal
    - Action Recognition in video

- Appearance + Motion
- Two-stream ConvNet architecture
- Competitive results with SOTA when it is published

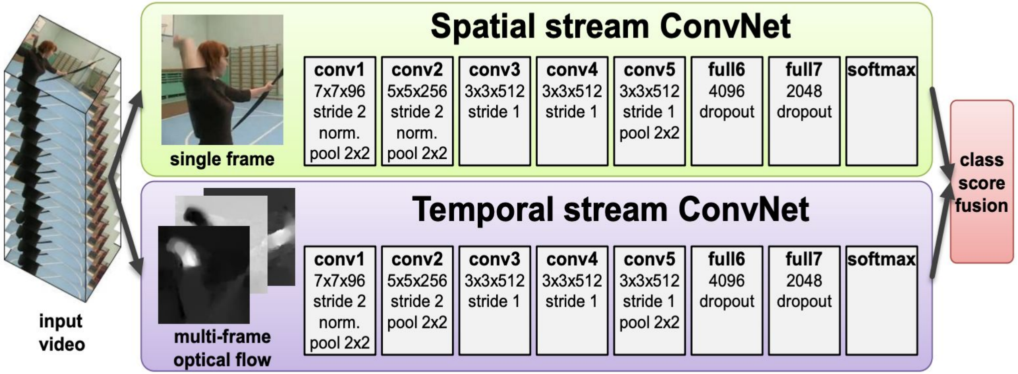


Figure:Two stream CNN architecture

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, *27*.

Dataset

Action Genome

Motivation:
Events : hierarchically structured to be perceived by humans.

Sitting on a sofa

↓

Person is next to the sofa

↓

Person is in front of the sofa

↓

Person is sitting on the sofa

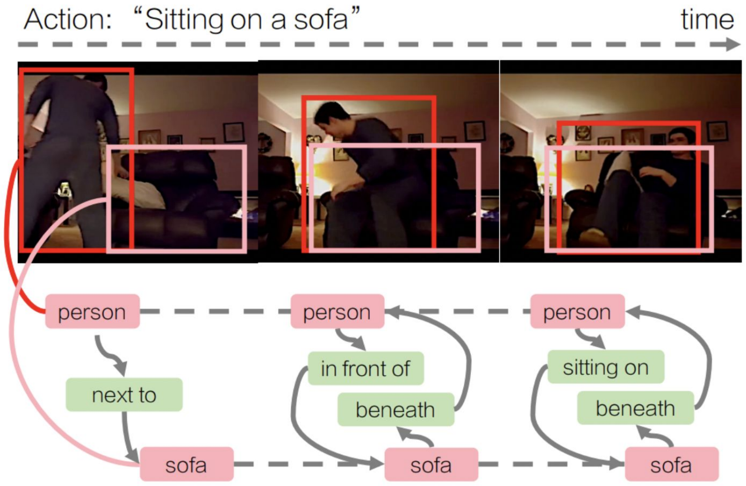


Figure: Action Genome Dataset Sample

Ji, J., Krishna, R., Fei-Fei, L., & Niebles, J. C. (2020). Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10236-10247).

Action Genome

Motivation:
No datasets includes dynamic changes in the relationships between objects to depict the event.

Goal: Understand action dynamics -> **relationship** between **object-subject pairs**

- **9848 videos** annotated with **action labels and spatio-temporal scene graph labels**
- **1.7 million human-object relations i**nstances of 25 categories
- **583K bounding boxes** of interacted objects of 35 classes.
- **265K frames** in the videos are labeled.

Ji, J., Krishna, R., Fei-Fei, L., & Niebles, J. C. (2020). Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10236-10247).

Action Genome

Relationships in Action Genome are splitted into 3 categories:

- Attention
- Spatial
- Contact

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

Table : Relationship types in Action Genome

- Attention relationships
  - Possible or ongoing interaction
- Spatial relationships
  - Spatial location
- Contact relationships
  - Type of interaction

Ji, J., Krishna, R., Fei-Fei, L., & Niebles, J. C. (2020). Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10236-10247).

Methodolgy

Input

Video Representation

Consists of several frames

- Each frame of the video at the timestamp t is represented as I_t
- Video with T frames is represented as $V = [I_1, I_2, I_3, \ldots, I_t]$

Relationship Representation

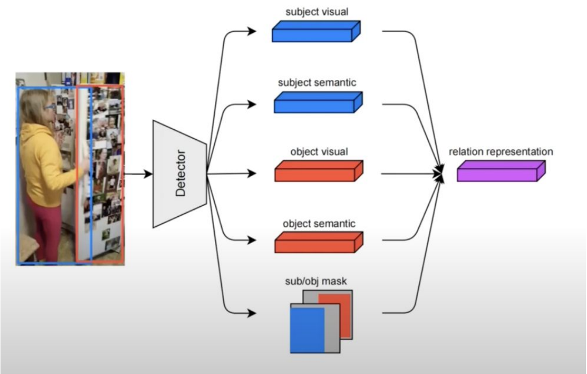
- FasterRCNN
  - generates bounding boxes.
  - extracts features of bounding boxes.

- For each frame at timestamp t, object detector proposes N(t) object proposals.
- Feature representations are depicted as :

$$[v_t^1, \ldots\ldots, v_t^{N(t)}] \quad \text{where} \quad \{v_t^1, \ldots\ldots, v_t^{N(t)}\} \in \mathbb{R}^{2048}$$

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Relationship Representation



- Between N(t) object proposals at the timestamp t,
  - There are K(t) relations :
  - $$R_t = \{r_t^1, r_t^2, \ldots, r_t^{K(t)}\}$$

- The representation vector x_t^k of the relation r_t^k represents the relationship between i-th and j-th objects proposals.

$$b_t^i,\ b_t^j \in \mathbb{R}^4 \rightarrow BBoxes$$

$$\boldsymbol{x}_t^k = \langle \boldsymbol{W}_s \boldsymbol{v}_t^i, \boldsymbol{W}_o \boldsymbol{v}_t^j, \boldsymbol{W}_u \varphi(\boldsymbol{u}_t^{ij} \oplus f_{box}(\boldsymbol{b}_t^i, \boldsymbol{b}_t^j)), \boldsymbol{s}_t^i, \boldsymbol{s}_t^j \rangle$$

Word2vec representation of class labels $\mathbb{R}^{200}$

$$W_s,\ W_o \in \mathbb{R}^{2048x512}$$

$\mathbb{R}^{2048}$

Feature Vector of object proposals i and j

flatten

$$W_U \in \mathbb{R}^{12544x512}$$

Element-wise multiplication

Convert Bounding boxes to $\mathbb{R}^{256x7x7}$

$$u_t^{ij} \in \mathbb{R}^{256x7x7}$$

Feature map of the union of the bounding boxes that is computed with ROI Align.

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Relationship Representation

$$\boldsymbol{x}_t^k = \left\langle \boldsymbol{W}_s \boldsymbol{v}_t^i, \boldsymbol{W}_o \boldsymbol{v}_t^j, \boldsymbol{W}_u \varphi(\boldsymbol{u}_t^{ij} \oplus f_{box}(\boldsymbol{b}_t^i, \boldsymbol{b}_t^j)), \boldsymbol{s}_t^i, \boldsymbol{s}_t^j \right\rangle$$

$W_s, W_o \in \mathbb{R}^{2048\text{x}512}$

$W_U \in \mathbb{R}^{12544\text{x}512}$

flatten

256x7x7=12544

Element-wise multiplication

256x7x7=12544

R^{12544}

$\mathbb{R}^{512}$ $\mathbb{R}^{512}$ $\mathbb{R}^{512}$ R^{200} R^{200}

$\mathbb{R}^{1936}$

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Spatio-Temporal Transformer

- Spatial Encoder
- Frame Encoding
- Temporal Decoder

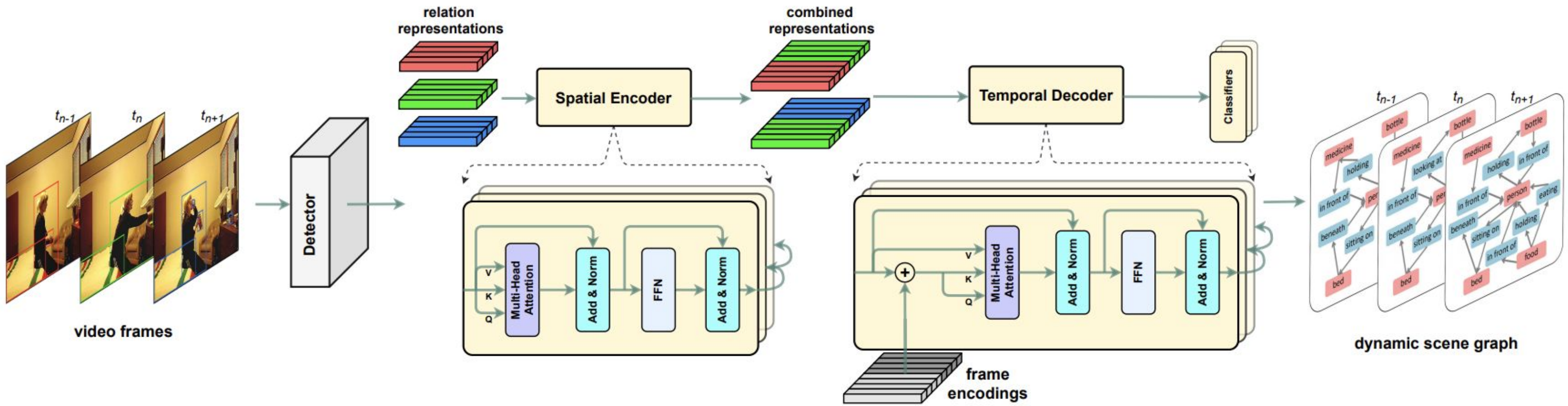


Figure: Spatio-Temporal Transformer Model Architecture

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Spatial Encoder

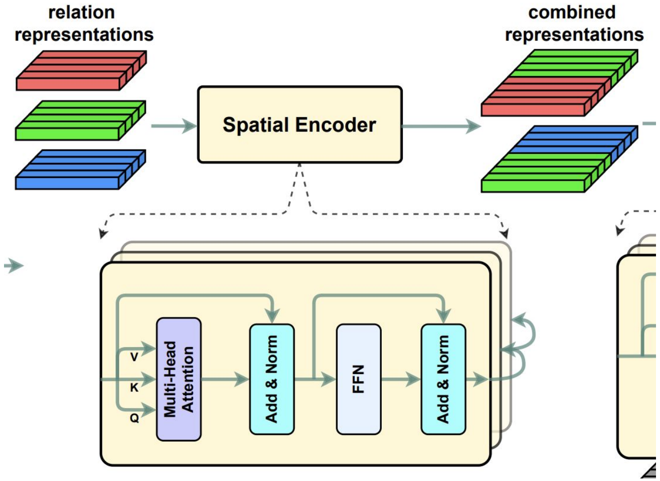


Figure: Spatial Encoder Architecture

Goal: Learning spatial context within a frame
Architecture: Classic Transformer Encoder Layer

- No positional embeddings is used

Input : $X_t = \{\boldsymbol{x}_t^1, \boldsymbol{x}_t^2, \ldots, \boldsymbol{x}_t^{K(t)}\}$

$\mathbb{R}^{1936}$

Number of Relationships at timestamp t

$$\boldsymbol{X}_t^{(n)} = Att_{enc.}(\boldsymbol{Q} = \boldsymbol{K} = \boldsymbol{V} = \boldsymbol{X}_t^{(n-1)})$$

Output of n-th encoder layer

Query

Key

Value

Output of (n-1)th encoder layer

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Frame Encodings

- **Motivation:**
  - Transformers are unaware of temporal dependencies
  - Model should leverage positional information
- **Goal :**
  - Inject the temporal position to the relationship representations
- Used only in the Temporal Decoder
- Custom learned embeddings
- Same size as relation representation vectors
- Number of embedding vectors is fixed and equal to sliding windows size

$$\boldsymbol{E}_f = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_\eta],$$

Frame Encodings

$\mathbf{e}_1 \in \mathbb{R}^{1936}$

Window size

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Temporal Decoder

- **Goal:**
  - Capture temporal dependencies **between frames**
- Sliding Window approach is used
  - Batch adjacent frames
  - **Motivation**
    - Reducing Memory consumption
    - Irrelevant information from far frames involves
- No masked decoder layer unlike original transformer decoder

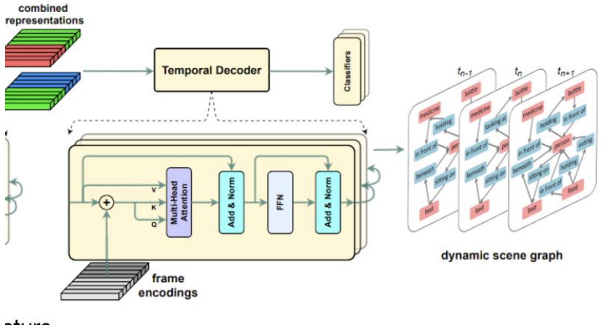


Figure: Spatial Decoder Architecture

$$\boldsymbol{Z}_i = [\boldsymbol{X}_i, \ldots, \boldsymbol{X}_{i+\eta-1}], i \in \{1, \ldots, T - \eta + 1\}$$

I-th generated input batch

Contextualized all representations in the i-th frame

Video size

Window size

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Temporal Decoder

$$\boldsymbol{Z}_i = [\boldsymbol{X}_i, \ldots, \boldsymbol{X}_{i+\eta-1}], i \in \{1, \ldots, T-\eta+1\}$$

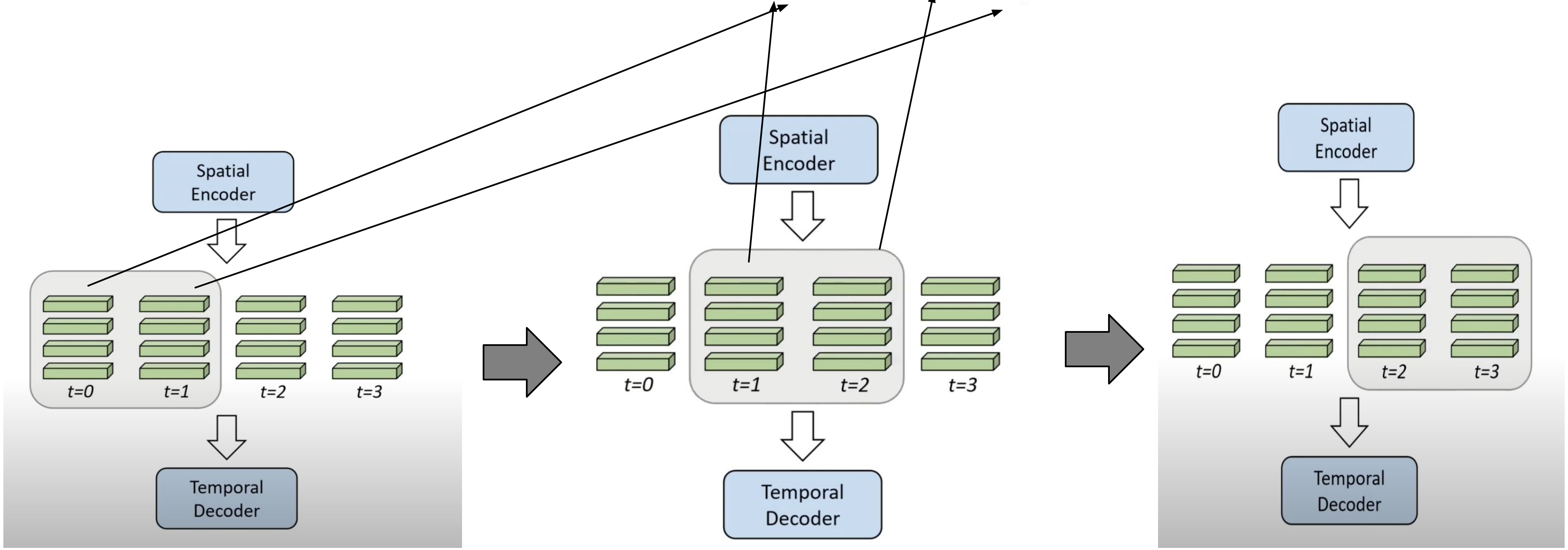


Figure: Decoder Sliding Window Approach

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

$$\boldsymbol{Z}_i = [\boldsymbol{X}_i, \ldots, \boldsymbol{X}_{i+\eta-1}], i \in \{1, \ldots, T - \eta + 1\}$$

I-th generated input batch

All contextualized representations in the i-th frame

Window size

Video size

Window size

Decoder Attention Computation

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{K} = \boldsymbol{Z}_i + \boldsymbol{E}_f, \\ \boldsymbol{V} &= \boldsymbol{Z}_i, \\ \hat{\boldsymbol{Z}}_i &= Att_{dec.}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}). \end{aligned}$$

$$\boldsymbol{E}_f = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_\eta],$$

Frame Encodings

$\mathbb{e}_1 \in \mathbb{R}^{1936}$

Window size

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Loss

Relationship Types

- Attention
- Spatial
- Contacting

- **Predication Classification**
  - Different linear transformations are applied to each relationship type

$$L_p(r, \mathcal{P}^+, \mathcal{P}^-) = \sum_{p \in \mathcal{P}^+} \sum_{q \in \mathcal{P}^-} max(0, 1 - \phi(r,p) + \phi(r,q))$$

Subject-Object Pair

Annotated Predicates

set of the predicates not in the annotation

Computed score of pth predicate

- **Classification Loss**
  - Object distribution -> two fully-connected layers with a ReLU activation and a batch normalization in between.
  - Cross entropy loss.

$$L_{total} = L_p + L_o$$

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Scene Graph Generation Strategies

- **With Constraint**
  - ➔ Only one predicate can be assigned to object-subject pair.
  - ➔ Assess predicting
    - ◆ **the most important relationship.**

Predicate	Confidence
eating	0.72
holding	0.21
standing	**0.88**

- **Without Constraint**
  - ➔ Multiple predicates can be assigned to object-subject pairs.
  - ➔ Possibility of adding noise and wrong information to the graph.

Predicate	Confidence
eating	0.72
holding	0.21
standing	**0.88**

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Scene Graph Generation Strategies

Semi Constraint

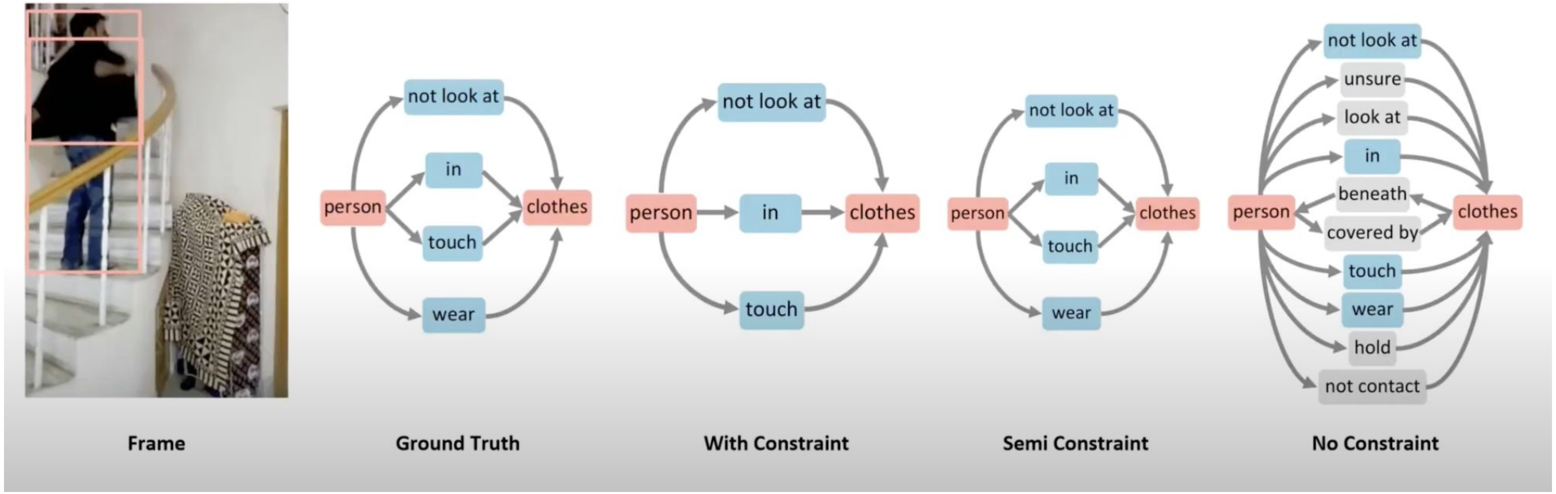
- Novel strategy
- Multiple predicates can be assigned to the subject-object pair.
  - the person (object), and food(subject) pair
    - <person "eating"   food>
    - <person “holding" food>
  - Threshold confidence of the predicates
  - Confidence > threshold -> Positive Predicate

Predicate	Confidence
eating	**0.72**
holding	0.21
standing	**0.88**

Threshold = 0.70

Bigger than the threshold

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).


not look at
in
touch
wear
person
clothes
not look at
in
touch
person
clothes
not look at
in
touch
wear
person
clothes
not look at
unsure
look at
in
beneath
covered by
touch
wear
hold
not contact
person
clothes
Frame
Ground Truth
With Constraint
Semi Constraint
No Constraint

Experiments

Evaluation Metrics

- **Predicate Classification (PREDCLS)**
  - Ground Truth Bounding Boxes and class information is given to model.
  - Model predict
    - Predicate labels
- **Scene Graph Classification (SGCLS)**
  - Ground Truth Bounding Boxes are given.
  - Model predicts :
    - Predicate labels
    - Class information of bounding boxes
- **Scene Graph Detection (SGDET)**
  - Model detects bounding boxes
  - Class Information of the bounding boxes
  - Predicate Labels
- **R@k: Recall for top k confident predictions (e.g. R@20, R@50)**

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Comparison to State of the art Single Image Based Methods

Method	With Constraint									No Constraint								
	PredCLS			SGCLS			SGDET			PredCLS			SGCLS			SGDET		
	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50
VRD	51.7	54.7	54.7	32.4	33.3	33.3	19.2	24.5	26.0	59.6	78.5	99.2	39.2	49.8	52.6	19.1	28.8	40.5
Motif Freq	62.4	65.1	65.1	40.8	41.9	41.9	23.7	31.4	33.3	73.4	92.4	**99.6**	50.4	60.6	64.2	22.8	34.3	46.4
MSDN	65.5	68.5	68.5	43.9	45.1	45.1	24.1	32.4	34.5	74.9	92.7	99.0	51.2	61.8	65.0	23.1	34.7	46.5
VCTREE	66.0	69.3	69.3	44.1	45.3	45.3	24.4	32.6	34.7	75.5	92.9	99.3	52.4	62.0	65.1	23.9	35.3	46.8
RelDN	66.3	69.5	69.5	44.3	45.4	45.4	24.5	32.8	34.9	75.7	93.0	99.0	52.9	62.4	65.1	24.1	35.4	46.8
GPS-Net	66.8	69.9	69.9	45.3	46.5	46.5	24.7	33.1	35.1	76.0	93.6	99.5	53.6	63.3	66.0	24.4	35.7	47.3
STTran	**68.6**	**71.8**	**71.8**	**46.4**	**47.5**	**47.5**	**25.2**	**34.1**	**37.0**	**77.9**	**94.2**	99.1	**54.0**	**63.7**	**66.4**	**24.6**	**36.2**	**48.8**

Table : Comparison of STTran with SOTA Models

- **Result:**
  - STTRan overperforms all other image based SOTA methods by using temporal relationships between frames.

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Hypothesis: Is using temporal relationship easy?

- **Setup:** Add LSTM/RNN on top of SOTA models.
- **Goal:** Using temporal information with LSTM/RNN
- **Result:** All methods improve their scene graph generation capability by leveraging temporal aspects.

Method	PredCLS-R@20	
	original	+LSTM
Motif Freq	65.1	65.2
MSDN	68.5	68.8
RelDN	69.5	69.7
GPS-Net	69.9	70.4

Table : Comparison of methods after adding LSTM on top.

CAMP

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Hypothesis: Does model leverage temporal dependencies?

- **Setup**: Shuffle or reverse ⅓ of training instances
- **Idea**: If model uses temporal information, adding noise to training samples will degrade the performance.
- **Result**: adding noise to the temporal information lowers the performance of the STTran.

Normal Video	Processed Video	Processing	PredCLS-R@20
2/3	1/3	shuffle	70.6
2/3	1/3	reverse	71.0
1	-	-	**71.8**

Table : Results of when 1/3 of training instances are shuffled or reversed

(a) spatial encoder only

(b) complete STTran

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Ablation Study

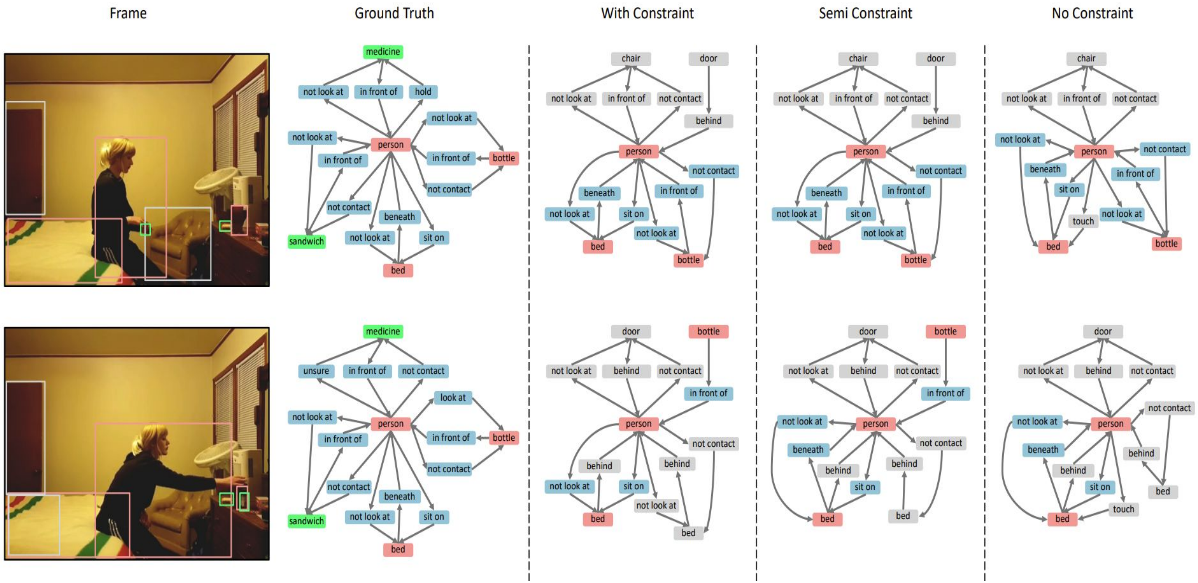
Spatial Encoder	Temporal Decoder	Frame Encoding	PredCLS-R@20 With	PredCLS-R@20 Semi	SGDET-R@20 With	SGDET-R@20 Semi
✓	-	-	69.6	78.7	32.9	35.1
-	✓	-	71.0	82.2	33.7	35.5
✓	✓	-	71.3	82.7	33.8	35.6
✓	✓	sinusoidal	71.3	82.8	33.9	35.7
✓	✓	learned	**71.8**	**83.1**	**34.1**	**35.9**

Table : Ablation Study

- Only Spatial Encoder w/o frame encodings -> similar performance to image-based models.
- Temporal decoder w/o frame encoding > only spatial encoder
- Spatial Encoder & Temporal Decoder -> Performance increase
- Learned embeddings > sinusoidal embeddings.

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Qualitative Results



Green boxes in the ground truth represent the objects that can **not be found** by the object detector.

Gray colors are **false positive detections** and therefore their relations are false positive.

The melons are the true positive boxes and correct relationships are represented with light blue.

Figure: Qualitative results of the model on the video where the woman tries to reach the medicine while sitting on the bed.

Cong, Y., Liao, W., Ackermann, H., Rosenhahn, B., & Yang, M. Y. (2021). Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 16372-16382).

Take Home Message

Take Home Message

- Temporal information helps to understand the relationship between the objects and subjects in the videos.
- Using temporal information leads to create more accurate scene graphs
- Multi-Label Visual Relationship Prediction
- SOTA results on dynamic Scene Graph Generation
- Having hypotheses and examining them with experiments make the paper more convincing.
- Qualitative experiments always gets the attraction and makes the paper more engaging.
- Ablations helps to understand the contributions of each modality.

- Why STTran does not overperform SOTA in some relations (holding) ?
- What might be the annotation problems that they mention in the supplementary material? And why they did not elaborate?

Thank you for listening …

Technische Universität München + Johns Hopkins Whiting School of Engineering