Co-Segmentation Aided Two-Stream Architecture for Video Captioning

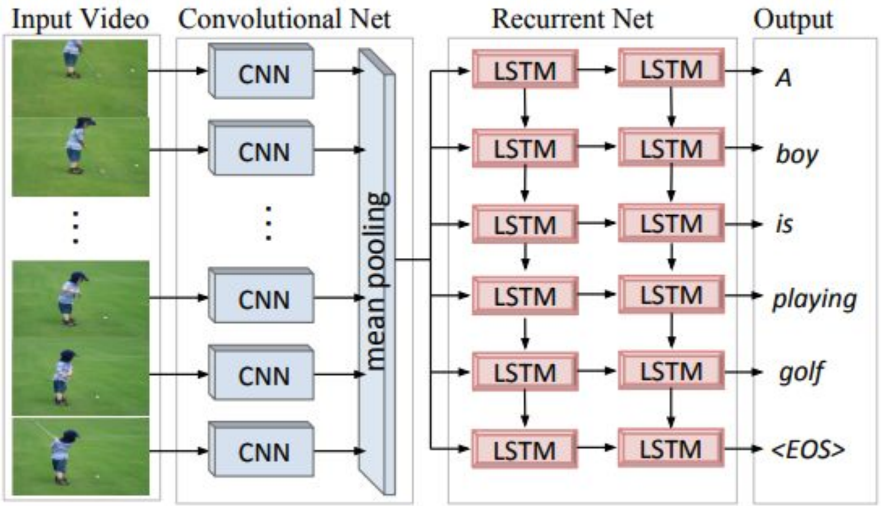
Jayesh Vaidya, Arulkumar Subramaniam, Anurag Mittal
Department of CSE, IIT-Madras, Chennai, India

Problem Overview

- Describing content of the video with natural language sentence.

Caption: A baseball player hits a baseball.

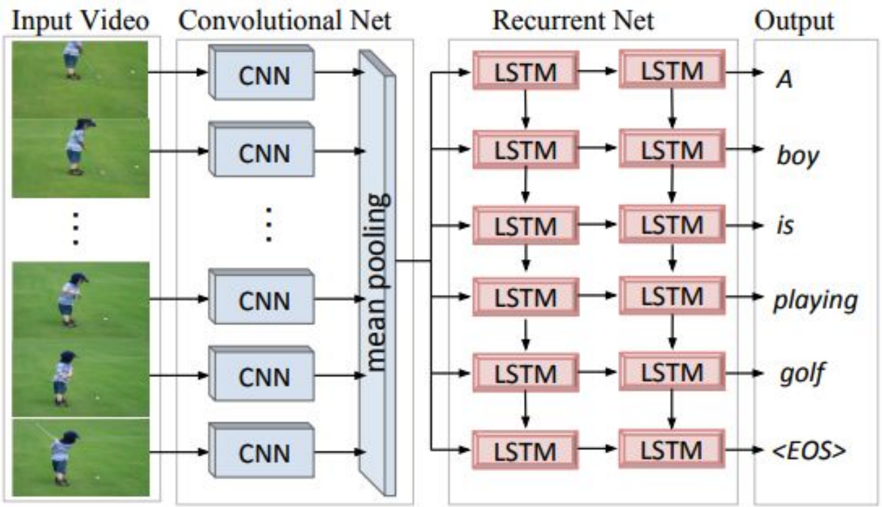
Earlier Work



[Venugopalan et al. 2014, Venugopalan et al. 2015, Chen et al. 2018, Pei et al. 2019]

Frame level features to generate captions

Earlier Work



[Venugopalan et al. 2014, Venugopalan et al. 2015, Chen et al. 2018, Pei et al. 2019]

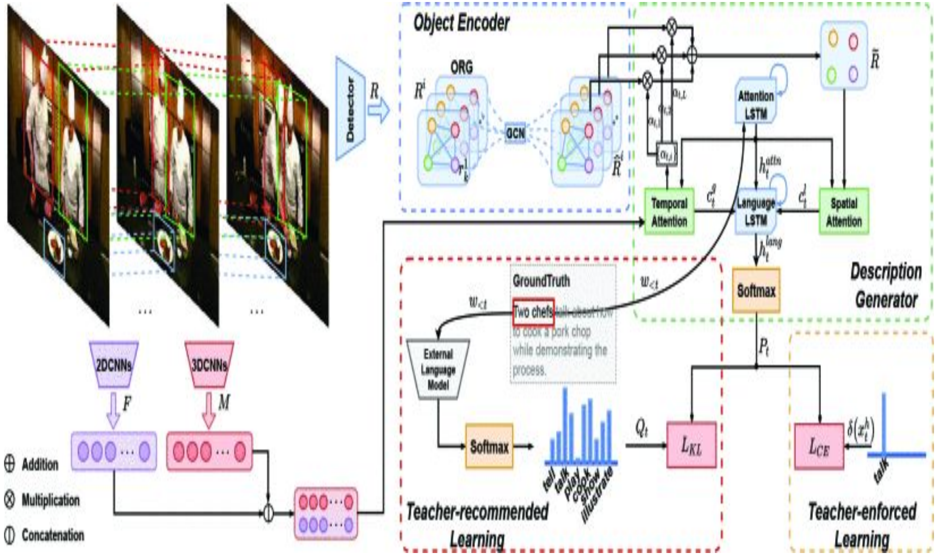
Frame level features

- Generating captions for video not only involves understanding of visual and temporal cues.
- But also object level features and interaction of these objects in spatio-temporal dimension.

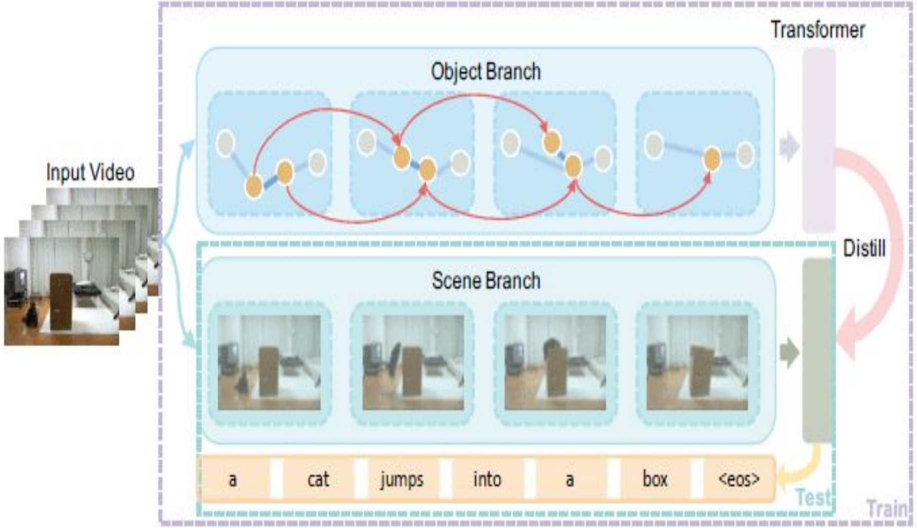
Recent Works

Enhance captions using BERT model.

Two steam Arch. with late fusion



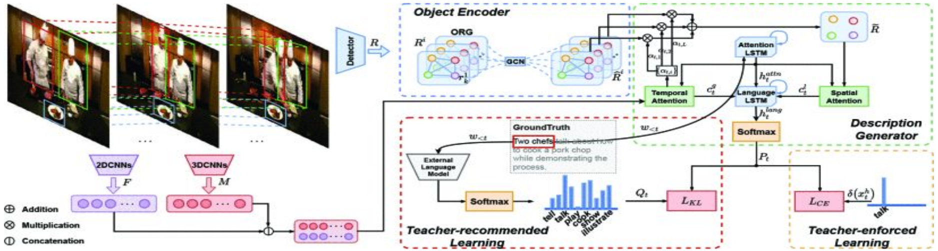
[Zhang et al. 2020]

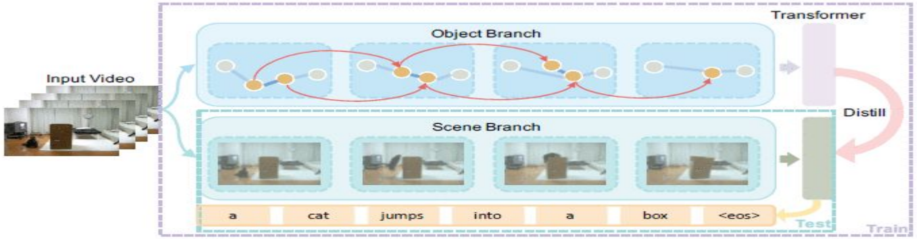


[Pan et al. 2020]

Object interactions using GCNs

Recent Works



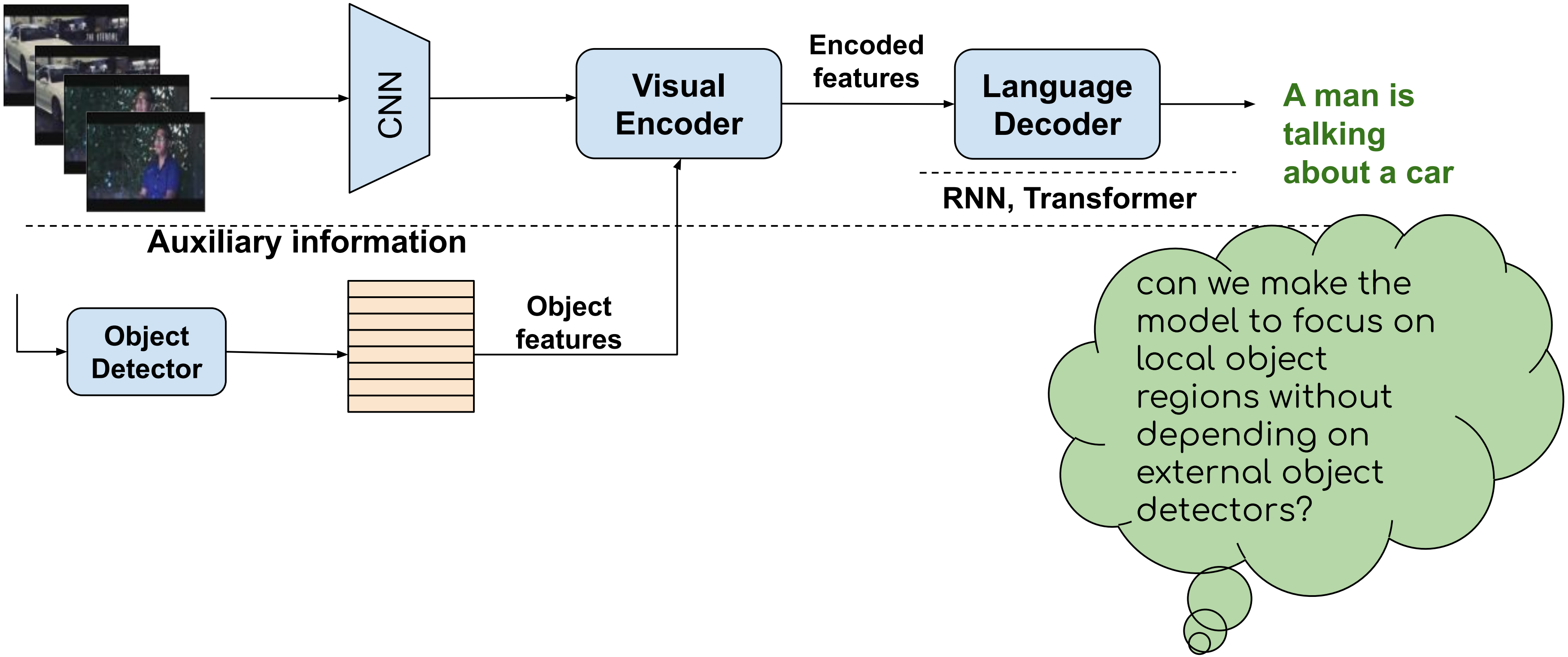


Zhang et al. 2020

Pan et al. 2020

- Object level information enhances visual encoding.
- But, features extracted using pretrained object detectors.
  - May not capture all object categories needed.
  - Can introduce bias.

Common Video Captioning Pipeline



CNN

Visual Encoder

Encoded features

Language Decoder

A man is talking about a car

RNN, Transformer

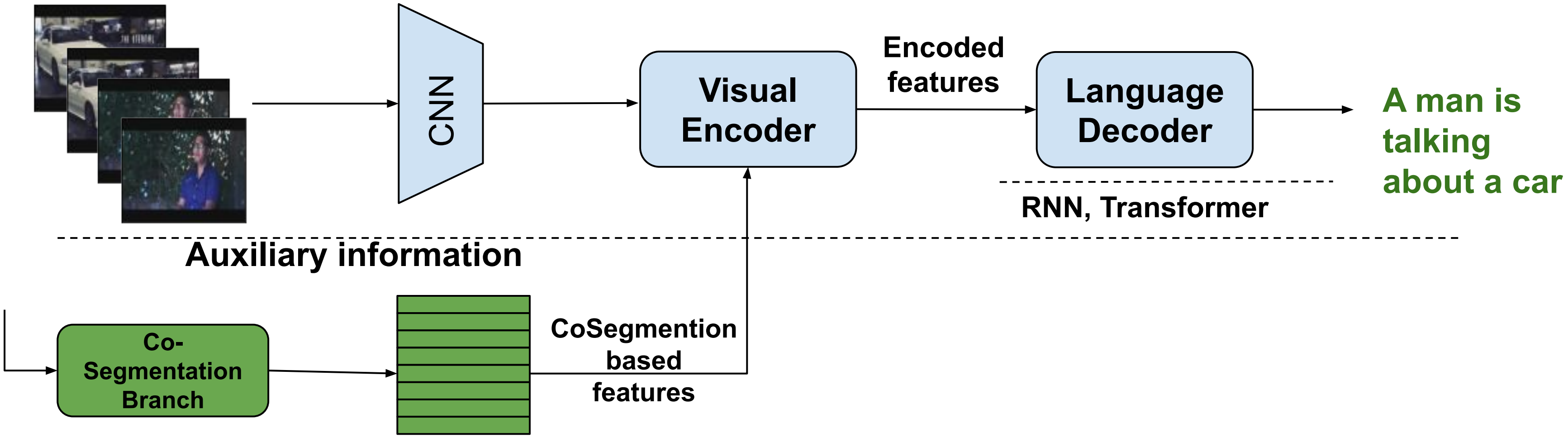
Auxiliary information

Object Detector

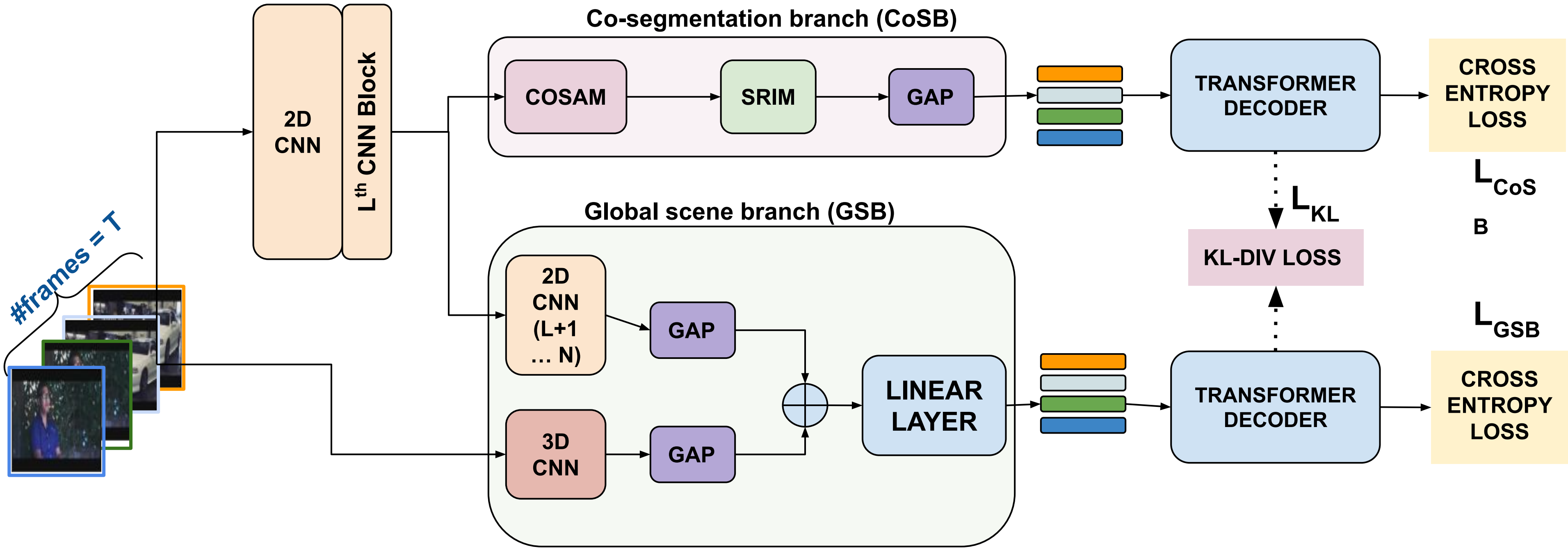
Object features

can we make the model to focus on local object regions without depending on external object detectors?

Common Video Captioning Pipeline

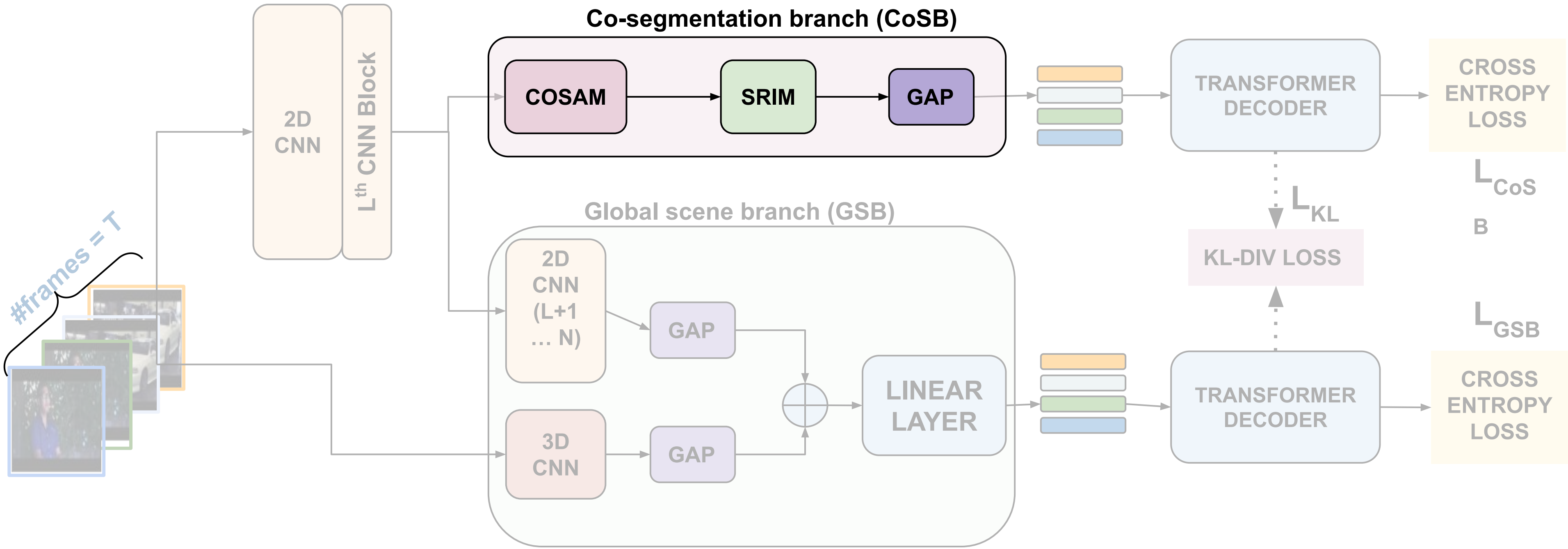


Architecture



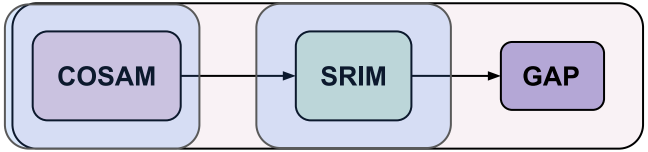
*Spatio-temporal graph for video captioning with knowledge distillation.,Pan et. al CVPR-2020

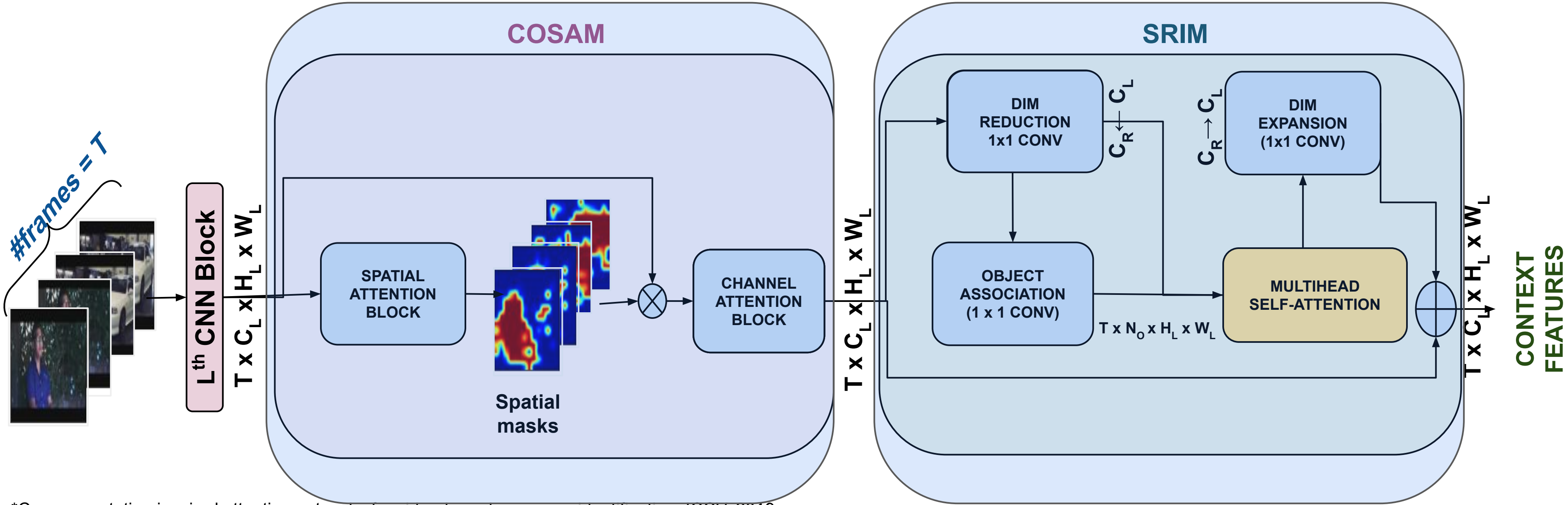
Architecture



Our Work

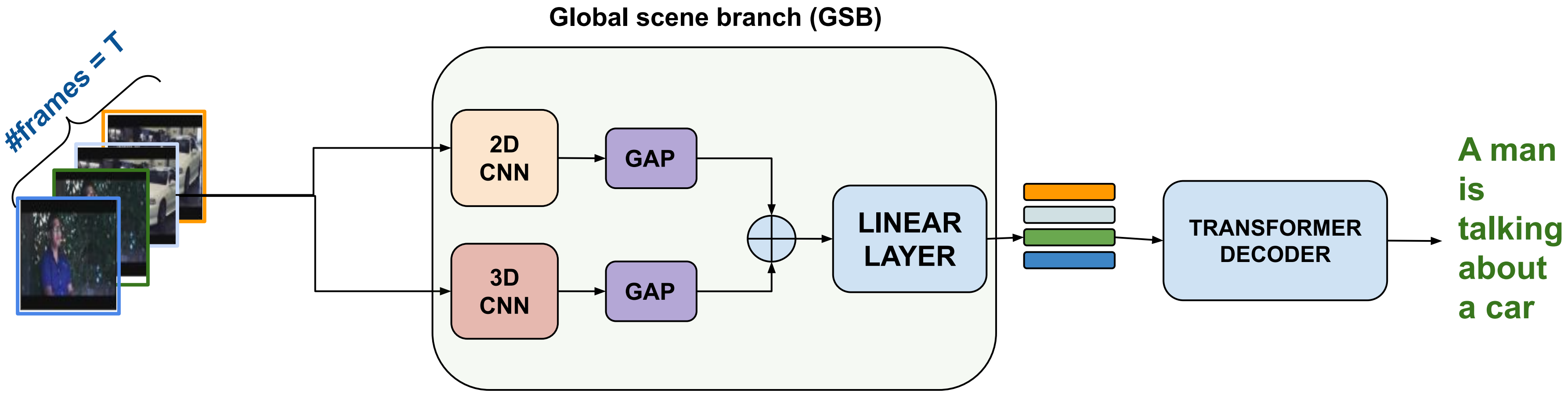
Co-segmentation branch (CoSB)





*Co-segmentation inspired attention networks for video-based person re-identification. ICCV, 2019.

Our Work - Testing

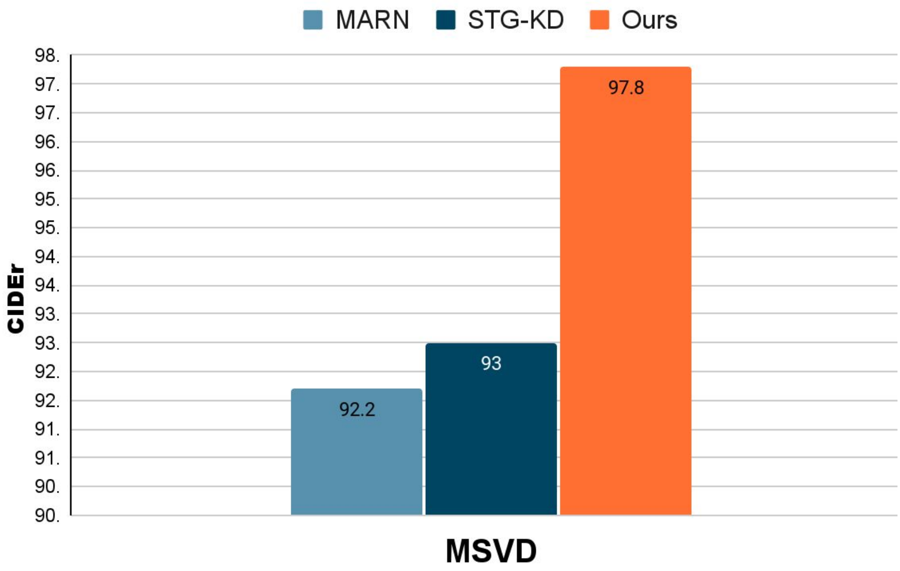


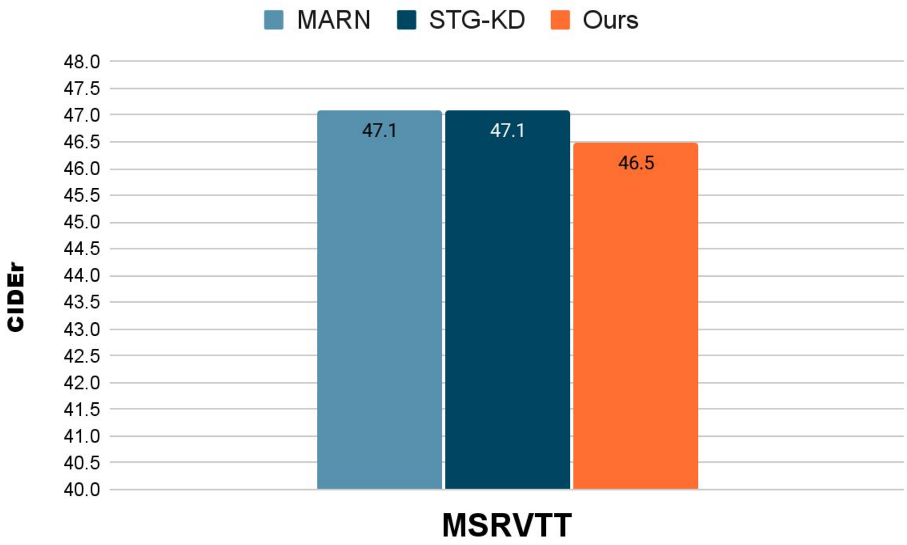
Dataset

	#Videos	Train/Val/Test	#Sentences/Video
MSVD	1970	1200/100/670	~40
MSR-VTT	10000	6513/497/2990	20

Microsoft Video-Description Corpus (MSVD)
Microsoft Research Video-to-Text (MSR-VTT)

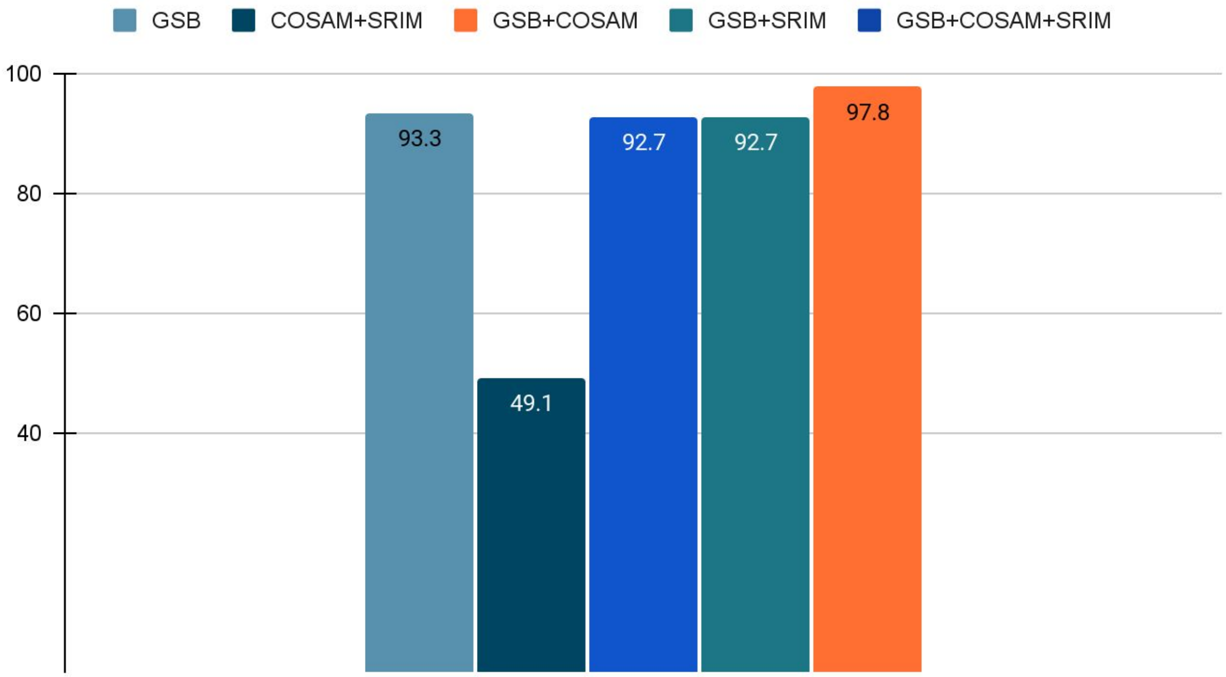
Quantitative Results





We achieve state-of-the-art performance on MSVD and get competitive results on MSRVTT.

Ablations

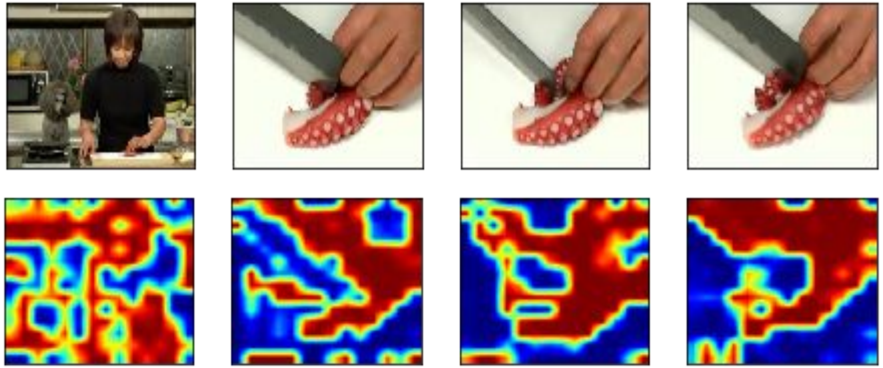


Ablation on MSVD dataset

Our complete model gives the best result.

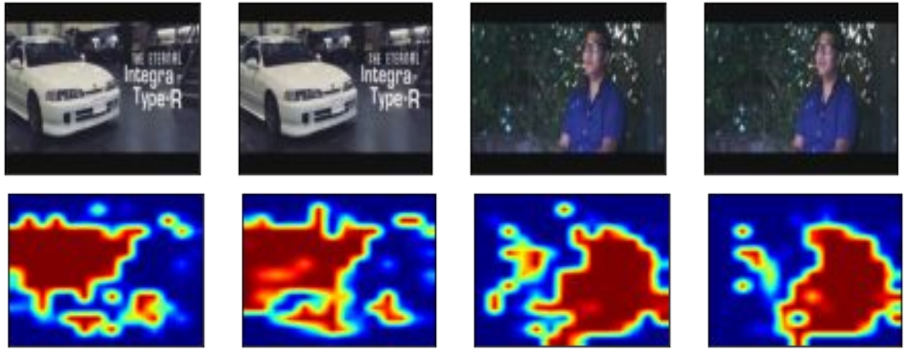
Qualitative Results - Salient Regions (COSAM)

- Better at localizing unusual objects.
- Our model correctly detects octopus, which is not usually in object detector datasets.

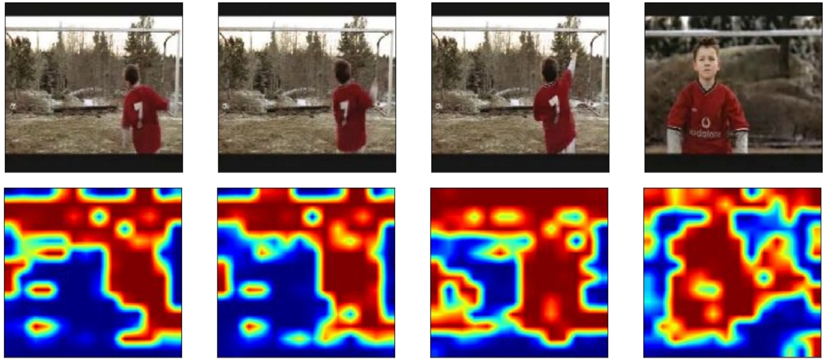


GT: A woman is slicing octopus.
Ours: A woman is slicing **octopus**.
STG-KD: A woman is slicing carrots.

Qualitative Results - Salient Regions (COSAM)

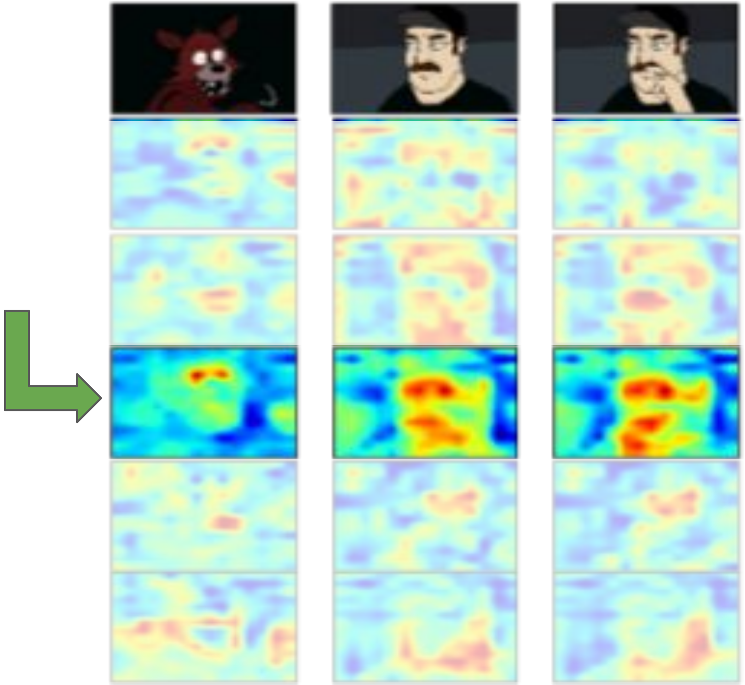


GT: A man is talking about a car.
Ours: A man is talking about a car.
STG-KD: A man is talking about a car.

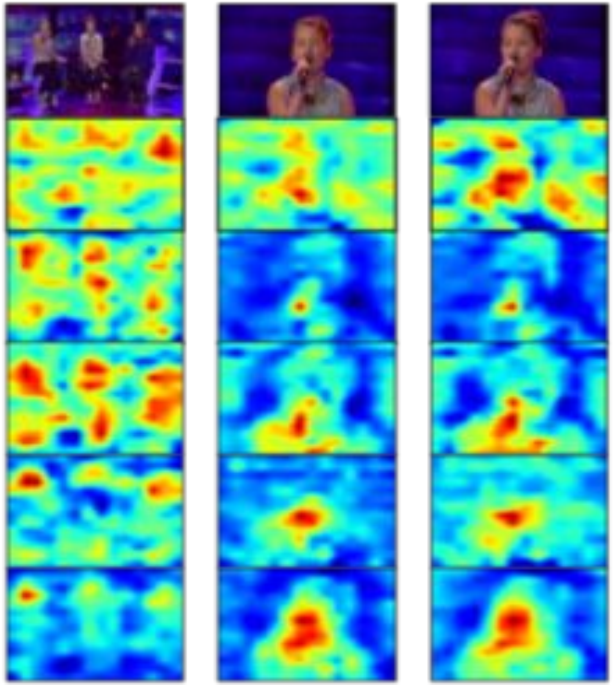


GT: A boy is kicking a soccer ball.
Ours: A boy is kicking a **soccer** ball.
STG-KD: A boy kicks a goal.

Qualitative Results - Object descriptors (SRIM)



Ours: A cartoon creature is talking to a man.
GT: A cartoon with a creature is running at a man.



Ours: A girl is singing on a stage.
GT: A girl is singing on a stage

Conclusion

- We proposed an end-to-end network to capture local salient regions in contrast to using pretrained object detectors.
- Visualizations show that co-segmentation is indeed able to capture salient regions including tail distribution objects.
- Competitive results on benchmarks without the usage of pretrained object detectors.

Thank you! ☺