Segmentation of **Moving** Objects

CS7.401.S22 Computer Vision - Final Project Presentation

May 3rd, 2022

ADITYA ● DHRUV ● ROHIT

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

IIIT HYDERABAD

Agenda

To share our ideas, approach, and results for implementing the course project: Segmentation of moving objects.

- Problem description
- Datasets available
- Segmentation under constraints
- MATNet
- Reciprocal Transformation Net
- TransVos
- Compiled Results
- Conclusion

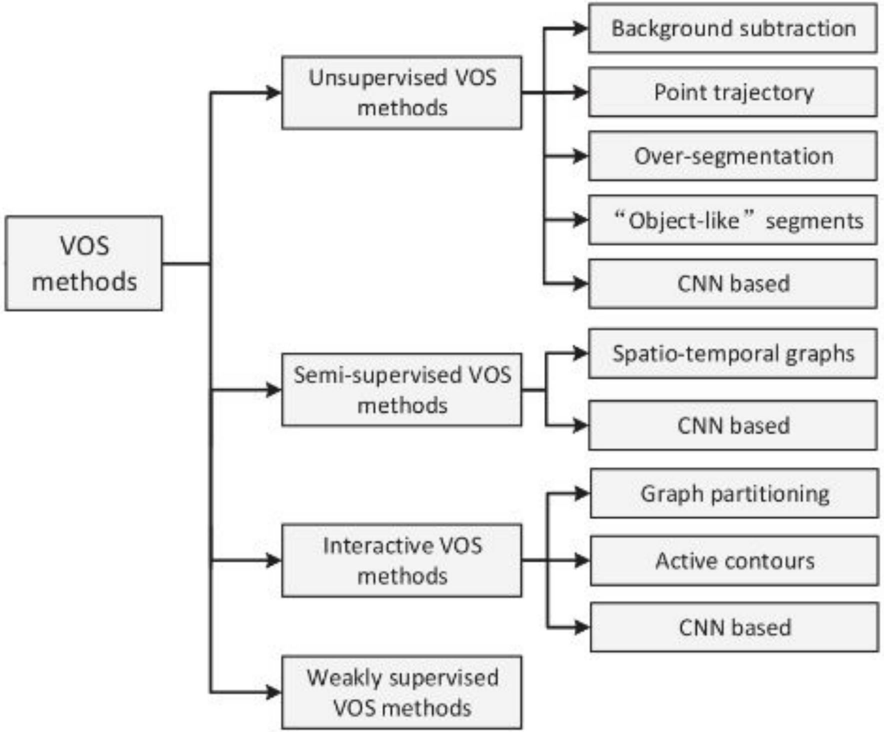
Problem **Description**

Brief **introduction** to Video Object Segmentation.

Segmentation of moving objects

Also known as VOS: Video Object Segmentation

- Aim is to segment the objects in a video that exhibit independent motion in at least one frame. The moving objects/pixels belong to foreground and the remaining pixels are the background. *This problem becomes challenging when background is also in motion since both foreground and background pixels interfere.*
- This problem is tackled in many ways and broadly categorized into methods based on training and type of architectures.
- Problem formulation-
  - Formally, let χ and y denote the input space and output segmentation space, respectively. Video segmentation solutions generally seek to learn an ideal video-to-segment space mapping $f^* : \mathcal{X} \longrightarrow \mathcal{Y}$.



Available **Datasets**

Overview of datasets **encountered** and **used** in this project.

VOS targeted datasets

Collected from multiple sources

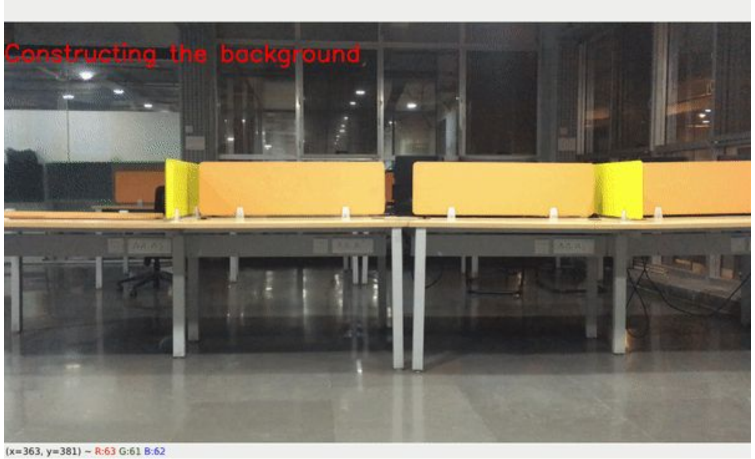
- **Davis16, 17, 18**: *DAVIS 2016* and *2017* datasets are some of the most popular datasets for training and evaluating video object segmentation algorithms. DAVIS 2016 dataset contains 50 full high quality video sequences with 3,455 annotated frames in total, and focuses on single-object video object segmentation, that is, there is only one foreground object per video. DAVIS 2017 dataset complements DAVIS 2016 dataset training and validation sets with 30 and 10 high-quality videos, respectively. It also provides an additional 30 development test sequences and 30 challenge test sequences. *DAVIS-18* adds 100 more videos with multiple objects per video.

- **FBMS**: FBMS-59 (Freiburg-Berkeley motion segmentation) are widely used for unsupervised and semi-supervised VOS methods. FBMS-59 is an extension of BMS-26 dataset that consists 26 videos with a total of 189 annotated image frames, with shots from movie stories and the 10 vehicles and 2 human sequence.

- **YoutubeVos:** This is a large-scale dataset series for VOS, with long-range video sequences. It considers more diverse objects and context and each video sequence in the datasets has a greater number of frames than any other datasets, allowing VOS methods to model and exploit long-range temporal dependency between frames.

VOS under constrained environment

- By constrained environment, we mean the camera is fixed and lighting does not change significantly throughout the deployment.
- With such constraints, we can develop a background using Weighted-Frame-Averaging and using this constructed background, we can subtract the incoming frames and extract the silhouettes of moving objects.
- This technique is widely used for vehicle detecting using existing CCTV infrastructure and local offices for tracking employee's activities.
- We implemented these on our own data and obtained the following results.

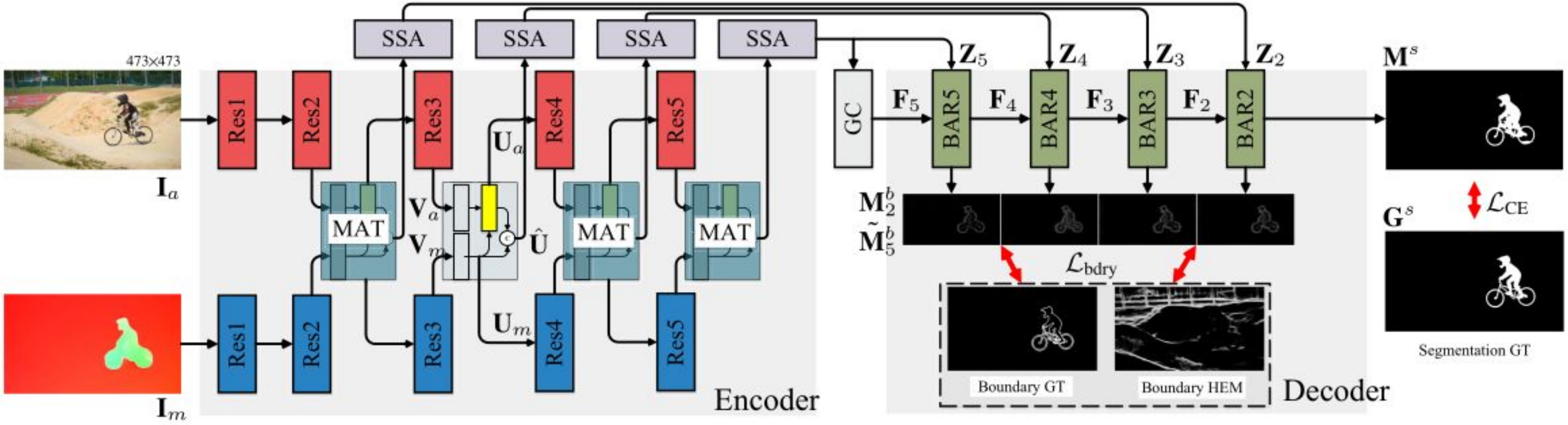


(x=363, y=381) ~ R:63 G:61 B:62

VOS under constrained environment (Result)

MATNet (UVOS)

- Its inspired by the human vision system (HVS) and leverages motion cues as bottom-up signal to guide the perception of object appearance.
- **Encoder**: Uses an asymmetric attention block within a two-stream encoder network to learn powerful spatio-temporal object representations for ZVOS.
- Introduce a **bridge network** that is responsible for selecting informative spatiotemporal features for the decoder. It is built upon several SSA modules.
- **Decoder**: A boundary-aware decoder that tries to obtain segmentation with crisp object boundaries.



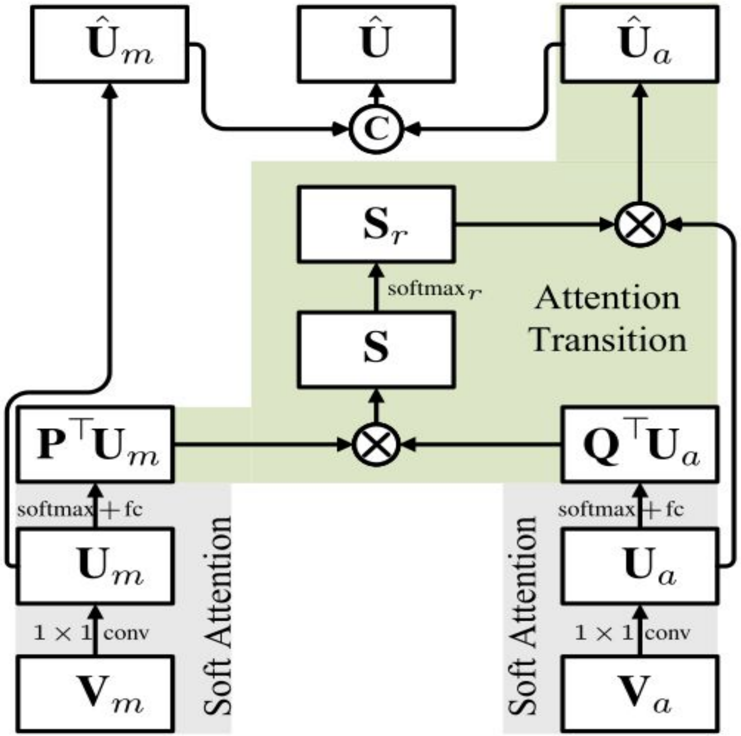


Fig. 2: MAT module

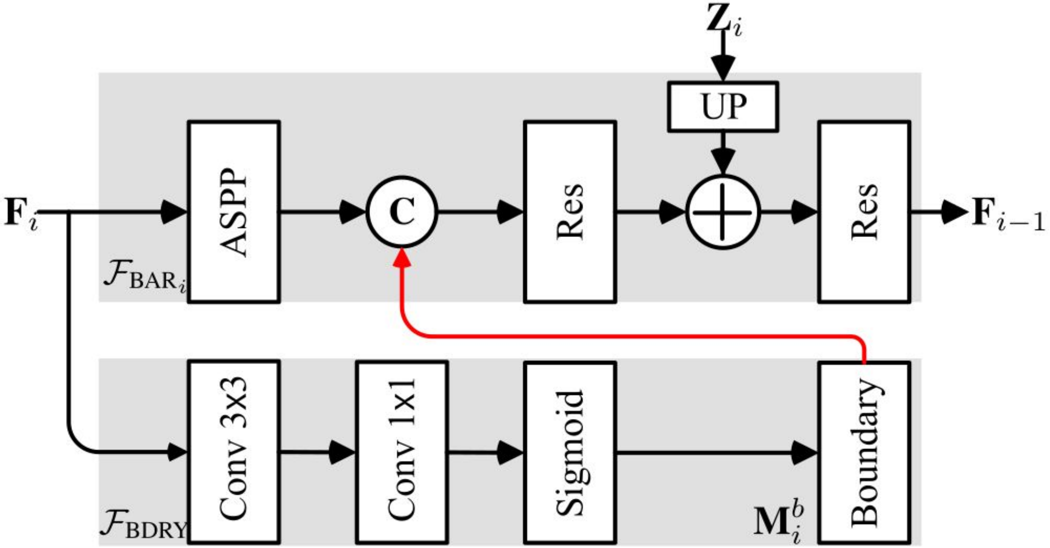
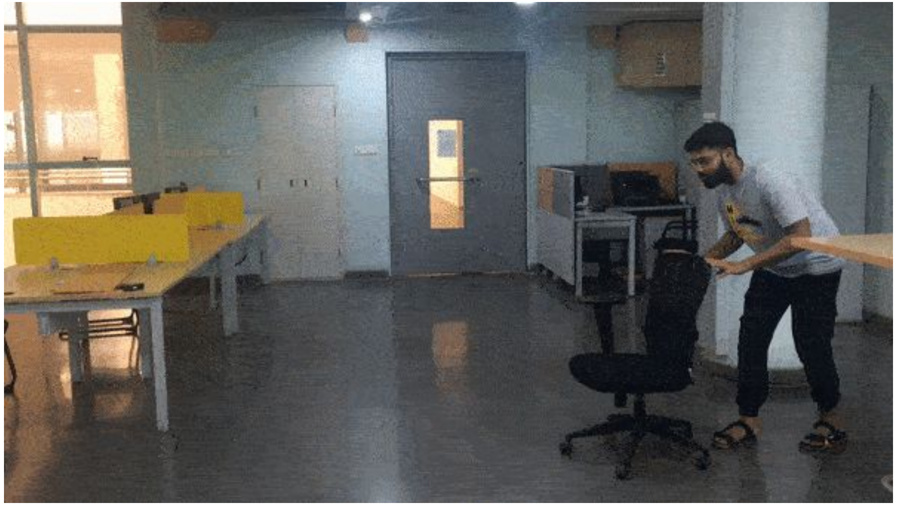
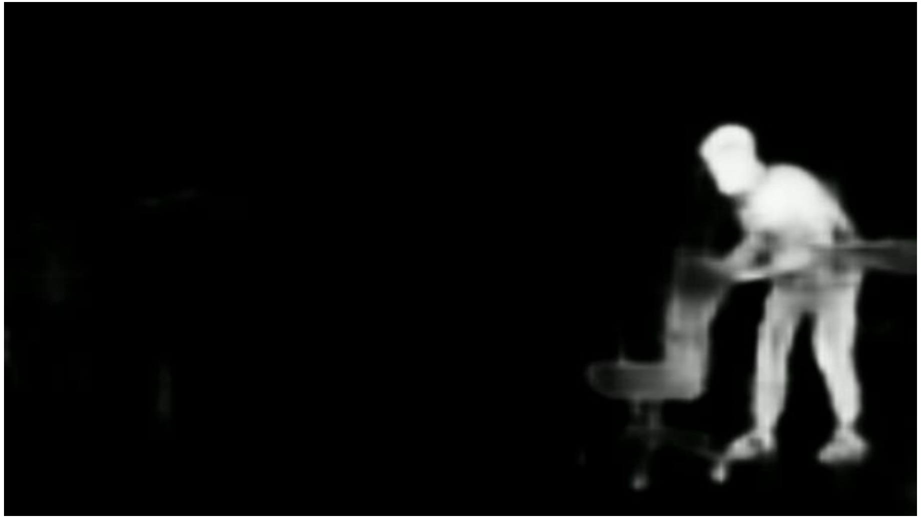


Fig. 3: BAR Module

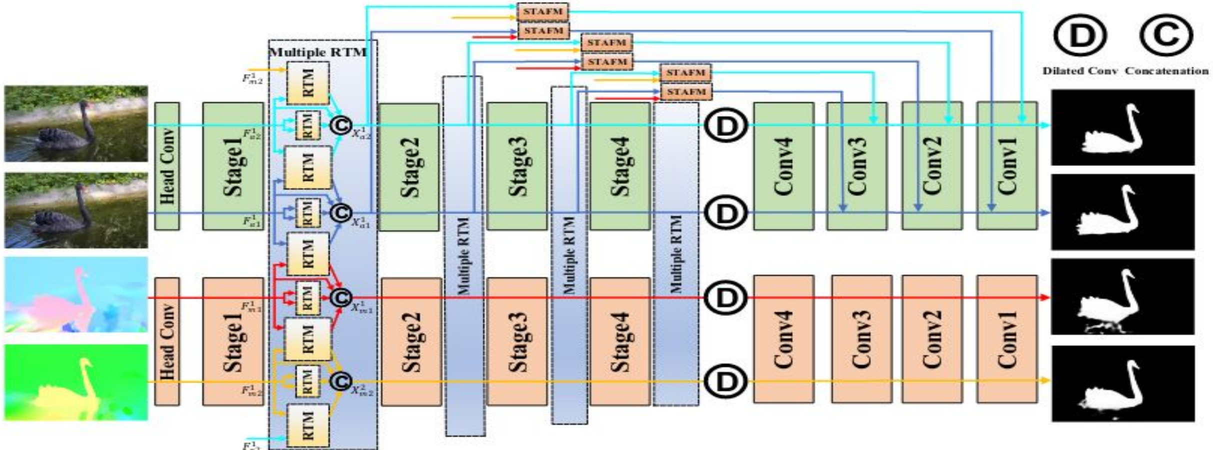
MATNet (Results)



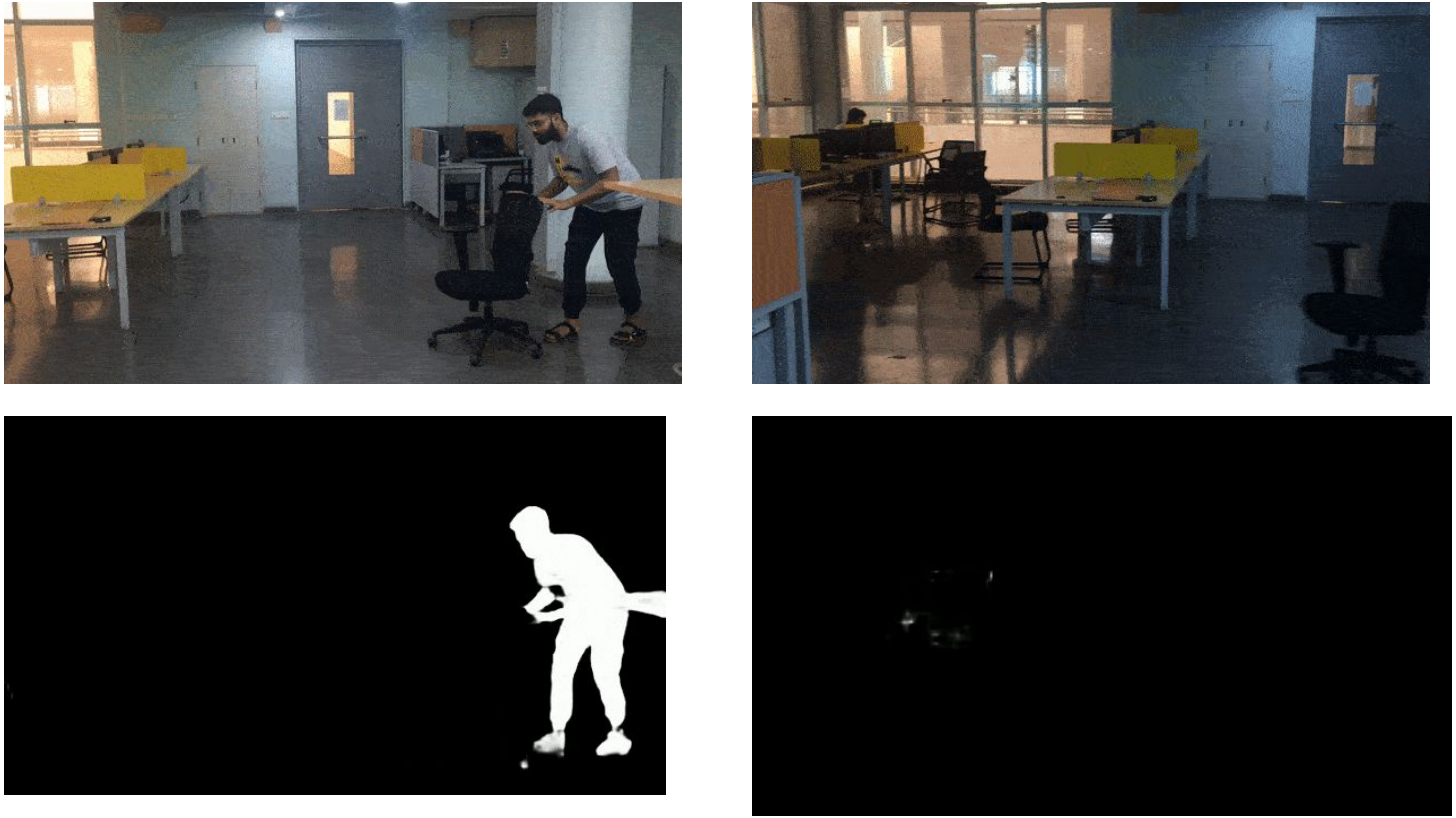


RT Net (Reciprocal Transformation Net)

- RTNet discovers the primary objects by correlating three key factors: the intra-frame contrast, the motion cues, and temporal coherence of recurring objects.
- Consists of two modules:
  - **RTM** (Reciprocal Transform Units): RTM allows in-domain and cross-domain feature interactions. This calculates similarities for all pairwise features, such as motion-motion, appearance-appearance, and appearance-motion.
  - **STAFM** (Spatial Temporal Attentive Fusion Module): STAFM is used to leverage the appearance and motion features from the corresponding encoder state, and segment spatio-temporally consistent primary objects.
- The model takes in the consecutive frames and also their forward and backward optical flows. These are passed through an encoder with RTM module in between. The skip connections between the Encoder-Decoder consist of STAFM module to retain the global information.

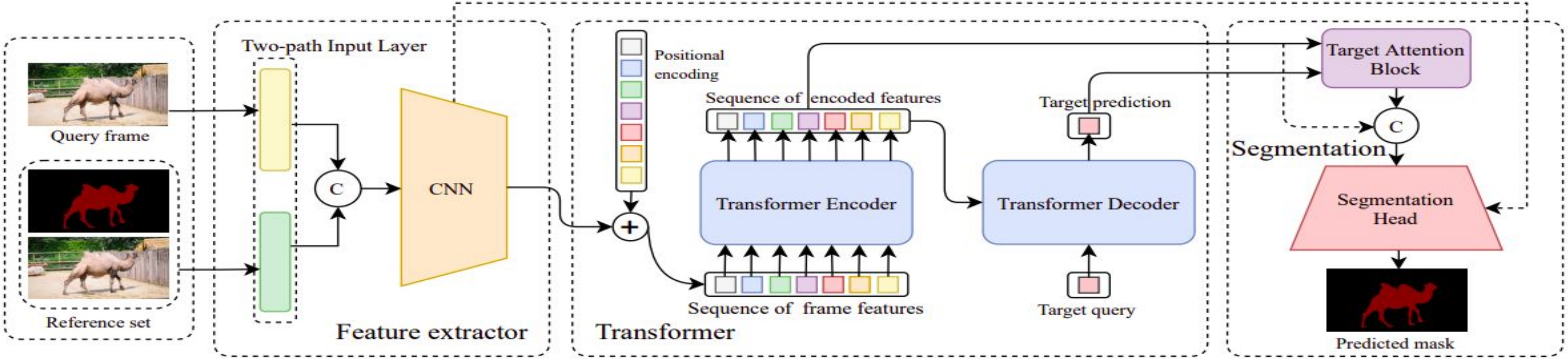


RT Net Results

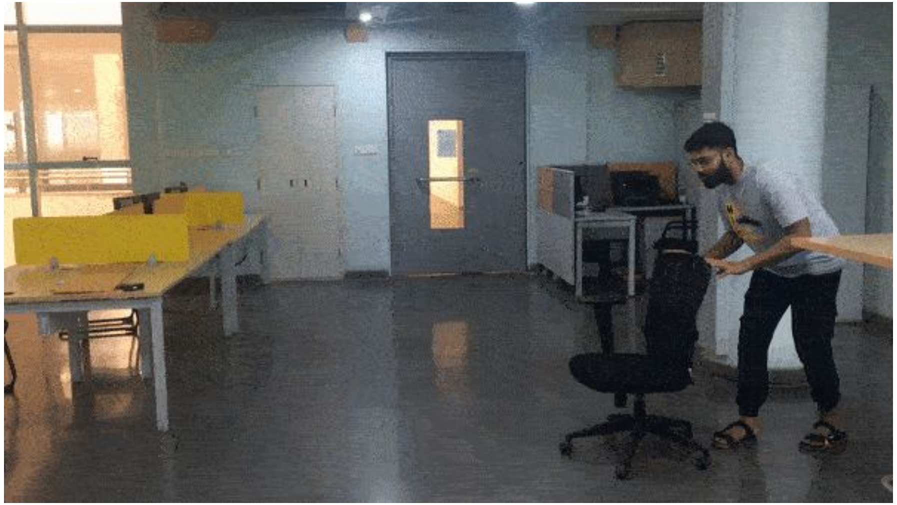


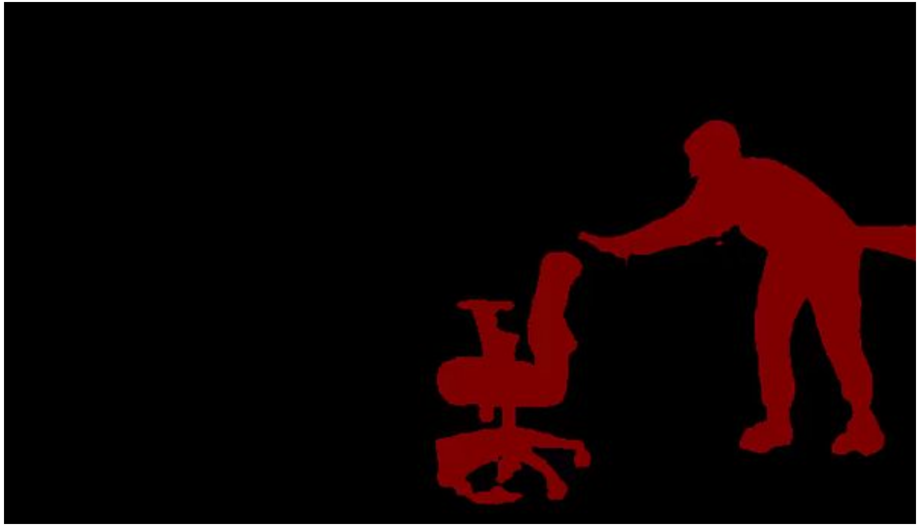
TransVOS

- **Feature Extractor:** This part consists of two-path input layer each of which encodes two types of inputs i.e, RGB frame(query image) and RGB frames with corresponding masks(reference set).This layer is designed to extract features of the target object that can be mapped to an embedding space.
- **Transformer:** This part of the architecture consists of Transformer Encoder and Decoder. **Encoder** is used to model spatial and temporal relationships among pixel features of the sequence.**Decoder** is used to predict the spatial position of the target object in the query frame using encoded features from the encoder and the target query.
- **Segmentation:** This part consists of **Target Attention Block** and **Segmentation Head.** Target Attention Block is needed to extract target's mask features of the query frame from the transformer output. The attention maps obtained from Target Attention Block are used to get the final predicted segmentation mask using the segmentation head.

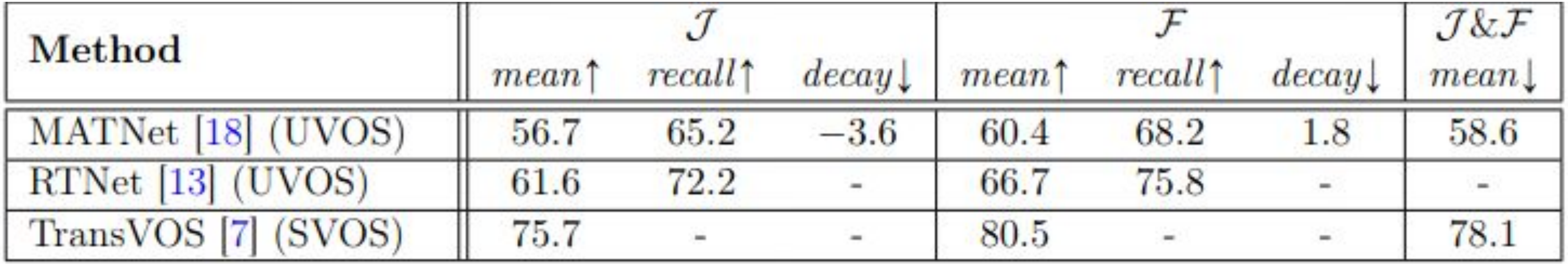


TransVOS (Results)





Compiled Results



Method	$\mathcal{J}$			$\mathcal{F}$			$\mathcal{J}\&\mathcal{F}$
	mean↑	*recall*↑	*decay*↓	*mean*↑	*recall*↑	*decay*↓	*mean*↓
MATNet [18] (UVOS)	56.7	65.2	−3.6	60.4	68.2	1.8	58.6
RTNet [13] (UVOS)	61.6	72.2	-	66.7	75.8	-	-
TransVOS [7] (SVOS)	75.7	-	-	80.5	-	-	78.1

- Average scores over all the videos.

Conclusion

- We implemented **four different approaches**:
  - One approach for covering the case of constrained environments, (Background construction and frame differencing)
  - Two unsupervised approaches which use optical flow information to detect and segment the object in motion. (MatNet and RtNet)
  - One semi-supervised approach with transformer based model. (TransVos)
- Among all the approaches, the semi-supervised TransVOS model resulted with the most crisp results.
- RtNet and MatNet had comparable results among which, the idea behind the RtNet is claimed to be the current state of the art (SoTA model)
- The constrained model has is most computationally efficient provided the constraints are satisfied.

Work Distribution

	Aditya	**Dhruv**	**Rohith**
Code	Frame Differencing, MATNet	Background Segmentation (baseline), RTNet	TransVOS
Report	Introduction, Dataset, MATNet	RTNet, Experimentation and Results	TransVOS, Challenges, Conclusion

Thank You