Temporal-aware Hierarchical Mask Classification for Video Semantic Segmentation

Zhaochong An, Guolei Sun*, Zongwei Wu, Hao Tang, Luc Van Gool
* Corresponding author

BMVC 2023

Introduction:

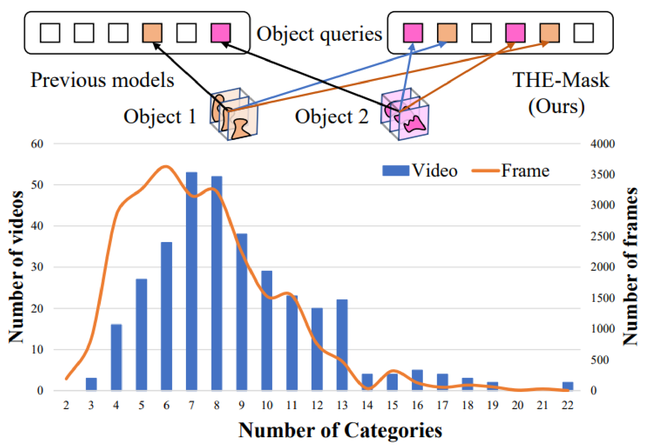
- Two paradigms of segmentation
  - per-pixel classification (semantic segmentation)
  - mask classification (instance-level segmentation)
- Mask classification paradigms
  - object query as object-centric representation
  - predict a pair of a binary mask and a class prediction
  - final segmentations are aggregated from all queries

Motivations:

- The number of object queries is set to be a large number (e.g., 100 in Mask2Former)
- Hungarian matching to assign the best-fit object query to each ground truth object

⇩

Only 8% utilization of queries during training

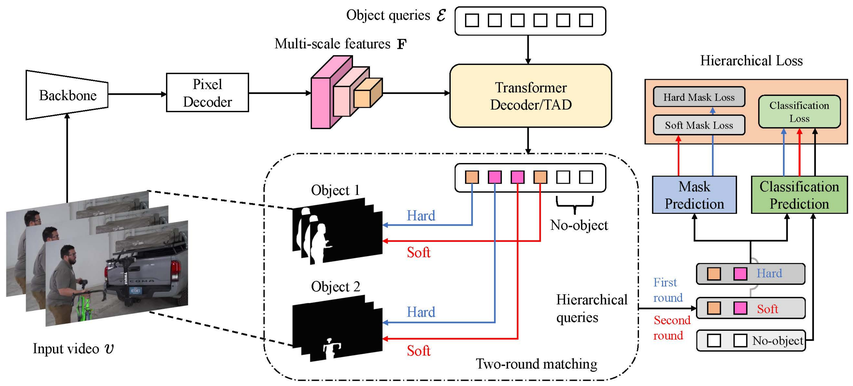


Goals:

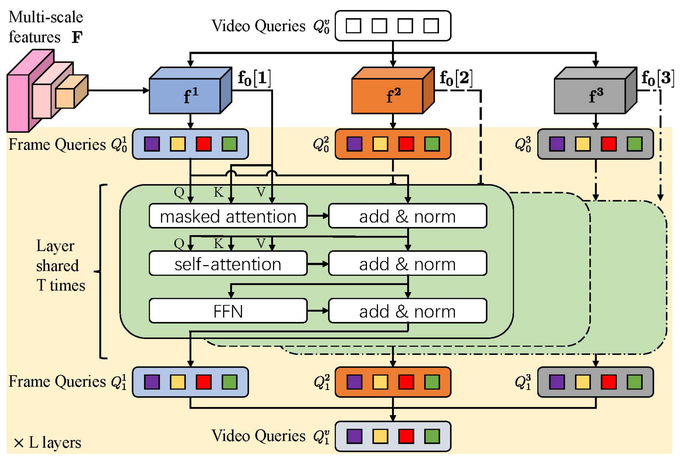
- A simple method to involve more queries during training without losing their own object representation abilities
- To effectively model the temporal interactions in the mask-classification based paradigm

Methods (THE-Mask):

- Hierarchical Mask Classification
  - Two-round queries
  - Hierarchical loss functions (Hard mask loss and soft mask loss)



- Temporal Aggregation Decoder (TAD)



Conclusions:

- We present THE-Mask, a simple and strong mask-classification-based model for VSS
- THE-Mask is the first model to renovate the traditional one-to-one matching and introduce parameter-free hierarchical structures into queries to fully exploit the representation ability
- We propose TAD to explicitly model the temporal interactions for cross-frame learning

Experiments:

Method	Backbone	Params (M) ↓	mIoU ↑	Weighted IoU ↑
DeepLabv3+ [9]	ResNet-101	62.7	34.7	58.8
UperNet [70]	ResNet-101	83.2	36.5	58.6
PSPNet [80]	ResNet-101	70.5	36.5	58.1
OCRNet [76]	ResNet-101	58.1	36.7	59.2
ETC [46]	PSPNet	89.4	36.6	58.3
ETC [46]	OCRNet	58.1	37.5	59.1
NetWarp [70]	PSPNet	89.4	37.0	57.9
NetWarp [70]	OCRNet	58.1	37.5	58.9
$\text{TCB}_{\text{st-ppm}}$ [52]	ResNet-101	70.5	37.5	58.6
$\text{TCB}_{\text{st-ocr}}$ [52]	ResNet-101	58.1	37.4	59.3
$\text{TCB}_{\text{st-ocr-mem}}$ [52]	ResNet-101	58.1	37.8	59.5
Video K-Net (Deeplabv3+) [41]	ResNet-101	–	37.9	–
Video K-Net (PSPNet) [41]	ResNet-101	–	38.0	–
SegFormer [71]	MiT-B1	13.8	36.5	58.8
SegFormer [71]	MiT-B2	24.8	43.9	63.7
CFFM ($t = 4$) [62]	MiT-B1	15.5	38.5	60.0
CFFM ($t = 4$) [62]	MiT-B2	26.5	44.9	64.9
MRCFA ($t = 4$) [63]	MiT-B1	16.2	38.9	60.0
MRCFA ($t = 4$) [63]	MiT-B2	27.3	45.3	64.7
Mask2Fomer ($t = 1$) [13]	MiT-B0	23.0	38.9	60.9
	MiT-B1	33.0	43.3	63.6
	MiT-B2	44.0	47.6	65.4
THE-Mask ($t = 1$)	MiT-B0	23.0	**39.8**	**61.3**
	MiT-B1	33.0	**44.1**	**64.2**
	MiT-B2	44.0	**48.5**	**66.2**
THE-Mask ($t = 2$)	MiT-B5	104.5	**52.1**	**67.2**

Ablation studies:

- Effects of training clip length

backbone	$t = 1$	$t = 2$	$t = 4$
MiT-B0	39.76	39.94	**40.68**
MiT-B1	44.06	44.68	**45.19**
MiT-B2	48.53	48.99	**49.11**

- Temporal aggregation ablation

temporal setting	$t = 2$	$t = 4$
one-to-video	43.40	43.90
one-to-frame	44.41	44.84
video-frame	**44.68**	**45.19**

- Ablation study on hierarchical loss

Matching	Loss	mIoU ↑	wIoU ↑
one round	original loss	43.26	63.56
two round	original loss	43.70	63.77
two round	hierarchical loss	**44.06**	**64.16**

Code:

- https://github.com/ZhaochongAn/THE-Mask