Mining Relations among Cross-Frame Affinities for Video Semantic Segmentation

Guolei Sun, Yun Liu*, Hao Tang, Ajad Chhatkuli, Le Zhang, Luc Van Gool

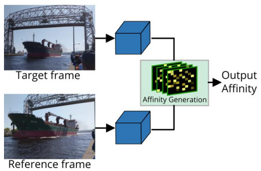
* corresponding author

Introduction

- Video semantic segmentation (VSS)
  - VSS aims at assigning a semantic class to all the pixels of all the frames in the video
  - Compared to image semantic segmentation, VSS is much less explored in literatures
  - There is no tremendous progress due to the lack of large-scale and fully-annotated dataset
- Importance of VSS
  - Real-life scenes are dynamic
  - Temporal information is naturally used by animals in the interaction with environments
- Recent development
  - Large-scale dataset VSPW [1] is proposed

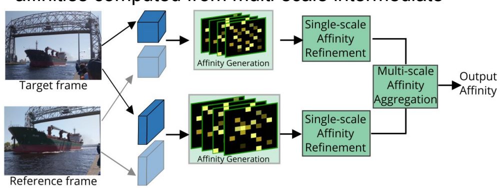
Motivation

- The core of VSS lies in how to exploit the temporal information for prediction
- Previous methods



  - Develop new techniques (e.g. optical flow) to compute the cross-frame affinities
  - Directly use affinities to refine the features

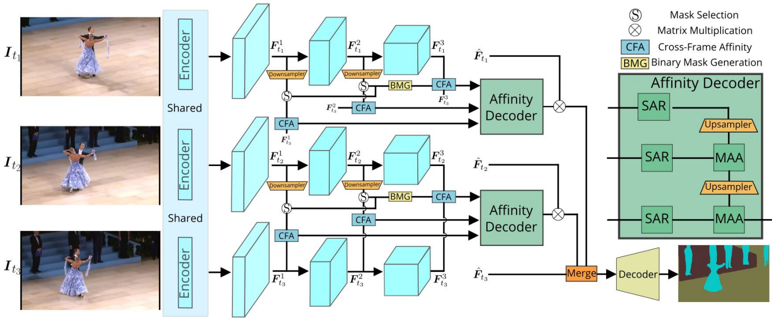
- **Our motivation**: mine the *relations* among multi-scale affinities computed from multi-scale intermediate



Problem Setting

- We are given video frames $\{\boldsymbol{I}_{t_i} \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^{T}$ with ground-truth masks $\{\boldsymbol{M}_{t_i} \in \mathbb{R}^{H\times W}\}_{i=1}^{T}$. The target frame is $\boldsymbol{I}_{t_T}$, while the remainings are reference frames
- Goal: train a segmentation network which can use information of all frames (reference and target) to segment the target one

Methodology



- *Mining Relations among Cross-Frame Affinities for video semantic segmentation: MRCFA*
  - Compute Cross-Frame Affinities (CFA)
  - Selective Token Masking
  - Single-scale Affinity Refinement (SAR)
  - Multi-scale Affinity Aggregation (MAA)
- *Compute Cross-Frame Affinities (CFA)*

$$\boldsymbol{Q}^l = f(\boldsymbol{F}^l_{t_T}; \boldsymbol{W}^l_{query}) \quad \boldsymbol{K}^l_{t_i} = f(\boldsymbol{F}^l_{t_i}; \boldsymbol{W}^l_{key}) \quad \boldsymbol{A}^l_{t_i} = \boldsymbol{Q}^l \times \boldsymbol{K}^{l\top}_{t_i}$$

  - Reference tokens for computing cross-frame affinities across scales should match
  - Computing CFA for large-scale features requires large computation
- *Selective Token Masking (STM)*
  - Downsample the multi-scale keys to the same spatial size
  - Further reduce the number of tokens in keys by selecting important tokens and discarding non-important ones
  - Use the affinity from the deepest features to determine the importance of tokens
- *Single-scale Affinity Refinement (SAR)*
  - Using 3D conv to learn the correlation within the single-scale affinity suffers from two weaknesses: large computational cost and non-meaningful 3D window
  - We propose to refine the affinities by common 2D convolutions
- *Multi-scale Affinity Aggregation (MAA)*
  - The affinity from the deep layers contains more semantic but more coarse info.
  - The affinity from the shallow layers has more fine-grained but less semantic info.
  - We propose a MAA module to aggregate the info. from small- to large-scale affinities
  - After obtaining refined affinities, we can reconstruct target features using reference features $\boldsymbol{O}_{t_i} = \boldsymbol{B}^1_{t_i} \times \hat{\boldsymbol{F}}_{t_i}$
- The final features

$$\boldsymbol{O}_{t_L} = \frac{1}{T-1}\Gamma\Big(\sum_{i=1}^{T-1} \boldsymbol{O}_{t_i}\Big) + \hat{\boldsymbol{F}}_{t_L}$$

Experiments

- Datasets: VSPW [1] dataset (largest-scale benchmark for VSS) and Cityscapes dataset
- Evaluation: mIoU & Video consistency (VC) [1,2]
  - VC: evaluate the temporal consistency of predictions
- Quantitative results

Methods	Backbone	mIoU ↑	Weighted IoU ↑	mVC_8 ↑	mVC_{16} ↑	Params (M) ↓	FPS (f/s) ↑
SegFormer [47]	MiT-B0	32.9	56.8	82.7	77.3	3.8	73.4
SegFormer [47]	MiT-B1	36.5	58.8	84.7	79.9	13.8	58.7
MRCFA (Ours)	MiT-B0	35.2	57.9	88.0	83.2	5.2	50.0
MRCFA (Ours)	MiT-B1	**38.9**	**60.0**	**88.8**	**84.4**	16.2	40.1
DeepLabv3+ [6]	ResNet-101	34.7	58.8	83.2	78.2	62.7	-
UperNet [46]	ResNet-101	36.5	58.6	82.6	76.1	83.2	-
PSPNet [52]	ResNet-101	36.5	58.1	84.2	79.6	70.5	13.9
OCRNet [50]	ResNet-101	36.7	59.2	84.0	79.0	58.1	14.3
ETC [33]	PSPNet	36.6	58.3	84.1	79.2	89.4	-
NetWarp [46]	PSPNet	37.0	57.9	84.4	79.4	89.4	-
ETC [33]	OCRNet	37.5	59.1	84.1	79.1	58.1	-
NetWarp [46]	OCRNet	37.5	58.9	84.0	79.0	58.1	-
$TCB_{st\text{-}ppm}$ [34]	ResNet-101	37.5	58.6	87.0	82.1	70.5	10.0
$TCB_{st\text{-}ocr}$ [34]	ResNet-101	37.4	59.3	86.9	82.0	58.1	5.5
$TCB_{st\text{-}ocr\text{-}mem}$ [34]	ResNet-101	37.8	59.5	87.9	84.0	58.1	5.5
SegFormer [47]	MiT-B2	43.9	63.7	86.0	81.2	24.8	39.2
SegFormer [47]	MiT-B5	48.2	65.1	87.8	83.7	82.1	17.2
MRCFA (Ours)	MiT-B2	45.3	64.7	90.3	86.2	27.3	32.1
MRCFA (Ours)	MiT-B5	**49.9**	**66.0**	**90.9**	**87.4**	84.5	15.7

- MRCFA achieves SOTA results and produces the temporally consistent masks
- MRCFA adds limited model complexity and latency
- MRCFA is effective in mining the relations among affinities between target and reference

Methods	Backbone	mIoU ↑	Params (M) ↓
FCN [41]	MobileNetV2	61.5	9.8
CC [42]	VGG-16	67.7	-
DFF [56]	ResNet-101	68.7	-
GRFP [36]	ResNet-101	69.4	-
PSPNet [52]	MobileNetV2	70.2	13.7
DVSN [48]	ResNet-101	70.3	-
Accel [22]	ResNet-101	72.1	-
ETC [33]	ResNet-18	71.1	13.2
SegFormer [47]	MiT-B0	71.9	3.7
MRCFA (Ours)	MiT-B0	72.8	4.2
SegFormer [47]	MiT-B1	74.1	13.8
MRCFA (Ours)	MiT-B1	**75.1**	14.9

Summary

- We propose a novel framework MRCFA for VSS by mining the relations among multi-scale cross-frame affinities in two aspects: single-scale intrinsic correlations and multi-scale relations
- STM is adopted to sample important tokens in keys of the reference frames to reduce computation and facilitate MAA
- Extensive experiments demonstrate the efficiency and effectiveness of MRCFA

Reference

[1] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. VSPW: A large-scale dataset for video scene parsing in the wild. In CVPR 2021

[2] Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, Luc Van Gool. Coarse-to-Fine Feature Mining for Video Semantic Segmentation. In CVPR 2022

Code: https://github.com/GuoleiSun/VSS-MRCFA