Prototypical Cross-Attention Networks for Multiple Object Tracking and Segmentation

More Info

(vis.xyz/pub/pcan)

Lei Ke[1,2], Xia Li[1], Martin Danelljan[1], Yu-Wing Tai[3], Chi-Keung Tang[2], Fisher Yu[1]

1. ETH Zürich 2. HKUST 3. Kuaishou Technology

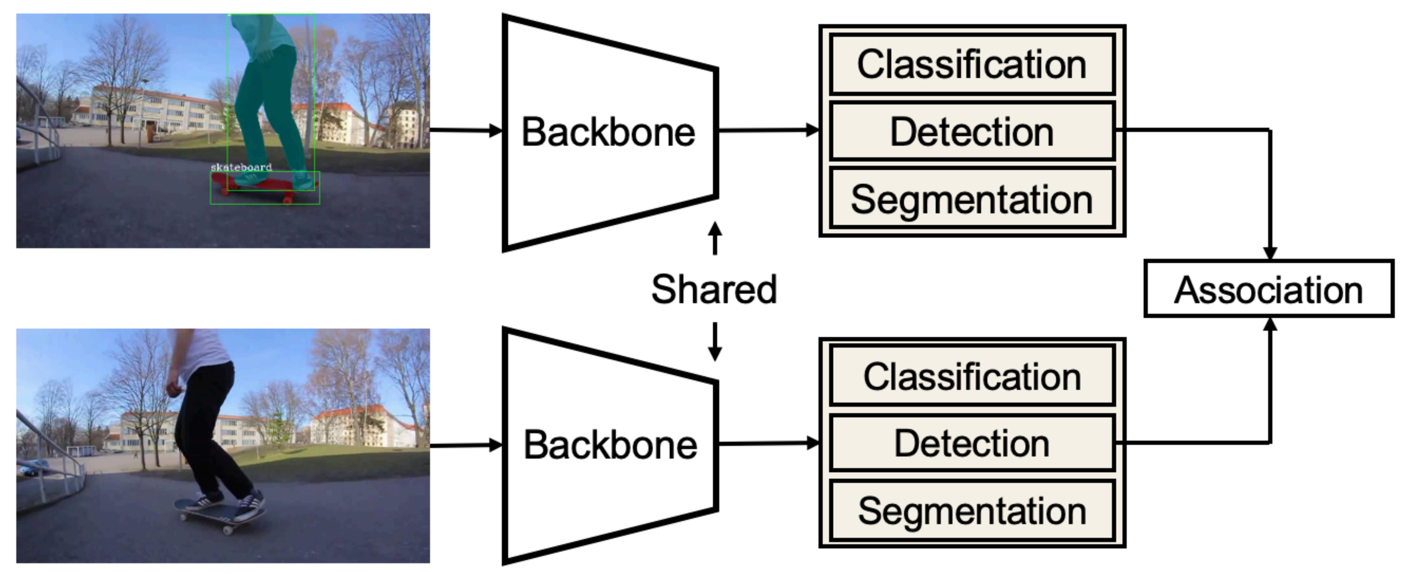
Introduction

➢ Problem Setup

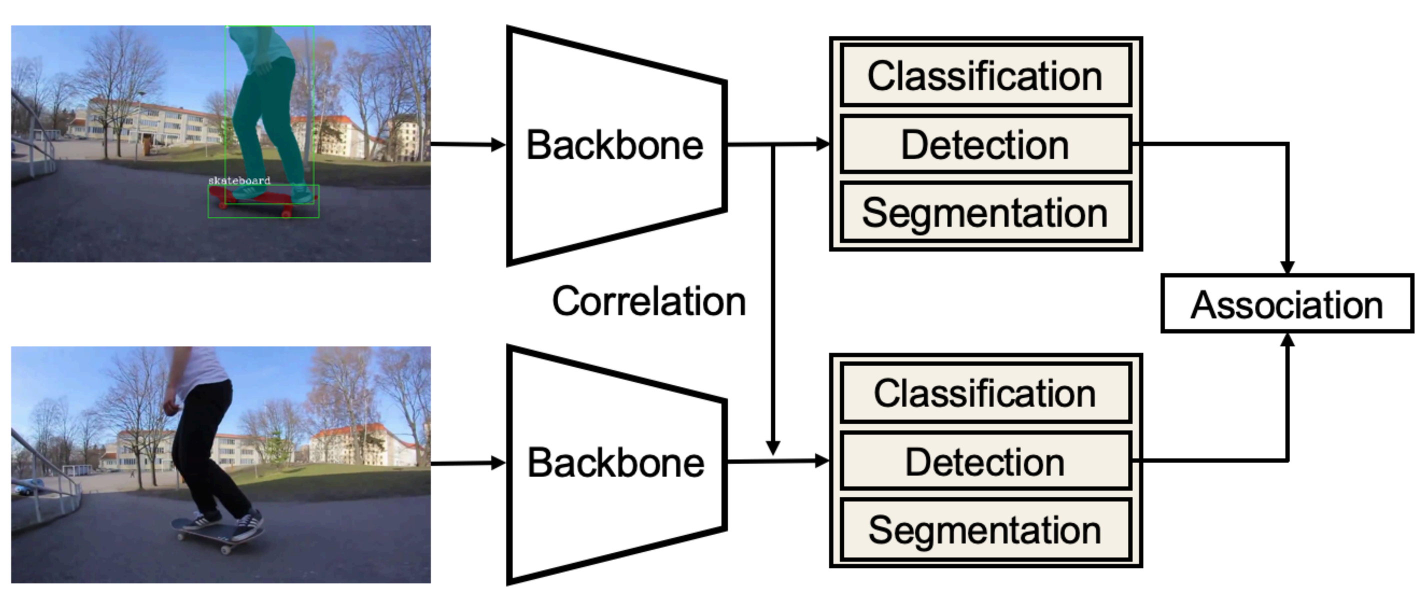
- Multiple object tracking and segmentation.
- Requires detecting, tracking and segmenting all interested objects in a video.

➢ Motivation

- The temporal dimension carries rich scene information while most previous methods are tracking by detection, which only exploit the temporal information to address the object association problem.



Previous MOTS paradigm **(a)**: Temporal information is limited to the object association phase.



Previous MOTS paradigm (**b**): temporal modeling is only between two adjacent frames.

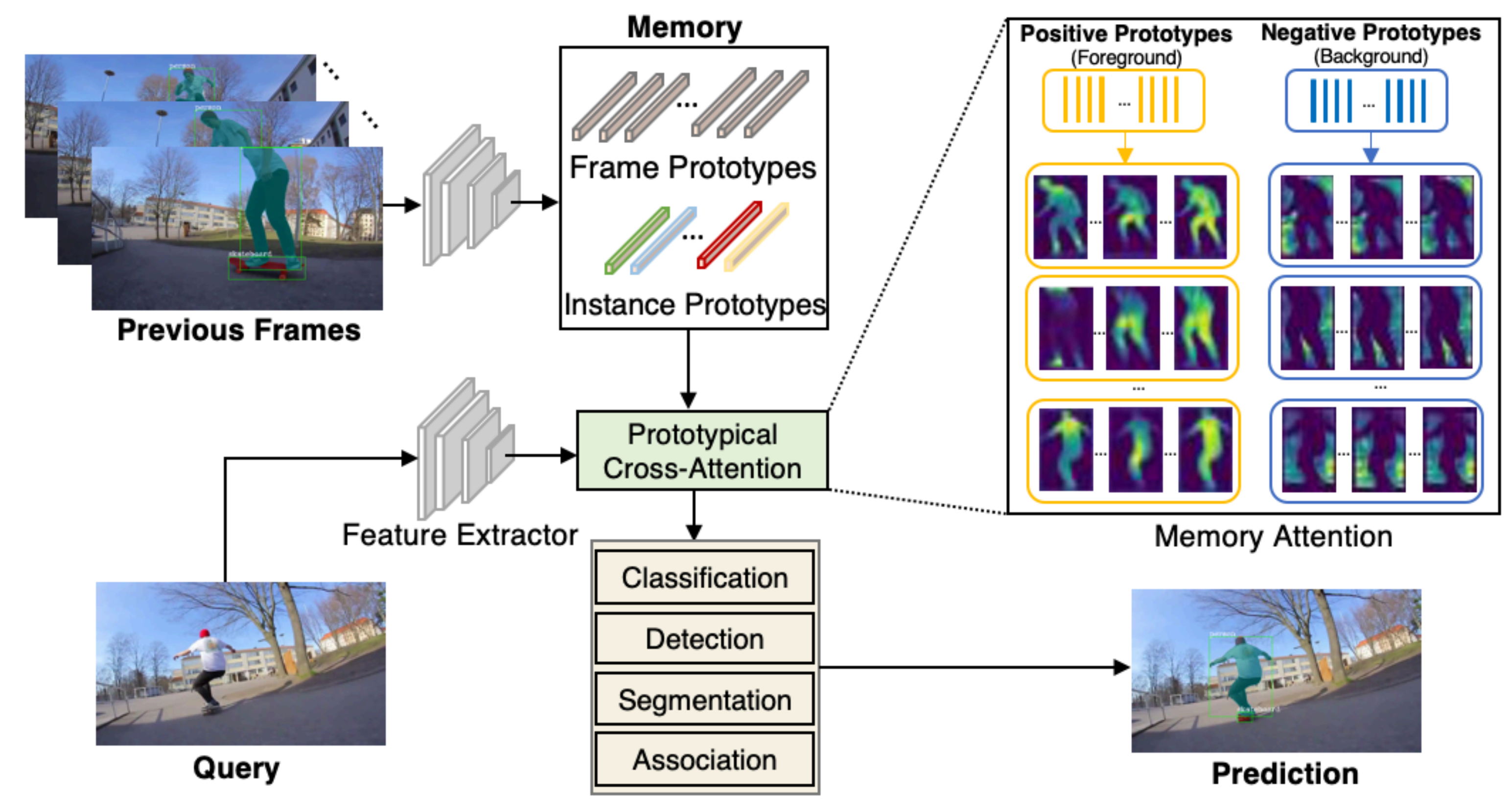
- Design an effective and efficient modeling for temporal information, which can be also used to improve temporal segmentation result.

Contribution

- Prototypical Cross-Attention Module (**PCAM**) for efficiently utilizing long-term spatio-temporal video information.
- An **online** MOTS approach Prototypical Cross-Attention Network (**PCAN**) that employs PCAM on **frame** and **instance-level**.
- The appearance of each video tracklet is encoded with contrastive foreground and background prototypes, which are propagated over time and updated recurrently.
- We extensively analyze our method on BDD100K and Youtube-VIS.

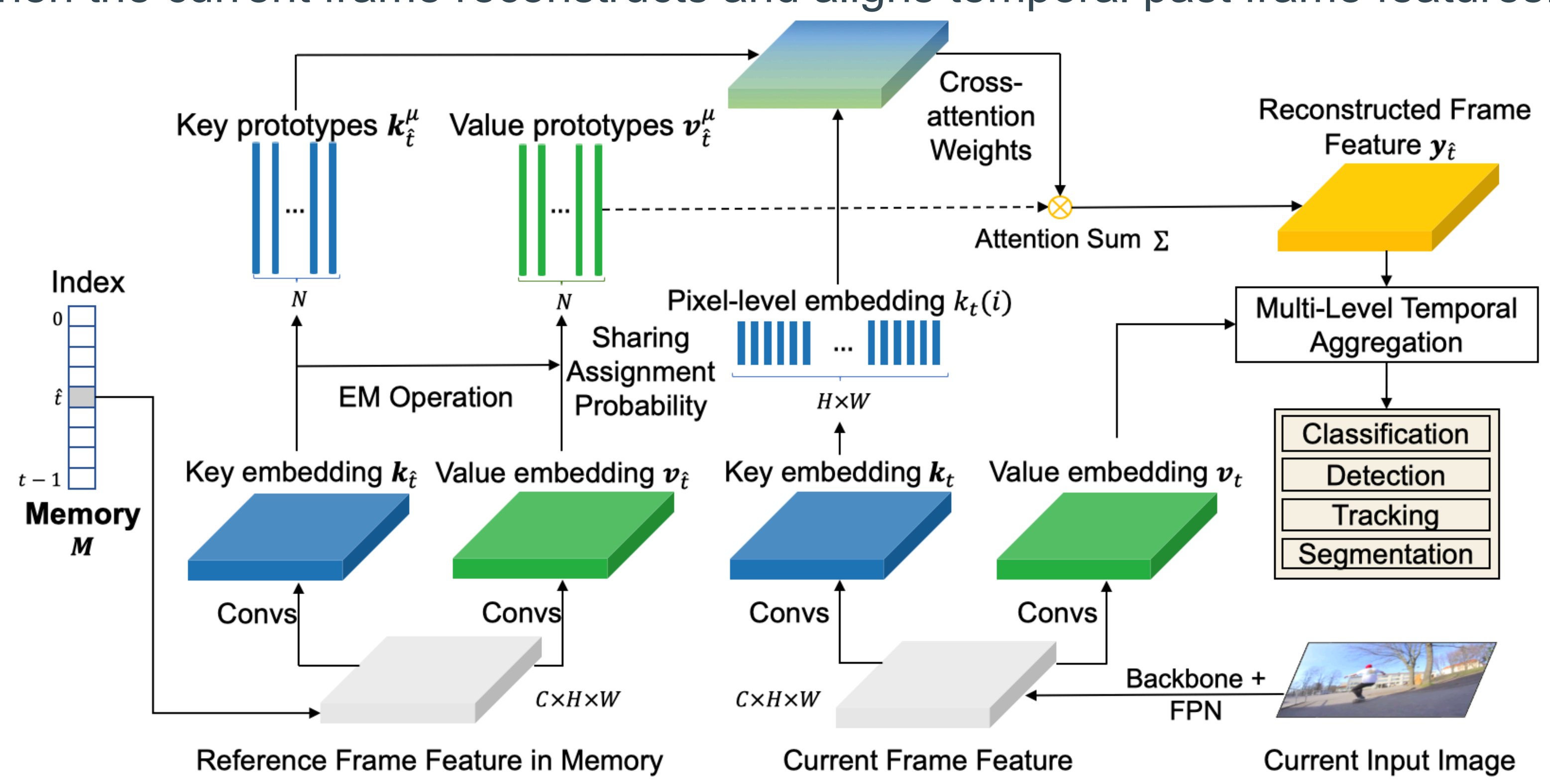
Prototypical Cross-Attention Network (PCAN)

- PCAN first condenses the space-time memory and high-resolution frame embedding into **frame-level and instance-level prototypes**.
- Prototypical cross-attention is employed to retrieve rich temporal information.



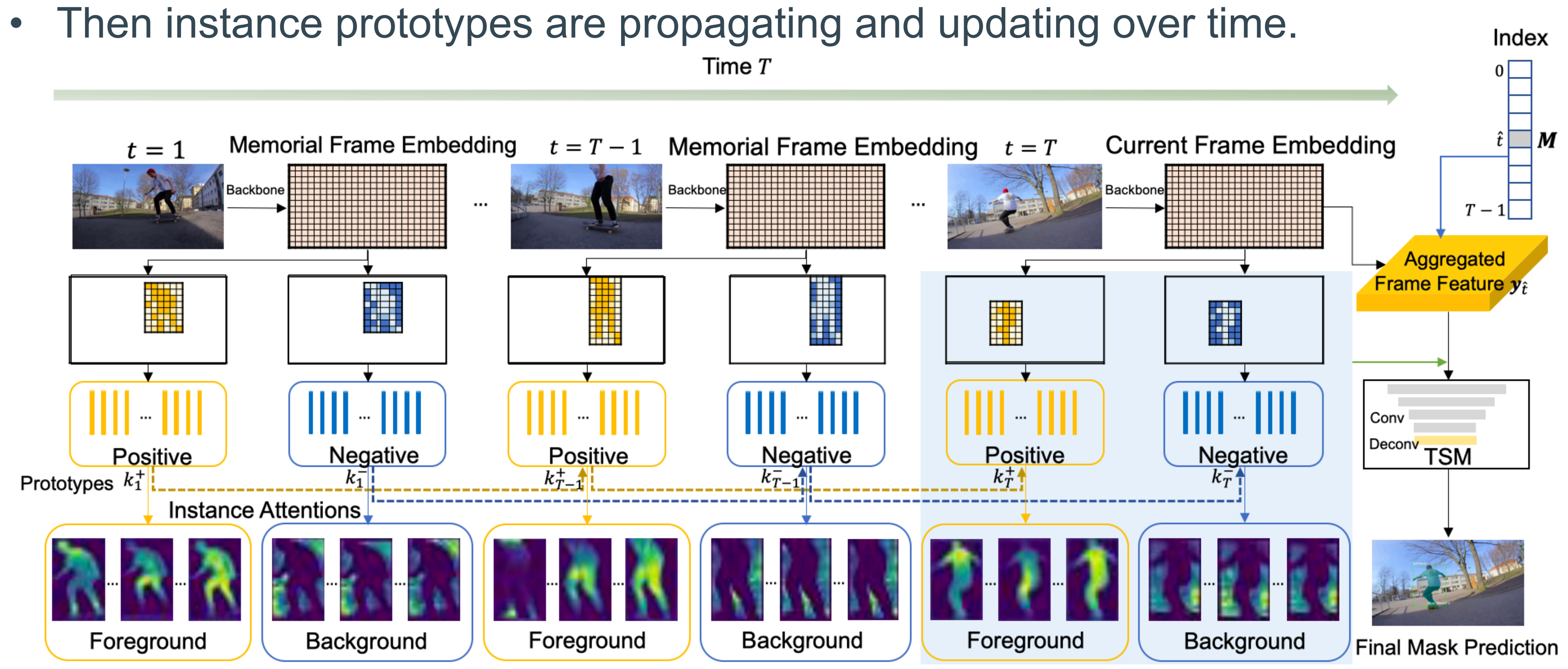
PCAN at **Frame-Level**

- Past frames are first reduced to sets of prototypes by GMM-based clustering.
- Then the current frame reconstructs and aligns temporal past frame features.



PCAN at **Instance-Level**

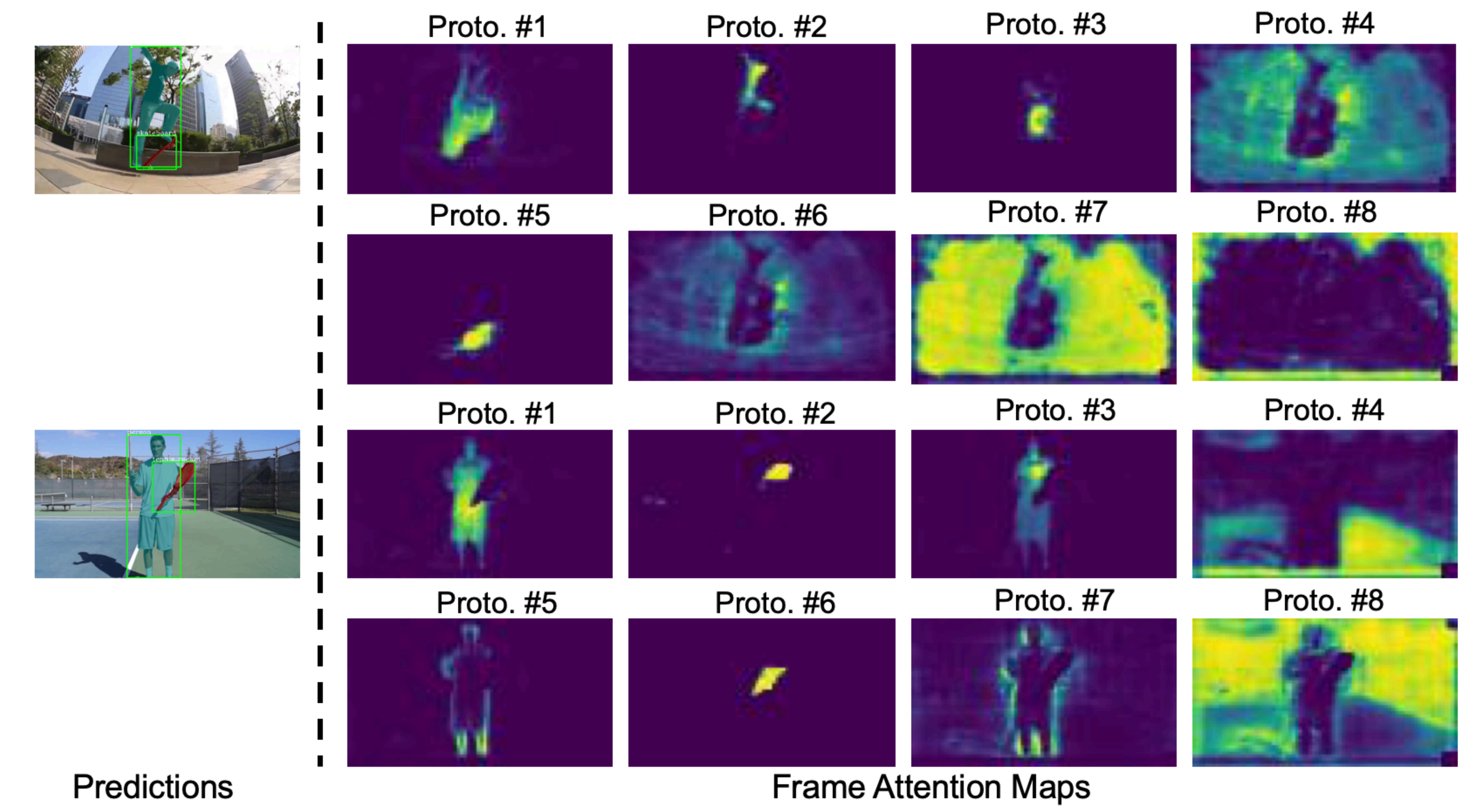
- Video tracklets are encoded by contrastive foreground and background prototypes.
- Then instance prototypes are propagating and updating over time.



Visualization of Sampled Prototypes

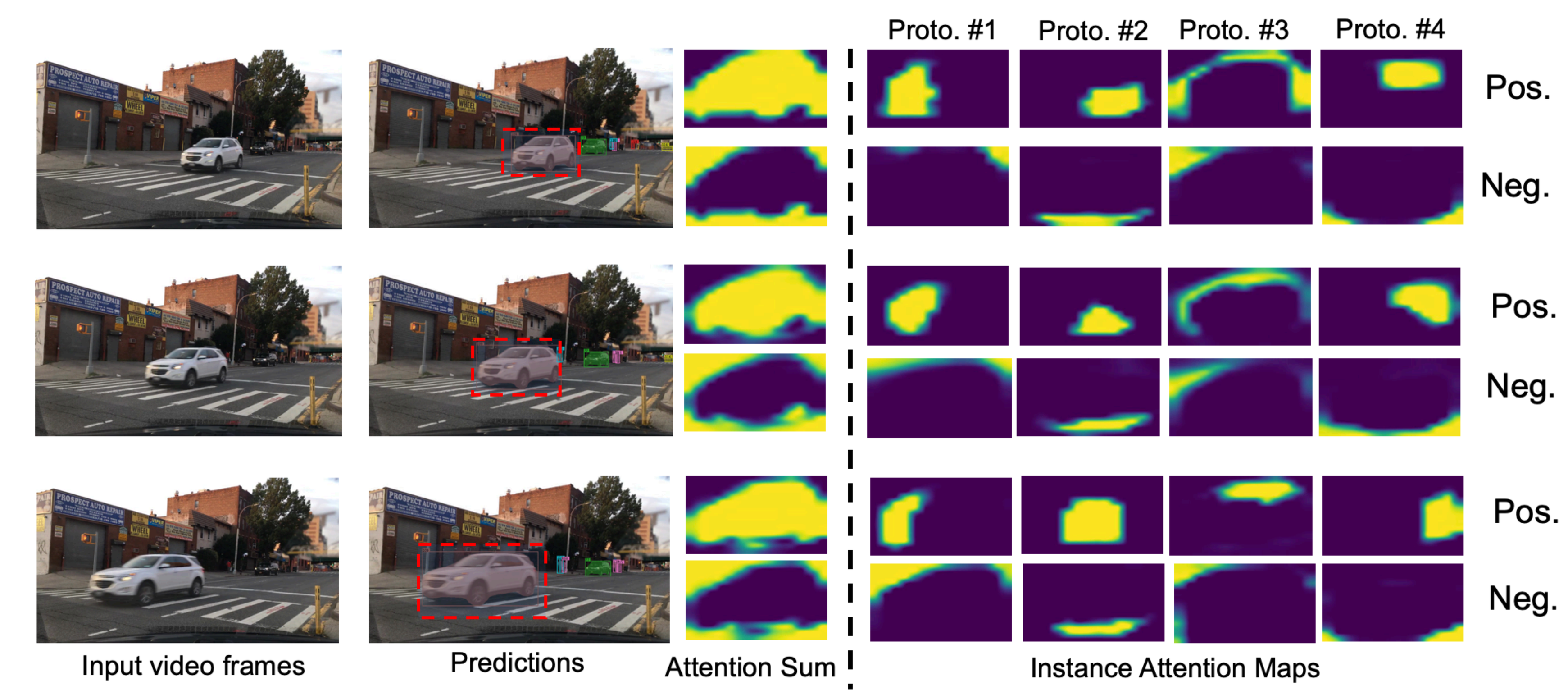
➢ Frame-level Attention

- Frame prototypes learn to correspond to semantic concepts of the whole image.



➢ Instance-level Attention

- Each instance prototype focuses on specific car sub-regions (foreground and background) with implicit unsupervised temporal consistency over time.



Experiment Results

PCAN achieves consistent large performance gain on BDD100K and Youtube-VIS.