Pyramid Constrained Self-Attention Network for Fast Video Salient Object Detection

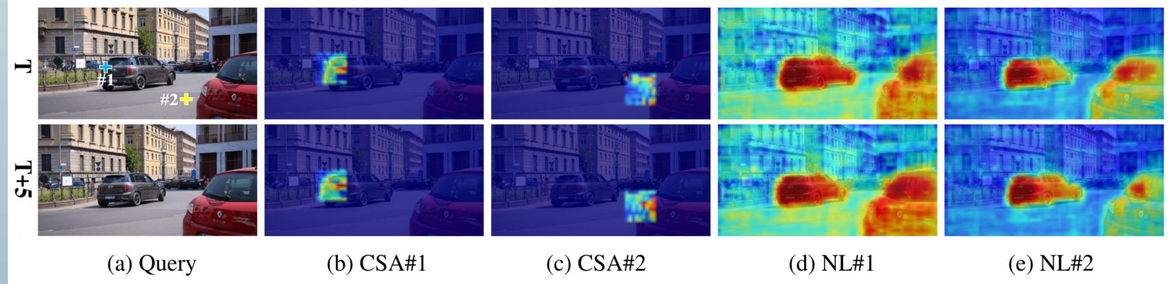
Yuchao Gu[1], Lijuan Wang[1], Ziqin Wang[2], Yun Liu[1], Ming-Ming Cheng[1], Shao-Ping Lu[1]

[1]Nankai University [2]The University of Sydney

Abstract

Spatiotemporal information is essential for video salient object detection (VSOD) due to the highly attractive object motion for human's attention. Previous VSOD methods usually use Long Short-Term Memory (LSTM) or 3D ConvNet (C3D), which can only encode motion information through step-by-step propagation in the temporal domain. Recently, the non-local mechanism is proposed to capture long-range dependencies directly. However, it is not straightforward to apply the non-local mechanism into VSOD, because i) it fails to capture motion cues and tends to learn motion-independent global contexts; ii) its computation and memory costs are prohibitive for video dense prediction tasks such as VSOD. To address the above problems, we design a Constrained Self-Attention (CSA) operation to capture motion cues, based on the prior that objects always move in a continuous trajectory. We group a set of CSA operations in Pyramid structures (PCSA) to capture objects at various scales and speeds. Extensive experimental results demonstrate that our method outperforms previous state-of-the-art methods in both accuracy and speed (110 FPS on a single Titan Xp) on five challenge datasets.

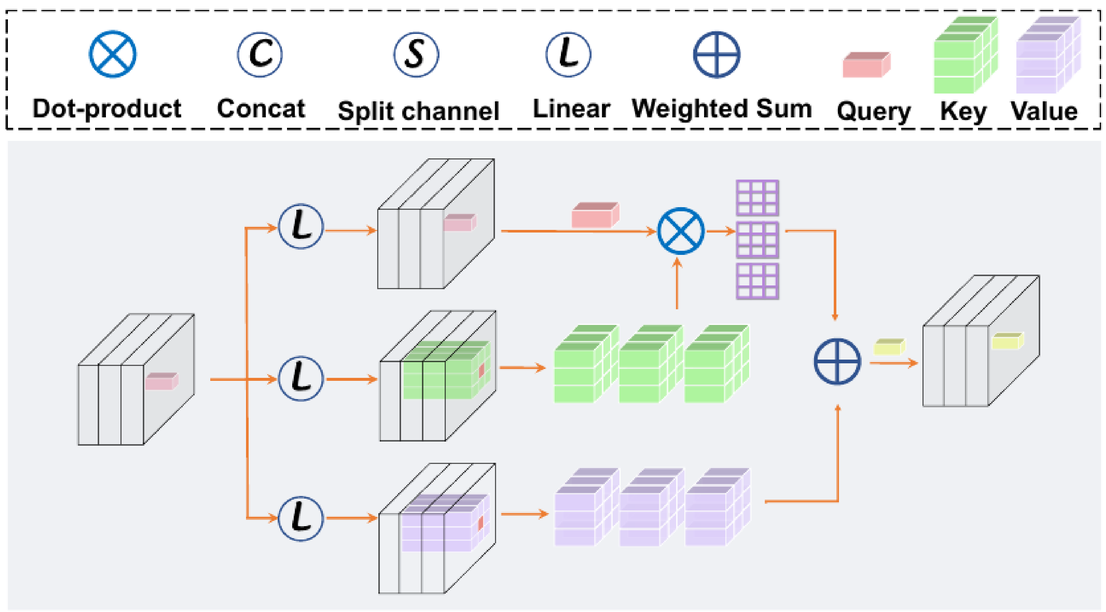
Motivation



(a) Query (b) CSA#1 (c) CSA#2 (d) NL#1 (e) NL#2

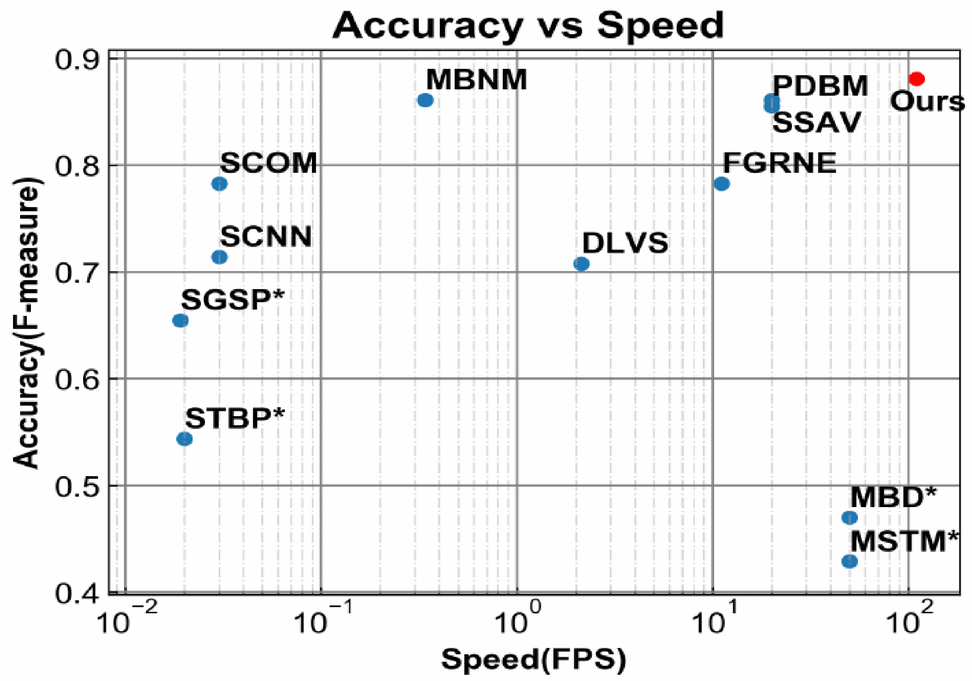
1) Capture long-range temporal information
2) Non-local tends to capture global context in VSOD
3) Design a efficient message passing method for VSOD

Constained Self-Attention



We constrain the self-attention area, thus the CSA can diliver message on all given frames but local area wrt. query position. Because object move in a continuous trajectory. CSA can well capture motion cues with less computation overhead,

Speed Comparsion



Evaluation on VSOD Dataset

Test Dataset		DLVS	FGRN	MBNM	PDBM	SSAV	Ours
DAVIS	max F↑	0.708	0.783	0.861	0.855	0.861	**0.880**
	S↑	0.794	0.838	0.887	0.882	0.893	**0.902**
	MAE↓	0.061	0.043	0.031	0.028	0.028	**0.022**
FBMS	max F↑	0.759	0.767	0.816	0.821	**0.865**	0.831
	S↑	0.794	0.809	0.857	0.851	**0.879**	0.866
	MAE↓	0.091	0.088	0.047	0.064	**0.040**	0.041
ViSal	max F↑	0.852	0.848	0.883	0.888	0.939	**0.940**
	S↑	0.881	0.861	0.898	0.907	0.943	**0.946**
	MAE↓	0.048	0.045	0.020	0.032	0.020	**0.017**
SegV2	max F↑	-	-	0.716	0.800	0.801	**0.810**
	S↑	-	-	0.809	0.864	0.851	**0.865**
	MAE↓	-	-	0.026	0.024	**0.023**	0.025
VOS	max F↑	0.675	0.669	0.670	0.742	0.742	**0.747**
	S↑	0.760	0.715	0.742	0.818	0.819	**0.827**
	MAE↓	0.099	0.097	0.099	0.078	0.073	**0.065**
DAVSOD	max F↑	0.521	0.573	0.520	0.572	0.603	**0.655**
	S↑	0.657	0.693	0.637	0.698	0.724	**0.741**
	MAE↓	0.129	0.098	0.159	0.116	0.092	**0.086**
Runtime (s)↓		0.47	0.09	2.63	0.05	0.05	**0.009**

Qualitive Results