Generalized Relation Modeling for Transformer Tracking

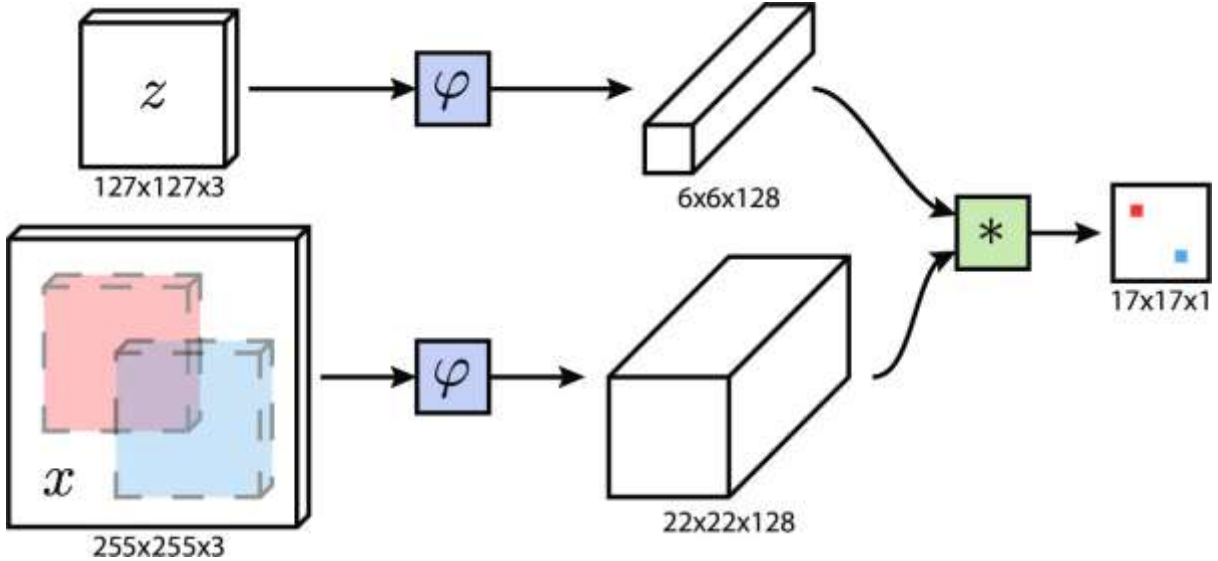
Shenyuan Gao
sygao@connect.ust.hk

Single Object Tracking

- Objective:
  - Given a target with bounding box annotation in the initial frame.
  - Localize the target in the successive frames.
- Requirements:
  - Single object: any other regions will be treated as the background.
  - Model free: no prior about the objects to track.
  - Real time: inference speed should be faster than the frame rate.

Typical Framework

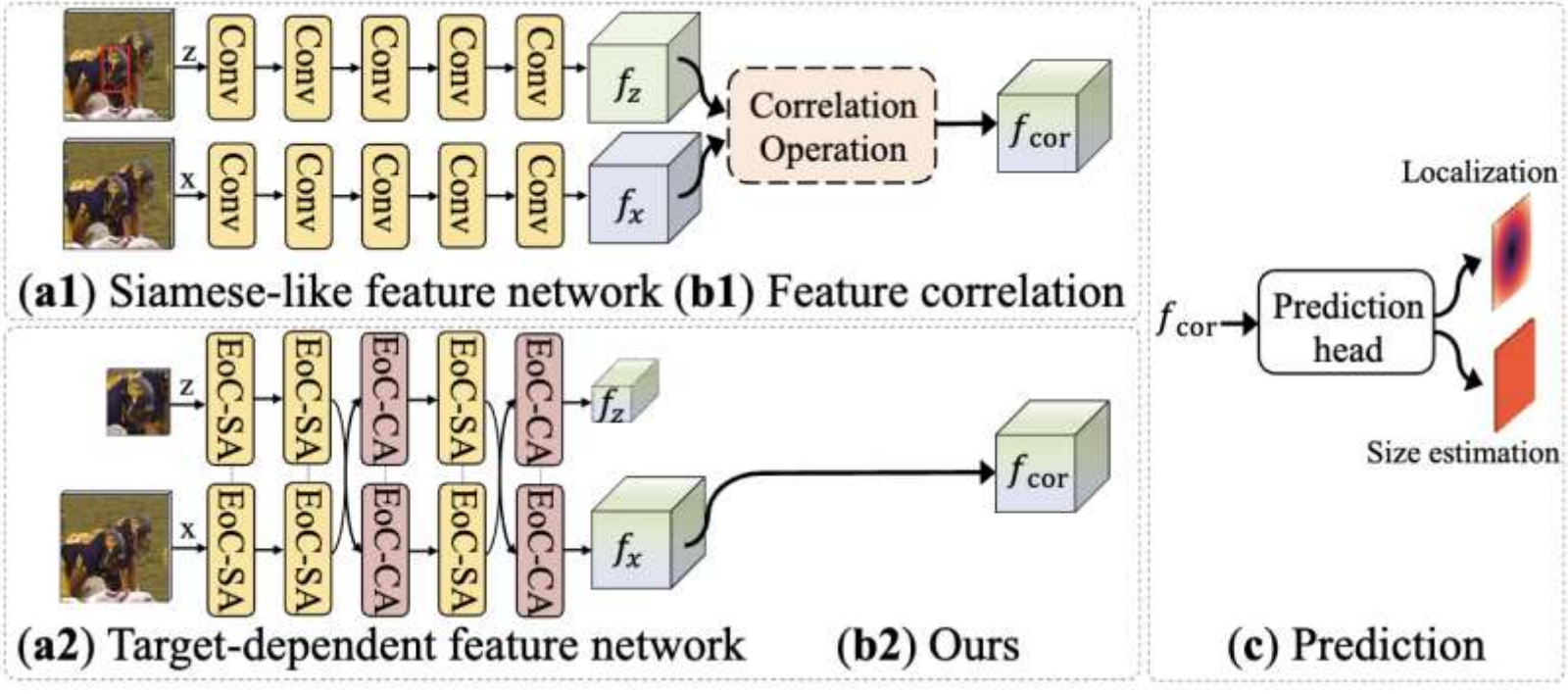
- Siamese network:
  - Learning to compare.
  - Popular in contrastive representation learning [1].
- Siamese tracker:
  - Typical framework for tracking.
  - One-shot learning (prompt learning).



[1] Chen, Xinlei, and Kaiming He. "Exploring simple siamese representation learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

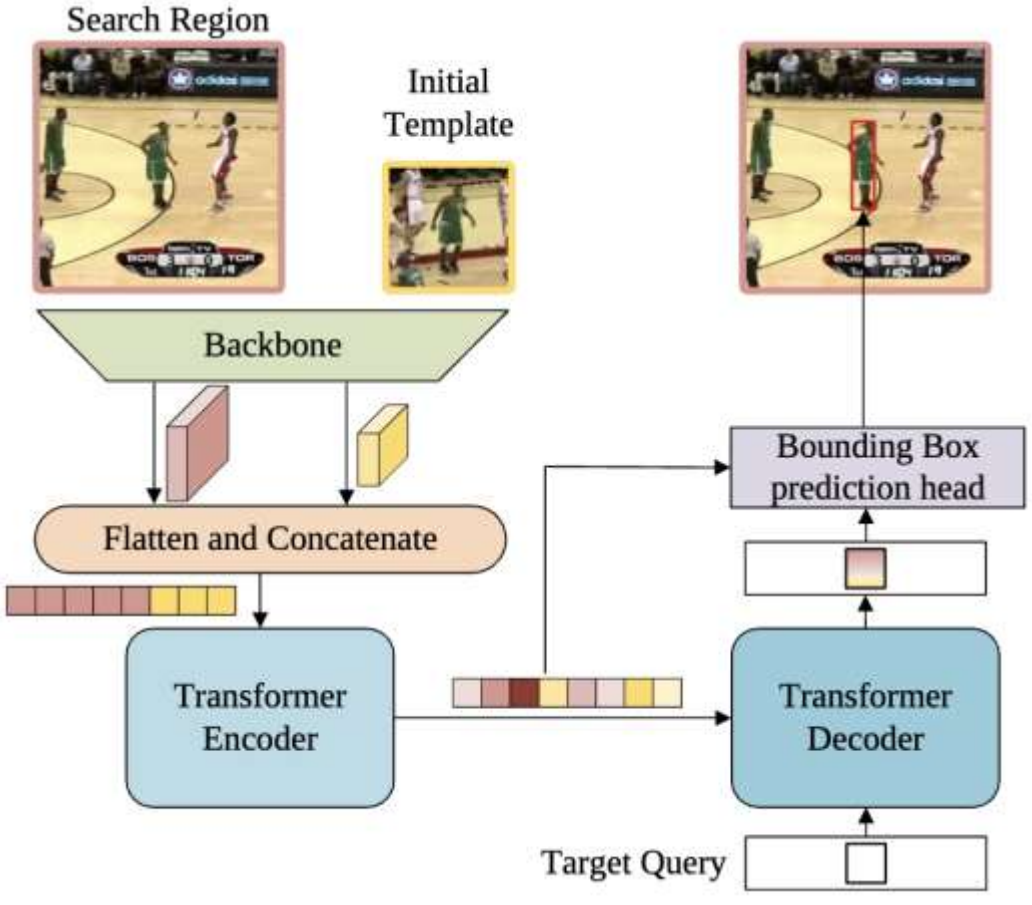
Recent Visual Tracking Development #1

- Feature interaction inside the backbone.

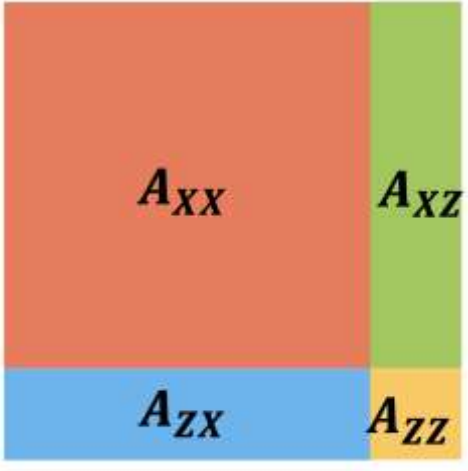


Recent Visual Tracking Development #2

- Concatenation-based feature interaction.

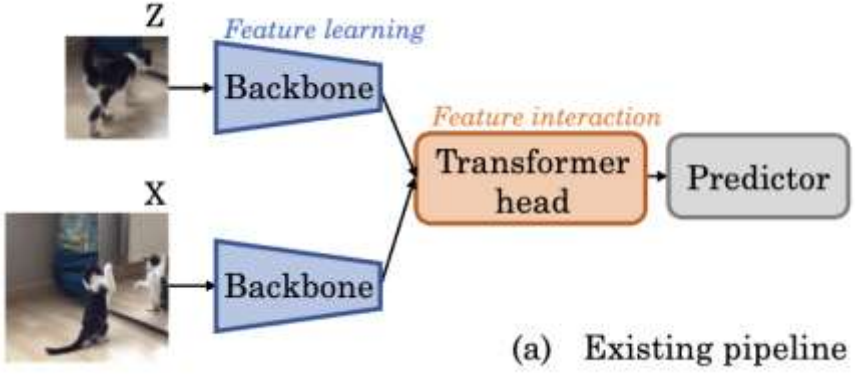


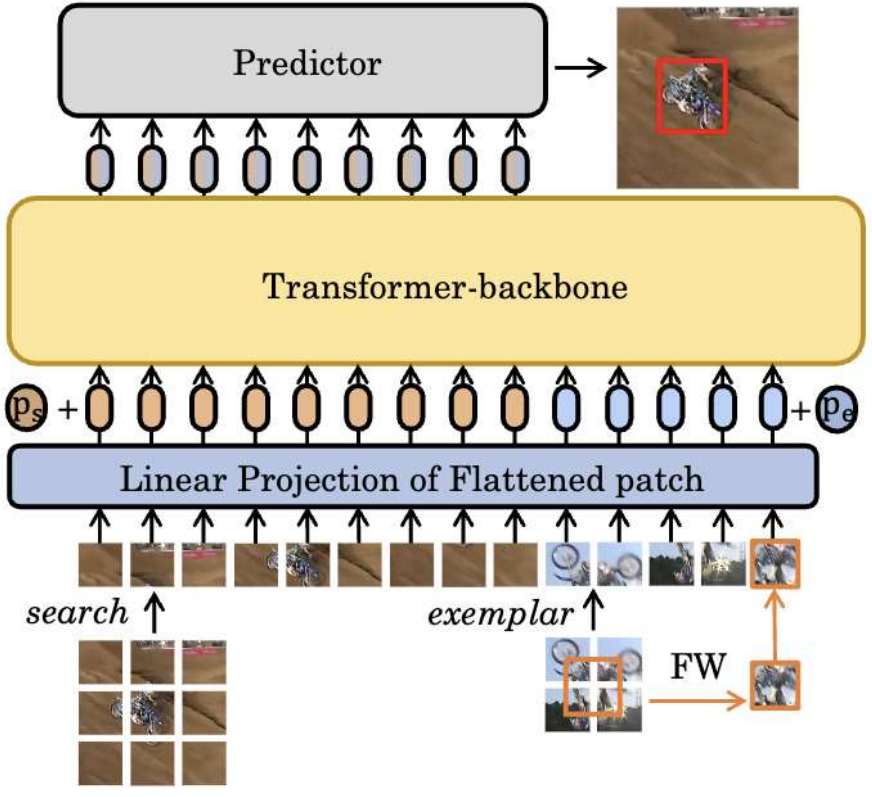
implicitly model
4 types of relations



Recent Visual Tracking Development #3

- Joint feature extraction and interaction.

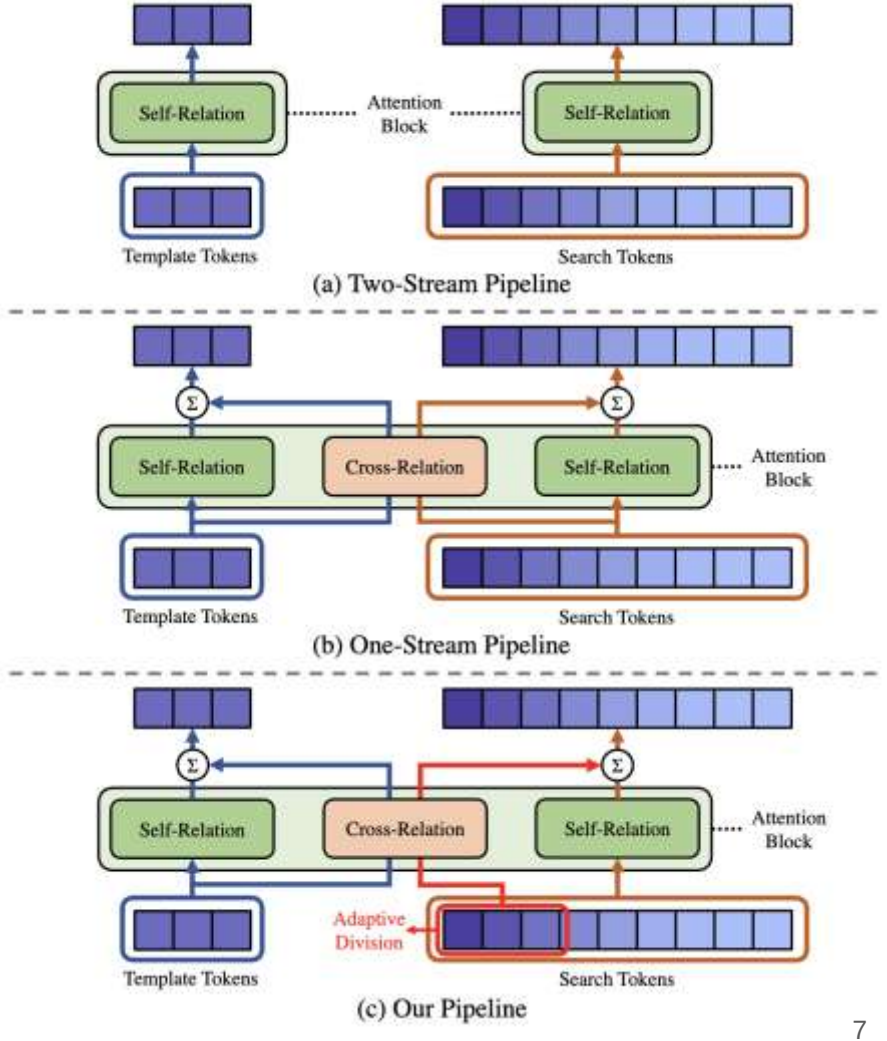




- What could be the next step?

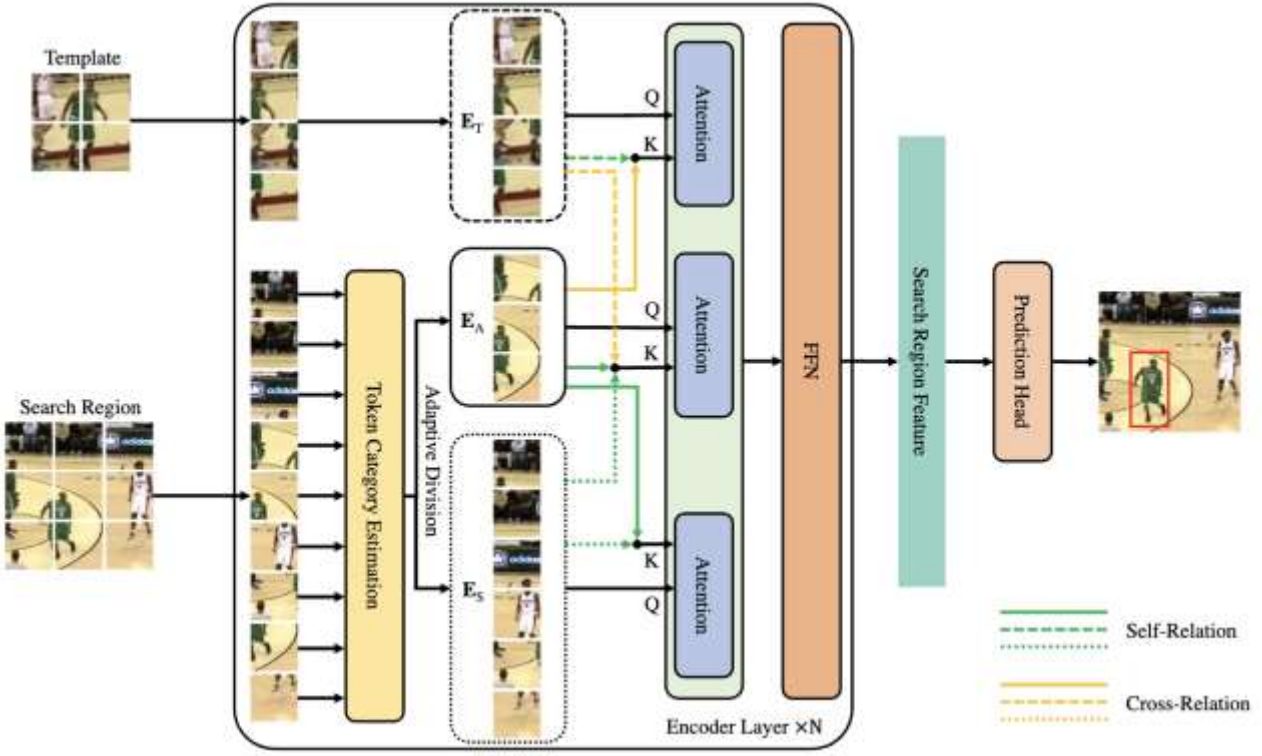
Motivation

- Concerns:
  - Early interaction is proved to be beneficial.
  - All previous works treat the search region as a union.
  - No evidence suggests that one-stream is always better than two-stream.
  - Only a portion of search tokens are suitable for cross-relation modeling with template.
  - Can we adaptively select these search tokens during inference?



A Generalized Formulation of Relation Modeling #4

- Enable more flexible relation modeling by adaptively selecting appropriate search tokens to interact with the template tokens.
- Both the two-stream and one-stream pipelines become the degenerated cases of our method.



Obstacles in Implementation

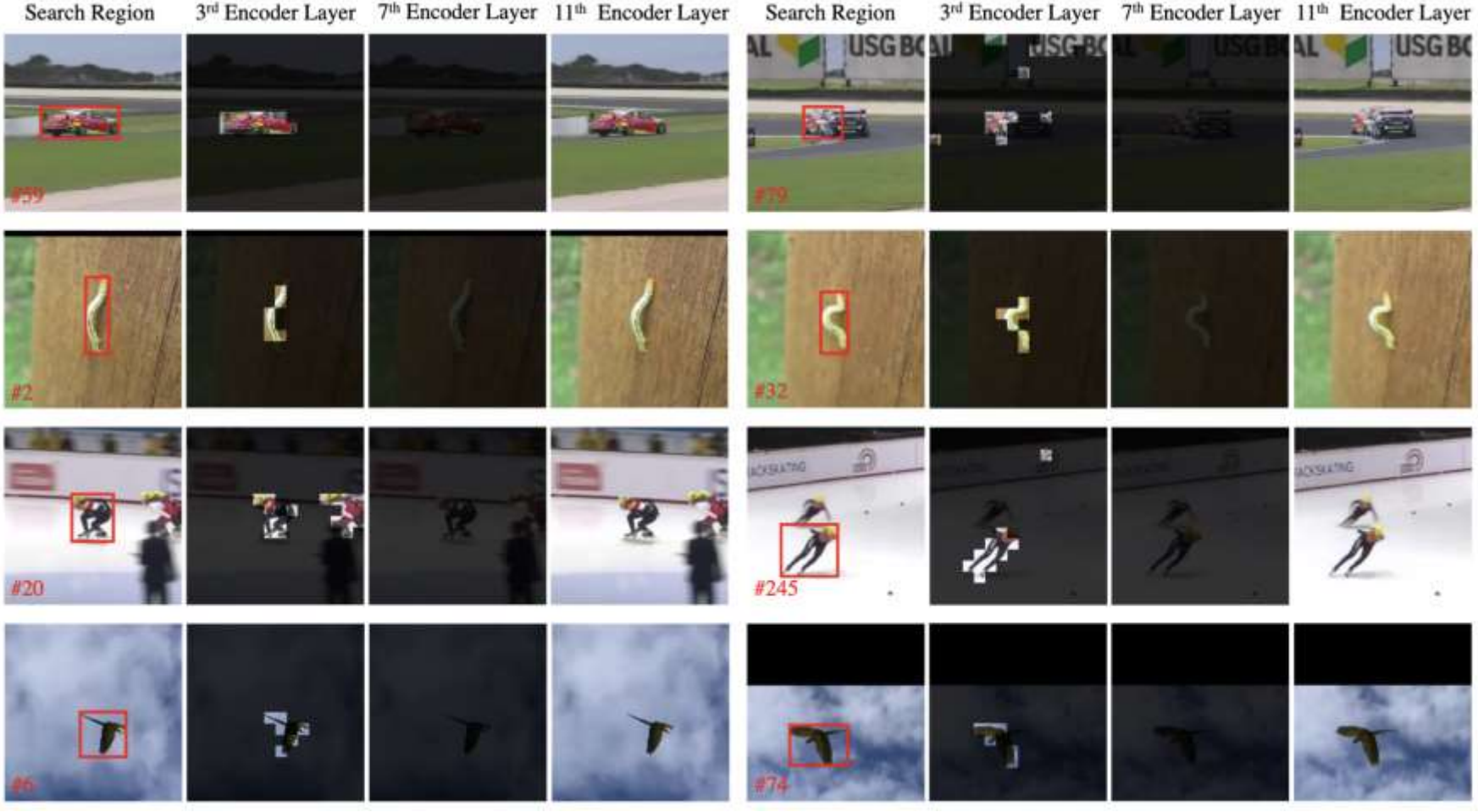
- Obstacle 1:
  - Parallel computation.
- Solution:
  - Attention masking strategy instead of separate attention operation.
- Effect:
  - 33 fps → 45 fps.

- Obstacle 2:
  - End-to-end optimization.
- Solution:
  - Gumbel-Softmax technique.
- Effect:
  - Search token division can be implicitly learned by the supervision from the target localization loss in a data-driven manner.

State-of-the-Art Comparison

Tracker	Source	GOT-10k* [16]			TrackingNet [31]			LaSOT [9]			AVisT [32]		
		AO	$SR_{0.5}$	$SR_{0.75}$	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AUC	OP50	OP75
GRM	Ours	73.4	82.9	70.4	84.0	88.7	83.3	69.9	79.3	75.8	54.5	63.1	45.2
OSTrack [44]	ECCV'22	71.0	80.4	68.2	83.1	87.8	82.0	69.1	78.7	75.2	-	-	-
AiATrack [11]	ECCV'22	69.6	80.0	63.2	82.7	87.8	80.4	69.0	79.4	73.8	-	-	-
SimTrack [3]	ECCV'22	68.6	78.9	62.4	82.3	86.5	-	69.3	78.5	74.0	-	-	-
RTS [33]	ECCV'22	-	-	-	81.6	86.0	79.4	69.7	76.2	73.7	50.8	55.7	38.9
Unicorn [40]	ECCV'22	-	-	-	83.0	86.4	82.2	68.5	76.6	74.1	-	-	-
MixFormer [5]	CVPR'22	70.7	80.0	67.8	83.1	88.1	81.6	69.2	78.7	74.7	53.7	63.0	43.0
ToMP [28]	CVPR'22	-	-	-	81.2	86.2	78.6	67.6	78.0	72.2	51.6	59.5	38.9
SBT [39]	CVPR'22	69.9	80.4	63.6	-	-	-	65.9	-	70.0	-	-	-
CSWinTT [36]	CVPR'22	69.4	78.9	65.4	81.9	86.7	79.5	66.2	75.2	70.9	-	-	-
STARK [41]	ICCV'21	68.0	77.7	62.3	81.3	86.1	78.1	66.4	76.3	71.2	51.1	59.2	39.1
KeepTrack [29]	ICCV'21	-	-	-	-	-	-	67.1	77.2	70.2	49.4	56.3	37.8
AutoMatch [48]	ICCV'21	65.2	76.6	54.3	76.0	-	72.6	58.3	-	59.9	-	-	-
TransT [4]	CVPR'21	67.1	76.8	60.9	81.4	86.7	80.3	64.9	73.8	69.0	49.0	56.4	37.2
Alpha-Refine [43]	CVPR'21	-	-	-	80.5	85.6	78.3	65.3	73.2	68.0	49.6	55.7	38.2
TMT [38]	CVPR'21	67.1	77.7	58.3	78.4	83.3	73.1	63.9	-	61.4	48.1	55.3	33.8
Ocean [50]	ECCV'20	61.1	72.1	47.3	-	-	-	56.0	65.1	56.6	38.9	43.6	20.5
PrDiMP [7]	CVPR'20	63.4	73.8	54.3	75.8	81.6	70.4	59.8	68.8	60.8	43.3	48.0	28.7
SiamAttn [46]	CVPR'20	-	-	-	75.2	81.7	-	56.0	64.8	-	-	-	-
DiMP [2]	ICCV'19	61.1	71.7	49.2	74.0	80.1	68.7	56.9	65.0	56.7	41.9	45.7	26.0
ATOM [6]	CVPR'19	-	-	-	70.3	77.1	64.8	51.5	57.6	50.5	38.6	41.5	22.2
SiamRPN++ [21]	CVPR'19	51.7	61.6	32.5	73.3	80.0	69.4	49.6	56.9	49.1	39.0	43.5	21.2

Visualization Results



Concluding Remarks

- We propose a generalized relation modeling method, which inherits the strengths of both the two-stream and one-stream pipelines while being more flexible.

- We devise a token division module with an attention masking strategy and the Gumbel-Softmax technique to adaptively classify the input tokens.

Paper

Code

Transformer Tracking