基于Transformer的视觉跟踪算法探索

Dong Wang

wdice@dlut.edu.cn

Dalian University of Technology

Visual Object Tracking

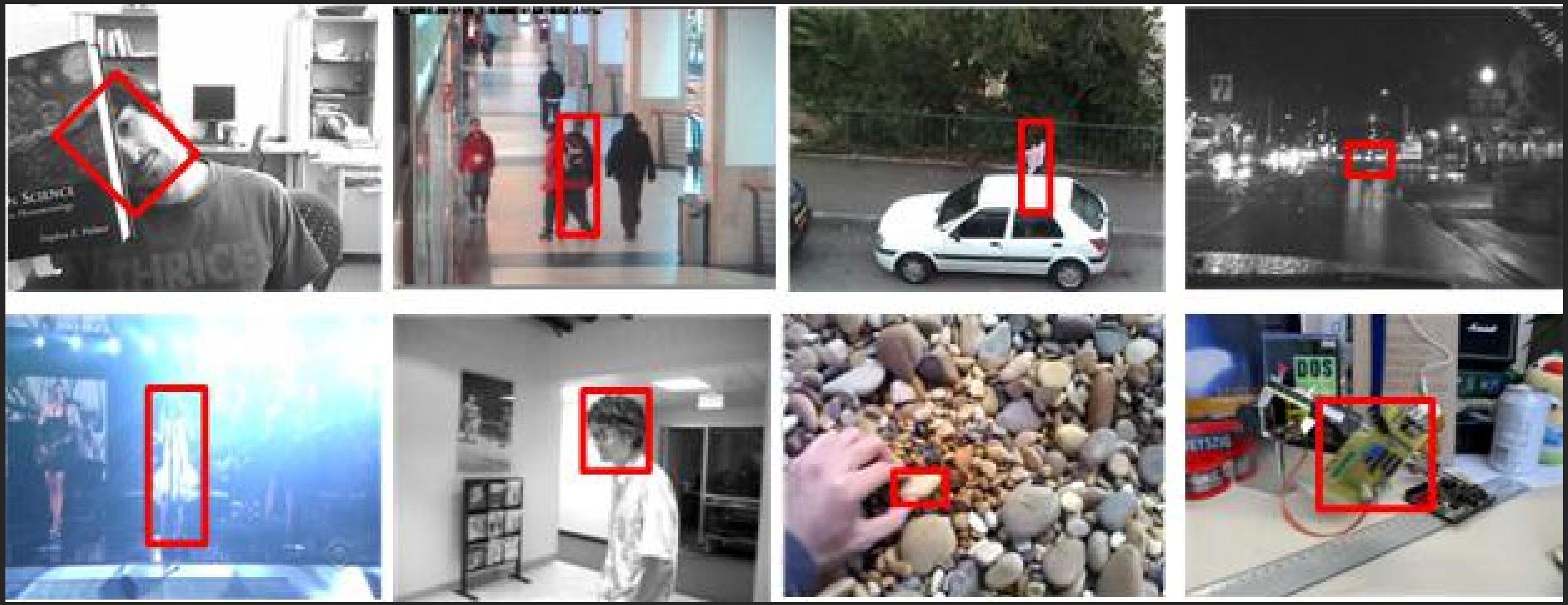
- **Goal**

  Track an arbitrary object in a video given its initial location

  Single-object, **Category-free**

- **Challenges**

  Occlusion, Light Change, Background Clutter, etc.



Visual Object Tracking Benchmarks

[LaSOT][2018] 1120 sequences for training and 280 for testing, 4M images, long-term.

[GOT-10k][2018] 10k sequences for training and 180 for testing, 3.5M images.

[TrackingNet][2018] 30k sequences for training and 511 for testing, 15M images.

[VOT Challenge][2013-2021] 60 challenging sequences for testing, 20k images.

[NfS][2017] 100 sequences for testing, 383k images, fast-moving objects.

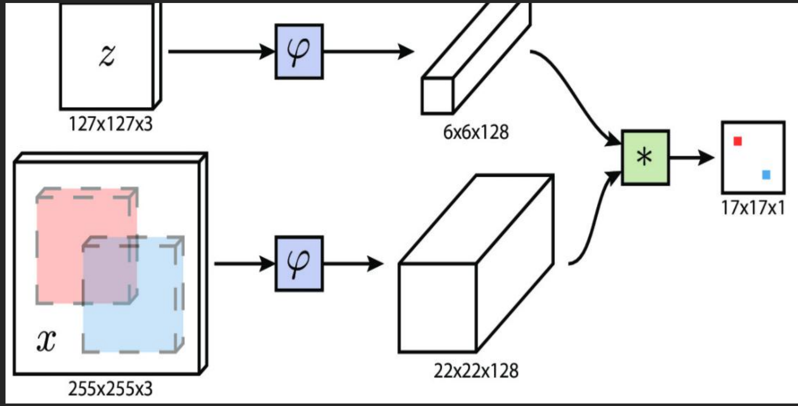
[UAV123][2016] 123 sequences for testing, 113k images, low altitude aerial videos.

[OTB100][2015] 100 sequences for testing, 59k images.

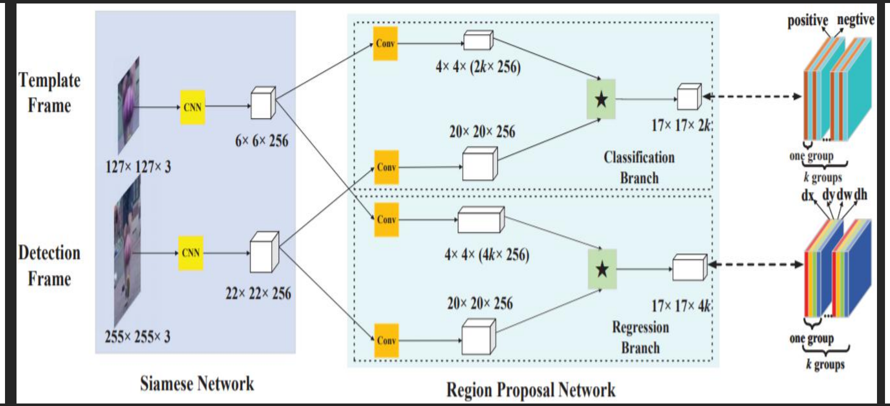
[TC128][2015] 128 sequences for testing, 55k images.

Visual Object Tracking: One-shot vs Online

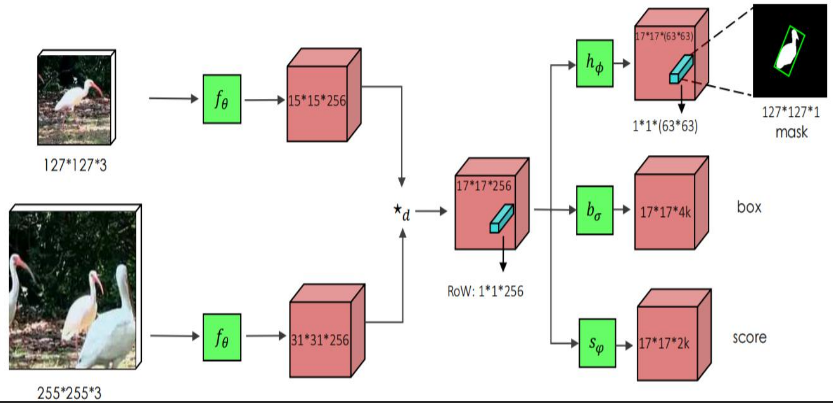
SiamFC (ECCVW16)



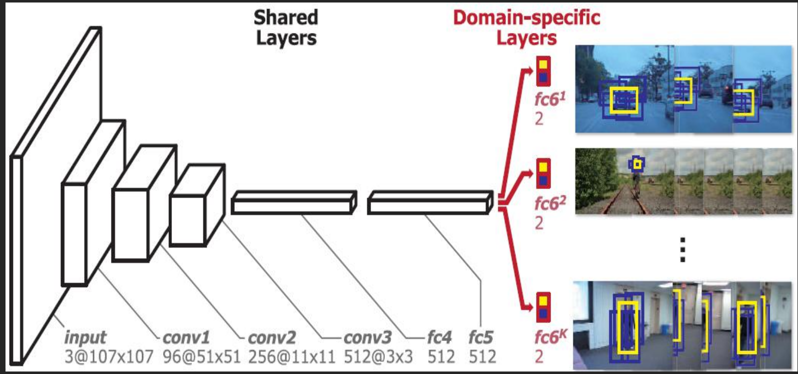
SiamRPN (CVPR18)



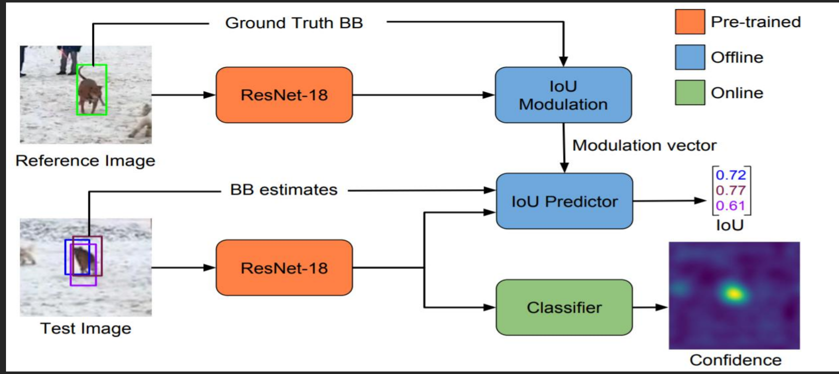
SiamMask (CVPR19)



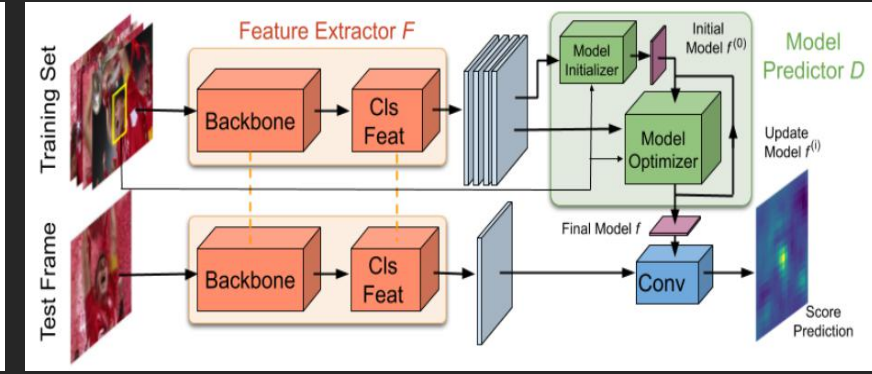
MDNet (CVPR16)



ATOM (CVPR19)

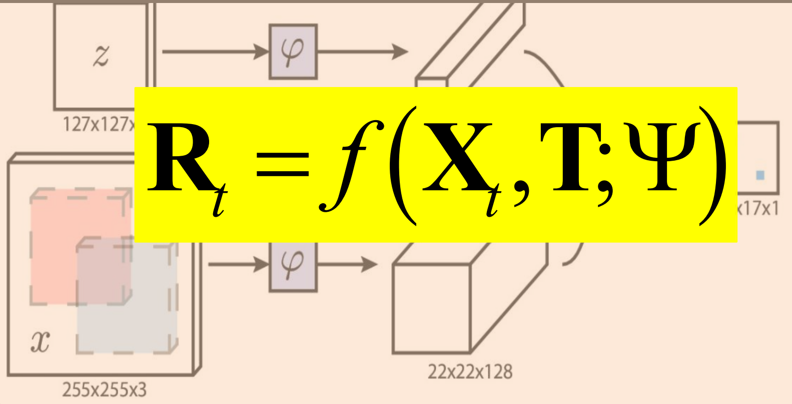


DiMP (ICCV19)



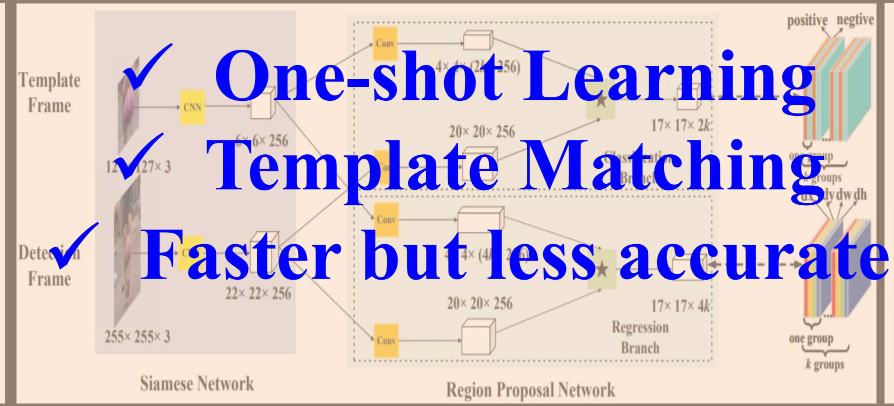
Visual Object Tracking: One-shot vs Online

SiamFC (ECCVW16)



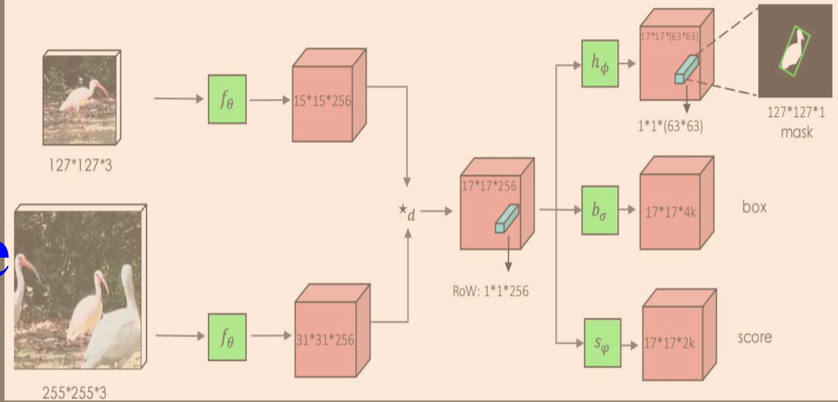
$$\mathbf{R}_t = f(\mathbf{X}_t, \mathbf{T}; \Psi)$$

SiamRPN (CVPR18)

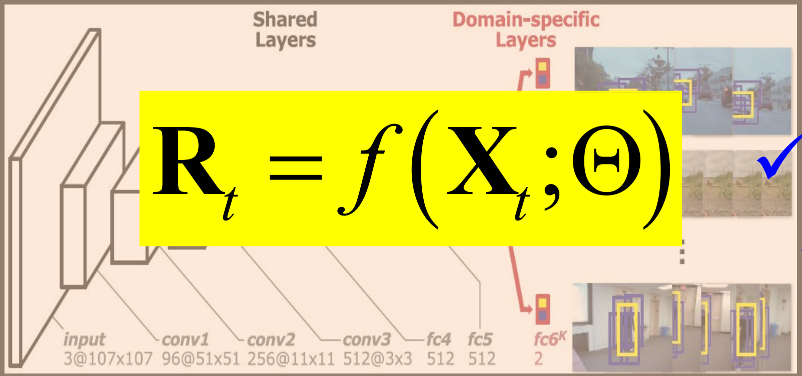


- ✓ One-shot Learning
- ✓ Template Matching
- ✓ Faster but less accurate

SiamMask (CVPR19)

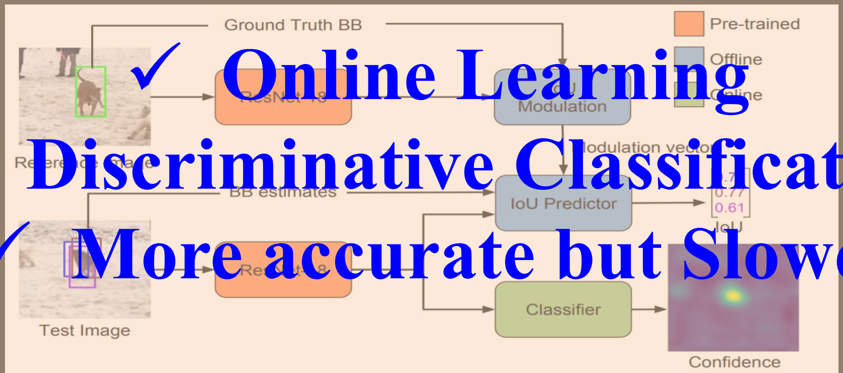


MDNet (CVPR16)



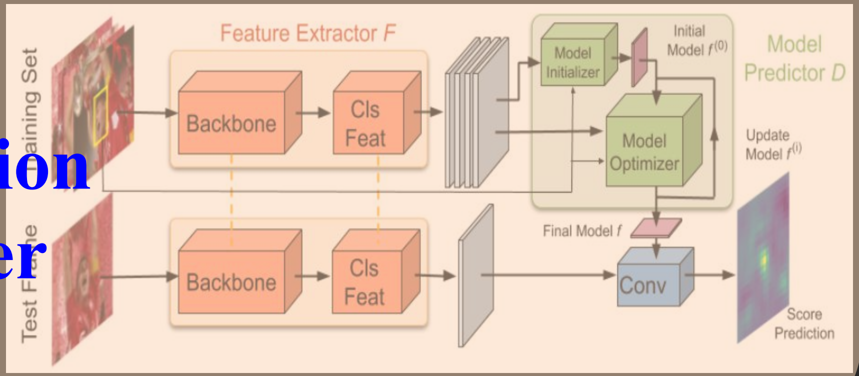
$$\mathbf{R}_t = f(\mathbf{X}_t; \Theta)$$

ATOM (CVPR19)

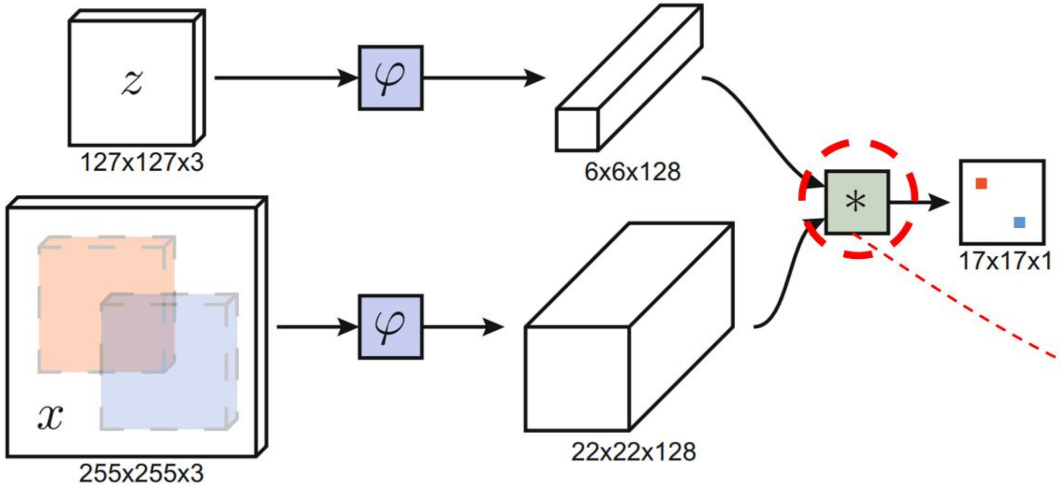


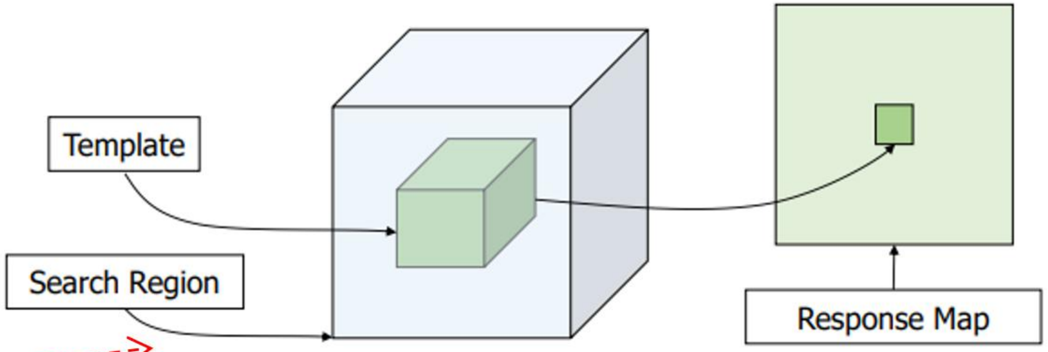
- ✓ Online Learning
- ✓ Discriminative Classification
- ✓ More accurate but Slower

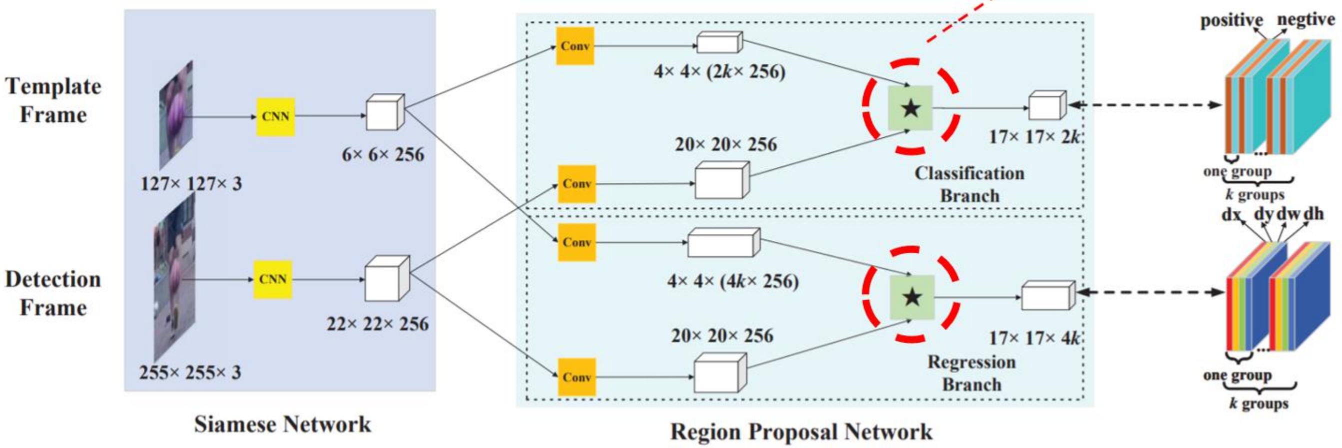
DiMP (ICCV19)



Correlation-based Siamese Tracking

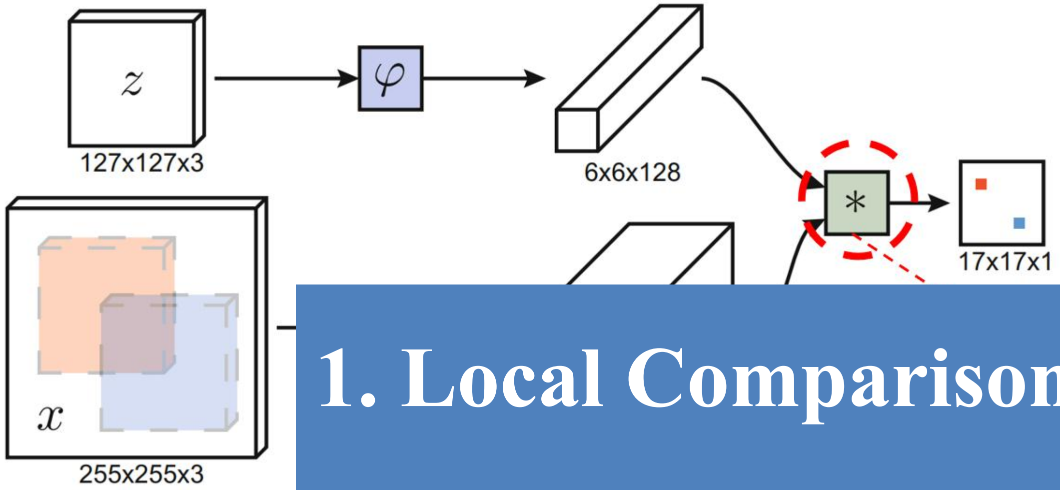


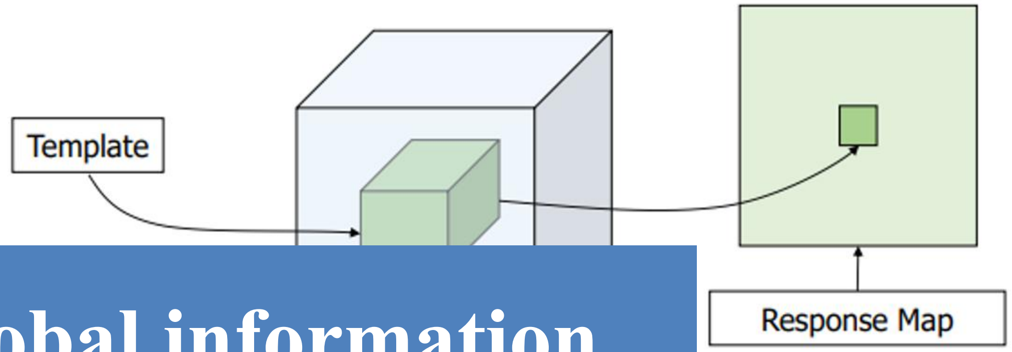




Correlation!

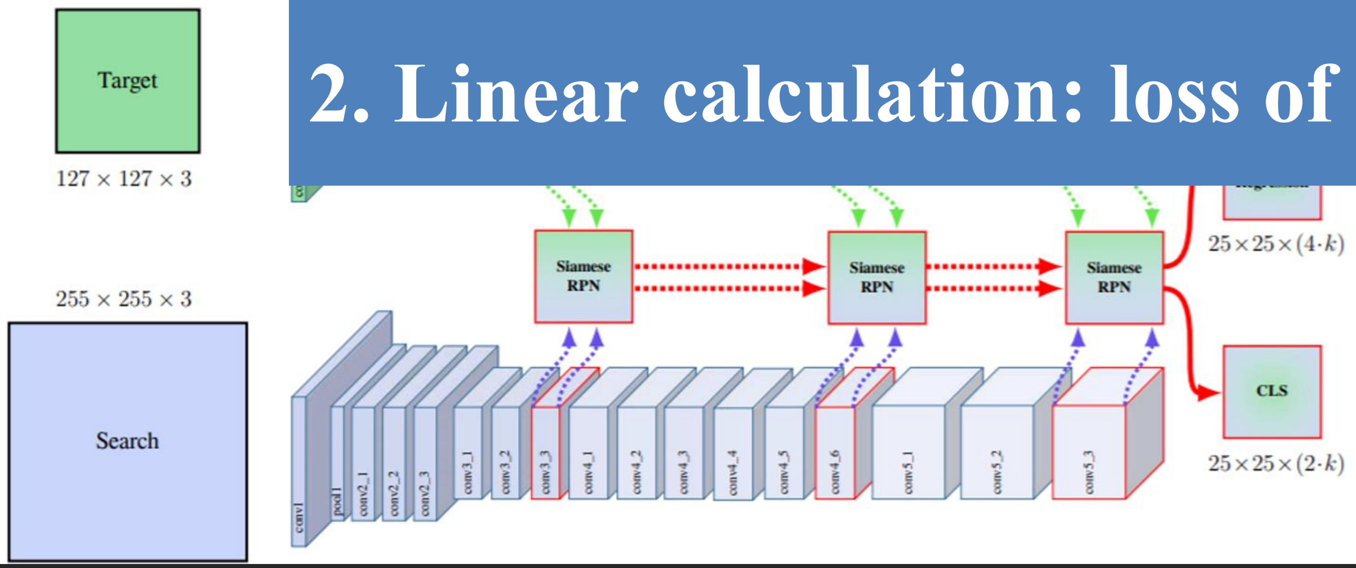
Correlation-based Siamese Tracking

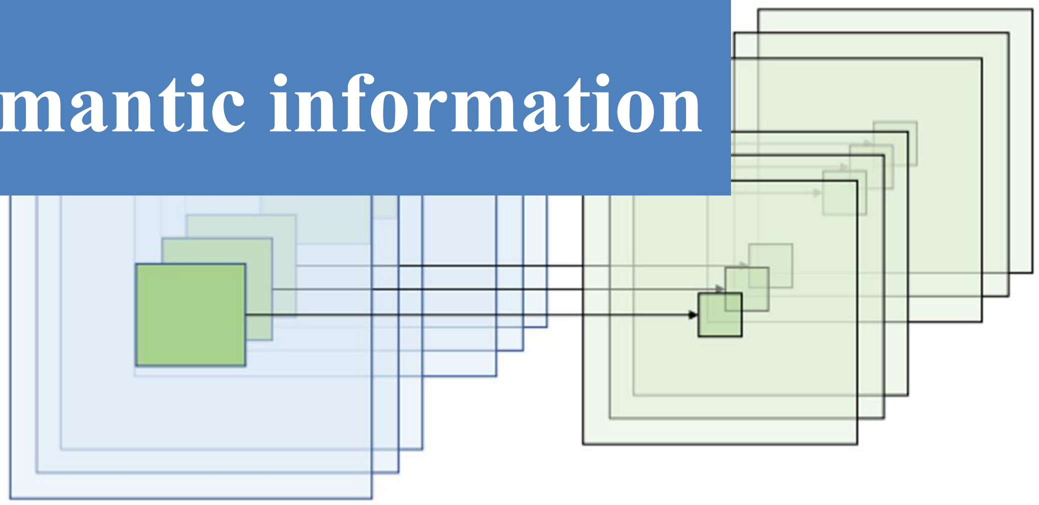




1. Local Comparison: lack of global information

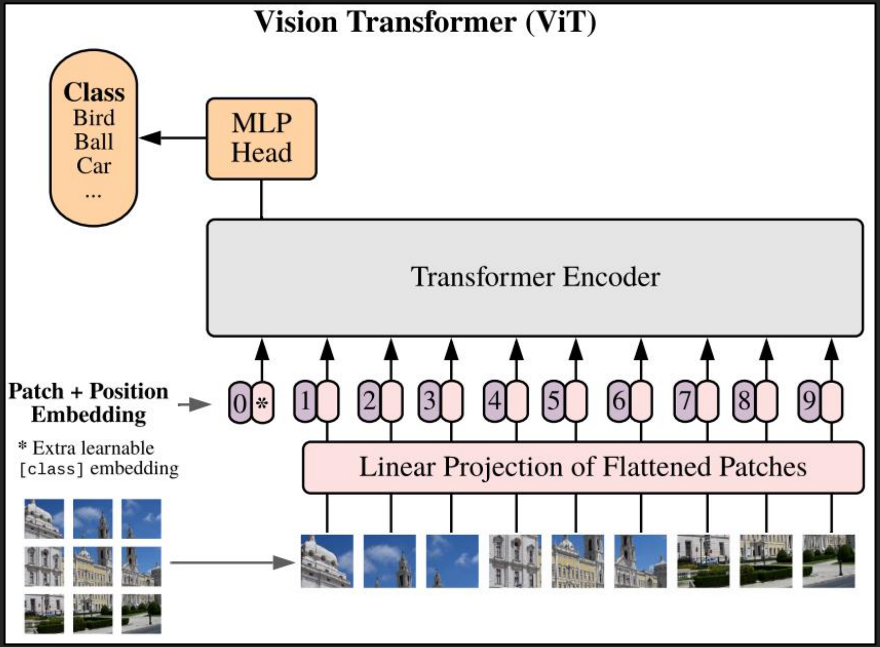
2. Linear calculation: loss of semantic information



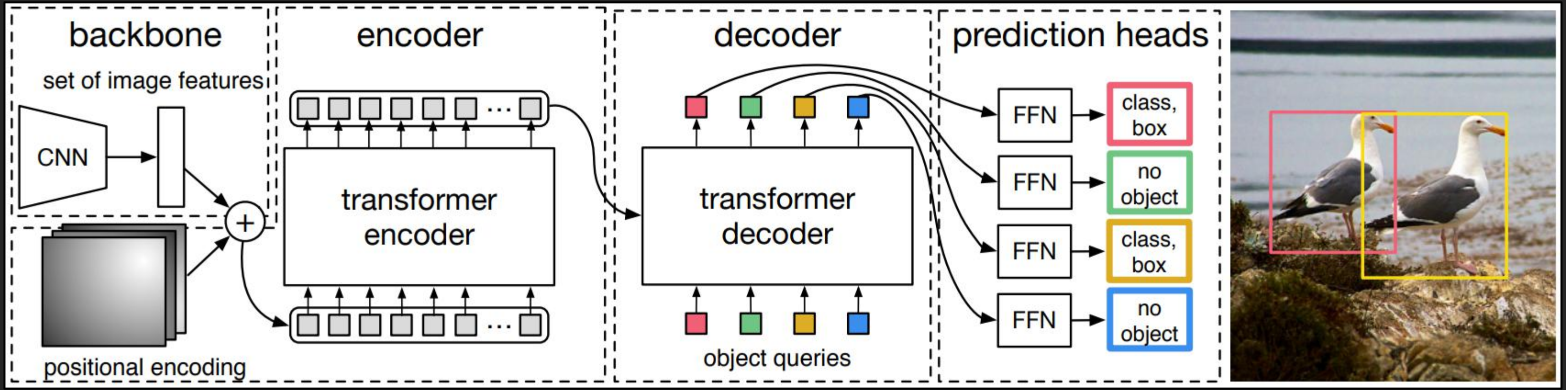


Transformer in Computer Vison

Google: ViT

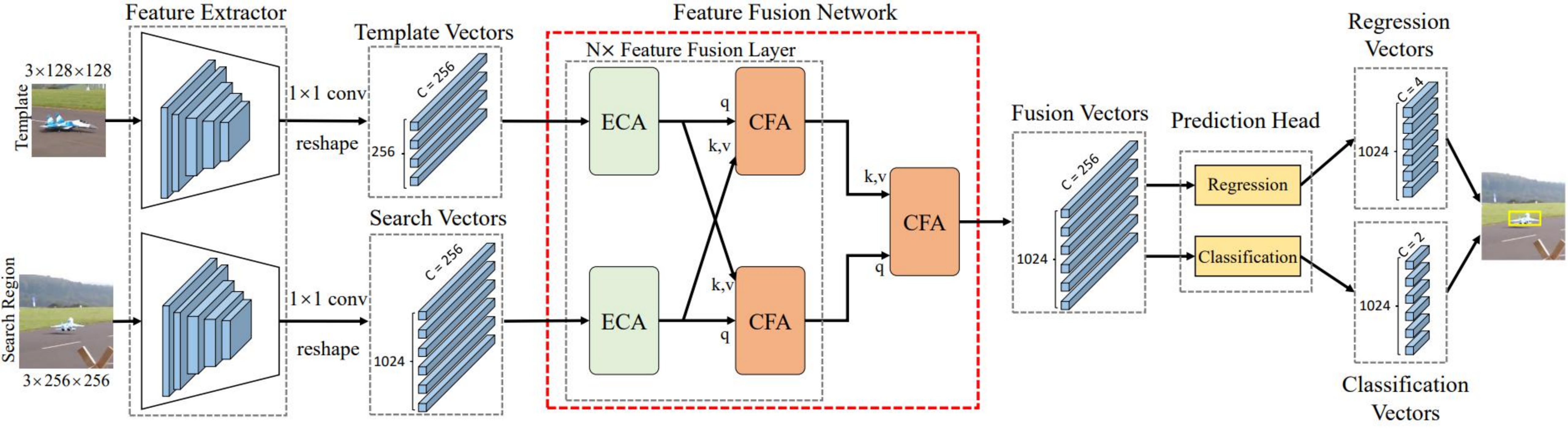


Facebook: DETR



What could Transformer bring to visual tracking?

Transformer Tracking



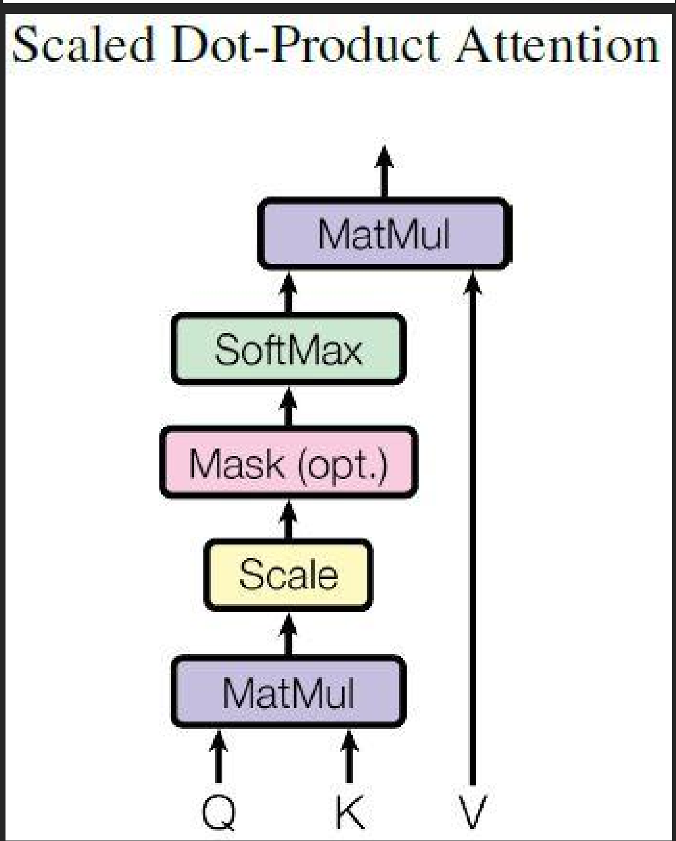
Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun yang, Huchuan Lu. Transformer Tracking. CVPR, 2021.

- Code: https://github.com/chenxin-dlut/TransT

Transformer Tracking

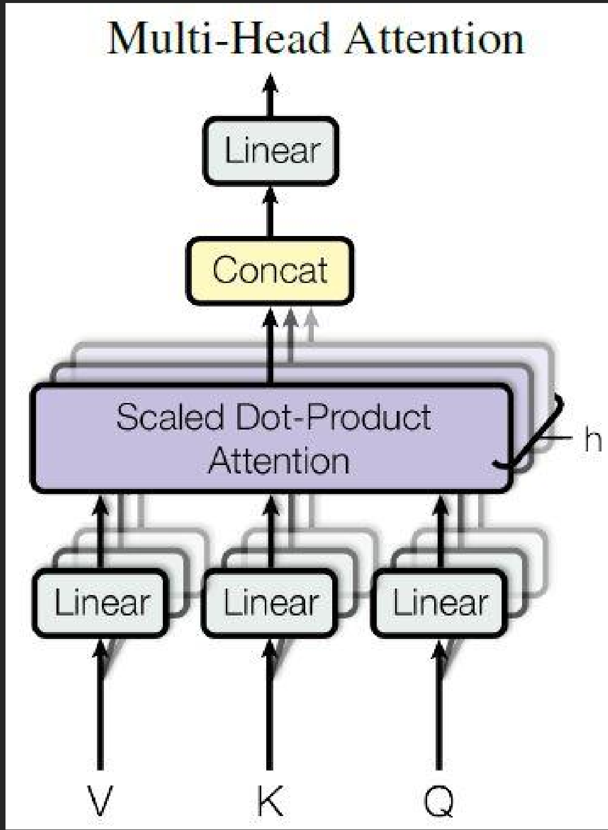
➢ Attention to Replace "Correlation"

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}})\mathbf{V}$$



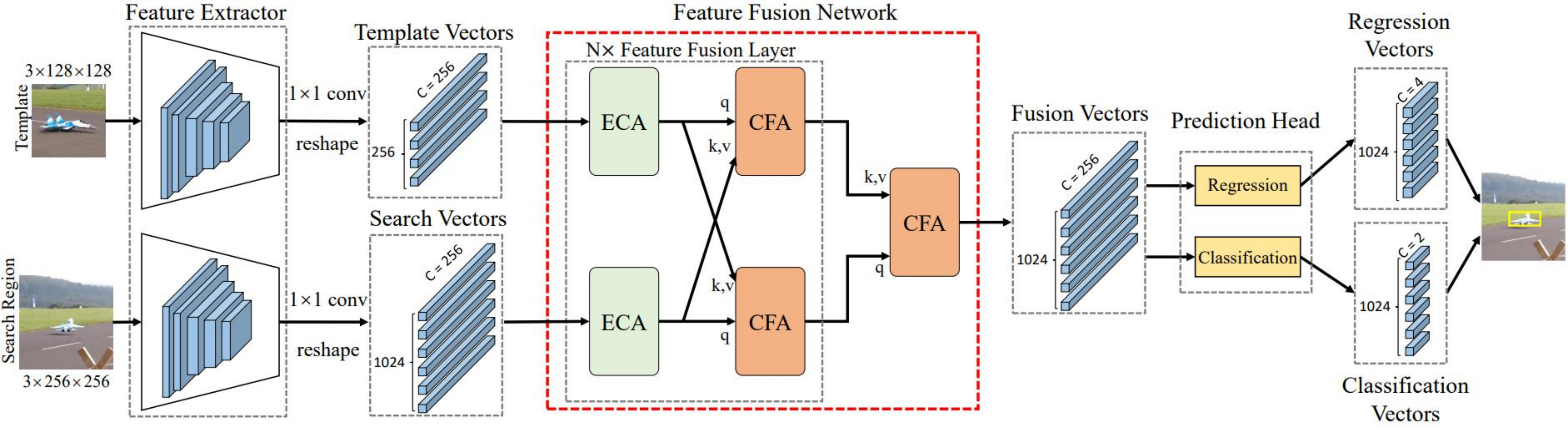
$$\mathrm{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Concat}(\mathbf{H}_1, ..., \mathbf{H}_{n_h})\mathbf{W}^O$$

$$\mathbf{H}_i = \mathrm{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V)$$



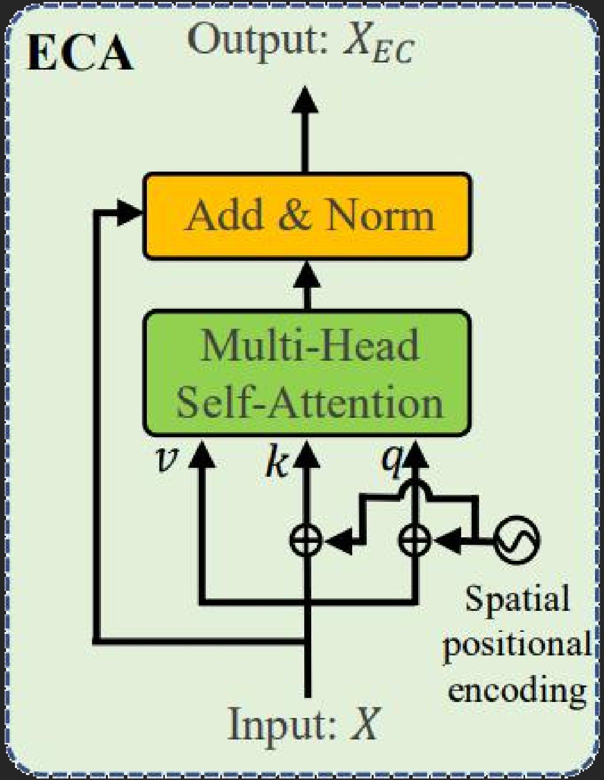
Transformer Tracking

Our TransT Framework

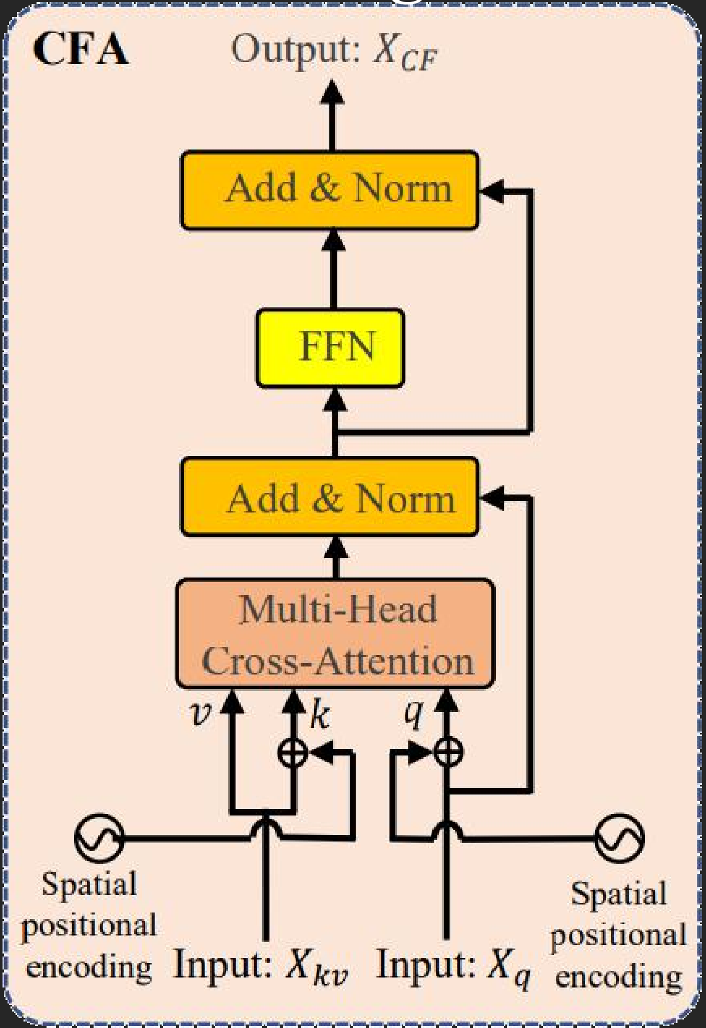


Transformer Tracking

Ego-Context Augment Module



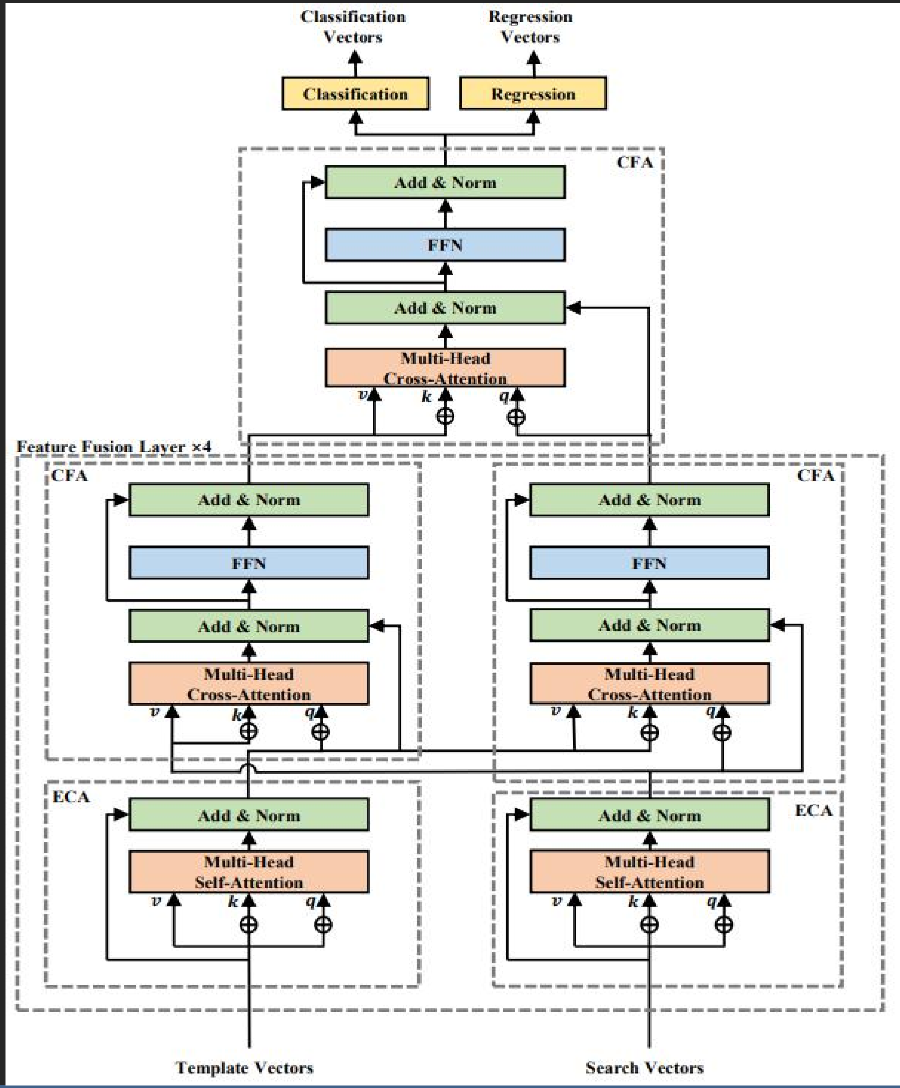
Cross-Feature Augment Module



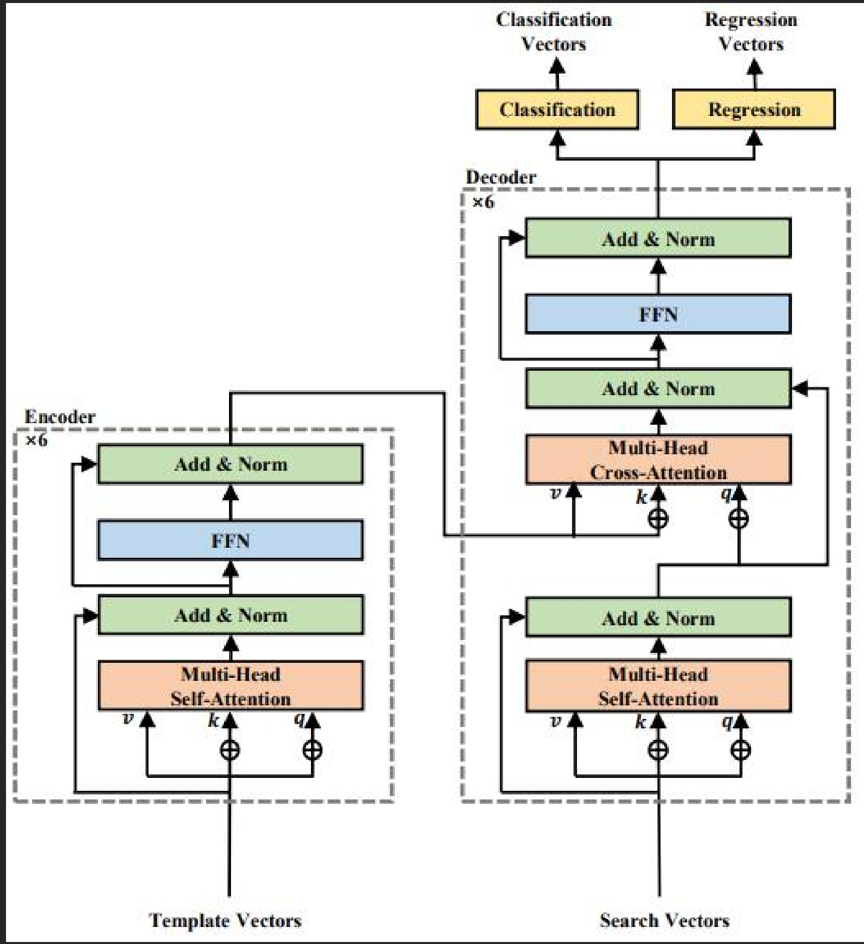
- ✓ ECA based on self-attention and CFA based on cross-attention
- ✓ CFA performs feature fusion, retaining rich semantic information
- ✓ ECA and CFA establish dependence between long distance features and aggregate global information

Transformer Tracking

Our Feature Fusion Network

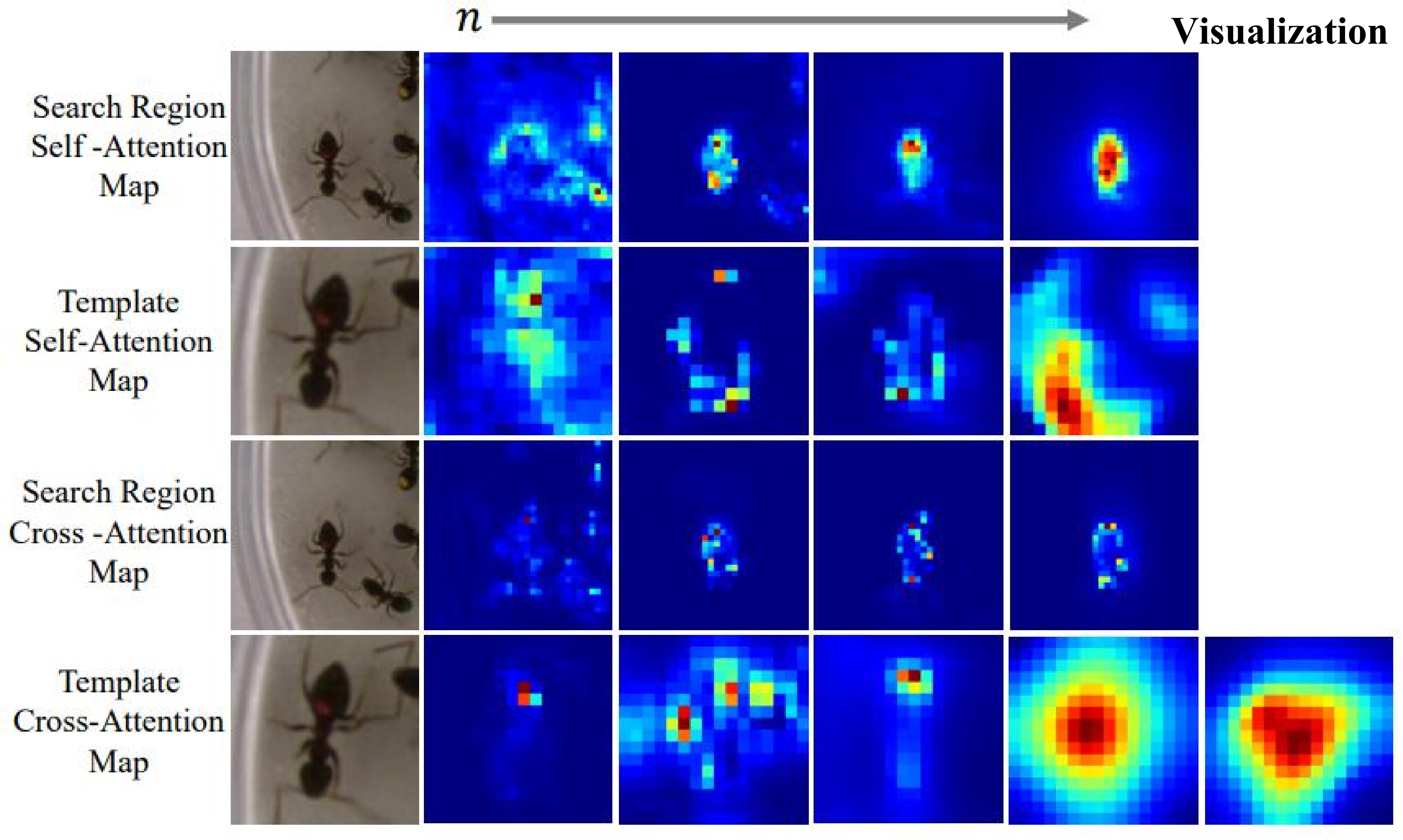


Original Transformer



Similar with DETR

Transformer Tracking



Transformer Tracking

https://github.com/chenxin-dlut/TransT

DUT, IIAU-LAB

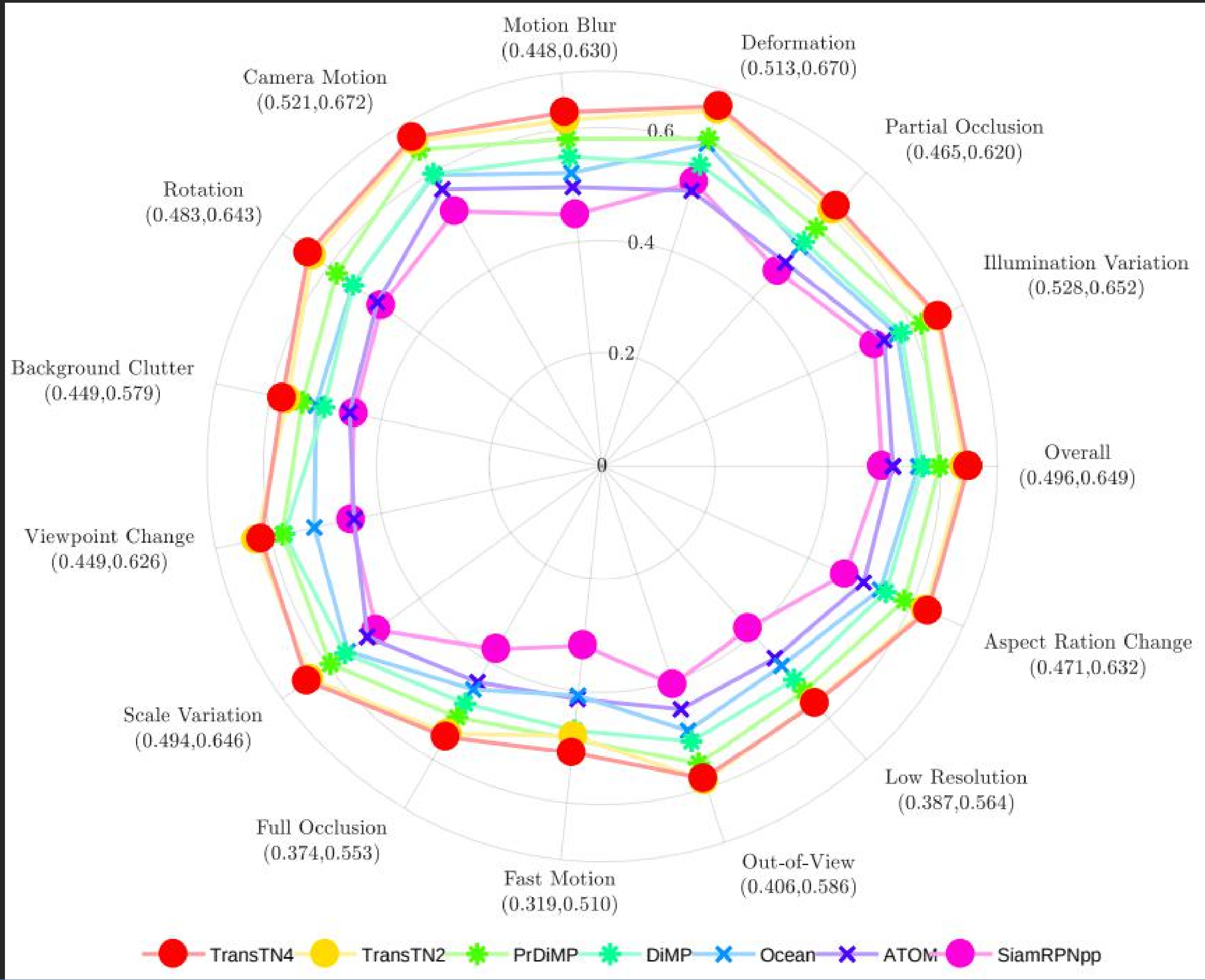
Transformer Tracking

Large-scale Benchmark Results

Method	Source	LaSOT [14]			TrackingNet [30]			GOT-10k [19]			
		AUC	P_{Norm}	P	AUC	P_{Norm}	P	AO	$SR_{0.5}$	$SR_{0.75}$	
TransT	Ours	**64.9**	**73.8**	**69.0**	**81.4**	**86.7**	**80.3**	**72.3**	**82.4**	**68.2**	46*fps*
TransT-N2	Ours	64.2	**73.5**	68.2	80.9	**86.4**	79.2	**69.9**	**80.1**	**65.9**	66*fps*
TransT-GOT	Ours	-	-	-	-	-	-	67.1	76.8	60.9	
SiamR-CNN [39]	CVPR2020	**64.8**	72.2	-	**81.2**	85.4	**80.0**	64.9	72.8	59.7	
Ocean [48]	ECCV2020	56.0	65.1	56.6	-	-	-	61.1	72.1	47.3	
KYS [3]	ECCV2020	55.4	63.3	-	74.0	80.0	68.8	63.6	75.1	51.5	
DCFST [49]	ECCV2020	-	-	-	75.2	80.9	70.0	63.8	75.3	49.8	
SiamFC++ [44]	AAAI2020	54.4	62.3	54.7	75.4	80.0	70.5	59.5	69.5	47.9	
PrDiMP [10]	CVPR2020	59.8	68.8	60.8	75.8	81.6	70.4	63.4	73.8	54.3	
CGACD [13]	CVPR2020	51.8	62.6	-	71.1	80.0	69.3	-	-	-	
SiamAttn [46]	CVPR2020	56.0	64.8	-	75.2	81.7	-	-	-	-	
MAML [40]	CVPR2020	52.3	-	-	75.7	82.2	72.5	-	-	-	
D3S [26]	CVPR2020	-	-	-	72.8	76.8	66.4	59.7	67.6	46.2	
SiamCAR [16]	CVPR2020	50.7	60.0	51.0	-	-	-	56.9	67.0	41.5	
SiamBAN [5]	CVPR2020	51.4	59.8	52.1	-	-	-	-	-	-	
DiMP [2]	ICCV2019	56.9	65.0	56.7	74.0	80.1	68.7	61.1	71.7	49.2	
SiamPRN++ [21]	CVPR2019	49.6	56.9	49.1	73.3	80.0	69.4	51.7	61.6	32.5	
ATOM [9]	CVPR2019	51.5	57.6	50.5	70.3	77.1	64.8	55.6	63.4	40.2	
ECO [8]	ICCV2017	32.4	33.8	30.1	55.4	61.8	49.2	31.6	30.9	11.1	
MDNet [31]	CVPR2016	39.7	46.0	37.3	60.6	70.5	56.5	29.9	30.3	9.9	
SiamFC [1]	ECCVW2016	33.6	42.0	33.9	57.1	66.3	53.3	34.8	35.3	9.8	

Transformer Tracking

LaSOT Attribute

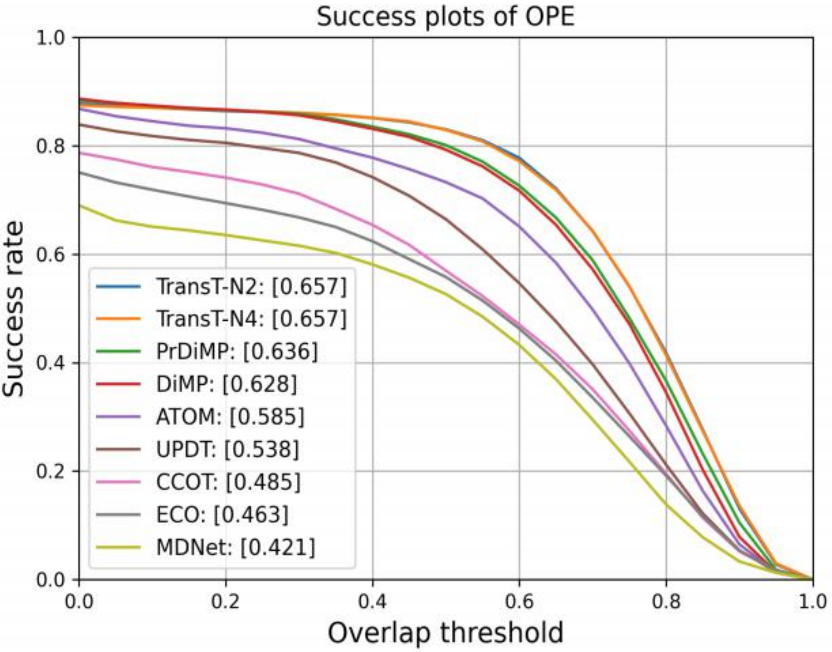


https://github.com/chenxin-dlut/TransT

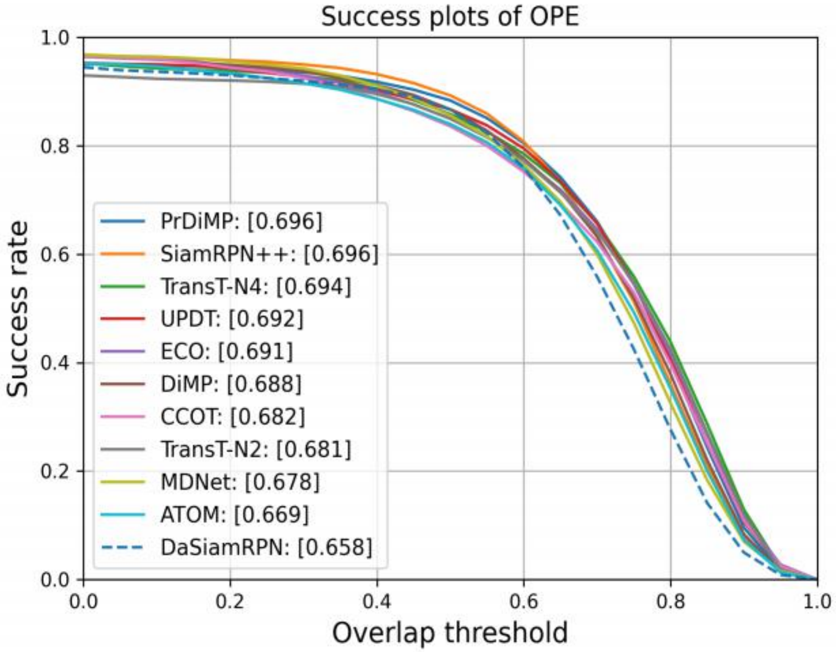
DUT, IIAU-LAB

Transformer Tracking

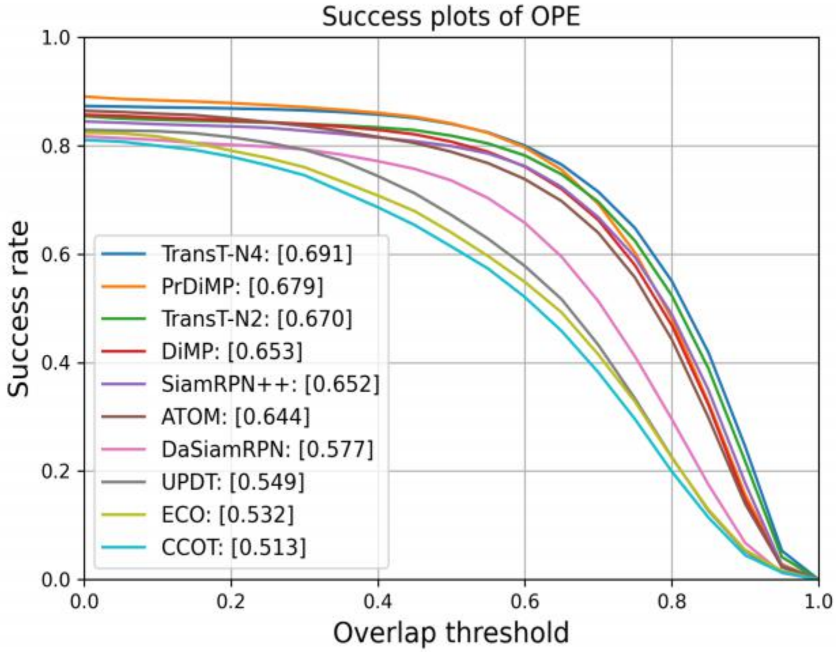
Results on NFS, OTB2015 and UAV123



(a) NFS



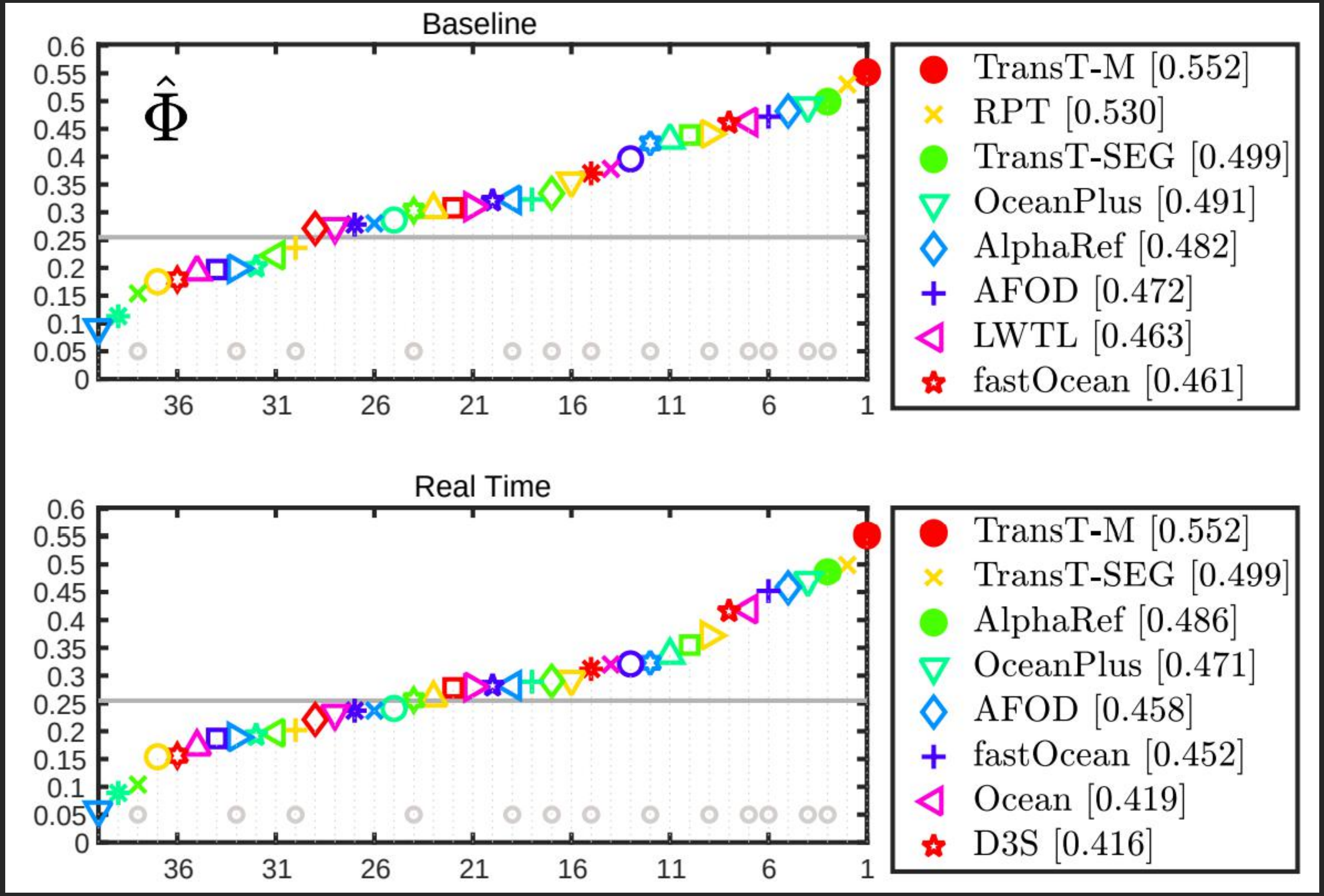
(b) OTB2015



(c) UAV123

Transformer Tracking

Results on VOT2020



Transformer Tracking

Ablation Study

Method	LaSOT [14]			TrackingNet [30]			GOT-10k [19]		
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AO	$SR_{0.5}$	$SR_{0.75}$
TransT	**64.9**	**73.8**	**69.0**	**81.4**	**86.7**	**80.3**	**72.3**	**82.4**	**68.2**
TransT-np	62.9	71.5	66.9	81.1	86.4	80.0	71.5	81.5	67.5
TransT(ori)	62.3	71.1	66.2	81.3	86.1	78.9	70.3	80.2	65.8
TransT(ori)-np	60.9	69.4	64.8	80.9	85.6	78.4	68.6	78.2	65.1

Transformer Tracking

Ablation Study

Method	ECA	CFA	Correlation	LaSOT [14]			TrackingNet [30]			GOT-10k [19]		
				AUC	P_{Norm}	P	AUC	P_{Norm}	P	AO	$SR_{0.5}$	$SR_{0.75}$
TransT	√	√		**64.9**	**73.8**	**69.0**	**81.4**	**86.7**	**80.3**	**72.3**	**82.4**	**68.2**
TransT		√		62.9	71.9	66.2	81.1	86.2	79.1	70.6	81.2	65.7
TransT	√		√	57.7	65.4	59.5	77.5	82.2	74.0	62.8	72.2	54.8
TransT			√	47.7	48.6	41.7	68.8	71.4	60.9	50.9	58.0	33.3
TransT-np	√	√		62.9	71.5	66.9	81.1	86.4	80.0	71.5	81.5	67.5
TransT-np		√		61.0	69.6	64.5	80.0	85.0	77.9	68.1	78.3	64.0
TransT-np	√		√	57.3	65.2	58.8	76.2	80.8	72.8	61.4	70.7	53.7
TransT-np			√	35.3	17.9	20.1	46.5	40.3	27.4	38.2	36.8	7.0

Conclusion

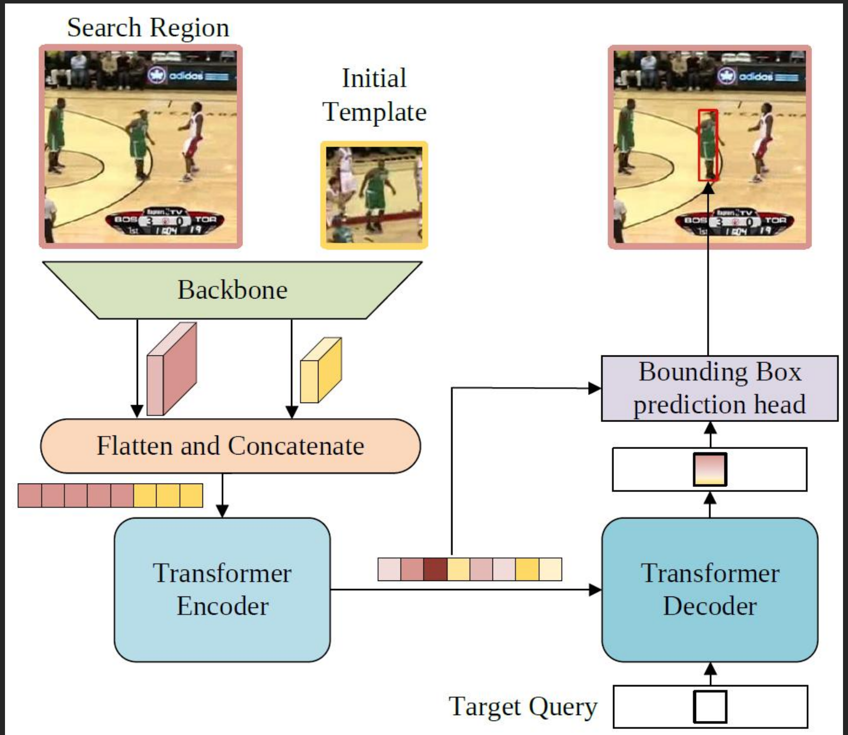
- A New Transformer-based tracking Framework
- Completely offline, high performance and real-time speed
- Code and Models: https://github.com/chenxin-dlut/TransT

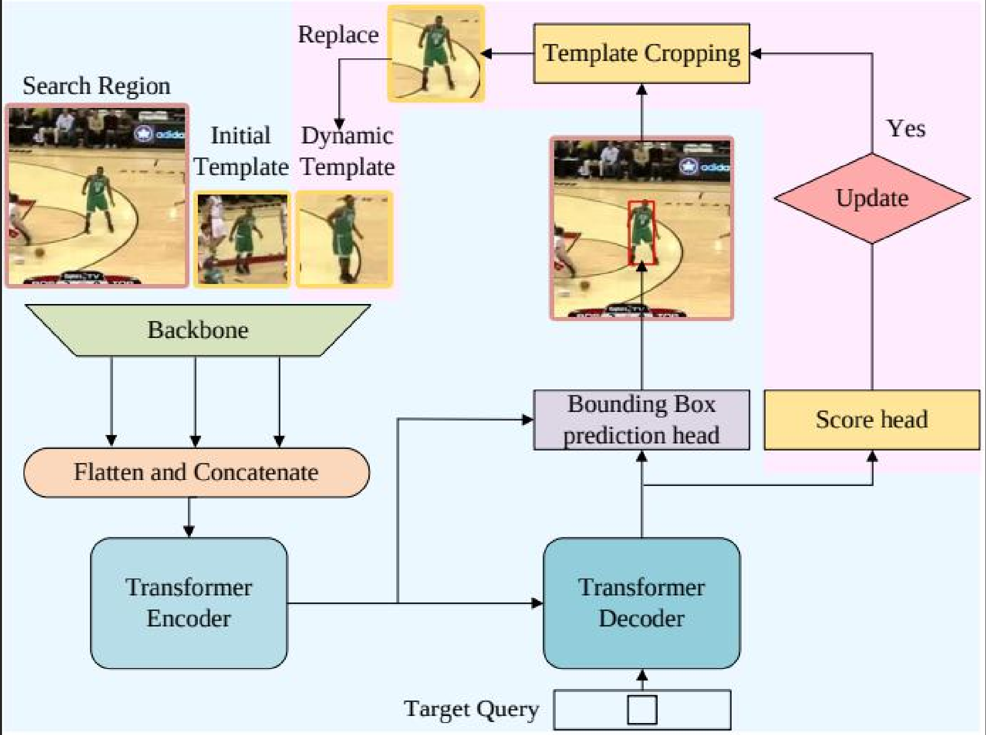
Transformer Tracking

■ Disadvantages

- ➢ TransT still hasn't gotten rid of post-processing completely
- ➢ TransT does not consider temporal information

Learning Spatio-Temporal Transformer for Visual Tracking



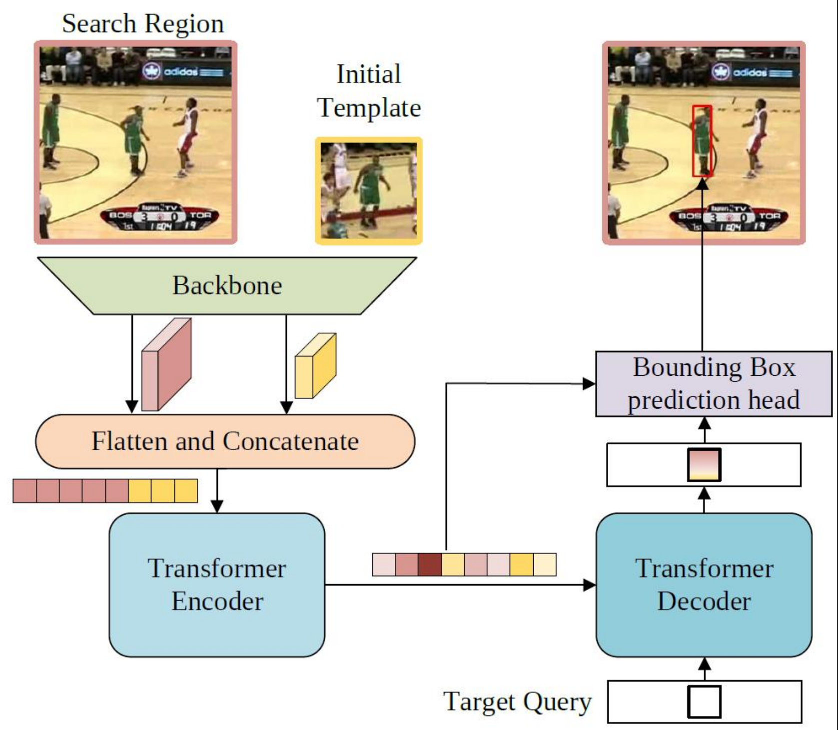


Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, Huchuan Lu. Learning Spatio-Temporal Transformer for Visual Tracking. ICCV, 2021.

- Code: https://github.com/researchmm/Stark

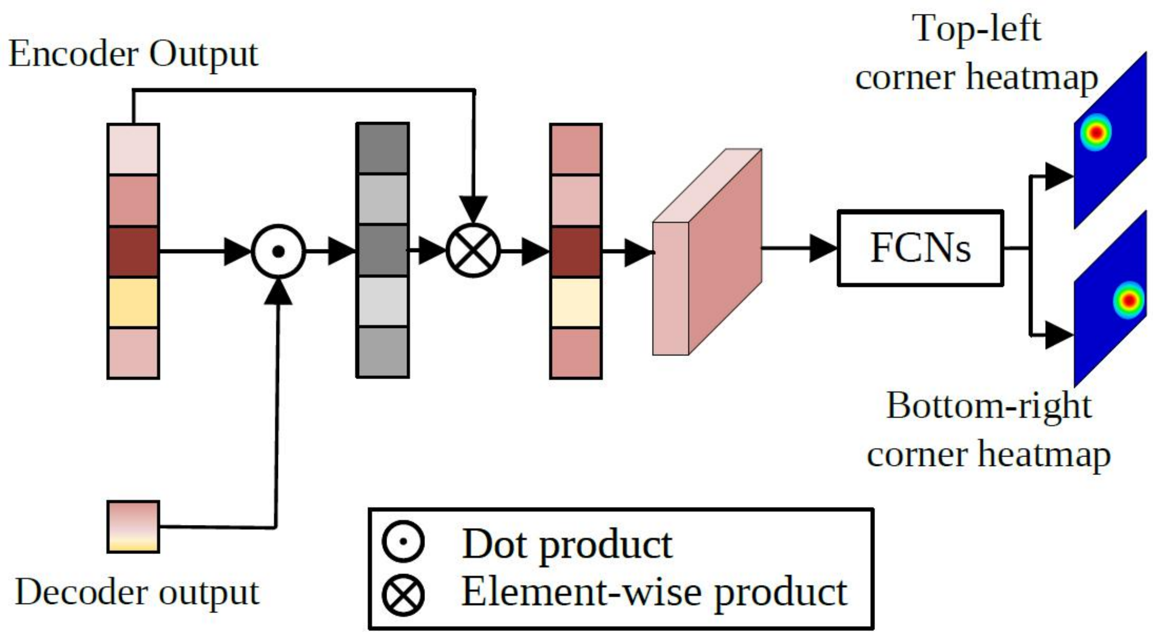
Learning Spatio-Temporal Transformer for Visual Tracking

Architecture



- ✓ Transformer architecture for feature integration

Corner prediction Head

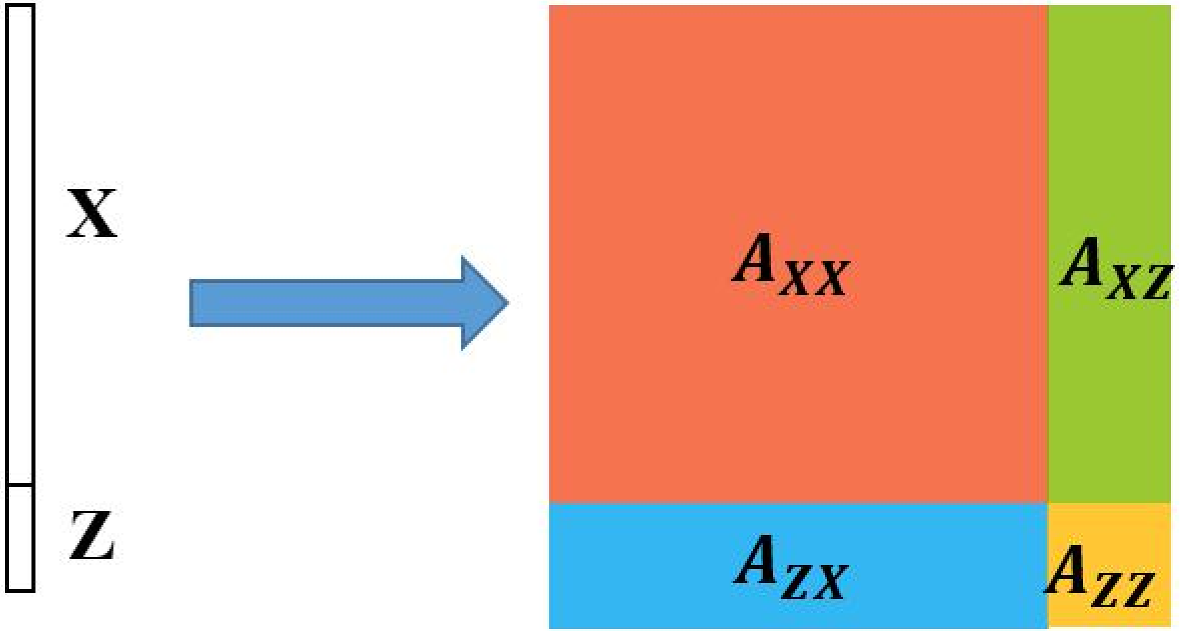


- ✓ Tracking as a direct end-to-end bounding box prediction problem
- ✓ Totally post-processing free

Learning Spatio-Temporal Transformer for Visual Tracking

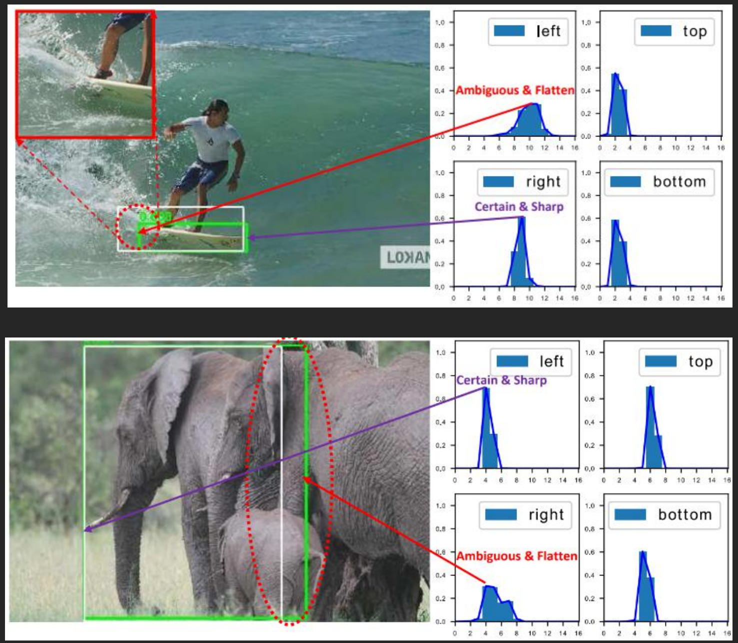
Insights behind the "concatenation" operation

$$\text{Attention}(Q,K,V) = \underbrace{\boxed{\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)}}_{\mathbf{A}} V$$



- ✓ **Implicitly modeling 4 types of feature interaction.**
- ✓ **Scalable to more inputs, such as more templates or more search regions**

Why predict heatmaps rather than directly predicting coords?

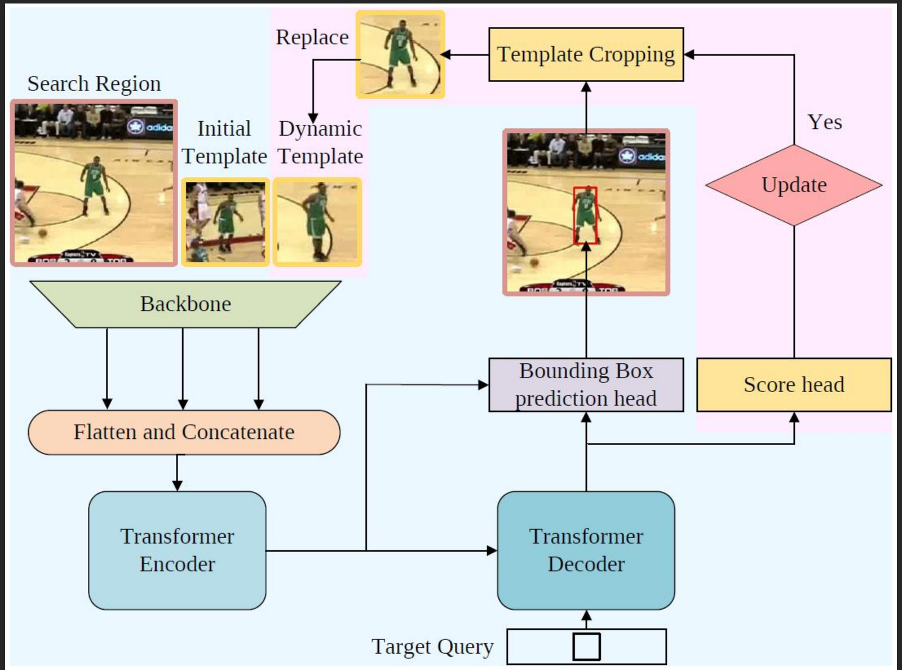


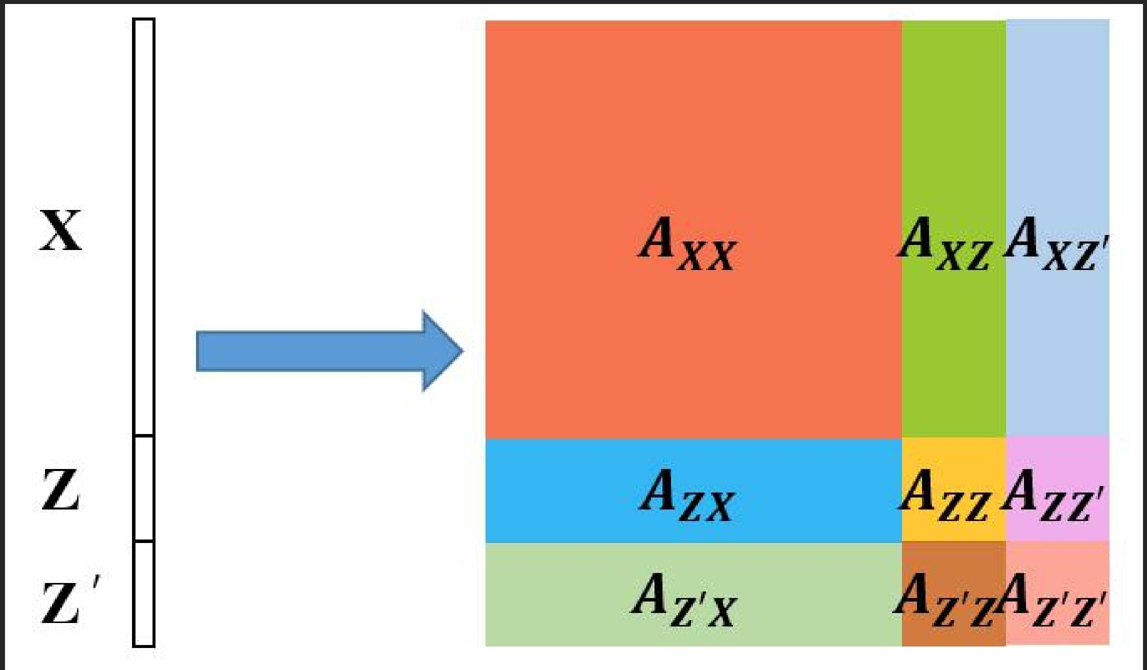
Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection (https://arxiv.org/pdf/2006.04388.pdf)

- ✓ Directly predicting coordinates is equivalent to fitting a Delta-Distribution
- ✓ However, there are many cases where the bounding box coordinates have large uncertainty (such as TrackingNet GTs)

Learning Spatio-Temporal Transformer for Visual Tracking

Dynamic Template





- ✓ **Dynamic template changes over time, bringing temporal information for the STARK tracker**
- ✓ **An update controller to control the update of the dynamic template**

Learning Spatio-Temporal Transformer for Visual Tracking

Experimental Results (Short-Term)

GOT-10K

	SiamFC [2]	SiamFCv2 [52]	ATOM [11]	SiamFC++ [59]	D3S [38]	DiMP50 [3]	Ocean [69]	PrDiMP50 [12]	SiamRCNN [54]	**STARK -S50**	**STARK -ST50**	**STARK -ST101**
AO(%)	34.8	37.4	55.6	59.5	59.7	61.1	61.1	63.4	64.9	**67.2**	**68.0**	**68.8**
SR0.5(%)	35.3	40.4	63.4	69.5	67.6	71.7	72.1	73.8	72.8	**76.1**	**77.7**	**78.1**
SR0.75(%)	9.8	14.4	40.2	47.9	46.2	49.2	47.3	54.3	59.7	**61.2**	**62.3**	**64.1**

TrackingNet

	DSiamRPN [70]	ATOM [11]	SiamRPN++ [28]	DiMP50 [3]	SiamAttn [65]	SiamFC++ [59]	MAML-FCOS [55]	PrDiMP50 [12]	SiamRCNN [54]	**STARK -S50**	**STARK -ST50**	**STARK -ST101**
AUC(%)	63.8	70.3	73.3	74.0	75.2	75.4	75.7	75.8	**81.2**	80.3	**81.3**	**82.0**
P_{norm}(%)	73.3	77.1	80.0	80.1	81.7	80.0	82.2	81.6	**85.4**	85.1	**86.1**	**86.9**

VOT-2020

	IVT [49]	KCF [19]	SiamFC [2]	CSR-DCF [39]	ATOM [11]	DiMP [3]	UPDT [4]	DPMT	SuperDiMP [1]	**STARK -S50**	**STARK -ST50**	**STARK -ST101**
EAO(↑)	0.092	0.154	0.179	0.193	0.271	0.274	0.278	0.303	0.305	0.280	0.308	0.303
Accuracy(↑)	0.345	0.407	0.418	0.406	0.462	0.457	0.465	0.492	0.477	0.477	0.478	0.481
Robustness(↑)	0.244	0.432	0.502	0.582	0.734	0.740	0.755	0.745	0.786	0.728	0.799	0.775
	STM [45]	SiamEM	SiamMask [57]	SiamMargin [25]	Ocean [69]	D3S [38]	FastOcean	AlphaRef [25]	OceanPlus [67]	**STARK -S50+AR**	**STARK -ST50+AR**	**STARK -ST101+AR**
EAO(↑)	0.308	0.310	0.321	0.356	0.430	0.439	0.461	0.482	**0.491**	0.462	**0.505**	**0.497**
Accuracy(↑)	0.751	0.520	0.624	0.698	0.693	0.699	0.693	0.754	0.685	**0.761**	**0.759**	**0.763**
Robustness(↑)	0.574	0.743	0.648	0.640	0.754	0.769	**0.803**	0.777	**0.842**	0.749	**0.817**	0.789

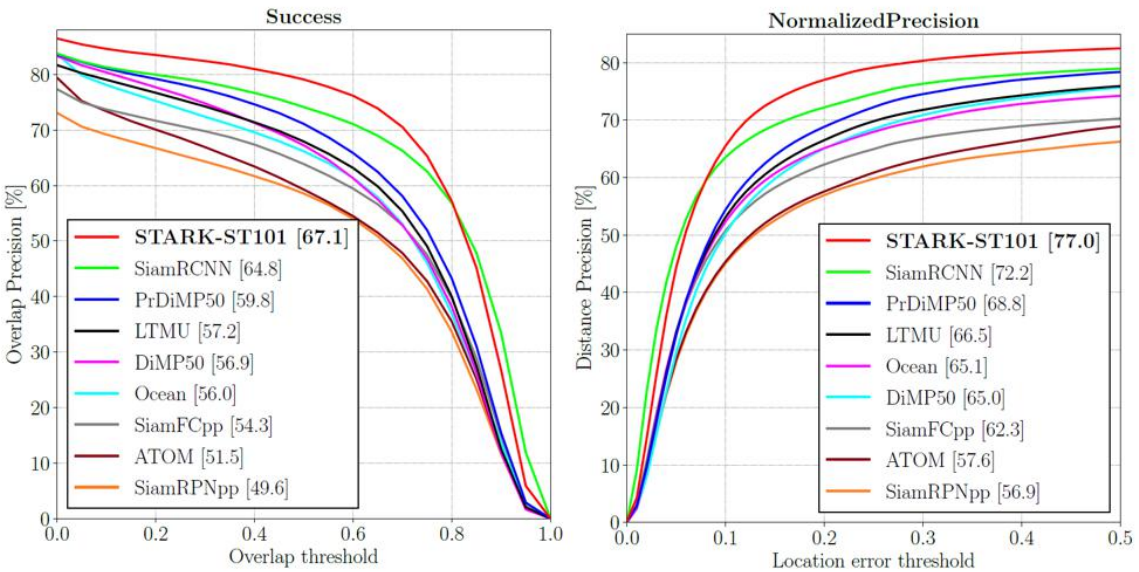
NOTU (NFS, OTB100, TC-128, UAV123)

	SiamFC [2]	RT-MDNet [23]	ECO [10]	Ocean [69]	LightTrack [60]	SiamRPN++ [28]	ATOM [11]	DiMP50 [3]	TransT [6]	**STARK-S50**	**STARK-ST50**	**STARK-ST101**
NOTU	47.2	52.9	56.7	56.7	57.4	59.8	61.5	63.4	**65.0**	64.9	**66.0**	**66.1**
NFS	37.7	43.3	52.2	49.4	49.3	57.1	58.3	61.8	**65.3**	64.3	**65.2**	**66.2**
OTB100	58.3	65.0	66.6	68.4	65.4	**68.7**	66.3	68.4	**69.5**	68.3	**68.5**	68.1
TC128	48.9	56.3	58.9	55.7	55.0	57.7	59.9	**61.2**	59.6	60.0	**62.6**	**63.1**
UAV123	46.8	52.8	53.5	57.4	62.6	59.3	63.2	64.3	68.1	**68.4**	**69.1**	**68.2**

Learning Spatio-Temporal Transformer for Visual Tracking

Experimental Results (Long-Term)

LaSOT



OxUVA

	#	User	Entries	Date of Last Entry	MaxGM ▲
	1	chmayer	4	03/05/21	0.812 (1)
STARK	2	AlphaBin	3	03/12/21	0.782 (2)
	3	ultio791	2	11/14/20	0.763 (3)
	4	MSRA_MSM	1	08/10/19	0.757 (4)
LTMU	5	Daikenan	1	11/07/19	0.751 (5)
	6	pmach	1	03/05/21	0.748 (6)
	7	bossxuan	2	07/15/19	0.741 (7)
Siam R-CNN	8	voigtlaender	3	11/12/19	0.723 (8)
	9	full	1	11/13/19	0.661 (9)
	10	doraiba2008	1	03/10/21	0.633 (10)

VOT2020-LT

	SPLT [62]	ltMDNet	SiamDW_LT [68]	RLT_DiMP	CLGS	Megtrack	LTMU_B [9]	LT_DSE	**STARK-ST50**	**STARK-ST101**
F-score(%)	56.5	57.4	65.6	67.0	67.4	68.7	69.1	**69.5**	**70.2**	**70.1**
Pr(%)	58.7	64.9	67.8	65.7	**73.9**	70.3	70.1	**71.5**	**71.0**	70.2
Re(%)	54.4	51.4	63.5	68.4	61.9	67.1	68.1	**67.7**	**69.5**	**70.1**

Learning Spatio-Temporal Transformer for Visual Tracking

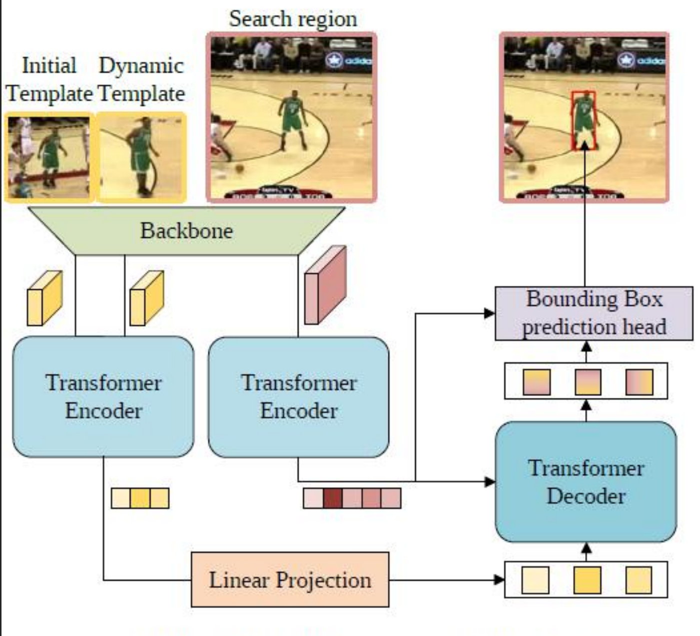
Component-wise Analysis

#	Enc	Dec	Pos	Corner	Score	Success	
1	✗					61.1	-5.3
2		✗				64.5	-1.9
3			✗			66.2	-0.2
4				✗		63.7	-2.7
5					✗	64.5	-1.9
6						66.4	

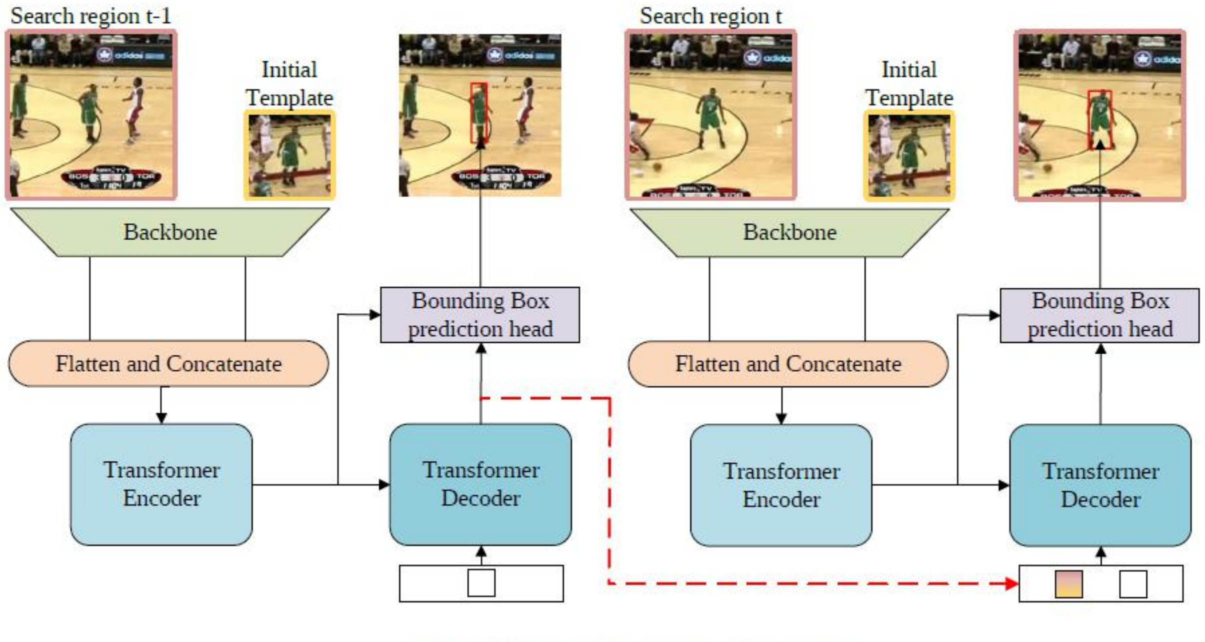
- ✓ **Transformer encoder and the corner prediction head are two most important component in STARK**
- ✓ **Positional encoding is the least important component in STARK**

Learning Spatio-Temporal Transformer for Visual Tracking

Comparison with other frameworks



(a) Taking templates images as the queries



(b) Updating the query embedding

	Template query	Hungarian	Update query	Loc-Cls Joint	Ours
Success	61.2	63.7	64.8	62.5	66.4

Learning Spatio-Temporal Transformer for Visual Tracking

Conclusion

- Tracking as a direct end-to-end bounding box prediction problem
- Dynamic template brings temporal information
- Code and Models: https://github.com/researchmm/Stark

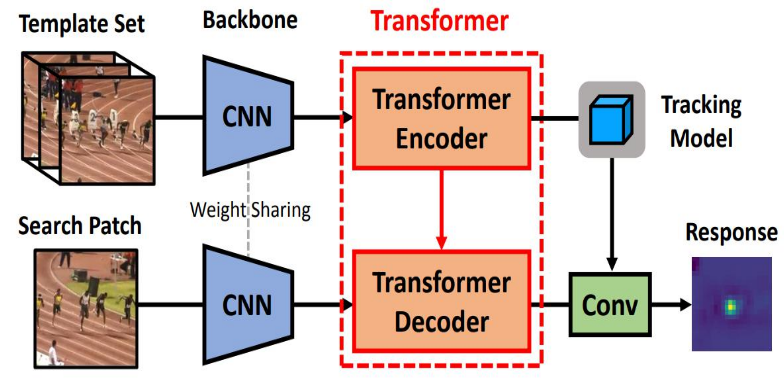
July 24, 2021

- We release an extremely fast version of STARK called **STARK-Lightning** ⚡ . It can run at **200~300 FPS** on a RTX TITAN GPU. Besides, its performance can beat DiMP50, while the model size is even less than that of SiamFC! More details can be found at STARK_Lightning_En.md/中文教程

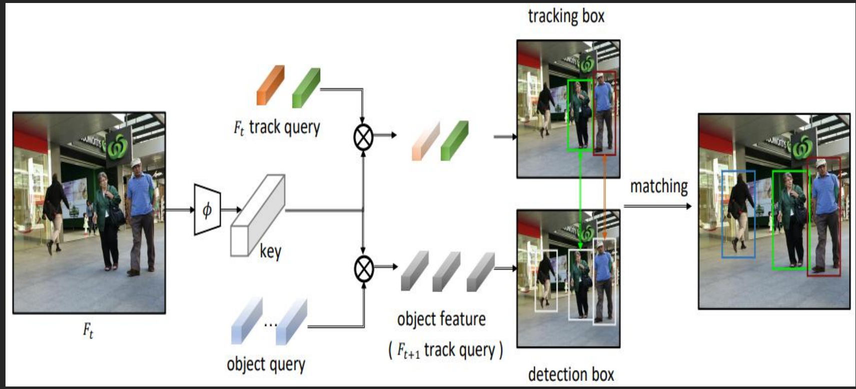
STARK

Transformer in Tracking

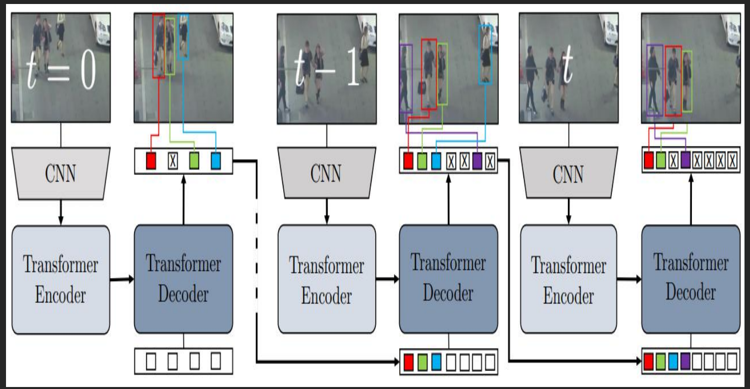
TMT (CVPR21)



TransTrack (CVPR21)



TrackFormer(CVPR21)



Thanks!

TransT

STARK

VOT SOTA