# Learning Spatio-Temporal Transformer for Visual Tracking

**Bin Yan**, Houwen Peng, Jianlong Fu, Dong Wang, Huchuan Lu

**ICCV2021**

Dalian University of Technology, Microsoft Research Asia

# Outline

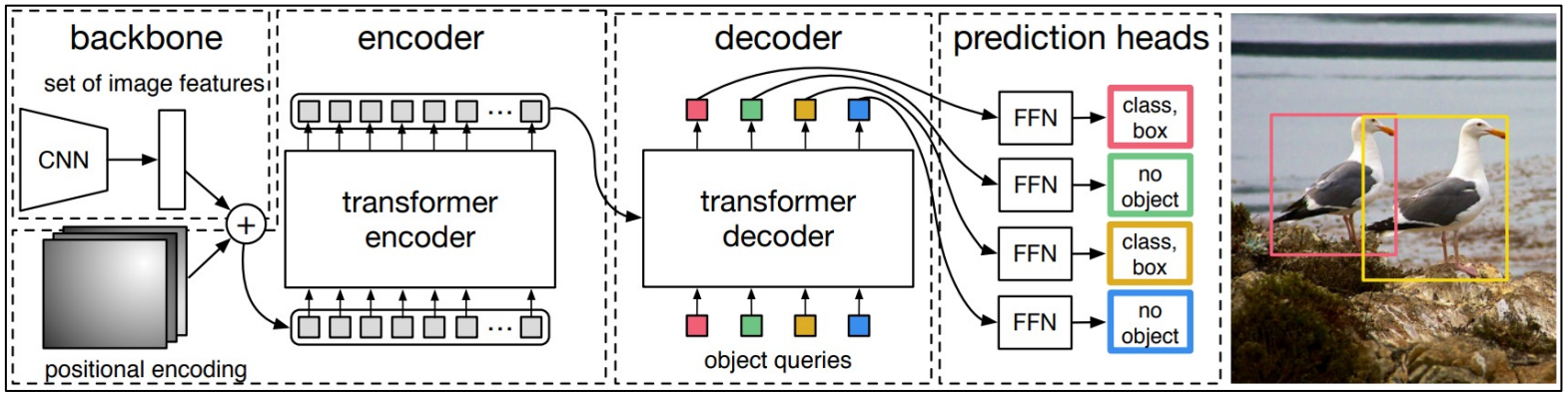
- Basic knowledge about STARK (ICCV2021)
- Better and Stronger: Winner of VOT2021 RGBD Challenge
- Towards Ultra-high Speed: STARK-Lightning

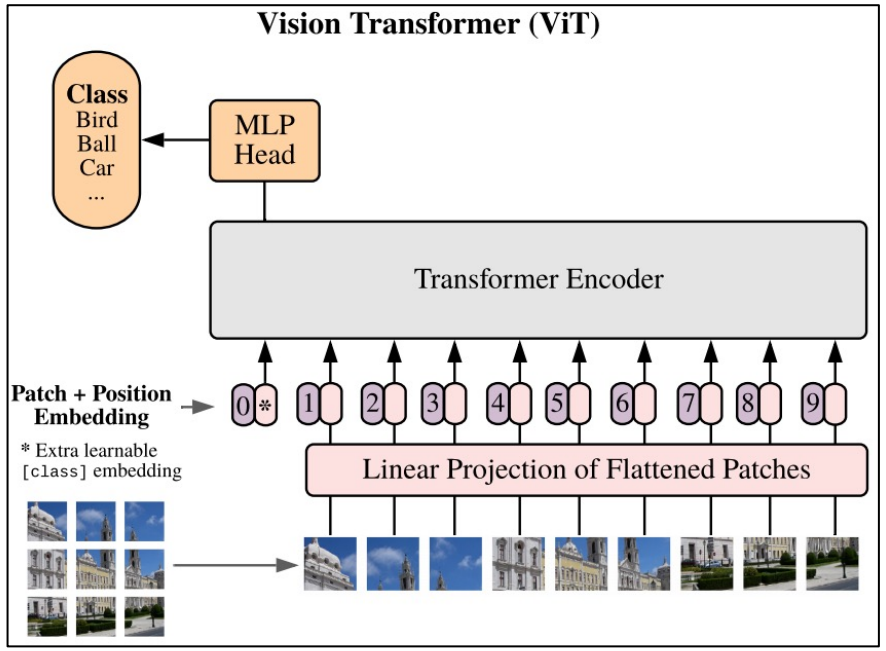
# Part1: Basic knowledge about STARK (ICCV2021)

- Background
- Insights behind our design
- A simple baseline based on Transformer
- Delving deep to the simple baseline
- Spatio-Temporal Transformer Tracking
- Implementation Details
- Experiments
- Component-wise Analysis

# Background

- Transformer has been successfully applied in many basic vision tasks
  - Facebook: Detection Transformer
  - Google: ViT for Image Recognition




## What could Transformer bring to visual tracking?

- End-to-End Object Detection with Transformers (https://arxiv.org/abs/2005.12872)
- AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (https://arxiv.org/abs/2010.11929)

# Insights behind our design

- Offline Siamese Tracking can be summarized as "**Localizing tracked targets by integrating information from the template and the search region**"
- This leads to two key problems: (1) **How to integrate information?** (2) **How to localize the target?**

## Information Interaction

Existing strategies:

- Naïve Correlation (SiamFC, SiamRPN)
- Depth-wise Correlation (SiamRPN++, Ocean)
- Point-wise Correlation (Alpha-Refine, LightTrack)

Disadvantages of existing strategies:

- Local comparison, without the global view
- Limited effective receptive fields (ERF)

These are exactly what Transformer is good at!

- Global information interaction
- Global receptive fields

## Target Localization

Existing strategies:

- Multi-scale (SiamFC)
- RPN (SiamRPN++)
- Anchor-free (SiamFC++, Ocean)

Disadvantages:

- Sensitive to hyper-parameters picking / Highly relying on hyper-parameter tuning
- Complex and redundant

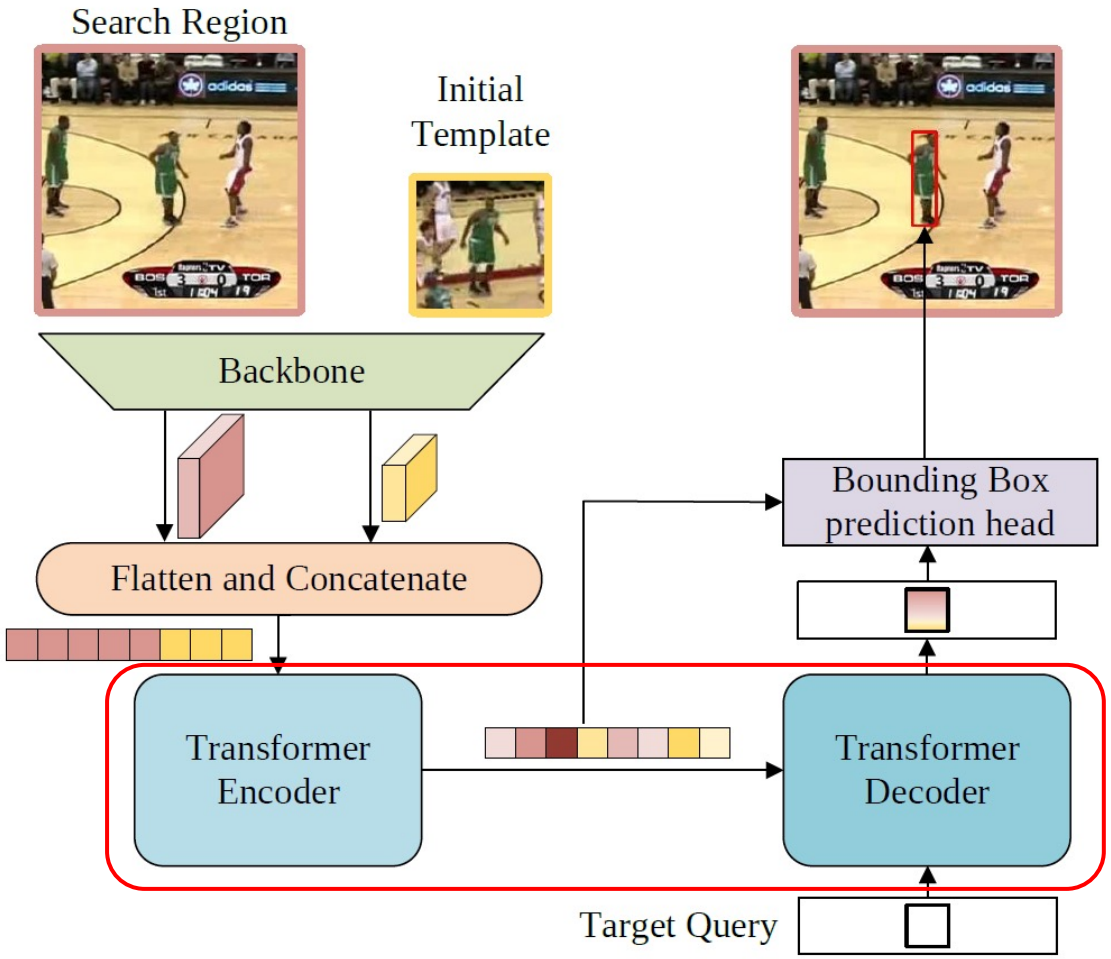
What if we directly predict one bounding box?

- Corner prediction techniques might be useful

# A Simple Baseline Based on Transformer

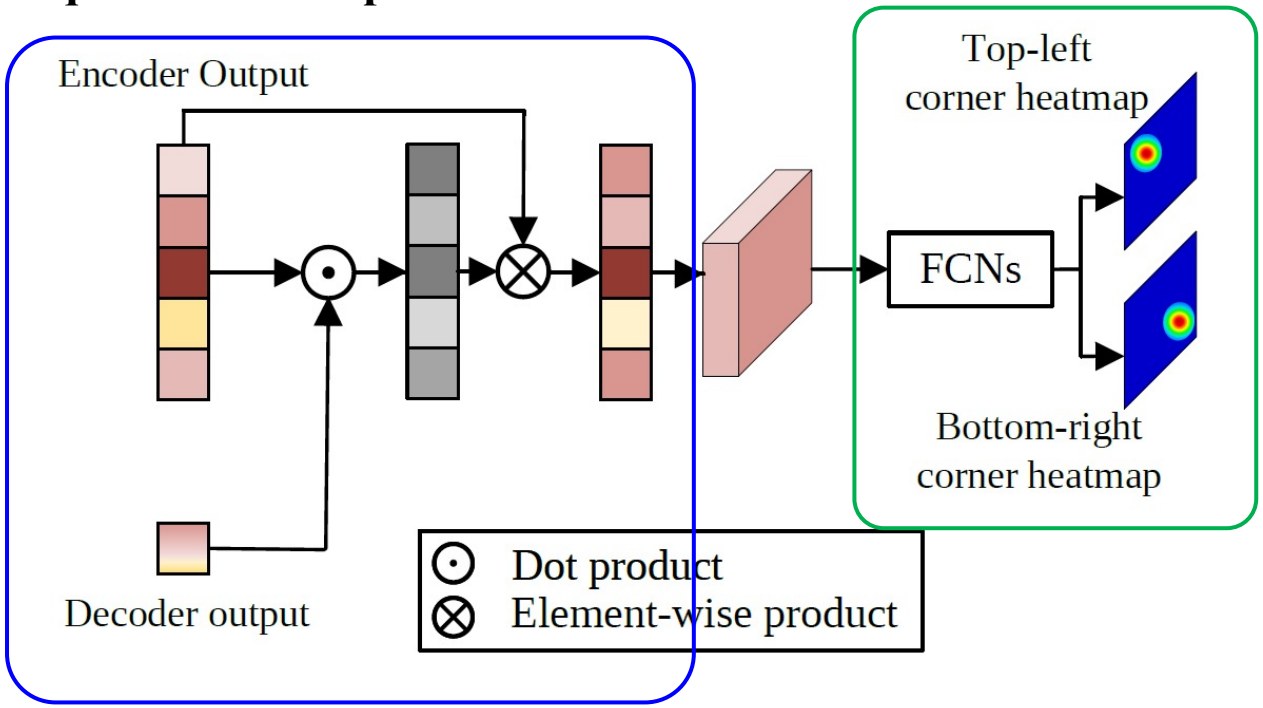
$$(\widehat{x_{tl}},\ \widehat{y_{tl}}) = (\sum_{y=0}^{H}\sum_{x=0}^{W} x \cdot P_{tl}(x,y),\ \sum_{y=0}^{H}\sum_{x=0}^{W} y \cdot P_{tl}(x,y)),$$

$$(\widehat{x_{br}},\ \widehat{y_{br}}) = (\sum_{y=0}^{H}\sum_{x=0}^{W} x \cdot P_{br}(x,y),\ \sum_{y=0}^{H}\sum_{x=0}^{W} y \cdot P_{br}(x,y)),$$




**Transformer architecture for feature integration**

**Equivalent to a spatial-attention**

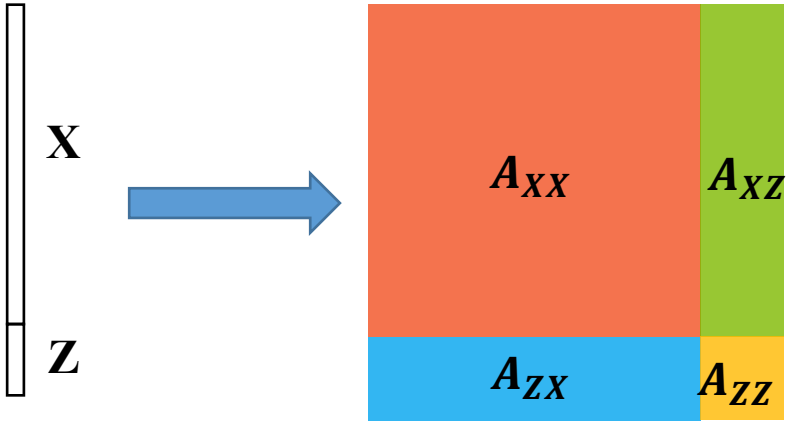



- **Tracking as a direct end-to-end bounding box prediction problem**
- **Totally post-processing free**

# Delving deep to the simple baseline

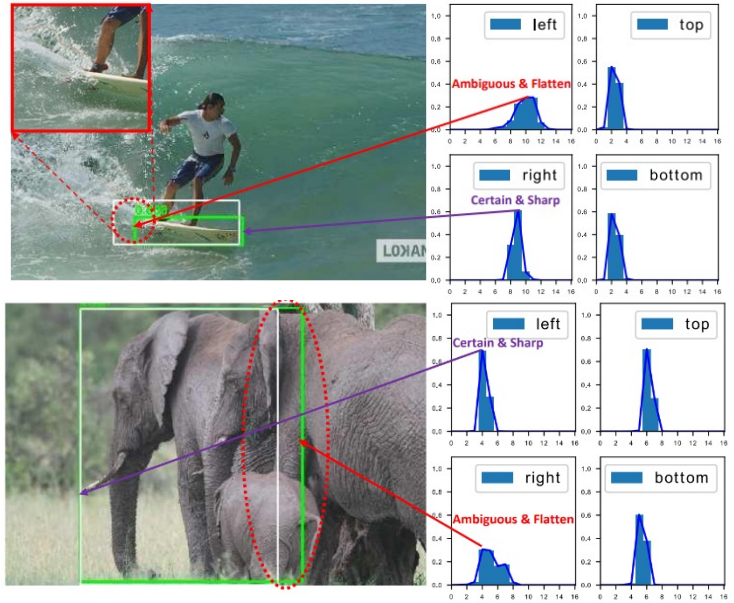
- **Insights behind the "concatenation" operation**

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




- 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?**
  - 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)



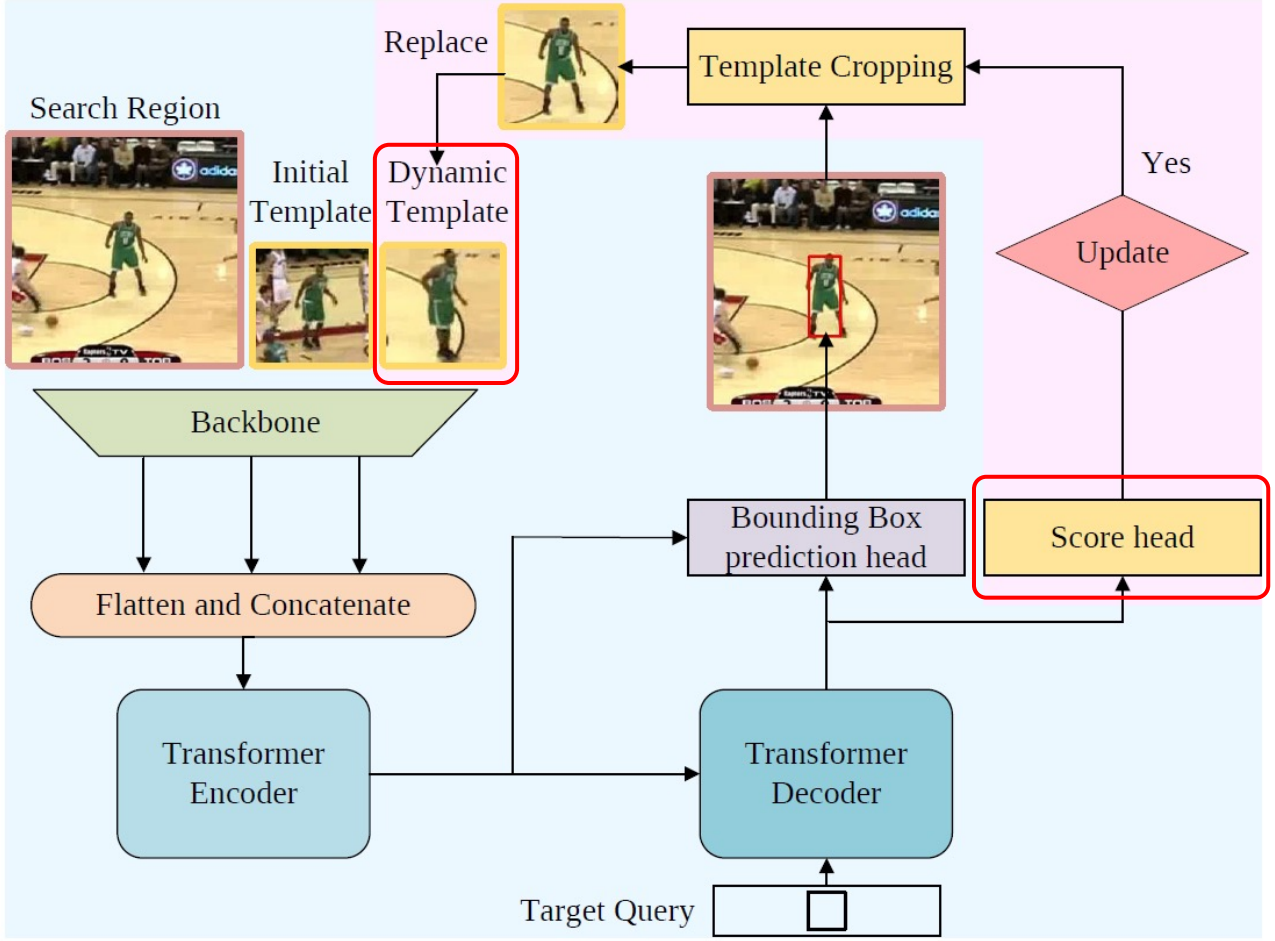

- **Difference between STARK and DETR**

| | Task | Inputs | Query | Head |
|---|---|---|---|---|
| **DETR** | Detection | Image Features | 100, Hungarian | MLP (coord) |
| **STARK** | Tracking | X & Z | 1, fixed | FCN (heatmap) |

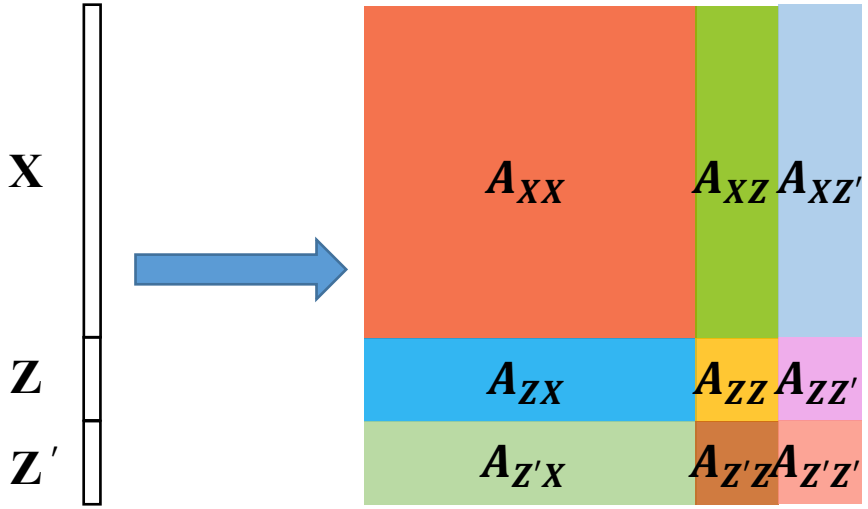
Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection (https://arxiv.org/pdf/2006.04388.pdf)

# Spatio-Temporal Transformer Tracking

Dynamic template changes over time, bringing temporal information for the STARK tracker




STARK-ST takes the current search region, the initial template, and a dynamic template as inputs, **jointly learning discriminative Spatio-Temporal representation**




Considering challenging bases (Occlusion, Disappearance…) during tracking, we introduce an **update controller** to control the update of the dynamic template

# Implementation Details

- Model
  - Inputs: Template (128x128), Search Region (320x320) (5x the target scale)
  - Backbone: ResNet-50, ResNet-101 (stride=16)
  - Transformer: 6 encoder layers, 6 decoder layers
  - Head: stacked Conv-BN-ReLU layers
- Training
  - Training set: LaSOT, GOT-10K, TrackingNet, COCO
  - 6w triplets per epoch, 500 epochs in total. Batchsize 16x8
  - Adam-W optimizer with base learning rate 1e-4 (backbone 1e-5)
  - Two-stage training for STARK-ST
- Inference
  - The inference pipeline is only one forward pass and a linear coordinate transform
  - The default update interval is 200 frames
  - The default update threshold is 0.5

# Experiments

## 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

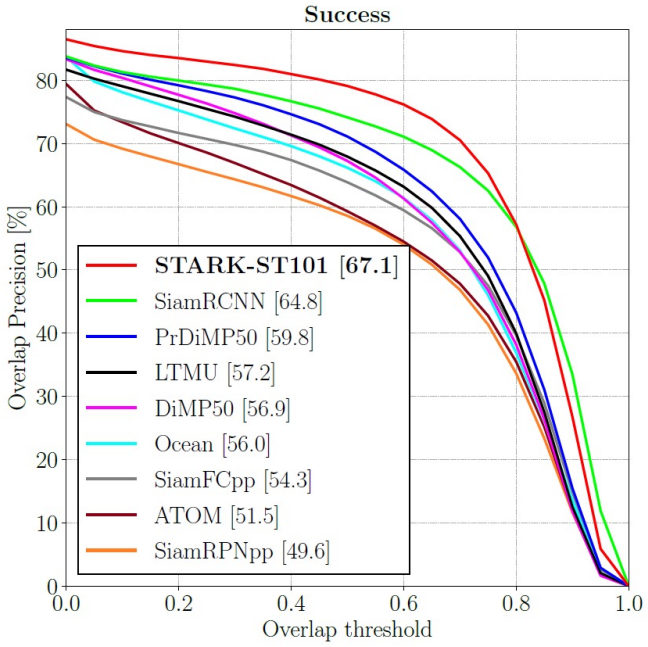
| | IVT [49] | KCF [19] | SiamFC [2] | CSR-DCF [39] | ATOM [11] | DiMP [3] | UPDT [4] | DPMT | SuperDiMP [1] | **STARK -S50** | **STARK -ST50** | **STARK -ST101** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EAO(↑) | 0.092 | 0.154 | 0.179 | 0.193 | 0.271 | 0.274 | 0.278 | 0.303 | 0.305 | 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 |

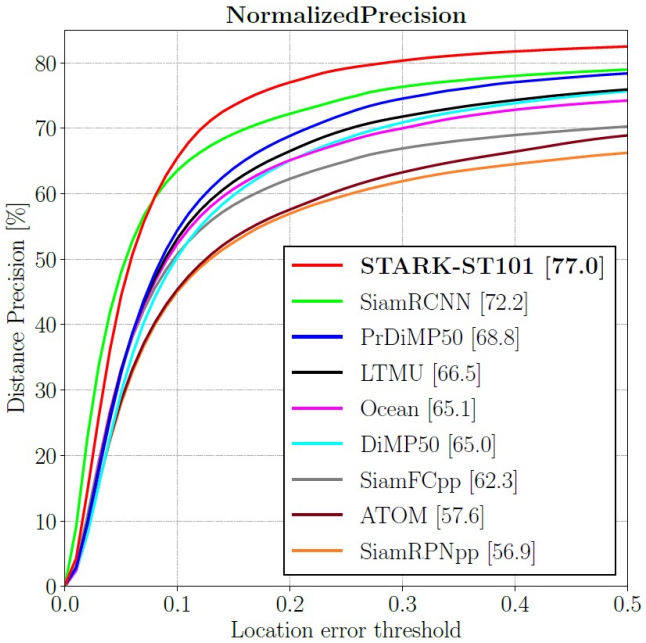
## NOTU (NFS, OTB100, TC-128, UAV123)

| | 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** |

# Experiments

## 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** |

# 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

# Part2: Better and Stronger: Winner of VOT2021 RGBD Challenge

- Brief Review of AlphaRefine
- Exploitation on STARK
- Failure Cases Analysis
- Ensemble DiMP with STARK
- Summary

# STARK in VOT Challenge 2021

## Real-Time Challenge

**Rank 2**

| Tracker | EAO | realtime A | realtime R |
|---|---|---|---|
| RPTMask | 0.368 | 0.659 | 0.704 |
| CFRPT | 0.325 | 0.612 | 0.684 |
| TransT_M | 0.550① | 0.742 | 0.869① |
| TregPlus | 0.440 | 0.706 | 0.777 |
| DualTFRon | 0.395 | 0.681 | 0.741 |
| F_TregPlus | 0.490 | 0.738 | 0.812 |
| DualTFRst | 0.512③ | 0.751② | 0.816 |
| STARK_RT | 0.531② | 0.780① | 0.829③ |
| DualTFR | 0.509 | 0.746 | 0.813 |
| RPT | 0.346 | 0.598 | 0.721 |
| SiamUSCP | 0.469 | 0.679 | 0.821 |
| D3Sv2 | 0.313 | 0.654 | 0.614 |
| LWL_B2S | 0.475 | 0.719 | 0.806 |
| TransT | 0.507 | 0.748③ | 0.817 |

## RGBD Challenge

**Rank 1**

| Tracker | Pr | Re | F-Score | Year |
|---|---|---|---|---|
| STARK_RGBD | 0.742① | 0.769① | 0.756① | 2021 |
| TALGD | 0.728② | 0.717② | 0.722② | 2021 |
| DRefine | 0.707 | 0.715③ | 0.711③ | 2021 |
| ATCAIS | 0.709 | 0.696 | 0.702 | 2020 |
| DDiMP | 0.703 | 0.689 | 0.696 | 2020 |
| CLGS_D | 0.725③ | 0.664 | 0.693 | 2020 |
| SLMD | 0.701 | 0.685 | 0.693 | 2021 |
| sttc_rgbd | 0.692 | 0.685 | 0.689 | 2021 |

### B.1. Fusing Complementary Trackers for Long-term Visual Tracking (mlpLT)

*M. Dunnhofer, K. Simonato, C. Micheloni*
*{matteo.dunnhofer, christian.micheloni}@uniud.it,*
*simonato.kristian@spes.uniud.it*

The idea behind the mlpLT tracker is to fuse the capabilities of different trackers. In particular, mlpLT implements a strategy that fuses the Stark [66] and SuperDiMP [2] trackers, which have been selected due to their complementary features. Indeed, the tracker Stark has been selected because of its ability to provide spatially accurate bounding-boxes, and to re-detect the target after disappearances. The meta-updater [14] controlled SuperDiMP was chosen due

## Long-Term Challenge

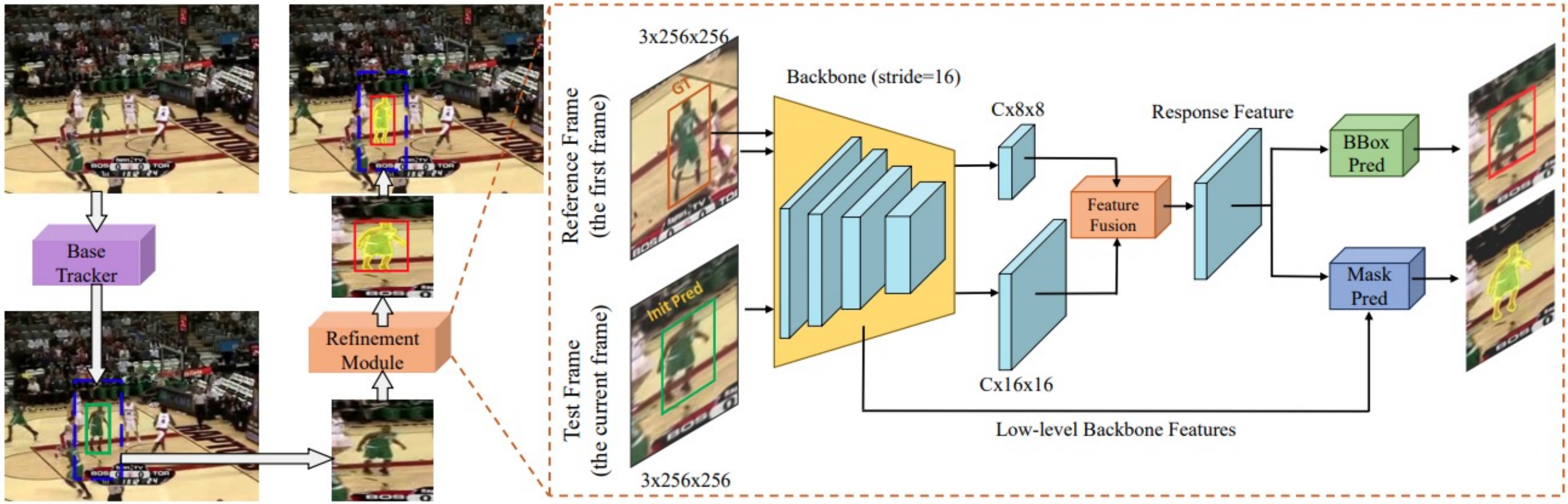
**Rank 2, 3**

| Tracker | Pr | Re | F-Score | Year |
|---|---|---|---|---|
| mlpLT | 0.741① | 0.729① | 0.735① | 2021 |
| STARK_LT | 0.721 | 0.725② | 0.723② | 2021 |
| STARK_RGBD_LT | 0.719 | 0.724③ | 0.721③ | 2021 |
| SLOT | 0.727③ | 0.711 | 0.719 | 2021 |
| keep_track_lt | 0.725 | 0.700 | 0.712 | 2021 |
| SuperD_MU | 0.738② | 0.680 | 0.708 | 2021 |
| RincTrack | 0.717 | 0.696 | 0.707 | 2021 |
| LT_DSE | 0.715 | 0.677 | 0.695 | 2020 |
| LTMU_B | 0.698 | 0.680 | 0.689 | 2020 |
| SuperDiMP | 0.675 | 0.660 | 0.667 | 2021 |
| SiamRCNN | 0.654 | 0.673 | 0.664 | 2021 |
| SION_LT | 0.640 | 0.456 | 0.533 | 2021 |
| TDIOT | 0.496 | 0.478 | 0.487 | 2021 |

Winner of the Long-Term Challenge also adopts STARK as its core component.

The Ninth Visual Object Tracking VOT2021 Challenge Results (ICCV2021 workshop)
https://openaccess.thecvf.com/content/ICCV2021W/VOT/papers/Kristan_The_Ninth_Visual_Object_Tracking_VOT2021_Challenge_Results_ICCVW_2021_paper.pdf

# Brief Review of Alpha-Refine




- Alpha-Refine boosts the quality of box estimation within a smaller search region, which is generated by expanding the base tracker's output.
- As Alpha-Refine can be viewed as a tracker running on a smaller search region, we can adapt STARK to a more powerful refinement module by setting a smaller search region for it.

"Alpha-Refine": Boosting Tracking Performance by Precise Bounding Box Estimation (**CVPR 2021**)
https://openaccess.thecvf.com/content/CVPR2021/papers/Yan_Alpha-Refine_Boosting_Tracking_Performance_by_Precise_Bounding_Box_Estimation_CVPR_2021_paper.pdf
Code: https://github.com/MasterBin-IIAU/AlphaRefine

# Explorations on STARK

**Improvement1: Changing the backbone**

| Methods | LaSOT (AUC) | VOT-RGBD (F) |
|---|---|---|
| STARK-ST (ResNet-50) | 66.4 | 68.1 |
| STARK-ST (DeiT-base 1000 ep) | **67.8** | **71.8** |

← Winner of VOT2020-RGBD (0.702)

**Improvement2: Using Alpha-Refine to get more accurate boxes**

| Methods | VOT-RGBD (F) |
|---|---|
| Stark-DeiT | 71.8 |
| Stark-DeiT+AR | 72.8 |
| Stark-DeiT+StarkAR | **73.4** |

← Using the standard Alpha-Refine (CVPR2021 version)

← Training another STARK-ST50 for accurate box estimation

Training data-efficient image transformers & distillation through attention (https://arxiv.org/abs/2012.12877)

# Failure Case Analysis

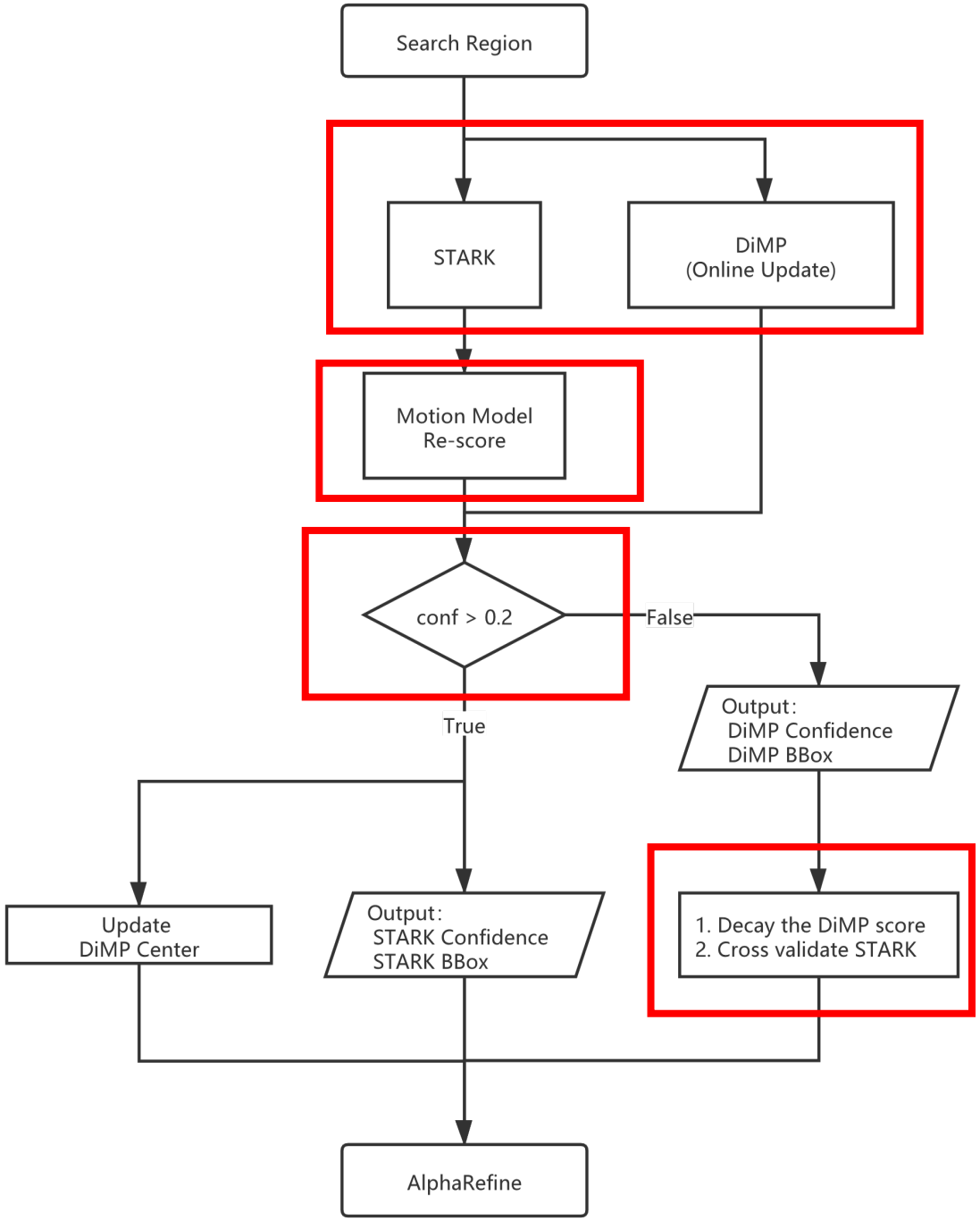
The result predicted by STARK is not robust to appearance change

Using a powerful online method DiMPsuper to strengthen STARK

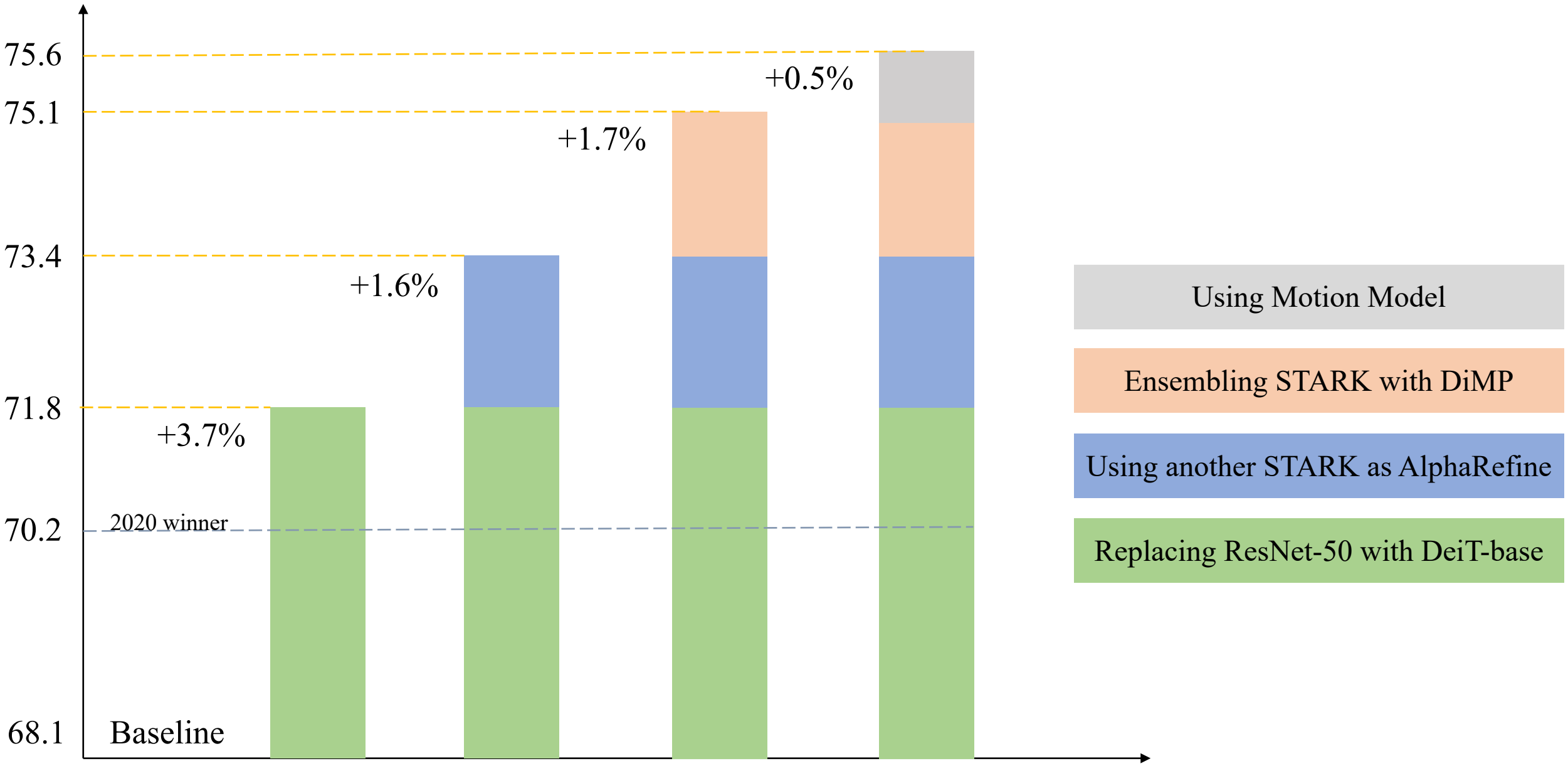
- We observed that the trackers with an online updating strategy are complementary to STARK
- It intuitively makes sense to combine STARK with a powerful online tracker
- We adopt the powerful DiMPsuper

# Ensemble DiMP with STARK

- *When the confidence of STARK is above the threshold*
  - Take the box predicted by STARK as the tracking result
  - Take the confidence of STARK as the confidence
  - DiMPsuper update
    - Perform tracking including online update
    - Follow the center position of STARK
    - Reset DiMP confidence Upper Boundary
- *When the confidence of STARK falls below the threshold*
  - Take the box predicted by DiMPsuper as the tracking result
  - Take the confidence of DiMP as the conficence
    - In case the DiMP focuses on distractor when the target vanishes, decay the upper boundary of DiMP score
    - If both of DiMP and STARK has low confidence, but the IOU of their prediction is high, boost the confidence
- Online update supervised by DiMPsuper



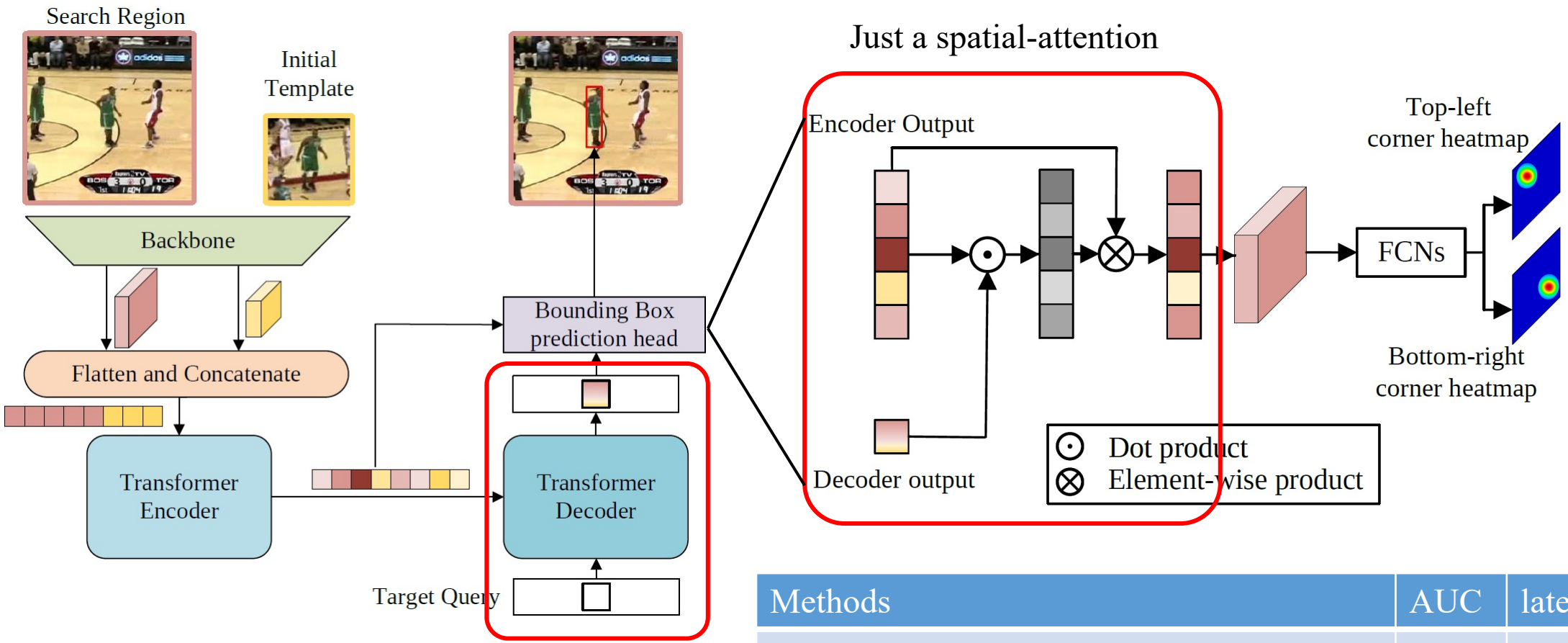
# Summary


75.6
75.1
73.4
71.8
70.2
68.1
+0.5%
+1.7%
+1.6%
+3.7%
2020 winner
Baseline
Using Motion Model
Ensembling STARK with DiMP
Using another STARK as AlphaRefine
Replacing ResNet-50 with DeiT-base

# Part3: Towards Ultra-high Speed: STARK-Lightning

- Revisiting STARK-S
- STARK-Lightning
- Experiments

# Revisiting STARK-S (STARK-Lite)

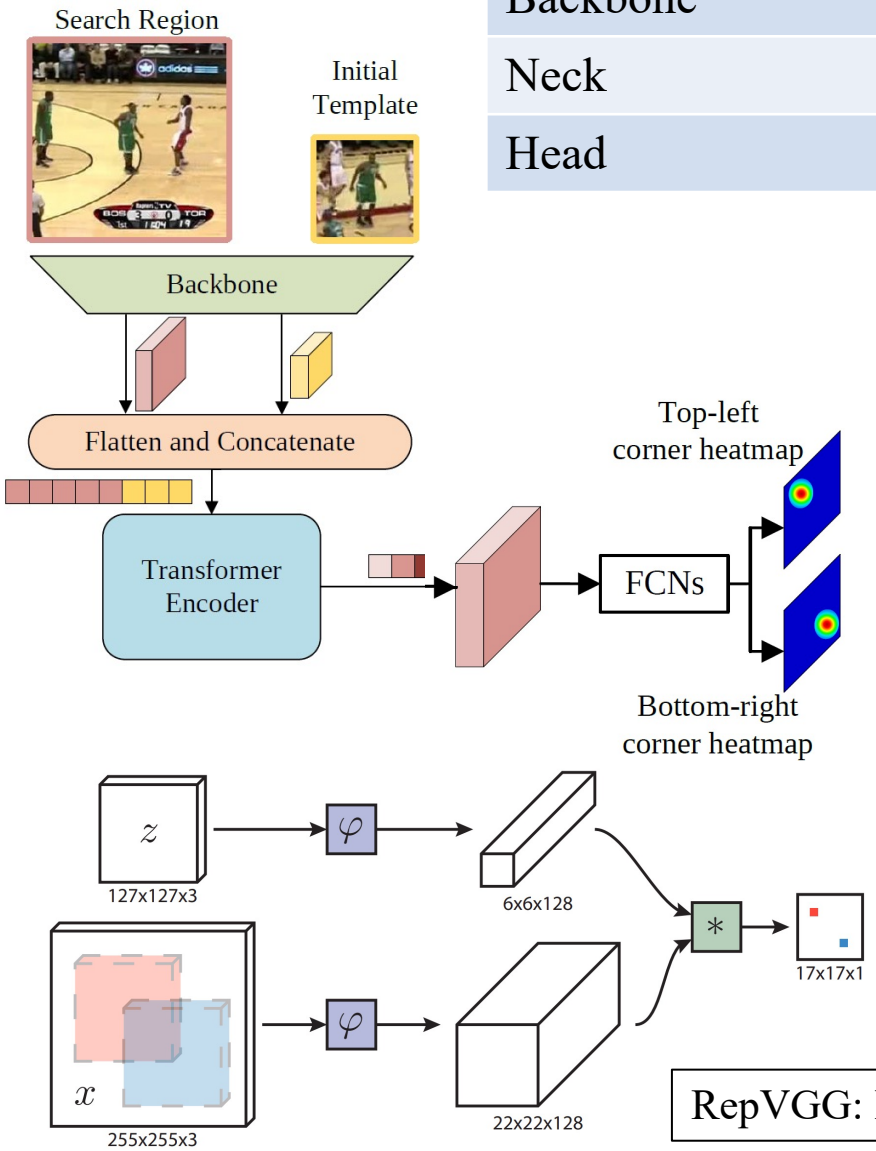


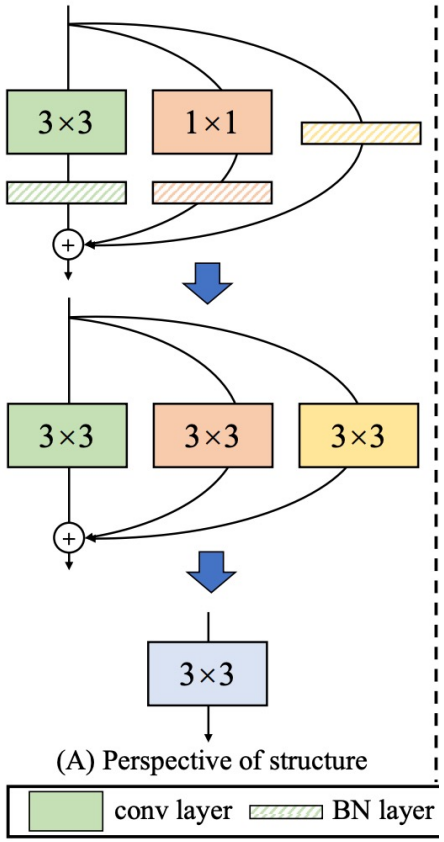

However, to learn this simple spatial-attention, we use 6 decoder layers, which is unnecessary.

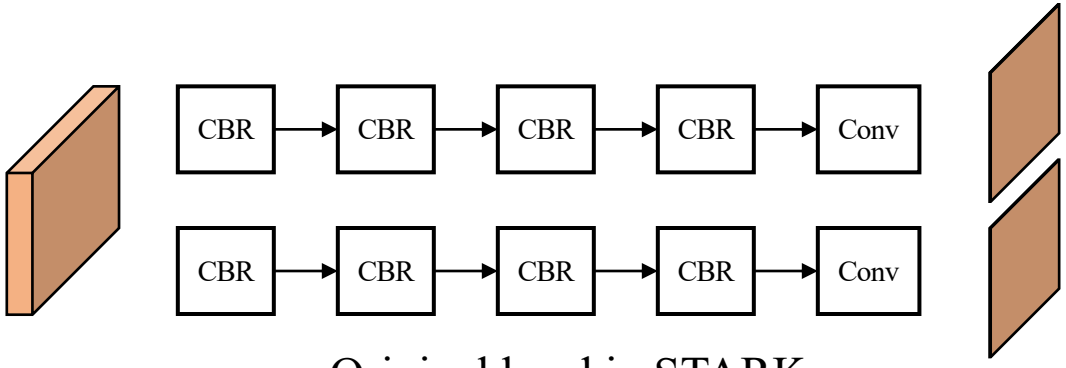
| Methods | AUC | latency | Params |
|---|---|---|---|
| STARK-S | 65.8 | 35ms | 23M |
| Removing decoder | 65.6 | 21ms | 17M |

# STARK-Lightning

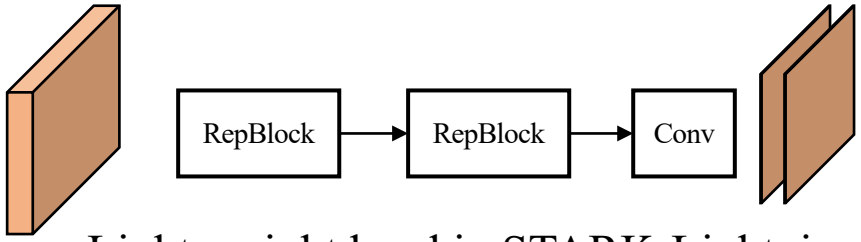
| | STARK-Lite | STARK-Lightning |
|---|---|---|
| Backbone | ResNet-50 | **RepVGG-A0 (4 layers in the last stage)** |
| Neck | 6 layer encoder | **1 layer encoder** |
| Head | 2 separate conv tower | **1 conv tower (width/2, depth/2)** |




Original head in STARK




Light-weight head in STARK-Lightning

RepVGG: Making VGG-style ConvNets Great Again (CVPR 2021) (https://arxiv.org/abs/2101.03697)

# Experiments

Note: Latency is tested on RTX TITAN GPU

| Methods | AUC | Size (Params) | Lat(torch) | Lat (onnx) |
|---|---|---|---|---|
| STARK (RepVGG-14) | 61.8 | 25.1MB | 5.23(ms) | 2.17 |
| STARK (RepVGG-10) | 61.6 | 20.1MB | 4.67(ms) | 2.11 |
| STARK (RepVGG-4) | 60.1 | 12.5MB | 4.11(ms) | 1.65 |
| **STARK (RepVGG-4-lite)** | 58.2 | **8.2MB** | **3.70(ms)** | **1.45/1.31** |

## STARK-Lightning v.s Other Trackers

| Tracker | LaSOT (AUC) | Speed (FPS) | Params (MB) |
|---|---|---|---|
| STARK-Lightning | 58.2 | ~200 | 8.2 |
| DiMP50 | 56.8 | ~50 | 165 |
| DaSiamRPN | 41.5 | ~200 | 362 |
| SiamFC | 33.6 | ~100 | 8.9 |

**Stronger** than DiMP50

As **fast** as DaSiamRPN

More **light-weight** than SiamFC

**Some Useful Links:**

- https://github.com/researchmm/Stark/blob/main/lib/tutorials/STARK_Lightning_Ch.md
- https://github.com/wangdongdut/Online-Visual-Tracking-SOTA

# Thanks for your listening

Hiring research interns for projects on **Object Tracking**, **Neural Architecture Search**, **Vision Transformer**, **Model Compression**, etc. If you have interest and experience on subjects mentioned above, you can contact houwen.peng@microsoft.com