STARK-RGBD

Xingyu Zhang*, Bin Yan*, Lijun Wang, Houwen Peng, Dong Wang,
Huchuan Lu, Xiaoyun Yang

{zhangxy71102, yan_bin}@mail.dlut.edu.cn

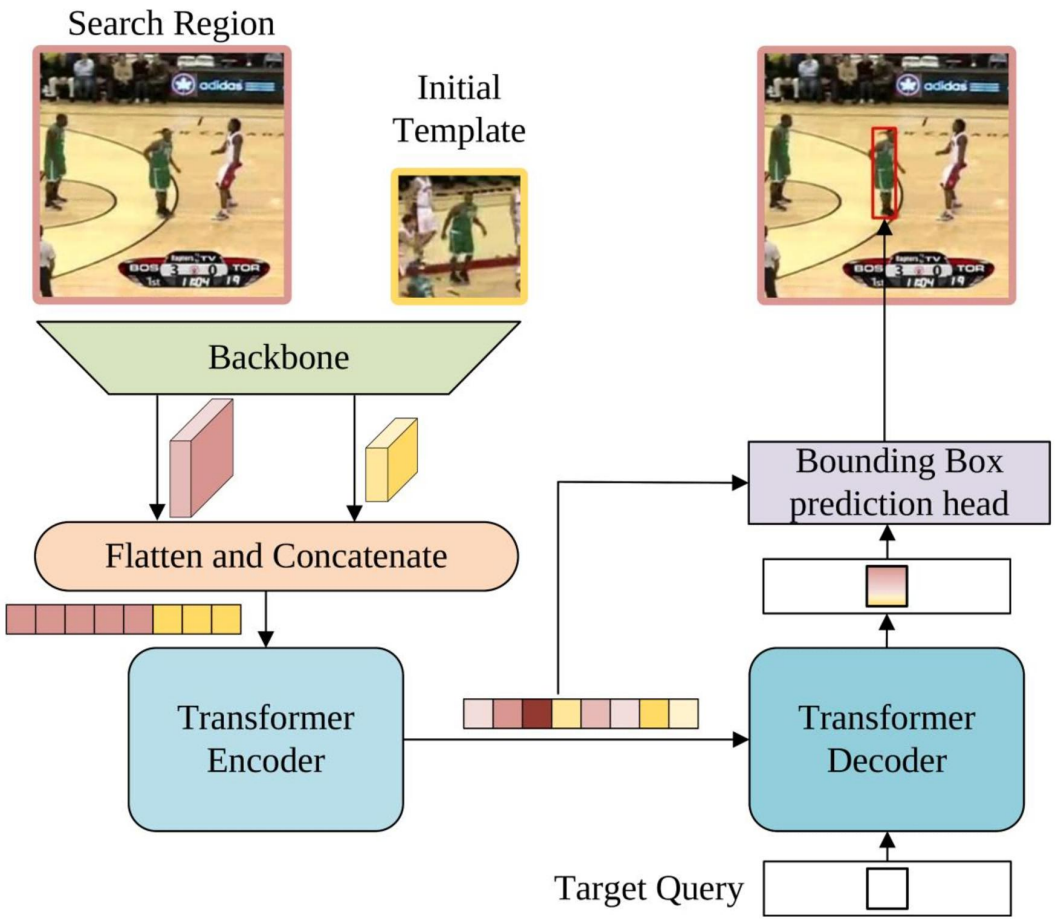
*contribute equally

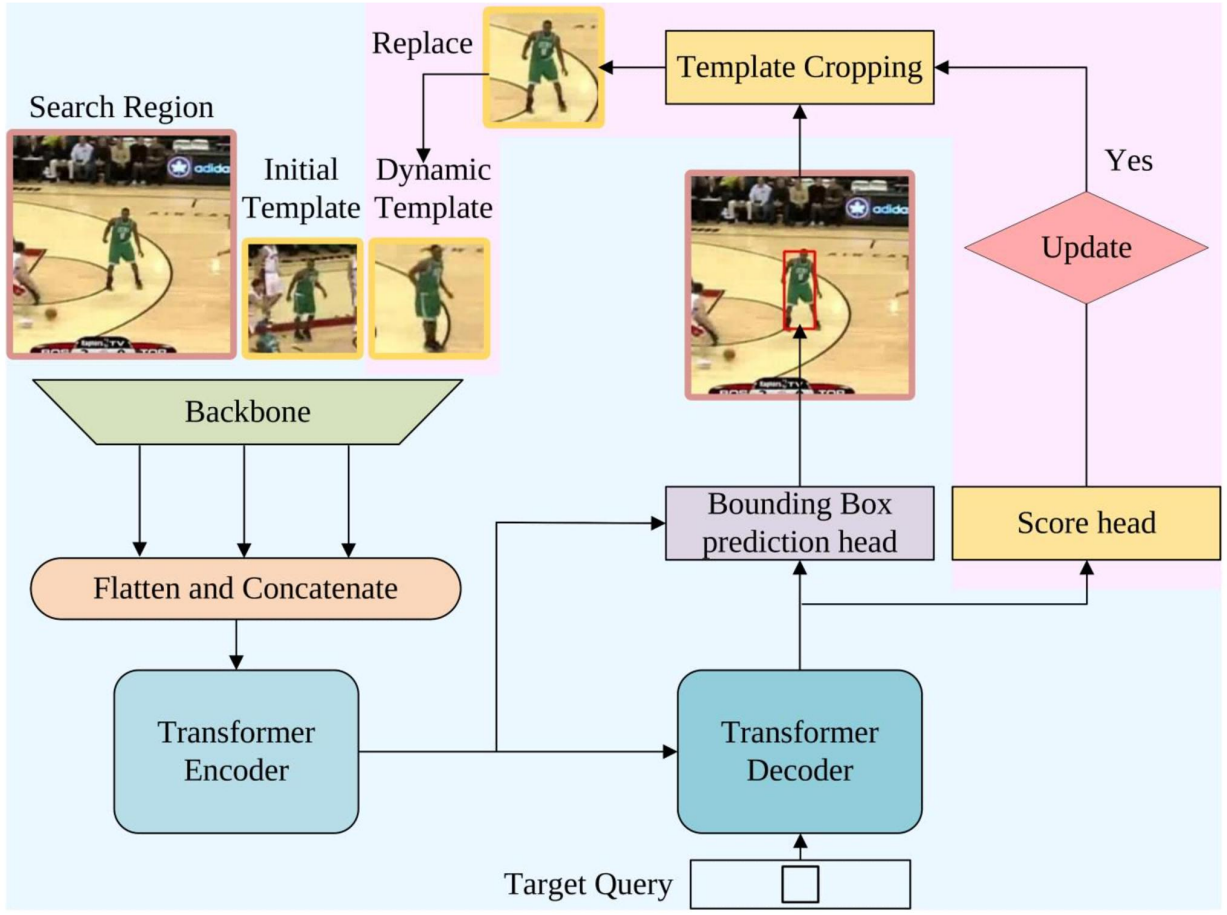
Outline

- A powerful baseline
- Our method
- The usage of the depth information
- Conclusion and inspiration

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

Breif Review of STARK



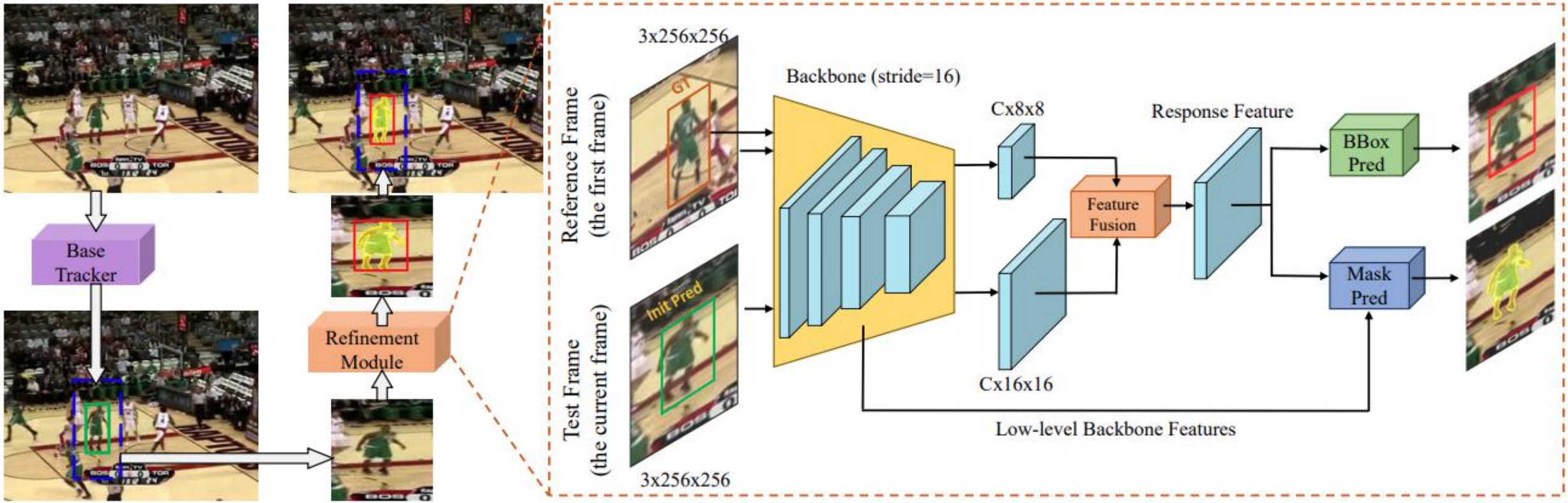


- STARK is a powerful transformer-based tracker
- We replace the original ResNet-50 backbone with DeiT

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

Breif Review of AlphaRefine

Code: https://github.com/MasterBin-IIAU/AlphaRefine



- AlphaRefine boosts the quality of box estimation within a smaller search region, which is generated by expanding the base tracker's output.
- As AlphaRefine 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.

Yan, Bin, Xinyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. "Alpha-Refine: Boosting Tracking Performance by Precise Bounding Box Estimation" CVPR 2021

Explorations on STARK

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)

Using AlphaRefine 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 AlphaRefine (CVPR version)

← Training another STARK-ST50 for accurate box estimation

By strengthen the tracker as a general tracker, we get a baseline with 73.4 F-Score

Failure Case Analysis

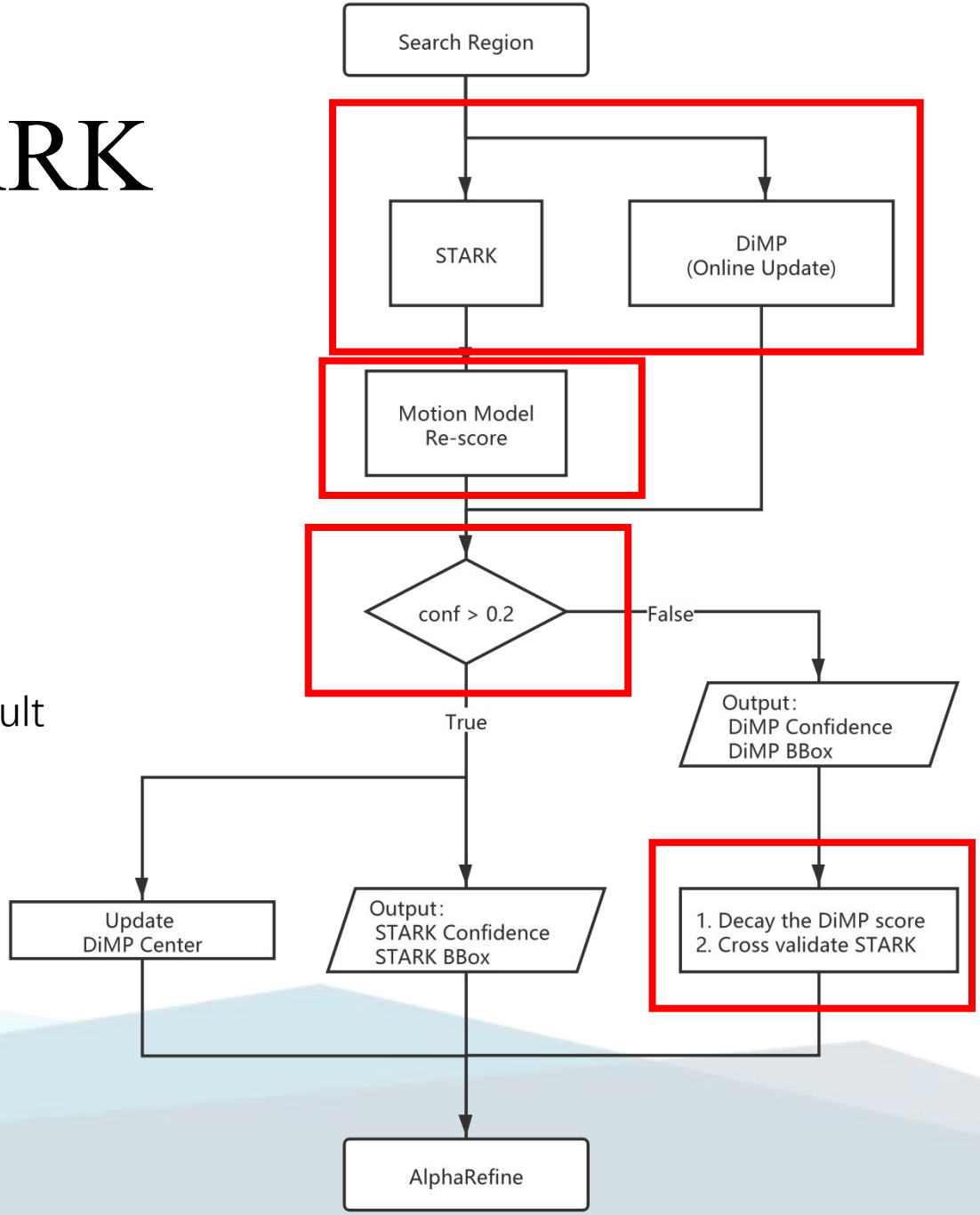
The result predicted by STARK is not robust to appearance change

Using a powerful online method DiMPsuper to strengthen STARK

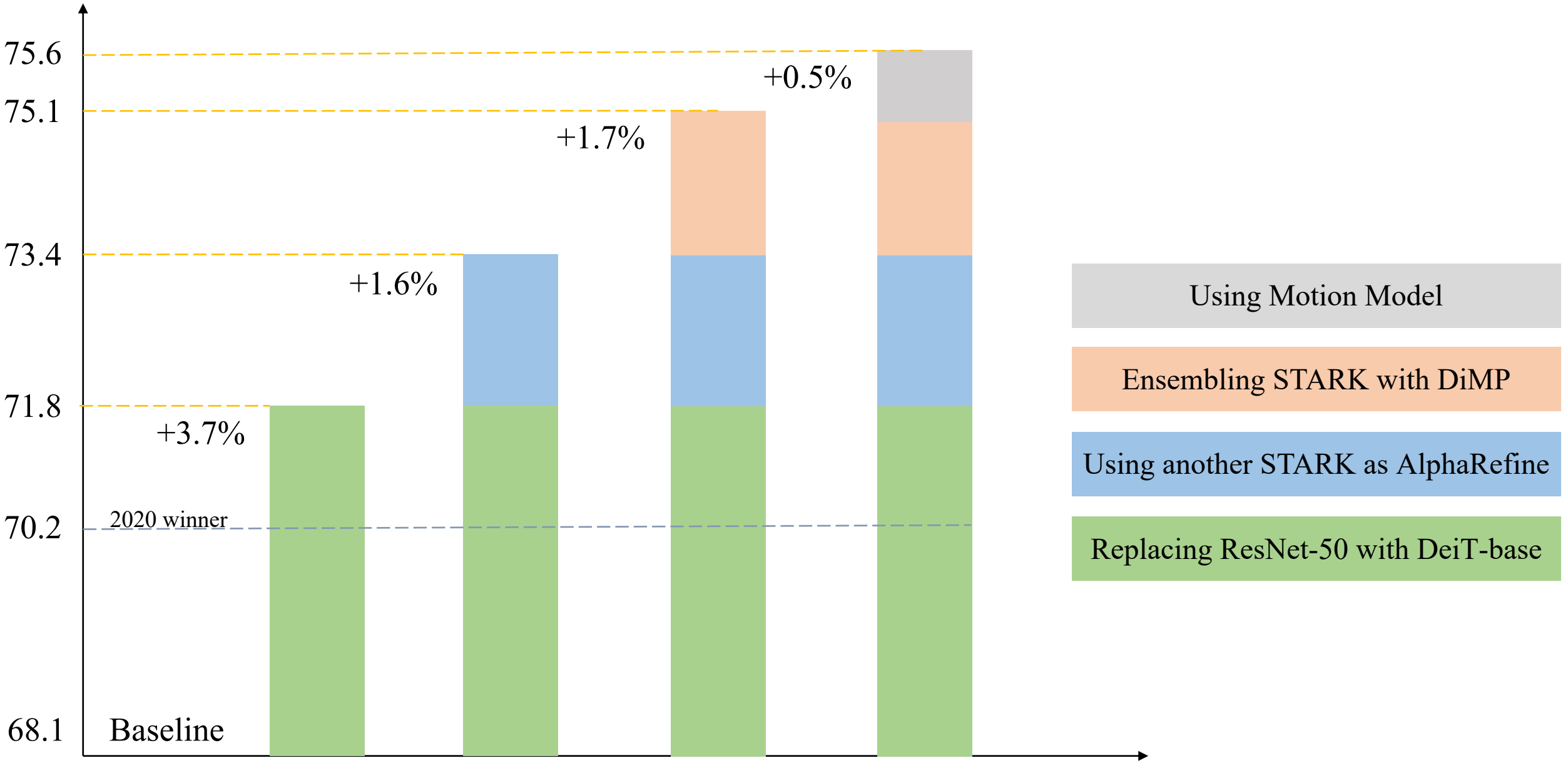
- We observed that the trackers with an online updating strategy are complementary to STARK
- It intuitively makes sense to combine STARK with a powerful online tracker
- We adopt the powerful DiMPsuper

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 fall 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 oftheir 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

About Depth Modelity

Valuable Depth Modality:

- Possibility of modeling the 3D appearance and 3D trajectory of the target
- The depth of the same object should not differ too much so that the depth can help us discriminate the target from the background.
- The depth modality also has the potential of modeling the environment

Difficulties:

- Many depth maps in the dataset only contain few discretized depth values (top row), squeezing points of different depth values into the same plane.
- For those depth maps with good quality (bottom row), the RGB modality is good enough, so that the depth information provides limited help

Observation of Depth

- The depth values of different parts of the same object differ too much, failing to reflect the shape of the object.
- The depth value changes discontinuously, making it difficult to model the 3D trajectory of objects
- There is mis-alignment between the depth modality and RGB modality. We can also observe that the contour of depth is not clean enough so that it is difficult to precisely localize the object by means of depth modality.

Depth Modelity with Better Quality will Help

An RGBD image of very high quality

An RGBD image available in the dataset

Conclusion and Inspiration

- Nowadays, as the RGB trackers grow more and more powerful, we need depth data with high quality to assist it.
- The trackers with/without online updating are complementary. Ensembling different types of trackers will boost the performance.
- How to ensemble them efficiently remains to be explored.

Thanks for Listening!

STARK: https://github.com/researchmm/Stark
AlphaRefine: https://github.com/MasterBin-IIAU/AlphaRefine
Online-Visual-Tracking-SOTA: https://github.com/wangdongdut/Online-Visual-Tracking-SOTA