DeTrack: In-model Latent Denoising Learning for Visual Object Tracking

Xinyu Zhou[1] **Jinglun Li**[2] **Lingyi Hong**[1] **Kaixun Jiang**[2] **Pinxue Guo**[2]

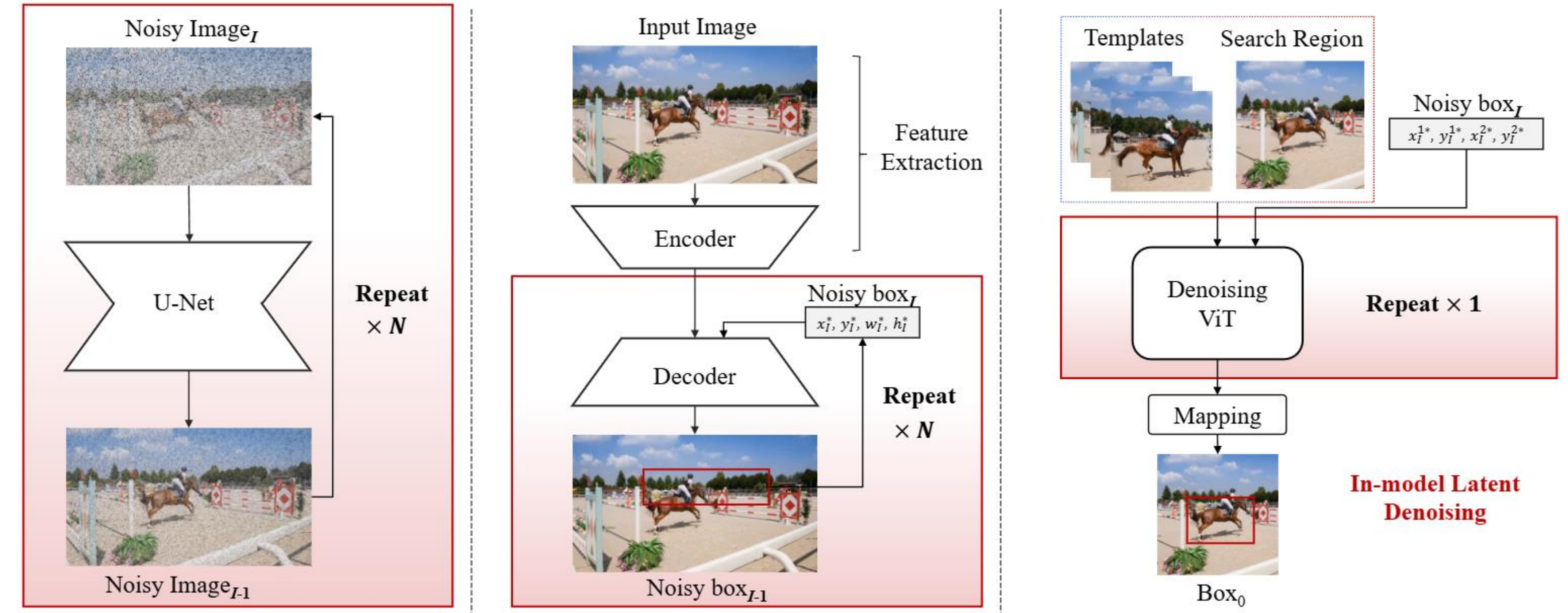
Weifeng Ge[1*] **Wenqiang Zhang**[1,2*]

[1]Shanghai Key Lab of Intelligent Information Processing,
School of Computer Science, Fudan University, Shanghai, China
[2]Shanghai Engineering Research Center of AI & Robotics,
Academy for Engineering and Technology, Fudan University, Shanghai, China
zhouxinyu20@fudan.edu.cn, jingli960423@gmail.com, lyhong22@m.fudan.edu.cn,
kxjiang22@m.fudan.edu.cn, pxguo21@m.fudan.edu.cn,
weifeng.ge.ic@gmail.com,wqzhang@fudan.edu.cn

DeTrack: In-model Latent Denoising Learning forVisual Object Tracking

Xinyu Zhou[1], Jinglun Li[2], Lingyi Hong[1], Kaixun Jiang[2], Pinxue Guo[2], Weifeng Ge[1] and Wenqiang Zhang[1,2]
[1]School of Computer Science, Fudan University, [2]Academy for Engineering and Technology, Fudan University,

Motivation



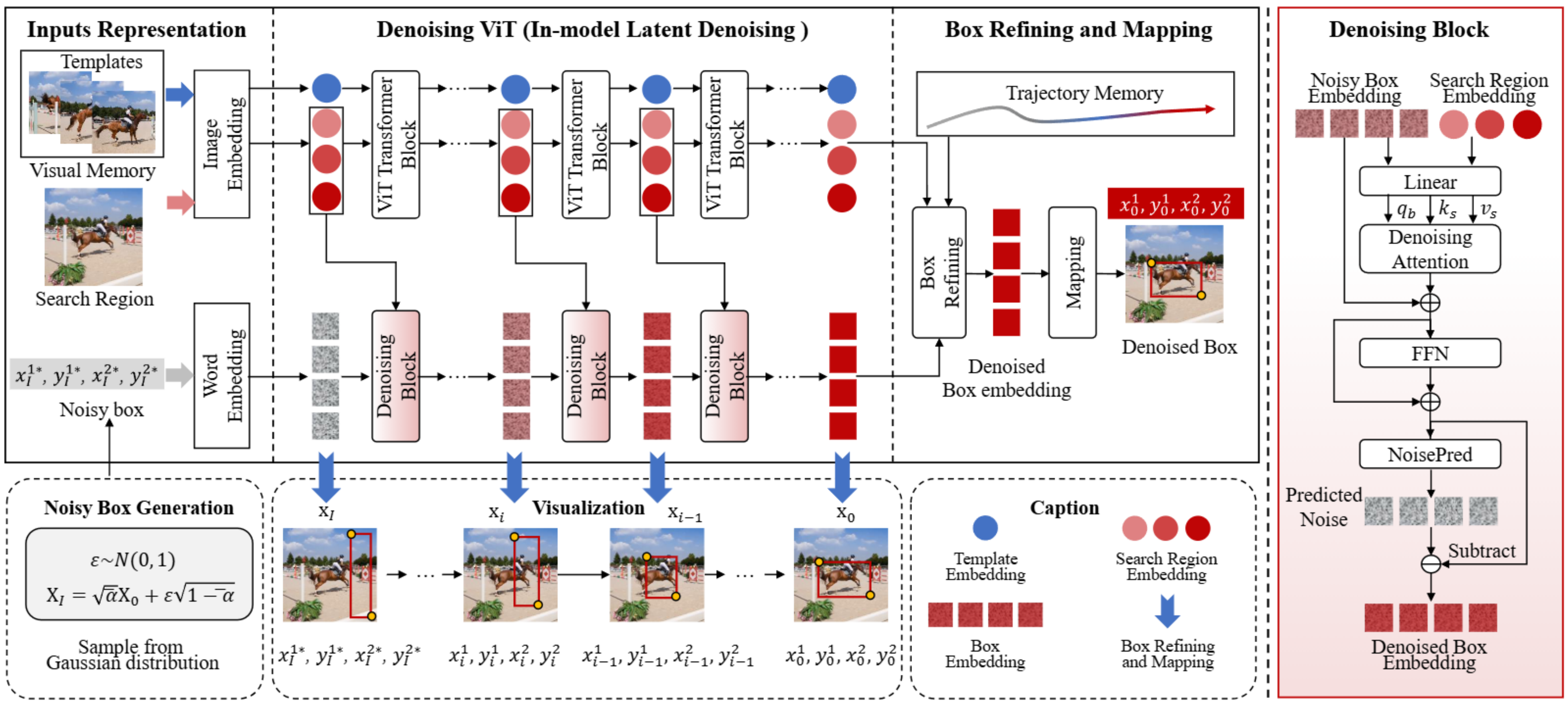
(a) Denoising Learning in Image Generation (b) Denoising Learning in Object Detection (c) The Proposed Denoising Learning Tracking Paradigm

☞ We propose a novel in-model latent denoising learning paradigm for visual object tracking,which provides a new perspective for the research community. It decomposes the classical explicit denosing process into several denoising blocks and solves the problem with atracking network in a single forward pass, which is valuable for real applications.

DeTrack: In-model Latent Denoising Learning forVisual Object Tracking

Xinyu Zhou[1], Jinglun Li[2], Lingyi Hong[1], Kaixun Jiang[2], Pinxue Guo[2], Weifeng Ge[1] and Wenqiang Zhang[1,2]
[1]School of Computer Science, Fudan University, [2]Academy for Engineering and Technology, Fudan University,

Overview

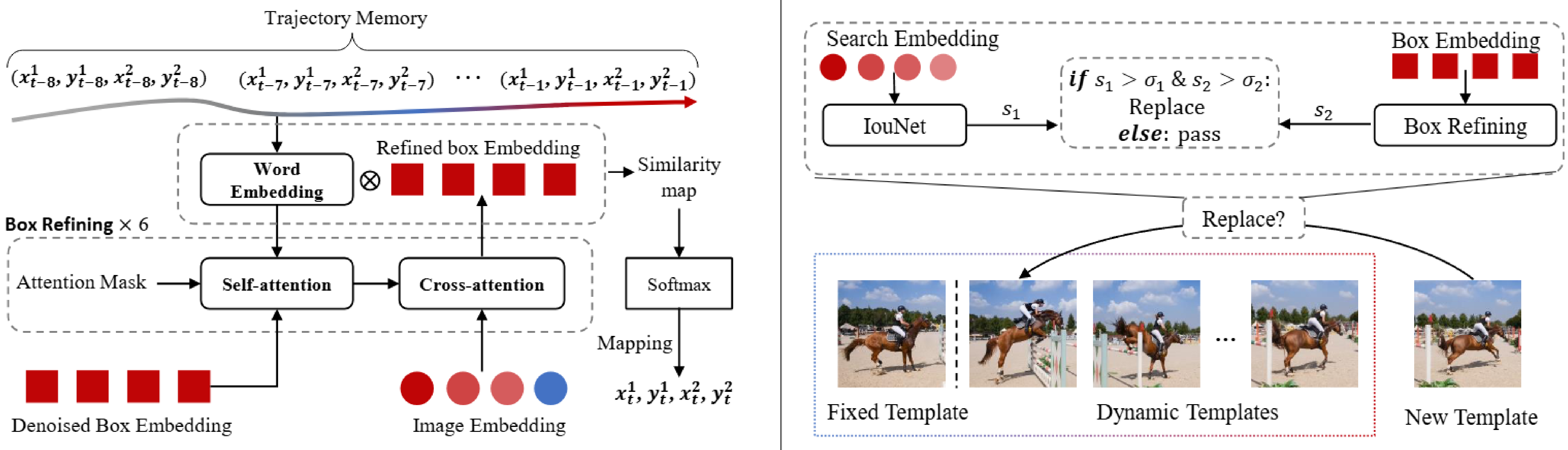


(a) (b)

We present a tracking model including a denoising ViT, comprised of multiple denoising blocks. The denoising process can be completed by progressively denoising through the denoising blocks within ViT. Furthermore, we construct a compound memory in the model that improve the tracking results using visual features and trajectory.

DeTrack: In-model Latent Denoising Learning forVisual Object Tracking

Xinyu Zhou[1], Jinglun Li[2], Lingyi Hong[1], Kaixun Jiang[2], Pinxue Guo[2], Weifeng Ge[1] and Wenqiang Zhang[1,2]
[1]School of Computer Science, Fudan University, [2]Academy for Engineering and Technology, Fudan University,



(a) Box refining and Mapping

(b) Visual memory

☞ We design a compound memory that includes both a visual memory and a trajectory memory. The visual memory enhances the model's ability to adapt to changes in the appearance of the target and the environment in the video. Besides, the trajectory memory enables the model to continue tracking the target even in the presence of occlusions or disappearances.

DeTrack: In-model Latent Denoising Learning forVisual Object Tracking

Xinyu Zhou[1], Jinglun Li[2], Lingyi Hong[1], Kaixun Jiang[2], Pinxue Guo[2], Weifeng Ge[1] and Wenqiang Zhang[1,2]
[1]School of Computer Science, Fudan University, [2]Academy for Engineering and Technology, Fudan University,

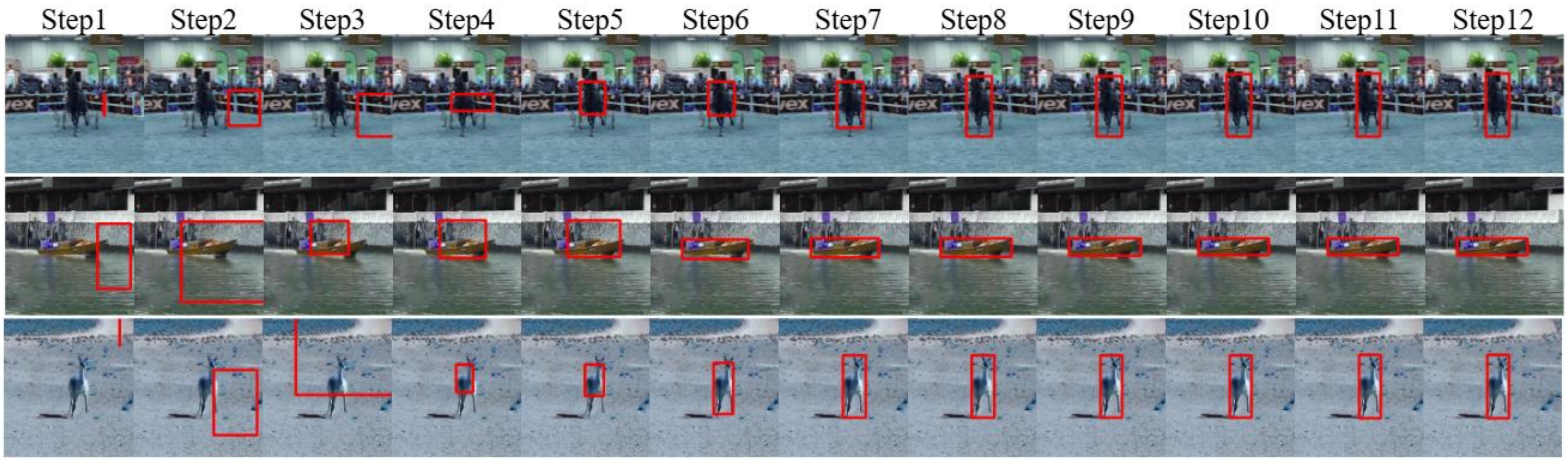
Method	AVisT			GOT-10k*			LaSOT			LaSOT_{ext}		
	AUC	OP50	OP75	AO	$SR_{0.5}$	$SR_{0.75}$	AUC	P_{Norm}	P	AUC	P_{Norm}	P
SiamPRN++_{255} [31]	39.0	43.5	21.2	51.7	61.6	32.5	49.6	56.9	49.1	34.0	41.6	39.6
DiMP_{288} [2]	-	-	-	61.1	71.7	49.2	56.9	65.0	56.7	39.2	47.6	45.1
ATOM_{288} [11]	38.6	41.5	22.2	-	-	-	51.5	57.6	50.5	37.6	45.9	43.0
PrDiMP_{288} [12]	43.3	48.0	28.7	63.4	73.8	54.3	59.8	68.8	60.8	-	-	-
Ocean_{255} [53]	38.9	43.6	20.5	61.1	72.1	47.3	56.0	65.1	56.6	-	-	-
$\text{Alpha-Refine}_{288}$ [48]	49.6	55.7	38.2	-	-	-	65.3	73.2	68.0	-	-	-
TransT_{256} [7]	49.0	56.4	37.2	67.1	76.8	60.9	64.9	73.8	69.0	-	-	-
ToMP_{288} [34]	51.9	59.5	38.9	-	-	-	67.6	78.0	72.2	45.9	-	-
DATT_{256} [52]	-	-	-	72.8	83.1	68.4	65.2	69.3	73.6	-	-	-
TATrack_{256} [24]	-	-	-	73.0	83.3	68.5	68.1	77.2	72.2	-	-	-
CTTrack_{256} [39]	56.3	66.1	44.8	71.3	80.7	70.3	67.8	77.8	74.0	-	-	-
TMT_{352} [42]	48.1	55.3	33.8	67.1	77.7	58.3	63.9	-	61.4	-	-	-
KeepTrack_{352} [35]	49.4	56.3	37.8	-	-	-	67.1	77.2	70.2	48.2	-	-
STARK_{320} [47]	51.1	59.2	39.1	68.8	78.1	64.1	67.1	77.0	-	-	-	-
AiATrack_{320} [17]	-	-	-	67.9	79.0		69.6	80.0	63.2	47.7	55.6	55.4
Mixformer_{320} [10]	56.5	66.3	45.1	70.7	80.0	67.8	69.2	78.7	74.7	-	-	-
OSTrack_{256} [51]	54.2	63.2	42.2	71.0	80.4	68.2	69.1	78.7	75.2	47.4	57.3	53.3
OSTrack_{384} [51]	57.7	67.3	48.3	73.7	83.2	70.8	71.1	81.1	77.6	50.5	61.3	57.6
SwinTrack_{224} [33]	-	-	-	71.3	81.9	64.5	67.2	70.8	-	47.6	53.9	-
SwinTrack_{384} [33]	-	-	-	72.4	80.5	67.8	71.3	76.5	-	49.1	55.6	-
ROMTrack_{256} [3]	57.8	67.6	48.6	72.9	82.9	70.2	69.3	78.8	75.6	-	-	-
ROMTrack_{384} [3]	59.1	68.7	50.5	74.2	84.3	72.4	71.4	81.4	78.2	-	-	-
F-BDMTrack_{256} [49]	-	-	-	72.7	82.0	69.9	69.9	79.4	75.8	47.9	57.9	54.0
F-BDMTrack_{384} [49]	-	-	-	75.4	84.3	72.9	72.0	81.5	77.7	50.8	61.3	57.8
GRM_{256} [18]	54.5	63.1	45.2	73.4	82.9	70.4	69.9	79.3	75.8	-	-	-
GRM_{320} [18]	55.2	64.2	46.8	73.4	82.9	70.5	69.9	79.3	75.8	-	-	-
SeqTrack_{256} [6]	56.8	66.8	45.6	74.7	84.7	71.8	69.9	79.7	76.3	49.5	60.8	56.3
SeqTrack_{384} [6]	57.8	67.4	48.0	74.8	81.9	72.2	71.5	81.8	77.8	50.5	61.6	57.5
ARTrack_{256} [43]	-	-	-	73.5	82.2	70.9	70.4	79.5	76.6	46.4	56.5	52.3
ARTrack_{384} [43]	-	-	-	75.5	84.3	74.3	72.6	81.7	**79.1**	51.9	62.0	58.5
DeTrack_{256} (ours)	60.1	**69.7**	**50.6**	77.1	86.1	73.5	71.3	80.1	76.8	47.9	56.6	52.1
DeTrack_{384} (ours)	**60.2**	69.1	50.2	**77.9**	**86.5**	**74.9**	**72.9**	**81.7**	**79.1**	**53.6**	**64.4**	**60.4**

Quantitative Results

- Our tracker demonstrates outstanding performance on AVisT, a dataset with extreme weather conditions and harsh environments. It outperforms SeqTrack384 by 2.4% in AUC, substantiating our tracker’s excellence in extreme environmental conditions
- Our method demonstrates superior performance on the GOT-10k. Our DeTrack256 achieves a significant improvement in AUC compared to SeqTrack256 , with increases of 3.0% and 2.4%, respectively. Our DeTrack384 outperforms the state-of-the-art method ARTrack384 by 2.4%.
- LaSOT is benchmark designed for long-term tracking, featuring a test collection consisting of 280 videos. Our DeTrack256 achieves an AUC of 71.3%, exhibiting performance improvement compared to other methods. Additionally, our DeTrack384 also demonstrates state-of-the-art performance. LaSOText demonstrates the strong generalization capability of our approach even with extended data, particularly manifesting notable advantages in the accuracy of bounding box center point.

DeTrack: In-model Latent Denoising Learning forVisual Object Tracking

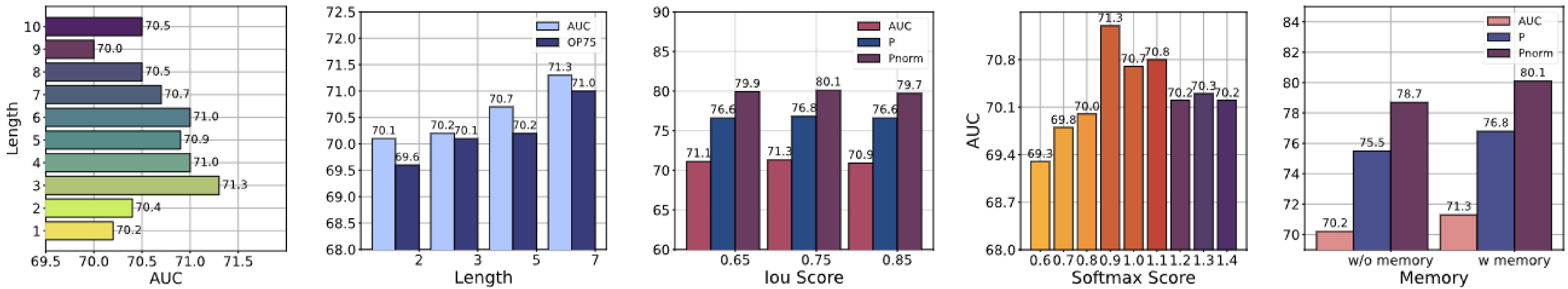
Xinyu Zhou[1], Jinglun Li[2], Lingyi Hong[1], Kaixun Jiang[2], Pinxue Guo[2], Weifeng Ge[1] and Wenqiang Zhang[1,2]
[1]School of Computer Science, Fudan University, [2]Academy for Engineering and Technology, Fudan University,



☞Here is the visualization of our denoising process from step 1 to step 12. It demonstrates the gradual denoising from the first to the twelfth step, ultimately resulting in the final bounding box.

DeTrack: In-model Latent Denoising Learning forVisual Object Tracking

Xinyu Zhou[1], Jinglun Li[2], Lingyi Hong[1], Kaixun Jiang[2], Pinxue Guo[2], Weifeng Ge[1] and Wenqiang Zhang[1,2]
[1]School of Computer Science, Fudan University, [2]Academy for Engineering and Technology, Fudan University,



(a) Visual Memory Length **(b) Trajectory Memory Length** **(c) IoU Threshold** **(d) Softmax Threshold** **(e) Compound memory**

☞ We also conducted ablation experiments on the lengths of the visual memoryand trajectory memory, as well as on the IoU threshold and Softmax threshold. Additionally, we performed an ablation study on the overall compound memory.