Tracking Meets LoRA: Faster Training, Larger Model, Stronger Performance

Liting Lin[1], Heng Fan[2], Zhipeng Zhang[3], Yaowei Wang[1#], Yong Xu[4], Haibin Ling[5#]

[1]Peng Cheng Laboratory [2]Department of CSE, University of North Texas [3]KargoBot

[4]School of Computer Science & Engineering, South China Univ. of Tech. [5]Department of Computer Science, Stony Brook University

GitHub

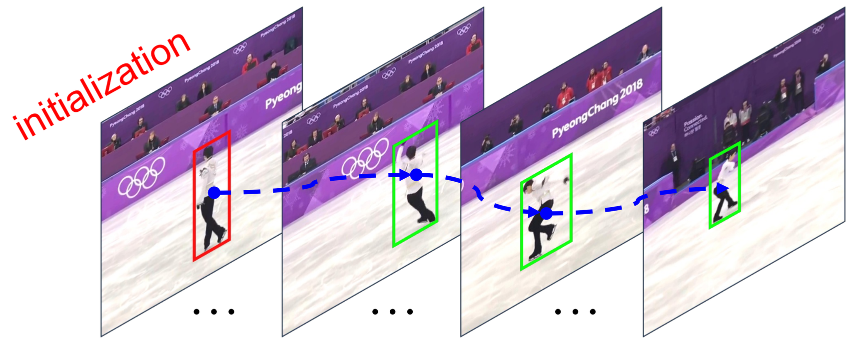
EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO 2024

Introduction

- Visual Object Tracking

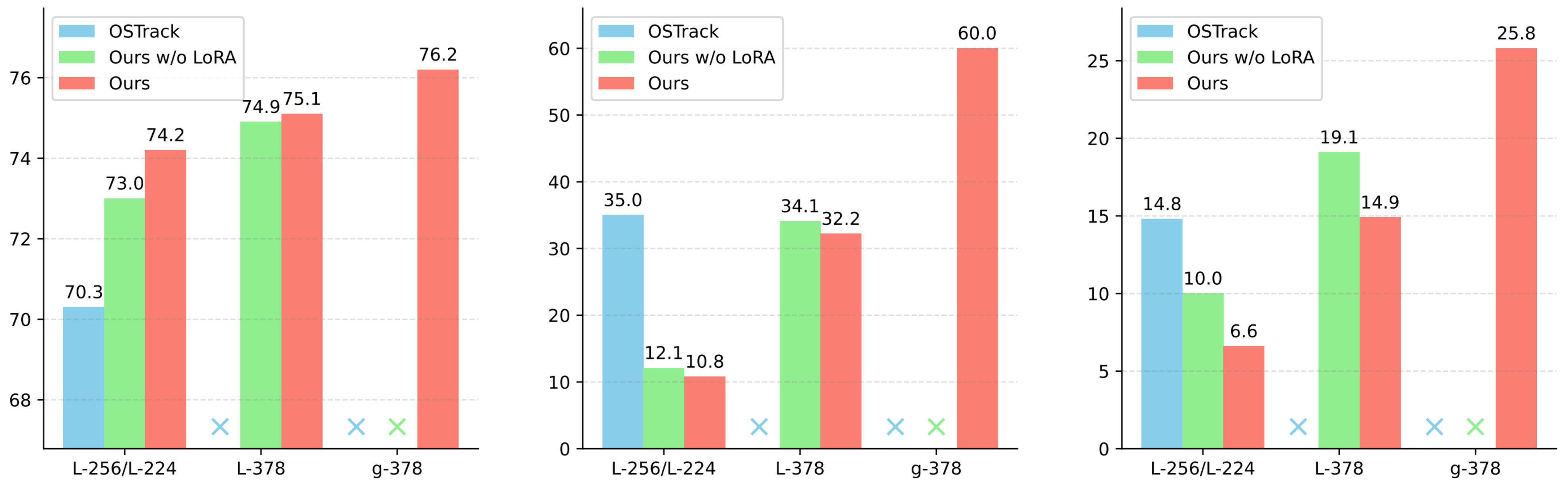
  Goal: Continuously localize object of interest in a video.



- Motivation
  - ➢ Exploring the scaling law in tracking to advance the field.
  - ➢ Applying LoRA to pure vision models is still lack of exploration, presenting unique challenges.
  - ➢ Training large-scale trackers with laboratory-manageable resources.

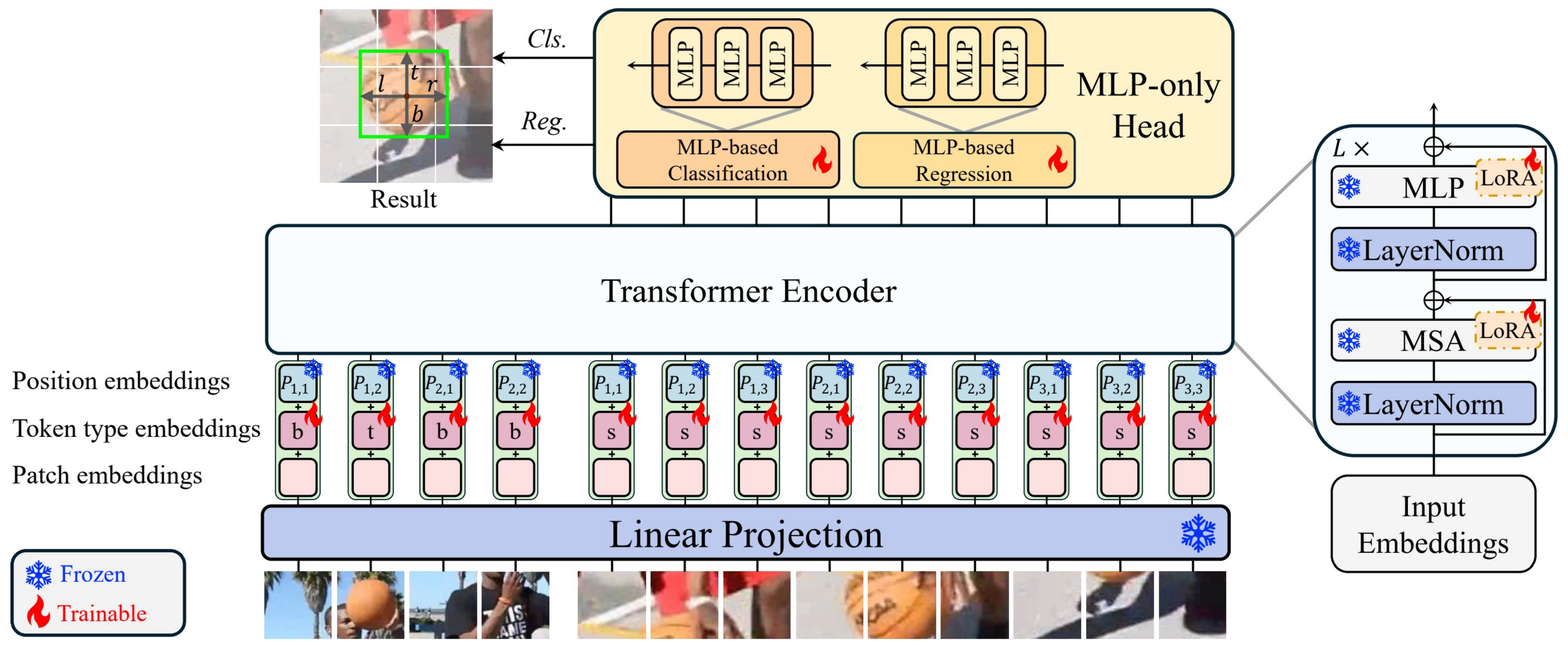
Contributions

- ❖ First generic object tracking model trained in a parameter-efficient way.
- ❖ Two simple yet effective designs enabling better adaption of LoRA for tracking.
- ❖ **LoRAT**: New state-of-the-art performance on multiple benchmarks with reasonable resource requirements.



(a) SUC on LaSOT (b) Training Time (h) (c) Training Memory (GB)

Challenges and Solutions



Our Finding

Existing one-stream trackers fail to converge when simply adopting LoRA on linear layers. Changes to model design essential to lower "optimization difficulty" (quantified by Gradient Norm of loss function during training). Bottlenecks include:

- Separate positional embeddings for template and search region tokens: *disrupt* the structure of the pre-trained ViT model
- Inductive bias introduced by convolutional heads: CNNs have much image-specific *inductive bias* but not applied in pre-training tasks of ViTs.

Our Solution

We apply **LoRA** for efficient training via two key solutions:

- **Decoupled Positional Embedding**: We separate positional embeddings into shared spatial embeddings (inherited from pre-trained backbones) and independent type embeddings (learned from scratch). This design preserves the structure of the pre-trained model, ensuring compatibility with LoRA for efficient fine-tuning. We also adopt type embeddings to explicitly annotating foreground and background parts within the template, further reducing the confusion during training.
- **MLP-Based Anchor-Free Head**: We replace the convolutional head with an MLP-based anchor-free head to eliminate inductive biases, enabling more flexible and efficient fine-tuning with LoRA.

Experimental Results

Comparison with state-of-the-arts.

Tracker	LaSOT			$LaSOT_{ext}$			TrackingNet			GOT-10k			TNL2K	
	SUC	P_{Norm}	P	SUC	P_{Norm}	P	SUC	P_{Norm}	P	AO	$SR_{0.5}$	$SR_{0.75}$	SUC	P
OSTrack	71.1	81.1	77.6	50.5	61.3	57.6	83.9	88.5	83.2	73.7	83.2	70.8	55.9	56.7
SwinTrack	71.3	-	76.5	49.1	-	55.6	84.0	-	82.8	72.4	80.5	67.8	55.9	57.1
DropTrack	71.8	81.8	78.1	52.7	63.9	60.2	-	-	-	75.9	86.8	72.0	56.9	57.9
SeqTrack	72.5	81.5	79.3	50.7	61.6	57.5	85.5	89.8	85.8	74.8	81.9	72.2	57.8	-
ARTrack	73.1	82.2	80.3	52.8	62.9	59.7	85.6	89.6	86.0	78.5	87.4	77.8	60.3	-
CiteTracker	69.7	78.6	75.7	-	-	-	84.5	89.0	84.2	74.7	84.3	73.0	57.7	59.6
MixViT	72.4	82.2	80.1	-	-	-	85.4	**90.2**	85.7	75.7	85.3	75.1	-	-
LoRAT-B-224	71.7	80.9	77.3	50.3	61.6	57.1	83.5	87.9	82.1	72.1	81.8	70.7	58.8	61.3
LoRAT-B-378	72.9	81.9	79.1	53.1	64.8	60.6	84.2	88.4	83.0	73.7	82.6	72.9	59.9	63.7
LoRAT-L-224	74.2	83.6	80.9	52.8	64.7	60.0	85.0	89.5	84.4	75.7	84.9	75.0	61.1	65.1
LoRAT-L-378	75.1	84.1	82.0	**56.6**	**69.0**	**65.1**	85.6	89.7	85.4	77.5	86.2	78.1	62.3	67.0
LoRAT-g-224	74.9	84.5	82.3	53.3	65.4	61.1	85.2	89.8	85.1	77.7	87.7	77.7	61.8	66.6
LoRAT-g-378	**76.2**	**85.3**	**83.5**	56.5	**69.0**	64.9	**86.0**	**90.2**	**86.1**	**78.9**	**87.8**	**80.7**	**62.7**	**67.8**

Inference efficiency.

Tracker	Speed (*fps*)	MACs (G)	#Params (M)
SwinTrack-B-384	45	69.7	91
OSTrack-256	130	21.5	-
OSTrack-384	68	48.3	-
SeqTrack-B256	38	66	89
SeqTrack-L384	6	524	309
LoRAT-B-224	209	30	99 (11, 2)
LoRAT-B-378	151	97	99 (11, 2)
LoRAT-L-224	119	103	336 (28, 4)
LoRAT-L-378	63	325	336 (28, 4)
LoRAT-g-224	50	378	1216 (71, 9)
LoRAT-g-378	20	1161	1216 (71, 9)

Training efficiency.

	Variant	Time(h)	Memory(GB)
LoRA	B-224	5.9	2.4
	B-378	12.2	5.7
	L-224	10.8	6.6
	L-378	32.2	14.9
	g-224	22.3	14.0
	g-378	60.0	25.8
Full Fine-tuning	B-224	5.9	3.2
	B-378	12.5	6.5
	L-224	12.1	10.0
	L-378	34.1	19.1
	g-224	29.3	27.1
	g-378	Out of Memory	

Ablations on proposed decoupled positional embedding.

	Shared P. Emb.	Type Emb.	Foreg. Indic.	Res.	LaSOT		$LaSOT_{ext}$		TNL2K	
					SUC	P	SUC	P	SUC	P
①				224	73.5	80.2	51.8	59.1	60.7	64.4
②	✓			224	73.8	80.6	53.7	61.4	60.6	64.5
③	✓	✓		224	74.0	80.7	52.4	59.6	60.7	64.7
④	✓	✓	✓	224	74.2	80.9	52.8	60.0	61.1	65.1
⑤	✓	✓		378	74.4	82.6	55.2	63.2	62.3	67.1
⑥	✓	✓	✓	378	75.1	82.0	56.6	65.1	62.3	67.0

Gradient norm (conv head vs. mlp head)