Divert More Attention to Vision-Language Tracking

Mingzhe Guo[1]* Zhipeng Zhang[2]* Heng Fan[3] Liping Jing[1]

[1]Beijing Jiaotong University [2]DiDi Chuxing [3]University of North Texas *Equal Contribution

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

➢Is Transformer the only path to SOTA?

No. Vision-Language (VL) multimodal tracking with pure CNN architecture is another way, which requires less training data and training time.

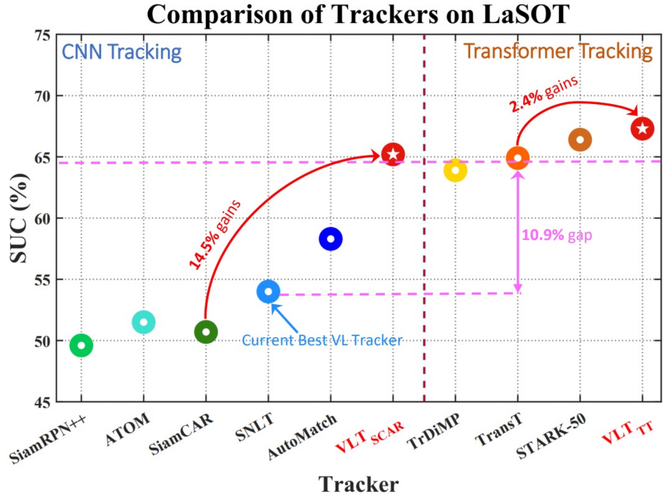
➢What is the bottleneck of VL tracking in achieving SOTA?

It's in the VL representation. Existing VL trackers treat vision and language *independently* and processed *distantly* until the final result fusion. This results in in a lower upper-bound for VL tracking.

➢Our Solution?

Learning a novel unified-adaptive vision-language representation, aiming for SOTA VL tracking without using Transformer.

- *unified*, needs deep interactions of multimodal.
- *adaptive*, requires to accommodate different scenarios of visual and linguistic information.

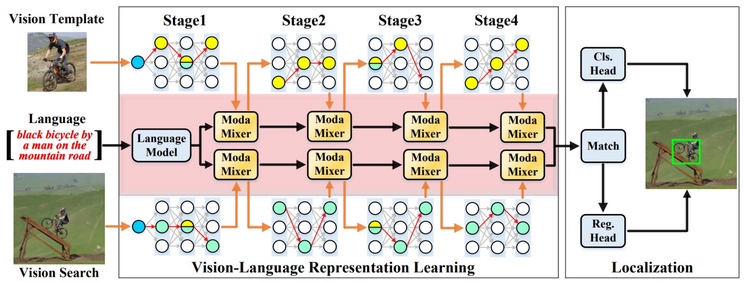


Contributions

- We introduce a novel unified-adaptive vision-language representation for SOTA VL tracking.
- We propose the embarrassingly simple yet effective ModaMixer for unified VL representation learning.
- We present ASS to adapt mixed VL representation for better tracking.
- With pure CNN architecture, we achieve SOTA results on multiple benchmarks.

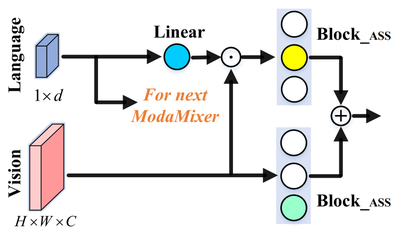
VL Multimodal Tracking Framework

- Adopting a hierarchical design to employ the ModaMixer.
- Arranging the proposed ModaMixer to different stages of the searched asymmetrical backbone to conduct multiple multimodal fusions.
- With the matching-based paradigm, both template and search backbone networks contain 4 stages, and the chosen NAS blocks of each stage are denoted with different colors.
- the asymmetry is revealed in not only the design of backbone networks, but also the ModaMixer.



ModaMixer with ASS

- Modality Mixer (ModaMixer) unifies multimodal information by considering language representation as selector to reweight channels of vision features.
- Asymmetrical Searching Strategy (ASS) learns an adaptive modeling structure for pairing with ModaMixer, which searches the optimal architects for different modalities.



$$\mathbf{f}_m = \text{Block}_{\text{ASS}}\left(\text{Linear}(\mathbf{f}_l) \odot \mathbf{f}_v\right) + \text{Block}_{\text{ASS}}\left(\mathbf{f}_v\right)$$

$$W_{\mathcal{A}} = \underset{W}{\text{argmin}}\ \mathbb{E}_{a\sim\Gamma(\mathcal{A})}\left[\mathcal{L}_{\text{train}}(\mathcal{N}(a, W(a)))\right]$$

$$a^* = \underset{a\in\mathcal{A}}{\text{argmax}}\ \text{SUC}_{\text{val}}\left(\mathcal{N}(a, W_{\mathcal{A}}(a))\right)$$

	Stem	Stage1	Moda Mixer	Stage2	Moda Mixer	Stage3	Moda Mixer	Stage4	Moda Mixer
Template	●	● ● ●	● ●	● ● ●	● ●	● ● ● ●	● ●	● ● ●	● ●
Search	●	● ● ●	● ●	● ● ●	● ●	● ● ● ●	● ●	● ● ●	● ●

Experiment

- Achieving considerable performance gains compared to the baselines, which also outperform current best VL tracker SNLT and some Transformer-based trackers.
- Transformer-based tracker could also be improved with the proposed ModaMixer and ASS.

Type	Method	LaSOT		$\text{LaSOT}_{\text{Ext}}$		TNL2K		GOT-10k			OTB99-L	
		SUC	P	SUC	P	SUC	P	AO	$\text{SR}_{0.5}$	$\text{SR}_{0.75}$	SUC	P
CNN-based	SiamRCNN [52]	64.8	68.4	-	-	52.3	52.8	64.9	72.8	59.7	70.0	89.4
	PrDiMP [13]	59.8	60.8	-	-	47.0	45.9	63.4	73.8	54.3	69.5	89.5
	AutoMatch [62]	58.3	59.9	37.6	43.0	47.2	43.5	65.2	76.6	54.3	71.6	93.2
	Ocean [64]	56.0	56.6	-	-	38.4	37.7	61.1	72.1	47.3	68.0	92.1
	KYS [6]	55.4	-	-	-	44.9	43.5	63.6	75.1	51.5	-	-
	ATOM [12]	51.5	50.5	37.6	43.0	40.1	39.2	55.6	63.4	40.2	67.6	82.4
	SiamRPN++ [32]	49.6	49.1	34.0	39.6	41.3	41.2	51.7	61.6	32.5	63.8	82.6
	C-RPN [18]	45.5	42.5	27.5	32.0	-	-	-	-	-	-	-
	SiamFC [5]	33.6	33.9	23.0	26.9	29.5	28.6	34.8	35.3	9.8	58.7	79.2
	ECO [11]	32.4	30.1	22.0	24.0	-	-	31.6	30.9	11.1	-	-
	SiamCAR [23]	50.7	51.0	33.9	41.0	35.3	38.4	56.9	67.0	41.5	68.8	89.1
CNN-VL	SNLT [20]	54.0	57.6	26.2	30.0	27.6	41.9	43.3	50.6	22.1	66.6	80.4
	$\text{VLT}^{0}_{\text{SCAR}}$ (Ours)	**65.2**	**69.1**	**41.2**	**47.5**	**48.3**	**46.6**	**61.4**	**72.4**	**52.3**	**72.7**	**88.8**
	$\text{VLT}^{3}_{\text{SCAR}}$ (Ours)	**63.9**	**67.9**	**44.7**	**51.6**	**49.8**	**51.1**	**61.0**	**70.8**	**52.2**	**73.9**	**89.8**
Trans-based	STARK [56]	66.4	71.2	47.8	55.1	-	-	68.0	77.7	62.3	69.6	91.4
	TrDiMP [54]	63.9	66.3	-	-	-	-	67.1	77.7	58.3	70.5	92.5
	TransT [8]	64.9	69.0	44.8	52.5	50.7	51.7	67.1	76.8	60.9	70.8	91.2
Trans-VL	$\text{VLT}^{0}_{\text{TT}}$ (Ours)	**66.3**	**70.5**	**45.4**	**52.1**	**52.2**	**52.1**	**68.4**	**81.5**	**62.4**	**74.7**	**91.2**
	$\text{VLT}^{4}_{\text{TT}}$ (Ours)	**67.3**	**72.1**	**48.4**	**55.9**	**53.1**	**53.3**	**69.4**	**81.1**	**64.5**	**76.4**	**93.1**

- Each component brings considerable improvement.
- VL multimodal fusion of ModaMixer is the key to SOTA.

#	Method	ModaMixer	ASS	LaSOT			TNL2K		
				SUC	P_{Norm}	P	SUC	P_{Norm}	P
①	Baseline	-	-	50.7	60.0	51.0	35.3	43.6	38.4
②	VLT_{SCAR}	√	-	57.6	65.8	61.1	41.5	49.2	43.2
③	VLT_{SCAR}	-	√	52.1	59.8	50.6	40.7	47.2	40.2
④	VLT_{SCAR}	√	√	**65.2**	**74.9**	**69.1**	**48.3**	**55.2**	**46.6**

- More specific description could help the tracker to distinguish the target from the background clutter.
- The mere description of the environment (the fourth column) may introduce interference instead.