ICML International Conference On Machine Learning

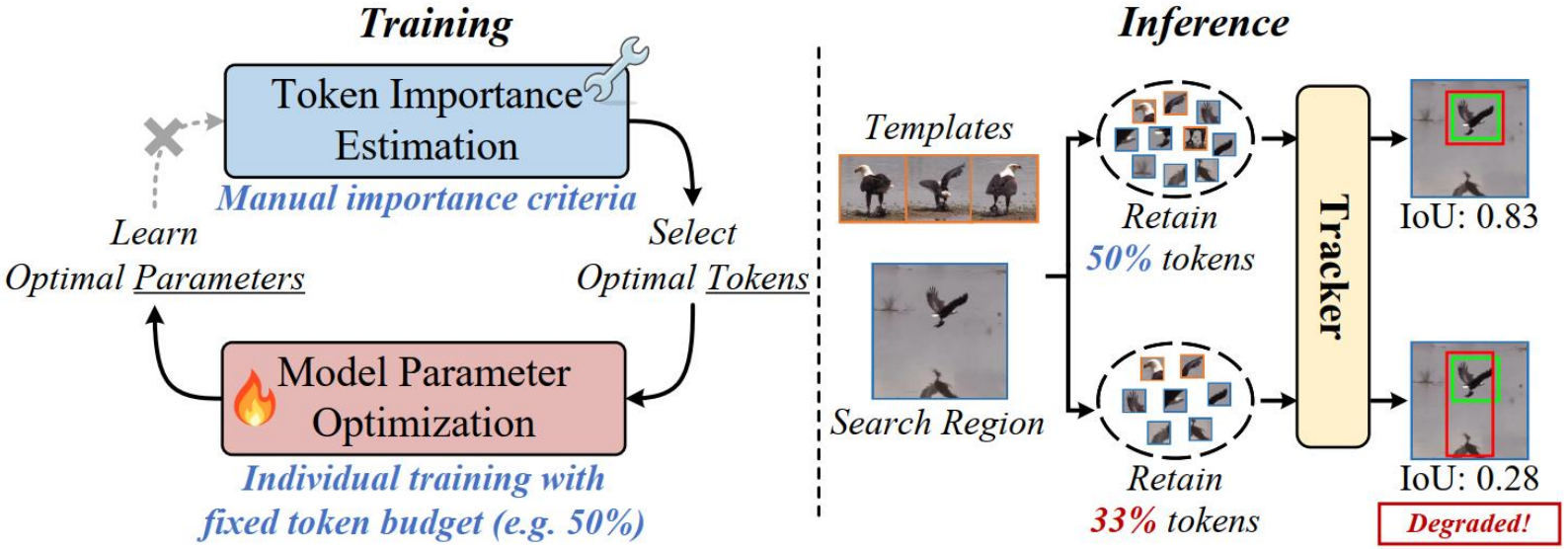
Proceedings of the 43rd International Conference on Machine Learning

Learning Generalized Trackers with Elastic Token Budgets

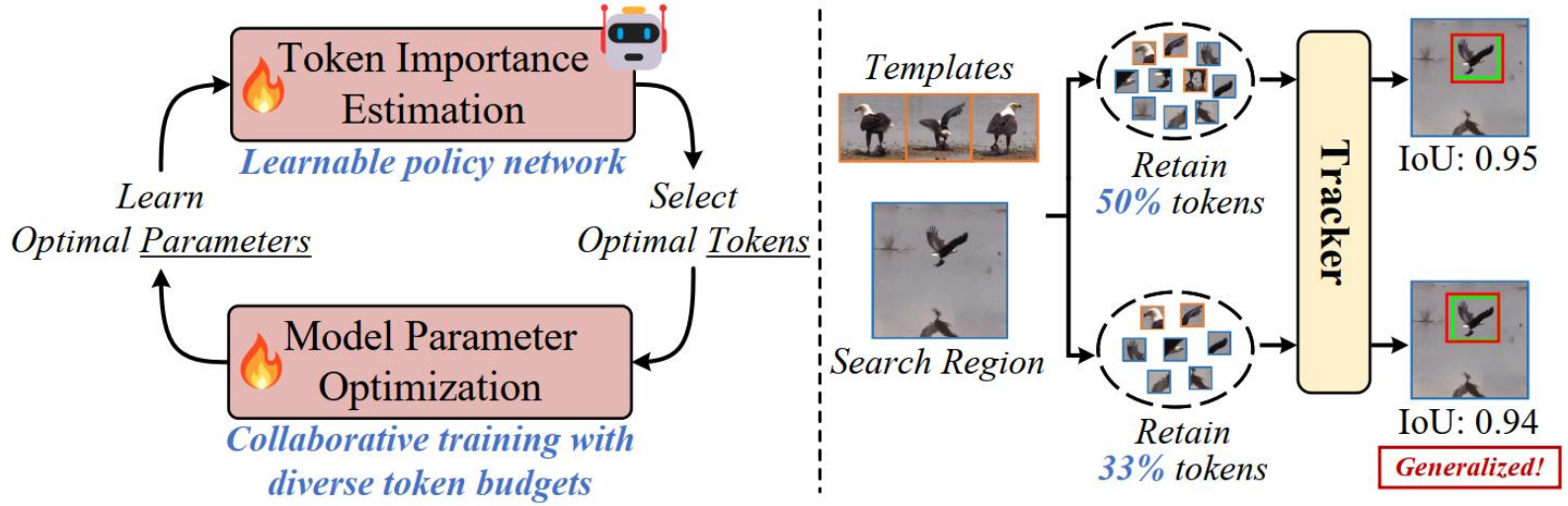
Yinchao Ma[1], Jianpeng Yang[1], Yuyang Tang[1], Jie Xiao[1], Dengqing Yang[1], Tianzhu Zhang[1]

[1]Deep Space Exploration Laboratory/School of Information Science and Technology, University of Science and Technology of China

Motivation

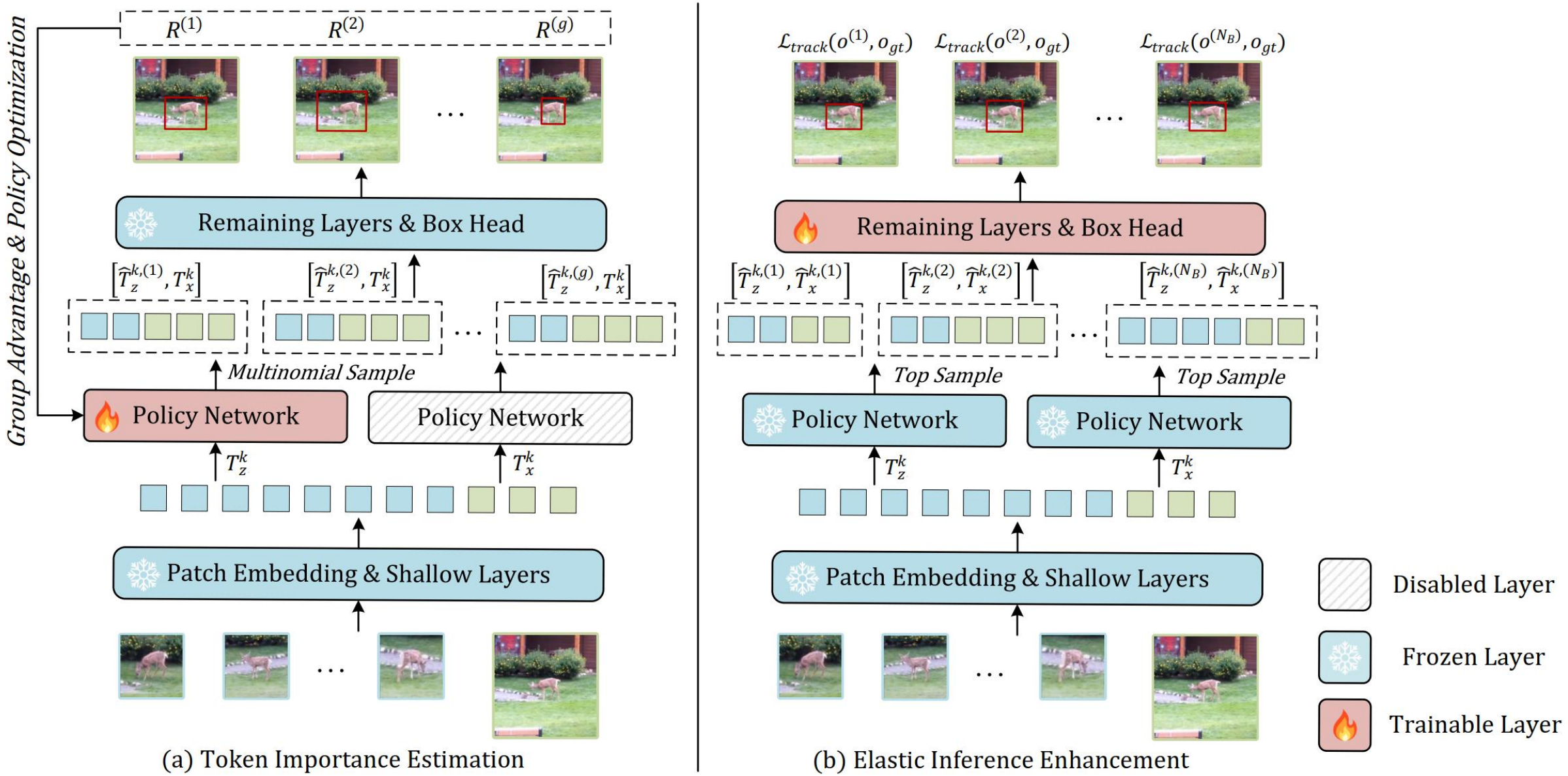


(a) Previous solution for tracking with pruned tokens



(b) Our solution for tracking with elastic token budgets

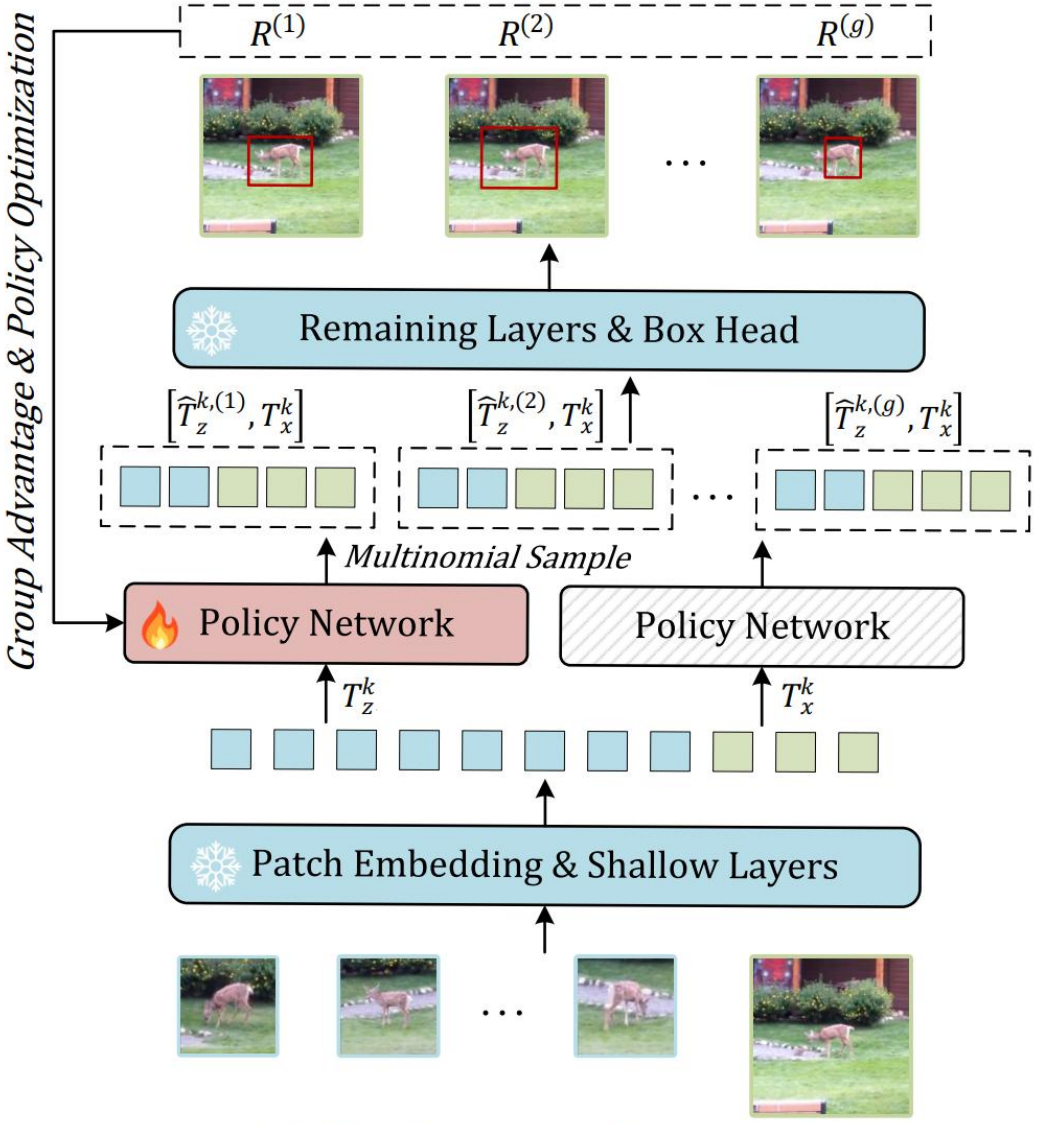
Framework


Group Advantage & Policy Optimization
$R^{(1)}$
$R^{(2)}$
$R^{(g)}$
Remaining Layers & Box Head
$[\hat{T}_z^{k,(1)}, T_x^k]$
$[\hat{T}_z^{k,(2)}, T_x^k]$
$[\hat{T}_z^{k,(g)}, T_x^k]$
Multinomial Sample
Policy Network
Policy Network
T_z^k
T_x^k
Patch Embedding & Shallow Layers
(a) Token Importance Estimation
$\mathcal{L}_{track}(o^{(1)}, o_{gt})$
$\mathcal{L}_{track}(o^{(2)}, o_{gt})$
$\mathcal{L}_{track}(o^{(N_B)}, o_{gt})$
Remaining Layers & Box Head
$[\hat{T}_z^{k,(1)}, \hat{T}_x^{k,(1)}]$
$[\hat{T}_z^{k,(2)}, \hat{T}_x^{k,(2)}]$
$[\hat{T}_z^{k,(N_B)}, \hat{T}_x^{k,(N_B)}]$
Top Sample
Top Sample
Policy Network
Policy Network
T_z^k
T_x^k
Patch Embedding & Shallow Layers
(b) Elastic Inference Enhancement
Disabled Layer
Frozen Layer
Trainable Layer

Token Importance Estimation

- We propose a novel result driven importance criteria to align the objectives of importance estimation and tracking precision.

$$P_z = \text{Policy}_z(T_z^k) \in \mathbb{R}^{N_z},$$
$$P_x = \text{Policy}_x(T_x^k) \in \mathbb{R}^{N_x}.$$



(a) Token Importance Estimation

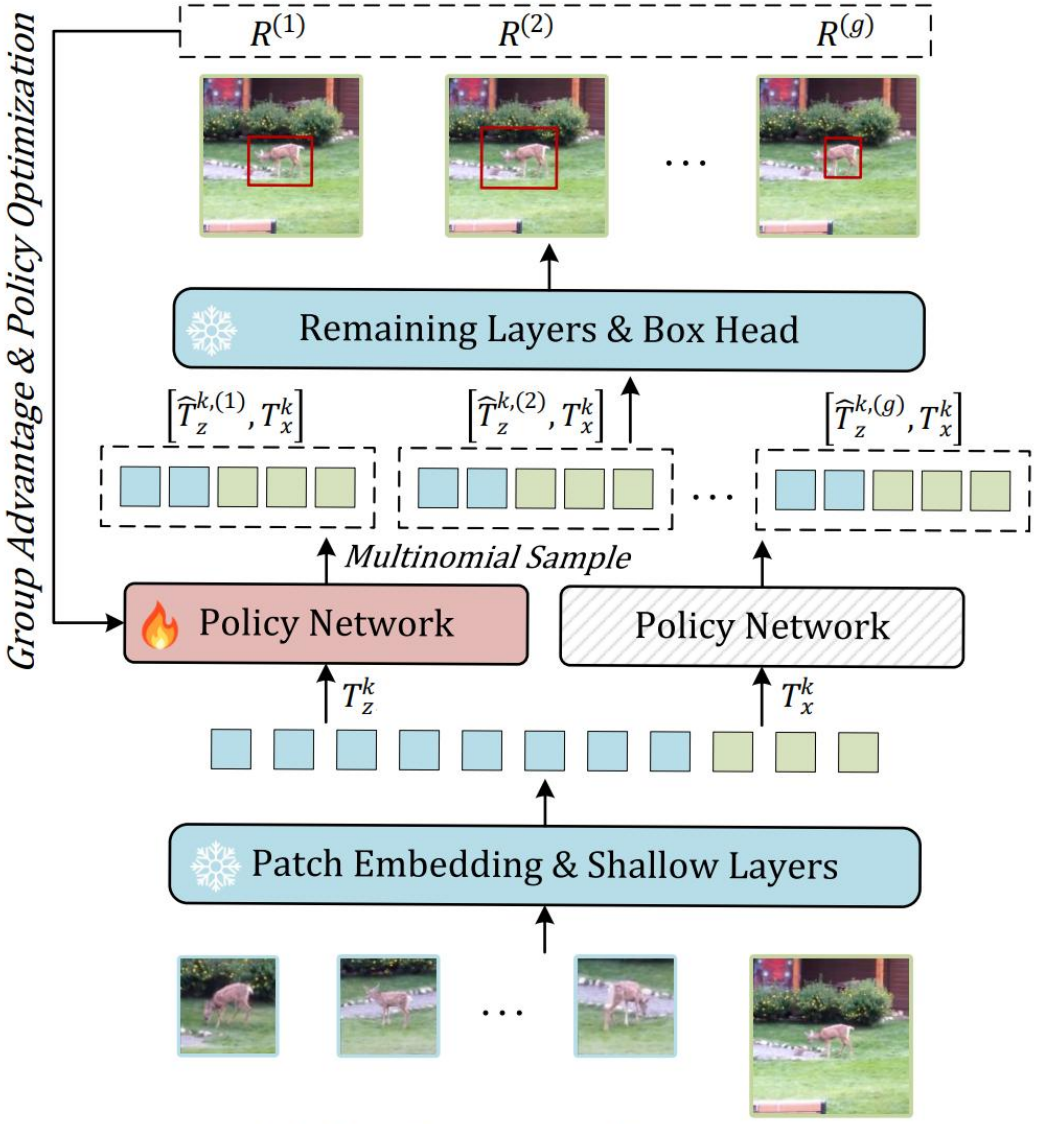
Token Importance Estimation

- We sample g group **template**/search tokens using a multinomial distribution parameterized by the importance scores output from the policy network.

$$p = \mathrm{Softmax}(P_z),$$
$$\mathcal{S}_z^{(i)} \sim \mathrm{Multinomial}\,(N_{sample},\, p)\,,$$
$$\hat{T}_z^{k,(i)} = T_z^k[\mathcal{S}_z^{(i)}, :] \in \mathbb{R}^{N_{sample} \times C}.$$

- We employ the IoU between predicted and groundtruth bounding boxes as the reward signal R and compute the advantage function A for these token groups.

$$R^{(i)} = \mathrm{IoU}(\hat{b}^{(i)}, b_{gt}),$$
$$A^{(i)} = \frac{R^{(i)} - \mathrm{mean}(\{R^{(i)}\}_{i=1}^{g})}{\mathrm{std}(\{R^{(i)}\}_{i=1}^{g})},$$



(a) Token Importance Estimation

Token Importance Estimation

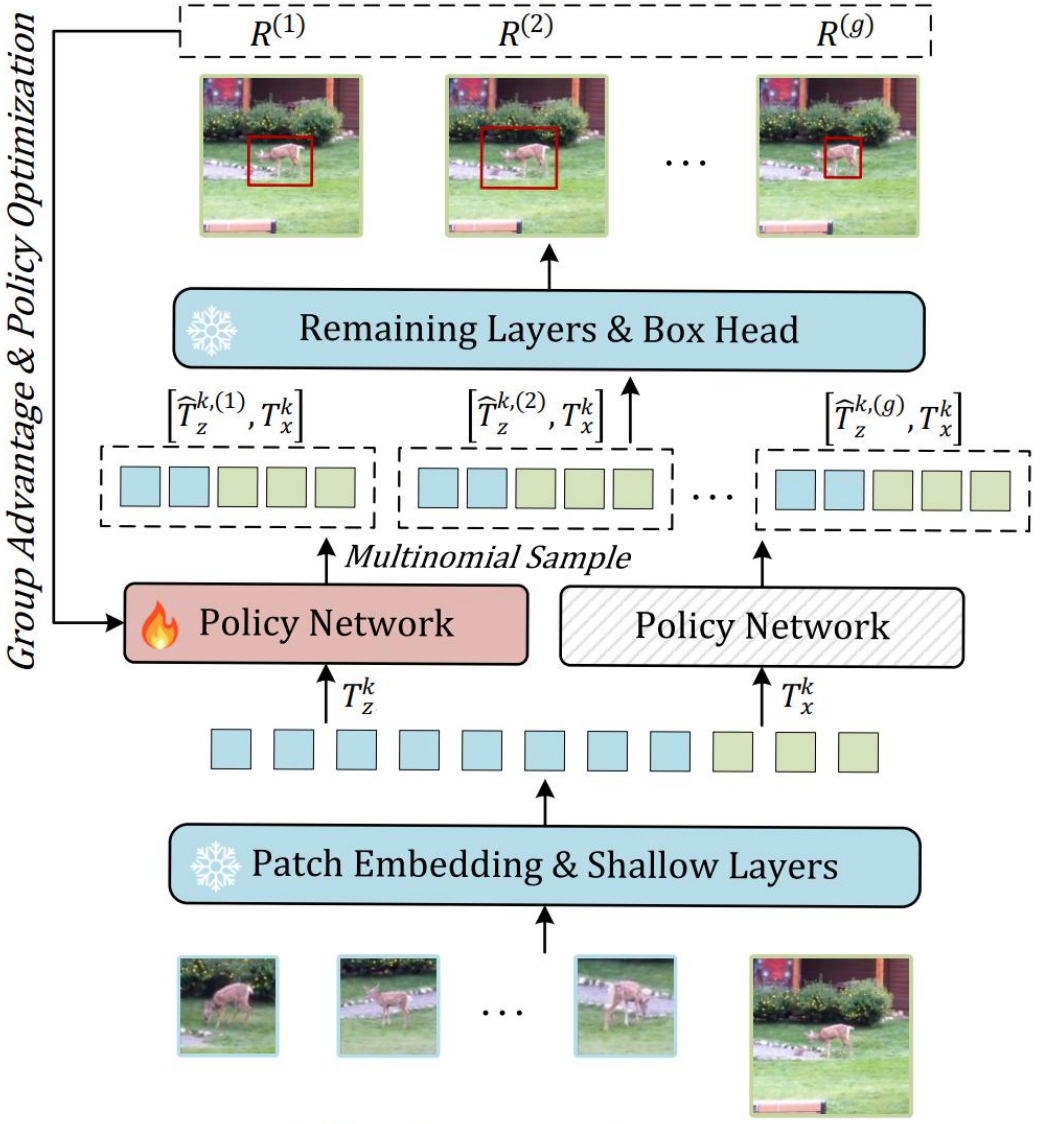
- For ease of exposition, we reformulate the template policy network with parameter θ as $\pi_\theta^z(T_x^k, \mathcal{S})$.

$$\pi_\theta^z(j|T_z^k, \mathcal{S}) = \begin{cases} \sigma(P_z), & \text{if } j \in \mathcal{S} \\ 1 - \sigma(P_z), & \text{if } j \in \mathcal{Z}_z \setminus \mathcal{S} \end{cases}$$

- We optimize the policy network by maximizing the following objective function.

$$\mathcal{J}^z(\theta) = \mathbb{E}_{j \in \mathcal{Z}_z} \left[\sum_{i=1}^{g} r_{i,j}^z(\theta) - \beta \mathbb{D}_{KL}(\pi_\theta^z || \pi_{\theta_{ref}}^z) \right]$$

$$r_{i,j}^z(\theta) = \pi_\theta^z \left(j|T_z^k, \mathcal{S}^{(i)} \right) A^{(i)}. \quad j \in \mathcal{Z}_z$$



(a) Token Importance Estimation

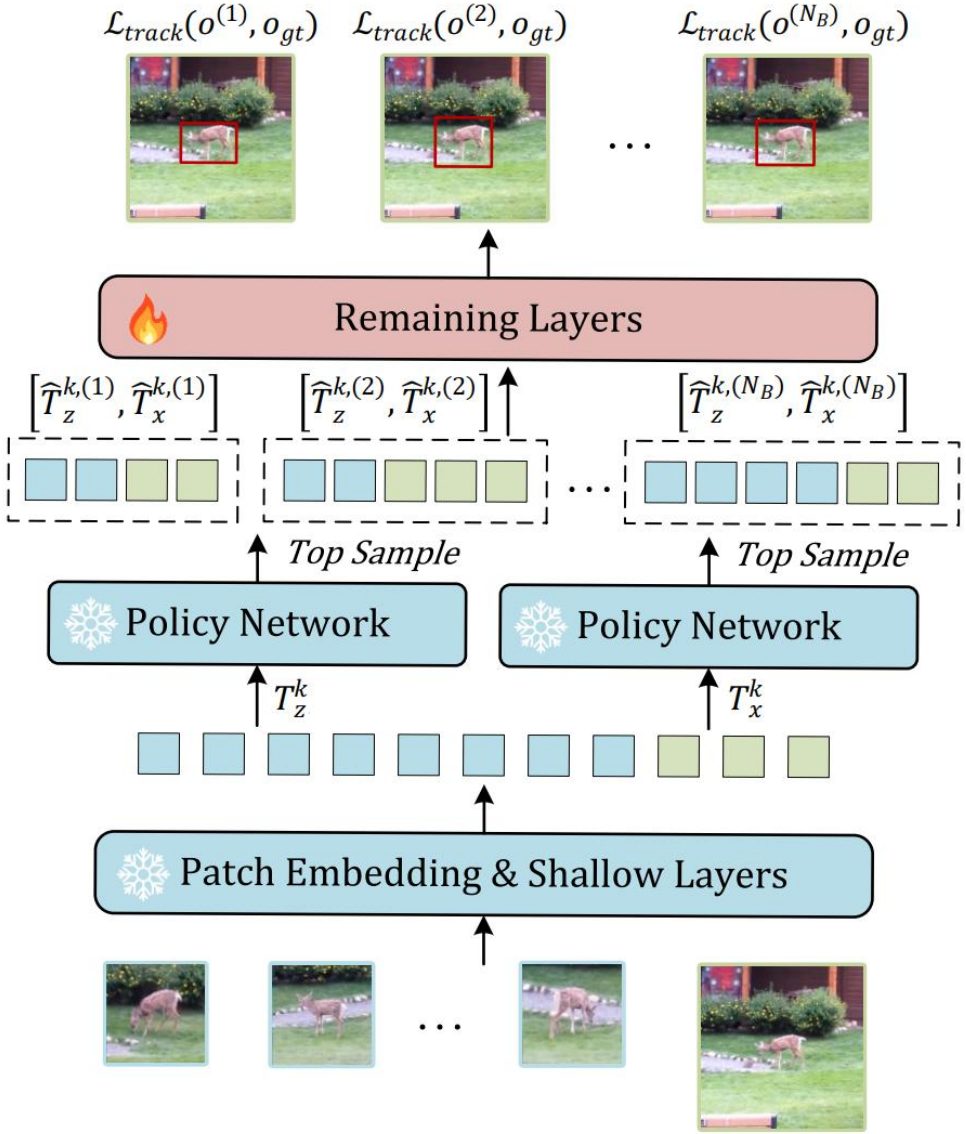
Elastic Inference Enhancement

- We freeze the policy network and all preceding layers, and utilize the trained policy network to select the most important tokens under different budgets $\{B_z^{(i)}, B_x^{(i)}\}_{i=1}^{N_B}$.

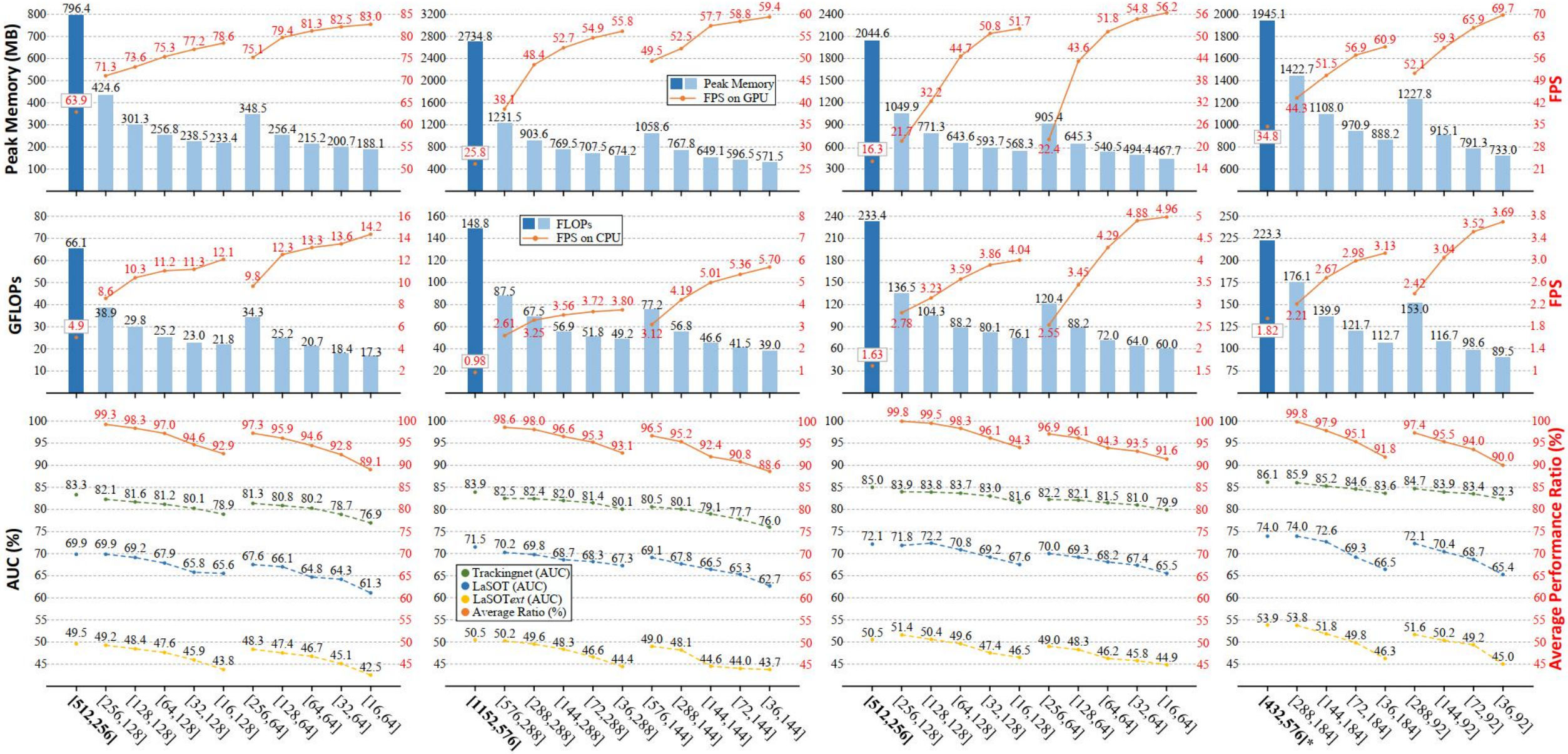
$$\mathcal{I}_z^{(i)} = \text{argtop}(P_z, B_z^{(i)}),$$
$$\mathcal{I}_x^{(i)} = \text{argtop}(P_x, B_x^{(i)}),$$
$$T_z^{(i)} = T_z[\mathcal{I}_z^{(i)}, :] \in \mathbb{R}^{B_z \times C},$$
$$T_x^{(i)} = T_x[\mathcal{I}_x^{(i)}, :] \in \mathbb{R}^{B_x \times C}.$$

- We use these sampled tokens under different budgets for collaborative optimization, thereby enhancing the tracking precision of the tracker under varying token budgets.

$$o^{(i)} = \mathcal{M}([T_z^{(i)}, T_x^{(i)}]),$$
$$\mathcal{L} = \sum_{i=1}^{N_B} \mathcal{L}_{track}(o^{(i)}, o_{gt}),$$



Quantitative Results



(a)Results on SeqTrack-B256 (b)Results on SeqTrack-B384 (c)Results on SeqTrack-L256 (d)Results on ODTrack-L384

Quantitative Results

Table 2. Comparison with state-of-the-art trackers on LaSOT and TrackingNet.

Method	LaSOT (AUC)	TrackingNet (AUC)	FPS
HIPTrack (Cai et al., 2024)	72.7	84.5	42.6
ARTrack-384 (Wei et al., 2023)	72.6	85.1	13.2
TATrack-L (Huang et al., 2024)	71.1	85.0	7.2
AQATrack-384 (Xie et al., 2024)	72.7	84.8	43.8
ROMTrack-384 (Cai et al., 2023)	71.4	84.1	28.0
EVPTrack-384 (Shi et al., 2024)	72.7	84.4	29.6
ARTrackV2-L384 (Bai et al., 2024)	73.6	86.1	43.2
LoRAT-g-224 (Lin et al., 2024)	74.9	85.2	41.6
SMTrack-M256 (Ma et al., 2026)	70.1	84.2	36.0
SMTrack-M384 (Ma et al., 2026)	71.9	85.2	34.0
MCITrack-T224 (Kang et al., 2025)	71.7	84.8	50.3
MCITrack-S224 (Kang et al., 2025)	73.8	85.6	40.7
ODTrack-L384 (Zheng et al., 2024)	74.0	86.1	34.8
ETBTrack(ODTrack-L384) [288, 184]	74.0	85.9	44.3
ETBTrack(ODTrack-L384) [288, 92]	72.1	84.7	52.1
ETBTrack(ODTrack-L384) [144, 184]	72.6	85.2	51.5

Ablation Studies

Table 3. Analysis of the sample strategy.

group size g	sample	[32,128]	[128,128]	[128,64]
16	multinomial	62.8	66.1	63.7
32	multinomial	63.2	66.3	64.1
64	multinomial	63.2	66.3	64.2
32	random	61.9	64.7	62.4

Table 4. Analysis of the regularization loss.

β	sync step	[32,128]	[128,128]	[128,64]
10.0	100	63.0	66.0	63.9
10.0	1,000	63.2	66.3	64.1
10.0	10,000	62.7	66.2	63.5
1.0	1,000	61.1	64.6	61.7
100.0	1,000	62.2	65.6	63.1

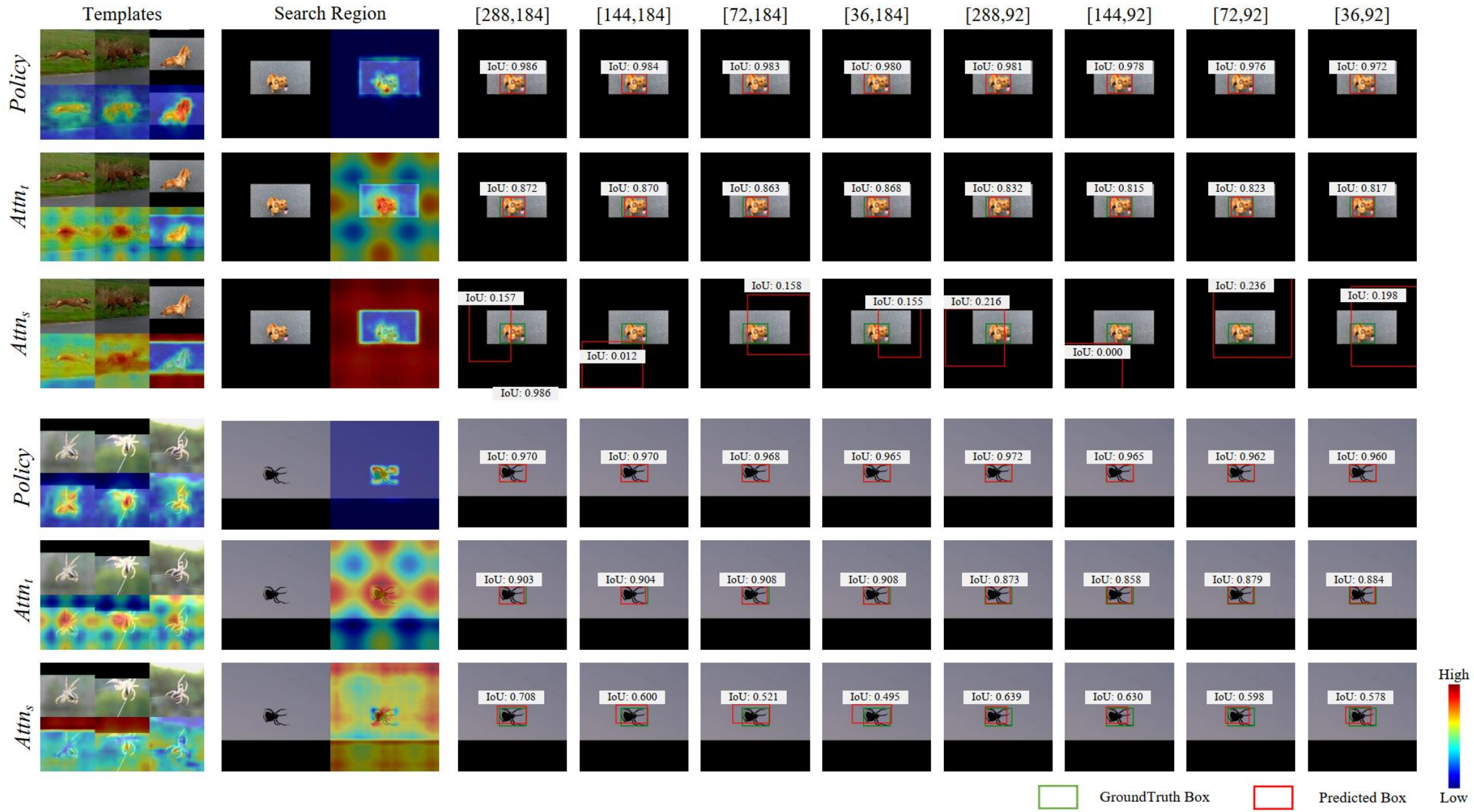
Table 5. Analysis of the policy network position.

layer	prune	[32,128]		[128,128]		[128,64]	
		AUC	GFLOPs	AUC	GFLOPs	AUC	GFLOPs
1	policy	62.6	18.725	65.8	26.206	63.4	21.189
2	policy	63.2	23.032	66.3	29.833	64.1	25.270
4	policy	63.3	31.646	66.3	37.087	64.3	33.431

Table 6. Analysis of the training strategies of elastic inference enhancement.

training strategy	[32,128]	[128,128]	[128,64]
w/o EIE	63.2	66.3	64.1
reduction	66.2	67.8	65.7
alternate	65.0	68.5	66.0
collaborative	65.8	69.2	66.1

Qualitative Results



Thank You!