Learn to Match: Automatic Matching Network Design for Visual Tracking
—— Supplementary Material ——

The supplementary material presents additional details of Sec.4 and Sec.5 in the main manuscript.

- **(Sec. 4) Derivation of the Binary Channel Manipulation.**
  We detail the derivation of Eq.14 $\rightarrow$ Eq.15 in the manuscript.
- **(Sec. 5) Ablation Experiments on Number of Retrained Operators.**
  We provide the ablation experiments on the number of retrained operator in matching networks.
- **(Sec. 5) Search with DARTS [2].**
  We provide the modification of searching a DARTS-like matching cell.

(Sec. 4) Derivation of the Binary Channel Manipulation.

In our paper, we simplify the derivation from Eq.14 to Eq.15 due to the space limit. Here we present the details. We rewrite the Eq.14,

$$y_k = \frac{\exp((\log(\pi_k) + g_k)/\tau)}{\sum_{c=1}^{2} \exp((\log(\pi_c) + g_c)/\tau)}. \tag{1}$$

$k = 1$ for binary case. Substituting $\pi_1 = \sigma(w_i^j)$, $\pi_2 = 1 - \sigma(w_i^j)$, we get,

$$y_1 = \frac{\exp((\log \sigma(w_i^j) + g_1)/\tau)}{\exp((\log \sigma(w_i^j) + g_1)/\tau) + \exp((\log(1 - \sigma(w_i^j)) + g_2)/\tau)} \tag{2}$$

$$= \frac{(\exp(\log \sigma(w_i^j) + g_1))^{\frac{1}{\tau}}}{(\exp(\log \sigma(w_i^j) + g_1))^{\frac{1}{\tau}} + (\exp(\log(1 - \sigma(w_i^j)) + g_2))^{\frac{1}{\tau}}} \tag{3}$$

$$= \frac{(\exp(\log \sigma(w_i^j)) \exp(g_1))^{\frac{1}{\tau}}}{(\exp(\log \sigma(w_i^j)) \exp(g_1))^{\frac{1}{\tau}} + (\exp(\log(1 - \sigma(w_i^j)) \exp(g_2))^{\frac{1}{\tau}}} \tag{4}$$

$$= \frac{(\sigma(w_i^j) \exp(g_1))^{\frac{1}{\tau}}}{(\sigma(w_i^j) \exp(g_1))^{\frac{1}{\tau}} + ((1 - \sigma(w_i^j)) \exp(g_2))^{\frac{1}{\tau}}} \tag{5}$$

$$= \frac{1}{1 + \frac{((1-\sigma(w_i^j)) \exp(g_2))^{\frac{1}{\tau}}}{(\sigma(w_i^j) \exp(g_1))^{\frac{1}{\tau}}}} \tag{6}$$

$$= \frac{1}{1 + (\frac{1-\sigma(w_i^j)}{\sigma(w_i^j)} \exp(g_2 - g_1))^{\frac{1}{\tau}}} \tag{7}$$

$$= \frac{1}{1 + (\frac{1 - \frac{1}{1+\exp(-w_i^j)}}{\frac{1}{1+\exp(-w_i^j)}} \exp(g_2 - g_1))^{\frac{1}{\tau}}} \tag{8}$$

$$= \frac{1}{1 + (\exp(-w_i^j) \exp(g_2 - g_1))^{\frac{1}{\tau}}} \tag{9}$$

$$= \frac{1}{1 + \exp(\frac{-w_i^i + g_2 - g_1}{\tau})} \tag{10}$$

$$= \sigma(\frac{w_i^i + g_1 - g_2}{\tau}) \tag{11}$$

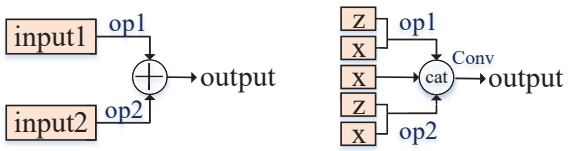
(Sec. 5) Number of Retrained Operators. After training of the search algorithm, we retain two operators for the final matching network. The decision refers to the rules in differentiable neural network search, *i.e.,* each node is assigned with two operators [2]. We conduct ablation experiments to analyze the influence of the retrained operators. As shown in Tab. 1,

when increasing the number of retrained operators, it does not bring additional gains. The conclusion is consistent with that in neural network search.

Table 1: Influence of retrained operators on LaSOT [1].

Number	1	2	3	4	5	6	7
AUC ↑	53.2	58.3	56.7	55.2	54.5	56.1	56.2
Prec. ↑	53.5	59.9	56.9	55.6	55.1	56.4	56.7

(Sec. 5) Search with DARTS [2]. In the section of "NAS-like Matching Cell", we present the ablation results of searching with DARTS. DARTS [2] defines the basic cell as a Directed Acyclic Graph (DAG) (see Fig.7 in the main paper). In the original DARTS, each node (the cyan block in Fig.7 of main paper) links with two output features from previous nodes. After processed by basic operators (*e.g.,* 3×3 convolution), the two inputs of a node are fused by addition (see Fig. 1(a)). However, this mechanism is not suitable for our matching network: 1) The output features of different matching operators have fundamentally different physical meanings. For instance, *Pairwise-Relation* models the global similarity of exemplar and candidate features. Each element in its output represents an affinity score. Nevertheless, the output of *Addition* is the combination of two visual features. The visual feature and similarity feature lie in the different embedding domains. It is thus unreasonable to direct fuse them by addition. In our work, we replace addition with concatenation, and then a 1×1 convolution layer is followed to explore complement of the fused feature. 2) The input of basic operator in DARTS requires only one tensor, yet, we need to feed two tensors into the proposed matching operators, *i.e.,* exemplar and candidate embeddings. The modification is illustrated in Fig. 1. It is worthy to note that we retain the original candidate feature during concatenation. The underlying reason is that involving the visual feature can prevent over-reliance on the relation feature when predicting target locations. Once the relation feature brings ambiguous prediction, *e.g.,* false-positives caused by distractors, the original visual feature with additional information may help to rectify it.



(a) DARTS Node (b) NAS-like Matching Node

Figure 1: Modify DARTS node for Siamese tracking.

References

[1] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5374–5383, 2019. 2

[2] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018. 1, 2