# Appendix

## A Positional Encoding

Transformer requires a positional encoding to identify the position of the current processing token [8]. Through a series of comparison experiments, we choose *untied positional encoding*, which is proposed in TUPE [6], as the positional encoding solution of our tracker. In addition, we generalize the *untied positional encoding* to arbitrary dimensions to fit with other components in our tracker.

The original transformer [8] proposes a absolute positional encoding method to represent the position: a fixed or learnable vector $p_i$ is assigned to each position $i$. Starting from the basic attention module, we have:

$$\mathrm{Atten}(Q,K,V) = \mathrm{softmax}\Big(\frac{QK^T}{\sqrt{d_k}}V\Big), \tag{1}$$

where $Q$,$K$,$V$ are the $query$ vector, $key$ vector and $value$ vector, which are the parameters of the attention function, $d_k$ is the dimension of $key$. Introducing the linear projection matrix and multi-head attention to the attention module (1), we get the multi-head variant defined in [8]:

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1, ..., \mathrm{head_h})W_O, \tag{2}$$

where $\mathrm{head_i} = \mathrm{Atten}(QW_i^Q, KW_i^K, VW_i^V)$, $W_i^Q \in \mathbb{R}^{d_{\mathrm{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\mathrm{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\mathrm{model}} \times d_v}$, $W_i^O \in \mathbb{R}^{hd_v \times d_{\mathrm{model}}}$ and $h$ is the number of heads. For simplicity, as in [6], we assume that $d_k = d_v = d_{\mathrm{model}}$, and use the single-head version of self-attention module. Denoting the input sequence as $x = x_1, x_2, \ldots, x_n$, where $n$ is the length of sequence, $x_i$ is the $i$-th token in the input data. Denoting the output sequence as $z = (z_1, z_2, \ldots, z_n)$. Self-attention module can be rewritten as

$$z_i = \sum_{j=1}^{n} \frac{\exp(\alpha_{ij})}{\sum_{j'=1}^{n} \exp(\alpha_{ij'})}(x_j W^V), \tag{3}$$

$$\text{where } \alpha_{ij} = \frac{1}{\sqrt{d}}(x_i W^Q)(x_j W^K)^T. \tag{4}$$

Obviously, the self-attention module is permutation-invariance. Thus it can not "understand" the order of input tokens.

**Untied absolute positional encoding.** By adding a learnable positional encoding [8] to the single-head self-attention module, we can obtain the following equation:

$$\begin{aligned} \alpha_{ij}^{Abs} &= \frac{((w_i + p_i)W^Q)((w_j + p_j)W^K)^T}{\sqrt{d}} \\ &= \frac{(w_i W^Q)(w_j W^K)^T}{\sqrt{d}} + \frac{(w_i W^Q)(p_j W^K)^T}{\sqrt{d}} \\ &+ \frac{(p_i W^Q)(w_j W^K)^T}{\sqrt{d}} + \frac{(p_i W^Q)(p_j W^K)^T}{\sqrt{d}}. \end{aligned} \tag{5}$$

The equation (5) is expanded into four terms: token-to-token, token-to-position, position-to-token, position-to-position. [6] discuss the problems exists in the equation and proposes the *untied absolute positional encoding*, which unties the correlation between tokens and positions by removing the token-position correlation terms in equation (5), and using an isolated pair of projection matrices $U^Q$ and $U^K$ to perform linear transformation upon positional embedding vector. The following is the new formula for obtaining $\alpha_{ij}$ using the *untied absolute positional encoding* in the $l$-th layer:

$$\begin{aligned} \alpha_{ij} &= \frac{1}{\sqrt{2d}}(x_i^l W^{Q,l})(x_j^l W^{K,l})^T \\ &+ \frac{1}{\sqrt{2d}}(p_i U^Q)(p_j U^K)^T. \end{aligned} \tag{6}$$

where $p_i$ and $p_j$ is the positional embedding at position $i$ and $j$ respectively, $U^Q \in \mathbb{R}^{d \times d}$ and $U^K \in \mathbb{R}^{d \times d}$ are learnable projection matrices for the positional embedding vector. When extending to the multi-head version, the positional embedding $p_i$ is shared across different heads, while $U^Q$ and $U^K$ are different for each head.

**Relative positional bias.** According to [7], relative positional encoding is a necessary supplement to absolute positional encoding. In [6], a relative positional encoding is applied by adding a relative positional bias to equation (6):

$$\begin{aligned}\alpha_{ij} &= \frac{1}{\sqrt{2d}}(x_i^l W^{Q,l})(x_j^l W^{K,l})^T \\ &+ \frac{1}{\sqrt{2d}}(p_i U^Q)(p_j U^K)^T + b_{j-i},\end{aligned} \tag{7}$$

where for each $j-i$, $b_{j-i}$ is a learnable scalar. The *relative positional bias* is also shared across layers. When extending to the multi-head version, $b_{j-i}$ is different for each head.

**Generalize to multiple dimensions.** Before working with our tracker's encoder and decoder network, we need to extend the *untied positional encoding* to a multidimensional version. One straightforward method is allocating a positional embedding matrix for every dimension and summing up all embedding vectors from different dimensions at the corresponding index to represent the final embedding vector. Together with *relative positional bias*, for an n-dimensional case, we have:

$$\begin{aligned}\alpha_{\underbrace{ij\ldots}_{\mathrm{n}},\underbrace{mn\ldots}_{\mathrm{n}}} &= \frac{1}{\sqrt{2d}}(x_{\underbrace{ij\ldots}_{\mathrm{n}}} W^Q)(x_{\underbrace{mn\ldots}_{\mathrm{n}}} W^K)^T \\ &+ \frac{1}{\sqrt{2d}}[\underbrace{(p_i^1 + p_j^2 + \ldots)}_{\mathtt{n}} U^Q][\underbrace{(p_m^1 + p_n^2 + \ldots)}_{\mathtt{n}} U^K]^T \\ &+ b_{\underbrace{m-i, n-j, \ldots}_{\mathrm{n}}}.\end{aligned} \tag{8}$$

**Generalize to concatenation-based fusion.** In order to work with *concatenation-based fusion*, the *untied absolute positional encoding* is also concatenated to match the real position, the indexing tuple of *relative positional bias* now appends with a pair of indices to reflect the origination of $query$ and $key$ involved currently.

Taking $l$-th layer in the encoder as the example:

$$\begin{aligned}\alpha_{ij,mn,g,h} &= \frac{1}{\sqrt{2d}}(x_{ij,g}^l W^{Q,l})(x_{mn,h}^l W^{K,l})^T \\ &+ \frac{1}{\sqrt{2d}}[(p_{i,g}^1 + p_{j,g}^2)U_g^Q][(p_{m,h}^1 + p_{n,h}^2)U_h^K]^T \\ &+ b_{m-i,n-j,g,h}\ ,\end{aligned} \tag{9}$$

where $g$ and $h$ are the index of the origination of $query$ and $key$ respectively, for instance, 1 for the tokens from the template image, 2 for the tokens from the search image. The form in the decoder is similar, except that $g$ is fixed. In our implementation, the parameters of *untied positional encoding* are shared inside the encoder and the decoder, respectively.

## B Figures on LaSOT Test set

Fig. 1 and Fig. 2 show the success plot and the precision plot respectively. The comparison includes our SwinTrack-T-224, our SwinTrack-B-384, DiMP [1], STMTrack[5], TiDiMP[9], TransT[2] and STARK[10].

## C Response Visualization

We provide the heatmap visualization of the response map generated by the IoU-aware classification branch head in our SwinTrack-B-384 in Fig. 3. The visualized sequences are from $\text{LaSOT}_{\text{ext}}$ [3], with challenges include fast motion, full occlusion, hard distractor, *etc*. The results demonstrate the great discriminative power of our tracker. Many trackers will show a multi-peak on the response map when the object is occluded or multiple similar objects exist. With the vision-motion integrated Transformer architecture, our tracker eases such phenomenon.

# References


[1] Bhat, G., Danelljan, M., Gool, L.V., Timofte, R., 2019. Learning discriminative model prediction for tracking, in: ICCV.

[2] Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H., 2021. Transformer tracking, in: CVPR.

[3] Fan, H., Bai, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Huang, M., Liu, J., Xu, Y., et al., 2021. Lasot: A high-quality large-scale single object tracking benchmark. IJCV 129, 439–461.

[4] Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H., 2019. Lasot: A high-quality benchmark for large-scale single object tracking, in: CVPR.

[5] Fu, Z., Liu, Q., Fu, Z., Wang, Y., 2021. Stmtrack: Template-free visual tracking with space-time memory networks, in: CVPR.

[6] Ke, G., He, D., Liu, T.Y., 2021. Rethinking positional encoding in language pre-training, in: International Conference on Learning Representations. URL: `https://openreview.net/forum?id=09-528y2Fgf`.

[7] Shaw, P., Uszkoreit, J., Vaswani, A., 2018. Self-attention with relative position representations. arXiv .

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: NeurIPS.

[9] Wang, N., Zhou, W., Wang, J., Li, H., 2021. Transformer meets tracker: Exploiting temporal context for robust visual tracking, in: CVPR.

[10] Yan, B., Peng, H., Fu, J., Wang, D., Lu, H., 2021. Learning spatio-temporal transformer for visual tracking, in: ICCV.

Figure 1: Comparison with state-of-the-art trackers on LaSOT [4] Test set using success (SUC) AUC score.

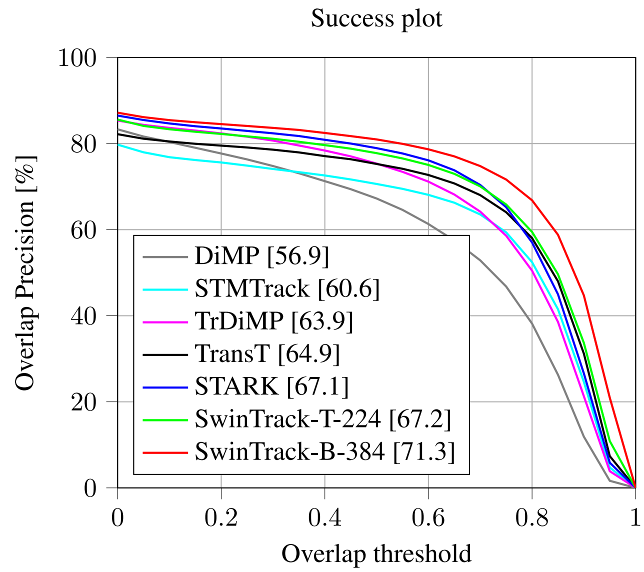


Figure 2: Comparison with state-of-the-art trackers on LaSOT [4] Test set using precision (PRE) AUC score.

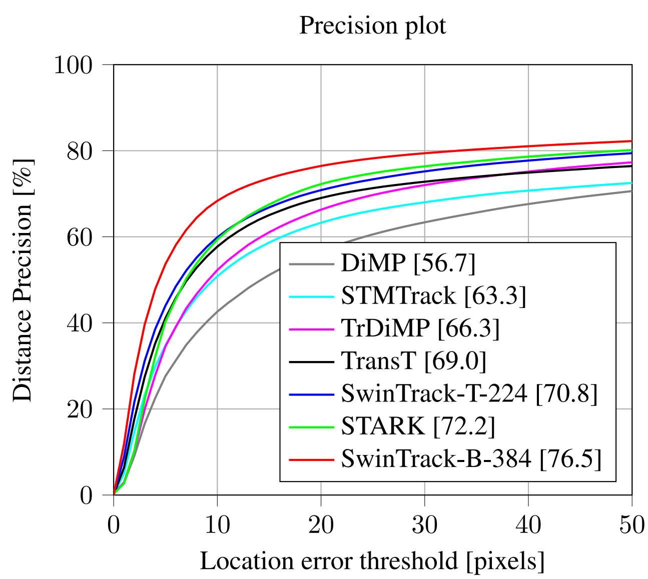

Figure 3: Heatmap visualization of the tracking response map of our SwinTrack-B-384 on $\text{LaSOT}_{\text{ext}}$ [3]. The odd rows visualize the search region patches with ground-truth bounding box (in red rectangles). The even rows visualize the search region patches blended with the heatmap visualization of the response map.