Supplementary Materials of "iKUN: Speak to Trackers without Retraining"

Yunhao Du[1], Cheng Lei[1], Zhicheng Zhao[1,2,3*], Fei Su[1,2,3]
[1]The school of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2]Beijing Key Laboratory of Network System and Network Culture, China
[3]Key Laboratory of Interactive Technology and Experience System Ministry of Culture and Tourism, Beijing, China
{dyh_bupt,mr.leicheng,zhaozc,sufei}@bupt.edu.cn

A. Details of KUM

Table 1. **Notations.**

Variable	explanation	defaults
B	batch size	8
T	temporal window size	8
L	description length	10
n	number of cross attention heads	4
k	convolution kernel size	1
p	dropout ratio	0.1
C_v	visual feature dimension	2048
C_t	textual feature dimension	1024
h, w	spatial size of local feature map	7,7
H, W	spatial size of global feature map	21,21

We detail the three designs of KUM in this section. The input of KUM includes visual global feature $f_{global} \in \mathbb{R}^{BT\times HW\times C_v}$, visual local feature $f_{local} \in \mathbb{R}^{BT\times hw\times C_v}$, and textual feature $f_t \in \mathbb{R}^{B\times 1\times C_t}$ (or $f'_t \in \mathbb{R}^{B\times L\times C_t}$ before squeezing). Then it outputs the unified feature $f_{uni} \in \mathbb{R}^{BT\times C_v}$. The notations are summarized in Table 1.

Cascade Attention. The cascade attention design is mainly implemented by two hierarchical cross attention layers with n heads. The first cross attention is utilized to aggregate the two visual features with residual adding by:

$$f_{vis} = f_{local} + CrossAtt(f_{local}, f_{global}, f_{global}) \in \mathbb{R}^{BT\times hw\times C_v}, \quad (1)$$

where f_{local} is taken as query and f_{global} is the key and value. The hidden embedding dimension is set to C_v. Then f'_t is repeated T times and transformed by a linear layer, resulting in $f_{txt} \in \mathbb{R}^{BT\times L\times C_v}$. f_{vis} and f_{txt} are aggregated by the second cross attention layer with residual multiplication by:

$$f'_{uni} = f_{local} * CrossAtt(f_{local}, f_{txt}, f_{txt}) \in \mathbb{R}^{BT\times hw\times C_v}. \quad (2)$$

Finally, a spatial global average pooling layer is applied to output the final unified feature by:

$$f_{uni} = AvgPool(f'_{uni}) \in \mathbb{R}^{BT\times C_v}. \quad (3)$$

Cross correlation. The cross correlation design mainly consists of a cross attention operation and a dynamic convolution operation. The first step is the same as "cascade attention" by:

$$f_{vis} = f_{local} + CrossAtt(f_{local}, f_{global}, f_{global}) \in \mathbb{R}^{BT\times hw\times C_v}, \quad (4)$$

Then f_t is repeated T times, and then processed by a convolution layer with kernel size k and a batch normalization layer to estimate the *dynamic kernel* [1] by:

$$f_{kernel} = BN(Conv(Repeat(f_t))) \in \mathbb{R}^{BT\times 1\times C_v}. \quad (5)$$

Afterwards, the cross correlation is performed on between f_{vis} and f_{kernel}. It is implemented by a depth-wise convolution by:

$$f_{cross} = DepthConv(f_{vis}; f_{kernel}) \in \mathbb{R}^{BT\times hw\times C_v}, \quad (6)$$

where f_{vis} is the input feature map and f_{kernel} is taken as the convolution kernels. Then a dropout layer with a ratio of p with residual adding is applied by:

$$f_{drop} = f_{cross} + Drop(f_{cross}) \in \mathbb{R}^{BT\times hw\times C_v}. \quad (7)$$

Another convolution layer with kernel size k and a batch normalization layer are followed by:

$$f'_{uni} = BN(Conv(f_{drop})) \in \mathbb{R}^{BT\times hw\times C_v}. \quad (8)$$

*Corresponding author

Finally, a spatial global average pooling layer is applied to output the final unified feature by:

$$f_{uni} = AvgPool(f'_{uni}) \in \mathbb{R}^{BT \times C_v}. \quad (9)$$

Text-first modulation. The text-first modulation design mainly includes a modulation operation and a cross attention operation. The f_t is first repeated T times, and then processed by a convolution layer with kernel size k and a batch normalization layer by:

$$f_r = BN(Conv(Repeat(f_t))) \in \mathbb{R}^{BT \times 1 \times C_v}. \quad (10)$$

Then it is repeated $H \times W$ times to modulate f_{global} with dropout of ratio p and residual multiplication by:

$$f'_{global} = f_{global} * Drop(Repeat(f_r)) \in \mathbb{R}^{BT \times HW \times C_v}. \quad (11)$$

The similar operation is utilized to modulate f_{local} by:

$$f'_{local} = f_{local} * Drop(Repeat(f_r)) \in \mathbb{R}^{BT \times hw \times C_v}. \quad (12)$$

Afterwards, a cross attention layer with residual adding is followed by:

$$f'_{uni} = f_{local} + CrossAtt(f'_{local}, f'_{global}, f_{global}) \in \mathbb{R}^{BT \times hw \times C_v}, \quad (13)$$

where f'_{local} is query, f'_{global} is key and the raw f_{global} is value. Finally, a spatial global average pooling layer is applied to output the final unified feature by:

$$f_{uni} = AvgPool(f'_{uni}) \in \mathbb{R}^{BT \times C_v}. \quad (14)$$

B. Refer-Dance

We introduce the details of Refer-Dance in this section. Refer-Dance is extended from DanceTrack [2] by adding description annotations and follows the same data split protocol. That is, 40 videos with 39 distinct descriptions are used for training and 25 videos with 17 distinct descriptions are used for testing. The dataset follows the open-set setting, in which test descriptions don't necessarily appear in the training set. The long-tail distribution of textual descriptions is visualized in Fig.1. Refer to Fig.2 for some visualization examples.

References

[1] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016. 1

[2] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20993–21002, 2022. 2

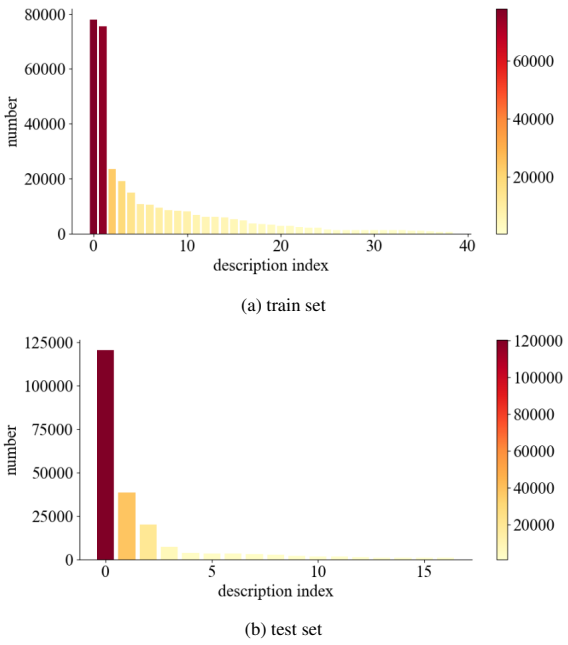

Figure 1. **The distribution of descriptions on Refer-Dance.**

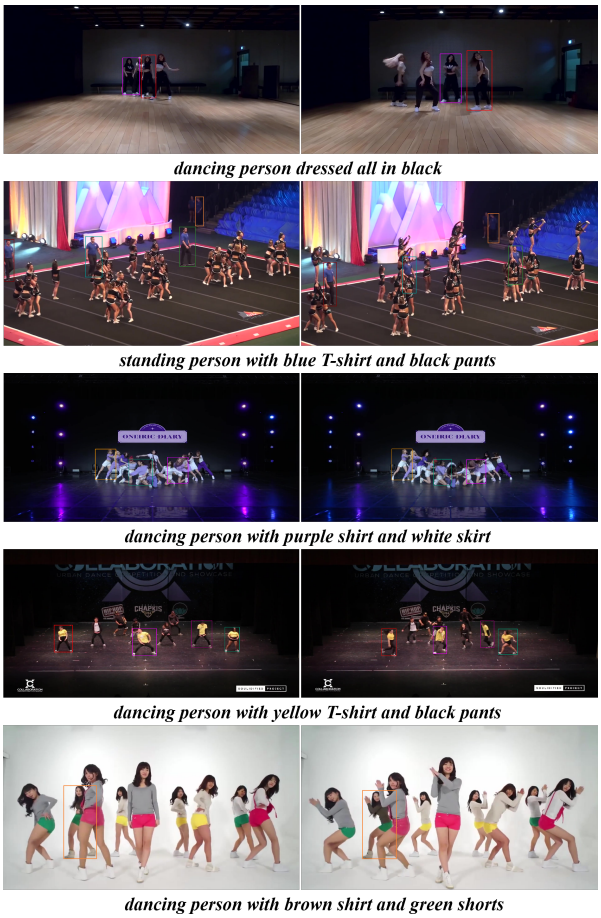

Figure 2. **Visualization examples of Refer-Dance.**