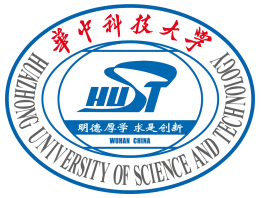


Focus On Details: Online Multi-object Tracking with Diverse Fine-grained Representation

Hao Ren[1], Shoudong Han[1,*], Huilin Ding, Ziwen Zhang, Hongwei Wang, Faquan Wang

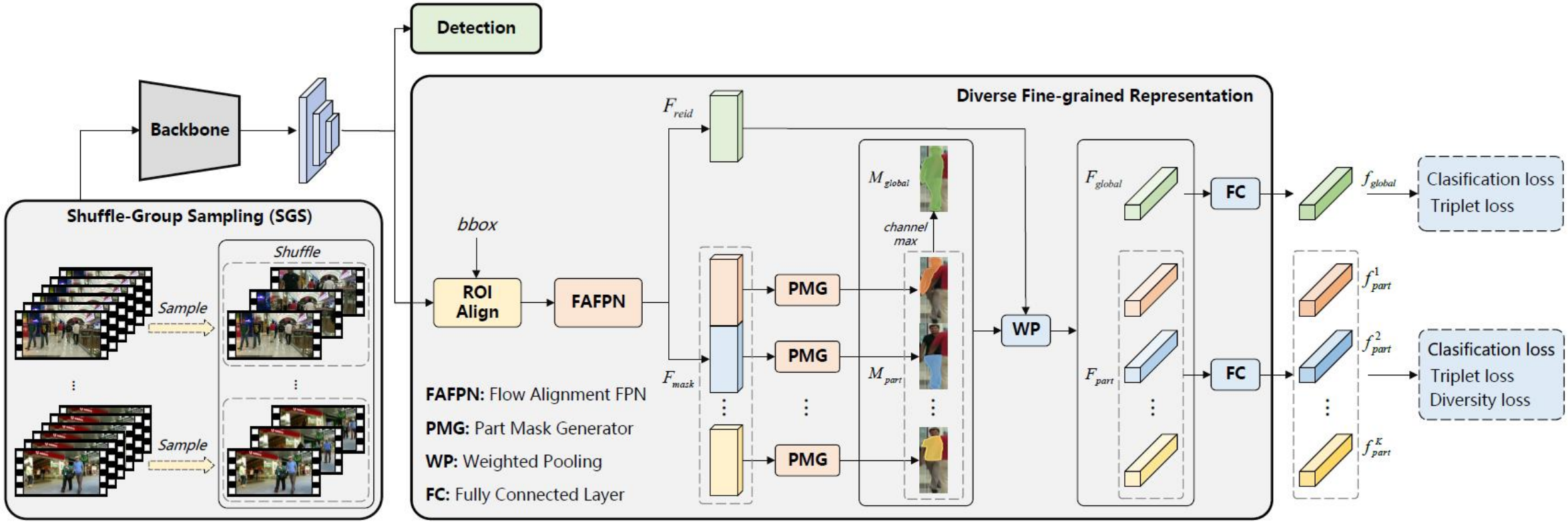
National Key Laboratory of Science and Technology on Multispectral Information Processing,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology

{haoren2000, shoudonghan}@hust.edu.cn

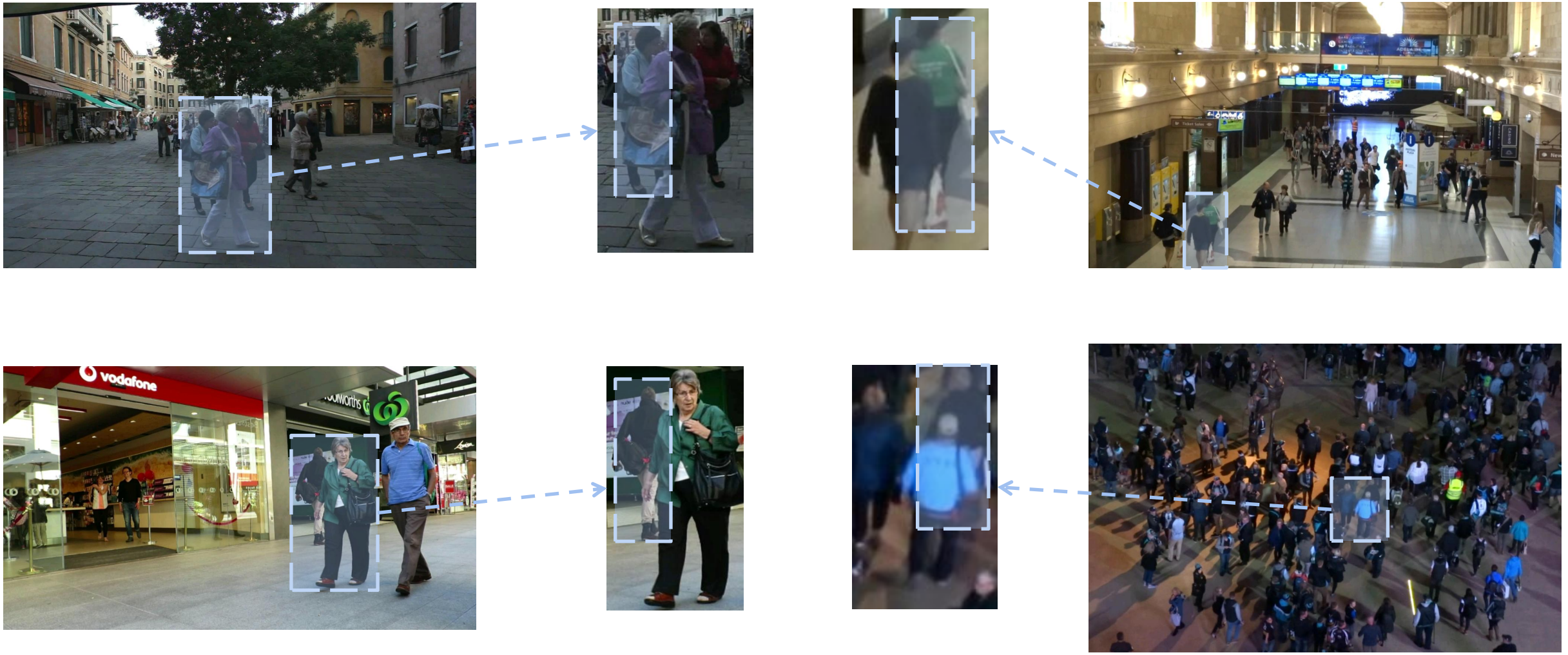
TAG：WED-AM-291

[1] Equal contribution
[*] Corresponding author

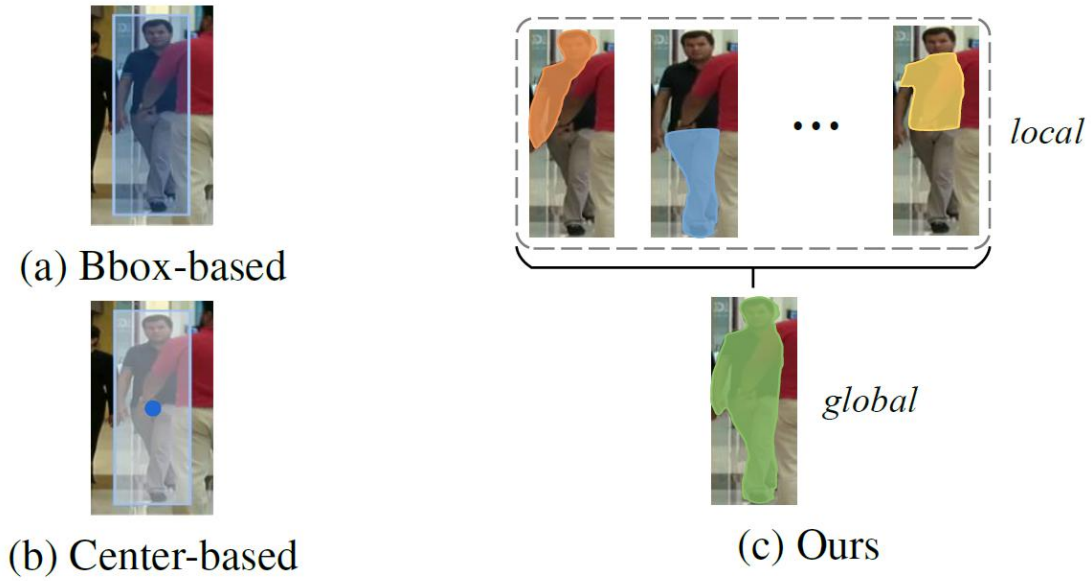
Preview of Our Work (FineTrack)



Background



Motivation



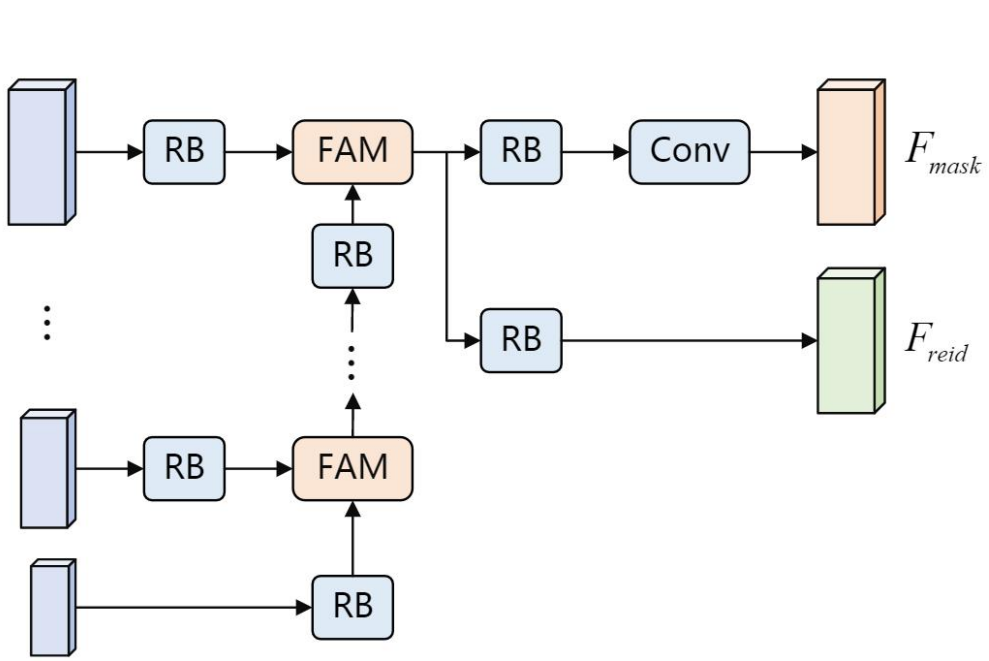
(a) Bbox-based

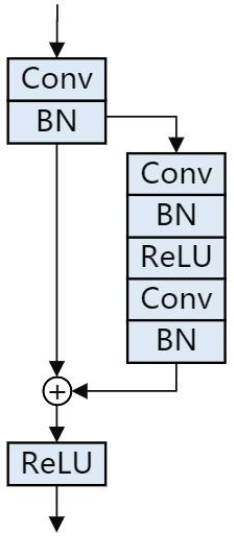
(b) Center-based

(c) Ours

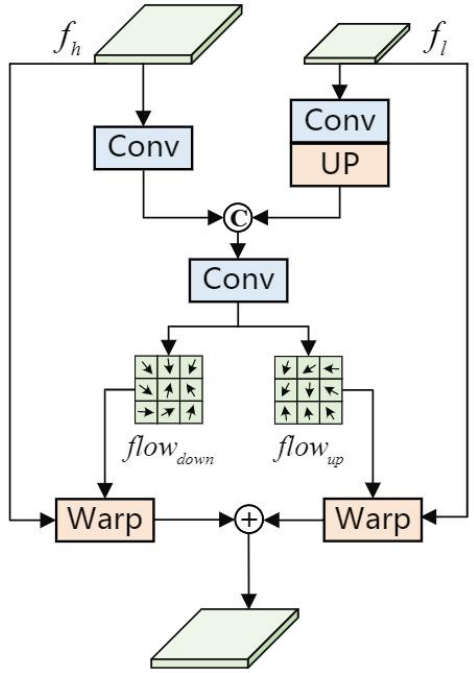
Our method focuses on different local **details** of targets, which is illustrated in (c). **Fine-grained** global and local representations complement each other and jointly describe appearance. When the target is **occluded**, our method can still identify it according to visible parts, similar to human judgment.

Flow Alignment FPN (FAFPN)





(a) RB

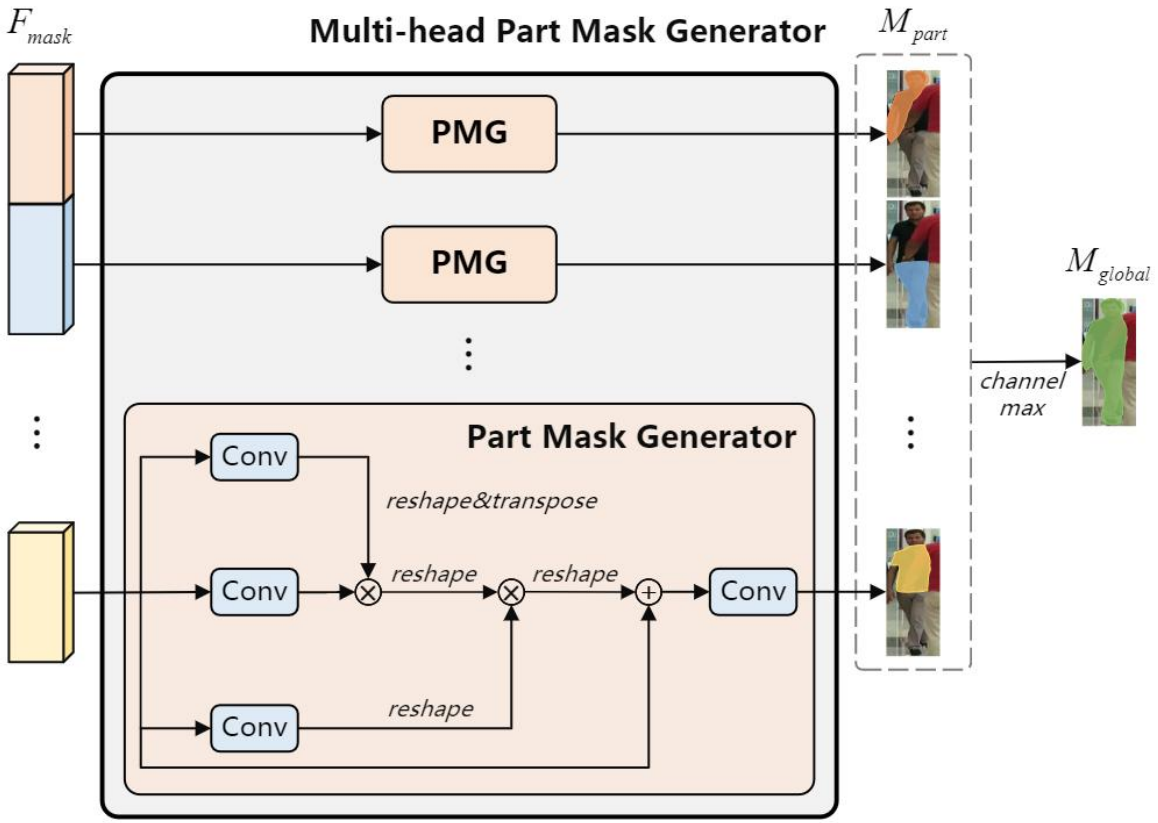


(b) FAM

To obtain **fine-grained features**, we employ the *Flow Alignment Module* (FAM) to generate semantic flow among feature maps of different resolutions. The semantic flow can guide alignment and eliminate spatial dislocation among feature maps from different scales. Furthermore, we utilize the FAM to optimize the aggregation process of FPN and then construct a ***Flow Alignment FPN*** **(FAFPN)**.

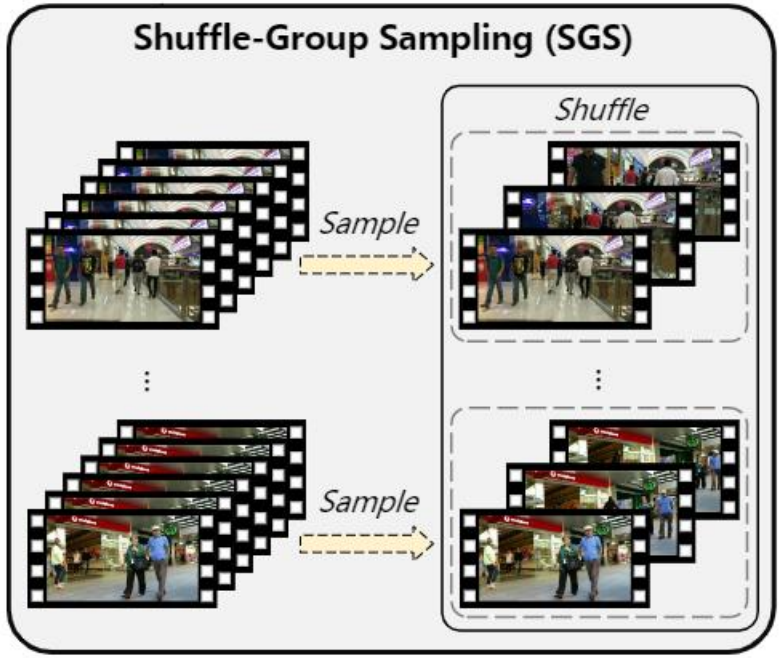
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Multi-head Part Mask Generator (MPMG)



We divide the input feature maps into blocks along their channel dimension, and each block can be considered as a different mapping of the target feature. These feature maps are fed into ***Multi-head Part Mask Generator*** **(MPMG)** to generate different part masks of the target without Ground Truth.

Shuffle-Group Sampling (SGS)



We construct the ***Shuffle-Group Sampling*** **(SGS)** training strategy. Different from random sampling, SGS adopts **sequential** sampling to group video frames and **disrupts** the order of grouped data to reduce convergence fluctuations. In this way, targets in the same batch hold positive samples with the same identity, thus alleviating the problem of **imbalanced positive and negative samples** caused by random sampling.

Training Loss

Due to the peculiarities of multi-branch structures, using only classification loss and Triplet loss does not ensure that the model focuses on different parts of the target. To avoid multiple branches gazing at similar details, we employ the **diversity loss** to distance different part features of the same target:

$$L_{div} = \frac{1}{N \cdot K(K-1)} \sum_{n=1}^{N} \sum_{k_i \neq k_j}^{K} \frac{\left\langle f_p^{n,k_i}, f_p^{n,k_j} \right\rangle}{\left\| f_p^{n,k_i} \right\|_2 \cdot \left\| f_p^{n,k_j} \right\|_2}$$

The purpose of diversity loss is intuitive, which is to keep the cosine similarity between different part features of the same target as low as possible. We combine the above losses into a final training loss:

$$L = \alpha \cdot (L_{cls}^{p} + L_{tri}^{p}) + \beta \cdot (L_{cls}^{g} + L_{tri}^{g}) + \gamma \cdot L_{div}$$

where α, β and γ are used to adjust the proportion of different losses and we set α = 0.5 ∗ K, β = 0.3, γ = 2.

Inference

Based on ByteTrack, we add a method similar to DeepSort that calculates Re-ID features into feature distance matrix. It is worth mentioning that we concatenate part features of targets with global features as Re-ID features.

$$d_{feat} = 1 - Similarity(\tilde{f}^{t-1}, f^{t})$$

$$d_{IoU} = 1 - IoU(b_{det}, b_{pre})$$

$$\tilde{d}_{feat} = 1 - (1 - d_{feat}) \cdot (d_{IoU} < 1)$$

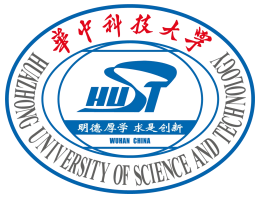
$$d = \sqrt{\tilde{d}_{feat} \cdot d_{IoU}}$$

Comparison with the State-of-the-art Methods

Method	HOTA↑	IDF1↑	MOTA↑	FP↓	FN↓	IDs↓
MOT17 private detection						
DAN [34]	39.3	49.5	52.4	25423	234592	8431
TubeTK [24]	48.0	58.6	63.0	27060	177483	4137
MOTR [51]	-	66.4	65.1	45486	149307	2049
CTracker [26]	49.0	57.4	66.6	22284	160491	5529
MAT [12]	53.8	63.1	69.5	30660	138741	2844
QuasiDense [25]	53.9	66.3	68.7	26589	146643	3378
TransTrack [33]	54.1	63.5	75.2	50157	86442	3603
TransCenter [46]	54.5	62.2	73.2	23112	123738	4614
GSDT [41]	55.2	66.5	73.2	26397	120666	3891
PermaTrackPr [36]	55.5	68.9	73.8	28998	115104	3699
SOTMOT [56]	-	71.9	71.0	39537	118983	5184
FUFET [29]	57.9	68.0	76.2	32796	98475	3237
MTrack [48]	-	72.1	73.5	53361	101844	2028
FairMOT [55]	59.3	72.3	73.7	27507	117477	3303
CSTrack [20]	59.3	72.6	74.9	23847	114303	3567
SiamMOT [31]	-	72.3	76.3	-	-	-
ReMOT [47]	59.7	72.0	77.0	33204	93612	2853
Semi-TCL [18]	59.8	73.2	73.3	22944	124980	2790
CorrTracker [39]	60.7	73.6	76.5	29808	99510	3369
RelationTrack [49]	61.0	74.7	73.8	27999	118623	1374
TransMOT [6]	61.7	75.1	76.7	36231	93150	2346
ByteTrack [54]	63.1	77.3	**80.3**	25491	**83721**	2196
FineTrack	**64.3**	**79.5**	80.0	**21750**	90096	**1272**

Method	HOTA↑	IDF1↑	MOTA↑	FP↓	FN↓	IDs↓
MOT20 private detection						
MLT [53]	43.2	54.6	48.9	45660	216803	2187
TransTrack [33]	48.5	59.4	65.0	27197	150197	3608
FairMOT [55]	54.6	67.3	61.8	103440	88901	5243
Semi-TCL [18]	55.3	70.1	65.2	61209	114709	4139
CSTrack [20]	54.0	68.6	66.6	25404	144358	3196
GSDT [41]	53.6	67.5	67.1	31913	135409	3131
SiamMOT [31]	-	69.1	67.1	-	-	-
RelationTrack [49]	56.5	70.5	67.2	61134	104597	4243
SOTMOT [56]	-	71.4	68.6	57064	101154	4209
ByteTrack [54]	61.3	75.2	77.8	26249	**87594**	1223
FineTrack	**63.6**	**79.0**	**77.9**	**24439**	89012	**980**

Method	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
CenterTrack [57]	41.8	35.7	86.8	22.6	**78.1**
FairMOT [55]	39.7	40.8	82.2	23.8	66.7
QuasiDense [25]	45.7	44.8	83.0	29.2	72.1
TransTrack [33]	45.5	45.2	88.4	27.5	75.9
TraDes [45]	43.3	41.2	86.2	25.4	74.5
ByteTrack [54]	47.7	53.9	89.6	32.1	71.0
FineTrack	**52.7**	**59.8**	**89.9**	**38.5**	72.4



Thanks for listening