Delving into the Trajectory Long-tail Distribution for Muti-object Tracking

Sijia Chen * En Yu * Jinyang Li Wenbing Tao †
Huazhong University of Science and Technology, China
{sijiachen, yuen, jinyangli, wenbingtao}@hust.edu.cn

Abstract

Multiple Object Tracking (MOT) is a critical area within computer vision, with a broad spectrum of practical implementations. Current research has primarily focused on the development of tracking algorithms and enhancement of post-processing techniques. Yet, there has been a lack of thorough examination concerning the nature of tracking data it self. In this study, we pioneer an exploration into the distribution patterns of tracking data and identify a pronounced long-tail distribution issue within existing MOT datasets. We note a significant imbalance in the distribution of trajectory lengths across different pedestrians, a phenomenon we refer to as "pedestrians trajectory long-tail distribution". Addressing this challenge, we introduce a bespoke strategy designed to mitigate the effects of this skewed distribution. Specifically, we propose two data augmentation strategies, including Stationary Camera View Data Augmentation (SVA) and Dynamic Camera View Data Augmentation (DVA) , designed for viewpoint states and the Group Softmax (GS) module for Re-ID. SVA is to backtrack and predict the pedestrian trajectory of tail classes, and DVA is to use diffusion model to change the background of the scene. GS divides the pedestrians into unrelated groups and performs softmax operation on each group individually. Our proposed strategies can be integrated into numerous existing tracking systems, and extensive experimentation validates the efficacy of our method in reducing the influence of long-tail distribution on multi-object tracking performance. The code is available at https://github.com/chen-si-jia/Trajectory-Long-tail-Distribution-for-MOT.

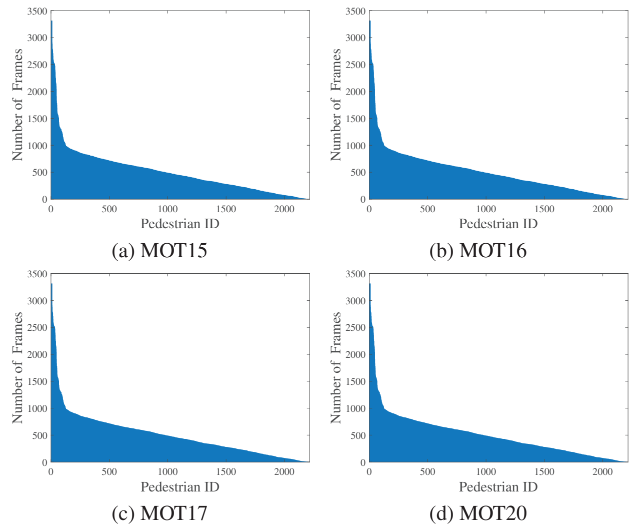

Figure 1. The number of frames of pedestrians with different identities in the MOTChallenge datasets. We stipulate that different pedestrian identities are regarded as different pedestrian classes.

1. Introduction

Multi-object tracking stands as one of the fundamental and most challenging tasks in computer vision. This technology involves the tracking of multiple objects of interest in video sequences, providing essential parameters such as location, trajectory and velocity. Multi-object tracking has found widespread applications in domains like autonomous driving, video analysis and smart transportation [2].

The prior research efforts in the field of multi-object tracking have primarily focused on the designs of tracking networks and post-processing strategies. However, the current MOT methods do not pay attention to the pedestrian long-tail characteristics of tracking data. Thus, we conducted an analysis experiment to count the number of frames of pedestrians with different identities in the MOTChallenge datasets. As shown in Fig. 1, we observe a large difference in the number of frames for different pedestrian identities. Because the characteristic of the long-tail distribution is that the head classes possess a substantial number of instances and the tail classes has only a few instances, we conclude that the number of pedestrian identities obeys the long-tail distribution.

There is a common problem with long-tail distribution datasets: The network is trained on long-tail distribution data often results in a bias towards learning features associated with the prevalent head classes, while neglecting those

*Equal contribution
†Corresponding author

in the less represented tail classes. At present, the problem of improving long-tail distribution can be divided into three aspects: class re-balancing, information augmentation and module improvement. In the image data collected by the camera, some people stay in the image for a long time, and some people move across the image in a hurry. Due to the reasons of the data itself, the network will learn less features of people who have hurriedly passed by. For the current Re-ID branch of multi-object trackers, most of them regard Re-ID as a classification problem and use the softmax module to calculate the classification probability. However, the softmax module has a huge flaw: the weights of classes with large weights become larger, and the weights of classes with small weights become smaller, which will intensify the long-tail distribution effect on the long-tail distribution data. Hence, to improve the issue, we propose our solution from two key perspectives: information augmentation and module improvement.

In the perspective of information augmentation, we classify camera data into two categories: stationary camera view data and dynamic camera view data, based on the motion status of the camera. For stationary camera view data, we propose the Stationary Camera View Data Augmentation (SVA) strategy that encompasses two techniques: backtracking continuation and prediction continuation. The backtracking continuation is applied to the pedestrians of tail classes in the middle frame of the training sequence data, while the prediction continuation is employed for the pedestrians of tail classes in the final frame of the training sequence data. This strategy can promote the network's learning of pedestrian trajectories in the tail classes. For dynamic camera view data, the Dynamic Camera View Data Augmentation (DVA) strategy is proposed. This strategy uses the diffusion model to perform style transformation on the scene background, improving the network's attention to the features of pedestrians areas.

In terms of module enhancement, we devise the Group Softmax (GS) module. The GS groups pedestrians with a similar number of training samples together, and then computes the softmax and cross-entropy loss for each group individually, preventing a significant suppression of tail classes by the weights of head classes, improving the network's ability to extract appearance features of tail classes.

We apply our tailored solution to the SOTA FairMOT [58] and CSTrack [30] of multi-object tracking networks and evaluate them on four public MOT benchmarks, ie., MOT15 [27], MOT16 [34], MOT17 [34] and MOT20 [12]. The experimental results clearly demonstrate that our approach appear to significant improvements. The performance of our network trained using only MOT20 data far exceeds the performance of the baseline trained using mixed data on the MOT20 test set. For example, FairMOT using our strategy is only trained on the MOT20 data, which is better than FairMOT trained using the mixed data. On the MOT20 test set, it increased 4.1% MOTA and 3.0% IDF1.

The main contributions of this work are as follows:

1. We serve as the first to discover the long-tail distribution problem in multi-object tracking and point out that this problem is caused by the imbalance of the number of frames for different pedestrians.
2. We propose the tailored data augmentation strategies, including SVA and DVA, from information augmentation perspective. SVA is used to backtrack and predict the pedestrians trajectory of tail classes, and DVA is used to change the background of the scene. Additionally, we design the GS module from module improvement perspective. The GS divides pedestrians with different identities into unrelated groups and performs a separate softmax operation on each group.
3. We apply our method to two SOTA Joint Detection and Tracking algorithms and evaluate on MOTChallenge datasets. The evaluation results showcase robust performance improvements, serving as a compelling validation of the efficacy of our method.

2. Related Work

Multiple Object Tracking. We review three main multiple object tracking frameworks: Tracking-by-Detection, Joint Detection and Tracking, and Transformer-based Tracking. Tracking-by-Detection [6, 15, 18, 20, 44, 59] primarily consists of two main components. In the detection phase, a detector is established to locate objects of interest. In the association phase, early methods used motion predictors to forecast the positions of objects in the next frame and relied on positional information to associate objects across consecutive frames. Joint Detection and Tracking[3, 30, 41, 43, 50, 58, 66], a unified network simultaneously produces detection results and the corresponding appearance features of the detected objects. Subsequently, association methods are employed to link objects between consecutive frames. Transformer-based Tracking [33, 39, 51, 52, 54, 60, 67] uses a transformer to combine detection queries with queries derived from the previous frame predictions for detecting and tracking objects in the current frame. This approach eliminates the need for post-processing steps, enabling end-to-end multi-object tracking.

Long-tail Distribution. Improving long-tail distribution issues can be categorized into three aspects: class re-balancing, information augmentation, and module improvement. Class re-balancing including resampling, cost-sensitive learning, and Logit adjustment. Resampling [17, 24, 35, 57] is one of the most widely used methods for addressing class imbalance. Cost-sensitive learning [7, 8, 11, 47] involves adjusting the loss weights of different classes. Logit adjustment [40, 42, 61] is a post-hoc technique that shifts the model's logits based on label frequen-

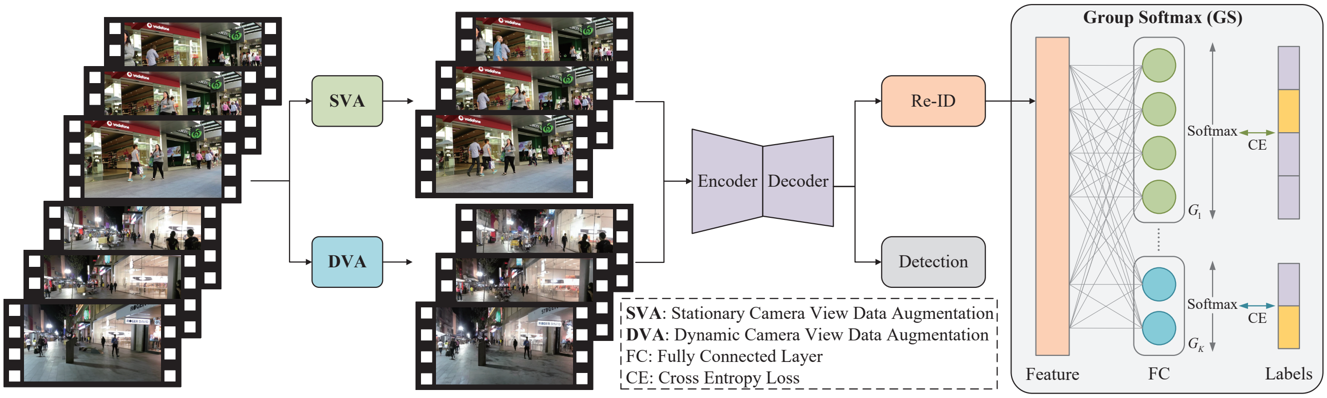

Figure 2. Overall pipeline of our strategies. Our strategies comprise 3 modules: (1) SVA: To backtrack and predict the pedestrians trajectory of tail classes. (2) DVA: To use diffusion model to change the background of the scene. (3) GS: To divide the pedestrians with different identities into unrelated groups and perform softmax operation on each group individually.

cies. Information augmentation including transfer learning and data augmentation. transfer learning aims to transfer knowledge from a source domain to enhance model training in a target domain. Transfer learning [21, 32, 45, 48, 69] primarily includes four approaches: head-tail knowledge transfer, model pre-training, knowledge distillation, and self-training. Data augmentation [9, 28, 53] utilizes an enhancement technique to increase the size and quality of the model training dataset, playing a significant role in optimizing especially limited training sets. Module improvement including representation learning, classifier design, decoupled training, and ensemble learning. Representation learning [10, 37, 63, 64, 68] entails refining network structures to facilitate a more effective acquisition of informative representations. Classifier design [22, 31] is designed by setting up the appropriate classifier so that it focuses more on the tail class. Decoupled training [13, 25, 56] untangles the learning process by segregating it into representation learning and classifier training, ensuring that they do not affect each other. Ensemble learning [19, 29, 65] to solve the issues of long-tail learning by deliberately generate and amalgamate multiple network modules.

3. Methodology

3.1. Overview

In this work, we choose the multi-object tracking algorithm of the Joint Detection and Tracking framework [43] to carry out our strategies, which is illustrated in Fig. 2. From the data perspective, we propose two bespoke data augmentation methods, including Stationary Camera View Data Augmentation (SVA) and Dynamic Camera View Data Augmentation (DVA), designed for viewpoint states, to simulate the pedestrians of tail classes motion trajectory and change the background style. In addition, we commence by addressing the similarity metric Re-ID used in the association and propose the Group Softmax (GS) module to improve the appearance recognition performance for the pedestrians of tail classes.

3.2. Camera View Data Augmentation

We observed the distinct characteristic in the data from the multi-object tracking, which the data was collected under varying camera motion conditions. Depending on the camera motion conditions during data collection, it can be categorized into data collected from static cameras and data collected from moving cameras. Consequently, we developed the customized data augmentation methods for data collected from stationary cameras and dynamic cameras.

In particular, we define the calculation formula for category division as shown in Eq. (1). Then, we use Eq. (1) to divide the pedestrian categories in each sequence in the dataset into head classes and tail classes.

$$C_i = \begin{cases} C_i^{tail} & \frac{1}{R_i} \geq T_j \\ C_i^{head} & \frac{1}{R_i} < T_j \end{cases} \tag{1}$$

where C_i represents the classes to which category i belongs, R_i represents the ratio of the number of category i to the number of all categories in j sequence, T_j represents the class threshold in j sequence for determining whether the classes of is tail or not.

3.2.1 Stationary Camera View Data Augmentation

For multi-object tracking data captured by stationary cameras, common data augmentation methods, such as image color transformation, image blending and image cropping, are currently available. Although these methods can be applied, they are not specifically designed for multi-object tracking tasks and lack customized designs for tracking targets. Therefore, we propose the Stationary Camera View Data Augmentation (SVA) strategy tailored for the

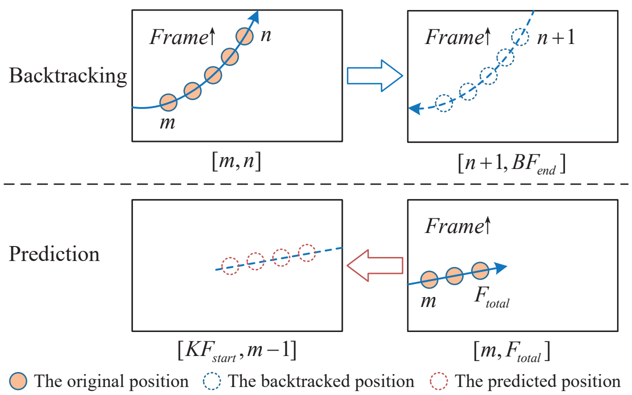

Figure 3. Illustration of the Stationary Camera View Data Augmentation (SVA).

multi-object tracking of data captured by stationary cameras, focusing on quantity augmentation for the pedestrians of tail classes. The SVA strategy includes backtracking continuation and prediction continuation, which is shown in Fig. 3. The backtracking continuation is to add the reversed original track in the subsequent frames after the end of the original track, applied to the pedestrians of tail classes in the middle frame of the training sequence data. The prediction continuation is to add the future trajectory predicted using the position information of the original trajectory to the previous frames at the beginning of the original trajectory, employed for the pedestrians of tail classes in the final frame of the training sequence data.

Backtracking continuation. For a training video with a total of F_{total} frames, if the pedestrian trajectory of tail classes appears in the m-th frame and disappears in the n-th frame, satisfying the condition $n < F_{total}$, we employ the Segment Anything Model (SAM) [26] algorithm to segment the image area of the pedestrian that appear in frames from the m-th to the n-th and then overlay these image areas in reverse order onto frames following the n-th frame. The backtracking continuation can be formulated as:

$$BP_j^k = P_i^k (m \leq i \leq n, n+1 \leq j \leq BF_{end}) \quad (2)$$

where BP_j^k represents the backtracked image position of the k-th pedestrian in the training data at the j-th frame, P_i^k represents the image position of the k-th pedestrian in the training data at the i-th frame, and BF_{end} is the backtracking continuation cutoff frame in the training dataset, and the value is the minimum value of F_{total} and $(2n - m)$.

Prediction continuation. For a training video with a total of F_{total} frames, if the pedestrian trajectory of tail classes appears in the final frame, we input the x and y image coordinates of the pedestrian appearing from the m-th frame to the F_{total}-th frame into a Kalman filter to predict the subsequent x and y image coordinates of the pedestrian, while ensuring that the predicted image coordinates fall in the image size range. In this pedestrian trajectory, we randomly select the pedestrian with visibility no less than the visibility threshold T_v, where $0 \leq T_v \leq 1$, using the SAM [26] algorithm to segment the pedestrian. The segmented image area is superimposed on the frame before the pedestrian trajectory appears based on the predicted x and y image coordinate randomly selected from the predicted image coordinates. The prediction continuation can be formulated as:

$$KP_j^k = R(KF\left(P_i^k\right))(m \leq i \leq F_{total}, KF_{start} \leq j < m) \quad (3)$$

where KP_j^k represents the Kalman filter predicted image position of the k-th pedestrian in the training data at the j-th frame, $R()$ represents the function that randomly selects an image position, $KF\left(P_i^k\right)$ represents the image positions predicted by the Kalman filter using the images coordinates of the m-th frame to F_{total} for the i-th pedestrian in the training data, and P_i^k represents the image position of the k-th pedestrian in the training data at the i-th frame. KF_{start} is the starting frame for applying the prediction continuation in the training dataset, and the value is the maximum value of 1 and $(2m - F_{total})$.

3.2.2 Dynamic Camera View Data Augmentation

Due to the characteristics of significant scene and subject size variations in data captured by dynamic cameras, the traditional data augmentation methods struggle to adapt to these changes. To address this issue, we propose the Dynamic Camera View Data Augmentation (DVA) strategy, as depicted in Fig. 4. The strategy comprises four main steps: image segmentation, image inpainting, image diffusion and image merging. This strategy, designed for input from dynamic camera perspectives, begins by using the image segmentation algorithm SAM [26] to separate pedestrians in the input image derived from the sequence, resulting in image with pedestrians removed, image with pedestrian mask, and image containing only the pedestrians area. Next, the image inpainting algorithm Navier-Stokes [5] is applied to repair the image with pedestrians removed, producing the repaired image. Following this, the Stable Diffusion [36] is used to process the repaired image, resulting in the diffused image. Finally, the image with pedestrians mask and the image containing only the pedestrians area obtained from the earlier segmentation step are merged with the diffused image to generate the output image.

Image Segmentation. SAM, which stands for Segment Anything Model, is the largest segmentation model ever released by Meta. This model segments objects by taking both a prompt and an image as inputs. Given that the multi-object tracking dataset provides bounding box annotations but lacks mask labels for pedestrians, we utilize the image

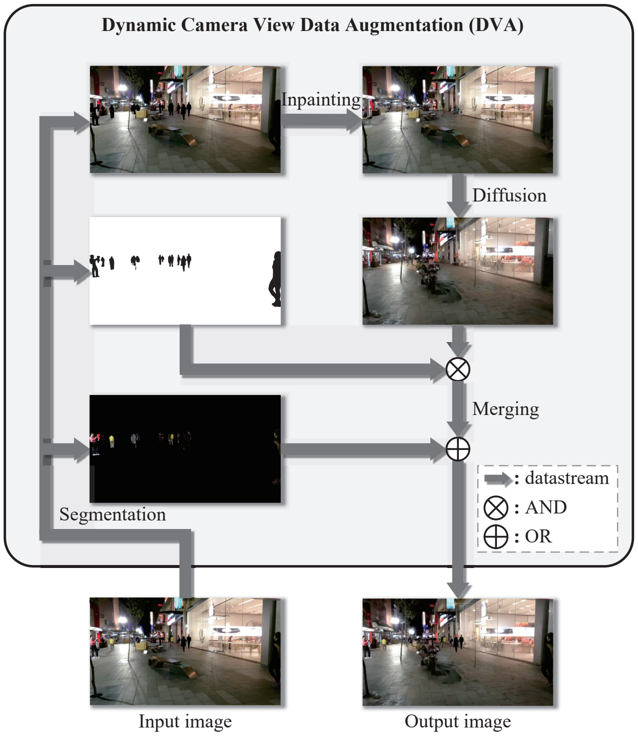

Figure 4. Illustration of the Dynamic Camera View Data Augmentation (DVA).

and its corresponding pedestrians bounding box labels as input for SAM to segment the pedestrians in the image.

Image Inpainting. In this paper, the algorithm used for image inpainting is based on the Navier-Stokes equation [5]. This algorithm aims to start repairing the image from the edges of the area to be patched, propagating image smoothness along the contour lines, and obtaining the repaired image after all information has been propagated.

Image Diffusion. Stable Diffusion is a type of Latent Diffusion Model (LDM) [36], which is a class of denoising diffusion probability models capable of generating new image. In principle, Stable Diffusion can model conditional distributions. This can be achieved by inputting text, semantic maps, or the other image-to-image transformation task information to control conditional denoising autoencoders. In this paper, we utilize input image to generate new image by adjusting prompt and enhancement coefficient.

Image Merging. We perform a bitwise AND operation between the image with pedestrian mask and the diffused image, effectively setting the pixel values in the diffused image corresponding to the original pedestrians area to 0, while leaving the pixel values in area outside the pedestrians area are unchanged. This results in the post-processed diffused image. Then, we perform a bitwise OR operation between the post-processed diffused image and the image containing only the pedestrians. Indeed, this operation involves setting the pixel values in the diffused image that correspond to the original pedestrians area to the corresponding pixel values from the image containing only pedestrians, resulting in the output image.

Training models on the augmented data from diffusion model often risks over-emphasizing spurious qualities [1]. The common solution assigns different sampling probabilities to original and augmented data to manage imbalance [23]. We apply a similar method to balance original images and the augmented images from DVA. Mathematically, the method can be formulated as:

$$I_i^n = \begin{cases} I_i & , P_i^n \leq T_s \\ \tilde{I}_i & , P_i^n > T_s \end{cases} \tag{4}$$

where I_i^n represents the image of index i in the n-th epoch, I_i represents the original image of index i, $\tilde{I}_i$ represents the augmented image of index i, P_i^n represents the probability of index i calling the original image in the n-th epoch, T_s represents the image selection threshold of calling the original image for each image selection. Given index i, with probability T_s the original image I_i is added to the epoch n, otherwise its augmented image $\tilde{I}_i$ is added.

3.3. Group Softmax Module

We observe the issue where the Re-ID have different degrees of feature learning for pedestrian categories with different quantities. It tends to perform better for categories with higher quantities (head classes) and less effectively for categories with fewer quantities (tail classes), which can negatively impact the performance of the Re-ID. To tackle this problem, we propose the Group Softmax (GS) module, as depicted in Fig. 2. The GS divides the pedestrian categories into several disjoint groups and performes softmax operation separately for each group. In this way, the pedestrian categories with similar quantities can compete in the same group. Thus, the GS can isolate categories with significant quantity differences, preventing the weights of tail classes from being heavily suppressed by the head classes.

Specifically, we divides the total of M pedestrian classes in the training dataset into K distinct groups according to their number in the training dataset, and the rule of partitioning group formula is: $T_j^l \leq N(i) \leq T_j^h$, where the value of i is from 1 to M, the value of j is from 1 to K, $N(i)$ is the quantity of the i-th pedestrian categories in the training dataset, T_j^l is the lowest quantity thresholds for the j-th group, T_j^h is the highest quantity thresholds for the j-th group, M represents the number of pedestrian categories, K represents the number of groups.

To ensure each pedestrian category is only assigned to one group and maintain ordered groups, we specify that the lowest quantity threshold for the $j+1$-th group is equal to the highest quantity threshold for the j-th group, i.e., T_{j+1}^l = $T_j^h + 1$. To facilitate grouping, we propose the set rule of T_j^h formula is as follows:

$$T_j^h = \frac{j}{K} max(N(i))(1 \leq i \leq M, 1 \leq j \leq K) \quad (5)$$

where T_j^h is the highest quantity thresholds for the j-th group, j represents the index of group.

Furthermore, we individually apply softmax processing to each group and utilize the Cross-Entropy Loss to compute the group loss. Then, we calculate the group loss mean as the Re-ID loss, the formula is as follows:

$$Loss_{\text{Re-ID}} = -\frac{1}{K}\sum_{j=1}^{K}\sum_{i \in G_j} y_i \log(p_i) \quad (6)$$

where $Loss_{\text{Re-ID}}$ represents the Re-ID loss, K represents the number of groups, j represents the index of group, G_j represents the j-th group, y_i represents the label in G_j, p_i represents the probability in G_j.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets: We conduct extensive experiments on four public MOT benchmarks, i.e., MOT15 [27], MOT16 [34], MOT17 [34] and MOT20 [12]. Specifically, MOT15 contains 22 sequences, 11 for training and the other 11 for testing, which includes 11286 frames. MOT16 contains 14 sequences, 7 for training and the other 7 for testing, which includes 11235 frames. Compared with MOT16, MOT17 adds the detection bounding boxes of three detectors that DPM, SDP, Faster-RCNN. MOT20 contains 8 sequences captured in the crowded scenes, 4 for training and the other 4 for testing, which includes 13410 frames. In some frames, more than 200 pedestrians are included simultaneously.

Evaluation metrics: To evaluate, we use the CLEAR metrics [4], including multiple object tracking accuracy (MOTA), ID F1 score (IDF1), Higher Order Tracking Accuracy (HOTA), mostly tracker rate (MT), mostly lost rate (ML) and identity switches (IDS). MOTA, IDF1 and HOTA are three important comprehensive metrics. MOTA focuses on detection performance, and IDF1 focuses on association performance. Compared with them, HOTA balances detection performance and association performance.

4.2. Implementation Details

All experiments are trained using an NVIDIA GeForce RTX 3090 GPU. All models are trained for 30 epochs. For MOT15, we set the class threshold T_j to 15 for all stationary camera view sequences, the visibility threshold T_v to 1.0, the image selection threshold T_p to 0.8, the prompt of diffusion to "A street" for all dynamic camera view sequences, the enhancement coefficient of diffusion to 0.4, and the group number K to 3 for FairMOT and 4 for CSTrack.

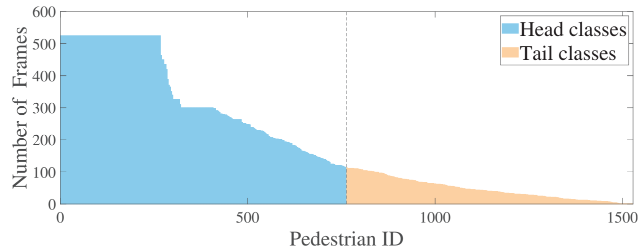

Figure 5. Division of head classes and tail classes is based on the class average principle on the MOT17 validation set.

Method	All classes		Head classes		Tail classes	
	MOTA↑	IDF1↑	MOTA↑	IDF1↑	MOTA↑	IDF1↑
Base	67.8	72.3	69.3	74.0	28.6	56.9
Base(+Focal Loss)	68.7	70.8	70.0	72.2	26.2	56.8
Base(+Triplet Loss)	68.4	65.7	69.7	67.7	23.3	51.3
Base(+CB Loss)	67.9	72.0	69.3	73.3	27.7	55.7
Base(+Logit Adjustment)	68.1	72.1	69.6	74.0	26.8	55.8
Base(+**Ours**)	**69.3**	**73.4**	**70.9**	**75.2**	**29.6**	**57.2**

Table 1. Comparison of different methods for improving long-tail distribution on the MOT17 validation set. The best results are shown in **bold**. Our method is highlighted in blue.

For MOT16 and MOT17, we set T_j to 120 for 02, 04 and 09 sequences, T_v to 1.0, T_p to 0.9, except T_p is 1.0 for CSTrack on MOT17, prompt to "A street" for 05, 10 and 13 sequences, "A mall" for 11 sequence, enhancement coefficient to 0.4, K to 3 for FairMOT on MOT16, K to 3 for FairMOT on MOT17, 4 for CSTrack on MOT16, and 2 for CSTrack on MOT17. For MOT20, since MOT20 is all stationary camera view data, we only need to set the SVA and GS parameters.We set T_j to 1000 for all sequences, T_v to 1.0, and K to 2. For a fair comparison, we adopt the same training settings as for the baseline to retrain each tracking model with and without our method, where the training datasets include MOT15, MOT16, MOT17 and MOT20.

4.3. Comparison of long-tail distribution solutions.

We follow the settings of the ablation experiment, count the number of frames of pedestrians with different identities on the MOT17 validation set, and divide all classes into head classes and tail classes according to the class average principle, as shown in Fig. 5. We evaluate various long-tail distribution solutions in multiple classes on the MOT17 validation set. As shown in Tab. 1, we can observe that some methods boost the MOTA metric but reduce the IDF1 metric on all classes, and the Logit Adjustment method boosts the performance on the all classes but reduces the performance on the tail classes. In comparison, our method achieves the best performance in all classes, head classes and tail classes.

4.4. Comparison with other SOTA

Due to the fact that Joint Detection and Tracking involve the concurrent learning of detectors and appearance extrac-

Method	HOTA↑	IDF1↑	MOTA↑	MT↑	ML↓	IDS↓
MOT15						
FairMOT [58]	45.2	59.4	53.1	39.0%	15.4%	911
FairMOT(**+Ours**)	**47.0**(+1.8)	**61.5**(+2.1)	**56.7**(+3.6)	**42.7%**	**14.0%**	**845**
MOT15						
CSTrack [30]	42.9	56.7	49.8	20.9%	27.2%	762
CSTrack(**+Ours**)	**46.4**(+3.5)	**60.2**(+3.5)	**55.0**(+5.2)	**46.7%**	**13.0%**	920
MOT16						
FairMOT [58]	57.7	70.8	71.0	38.2%	21.6%	1,274
FairMOT(**+Ours**)	**57.8**(+0.1)	**71.3**(+0.5)	**71.0**(+0)	**38.9%**	**21.5%**	**1,270**
MOT16						
CSTrack [30]	53.1	67.5	65.7	30.4%	25.0%	1,265
CSTrack(**+Ours**)	**53.3**(+0.2)	67.5(+0)	**66.0**(+0.3)	29.5%	25.4%	1,303
MOT17						
FairMOT [58]	57.1	70.2	70.2	40.9%	18.6%	4,329
FairMOT(**+Ours**)	**57.3**(+0.2)	70.1(-0.1)	69.9(-0.3)	**41.4%**	**17.8%**	4,776
MOT17						
CSTrack [30]	52.6	66.1	65.2	30.7%	23.8%	4,605
CSTrack(**+Ours**)	**53.1**(+0.5)	**66.7**(+0.6)	65.1(-0.1)	**32.0%**	25.0%	**4,341**
MOT20						
FairMOT [58]	52.3	65.0	56.8	67.2%	7.3%	6,108
FairMOT(**+Ours**)	**54.4**(+2.1)	**70.3**(+5.3)	**65.9**(**+9.1**)	49.7%	12.5%	**3,548**
MOT20						
CSTrack [30]	45.7	59.9	58.1	33.7%	21.8%	3,645
CSTrack(**+Ours**)	**47.1**(+1.4)	**60.4**(+0.5)	**58.1**(+0)	**35.5%**	**20.9%**	4,358

Table 2. State-of-the-art comparisons on four public MOT benchmarks, i.e., MOT15, MOT16, MOT17 and MOT20. Performance under the private detection on the test set of four public MOT benchmarks, only using themselves train set. All the results are obtained from the official MOT challenge evaluation server. Our better results are marked in **bold**. The gain vlaues are marked in red. The best gain vlaue is marked in **red**. Our method is highlighted in blue.

tors, it is highly suitable for evaluating our method aimed at training data and Re-ID. To evaluate the effectiveness of our method, we apply them to two state-of-the-art trackers of the Joint Detection and Tracking, FairMOT [58] and CSTrack [30]. We evaluate FairMOT and CSTrack on four public MOT benchmarks, i.e., MOT15, MOT16, MOT17 and MOT20. The results are reported in Tab. 2. According to the results, our method can improve the performance of the algorithm on MOTA, IDF1, HOTA and other metrics, especially on the MOT15 and MOT20 benchmarks.

MOT15: As shown in Tab. 2, the superiority of our method can be fully reflected on the benchmark MOT15. FairMOT is improved by 1.8% HOTA, 2.1% IDF1 and 3.6% MOTA. CSTrack is improved by 3.5% HOTA, 3.5% IDF1, 5.2% MOTA and 25.8% MT, decreased by 14.2% ML. This demonstrates that our method excellently improves the performance of detection and appearance feature extraction.

MOT16 and MOT17: Compared with MOT15, MOT16 and MOT17 contain more data and more precised annotations. The results in Tab. 2 show that FairMOT is improved by 0.1% HOTA on MOT16 and 0.2% HOTA on MOT17,

Method	Venue	MOTA↑	IDF1↑	HOTA↑	IDS↓
Joint Detection and Tracking framework:					
FairMOT [58]	IJCV 2021	61.8	67.3	54.6	5,243
CSTrack [30]	TIP 2022	66.6	68.6	54.0	**3,196**
RelationTrack [50]	TMM 2022	67.2	70.5	55.1	4,243
MTrack [49]	CVPR 2022	63.5	69.2	-	6,031
FairMOT(**+Ours**)	-	**67.8**	**70.7**	**55.4**	3,505

Table 3. Comparison with the SOTA methods of the Joint Detection and Tracking framework on the MOT20 test set. The best results are shown in **bold**. Our method is highlighted in blue.

Method	Training Data	Images	Identities	MOTA↑	IDF1↑	IDS↓
FairMOT [58]	MIX	77K	10.4K	61.8	67.3	5,243
FairMOT(**+Ours**)	MIX	77K	10.4K	**67.8**	**70.7**	**3,505**
FairMOT(**+Ours**)	MOT20	**9K**	**2.2K**	**65.9**	**70.3**	**3,548**

Table 4. Results of the MOT20 test set when using different methods and different datasets for training. "MIX" represents the mixed datasets, including MOT20 [12], ETH [16], CityPerson [55], CalTech [14], CUHK-SYSU [46], PRW [62] and CrowdHuman [38] dataset. Our better results are shown in **bold**. Our method is highlighted in blue.

and CSTrack is improved by 0.2% HOTA on MOT16 and 0.5% HOTA on MOT17, after adding our method. HOTA is a comprehensive index that balances detection performance and association performance. This shows that our method can improve the comprehensive ability of the network.

MOT20: Compared with previous MOT benchmarks, MOT20 is more crowded. As shown in Tab. 2, our method delivers extremely outstanding results, with improvements of 2.1% HOTA, 5.3% IDF1 and 9.1% MOTA for FairMOT and 1.4% HOTA for CSTrack. We speculate that it is because the number of head classes and tail classes in the MOT20 dataset is very different, and our method alleviates the negative impact of the long tail distribution of the MOT20 dataset. Our method achieves extremely superior performance in dense pedestrian scenes.

Then, we evaluate the SOTA methods of the Joint Detection and Tracking framework on the MOT20 test set. As shown in Tab. 3, our method achieves the best performance on comprehensive metrics.

Furthermore, to explore the data efficiency benefits of our approach. We used different methods to train models on different data, and the results are shown in Tab. 4. Notably, our method trained using only the MOT20 data is 4.1% higher in MOTA and 3.0% higher in IDF1 than the baseline method trained using MIX data, indicating that our method is particularly effective for data efficiency.

4.5. Ablation Study

Using FairMOT [58] as the baseline tracker, we perform a series of ablation study on MOT17 dataset to demonstrate the effectiveness of our method from different as-

Method	SVA	DVA	GS	MOTA↑	IDF1↑
1 Base	✘	✘	✘	67.8	72.3
2 Base(+SVA)	✔	✘	✘	68.3	72.8
3 Base(+DVA)	✘	✔	✘	68.8	72.7
4 Base(+SVA+DVA)	✔	✔	✘	69.0	73.0
5 Base(+SVA+DVA+GS)	✔	✔	✔	**69.3**	**73.4**

Table 5. Impact of each proposed component on validation set of MOT17. (SVA: Stationary Camera View Data Augmentation, DVA: Dynamic Camera View data Augmentation, GS: Group Softmax. The best results are marked in **bold**.)

T_s	MOTA↑	IDF1↑	FP↓	FN↓	IDS↓
0.0	68.0	70.2	2,511	14,329	475
0.1	68.0	70.2	2,945	13,868	462
0.2	68.4	70.2	2,778	13,810	461
0.3	68.1	69.7	2,670	14,090	479
0.4	68.1	71.3	2,991	13,787	451
0.5	68.5	70.8	2,663	13,844	495
0.6	68.0	69.8	2,785	14,016	515
0.7	69.0	72.5	2,685	13,591	470
0.8	69.0	70.6	2,670	13,609	453
1.0	68.3	71.0	3,224	13,426	458
0.9	69.3	73.4	2,640	13,504	455

Table 6. Comparison of different image selection thresholds T_s. The top two results are highlighted with red and blue.

pects. Since MOTChallenge does not provide the validation set, we divide the training datasets of MOT17 into two parts, the first one half as training set and the other half as validation set. the training datasets of MOT17 include 02, 04, 05, 09, 10, 11 and 13 sequences. 02, 04 and 09 sequences belong to stationary camera view data. 05, 10, 11 and 13 sequences belong to dynamic camera view data. Thus, we apply SVA to 02, 04 and 09 sequences, and DVA to 05, 10, 11 and 13 sequences. All the models are trained for 30 epochs on the training set of MOT17.

Impact of each component. As shown in Tab. 5, all the components have boosted the tracking performance effectively. Our method can obtain 1.5% MOTA and 1.1% IDF1 gains (1 vs. 5). Among them, the SVA increases 0.5% MOTA and 0.5% IDF1, and DVA can further improve MOTA to 69.0% and IDF1 to 73.0%. Due to GS, our method boosts the performance to 69.3% on MOTA and 73.4% on IDF1.

Analysis of DVA. As introduced in Sec. 3.2.2. We select the image selection threshold T_p in Eq. (4) to analyze its impact. We change the image selection threshold T_s from 0.1 to 1.0 at intervals of 0.1. The results by using different image selection threshold T_s are illustrated in Tab. 6. We can observe that the best experimental results occur when the image selection threshold T_s is 0.9. This phenomenon shows that using a small amount of enhanced images by DVA can help improve the performance of network. We speculate that the reason is that the network learns too much about the features of the enhanced image by DVA, which leads to ignoring the scene features of the original image.

(a) Baseline

(b) Baseline+GS

Figure 6. Visualization about the cosine metric matrix between appearance vectors of the current example frame and the track templates. Red color indicates a higher association, and blue indicates a lesser association.

Analysis of GS. As introduced in Sec. 3.3. For a intuitive comparison, we construct the cosine metric matrix between appearance vectors of the current example frame and the track templates in Fig. 6. During the matching process, the ideal situation is that there is at most one red color in each row and column of the cosine metric matrix, and the rest are blue. We can observe that the GS can significantly improve the association ability.

5. Conclusion

In this study, we note a significant imbalance in the distribution of trajectory lengths across different pedestrians, revealing the long-tail distribution issue within existing MOT datasets. To address it, we propose our method, which focuses on two key aspects: information augmentation and module improvement. Specifically, we introduce two data augmentation approach tailored for viewpoint states, including SVA and DVA, and the GS module for Re-ID. Notably, our work represents the pioneering effort in tackling the long-tail distribution in the realm of MOT. Using two SOTA multi-object trackers, we have verified the effectiveness of our method on MOTChallenge benchmarks. The experimental results demonstrate that our method effectively mitigates the impact of long-tail distribution on MOT.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62176096.

References

[1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017. 5

[2] Mk Bashar, Samia Islam, Kashifa Kawaakib Hussain, Md Bakhtiar Hasan, ABM Rahman, and Md Hasanul Kabir. Multiple object tracking in recent times: a literature review. *arXiv preprint arXiv:2209.04796*, 2022. 1

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 2

[4] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 6

[5] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pages I–I. IEEE, 2001. 4, 5

[6] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 2

[7] Dong Cao, Xiangyu Zhu, Xingyu Huang, Jianzhu Guo, and Zhen Lei. Domain balancing: Face recognition on long-tailed domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5671–5679, 2020. 2

[8] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 2

[9] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 95–110. Springer, 2020. 3

[10] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 715–724, 2021. 3

[11] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 2

[12] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 2, 6, 7

[13] Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. Learning of visual relations: The devil is in the tails. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15404–15413, 2021. 3

[14] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE conference on computer vision and pattern recognition*, pages 304–311. IEEE, 2009. 7

[15] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023. 2

[16] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. A mobile vision system for robust multi-person tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 7

[17] Chengjian Feng, Yujie Zhong, and Weilin Huang. Exploring classification equilibrium in long-tailed object detection. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 3417–3426, 2021. 2

[18] Tobias Fischer, Thomas E Huang, Jiangmiao Pang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[19] Hao Guo and Song Wang. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15089–15098, 2021. 3

[20] Shoudong Han, Piao Huang, Hongwei Wang, En Yu, Donghaisheng Liu, and Xiaofeng Pan. Mat: Motion-aware multi-object tracking. *Neurocomputing*, 476:75–86, 2022. 2

[21] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019. 3

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[23] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 5

[24] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 2

[25] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020. 3

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 4

[27] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015. 2, 6

[28] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5212–5221, 2021. 3

[29] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10991–11000, 2020. 3

[30] Chao Liang, Zhipeng Zhang, Xue Zhou, Bing Li, Shuyuan Zhu, and Weiming Hu. Rethinking the competition between detection and reid in multiobject tracking. *IEEE Transactions on Image Processing*, 31:3182–3196, 2022. 2, 7

[31] Bo Liu, Haoxiang Li, Hao Kang, Gang Hua, and Nuno Vasconcelos. Gistnet: a geometric structure transfer network for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8209–8218, 2021. 3

[32] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2970–2979, 2020. 3

[33] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022. 2

[34] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 2, 6

[35] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020. 2

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4, 5

[37] Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9495–9504, 2021. 3

[38] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 7

[39] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 2

[40] Yingfan Tao, Jingna Sun, Hao Yang, Li Chen, Xu Wang, Wenming Yang, Daniel Du, and Min Zheng. Local and global logit adjustments for long-tailed learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11783–11792, 2023. 2

[41] Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10860–10869, 2021. 2

[42] Yuchao Wang, Jingjing Fei, Haochen Wang, Wei Li, Tianpeng Bao, Liwei Wu, Rui Zhao, and Yujun Shen. Balancing logit variation for long-tailed semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19561–19573, 2023. 2

[43] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. *The European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[44] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 2

[45] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 247–263. Springer, 2020. 3

[46] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3415–3424, 2017. 7

[47] Han-Jia Ye, Hong-You Chen, De-Chuan Zhan, and Wei-Lun Chao. Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprint arXiv:2001.01385*, 2020. 2

[48] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5704–5713, 2019. 3

[49] En Yu, Zhuoling Li, and Shoudong Han. Towards discriminative representation: Multi-view trajectory contrastive learning for online multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8834–8843, 2022. 7

[50] En Yu, Zhuoling Li, Shoudong Han, and Hongwei Wang. Relationtrack: Relation-aware multiple object tracking with decoupled representation. *IEEE Transactions on Multimedia*, 2022. 2, 7

[51] En Yu, Songtao Liu, Zhuoling Li, Jinrong Yang, Zeming Li, Shoudong Han, and Wenbing Tao. Generalizing multiple object tracking to unseen domains by introducing natural language representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3304–3312, 2023. 2

[52] En Yu, Tiancai Wang, Zhuoling Li, Yuang Zhang, Xiangyu Zhang, and Wenbing Tao. Motrv3: Release-fetch supervision for end-to-end multi-object tracking. *arXiv preprint arXiv:2305.14298*, 2023. 2

[53] Yuhang Zang, Chen Huang, and Chen Change Loy. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3457–3466, 2021. 3

[54] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pages 659–675. Springer, 2022. 2

[55] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3221, 2017. 7

[56] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021. 3

[57] Xing Zhang, Zuxuan Wu, Zejia Weng, Huazhu Fu, Jingjing Chen, Yu-Gang Jiang, and Larry S Davis. Videolt: Large-scale long-tailed video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7960–7969, 2021. 2

[58] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021. 2, 7

[59] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022. 2

[60] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pre-trained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22056–22065, 2023. 2

[61] Yan Zhao, Weicong Chen, Xu Tan, Kai Huang, and Jihong Zhu. Adaptive logit adjustment loss for long-tailed visual recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3472–3480, 2022. 2

[62] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1367–1376, 2017. 7

[63] Yaoyao Zhong, Weihong Deng, Mei Wang, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang. Unequal-training for deep face recognition with long-tailed noisy data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7812–7821, 2019. 3

[64] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021. 3

[65] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020. 3

[66] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, pages 474–490. Springer, 2020. 2

[67] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8771–8780, 2022. 2

[68] Linchao Zhu and Yi Yang. Inflated episodic memory with region self-attention for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4344–4353, 2020. 3

[69] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33:3833–3845, 2020. 3