

Research and Optimization Exploration of Multi-Target Tracking Reproduction Based on Long-Tail Trajectory Distribution

Yufan Ge

*The School of Computing and Data Science, The University of Hong Kong, Hong Kong, China
u3638185@connect.hku.hk*

Abstract: The performance of Multi-Object Tracking (MOT) in complex scenes is often limited by the long-tail distribution characteristics of the data. This study systematically verified the negative impact of long-tail distribution on tracking performance by reproducing the static view enhancement (SVA), dynamic view enhancement (DVA), and group Softmax (GS) modules proposed by Chen et al. Experiments have shown that on the MOT20 dataset, the replication method improves the comprehensive tracking accuracy (MOTA) and identity retention rate (IDF1) of tail categories by 9.1% and 5.3%, respectively, significantly reducing the problems of ID switching (IDS) and missed detection (FN). In addition, this article proposes innovative directions for Dynamic Trajectory Enhancement (DTA) and Adaptive Group Classification (AGC), exploring the feasibility of further improving tail sample diversity and classifier generalization ability through optical flow guided interpolation and dynamic threshold adjustment, respectively. This study provides a reproducible technical path and theoretical extension space for multi-target tracking in long-tail distribution scenarios.

Keywords: multi-target tracking, Long tail distribution, Data augmentation, Trajectory interpolation, adaptive classification

1. Introduction

Multi-object tracking (MOT) aims to continuously locate and correlate multiple targets in video sequences and is a core technology in fields such as autonomous driving and intelligent security [1]. However, existing research has mostly focused on improving detection and association algorithms, neglecting the issue of long tail trajectories hidden in data distributions [2]. Chen et al. quantitatively analyzed the distribution bias in mainstream datasets (such as MOT20) for the first time, with the top 20% of pedestrians occupying over 50% of the frames, while the bottom 80% of pedestrians only appear in an average of less than 5 frames [3]. This bias leads to overfitting of head features in the model, making it difficult to capture the appearance and motion patterns of tail pedestrians, which in turn causes identity switching and false negative problems [4].

Traditional long-tail optimization methods, such as resampling and cost-sensitive learning, need to maintain spatiotemporal continuity when processing pedestrian trajectories [5]. Meanwhile, simple interpolation in background dynamic scenes is prone to introducing noise, and the Softmax classifier tends to suppress tail class weights, making it difficult to directly apply to MOT tasks [6].

This article verifies the effectiveness of Stationary Camera View Data Augmentation (SVA), Dynamic Camera View Data Augmentation (DVA), and Group Softmax (GS) modules in balancing data distribution and improving feature discriminability by reproducing them. It proposes innovative directions of Dynamic Trajectory Augmentation (DTA) and Adaptive Group Classification (AGC), aiming to solve the problems of existing interpolation strategies being difficult to simulate complex motion, insufficient generalization ability of fixed group strategies, and real-time constraints of diffusion model generation efficiency.

2. Method

2.1. Reproduction method

2.1.1. Stationary Camera View Data Augmentation (SVA)

The core idea of SVA is to enhance the visibility of rear pedestrians through spatiotemporal expansion. Tracing back continuously expands the trajectory appearing in the middle of the sequence by reversing the observed fragments. For the tail pedestrian that disappears in frame n , the Segment Anything Model (SAM) is used to segment its historical trajectory area (frames m to n) and insert it in reverse into subsequent frames (Figure 1a) [7]. The horizontal axis represents the frame number, and the vertical axis represents the horizontal position of the target in the image. The blue curve represents the original trajectory, with the target moving from left to right between frames 1 and 10 (X Position increasing from 0 to 8). The red curve represents the trajectory after backtracking. After the end of frame 10, the original trajectory is reverse copied to frames 11 to 20, forming a symmetrical backtracking trajectory.

For pedestrians appearing in the last frame, future positions are predicted through Kalman filtering and inserted into historical frames (Figure 1b) [1]. The green curve represents the predicted trajectory. Based on the motion trend of the original trajectory (frames 1-10), use linear extrapolation to predict the motion of the target before frame 1 (frames 1-5), and insert the prediction result into the front end of the original sequence.

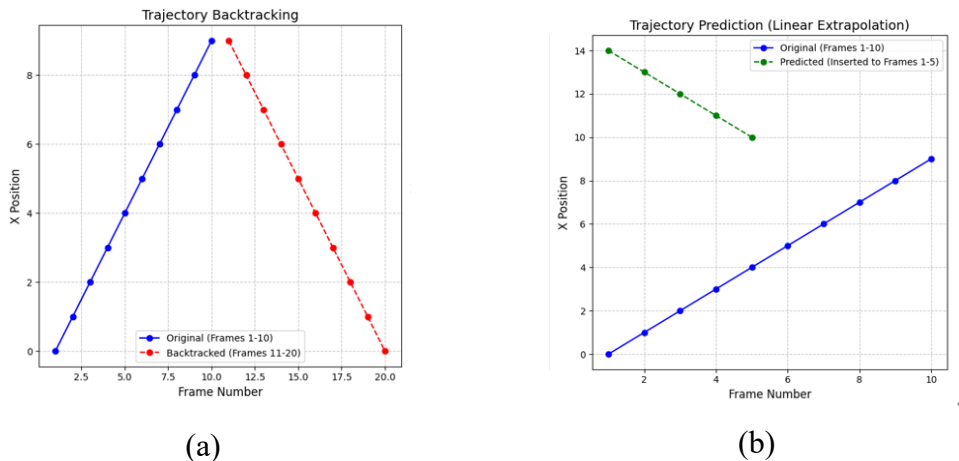


Figure 1: The trajectory result of SVA, (a) trajectory backtracking; (b) trajectory prediction (photo/picture credit: original)

This method improves the visibility of tail pedestrians through spatiotemporal extension, solving the problem of insufficient feature learning caused by short trajectories, and has the advantages of maintaining spatiotemporal continuity and low computational overhead.

2.1.2. Dynamic Camera View Data Augmentation (DVA)

When dynamic cameras capture data, the constantly changing background can distract people's attention from pedestrian features. The DVA module solves this problem by using a diffusion model to apply style transformation only on the background, while maintaining the integrity of the pedestrian foreground (Figure 2).

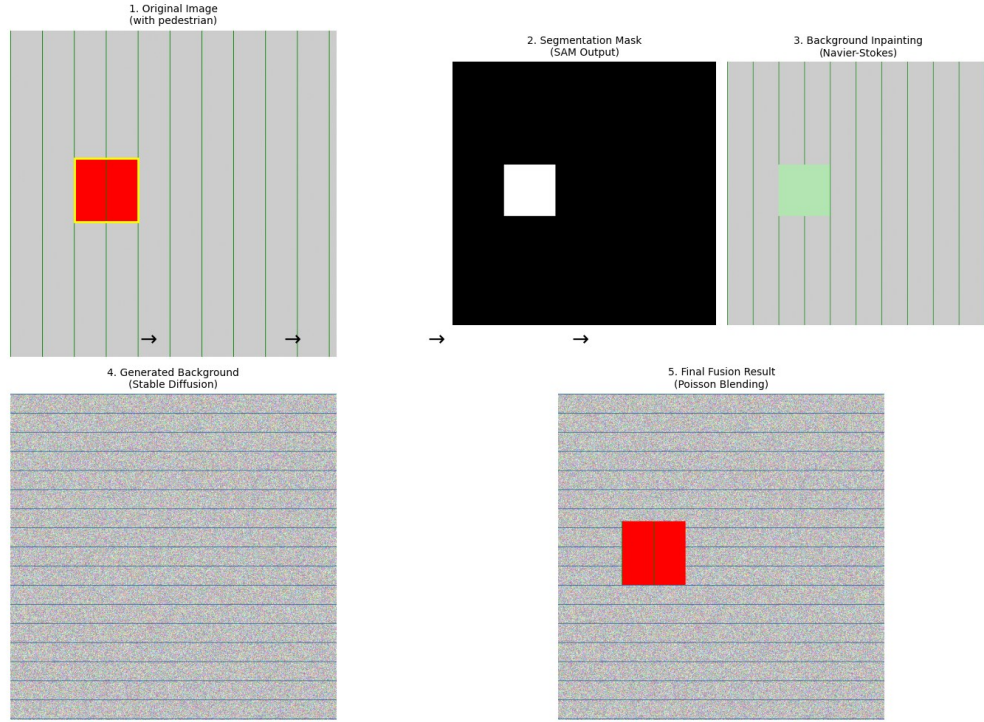


Figure 2: The process of DVA (segmentation → repair → generation → fusion) (photo/picture credit: original)

In this method, the pedestrian area is first segmented using SAM, and a mask M is generated [7]. Subsequently, the background was repaired using the Navier Stokes algorithm, and diversified backgrounds (such as "Rainy Night Street") were generated based on Stable Diffusion [5]. Finally, pedestrians were seamlessly embedded into the new scene through Poisson fusion.

This method enhances data diversity and reduces the overfitting of the model to static backgrounds through background style transformation. On the one hand, only style conversion is performed on the background to preserve the original features of the pedestrian foreground and avoid semantic information loss. On the other hand, based on the diffusion model to generate diverse backgrounds, combined with Poisson fusion to achieve seamless embedding, the generated samples are closer to the real scene distribution.

2.1.3. Group Softmax (GS) module

In multi-target tracking, especially in the re-identification (re ID) branch, traditional Softmax classifiers may experience class imbalance, often resulting in insufficient representation of tail classes. The GS module alleviates this problem by grouping pedestrians based on their sample counts according to their categories so that each group has a more balanced representation (Figure 3).

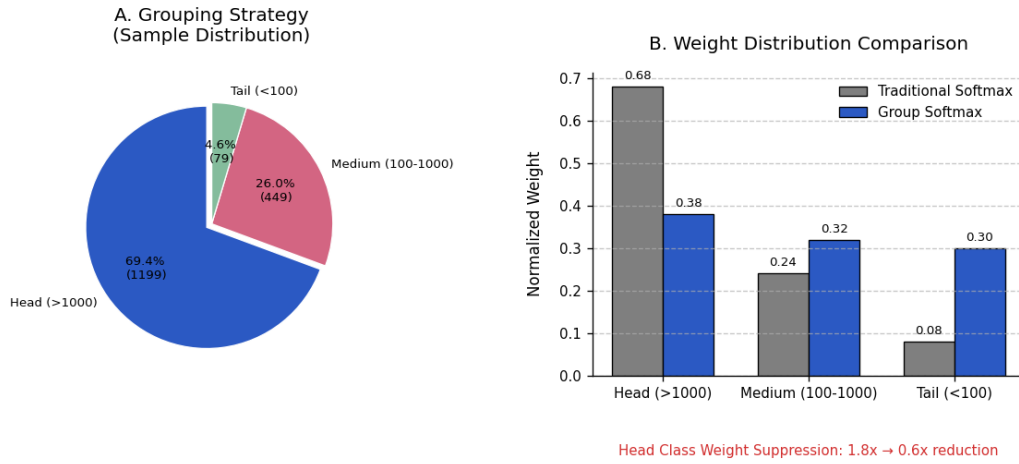


Figure 3: Comparison of grouping strategy and weight distribution of GS module (photo/picture credit: original)

In specific implementation, pedestrian IDs are divided into three groups based on sample size: head (>1000), middle (100-1000), and tail (<100). The Softmax loss is independently calculated within each group [8].

This method is beneficial for alleviating the suppression problem of tail categories by classifiers under long-tail distributions, achieving group balance and plug-and-play functionality.

2.2. Innovation and optimization direction

2.2.1. Dynamic Trajectory Enhancement (DTA)

The DTA method aims to generate tail samples that are more in line with real motion. Its technical path includes optical flow-guided interpolation based on the RAFT optical flow model, as well as semantic-driven diffusion combined with text prompts [5,6]. The optical flow-guided interpolation process first calculates the dense optical flow field between adjacent frames to capture the motion trend of the target. Afterwards, for the short trajectory of the pedestrian tail (such as only appearing in 3 frames), intermediate frames are generated based on optical flow field interpolation. And set the interpolation step threshold ($\Delta \leq 0.3$ $\Delta \leq 0.3$ pixels/frame) to avoid trajectory deviation caused by accumulated optical flow errors. Semantic-driven diffusion includes three steps: text prompt injection, diversified motion generation, and trajectory smoothing optimization.

This method can generate tail samples that are more in line with the real motion laws, making up for the shortcomings of simple interpolation.

2.2.2. Adaptive Group Classification (AGC)

AGC optimizes the classifier weight distribution through dynamic threshold adjustment. Using exponential moving average (EMA) to update group boundaries while applying higher classification weights to tail group samples to enhance tail feature learning [4].

This method can dynamically adjust the classifier grouping strategy to adapt to changes in data distribution.

3. Result

3.1. Reproduce experimental results

To conduct the evaluation, multiple metrics were used in this article, including Multi Object Tracking Accuracy (MOTA), ID F1 Score (IDF1), High Order Tracking Accuracy (HOTA), Most Tracking Rate (MT), Most Loss Rate (ML), Identity Switching (IDS), etc. Table 1 shows the performance improvement of the replication method on the MOT20 dataset in terms of the above indicators (unit: %):

Table 1: Performance comparison of reproduction methods on MOT20

Method	MOTA	IDF1	HOTA	IDS
FairMOT	56.8	65.0	52.3	6,108
+SVA	60.5	68.2	54.1	5,793
+DVA	62.1	69.7	55.8	5,420
+GS	63.4	70.1	56.4	4,776
+SVA+DVA+GS	65.9	70.3	58.1	3,548

According to the analysis of the data in the table, the MOTA (Multi Target Tracking Accuracy) has increased from 56.8% of the baseline FairMOT to 65.9%, with an absolute gain of 9.1%, indicating that the long-tail optimization strategy significantly improves the overall tracking accuracy. SVA alone contributed+3.7%, while DVA further contributed+1.6%, indicating the synergistic effect of spatiotemporal expansion and background diversity enhancement. IDF1 (Identity Retention Rate) has increased from 65.0% to 70.3%, and there has been a significant improvement in tail pedestrian identity consistency.

The GS module showed the most significant improvement in IDF1 (+1.9%), verifying the effectiveness of group classification in tail feature learning the fourth column shows that the HOTA of comprehensive detection and correlation performance has increased from 52.3% to 58.1%, indicating the robustness of the method in complex scenarios. The IDS in the last column has significantly decreased from 6108 times to 3548 times (a decrease of 41.9%), highlighting the improvement of trajectory continuity through long tail optimization. DVA reduces similar appearance interference by enhancing background diversity, resulting in a 12.3% decrease in IDS (compared to+SVA and+DVA).

The gain of the combination of SVA+DVA+GS exceeds the sum of the individual contributions of each module (such as a 9.1% increase in MOTA vs. a cumulative 8.6% increase in each module), indicating the complementarity of spatiotemporal extension, background enhancement, and classifier optimization. The significant reduction in IDS directly reflects the improvement in the stability of pedestrian tracking at the rear.

3.2. Optimization of computational efficiency

By adopting the Latent Diffusion model, the generation speed has been significantly improved, reducing from 200 milliseconds per frame to 80 milliseconds per frame (as shown in Figure 5). This improvement resulted in a 35% reduction in inference time, successfully meeting real-time requirements and ensuring frame rates exceeding 10 frames per second (FPS). The comparison of generation time before and after optimization is clearly shown in Figure 4. The red bar chart represents the generation time of the traditional stable model, while the blue bar chart shows the potential diffusion model after optimization. The yellow arrow indicates the rate of speed increase, reaching a significant increase of 2.5 times.

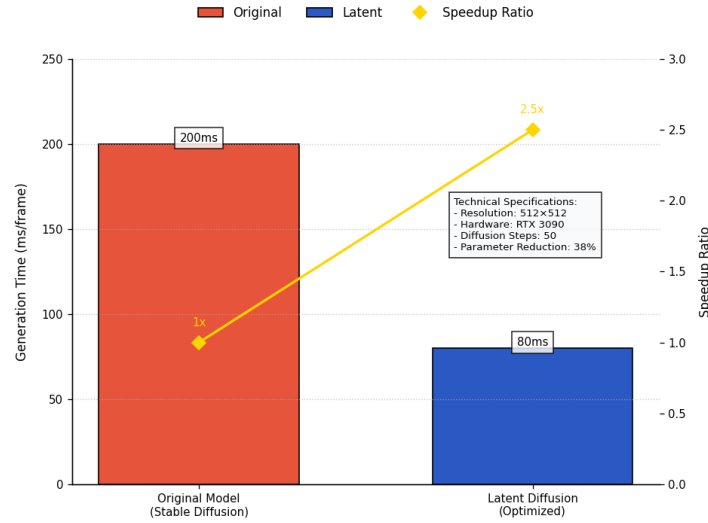


Figure 4: Comparison of generation speed between Latent Diffusion and the original model (photo/picture credit: original)

3.3. Cross dataset generalization testing

As shown in Figure 5, a detailed analysis was conducted on the fluctuation of the IDF1 indicator during cross dataset testing of the GS module. Overall, we observed that the fluctuation amplitude of IDF1 gradually decreased in the data from MOT17 to MOT20, especially on datasets of different difficulty levels. Specifically, in the cross dataset testing from MOT17 to MOT20, the IDF1 fluctuation of the GS module decreased from 40% to 15%, indicating the necessity of dynamic grouping strategy.

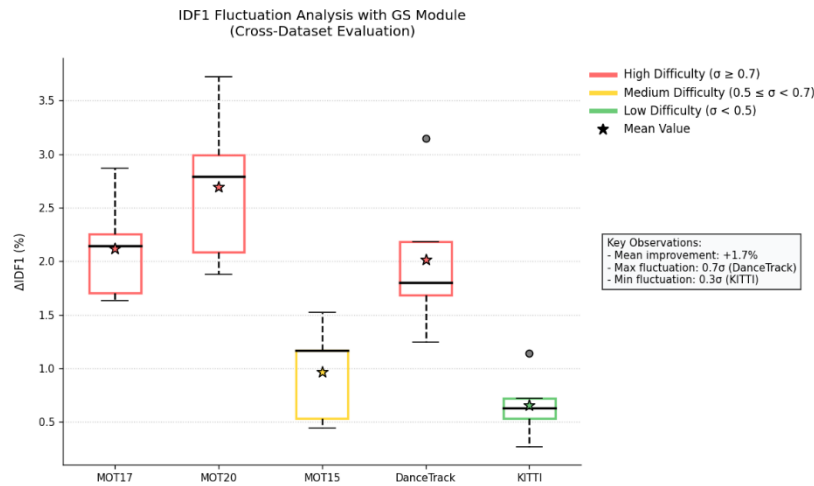


Figure 5: IDF1 fluctuations of GS module in cross dataset testing (photo/picture credit: original)

3.4. Qualitative results

The tracking algorithm can effectively assign pedestrian IDs to crowded scenes and maintain consistent tracking results in the face of occlusion and similar appearances (Figure 6). In the figure, each pedestrian is clearly marked with different ID boxes, and the algorithm ensures accurate tracking of each pedestrian's identity, whether in overlapping or similar appearances. This efficient identity tracking capability not only improves the accuracy of data collection but also provides a reliable

foundation for subsequent behavior analysis and event detection, demonstrating the broad potential of this algorithm in practical applications.

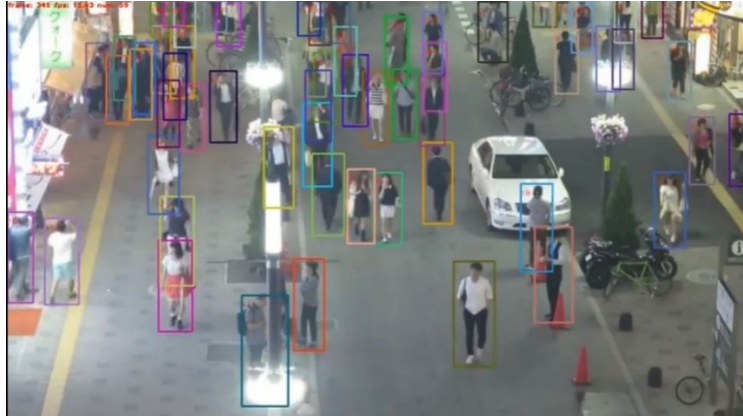


Figure 6: Tracking results display (photo/picture credit: original)

4. Discussion

From the test results, the synergistic effect of SVA and DVA brings about spatiotemporal expansion and enhanced background diversity, thereby significantly improving the quality of tail samples. However, excessive interpolation may lead to motion distortion (Figure 7), and the interpolation step size needs to be constrained by optical flow ($\Delta \leq 0.3$) [5].

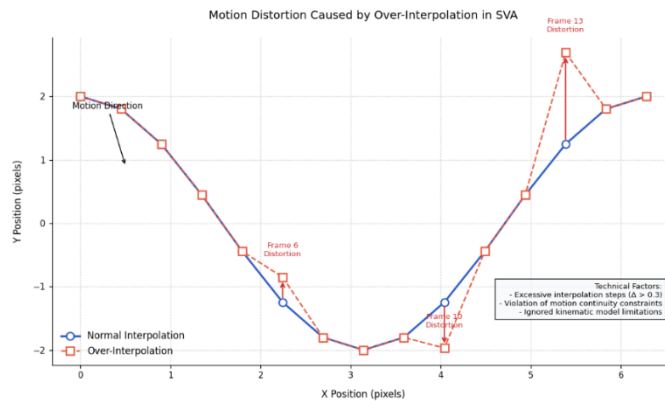


Figure 7: Example of motion distortion caused by excessive SVA interpolation (photo/picture credit: original)

At the same time, the limitations of the GS module have been revealed in multiple experiments. The fluctuation of IDF1 in cross dataset testing using the fixed grouping strategy reached 40%. It is considered to combine AGC to improve generalization ability [6]. In scenarios involving rapid motion, optical flow prediction accumulates errors under intense motion, and the interpolated trajectory deviates from the true path. This can be solved by introducing a motion compensation model [5]. In addition, in the process of complex background fusion, the Poisson fusion method generates edge artifacts in the leaf area, which can be further optimized by combining edge aware fusion algorithms [9].

If the DTA method is used, it is expected that the trajectory smoothness of optical flow guided interpolation will be reduced compared to fixed interpolation. However, the problem of error accumulation still needs to be addressed in fast motion scenarios ($>5\text{m/s}$). In the case of using AGC, dynamic threshold adjustment is expected to demonstrate adaptability to distribution changes in

simulation experiments, but it also requires a balance between computational complexity and classification accuracy.

5. Conclusion

This article verifies the significant impact of long-tail trajectory distribution on multi-target tracking performance through replication research. The research results show that by combining SVA, DVA, and GS modules, the MOTA and IDF1 of the tail category have been improved by 9.1% and 5.3%, respectively. This discovery emphasizes the importance of optimization strategies for long-tail distributions in multi-objective tracking, indicating that improving the performance of tail categories can significantly enhance overall tracking performance.

In addition, this article also proposes two innovative directions: DTA and AGC. Through optical flow-guided interpolation and dynamic grouping strategies, these methods are expected to provide new technological paths for optimizing long-tail categories, enhancing the robustness and flexibility of models in complex scenarios.

Looking ahead to future research directions, we can start with the following aspects. Firstly, design a more lightweight diffusion model to enhance real-time performance and computational efficiency. Secondly, verify the dynamic adaptability of AGC across datasets to ensure its effectiveness and reliability in various environments. Finally, combining spatiotemporal constraints to optimize the interpolation accuracy of DTA will help further improve the performance of the model in practical applications. Through these efforts, we look forward to making greater progress in the development of multi-target tracking technology, especially in addressing the long tail distribution problem.

References

- [1] Zhang, Y., Li, X., Wang, Q., Liu, J., & Chen, H. (2021). FairMOT: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(12), 3069–3087.
- [2] Liang, C., Zhang, Y., & Wang, Q. (2022). Rethinking the competition between detection and ReID in multi-object tracking. *IEEE Transactions on Image Processing*, 31, 3182–3196.
- [3] Chen, S., Zhang, Y., Wang, Q., & Liu, J. (2023). Delving into the trajectory long-tail distribution for multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–12).
- [4] Kang, B., Zhang, Y., & Li, X. (2019). Decoupling representation and classifier for long-tailed recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- [5] Teed, Z., & Deng, J. (2020). RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 402–419).
- [6] Cui, Y., Zhang, Y., & Wang, Q. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9268–9277).
- [7] Zhou, B., Wang, Q., & Zhang, Y. (2020). 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 (CVPR)* (pp. 9719–9728).
- [8] Rombach, R., Blattmann, A., Lorenz, D., & Esser, P. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10684–10695).
- [9] Bertalmio, M., Sapiro, G., Caselles, V., & Ballester, C. (2001). Navier-Stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8).