Mechatronics and Intelligent Transportation Systems
https://www.acadlore.com/journals/MITS

Enhancing Occluded Pedestrian Re-Identification with the MotionBlur Data Augmentation Module

Zhen Xue[1*], Teng Yao[2]

[1] School of Electronics and Information Engineering, Soochow University, 215006 Suzhou, China
[2] Fok Ying Tung Research Institute, The Hong Kong University of Science and Technology, 999077 Hong Kon, China

* Correspondence: Zhen Xue (20214228018@stu.suda.edu.cn)

Received: 02-12-2024 **Revised:** 03-16-2024 **Accepted:** 03-23-2024

Citation: Z. Xue and T. Yao, "Enhancing occluded pedestrian re-identification with the MotionBlur data augmentation module," *Mechatron. Intell Transp. Syst.*, vol. 3, no. 2, pp. 73–84, 2024. https://doi.org/10.56578/mits030201.

Abstract: In the field of pedestrian re-identification (ReID), the challenge of matching occluded pedestrian images with holistic images across different camera views is significant. Traditional approaches have predominantly addressed non-pedestrian occlusions, neglecting other prevalent forms such as motion blur resulting from rapid pedestrian movement or camera focus discrepancies. This study introduces the MotionBlur module, a novel data augmentation strategy designed to enhance model performance under these specific conditions. Appropriate regions are selected on the original image for the application of convolutional blurring operations, which are characterized by predetermined lengths and frequencies of displacement. This method effectively simulates the common occurrence of motion blur observed in real-world scenarios. Moreover, the incorporation of multiple directional blurring accounts for a variety of potential situations within the dataset, thereby increasing the robustness of the data augmentation. Experimental evaluations conducted on datasets containing both occluded and holistic pedestrian images have demonstrated that models augmented with the MotionBlur module surpass existing methods in overall performance.

Keywords: Pedestrian re-identification (ReID); Intelligent video surveillance; Occluded pedestrian re-identification (occluded-ReID); Vision transformer (ViT)

1 Introduction

In real-world video perception scenarios, occlusion is a prevalent issue in pedestrian images captured by cameras. This problem significantly affects the performance of existing ReID algorithms in practical settings [1–3]. Generally, occlusions in real-world scenarios are diverse, ranging from trees, buildings, and vehicles, where any surrounding object can potentially occlude the target pedestrian. Moreover, these occurrences are more frequent and widespread in locations where ReID technology is extensively applied, such as shopping malls, train stations, hospitals, and schools [4–6]. Therefore, there is a need to design a more robust model structure for occluded-ReID.

Currently, common data augmentation techniques used for ReID include random erasing, color jittering, random cropping, and rotation. These methods are aimed at reducing the risk of overfitting and enhancing the model's robustness to occlusion. However, when dealing with occluded pedestrian images, the diversity and randomness of occlusions often pose challenges for achieving satisfactory results using these generic augmentation methods. Therefore, tailored designs specifically addressing occlusion are often more effective in meeting the requirements, thereby further enhancing the model's ability to handle occlusion.

The main focus of this paper is to investigate the challenges encountered in tracking and locating target pedestrians in real-world scenarios, particularly when motion blur occurs due to the high-speed movement of the target pedestrian or focusing issues with the camera. We aim to address this issue by employing data augmentation methods to introduce random motion blur to images in the dataset, thereby enabling the model to better adapt to such scenarios. As illustrated in Figure 1, motion blur resembles a visual residue phenomenon resulting from small incremental displacements in a particular direction of certain parts of the target pedestrian's body. Generally, humans can easily discern specific motion changes and make informed judgments, whereas machines often struggle with this task. For instance, in the scenarios described, the presence of motion blur not only affects ReID but also introduces

corresponding flaws in keypoint detection techniques [7–9], leading to confusion of the human body's specific structure and erroneous results, consequently resulting in a decrease in model accuracy.

Figure 1. The motion blurred image after the human key point detection

As previously mentioned, enhancing the model's capability to handle motion blur is crucial. However, current data augmentation techniques often struggle to achieve this, leading to limited improvements in model performance. Traditional motion blur methods typically apply blur uniformly across the entire image, and adjusting the associated coefficients usually results in minor variations over a broad scale. This approach often adversely affects the background and fails to effectively address real-world scenarios. To address this challenge, we have developed an enhancement occlusion module specifically designed for motion blur, allowing for significant blur operations within localized regions. Specifically, by judiciously extracting and expanding a portion of the pedestrian's body to simulate motion blur, our data augmentation approach introduces motion blur effects to images in the dataset, thereby improving the model's ability to handle such scenarios. Through extensive experimentation on occluded pedestrian datasets (Occluded-DukeMTMC, Partied-REID [10], and Occluded-REID [11]) and comprehensive datasets (Market1501 [12] and DukeMTMC-reID [13]), we have validated the effectiveness of our proposed method. When compared to existing methods, our approach achieves higher Rank-1 and mean Average Precision (mAP) accuracies, demonstrating its superiority in handling motion blur challenges.

2 Related Works

In this section, we will provide a brief overview of existing methods in the fields of general ReID and occluded-ReID.

2.1 Holistic Person Re-Identification

ReID aims to retrieve target individuals of interest from different camera views and has made significant strides in recent years. Existing ReID methods can be broadly categorized into three types: manually annotated methods [14, 15], metric learning methods [16, 17], and deep learning methods [18–20]. With the advent of large-scale datasets and the proliferation of Graphics Processing Units (GPUs), deep learning-based approaches have become predominant in today's pedestrian ReID domain. Recent efforts have primarily focused on leveraging part-based features to achieve state-of-the-art performance in overall pedestrian ReID. Zhang et al. [21] achieved automatic alignment of part features through shortest path loss during the learning process, eliminating the need for additional supervision or explicit pose information. A generic method for learning features at the part level was proposed, which can be adapted to different strategies for partitioning parts. This method incorporates attention mechanisms to ensure that the model emphasizes the human body region, leading to the extraction of more effective features [22, 23]. However, these methods often struggle to achieve high accuracy in the presence of occlusion. These limitations hinder the practical applicability of the methods, particularly in common, crowded scenarios.

2.2 Occluded Person Re-Identification

Research on occluded-ReID introduced a novel approach. The training set and gallery set are generally constructed from images of unobstructed pedestrians, while the query set is constructed from images of occluded pedestrians. Currently, research methods in this field can be divided into two categories: pose estimation-assisted [24, 25] and manually annotated parsing [26, 27]. Gao et al. [8] proposed a Pose-guided Visible Part Matching (PVPM) method, which jointly learns discriminative features with a pose-guided attention mechanism and self-discovers the visibility of part-level features within an end-to-end framework. He and Liu [24] introduced a new method called Pose-Guided Feature Alignment (PGFA), which separates useful information from occlusion noise using pose landmarks. Zhao et al. [28] proposed a model called HPNet for extracting part-level features and predicting the visibility of each part based on manual parsing. This method extracts features from semantic regions, compares them considering visibility, reduces background noise, and achieves pose alignment.

In contrast to the aforementioned methods, our approach does not rely on additional models. By addressing potential motion blur phenomena that may occur in real-world scenarios, our model is better equipped to handle such occurrences, enhancing its robustness to motion blur and effectively improving its performance in addressing this issue.

3 Motion Blur Enhancement Module

The overall architecture of the network is illustrated in Figure 2. The original image passes through our motion blur module, and then the enhanced image is fed into our feature extractor. In our experiments, we simply employ a ViT [29, 30] as the feature extractor. For an input image $x \in \mathbb{R}^{3*h*w}$, it is first divided into N sequences of patches, where each patch has a resolution of (P, P), and then positional embeddings and a classification [cls] token are attached to the input image. The output features for each image are represented as $f \in \mathbb{R}^{(N+1)*c}$, where $\mathrm{N}+1$ denotes the image tokens and one [cls] token, and c represents the channel dimension. In our experimental setup, n and c are set to 128 and 768, respectively. The extracted features undergo computation through fully connected layers to calculate both the triplet loss for feature embedding and the classification ID loss, thereby deriving the final results.

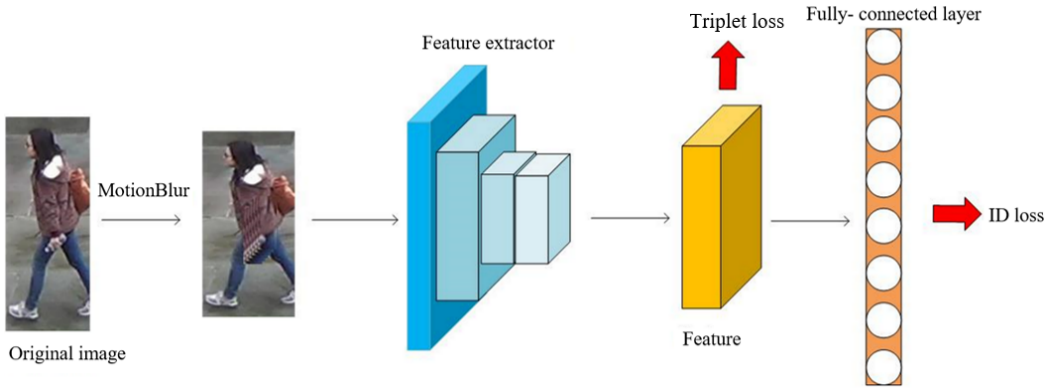

Figure 2. Overall network framework

Inspired by techniques such as light painting and other photography methods utilizing multiple shutter releases within a short time frame to capture the dynamic essence of moving objects, we aim to replicate a similar effect akin to visual residual phenomena. By partially displacing portions of the image at the pixel level, we simulate an effect more closely resembling motion blur. This approach mimics the scenario where multiple images of a pedestrian in motion are captured within the same time interval, akin to the effect generated by stacking multiple exposures of the same subject within a single frame. The specific procedure for this operation will be outlined below.

Given an image $x \in \mathbb{R}^{3*h*w}$, where h and w represent the height and width of the image, respectively, in typical experiments, unless otherwise specified, the shape of the image will be uniformly reshaped to 256×128. First, we need to determine the parts of the image that require motion blur. Since surveillance cameras in reality often maintain a fixed perspective over short periods of time, background images tend to remain static without blur. Therefore, when simulating motion blur in images, we aim for the blur effect to predominantly appear in the human body rather than the background. According to the dataset, due to prior manual processing and annotations, pedestrian parts in images tend to occur closer to the center. Moreover, based on the physical shape of the human body, it tends to have a slender appearance, with the upper body occupying a larger proportion of the image. Thus, the extraction range will have fewer portions removed from the left and right sides and more from the top and bottom to ensure sufficient extraction of the body parts. Therefore, motion blur extraction range can be obtained from the following Eq. (1) and Eq. (2):

$$P_{h\,\min} = h * \omega_h, P_{h\,\max} = h * (1 - \omega_h) \tag{1}$$

$$P_{w\,\min} = w * \omega_w, P_{h\,\max} = w * (1 - \omega_w) \tag{2}$$

where, $0.5 > \omega_h > \omega_w$, ensuring that the width of the extraction range is larger than the height, aligning with the distribution of human bodies in the dataset images. Therefore, the final determined range is determined by these four values, namely $(P_{h\,\min}, P_{w\,\min})$, $(P_{h\,\max}, P_{w\,\min})$, $(P_{h\,\min}, P_{w\,\max})$, $(P_{\mathrm{hmax}}, P_{\mathrm{wmax}})$, forming a plane space enclosed by four points. By confining the extraction range in this manner, more body parts are included while minimizing excessive background portions. After obtaining the extraction range, considering subsequent operations

such as displacement, it's essential to ensure that the operations do not extend beyond the original boundaries of the image. Otherwise, it would be futile and counterproductive. Therefore, not the entire extraction range is designated for motion blur. Instead, within the selected extraction range, heights and widths are cropped to:

$$P_H = \frac{1}{2}\left(P_{h\max} - P_{h\min}\right) \tag{3}$$

$$P_W = \frac{1}{2}\left(P_{w\max} - P_{w\min}\right) \tag{4}$$

The area ultimately selected based on the aforementioned formulas is $P_H * P_W$ This ensures that a sufficiently large area is chosen to encompass an ample portion of the human body, thereby guaranteeing the desired effect of motion blur. Additionally, it prevents the final image from exceeding the original boundaries after displacement, thereby preserving the intended effect of the enhancement strategy.

Having determined the extraction range and the size of the extracted portion, the next step is to decide the direction of displacement for the blurred portion and the selection of the initial displacement point. The displacement direction should align with the likely movement direction of pedestrians captured by the camera at the time. Upon analyzing images in the dataset, it's observed that most pedestrians predominantly move either left or right due to the camera's perspective. Additionally, due to variations in camera height, pedestrians may also move in directions such as upper left, lower left, upper right, and lower right. These six directions encompass the majority of pedestrian movement directions. Hence, we opt to use a probability, denoted as pdis, to randomly determine the displacement direction of the extracted portion in the image. Considering the varying proportions of different directions in the image, we assign a higher probability to left and right directions $p_{\text{left}} = p_{\text{right}} = 0.4$, and lower probabilities to upper left, lower left, upper right, and lower right directions $p_{\text{top left}} = p_{\text{bottleleft}} = p_{\text{topright}} = p_{\text{bottleright}} = 0.05$, such that $p_{\text{all}} = p_{\text{left}} + p_{\text{right}} + p_{\text{topleft}} + p_{\text{bottleleft}} + p_{\text{topright}} + p_{\text{bottleright}} = 1$. Furthermore, to ensure the rationality of the blurred image after displacement, the initial displacement point should also vary according to the chosen displacement direction. Given that there is relatively more space left in the vertical direction within the extraction range, we only consider the influence of the initial displacement point in the horizontal direction. In summary, the specific selection of the initial displacement point is as shown in Eq. (5) and Eq. (6):

$$x_{\text{init}} = \begin{cases} \text{Random}\left(P_{\text{wmin}}, \frac{1}{2}\left(P_{\text{win}} + P_{\text{wmax}}\right)\right), & p_{\text{dis}} = p_{\text{right}}/p_{\text{topright}}/p_{\text{bottleright}} \\ \text{Random}\left(\frac{1}{2}\left(P_{\text{wmin}} + P_{\text{wmax}}\right), P_{\text{wmax}}\right), & p_{\text{dis}} = p_{\text{left}}/p_{\text{topleft}}/p_{\text{bottleleft}} \end{cases} \tag{5}$$

$$y_{\text{init}} = \text{Random}\left(P_{\text{hmin}}, P_{\text{hmax}}\right) \tag{6}$$

From the aforementioned formulas, it can be concluded that the extracted portion is defined by a rectangular area enclosed by four points: $(x_{\text{init}}, y_{\text{init}})$, $(x_{\text{init}} + P_W, y_{\text{init}})$, $(x_{\text{init}}, y_{\text{init}} + P_{WH})$, $(x_{\text{init}} + P_W, y_{\text{init}} + P_H)$.These equations allow for the generation of the required initial displacement points within the specified range, ensuring that the resulting blurred image remains within the original boundaries of the image while accommodating various potential displacement directions as observed in the dataset. As vertical blur effects are largely unaccounted for in the scenario, a traditional motion blur convolution kernel is subsequently applied to the displacement portion to further enhance the blur effect, as depicted in Figure 3. In summary, the final creation of motion-blurred displacement segments aims to closely replicate real-world motion blur phenomena.

$$\begin{bmatrix} 0 & 0 & 1/5 & 0 & 0 \\ 0 & 0 & 1/5 & 0 & 0 \\ 0 & 0 & 1/5 & 0 & 0 \\ 0 & 0 & 1/5 & 0 & 0 \\ 0 & 0 & 1/5 & 0 & 0 \end{bmatrix}$$

Figure 3. Vertical motion blur convolution kernel

Finally, consideration must be given to the length of displacement l, frequency f, and mode. Regarding length and frequency, they are related to the selected displacement direction. If the direction is left or right, only displacement along the x-axis is involved, excluding movement along the y-axis. However, for other directions, displacement along the y-axis accompanies movement along the x-axis. Further methods and selections require emulation to closely resemble real-world effects, necessitating additional experimentation to determine the appropriate proportions between length and frequency. Specific experimental designs and results will be presented in the next chapter's

parameter analysis. As for the mode, it involves how displacement is realistically depicted in the image. One approach is to sequentially displace the selected portion of the image from the initial point by the determined length for each frequency. However, this method results in the gradual advancement of the image, with the final position overlaying a complete original image, obscuring the original position with various objects. Observing common motion blur images resulting from high-speed movement, the visual residual effect typically appears behind the moving subject, resembling a trailing effect. Thus, an alternative method is selected here. The image is gradually displaced and pasted from the furthest position (e.g., $(x_{\text{init}} + l * f, y_{\text{init}} + l * f)$) towards the original position, creating a partial body due to visual persistence. This approach minimizes the impact on the original body position, ensuring a degree of image integrity and simulating the lag effect associated with motion blur more accurately.

In summary, the specific motion blur enhancement module will encompass aspects such as extraction range determination, extracted portion size determination, and displacement direction selection. It is through the aforementioned steps that optimal experimental results can be achieved. The overall process is illustrated in Figure 4.

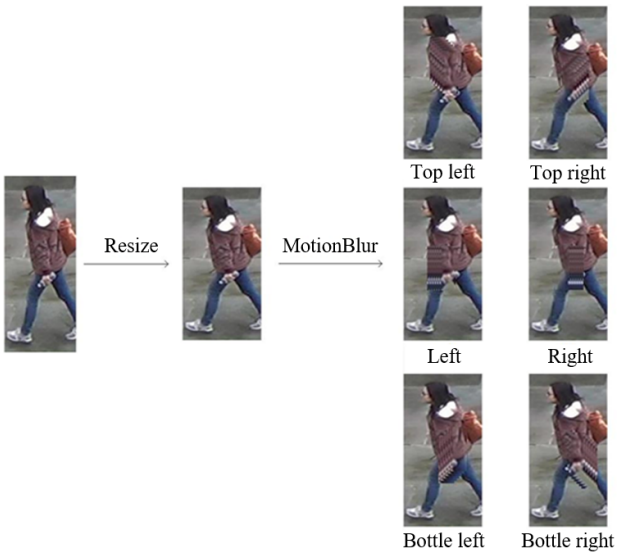

Figure 4. The results of the motion blur module are displayed

4 Experimental Results

4.1 Dataset and Evaluation Metrics

Occluded-dukemtmc consists of 15,618 training images of 702 individuals, 2,210 query images of 519 individuals, and 17,661 gallery images of 1,110 individuals. Due to the diversity of scenes and interferences, it is the most challenging occluded-ReID dataset.

Occluded-ReID is a dataset for occluded-ReID captured by a moving camera. It comprises 2000 images belonging to 200 identities, with each identity having 5 full-body images and 5 heavily occluded images with different viewpoints and types of occlusions.

Partial-ReID is a specially designed ReID dataset consisting of images of pedestrians with occlusions, partial views, and full views. It consists of 600 images of 60 individuals. We conduct experiments using the occluded person query set and the full-body person gallery set.

Market-1501 is a well-known dataset for whole-body person ReID. It includes 12,936 training images of 751 individuals, 19,732 query images, and 3,368 gallery images of 750 individuals captured by 6 cameras. This dataset contains very few occluded images.

DukeMTMC-reID comprises 16,522 training images, 2,228 query images, and 17,661 gallery images of 702 individuals. These images were captured by eight different cameras, making it more challenging. Since this dataset contains more whole-body images than occluded images, it can be considered a whole-body ReID dataset.

Evaluation Metrics: To ensure fair comparison with existing person identification methods, all methods are evaluated under Cumulative Matching Characteristics (CMC) and mAP. All experiments are conducted in a single-query setting.

4.2 Implementation Details

If not otherwise specified, all images in the datasets are resized to 256×128. We train our network end-to-end using the SGD optimizer with a momentum of 0.9 and a weight decay of 1e-4. The initial learning rate is set to

0.008, with cosine learning rate decay. For each input branch, the batch size is set to 64, comprising 16 labels with 4 samples per label. All testing experiments are conducted on a single RTX 3090 GPU.

4.3 Comparison of Occluded Pedestrian Datasets

The results for Occluded-DukeMTMC (O-Duke), Occluded-REID (O-REID), and Partial-REID (P-REID) are shown in Table 1. Since O-REID and P-REID do not have corresponding training sets, we directly test models trained on Market-1501.

Table 1. Performance comparison with other methods on the occluded-person dataset

Method	O-Duke		O-REID		P-REID	
	R@1	mAP	R@1	mAP	R@1	mAP
PCB	42.6	33.7	41.3	38.9	66.3	63.8
RE	40.5	30.0	-	-	54.3	54.4
FD-GAN	40.8	-	-	-	-	-
DSR	40.8	30.4	72.8	62.8	73.7	68.1
SFR	42.3	32	-	-	56.9	-
FRR	-	-	78.3	68.0	81.0	76.6
PVPM	47	37.7	70.4	61.2	-	-
PGFA	51.4	37.3	-	-	69.0	61.5
HOReID	55.1	43.8	80.3	70.2	85.3	-
OAMN	62.6	46.1	-	-	86.0	-
PAT	64.5	53.6	81.6	72.1	88.0	-
ViT	60.5	53.6	81.6	72.1	73.3	74.0
TransReID	64.2	55.7	70.2	67.3	71.3	68.6
Denseformer	63.8	55.6	-	-	-	-
ResT-ReID	59.6	51.9	-	-	-	-
DRL-Net	65.0	50.8	-	-	-	-
DAAT	63.3	57.1	-	-	-	-
FED	67.9	56.3	86.3	79.3	83.1	80.5
MB	68.1	56.6	86.8	81.2	83.3	80.4

Specifically observing the results in the table, PAT [31] adopts ResNet50 [32] as the backbone and employs a transformer-based encoder-decoder structure for multi-part discovery. The prototypes in the network act as specific feature detectors, crucial for improving the network's performance on occluded data. TransReID is the first pure Transformer-based ReID architecture. The results utilize ViT as the main framework without setting sliding windows as the backbone, with image sizes also adjusted to 256×128.

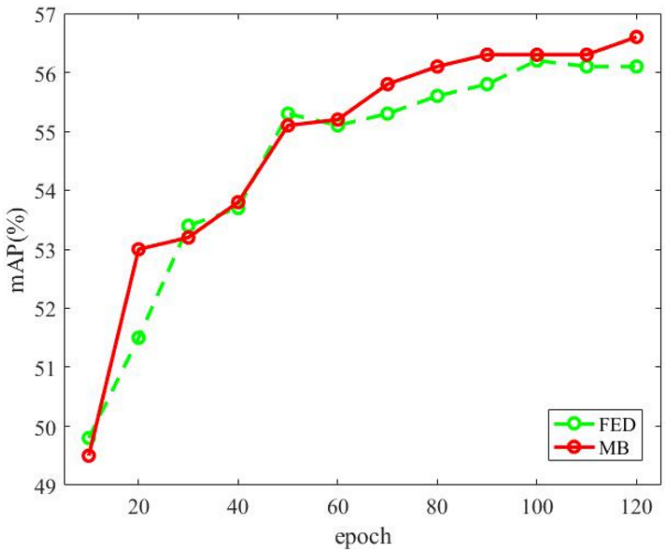

Figure 5. mAP curve on the O-Duke

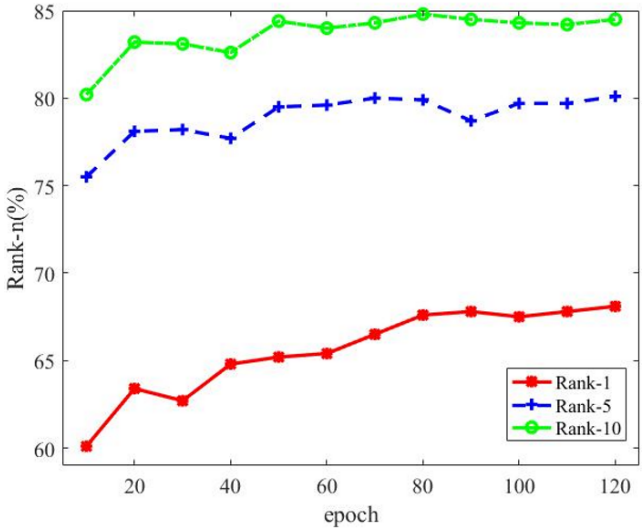

Figure 6. Rank-N curve on the O-Duke

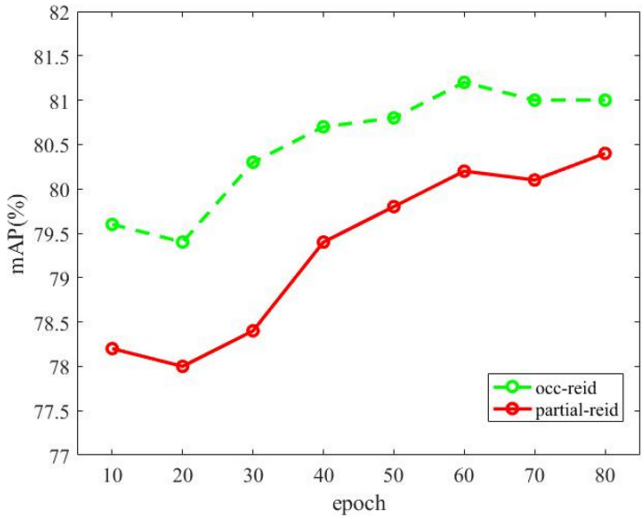

Figure 7. mAP curve on the O-REID and P-REID

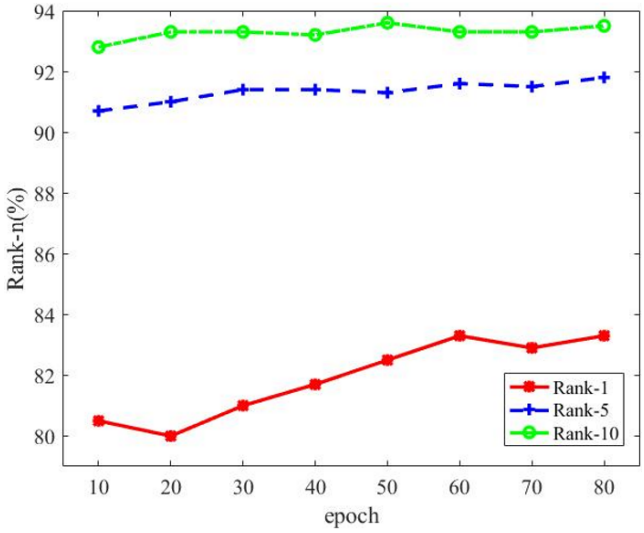

Figure 8. mAP curve on the O-REID

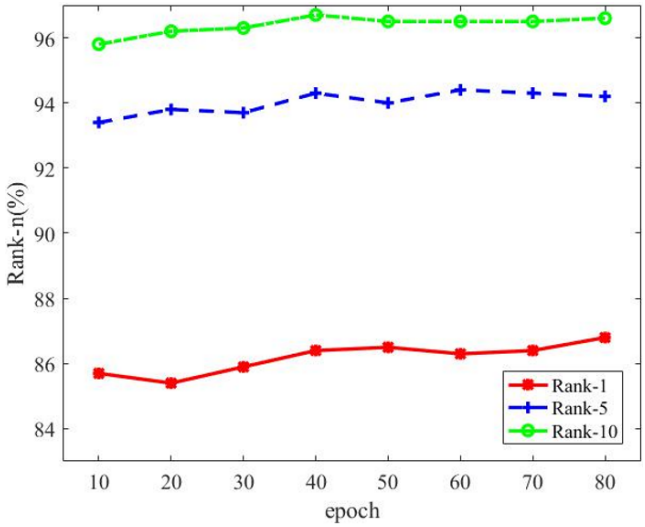

Figure 9. Rank-N curve on the P-REID

The ViT Baseline performs better on O-REID and P-REID datasets compared to TransReID, as TransReID uses many dataset-specific tokens, reducing the model's cross-domain generalization and increasing the risk of overfitting, leading to decreased performance in cases where valid information cannot be effectively extracted for the dataset. When comparing with our method, it is evident that we achieve the best performance on Occluded-REID datasets in terms of both Rank-1 and mAP metrics. Particularly on the Occluded-REID dataset, there are improvements of 0.5% and 1.9% in Rank-1 and mAP respectively, considering significant improvements over previous performances. Additionally, in Figures 5-9, we demonstrate the mAP curves and Rank-N curves on Occluded-DukeMTMC, Occluded-REID, and Partial-REID, confirming the effectiveness of our method on all three datasets. It can also be observed that the changes in Rank-5 and Rank-10 are relatively small; thus, the previous table only selected the values of mAP and Rank-1 without listing all numerical values.

4.4 Comparison of Holistic Pedestrian Datasets

We also conducted experiments on the holistic person ReID datasets, including Market-1501 (MARKET) and DukeMTMC-reID (MTMC). The results are presented in Table 2. In contrast to the performance on occluded pedestrian datasets, we achieved relatively better performance compared to other existing methods, but there is some gap compared to methods specifically designed for overall person datasets. TransReID, without sliding window settings, utilized image sizes of 256×128. It is evident that TransReID outperformed our method on overall datasets. This is because TransReID is specifically designed for these whole-body person datasets and encodes additional information regarding camera viewpoints and identity labels during training.

Table 2. Performance comparison with other methods on the holistic-person dataset

Method	**Market-1501**		**DukeMTMC-reID**	
	R@1	mAP	R@1	mAP
PT	87.7	68.9	78.5	56.9
PGFA	91.2	76.8	82.6	65.5
PCB	92.3	77.4	81.8	66.1
OAMN	92.3	79.8	86.3	72.6
BoT	94.1	85.7	86.4	76.4
HOReID	94.2	84.9	86.9	75.6
ViT	94.7	86.8	88.8	79.3
TransReID	95.0	88.2	89.6	80.6
DRL-Net	94.7	86.9	88.1	76.6
PAT	95.4	88.0	88.8	78.2
FED	95.0	86.3	89.4	78.0
MB	95.0	86.7	89.6	78.4

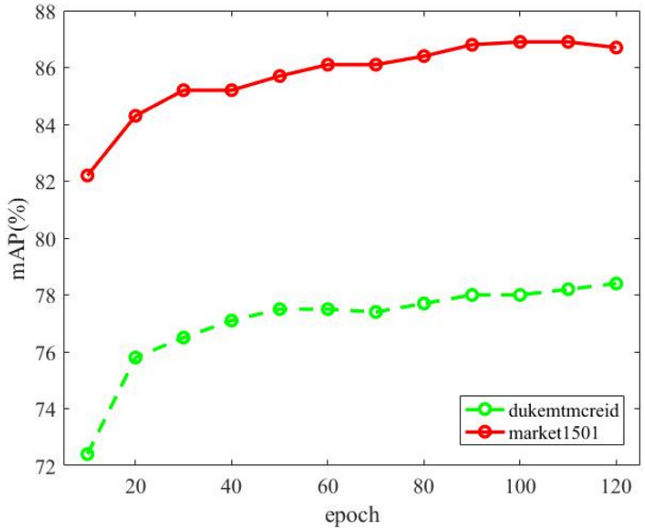

Figure 10. mAP curve on the Market-1501 and DukeMTMC-reID

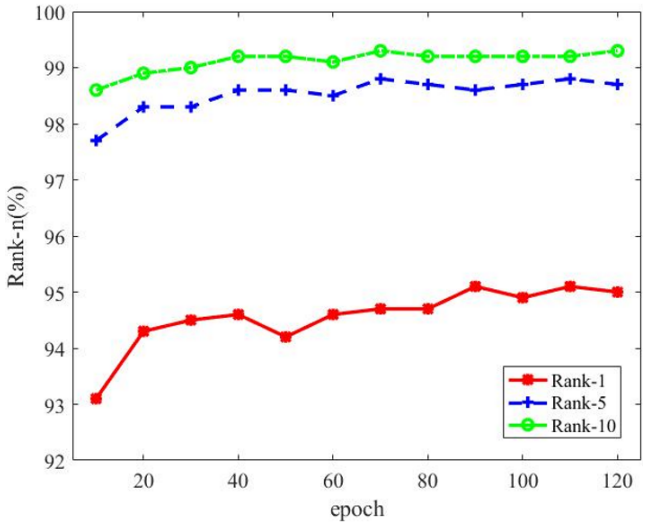

Figure 11. mAP curve on the MARKET

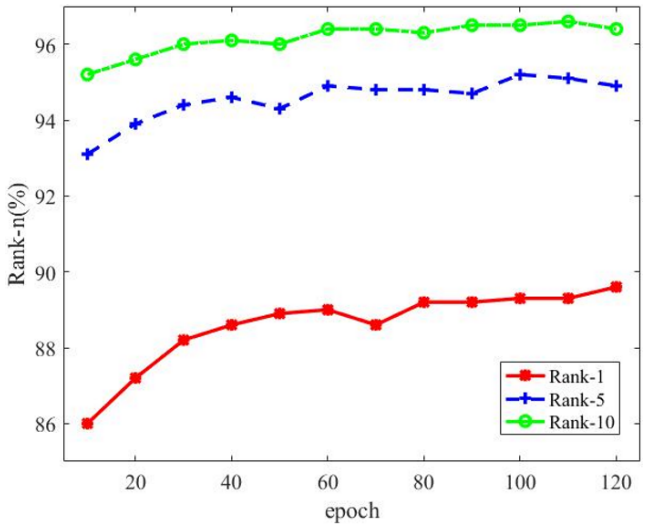

Figure 12. Rank-N curve on the MTMC

While our proposed method may still have room for improvement on overall person datasets, simultaneously, our method also achieved a Rank-1 accuracy of 89.6% on the DukeMTMC-reID dataset, surpassing other CNN-based methods and reaching parity with TransReID. Similar to the experimental results on occluded pedestrian datasets, we also demonstrate the mAP curves and Rank-N curves on Market-1501 and DukeMTMC-reID in Figures 10-12, confirming the effectiveness of our method on these two overall person datasets.

4.5 Qualitative Analysis

Meanwhile, we conducted further experiments on the displacement length l and frequency f mentioned earlier. To ensure a closer approximation to motion blur effects encountered in real-world scenarios, l should not be set too short, as this would result in adjacent pixel values being identical, affecting the model's judgment. Conversely, setting l too high would deviate significantly from realistic simulated scenes. The frequency f primarily controls the size of the simulated parts, ensuring they remain perceptible to the model without exceeding the original image boundaries, thus aiding in enhancing the model's performance. Based on the above considerations, we conducted the following experiments on the Occluded-DukeMTMC (O-Duke) dataset, with results detailed in Table 3.

Table 3. Performance of different lengths l and frequency f on the O-Duke dataset

Datasets	**Occluded-DukeMTMC (O-Duke)**			
l	1	2	3	4
f	8	8	7	6
mAP	55.8	56.1	56.6	56.4
R@1	67.5	67.9	68.1	68.0
R@5	79.8	80.2	80.1	79.7

The experimental data in the table shows that the differences in performance based on different values of l and f are not substantial. Therefore, we utilized the configuration where $1 = 3$ and $f = 7$, which yielded relatively optimal results. The data from the table also indicates that overall, our approach achieved favorable results on the remaining occluded pedestrian datasets and overall person datasets.

5 Conclusion

In this study, we address the challenge of motion blur arising from high-speed motion of target pedestrians or focus issues with the camera by proposing a novel data augmentation module called MotionBlur to enhance the model's robustness to this problem. Specifically, we analyze the initial images to select appropriate regions for blurring, capturing suitable occlusion sizes. We simulate various directions of motion blur to mimic real-world scenarios and integrate them with traditional motion blur methods. This enables us to effectively simulate motion blur on body parts within a small range, enhancing the model's performance. We conducted experiments on multiple occluded pedestrian datasets and overall person datasets, achieving relatively favorable results compared to existing methods on conventional ReID benchmarks, thus demonstrating the effectiveness of our approach.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, 2021. https://doi.org/10.1109/TPAMI.2021.3054775

[2] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Learning to adapt invariance in memory for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2723–2738, 2020. https://doi.org/10.1109/TPAMI.2020.2976933

[3] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian, "Unsupervised person re-identification via softened similarity learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 3387–3396. https://doi.org/10.1109/CVPR42600.2020.00345

[4] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, 2018. https://doi.org/10.1109/TCSVT.2018.2873599

[5] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan, "Pose-guided multi-granularity attention network for text-based person search," *AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 11 189–11 196, 2020. https://doi.org/10.1609/aaai.v34i07.6777

[6] J. Yang, W. S. Zheng, Q. Yang, Y. C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 3286–3296. https://doi.org/10.1109/CVPR42600.2020.00335

[7] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 5686–5696. https://doi.org/10.1109/CVPR.2019.00584

[8] S. Gao, J. Wang, H. Lu, and Z. Liu, "Pose-guided visible part matching for occluded person ReID," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 11 744–11 752. https://doi.org/10.1109/CVPR42600.2020.01176

[9] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019, pp. 542–551. https://doi.org/10.1109/ICCV.2019.00063

[10] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," *Pattern Recognit.*, vol. 86, pp. 143–155, 2019. https://doi.org/10.1016/j.patcog.2018.08.015

[11] J. Zhuo, Z. Chen, J. Lai, and G. Wang, "Occluded person re-identification," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, San Diego, CA, USA, 2018, pp. 1–6. https://doi.org/10.1109/ICME.2018.8486568

[12] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1116–1124. https://doi.org/10.1109/ICCV.2015.133

[13] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 3774–3782. https://doi.org/10.1109/ICCV.2017.405

[14] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image Vis. Comput.*, vol. 32, no. 6-7, pp. 379–390, 2014. https://doi.org/10.1016/j.imavis.2014.04.002

[15] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Computer Vision–ECCV 2014: 13th European Conference*, Zurich, Switzerland, 2014, pp. 536–551. https://doi.org/10.1007/978-3-319-10590-1_35

[16] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 1320–1329. https://doi.org/10.1109/CVPR.2017.145

[17] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 3652–3661. https://doi.org/10.1109/CVPR.2017.389

[18] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 480–496. https://doi.org/10.1007/978-3-030-01225-0_30

[19] Z. Wang, L. He, X. Gao, and J. Shen, "Robust person re-identification through contextual mutual boosting," *arXiv preprint arXiv:2009.07491*, 2020. https://doi.org/10.48550/arXiv.2009.07491

[20] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 2119–2128. https://doi.org/10.1109/CVPR.2018.00226

[21] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "AlignedReID: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017. https://doi.org/10.48550/arXiv.1711.08184

[22] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 2285–2294. https://doi.org/10.1109/CVPR.2018.00243

[23] C. P. Tay, S. Roy, and K. H. Yap, "AANet: Attribute attention network for person re-identifications," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 7127–7136. https://doi.org/10.1109/CVPR.2019.00730

[24] L. He and W. Liu, "Guided saliency feature learning for person re-identification in crowded scenes," in *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, 2020, pp. 357–373. https://doi.org/10.1007/97

8-3-030-58604-1_22

[25] L. He, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019, pp. 8449–8458. https://doi.org/10.1109/ICCV.2019.00854

[26] H. Huang, X. Chen, and K. Huang, "Human parsing based alignment with multi-task learning for occluded person re-identification," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, London, UK, 2020, pp. 1–6. https://doi.org/10.1109/ICME46284.2020.9102789

[27] S. Yu, D. Chen, R. Zhao, H. Chen, and Y. Qiao, "Neighbourhood-guided feature reconstruction for occluded person re-identification," *arXiv preprint arXiv:2105.07345*, 2021. https://doi.org/10.48550/arXiv.2105.07345

[28] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 6230–6239. https://doi.org/10.1109/CVPR.2017.660

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. https://doi.org/10.48550/arXiv.2010.11929

[30] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 14 993–15 002. https://doi.org/10.1109/ICCV48922.2021.01474

[31] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021, pp. 2897–2906. https://doi.org/10.1109/CVPR46437.2021.00292

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015. https://doi.org/10.48550/arXiv.1512.03385