Seq-Masks: Bridging the gap between appearance and gait modeling for video-based person re-identification

Zhigang Chang*, Zhao Yang*, Yongbiao Chen†, Qin Zhou‡, Shibao Zheng*
*Institute of Image Processing and Network Engineering, Shanghai Jiao Tong University, Shanghai, China
‡School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
†Shanghai Key Laboratory of Scalable Computing and Systems, Shanghai Jiao Tong University, Shanghai, China

2021 International Conference on Visual Communications and Image Processing (VCIP) | 978-1-7281-8551-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/VCIP53242.2021.9675368

***Abstract*—Video-based person re-identification (Re-ID) aims to match person images in video sequences captured by disjoint surveillance cameras. Traditional video-based person Re-ID methods focus on exploring appearance information, thus, vulnerable against illumination changes, scene noises, camera parameters, and especially clothes/carrying variations. Gait recognition provides an implicit biometric solution to alleviate the above headache. Nonetheless, it experiences severe performance degeneration as camera view varies. In an attempt to address these problems, in this paper, we propose a framework that utilizes the sequence masks (SeqMasks) in the video to integrate appearance information and gait modeling in a close fashion. Specifically, to sufficiently validate the effectiveness of our method, we build a novel dataset named MaskMARS based on MARS. Comprehensive experiments on our proposed large wild video Re-ID dataset MaskMARS evidenced our extraordinary performance and generalization capability. Validations on the gait recognition metric CASIA-B dataset further demonstrated the capability of our hybrid model. Our codes and dataset MaskMARS will be open-sourced as a strong baseline.**

***Index Terms*—Multi-modal Fusion, Video-based Person Re-ID, Gait Recognition, appearance model**

I. Introduction

Video-based Person re-identification (Re-ID) [1]–[6] has been drawing proliferating attention from researchers worldwide because of its ubiquitous presence in diversified scenarios ranging from cross-camera object tracking, pedestrian behavior analysis, video surveillance to criminal investigations. Traditional RGB-based, Video-based Person Re-ID methods seeking to build a discriminative appearance model has gained growing momentum in recent years. But the prerequisite is harsh and fatal. It is vulnerable against illumination changes, camera parameters, and scene clutter; the time-effectiveness of the model is short due to clothing change; uniforms and camouflage can easily paralyze the system. Consequently, it is rather reasonable to ponder the possibility of investigating auxiliary information or characteristics.

Gait [7]–[9] which manifests the walking style of pedestrians comes to rescue, given its unique advantages ranging from being insensitive to resolution variation to being extremely challenging to impersonate, thus, highlighting the necessity of employing gait feature into performing person re-id tasks [8]. Some gait recognition methods [10] aim to extract implicit features from videos based on contour or articulated body representations(e.g., UV maps and 2d keypoints), where semantic segmentation (Mask-RCNN [11]) and human pose estimation (DensePose [12] and OpenPose [13]) can be utilized to model the discriminative representations. However, this stream of works is limited by pedestrian speed, camera perspective, video frame rate, and other factors, leading to low gait recognition performance, especially in practical scenarios. Since video-based person Re-ID methods regard re-identification as sequence matching, which is also challenging due to the random camera view angles and wild scenarios. The goals and settings of these two branches have strong consistency, which naturally leads us to integrate them for better Re-ID performance in wild scenarios.

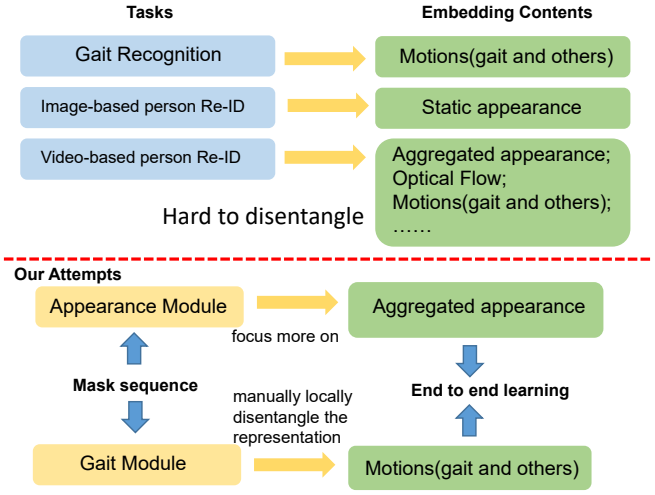

Fig. 1. Motivation

In traditional video-based Re-ID, 3D convolution, RNN models, and temporal pooling are commonly used to build an embedding for a video sequence, which integrates multiple information like appearance, motions (gait and other motions), optical flow, etc,. These representations are tightly bound and hard to disentangle. This representation lacks interpretability and may be hard to learn. Our motivation is somehow trying to locally disentangle the gait representation manually by adding a gait branch to a video Re-ID framework. As shown in Fig. 1, the enhanced gait representation serves as strong prior to joining the final embedding.

In this paper, we utilize the foreground masks of the video

Funded by NSFC: 62071292, 61771303 and STCSM 18DZ2270700.

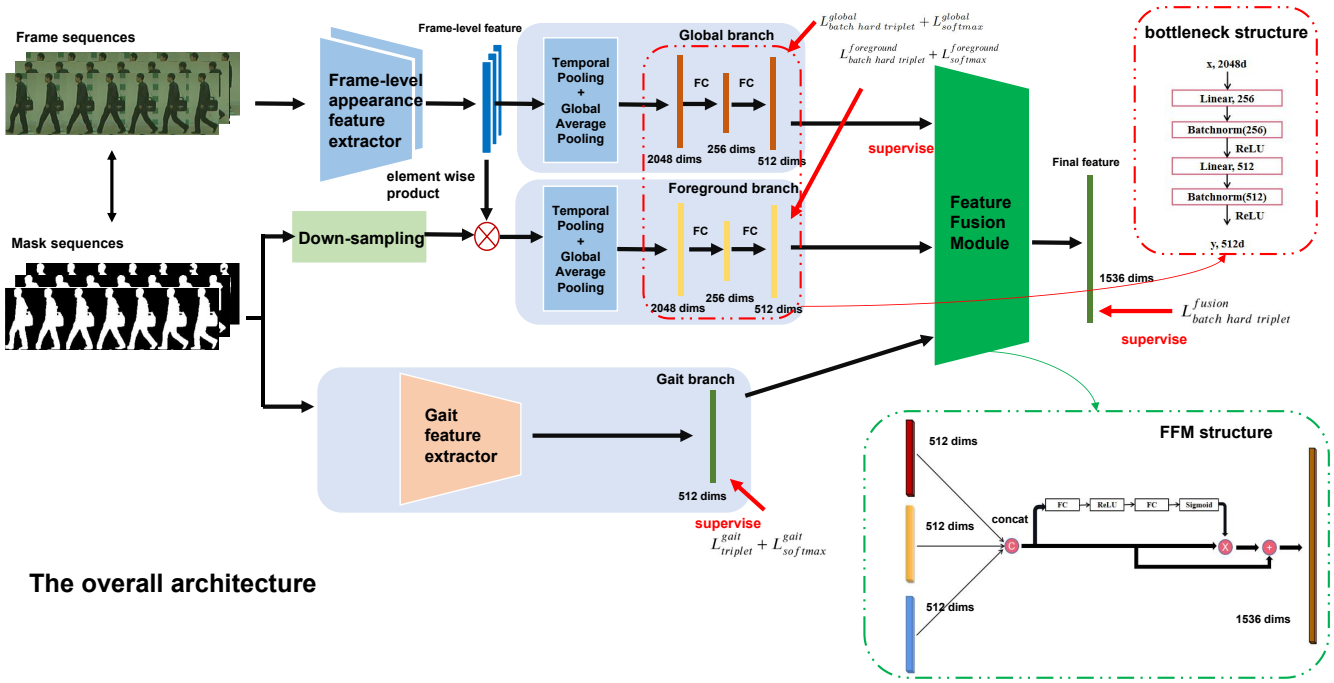

Fig. 2. Detailed structure of our Framework

sequence to bridge the appearance branch and the gait branch. For the gait branch, we adopt a variant of an advanced gait recognition network Gaitset [10] where the foreground sequence masks serve as the inputs of the gait branch. For the appearance branch, the foreground sequence masks can be regarded as the saliency map to highlight the foreground human semantic and eliminate background interference. Meanwhile, we also keep the original global branch to reserve global information of the sequence. Finally, we concatenate all the features after each branch and integrate them into a fused representation.

In summary, our contributions are three-fold: First, we proposed a novel end-to-end video person id framework which exploits both appearance and gait information. Second, we built a novel large-scale video Re-id dataset named **Mask-MASRS**. Third, extensive experiments have demonstrated the validity and effectiveness of our proposed model on both **Mask-MARKS** and **CASIA-B** datasets.

II. Method

The overall network architecture is shown in Fig 2. The network contains 3 modules in total: Appearance Module, Gait Module, and Feature Fusion Module. The appearance module is comprised of the global branch and the foreground branch. For each input video sequence (each sequence contains T frames), we use foreground extraction methods (such as Mask RCNN) to extract the pedestrian foreground sequence masks.

In Appearance Module global branch, the backbone neural network extracts the feature maps of each frame of the sequence; then the global average pooling and temporal average pooling operations flatten and aggregate these feature maps of one sequence to a 2048 dimensions vector, after which a designed bottleneck block reduces the vector to 512 dimensions. As for the foreground branch, the output feature maps of the global branch backbone network are re-used and conducted dot multiplication with the resized foreground masks, which solely average pool the foreground region of the heatmaps and represent the feature of foreground (also 2048 dimensions). Another bottleneck block reduces the foreground vector to 512 dimensions in the same way.

Meanwhile, Gait Module randomly samples K frames of foreground sequence masks as the inputs of the module. After the Gait Module (a variant of Gaitset [10]), we obtain a gait feature (a 512-dimension vector) of each sequence masks.

Finally, the global appearance feature, the foreground appearance feature, and gait feature are concatenated to a 1536-dimension feature vector. We use the Feature Fusion Module to get the fusion feature vector (1536 dimensions). Finally, the fusion feature vector is used for the final representation of the video. We will further introduce the details of each module in the following subsections.

A. Backbone network and bottleneck structure

We adopt ResNet-50 [14] as the backbone network. By removing the final fully connected layer and changing the last stride from 2 to 1 in res_conv5, we can obtain a larger feature map incorporating richer feature information. For a typical image input size of 256 $\times$ 128, the output of the modified ResNet-50 is 16 $\times$ 8.

Both the global branch and the foreground branch adopt the bottleneck structure. The structure of the bottleneck is depicted in Fig. 2 (3), which is comprised of two fully connected layers, each followed by a normalization layer and a ReLU function. Compared to a fully connected layer (2048 $\times$ 512), the bottleneck structure results in notable network parameter reduction.

B. Gait Branch

We adopt and modify an advanced state-of-the-art gait recognition method GaitSet to obtain the gait feature. Different from other template-based and sequence-based methods, GaitSet treats the input as a set of disordered pedestrian contour images. Since pedestrian contours at different time

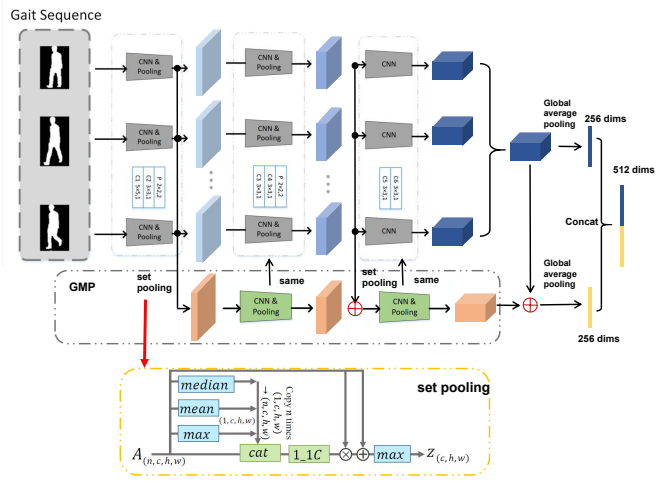

Fig. 3. The design of the Gaitset

flames exhibit different shapes intuitively, even if the contour sequence is re-shuffled, they can be rearranged into correct orders according to the shapes. The set-based GaitSet method inherently has a processing requirement for sets: sort independence, where the result has nothing to do with the order of input contours.

Compared with the original GaitSet, the modified GaitSet network reduces the number of horizontal strips at the end of the network, achieving the purpose of reducing the dimension of feature, and the performance is slightly reduced. The modified GaitSet architecture is shown in Fig. 3. The architecture settings before Global Pooling are the same as the original network (also shown detailly in fig. 2). The Set Pooling (SP) module is used to integrate each frame-level feature to form a set-level feature. As we know, a deeper convolutional layer has a larger receptive field. Shallow feature maps contain more local and fine-grained information, while deep feature maps contain global and coarse-grained information. Therefore, MGP(Multilayer Global Pipeline) is designed for aggregating the set-level features from different layers' outputs. The MGP structure is consistent with the main branch, with the same convolutional layer and pooling layer design, but the parameters are not shared with the main branch.

We obtain two 128-dimension vectors from the main branch and MGP after global pooling. After two independent fully connected layers, vectors are mapped to 256 dimensions. During the training phase, these two feature vectors are associated with two separate training losses. In the inference stage, they will be concatenated into a 512-dimensional feature vector.

Set Pooling is designed as Fig. 3. It uses 3 kinds of statistical operations that are independent of order: $max()$, $mean()$ and $median()$. 1_1C represents a 1×1 convolutional layer, which is used to fuse cascaded statistical features. Attention mechanism and residual structure are also designed for aggregation and maintaining the information capacity while accelerating and stabilizing convergence. Global Pooling computes the sum of Global Average Pooling and Global Max Pooling.

C. Feature Fusion Module (FFM)

We refer to the channel attention mechanism designed in SeNet to design a feature fusion network, shown as Fig. 2. The global appearance feature, the foreground appearance feature, and the gait feature are concatenated to a 1536-dimension vector. Then we use the Bottleneck mechanism (ratio=8) to construct a channel attention mechanism to promote the exchange of information between different features. At the same time, the residual structure is used to accelerate and stabilize convergence. Finally, a 1526-dimensional fusion feature vector is obtained.

D. Loss Functions

During training phase, batch-all triplet [15], batch-hard triplet loss [15] and SoftMax loss with Label-Smoothing Regularization (LSR) [16] are employed to train the network. The loss of the Appearance Module ($L_{appearance}$) is calculated as shown in Eq. 1; the loss of the Gait Module (L_{gait}) and the Feature Fusion Module (L_{fusion}) are shown as Eq. 2.

$$L_{appearance} = L^{global}_{batch\ hard\ triplet} + L^{global}_{softmax} + L^{foreground}_{batch\ hard\ triplet} + L^{foreground}_{softmax} \tag{1}$$

$$L_{gait} = L^{gait}_{batch\ all\ triplet} + L^{gait}_{softmax};\ L_{fusion} = L^{fusion}_{batch\ hard\ triplet} \tag{2}$$

And the total loss is shown as as Eq. 3.

$$L_{total} = \lambda_1 L_{fusion} + \lambda_2 L_{appearance} + \lambda_3 L_{gait} \tag{3}$$

III. Datasets and Data Preprocessing

In this paper, both sequence images and foreground sequence masks are needed as inputs for experiments. To obtain data in wild scenarios, we build a dataset named Mask-MARS based on the MARS dataset [3], a large-scale dataset for video-based person Re-ID. We also conduct experiments on the CASIA-B dataset [7].

- Mask-MARS: We create the original Mask-MARS dataset by computing the foreground mask of each RGB image in MARS dataset by a strong instance segmentation method. We require a video sequence to contain at least 8 effective foreground masks (not necessarily continuous), where an effective mask is simply defined as the proportion of the foreground area to the original image is not less than 15%. After screening by these two simple rules, the Mask-MARS data set contains a total of 1250 IDs and 14764 video sequences. The training set contains 624 IDs and 5726 video sequences; the query set contains 626 IDs and 1819 video sequences; the gallery set contains 621 IDs and 7170 video sequences. The length of the video sequence varies from 8 to 920 frames, with an average of 70 frames.
- CASIA-B: is a popular gait dataset. CASIA-B [7] has a total of 124 IDs, containing 11 angles (0, 18, 36,..., 180 degrees) and 3 different walking conditions. Walking conditions include normal (NM), each pedestrian contains 6 sequences; carrying bag (BG), each pedestrian contains 2 sequences; wearing a jacket or jacket (CL), each pedestrian contains 2 sequences. so each pedestrian contains 110 sequences. We use the first 74 IDs as the training set and the last 50 IDs as the test set. In the test environment, the first 4 sequences (NM1-4) under NM conditions are retained in the gallery subset, and the remaining 6 sequences are divided into 3 query subsets (NM5-6, BG1-2, and CL1-2).

We adopt the method in [9] to preprocess the foreground mask to achieve alignment. During the experiment, the size of the aligned mask image is set to 64×64. We crop 10 pixels on both the left and right sides of the horizontal direction to obtain the size of 64×44 as the input to the GaitSet network.

In the training phase, we preprocess the color image sequence and its corresponding masks in the following way: (1) Random sequence crop: We set the size of the cropped color image to 256×128, and the corresponding mask size of the output feature map of the last layer of ResNet-50 is 16×8, and the random probability $p = 0.5$. (2) Random sequence flip: we set random probability $p = 0.5$. (3) Image standardization: we use the mean and variance statistics of the 3 channels of the ImageNet dataset to normalize the image.

IV. Experiments

A. Progressive results and analysis

For simplicity and clarity, we name the baseline along with other module combines as follows:

Appearance Baseline: the global branch of Appearance Module, training individually, 512-dimension feature (without FFM). **Appearance Baseline + Foreground Branch**: whole Appearance Module, joint training, 1024-dimension feature (without FFM). **Modified GaitSet**: whole Gait Module, training individually, 512-dimension feature (without FFM). **Fusion Network**: our full system, joint end-to-end training, 1536-dimension feature (with FFM), with two version: each pretrained parts finetuning version and end-to-end trained version (**end2end**) .

As shown in Tab. I, in Mask-MARS, compared with the baseline model, the Rank1 index of the model with the foreground branch increased from 84.7% to 86.5%, and the mAP increased from 78.9% to 80.7%. Because MARS is a large-scale wild scene dataset with random camera views and large background clutters. Only use the gait branch, the performance is very poor (MAP 10.7% and rank1 17.7%). Even if the gait model does not perform well, our Fusion Network still outperforms Baseline + Foreground Branch with a notable margin. The two indicators of Rank1 and mAP have increased by 0.8% and 3.0% respectively. Compared to the Baseline, the improvement is even more significant (Rank1 +2.6% and mAP +4.8%).

In the CASIA-B dataset, the Tab. II records the experimental results of including and excluding the same view sequence as the query in the gallery. Adding appearance features makes the model more robust to view angle variation. Since the dataset is quite simple, there is basically no background clutter. Therefore, we can see that in the case of NM and BG, the appearance-based model can perform very well. However, when clothes changed (CL case), the performance of our Fusion Network (end2end) dominates all other models, achieving 72.891% rank1 accuracy, far exceeding the Modified GaitSet (+19.86%) and Appearance Baseline (+29.49%). This reflects that when the appearance of pedestrians changes significantly, the fusion features are often more discriminative than single-modal features. In any case in the CASIA-B dataset (NM, BG, CL), we achieve the best performance compared to a single-modal model.

TABLE I
PROGRESSIVE RESULTS ON MASK-MARS DATASET

Model	Rank1	Rank5	Rank10	Rank20	MAP
Appearance Baseline	84.7	94.6	95.9	97.2	78.9
Appearance Baseline + Foreground Branch	86.5	**95.5**	96.3	97.1	80.7
Modified GaitSet	17.7	34.1	43.1	51.8	10.7
FusionNetwork(finetune)	**87.1**	95.2	**96.3**	**97.2**	**81.1**
FusionNetwork(end2end)	**87.3**	**95.6**	**96.5**	**97.9**	**83.7**

TABLE II
COMPARISON OF CONCATENATION AND FUSION EFFECTS OF DIFFERENT FEATURES

Model	Including the same angle			Excluding the same angle		
	NM	BG	CL	NM	BG	CL
AppearanceBaseline	98.669	96.630	45.884	98.536	96.475	45.400
AppearanceBaseline+Foreground Branch	98.405	95.960	48.843	98.245	95.729	48.364
Modified GaitSet	83.570	71.284	55.934	81.964	69.168	55.036
FusionNetwork(finetune)	99.455	96.782	73.634	99.437	96.579	69.495
FusionNetwork(end2end)	**99.620**	**96.822**	**75.950**	**99.582**	**96.623**	**74.891**

B. Comparison of concatenation and fusion effects of different features

This experiment is to verify the effectiveness of the feature fusion module (FFM). We name models with different setting as follows: **GGConcat**: concatenated feature from 2 branches: global branch and gait branch; **GGFusion**: fusion feature from 2 branches with FFM; **AGConcat**: concatenated feature from 3 branches: global branch, foreground branch and gait branch without FFM; **AGFusion**: fusion feature from 3 branches with FFM. The experimental results on the Mask-MARS and CASIA-B datasets are shown in Tab. III.

C. Compared with other advanced video-based ReID methods

We also reproduce some recent advanced algorithms [2], [4], [17] to validate the effectiveness of our hubrid model on the dataset we created in Tab. IV. The performance of our method is far superior compared to methods that model only the appearance features.

TABLE III
COMPARISON OF CONCATENATION AND FUSION EFFECTS OF DIFFERENT FEATURES

Model	Mask-MARS				CASIA-B		
	rank1	rank5	rank10	mAP	NM	BG	CL
AGConcat	86.6	95.2	96.4	81.0	99.736	97.905	68.785
AGFusion	87.1	95.2	96.3	81.1	99.620	96.822	75.950
GGConcat	87.0	95.4	96.5	81.0	99.810	97.806	75.099
GGFusion	87.3	95.3	96.2	80.7	99.149	96.740	75.992

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART VIDEO-BASED RE-ID METHODS ON MASK-MARS DATASET.

Model	Rank1	Rank5	Rank10	Rank20	MAP
Non-local+C3D [2]	84.1	94.5	96.0	97.3	77.2
STAN [17]	82.3	92.9	94.6	96.8	65.7
Snipped [4]	81.2	92.1	94.6	96.5	69.4
Snipped+OF [4]	86.3	94.7	95.7	**98.2**	76.1
Ours (end2end)	**87.3**	**95.6**	**96.5**	97.9	**83.7**

V. Conclusion

This paper propose an end-to-end framework which utilizes the sequence masks (SeqMasks) in each video to jointly exploit the power of appearance and gait in video Re-ID. Experiments on Mask-MARS dataset evidence the favorable performance and generalization ability of the proposed algorithm. Further validations on gait recognition metric CASIA-B dataset highlight the performance of our hybrid model.

REFERENCES

[1] Jiyang Gao and Ram Nevatia, "Revisiting temporal modeling for video-based person reid," *arXiv preprint arXiv:1805.02104*, 2018.

[2] Xingyu Liao, Lingxiao He, Zhouwang Yang, and Chi Zhang, "Video-based person re-identification via 3d convolutional networks and non-local attention," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 620–634.

[3] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian, "Mars: A video benchmark for large-scale person re-identification," in *European Conference on Computer Vision*. Springer, 2016, pp. 868–884.

[4] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1169–1178.

[5] Zhigang Chang, Zhou Qin, Heng Fan, Hang Su, Hua Yang, Shibao Zheng, and Haibin Ling, "Weighted bilinear coding over salient body parts for person re-identification," *Neurocomputing*, vol. 407, pp. 454–464, 2020.

[6] Zhao Yang, Zhigang Chang, and Shibao Zheng, "Large-scale video-based person re-identification via non-local attention and feature erasing," in *International Forum on Digital TV and Wireless Multimedia Communications*. Springer, 2019, pp. 327–339.

[7] Shiqi Yu, Daoliang Tan, and Tieniu Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *18th International Conference on Pattern Recognition (ICPR 2006), 20-24 August 2006, Hong Kong, China*, 2006.

[8] Athira Nambiar, Alexandre Bernardino, and Jacinto C Nascimento, "Gait-based person re-identification: A survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 2, pp. 1–34, 2019.

[9] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *Ipsj Transactions on Computer Vision and Applications*, vol. 10, no. 1, pp. 4, 2018.

[10] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," 2018.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] Rza Alp Güler, Natalia Neverova, and Iasonas Kokkinos, "Densepose: Dense human pose estimation in the wild," 2018.

[13] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," 2015.

[17] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 369–378.