

Horse ReIDing: Addressing Re-Identification in Horse Racing Scenarios

Original

Horse ReIDing: Addressing Re-Identification in Horse Racing Scenarios / Rossi, L.F., Sanna, A., Manuri, F., Donna Bianco, M.. - ELETTRONICO. - 15925:(2026), pp. 209-217. (The 2nd International Sports Analytics Conference and Exhibition (ISACE) Shanghai (CN) 26-27 September 2025) [10.1007/978-3-032-06167-6_16].

Availability:

This version is available at: 11583/3001715 since: 2026-08-05T07:09:42Z

Publisher:

Springer Nature

Published

DOI:10.1007/978-3-032-06167-6_16

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

Springer postprint/Author's Accepted Manuscript

This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/978-3-032-06167-6_16

(Article begins on next page)

Horse ReIDing: Addressing Re-Identification in Horse Racing Scenarios

✉ Luca Francesco Rossi^{1,2}[0009–0008–1571–5590],
Andrea Sanna¹[0000–0001–7916–1699], Federico Manuri¹[0000–0002–6599–9949], and
Mattia Donna Bianco²

¹ Department of Control and Computer Engineering (DAUIN)
Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, TO, Italy
lucafrancesco.rossi@polito.it

² netventure R&D S.r.l., Via della Consolata 1/bis, 10122 Torino, TO, Italy

Abstract. Re-Identification (ReID) tasks, traditionally employed in person tracking across diverse camera views, face unique challenges in the domain of horse racing due to frequent occlusions, dynamic motion, and varying environmental conditions. This study addresses these complexities by developing a custom pipeline and dataset for jockey ReID, specifically collected from horse racing footage. A ResNeXt-based architecture is employed to process input data, with additional experiments exploring the inclusion of segmentation mask information for improved performance. Empirical evaluations demonstrate the model’s efficacy in both closed-set and open-set scenarios, showcasing significant gains in mean Average Precision (mAP) and top-k Cumulative Matching Characteristic (CMC) rank metrics when segmentation masks are incorporated. Comparative analysis across different ResNeXt configurations underscores the robustness and scalability of the proposed approach, contributing as a pioneering framework for ReID in high-motion sports contexts and advancing the application of computer vision technologies in horse racing scenarios.

Keywords: Re-identification · Computer vision · Artificial intelligence
· Sport · Broadcast footage · Horse racing.

1 Introduction

Person Re-Identification (ReID) is the task of recognizing whether a person seen in one camera view or image has appeared in another, effectively matching people across non-overlapping camera views [1]. This problem is important in many real-world settings such as video surveillance, security, and smart-city applications, where one might track a person of interest through a network of cameras [2]. In a sense, it can be thought of as a specialized image-retrieval problem: given a “query” image of a person, the system searches a database of other images to find the same individual [3]. A central difficulty is that the same entity can have very dissimilar representations in different camera views or at distinct times,

with changes in lighting, viewpoint, pose or clothing making matching way more complicate [4]. Early approaches to person ReID relied on hand-crafted appearance features, i.e., color histograms, texture descriptors or edge-based features, along with traditional matching metrics [5]. However, as large labeled datasets became available starting from last decade, deep learning methods – especially convolutional neural networks, or CNNs – became the standard, automatically learning powerful appearance features end-to-end from the data. Since then, deep architectures have dominated the literature, greatly outperforming earlier hand-engineered methods. In the past few years, even more flexible Transformer-based [6] architectures have emerged: these models (originally developed for language processing) can learn to focus on different parts of an image and have quickly set new state-of-the-art records in ReID tasks. In sports, person ReID can therefore be used to track players across multiple broadcast cameras, enabling richer analytics and improved viewing experiences [7]: in such a way, coaches can monitor each player’s movements and performance over time, and broadcasters can overlay real-time statistics and highlights onto live video. However, despite the growing research attention devoted to widely followed sports such as soccer [8], basketball [9], or hockey [10], relatively limited investigation has been directed toward the domain of horse racing. This study thus aims to bridge such a gap by addressing the re-identification task within the context of horse racing, a highly-dynamic environment characterized by its fast-paced nature and frequent occlusions.

The paper is organized as follows: Section 2 provides a comprehensive review of related work relevant to computer vision applications specifically tailored for horse racing; Section 3 details the proposed methodology, including the architectural design and implementation specifics for incorporating segmentation masks into ReID models; Section 4 presents an empirical evaluation of the proposed approach, analyzing its performance gains relative to baseline models; and Section 5 concludes the study by summarizing key contributions and outlining potential directions for future research.

2 Related Work

Binning et al. [11] addressed the challenges of horse racing video analysis, specifically focusing on the detection and re-identification (ReID) of jockeys through helmet identification. Horse racing presents unique challenges such as occlusion, motion blur, and varying illumination, which complicate the detection and tracking of both horses and jockeys. The authors propose a helmet detection framework [20] that aims to address these issues and serve as a preliminary step for ReID. The evaluation dataset features 12 jockeys, but the study focuses on five distinct helmet classes for semi-automated ground truth labeling. The model was trained on data from one camera and tested on the same camera as well as unseen data from others, showing good performance with higher accuracy for simple helmet designs compared to more intricate ones. Hedayati et al. [13] presented a novel approach for tracking jockeys in a crowded horse racing

environment, focusing on the challenges of occlusion and group dynamics. The primary goal is to track jockeys at the turning point of the race, where their motion changes direction [14]. Tracking in this scenario is particularly difficult due to the high levels of occlusion, as jockeys race closely together, often obscuring one another. The proposed solution leverages the behavior of jockeys as they move in a relatively uniform manner, even when partially obscured, to improve tracking accuracy. By integrating group dynamics into the tracking system, the approach helps estimate the positions of jockeys who are not fully visible. The system consists of three main components: sampling, localization, and data association. Sampling involves capturing motion and color cues from the jockey’s cap at multiple levels—point, object, and group—to track the jockeys even under occlusion. Localization uses two strategies—object-based localization when the jockey’s cap is clearly visible, and group-based localization when occlusion occurs. The latter estimates the position of a jockey by considering the movement of the entire group of jockeys. Data association is employed to link new detections with existing tracks [15], ensuring consistent tracking even when multiple objects overlap. Empirical results on video footage of horse races show that, by incorporating group dynamics, the method effectively handles occlusion, achieving a high correct tracking ratio. Wing et al. [16] similarly examined the challenges associated with multi-object tracking (MOT) in the context of horse racing, given the well-known difficulties of occlusion, trajectory overlap, frequent camera switching, and motion blur caused by horses sprinting. The authors introduce an augmentation-based multi-object tracking method (GMOT) designed to address these challenges. GMOT leverages an auxiliary classifier generative adversarial network (ACGAN) to augment horse racing data [17], thereby enhancing target detection and re-identification by generating synthetic images that simulate difficult scenarios, such as blurred sprinting and occlusion. Building upon the FairMOT framework [18], which employs an anchor-free approach with separate branches for position detection, GMOT is particularly well-suited for handling issues such as camera switching and occlusion, which typically hinder traditional MOT methods. The model is trained on both day- and night-time race datasets, with additional synthetic training data generated through ACGAN to further strengthen the model’s ability to cope with real-world complexities in horse racing. Experimental results indicate that GMOT outperforms FairMOT in a variety of challenging horse racing conditions, demonstrating superior accuracy in tracking. The interested reader is finally referred to [19] for a broader analysis on how AI is impacting the world of horse racing, aside from computer vision applications only.

3 Method

Due to the lack of availability of horse racing data for computer vision, a custom ReID dataset is privately collected from multiple videos provided by The Hong Kong Jockey Club. Specifically, six full-race replays from the Sha Tin Racecourse under different environmental conditions are used for training, while a full-race

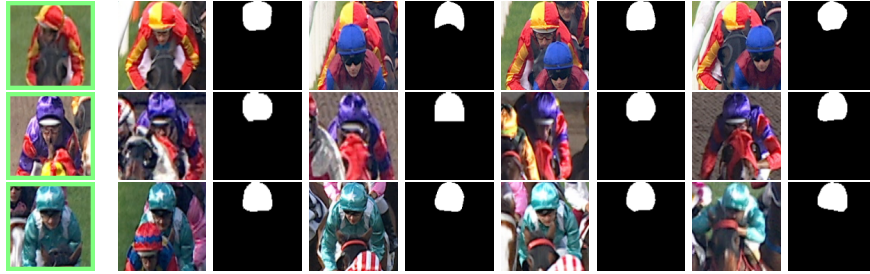


Fig. 1. Qualitative visualization of dataset samples with paired helmet masks.

Table 1. Summary statistics concerning the collected horse racing ReID dataset.

<i>Race</i>	<i>Racecourse</i>	<i>Time</i>	<i>Condition</i>	<i>Jockeys</i>	<i>Frames</i>	<i>Images</i>	<i>Split</i>
#1	Sha Tin	Day	Grass	11	89	954	Train
#2	Sha Tin	Day	Grass	14	74	1014	Train
#3	Sha Tin	Night	Grass	14	86	1171	Train
#4	Sha Tin	Night	Grass	14	78	1059	Train
#5	Sha Tin	Night	Dirt	12	76	875	Train
#6	Sha Tin	Night	Dirt	12	107	1235	Train
#7	Happy Valley	Day	Grass	12	104	1216	Test

from the Happy Valley Racecourse is separately used for testing. To construct such a dataset, a semi-automatic pipeline is instantiated in the following way:

1. a frame is extracted every single second from each full-race video;
2. bounding boxes for jockeys’ helmets are retrieved by deploying a GroundingDINOv1.5 model [20];
3. helmets’ segmentation mask are determined via a Segment Anything Model 2 (SAM2) [21] by prompting the previously-detected boxes;
4. hybrid manual-automatic refinement [22] is performed to edge cases, followed by track identity assignment;

Since silks (i.e., the brightly colored shirts worn by jockeys) convey as much information concerning identification as helmets do, the final image for each jockey is cropped by zooming out from helmet detection only. Roughly following the heuristics of the canon of proportions in ancient Greek sculpture, if one defines the helmet’s bounding box in its $tlwh$ format, that is, by the (x, y) -coordinates of its top-left corner, width and height, optimal empirical results for the final image are retrieved by cropping the original one at $< x_{\text{helmet}} - w_{\text{helmet}}, y_{\text{helmet}}, 3 \times w_{\text{helmet}}, 3 \times h_{\text{helmet}} >$, eventually clipped to image borders, which captures the whole jockey’s upper body. Figure 1 qualitatively depicts some output examples from such pipeline, while Table 1 provide a more quantitative summary concerning the constructed dataset. Due to the real-time requirements often imposed by computer vision applications for sports, a ResNeXt architecture [23] is instantiated.

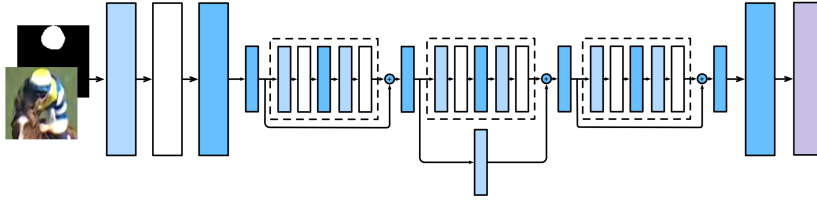


Fig. 2. Illustrative visualization of the residual architecture, with the final layer embedding the high-dimensional representation of the input.

ResNeXt is a lightweight convolutional neural network architecture that builds on the principles of ResNet [24] by introducing a new structural dimension called “cardinality”, which refers to the number of parallel transformations within each block. While maintaining the simplicity and modularity of ResNet through repeating blocks with identical topology, ResNeXt enhances representational power by aggregating multiple homogeneous transformations (each operating on a low-dimensional embedding of the input) through summation. For this study, input images have been reshaped as 128×128 before being fed to the ResNeXt model, with the final classification head computing a 256-dimensional embedding representation. Following literature [25], the network is optimized for a total of 10 epochs jointly using cross-entropy and hard-mining triplet loss [26]. Similarly to other works in occluded ReID [27], the inclusion of segmentation mask information is investigated to assess whether it could guide the model towards focusing on more relevant parts of the image [28–31]. To do so, the architecture is extended in order to accept a dilated representation of the input image, obtained by concatenating the jockey’s helmet mask to the RGB channels.

4 Evaluation and Discussion

In a practical setting, it is expected that such a model will be deployed over two different scenarios: known jockeys and unknown jockeys. The former consists in a deployment scenario where all athletes have already been clustered during training (closed-set), while the latter deals with generalization, that is, how well the model behaves when encountering jockeys never seen before (open-set). For such a reason, Race #4 is employed for validation during the training phase, as it features the same jockeys as Race #3, albeit recorded by a different patrol camera. On the contrary, the Happy Valley race features completely different, never-seen-before jockeys, and it is used to test its generalization capabilities concerning open-set ReID. Following common practices for ReID tasks, such test race is further subject to a random 90-10 split in order to provide the standard gallery-query setting. Three different ResNeXt configurations are considered: depth-50, depth-101 and depth-152. Pretrained on the ImageNet-1k dataset [33], these configurations are fine-tuned on the collected dataset both with and without injecting helmet’s segmentation mask inside the input tensor.

Table 2. Empirical results on the test set for all proposed configurations.

<i>Method</i>	<i>Depth</i>	<i>mAP</i>	<i>R@1</i>	<i>R@5</i>	<i>R@10</i>	<i>R@20</i>	<i>FPS</i>
Vanilla	50	0.7690	0.9580	0.9830	0.9830	0.9920	906.8
	101	0.8380	0.9580	0.9830	1.0000	1.0000	894.4
	152	0.8560	0.9500	1.0000	1.0000	1.0000	882.4
Segmentation	50	0.7550	0.9330	0.9750	0.9920	0.9920	685.5
	101	0.8960	0.9670	1.0000	1.0000	1.0000	676.2
	152	0.8890	0.9750	1.0000	1.0000	1.0000	670.4

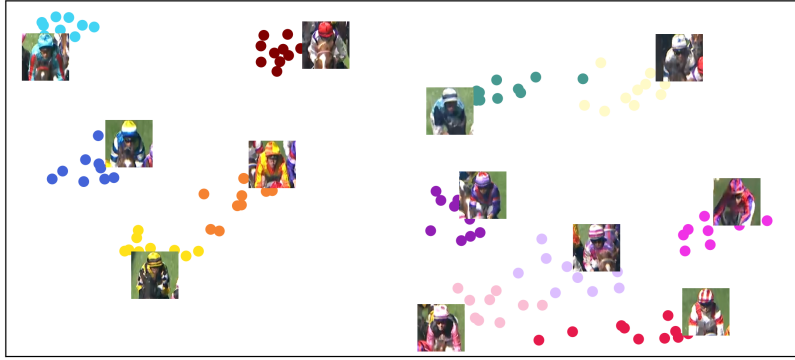
**Fig. 3.** Qualitative t-SNE [32] visualization of query jockeys clustering.

Table 2 lists mean Average Precision (mAP) and top- k Cumulative Matching Characteristic (CMC) rank metrics ($R@k$) on the test set. Aside from the most shallow configuration, empirical results validate the advantage of incorporating segmentation knowledge for ReID purposes, with boosts up to 6.92% in mAP and 2.63% in $R@1$. FPS (averaged over 10 test iterations) on customer-level hardware are also reported in Table 2, confirming the suitability of the proposed architecture for real-time purposes.

5 Conclusion and Future Work

This study demonstrates the effectiveness of integrating ReID dynamics in the context of horse racing for jockey ReID in broadcast footage, with the incorporation of helmet segmentation masks significantly improving re-identification performance. Future work could explore expanding the dataset to include a more diverse range of racing environments and conditions to enhance model generalization. Additionally, the integration of temporal information and the exploration of transformer-based architectures may further improve the robustness of the system, paving the way for broader applications in sports analytics and beyond.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Asperti, A., Fiorilla, S., Nardi, S., Orsini, L.: A review of recent techniques for person re-identification. *Machine Vision and Applications* **36**(1), pp. 25–53 (2025). <https://doi.org/10.1007/s00138-024-01622-3>
2. Kansal, K., Wong, Y., Kankanhalli, M.: Privacy-Enhancing Person Re-identification Framework – A Dual-Stage Approach. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 8528–8537. Waikoloa, HI, USA (2024). <https://doi.org/10.1109/WACV57701.2024.00835>
3. Saad, R. S. M., Moussa, M. M., Abdel-Kader, N. S., Farouk, H., Mashaly, S.: Deep video-based person re-identification (Deep Vid-ReID): comprehensive survey. *EURASIP J. Adv. Signal Process.* **2024**(1), 63–111 (2024). <https://doi.org/10.1186/s13634-024-01139-x>
4. Ming, Z., Zhu, M., Wang, X., Zhu, J., Cheng, J., Gao, C., Yang, Y., Wei, X.: Deep learning-based person re-identification methods: A survey and outlook of recent works. *Image and Vision Computing* **119**, 104394 (2022). <https://doi.org/10.1016/j.imavis.2022.104394>
5. Chahla, C., Snoussi, H., Abdallah, F., Dornaika, F.: Learned versus Hand-crafted Features for Person Re-identification. *International Journal of Pattern Recognition and Artificial Intelligence* **34**(4), p. 2055009 (2020). <https://doi.org/10.1142/S0218001420550095>
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pp. 6000–6010. Curran Associates Inc., Red Hook, NY, USA (2017). <https://papers.nips.cc/paper/7181-attention-is-all-you-need>
7. Fujii, K.: Computer Vision for Sports Analytics. *Machine Learning in Sports, SpringerBriefs in Computer Science*, 21–57 (2025). https://doi.org/10.1007/978-981-96-1445-5_2
8. Delière, A., Cioppa, A., Giancola, S., Seikavandi, M. J., Dueholm, J. V., Nasrollahi, K., Ghanem, B., Moeslund, T. B., Van Droogenbroeck, M.: SoccerNet-v2: A Dataset and Benchmarks for Holistic Understanding of Broadcast Soccer Videos. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4503–4514. Nashville, TN, USA (2021). <https://doi.org/10.1109/CVPRW53098.2021.00508>
9. Van Zandycke, G., Somers, V., Istasse, M., Del Don, C., Zambrano, D.: DeepSportradar-v1: Computer Vision Dataset for Sports Understanding with High Quality Annotations. In: *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports (MMSports '22)*, pp. 1–8. Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3552437.3555699>
10. Koshkina, M., Elder, J. H.: Towards Long-Term Player Tracking with Graph Hierarchies and Domain-Specific Features. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), pp. 1175–1185. Tucson, AZ, USA (2025). <https://doi.org/10.1109/WACVW65960.2025.00140>
11. Binning, W., Rahmaniboldaji, S., Dong, X., and Gilbert, A.: Detection and re-identification in the case of horse racing. In: *Information Sciences* (2024). <https://openresearch.surrey.ac.uk/esploro/outputs/conferenceProceeding/Detection-and-Re-Identification-in-the-case/99925861902346>

12. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., and Zhang, L.: Grounded SAM - assembling open-world models for diverse visual tasks. In: arXiv cs.CV (2024). <https://arxiv.org/abs/2401.14159>
13. Hedayati, M., Cree, M. J., and Scott, J. B.: Tracking jockeys in a cluttered environment with group dynamics. In: Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports (2019). <https://doi.org/10.1145/3347318.3355518>
14. Hedayati, M., Cree, M. J., and Scott, J. B.: Scene structure analysis for sprint sports. In: 2016 International Conference on Image and Vision Computing New Zealand (2016). <https://doi.org/10.1109/IVCNZ.2016.7804429>
15. Murty, K. G.: An algorithm for ranking all the assignments in order of increasing cost. In: Operations Research 16 (1968). <http://www.jstor.org/stable/168595>
16. Ng, W. W. Y., Liu, X., Yan, X., Tian, X., Zhong, C., and Kwong, S.: Multi-object tracking for horse racing. In: Information Sciences 638 (2023). <https://doi.org/10.1016/j.ins.2023.118967>
17. Odena, A., Olah, C. and Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: Proceedings of the 34th International Conference on Machine Learning (2017). <https://proceedings.mlr.press/v70/odena17a.html>
18. Zhang, Y., Wang, C., Wang, X., Zeng, W., and Liu, W.: FairMOT - on the fairness of detection and re-identification in multiple object tracking. In: Int. J. Comput. Vis. 129, 3069–3087 (2021). <https://doi.org/10.1007/s11263-021-01513-4>
19. Colle, P.: What AI can do for horse-racing? In: arXiv cs.LG (2022). <https://arxiv.org/abs/2207.04981>
20. Ren, T., Jiang, Q., Liu, S., Zeng, Z., Liu, W., Gao, H., Huang, H., Ma, Z., Jiang, X., Chen, Y., Xiong, Y., Zhang, H., Li, F., Tang, P., Yu, K., Zhang, L.: Grounding DINO 1.5: Advance the "Edge" of Open-Set Object Detection. arXiv 2405.10300, cs.CV (2024). <https://arxiv.org/abs/2405.10300>
21. Ravi, N., Gabeur, V., Hu, Y., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C., Girshick, R., Dollár, P., Feichtenhofer, C.: SAM 2: Segment Anything in Images and Videos. arXiv 2408.00714, cv.CV (2024). <https://arxiv.org/abs/2408.00714>
22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W., Dollár, P., Girshick, R.: Segment Anything. arXiv 2304.02643, cv.CV (2023). <https://arxiv.org/abs/2304.02643>
23. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated Residual Transformations for Deep Neural Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987–5995. Honolulu, HI, USA (2017). <https://doi.org/10.1109/CVPR.2017.634>
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. Las Vegas, NV, USA (2016). <https://doi.org/10.1109/CVPR.2016.90>
25. Yuan, Y., Chen, W., Yang, Y., Wang, Z.: In Defense of the Triplet Loss Again: Learning Robust Person Re-Identification with Fast Approximated Triplet Loss and Label Distillation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1454–1463. Seattle, WA, USA (2020). <https://doi.org/10.1109/CVPRW50498.2020.00185>
26. Xuan, H., Stylianou, A., Liu, X., Pless, R.: Hard Negative Examples are Hard, but Useful. In: Computer Vision – ECCV 2020: 16th European Conference, Springer-Verlag, pp. 126–142. Glasgow, UK (2020). https://doi.org/10.1007/978-3-030-58568-6_8

27. Sun, Z., Wei, G., Fu, W., Ye, M., Jiang, K., Liang, C., Zhu, T., He, T., Mukherjee, M.: Multiple Pedestrian Tracking Under Occlusion: A Survey and Outlook. *IEEE Transactions on Circuits and Systems for Video Technology* **35**(2), pp. 1009–1027 (2025). <https://doi.org/10.1109/TCSVT.2024.3481425>
28. Meng, H., Zhao, Q.: A Lightweight Model Based on Co-segmentation Attention for Occluded Person Re-identification. In: *Proceedings of 2021 Chinese Intelligent Automation Conference. Lecture Notes in Electrical Engineering*, vol 801., pp. 692–701. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-6372-7_74
29. Zhou, Y., Qin, Y., Wang, S., Huang, Z., Zhang, D.: Person Re-Identification Based on Instance Segmentation and Pose Estimation. In: *IEEE 6th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, pp. 584–589. Haikou, China (2023). <https://doi.org/10.1109/PRAI59366.2023.10332122>
30. Qin, P., Gao-hua, C.: Occluded person re-identification based on foreground segmentation and multi-scale feature fusion. In: *Proc. SPIE 13552, International Conference on Physics, Photonics, and Optical Engineering (ICPPOE 2024)*, 135522D (2025). <https://doi.org/10.1117/12.3060412>
31. Liu, Y., Wang, Z., Zhang, W., Li, Z.: DGSN: Learning how to segment pedestrians from other datasets for occluded person re-identification, *Image and Vision Computing* **140**, 104844 (2023). <https://doi.org/10.1016/j.imavis.2023.104844>
32. van der Maaten, L., Hinton, G.: Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**(86), pp. 2579–2605 (2008). <http://jmlr.org/papers/v9/vandermaaten08a.html>
33. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Miami, FL, USA (2009). <https://doi.org/10.1109/CVPR.2009.5206848>