MEMORY-DRIVEN PERSON RE-IDENTIFICATION FOR IDENTITY CONSISTENCY IN MULTI-OBJECT TRACKING

Tista Pal, Trinh Quoc Nguyen, Oky Dicky Ardiansyah Prima

ACHI 2026 | May 2026
Email: s231x018@s.iwate-pu.ac.jp

岩手県立大学

IARIA

TISTA PAL

Tista Pal is currently a master's student majoring in artificial intelligence at Graduate school of Software and Information Science, Iwate Prefectural University, Iwate, Japan.

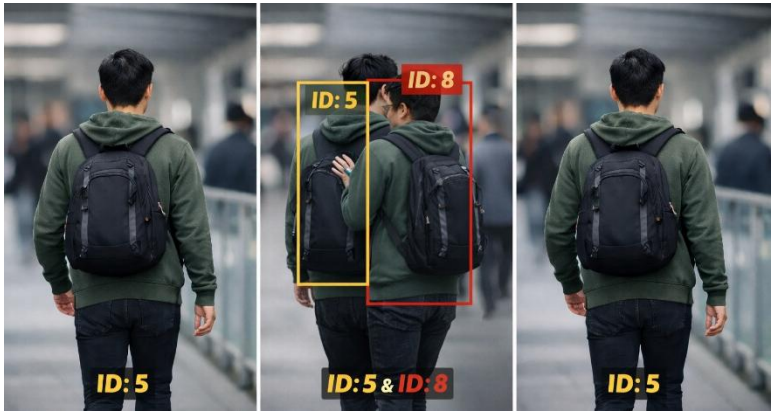
Her research interests lies in image processing, computer Vision, object detection and deep learning.

RESEARCH BACKGROUND

Tracking people across frames or camera views is difficult because traditional MOT methods rely on short-term cues and often lose identities during occlusions or re-entries.

Goal: Design a tracking pipeline that maintains **global identity consistency** across time using long-term appearance memory.

Traditional tracking: lose identity during occlusion



Improved tracking: consistent identity after occlusion

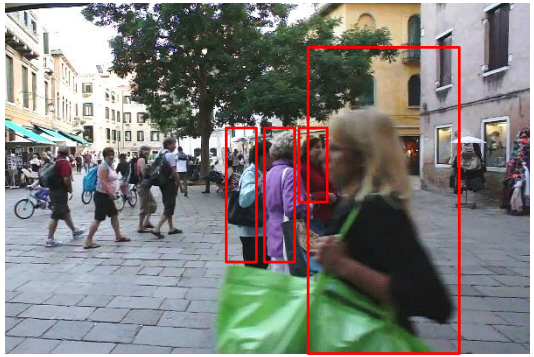
RESEARCH BACKGROUND

Challenges encountered

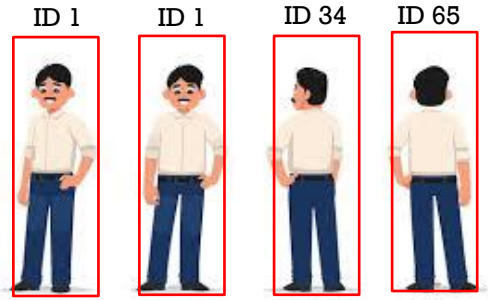
- Similar clothes.
- Environmental factors such as bad lighting, shaky camera, etc.
- Occlusion.
- Change in pose and orientation.

Similar clothes

Blurry image source



MOT17-02 frame: depicting occlusion



Change in human orientation leads to identity inconsistency

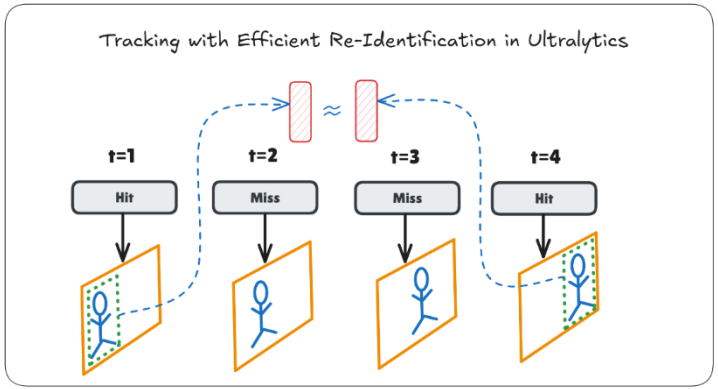
Solution

Person Re-ID

RESEARCH BACKGROUND

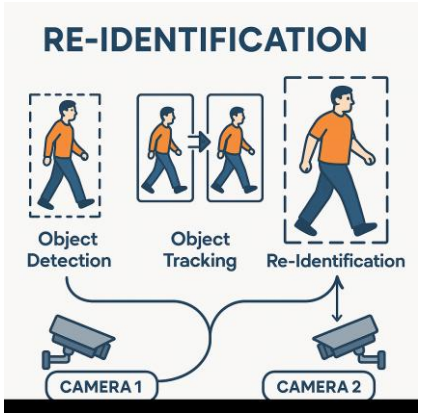
Goal: Design a tracking pipeline that maintains **global identity consistency** across time using long-term appearance memory.

Single camera



https://y-t-g.github.io/tutorials/yolo-reid/

Multi-camera

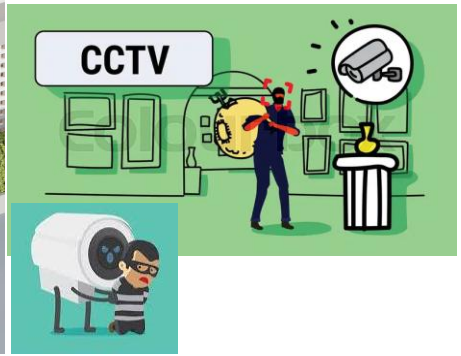


Detect Object → Track + Assign ID → If missed in CAM1 and again appears in CAM2 → Re-Identify with correct Id

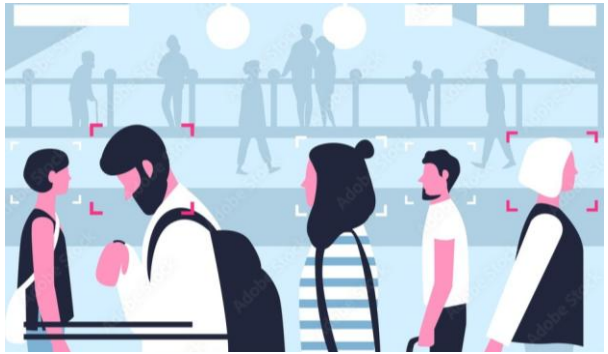
RESEARCH BACKGROUND

Surveillance and public safety

Track a person across multiple non-overlapping cameras



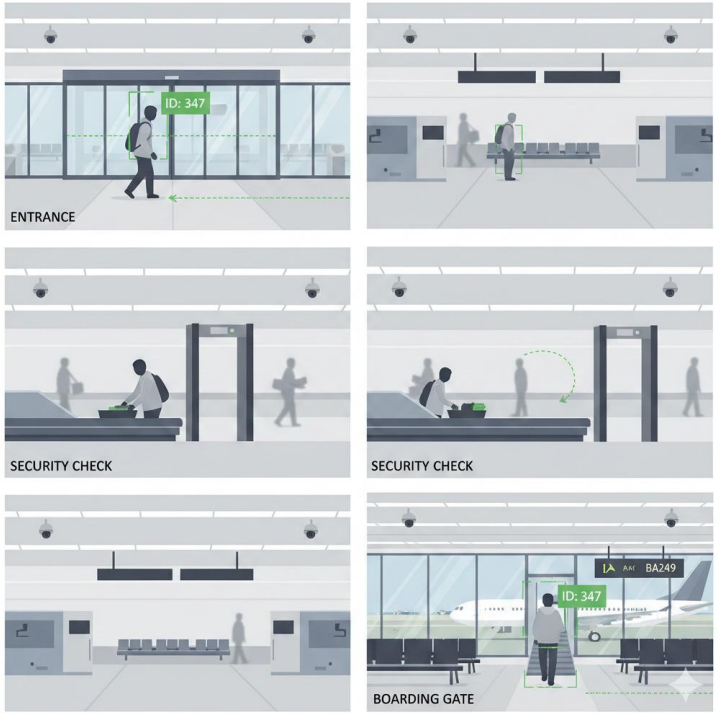
Suspect tracking

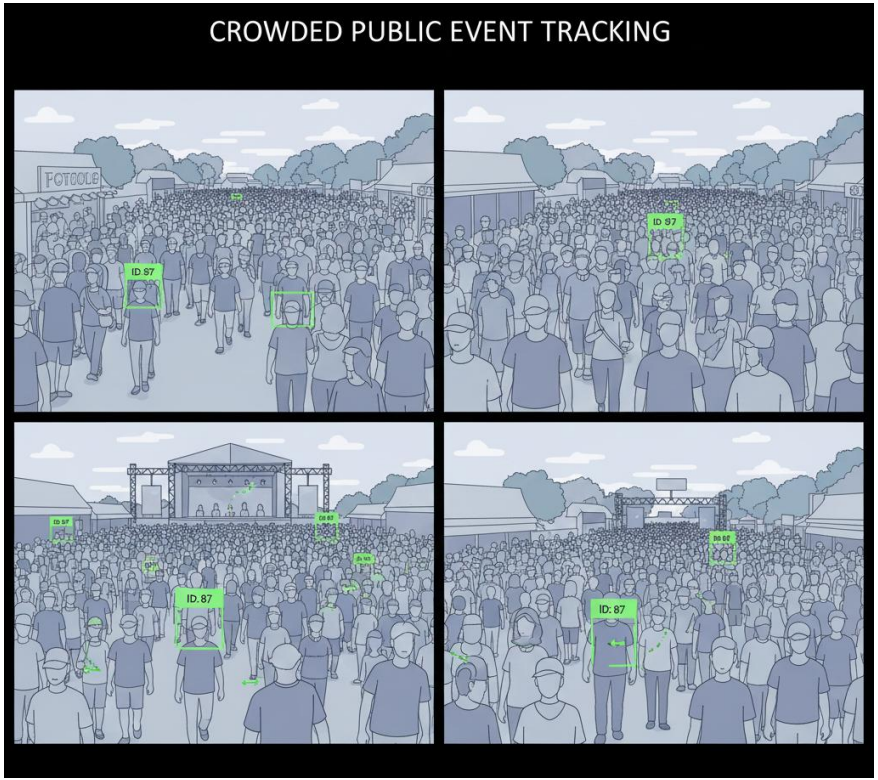


Missing person tracking

RESEARCH BACKGROUND

Crowd and event management





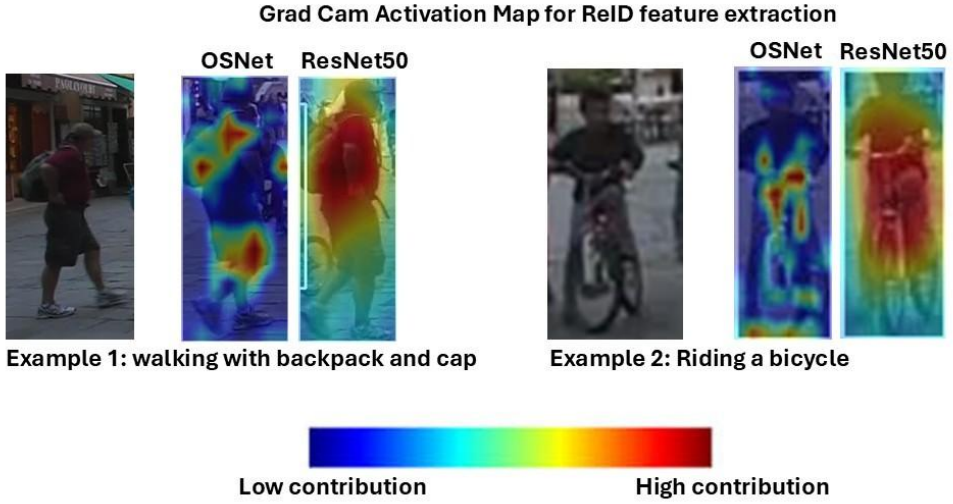
Crowd flow analysis, Stadium & concert monitoring, track staff, security, or VIP movement across zones

RELATED WORK :: MULTI-OBJECT DETECTION (MOT)

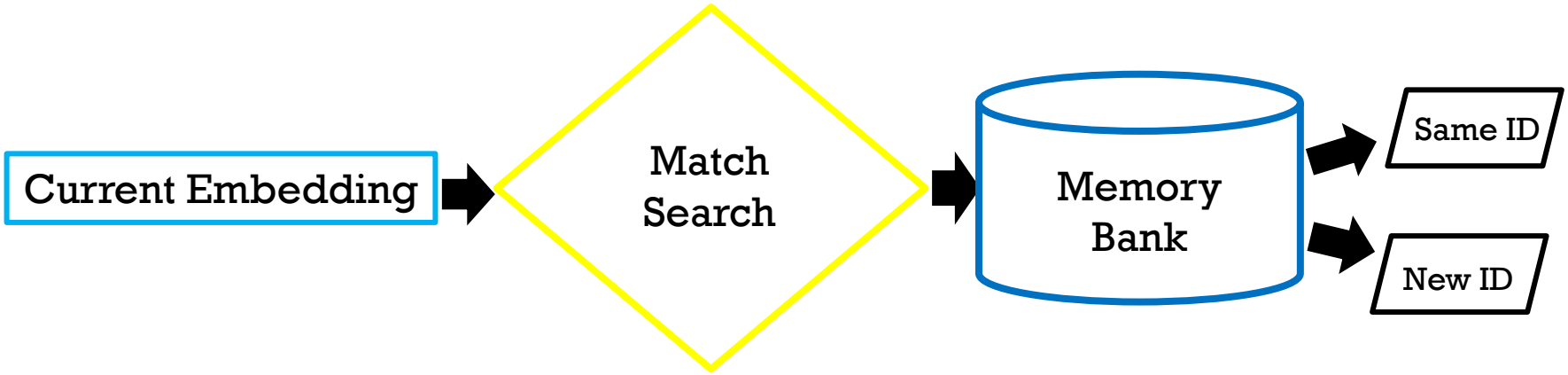
Feature	SORT	DeepSORT
Core Idea	Motion-based tracking (Kalman filter)	SORT with a deep appearance descriptor
Association Metric	IoU distance	Mahalanobis(motion) + Cosine(appearance) distance
Occlusion Handling	Poor	Strong
Computational cost	Extremely low(very fast)	Moderate (slower due to CNN feature extraction)
Data Association	Simple Hungarian Algorithm	Matching Cascade + Hungarian Algorithm

RELATED WORK :: FEATURE EXTRACTION

Feature	OSNet	ResNet-50
Feature focus	Omni scale: Fine details + Global structure	Global: Overall shape and dominant colors
Capability	Distinguish by small unique visual cues	Struggle with two people wearing similar clothes
Model size	Lightweight (~2.2M parameters)	Heavyweight (~23.5M-25M parameters)
Goal	Person ReID	General purpose image recognition



RELATED WORK :: MEMORY BASED RE-ID

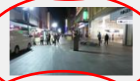


- Gallery-based matching: Commonly used in cross-camera ReID scenarios.
- Memory size is often fixed.
- Explicit global identity memory for online tracking remains underexplored.

DATASET

Consists of multiple pedestrian video sequences captured under varying illumination, crowd density, and camera motion conditions. Each sequence provides ground truth annotations in the MOTChallenge format, including bounding boxes and identity labels.

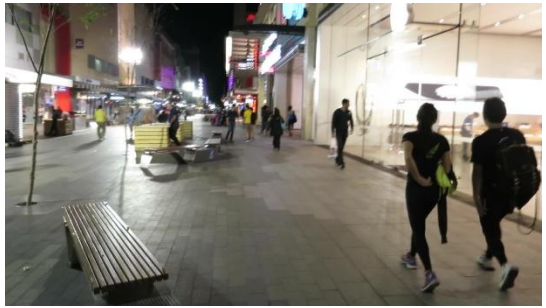
A representative subset of the MOT17 sequences was selected to perform inference on it to ensure balanced coverage of diverse motion and occlusion scenarios.

Sample	Name	FPS	Resolution	Length	Tracks	Boxes	Density	Description
	MOT17-13-SDP	25	1920x1080	750 (00:30)	110	11642	15.5	Filmed from a bus on a busy intersection
	MOT17-11-SDP	30	1920x1080	900 (00:30)	75	9436	10.5	Forward moving camera in a busy shopping mall
	MOT17-10-SDP	30	1920x1080	654 (00:22)	57	12839	19.6	A pedestrian scene filmed at night by a moving camera
	MOT17-09-SDP	30	1920x1080	525 (00:18)	26	5325	10.1	A pedestrian street scene filmed from a low angle.
	MOT17-05-SDP	14	640x480	837 (01:00)	133	6917	8.3	Street scene from a moving platform
	MOT17-04-SDP	30	1920x1080	1050 (00:35)	83	47557	45.3	Pedestrian street at night, elevated viewpoint
	MOT17-02-SDP	30	1920x1080	600 (00:20)	62	18581	31.0	People walking around a large square.

DATASET

MOT17-11-SDP

MOT17-10-SDP



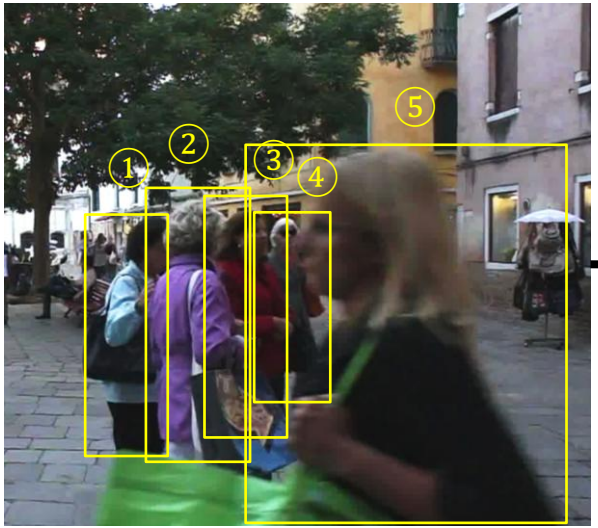
MOT17-09-SDP

MOT17-04-SDP

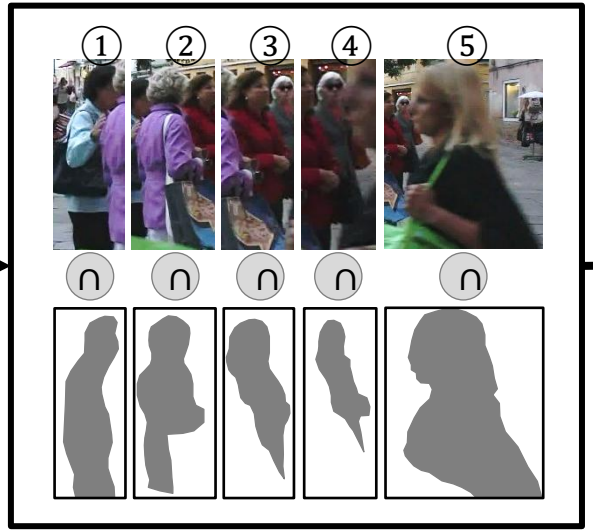
MOT17-02-SDP

METHODOLOGY :: OBJECT DETECTION

Input image



Detection using YOLOv8
(Bbox; Segmentation masks)

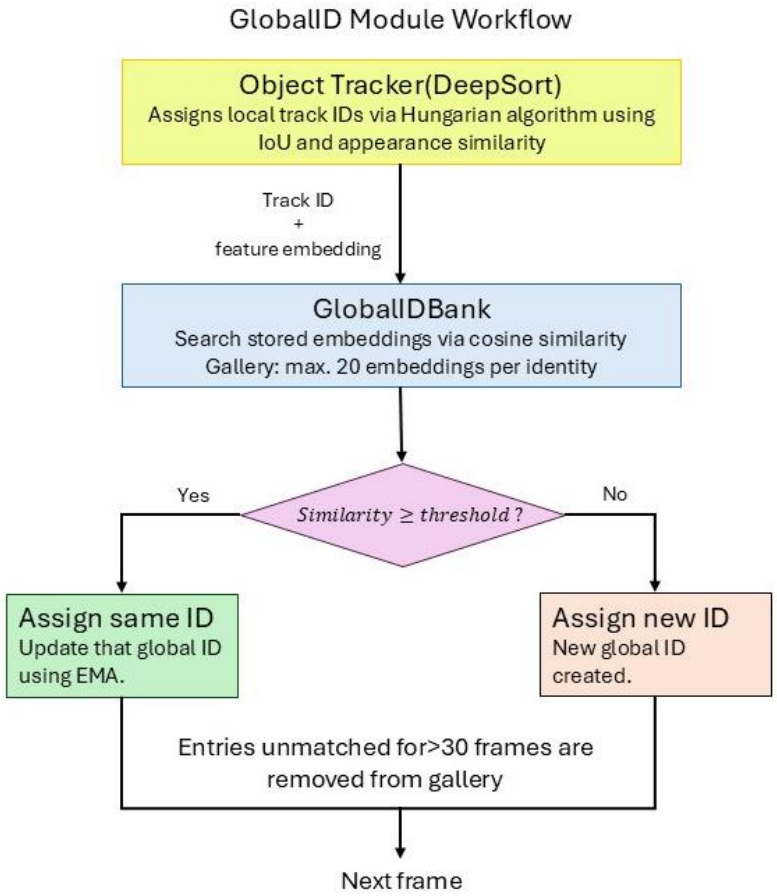


Bbox-aligned crops

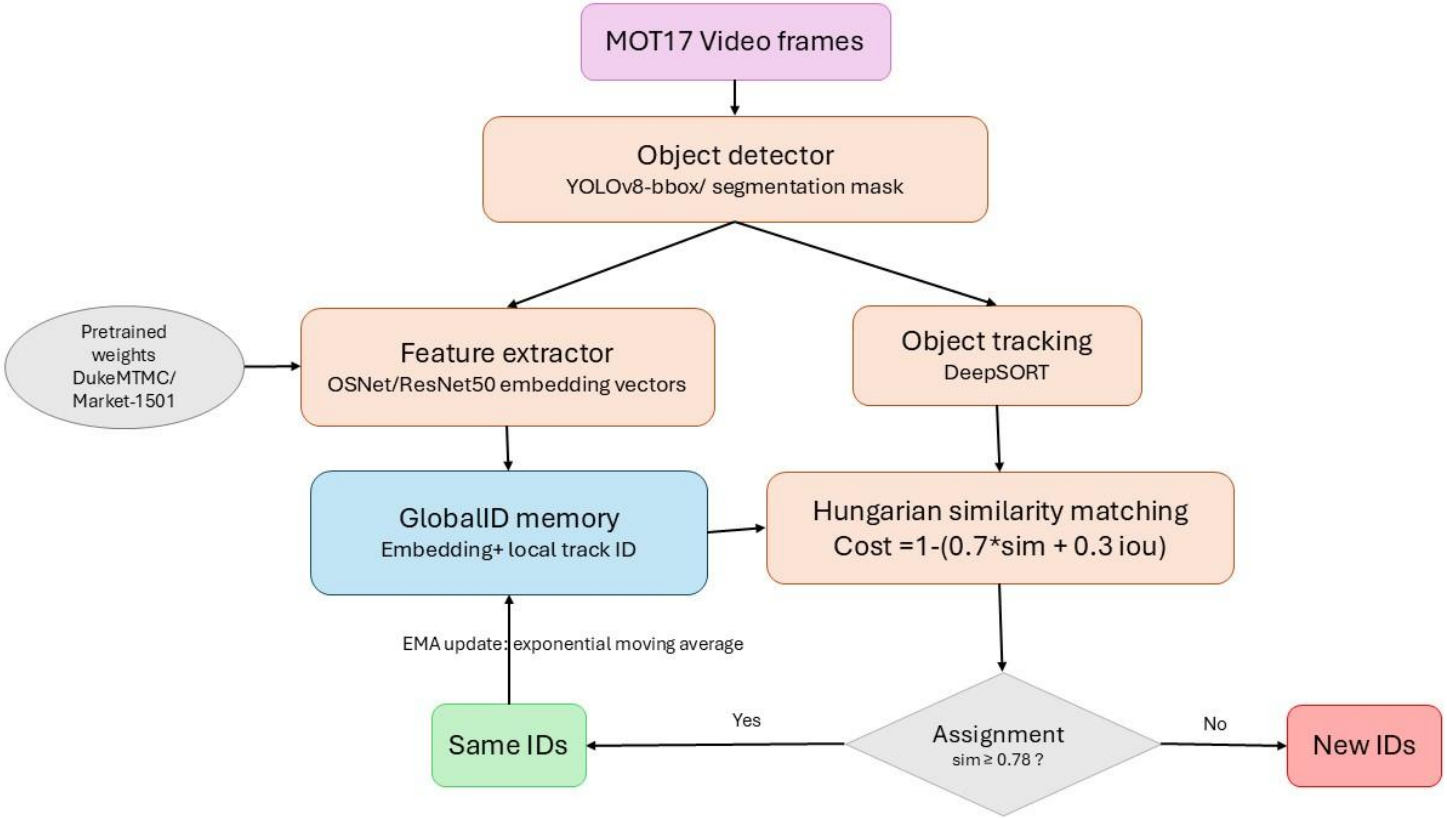
Feature extractor

*Segmentation masks used only for region refinement (not mask-aware embeddings)

METHODOLOGY :: GLOBAL-ID MODULE



COMPLETE PIPELINE



RESULTS

MODEL	Performance comparison		
	IDF1	F-IDF1	FPS
Baseline (YOLOv8+ DeepSORT)	0.2916	0.6412	8.51
OSNet-DukeMTMC (YOLOv8 seg)	0.3918	0.6571	1.38
OSNet-Market-1501 (YOLOv8 seg)	0.3871	0.6583	1.44
ResNet50-DukeMTMC (YOLOv8 seg)	0.3479	0.6354	2.23
ResNet50-Market-1501 (YOLOv8 seg)	0.3479	0.6354	2.46
OSNet-DukeMTMC (YOLOv8)	0.3783	0.6514	7.13
OSNet-Market-1501 (YOLOv8)	0.3864	0.6531	6.55
ResNet50-DukeMTMC (YOLOv8)	0.3084	0.6102	6.03
ResNet50-Market-1501 (YOLOv8)	0.3140	0.6203	5.73

$$IDF1 = \frac{|TP|}{|TP| + |FN| + |FP|}$$

$$F\text{-}IDF1 = \frac{2 \times IDTP_{det}}{2 \times IDTP_{det} + IDFP_{det} + IDSW_{det}}$$

RESULTS :: MOTIVATION FOR F-IDF1

1. IDF1 measures how well identities are preserved over time.
2. Ground Truth(GT) : Manually annotated bounding boxes and identity labels for all visible people in each frame is provided in the dataset.
3. Detected People: Individuals detected by YOLO in model pipeline which differs from the GT. When YOLO misses a person, that person does not enter DeepSORT or OSNet/ResNet, but IDF1 still penalizes the system.

$$IDF1 = \frac{|TP|}{|TP| + |FN| + |FP|}$$

If a person in GT is never detected, the tracker is still penalized. This includes identity switches that DeepSORT could never fix.

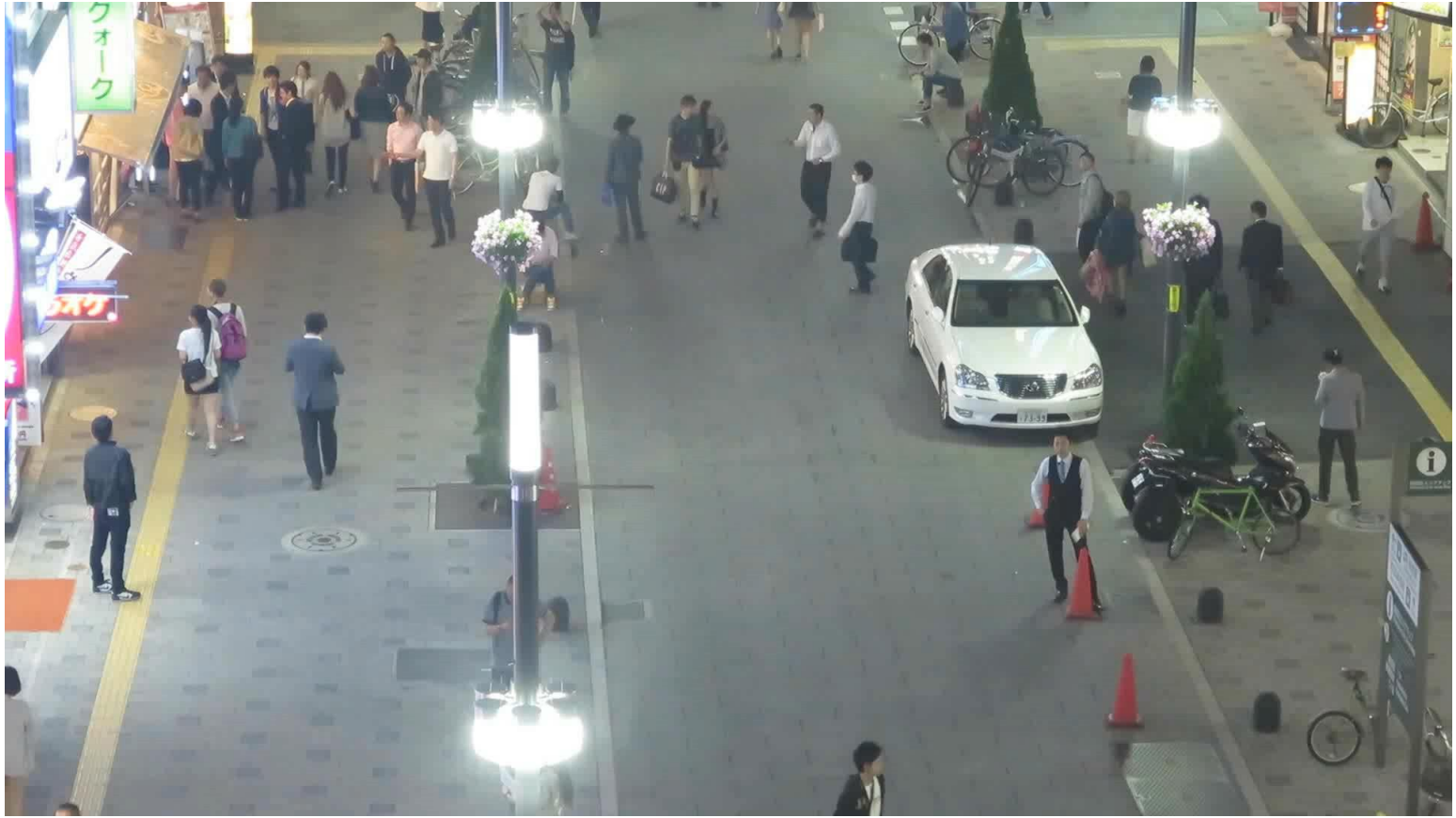
RESULTS :: F-IDF1

Idea: Calculate IDF1 only for people detected by YOLO, to isolate the ReID/tracking performance.

$$\text{Filtered IDF1} = \frac{2 \times \text{IDTP}_{\text{det}}}{2 \times \text{IDTP}_{\text{det}} + \text{IDFP}_{\text{det}} + \text{IDSW}_{\text{det}}}$$

Where instead of False Negative, we change it ID switches of only the detected IDs.
This avoids punishing tracker for detector weakness.

RESULTS



CONCLUSION

① Feature embeddings plays a very important role in person ReID, therefore managing and storing them makes a difference in ID consistency.

② Segmentation masks does not play a significant role in improving the Identity consistency if the feature extractor is not mask aware.

③ A strong GlobalID bank keeps the feature vectors and time to time update the features so that the features of that identity can be more generalized leading to more stable identity assignment.

FUTURE WORK

Limitations and Solutions

	Limitation	Solution
1	The current ReID backbones doesn't work well with segmentation masks.	Make the ReID backbones mask aware.
2	The detector causes the original IDF1 score to decrease.	Try out different detectors to improve the IDF1 score.
3	The current pipeline is computationally costly because of the extra globalID module and DeepSORT CNN feature extractor.	Try other alternatives for tracking and mainitaing the feature vectors.

REFERENCES

[1] W. Luo, J. Xing and X. Zhang, "Multiple object tracking: A review," arXiv preprint arXiv:1409.7618, 2014.

[2] H. Wang, S. Ullah, D. Li and Y. Liu, "Recent advances in deep learning-based person re-identification," Applied Sciences, vol. 9, no. 8, p. 1535, 2019

[3] 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 Proc. European Conf. Computer Vision (ECCV), 2018, pp. 269-286.

[4] R. E. Kalman, "A new approach to linear filtering and prediction problems," Journal of Basic Engineering, vol. 82, no. 1, pp. 35-45, 1960.

[5] A. Bewley, Z. Ge, L. Ott, F. Ramos and B. Upcroft, "Simple online and realtime tracking," in Proc. IEEE Int. Conf. Image Processing (ICIP), 2016, pp. 2956-2960.

[6] N. Wojke, A. Bewley and D. Paulus, "Simple online and realtime tracking with a deep association metric," in Proc. IEEE Int. Conf. Image Process. (ICIP), Beijing, China, 2017, pp.3645-3649.

[7] Z. Zhang, L. Sun, Q. Leng and S. Liao, "Towards real-time multi-object tracking with adaptive appearance models," Pattern Recognition Letters, vol. 136, pp. 213-220, 2020.

[8] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp.770-778.

[9] K. Zhou, Y. Yang, A. Cavallaro and T. Xiang, "Omni-scale feature learning for person re-identification," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2019.

[10] Y. Li, X. Zhu and S. Gong, "Unsupervised person re-identification with stochastic training strategy," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2020.

[11] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in Proc. Int. Joint Conf. Artificial Intelligence (IJCAI), 1981.

[12] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016.

REFERENCES

[13] A. Bochkovskiy, C.-Y. Wang and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[14] G. Wang, Y. Yuan, X. Chen, J. Li and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in Proc. ACM Int. Conf. Multimedia (ACM MM), 2018.

[15] S. He, H. Luo, P. Wang, F. Wang, H. Li and W. Jiang, "TransReID: Transformer-based object re-identification," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2021.

[16] P. Bergmann, T. Meinhardt and L. Leal-Taixé, "Tracking without bells and whistles," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2019.

[17] Y. Zhang, C. Wang, X. Wang, W. Zeng and W. Liu, "FairMOT: On the fairness of detection and re-identification in multiple object tracking," Int. J. Comput. Vis., vol. 129, no. 11, p. 3069–3087, 2021.

[18] Z. Wang, L. Zheng, Y. Liu and S. Wang, "Towards real-time multi-object tracking," in Proc. European Conf. Computer Vision (ECCV), 2020.

[19] L. Jin, Z. Zheng and Y. Sun, "Learning a self-adaptive gallery for unsupervised person re-identification," IEEE Trans. Image Process., vol. 31, p. 5282–5294, 2022.

[20] Z. Zheng, L. Zheng and Y. Yang, "Open-world person re-identification," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 8, p. 2630–2647, 2021.

[21] A. Milan, L. L. Taixé, I. Reid, S. Roth and K. Schindler, "MOT16: A benchmark for multi-object tracking," arXiv spreprint arXiv:1603.00831, 2016.

[22] E. Ristani, F. Solera, R. Zou, R. Cucchiara and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in Proc. European Conf. Comput. Vis. (ECCV), 2016.

[23] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. L. Taixé and B. Leibe, "HOTA: A higher order metric for evaluating multi-object tracking," Int. J. Comput. Vis., vol. 129, p. 548–578, 2021.

[24] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," Ultralytics, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.

[25] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian, "Scalable person re-identification: A benchmark," in Proc. IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1116-1124.

Thankyou