M.S. Project Defense

Multi-target Multi-camera Person Tracking

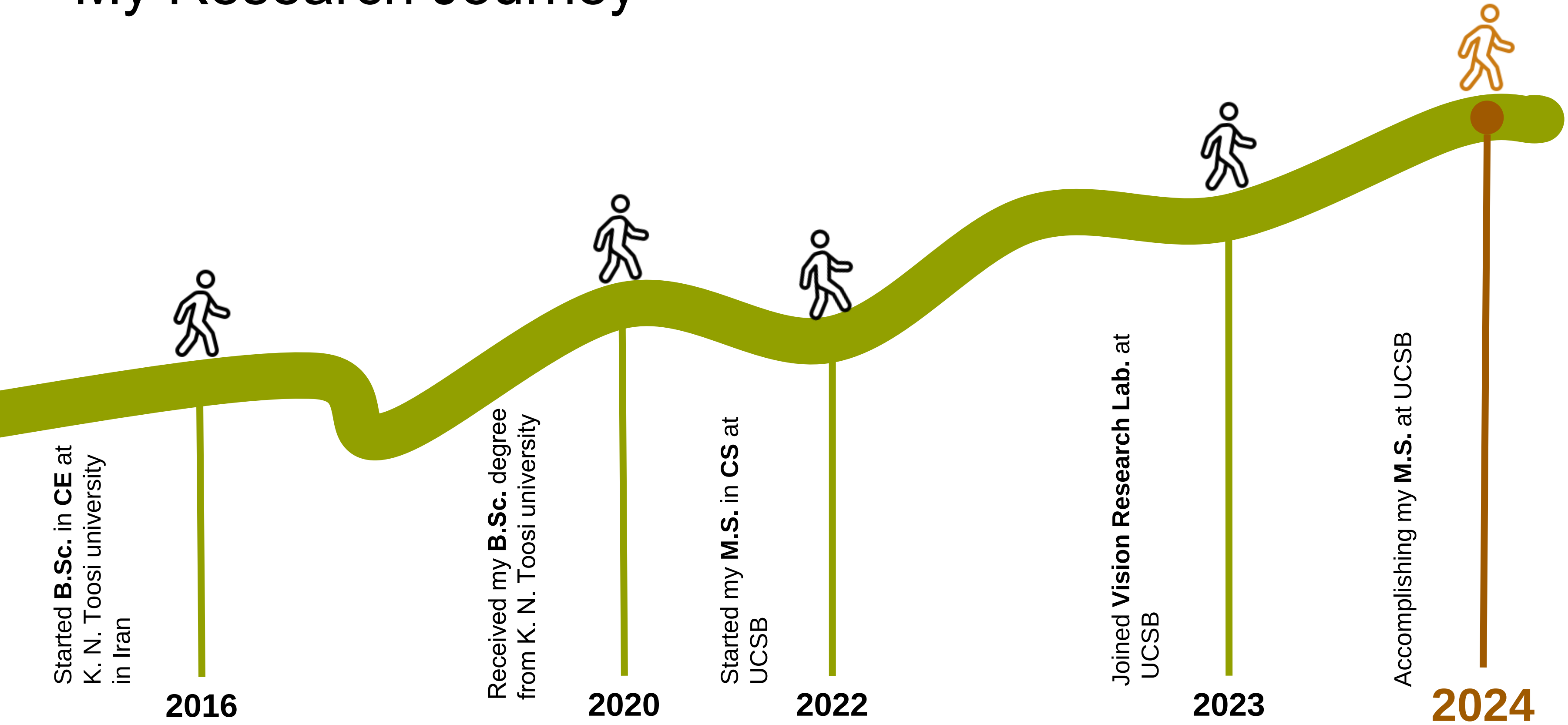
Kimia Afshari

Committee members:
Prof. B.S. Manjunath (Chair)
Prof. Tobias Höllerer

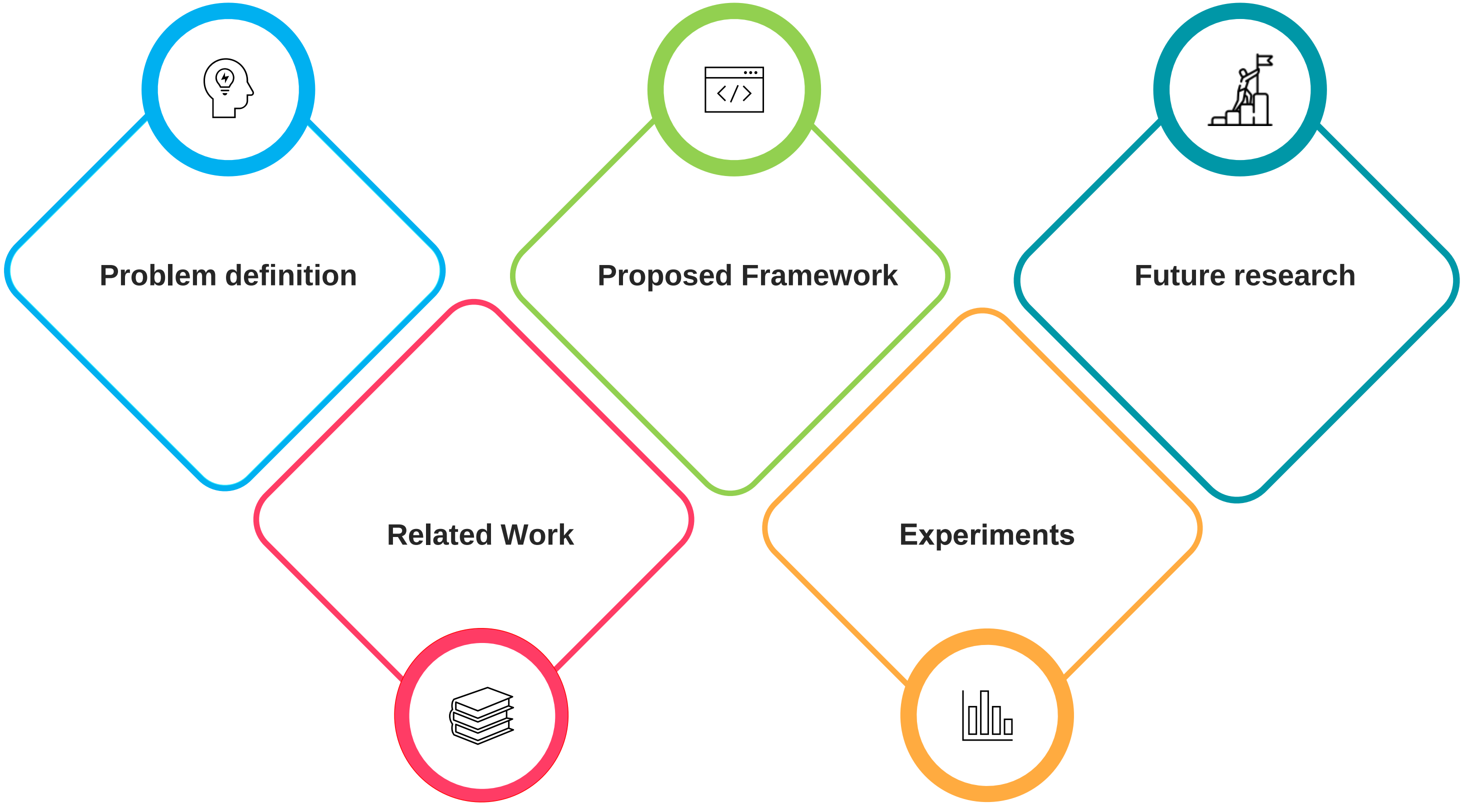
Summer 2024

Vision Research Lab. (VRL)

My Research Journey

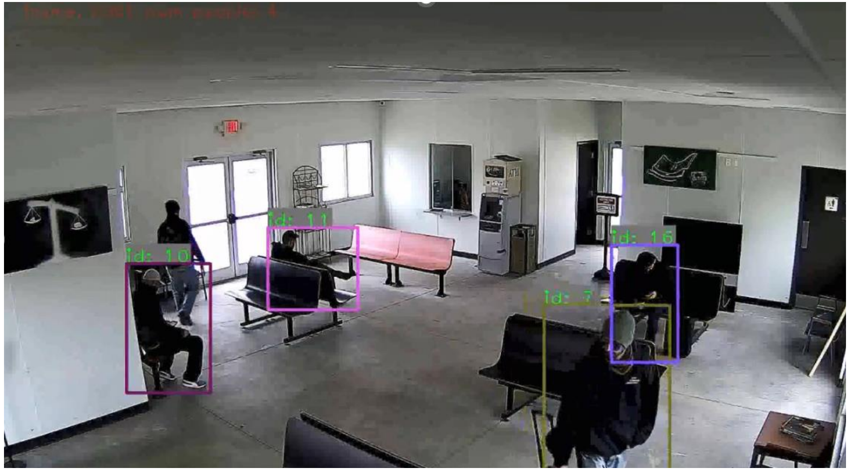


Presentation Outline



Problem Definition

- Consistently track and re-identify people across multi-view cameras.
- Generalize to both overlapping and non-overlapping cameras in indoor and outdoor.
- Address challenges in viewpoints variance, illumination changes and frames quality through refinement techniques.



2018-03-05.13-15-01.13-20-01.bus.G331

2018-03-05.13-15-00.13-20-00.bus.G506

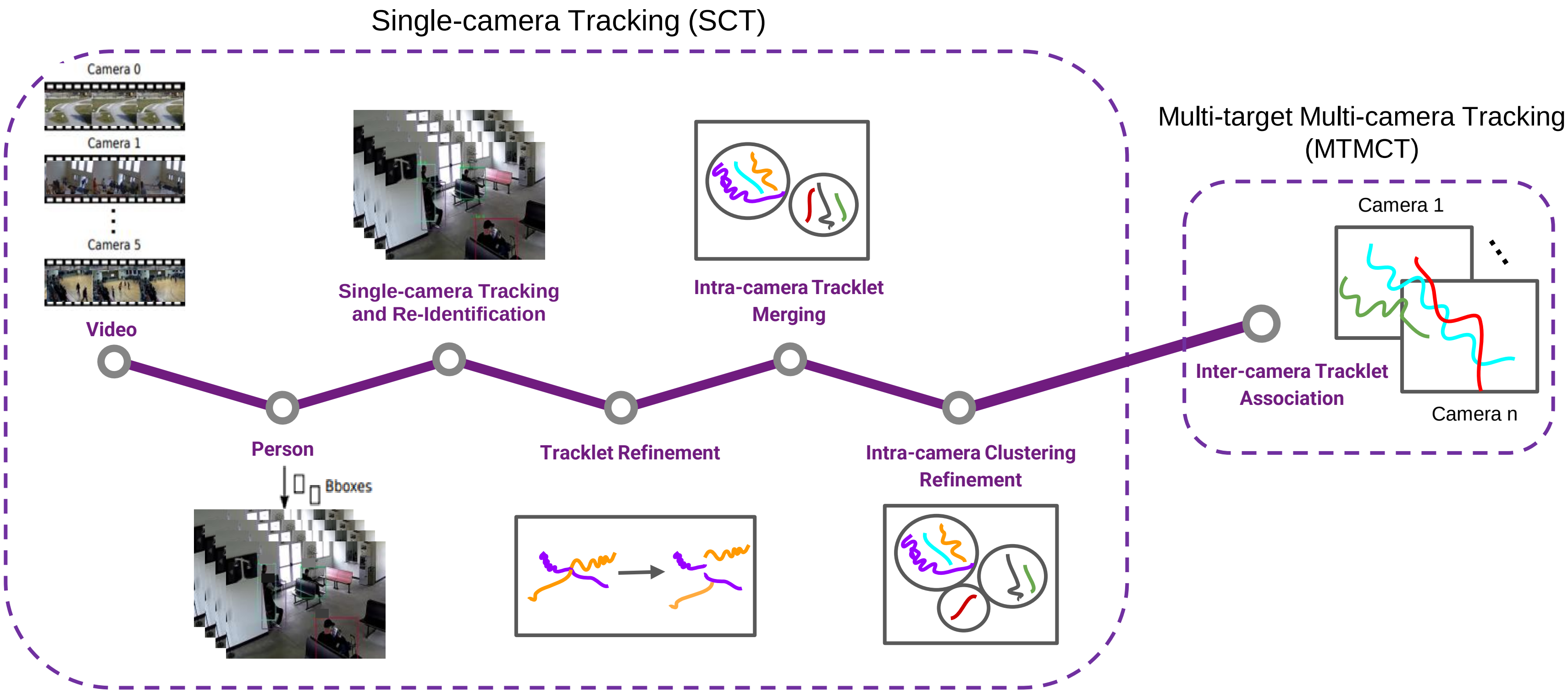
Importance of the Problem

- Public safety and security
  - detect and prevent anomalies and suspicious activities in surveillance systems
- Traffic monitoring and management
  - detect traffic violations, accident and congestion, law enforcement, etc.
- Healthcare and elderly care
  - patient's movement monitoring

Person **A** hands off a box to Person **B**

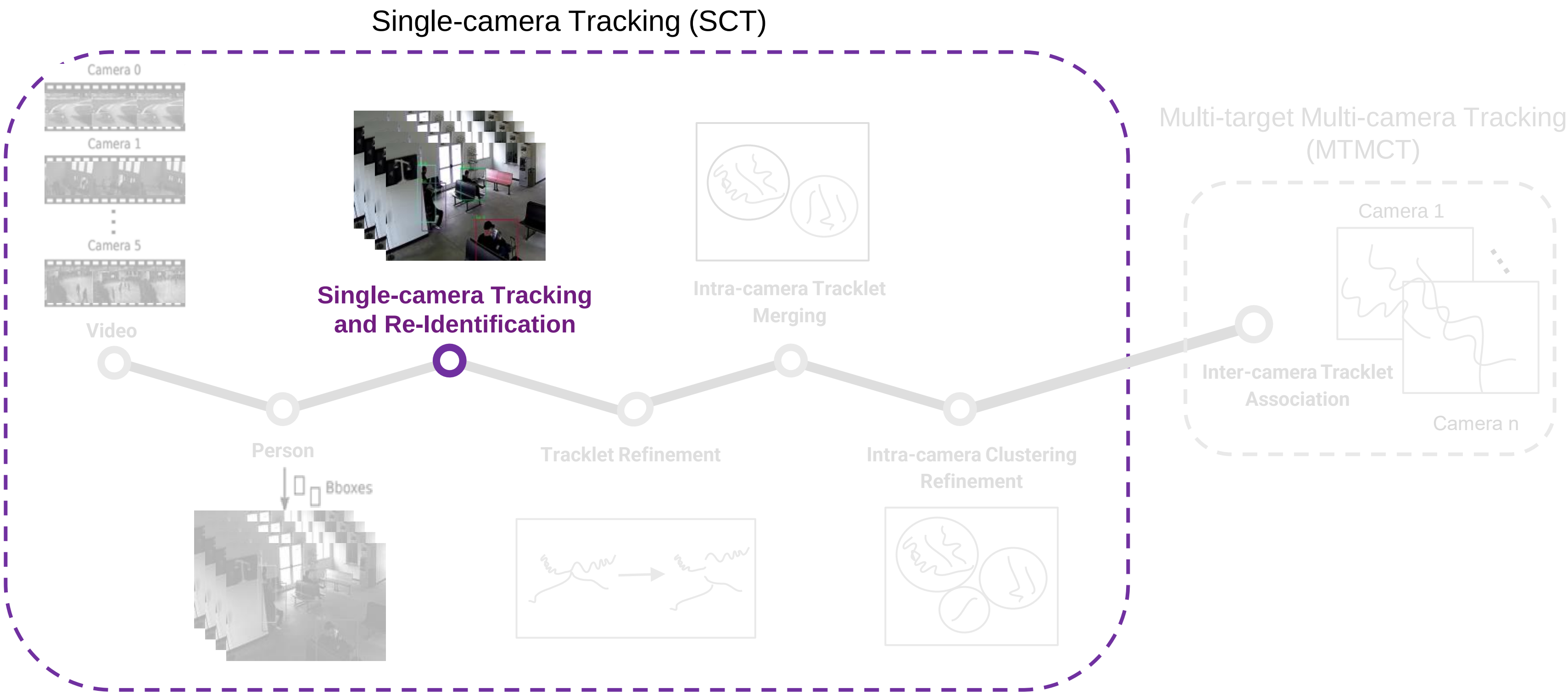
Person **B** enters a building carrying the box handled by Person **A**

Proposed Tracking Framework



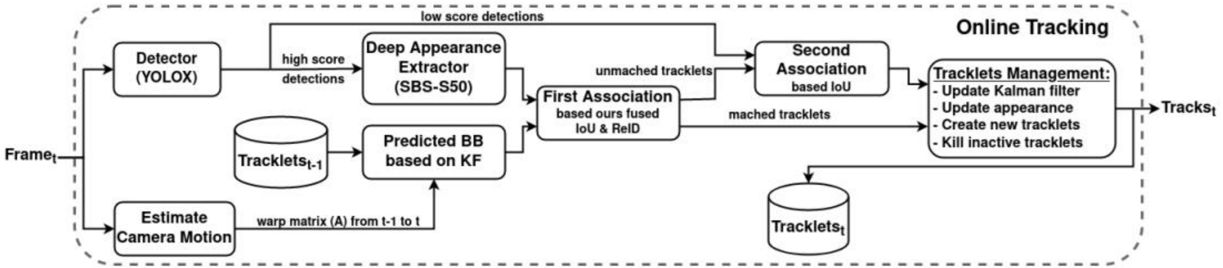
* Tracklet is a sequence of consecutive frames in which an object is consistently detected and tracked.

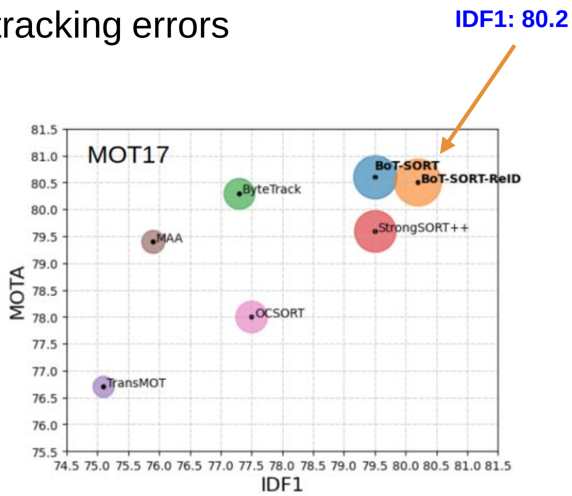
First Step: Single Camera Tracking



Single-view Multi-Object Tracking

- [1] **BoT-SORT**: Robust Associations Multi-Pedestrian Tracking
- Fuses motion and appearance features
- Uses **Kalman filter** motion-based future position estimator
- Employs [2] **BoT (SBS)** appearance-based feature extractor to reduce tracking errors

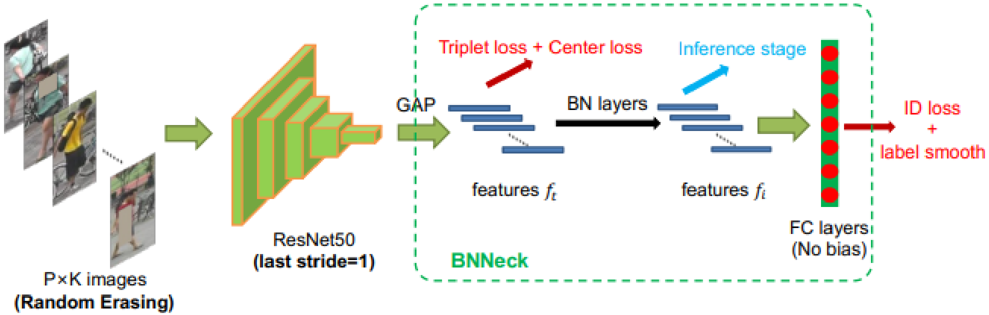




[1] Aharon, Nir, Roy Orfaig, and Ben-Zion Bobrovsky. "BoT-SORT: Robust associations multi-pedestrian tracking." *arXiv preprint arXiv:2206.14651* (2022).
[2] Luo, Hao, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. "Bag of tricks and a strong baseline for deep person re-identification." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 0-0. 2019.

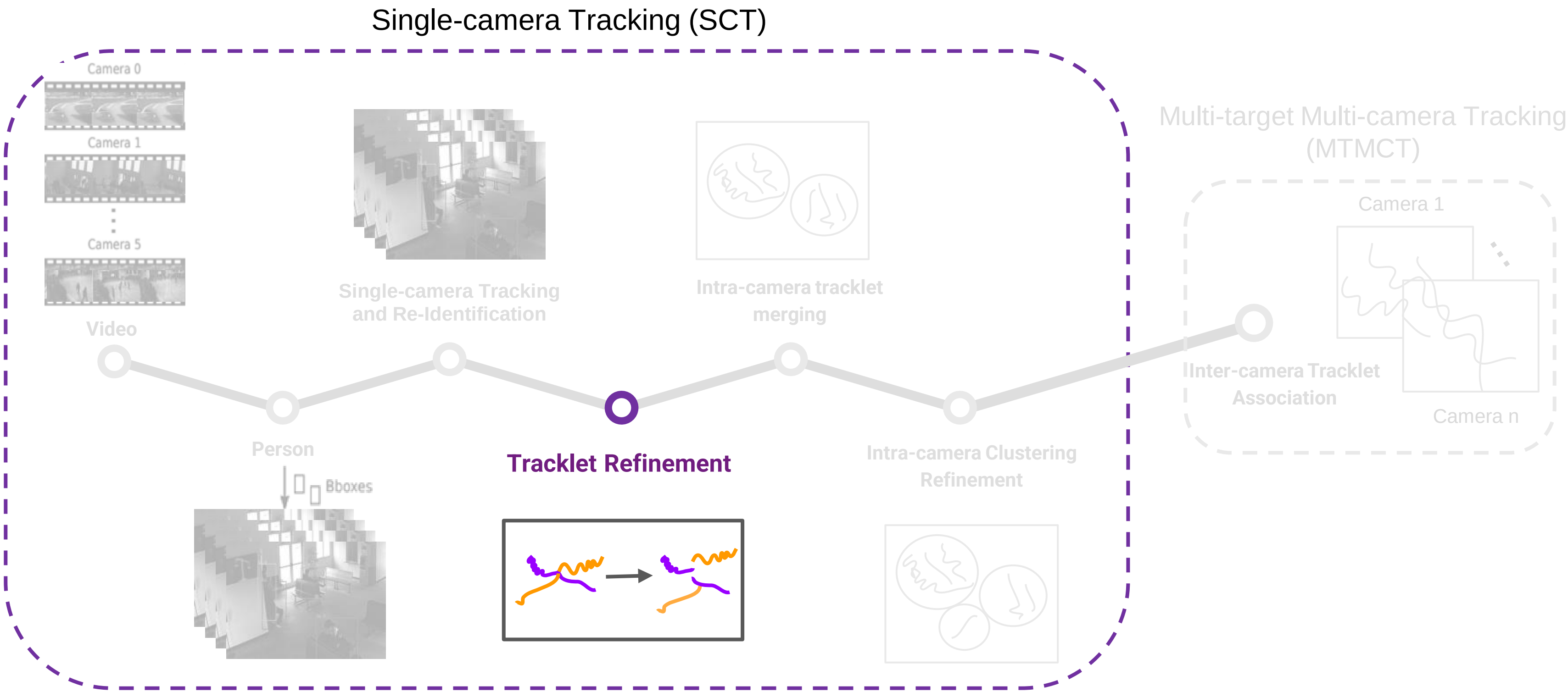
BoTSORT Deep Appearance Feature Extractor

- Adopts the strong baseline on top of [2] **BoT (SBS)** from [3] **FastReID** Tool
- Uses [4] **ResNest50** as the backbone
- Features are result of the batch normalization layer
- Features have 1024 dimensions



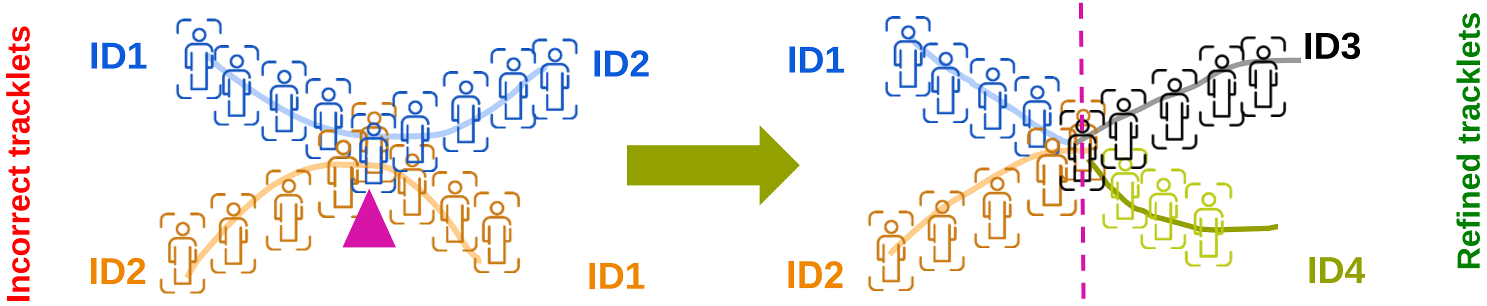
The pipeline is taken from [2] .
[3] He, Lingxiao, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. "Fastreid: A pytorch toolbox for general instance re-identification." In Proceedings of the 31st ACM International Conference on Multimedia, pp. 9664-9667. 2023.
[4] Zhang, Hang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun et al. "Resnest: Split-attention networks." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2736-2746. 2022.

Next Step: Tracklet Refinement



Tracklet Refinement

- Splits tracklets containing different identities
- Use intra-variance of tracklets to find should-split tracklets
- Higher [5] **intra-variance** → Higher **appearance variation** → tracklet has different identities
- Applies **K-Means** clustering to split the tracklet
- Reduces errors caused by single-camera trackers



[5] The intra-variance method is adopted from Li, Zongyi, Runsheng Wang, He Li, Bohao Wei, Yuxuan Shi, Hefei Ling, Jiazhong Chen, Boyuan Liu, Zhongyang Li, and Hanqing Zheng. "Hierarchical clustering and refinement for generalized multi-camera person tracking." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5520-5529. 2023.

Intra-variance Calculation for Tracklet *t*

- [5] Intra-variance is considered as the cosine distance between each * appearance feature and the mean of all appearance features in a tracklet.
- Tracklets with intra-variance greater than a threshold will be split.

For tracklet $\boldsymbol{t}$ with $\boldsymbol{N}$ frames:

$$\overline{f^{t}} = \frac{1}{N}\sum_{i=1}^{N} f_i^t$$

$$V_{\text{intra}} = \frac{1}{N}\sum_{i=1}^{N} D_{\text{cosine}}(f_i^t, \overline{f}^t)$$

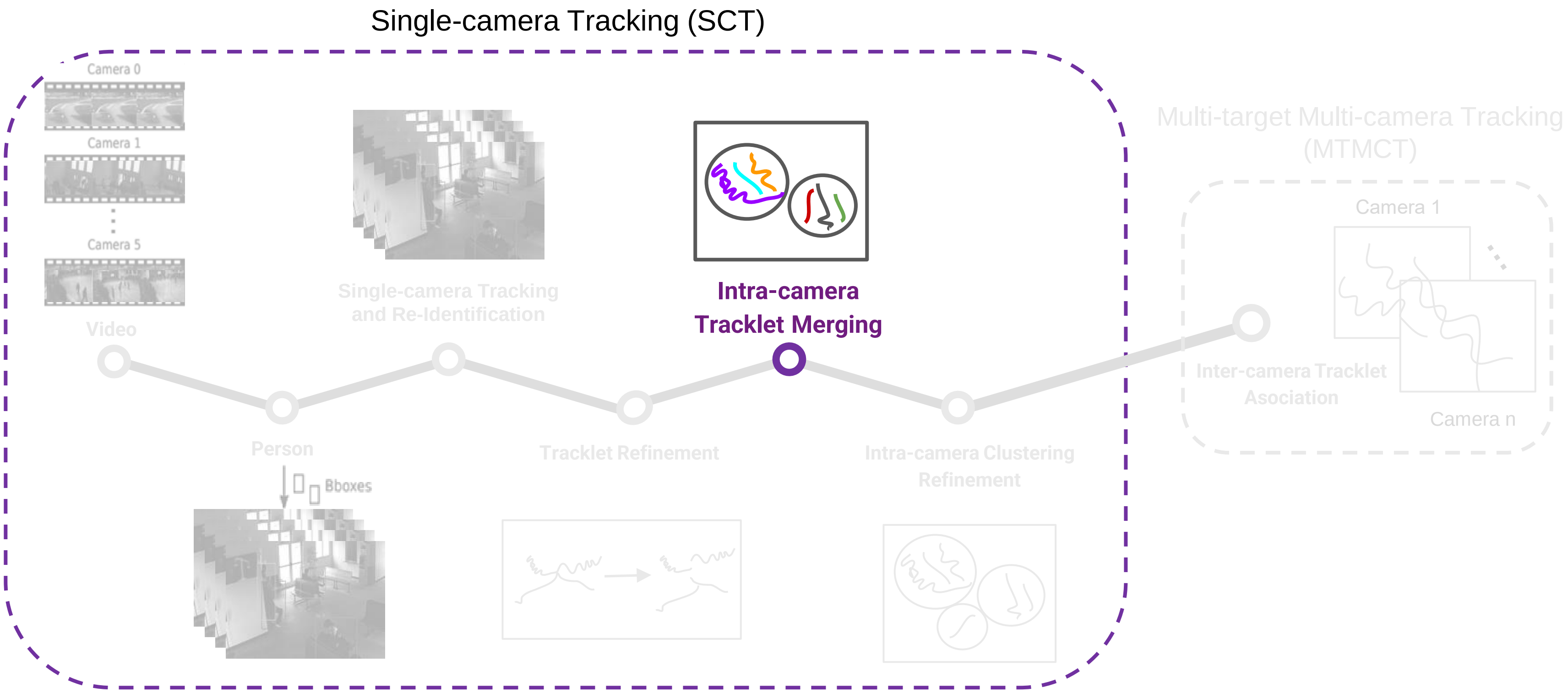
f_i: *i*-th appearance feature of the tracklet.
$\overline{f}$: mean of all appearance features in a tracklet.
D_{cosine}: cosine distance matrix

$$D_{\text{cosine}}\left(f_i, f_j\right) = 1 - \frac{f_i \cdot f_j}{||f_i|| \, ||f_j||}$$

* Note: we use the same appearance features extracted from [2] BoT during tracking.

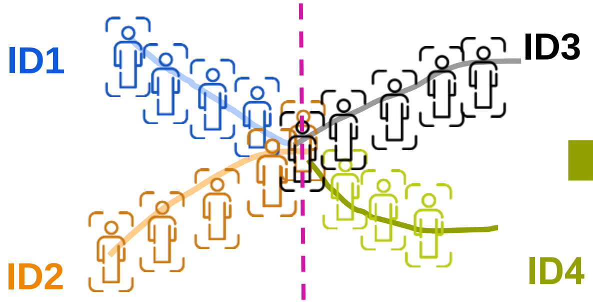
[5] The intra-variance method is adopted from Li, Zongyi, Runsheng Wang, He Li, Bohao Wei, Yuxuan Shi, Hefei Ling, Jiazhong Chen, Boyuan Liu, Zhongyang Li, and Hanqing Zheng. "Hierarchical clustering and refinement for generalized multi-camera person tracking." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5520-5529. 2023.

Next Step: Intra-camera Tracklet Association

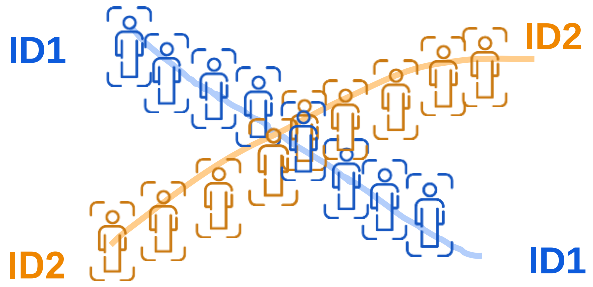


Hierarchical Constrained Tracklet Association

- Uses **Agglomerative** clustering method to group tracklets of the same identity through their appearance features.
- Calculate **aggregated distance matrix** between each pairs of tracklets
- Use the aggregated distance matrix to cluster tracklets



Split tracklets



Reunited tracklets

Aggregated Distance Matrix

- Appearance Feature Distance Matrix
  - Use mean of appearance features across all frames for each individual.
  - Apply cosine dissimilarity metric to find appearance feature distance between individuals.

$$D_{\text{appearance}}\left(f_i, f_j\right) = D_{\text{cosine}}\left(f_i, f_j\right)$$

- Temporal Distance Matrix
  - Intra-camera
    - In each frame, not two people are allowed to be matched.
  - Inter-camera
    - Across time-overlapping videos, not two people are allowed to be matched.

$$D^{i,j}_{\text{temporal}} = \begin{cases} 1, \ \textit{if} \ \{t^i_i, t^i_o\} \cap \{t^j_i, t^j_o\} \neq \varnothing \\ 0, \ \text{else} \end{cases}$$

i, *j*: tracklets *i* and *j*
t_i^i: the time that tracklet *i* enters the scene
t_o^i: the time that trackelet *i* exits the scene

- Aggregated Distance Matrix

$$\boxed{D = \alpha D_{\text{appearance}} + \beta D_{\text{temporal}}}$$

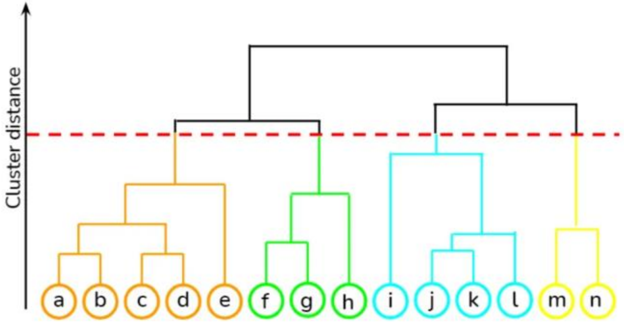
* **Greater** ***β*** ensures stronger adherence to temporal constraints.

Agglomerative Clustering

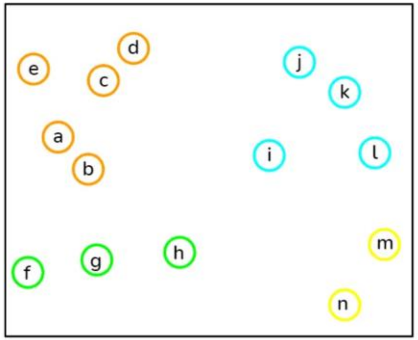
- Recursively merges pair of clusters of sample data.
- Uses linkage distance to stop iteration.
- Feed precomputed constrained distance matrix as input data.
- Use *dendrogram plot to analyze hierarchical merging distances.

$$C_i = \bigcup_{p,q} \left\{ C_p \cup C_q \mid D_{pq} \le \tau \right\}$$

C_i: cluster *i*
U: union
D_{pq}: distance between *p, q*
τ: merging distance threshold



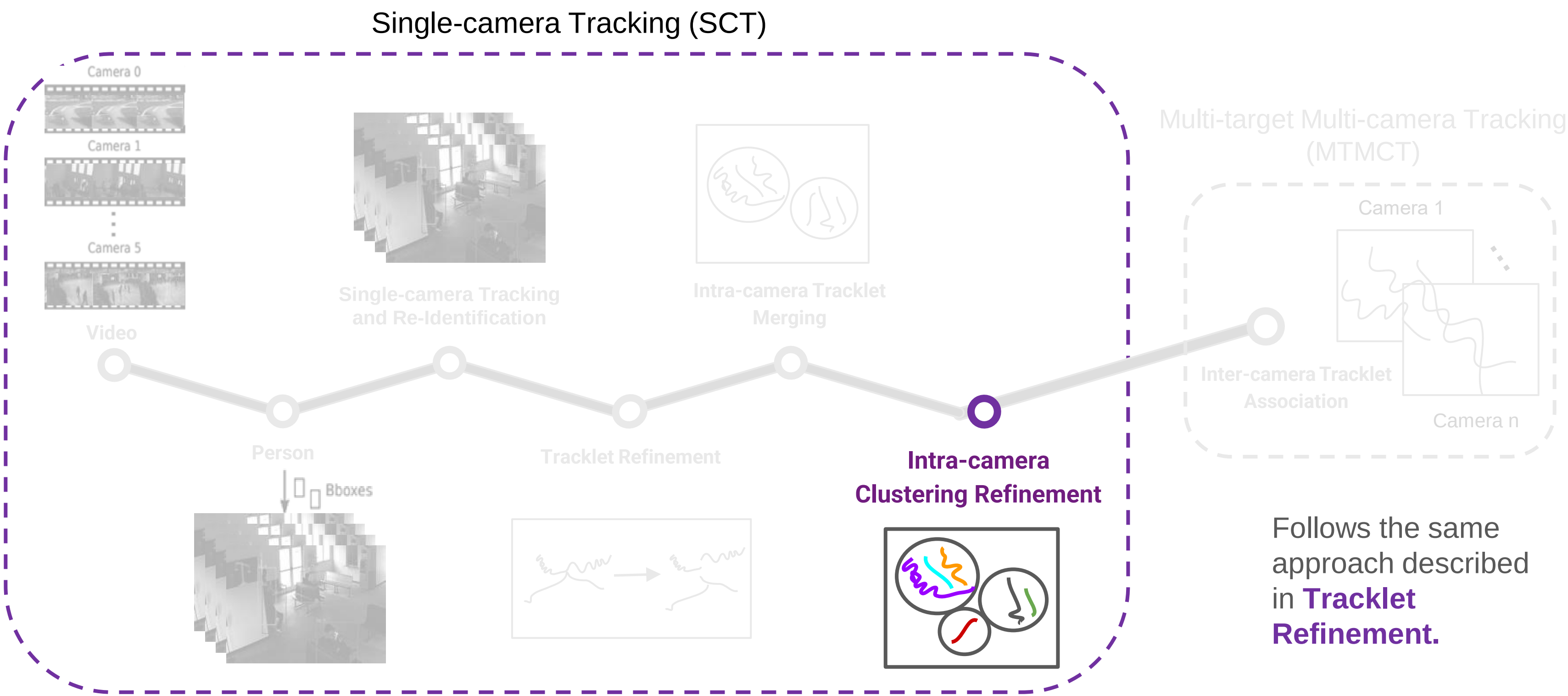
Data Samples



Clusters

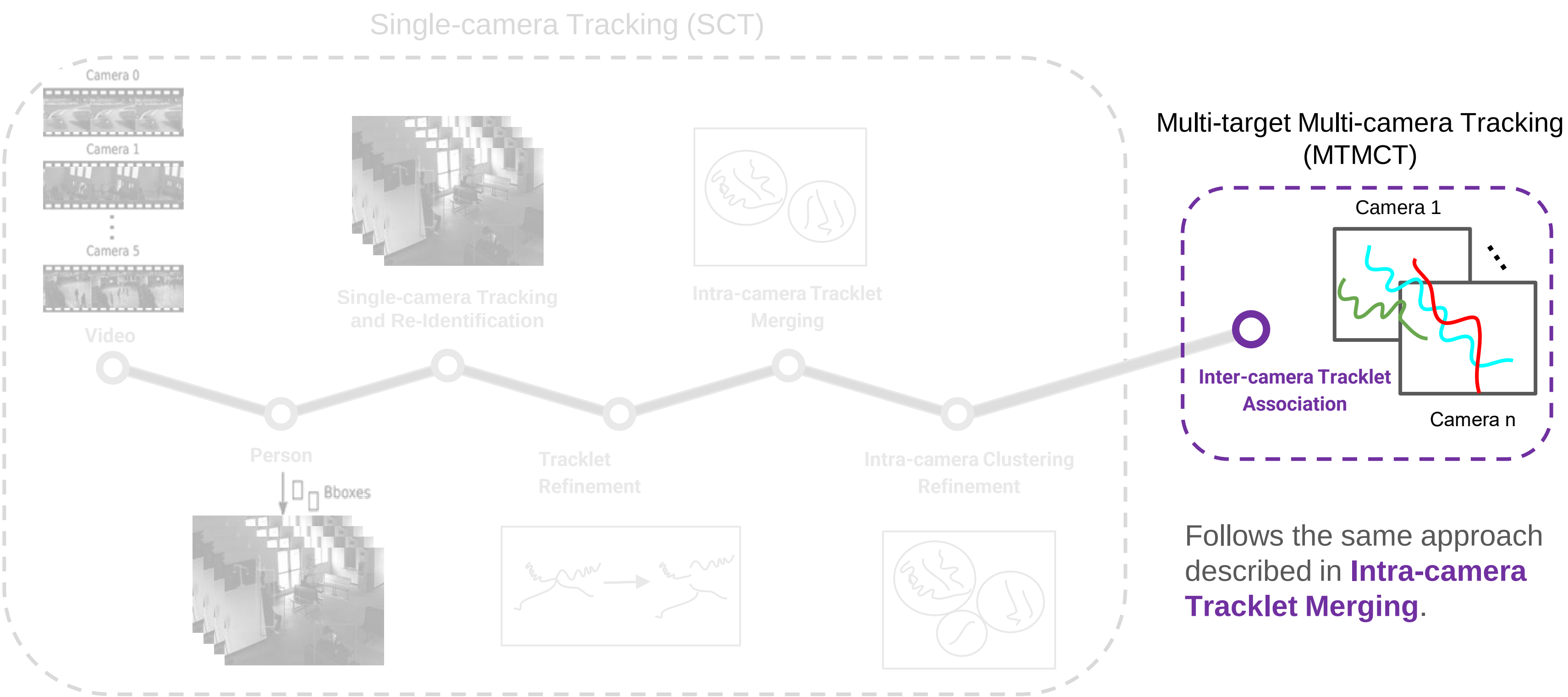
* A dendrogram is a tree-like diagram that displays the arrangement of the clusters produced by hierarchical clustering.
* Dendrogram is taken from https://towardsdatascience.com/hierarchical-clustering-explained-e59b13846da8.

Next Step: Intra-camera Clustering Refinement



Follows the same approach described in **Tracklet Refinement.**

Final Step: Inter-camera Tracklet Association



Follows the same approach described in **Intra-camera Tracklet Merging**.

Experiments

Single-camera Tracking	• Default [1] BoT-SORT with tuned parameters.
Hierarchical Clustering	• Intra-camera distance threshold τ = 0.15 • Inter-camera distance threshold τ = 0.20
Pre-clustering Tracklet Refinement	• Intra-variance threshold τ = 0.10
Post-clustering Tracklet Refinement	• Intra-variance threshold τ= 0.10
Dataset	• MEVA, a large-scale multi-view activity recognition dataset
Results	• Report quantitative and qualitative results on MEVA dataset

$$C_i = \bigcup_{p,q} \left\{ C_p \cup C_q \,\middle|\, D_{pq} \leq \tau \right\}$$

$$V_{\text{intra}} > \tau$$

[4] MEVA Dataset

- A challenging **large-scale video** dataset designed for **activity recognition** in **multi-camera** environments

- Contains over 9,300 hours of untrimmed videos with diverse backgrounds, camera poses, illuminations and indoor/outdoor scenes

- Each camera has lots of videos taken during different times of the day, month and year, each split into 5-min length

- Has **158** unique people wearing **598** outfits taken seen in **33** camera views

- We focus on **5** connected cameras:
  - 14 videos
  - Each video having 9,000 frames (5 mins)
  - Containing 20 unique people across cameras

[4] Dataset from Corona, Kellie, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. "Meva: A large-scale multiview, multimodal video dataset for activity detection." In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 1060-1068. 2021.

MEVA Dataset Sample Data

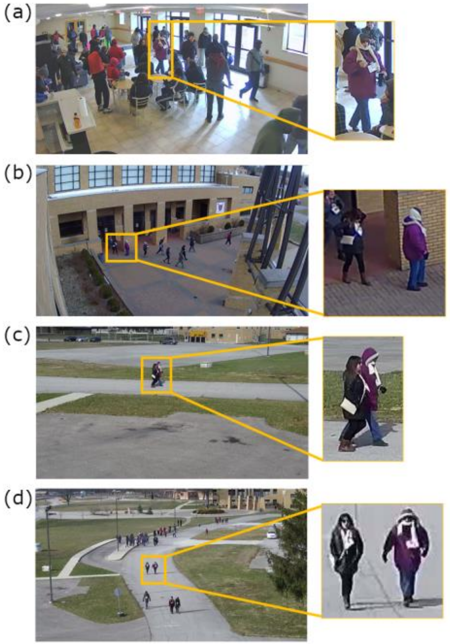
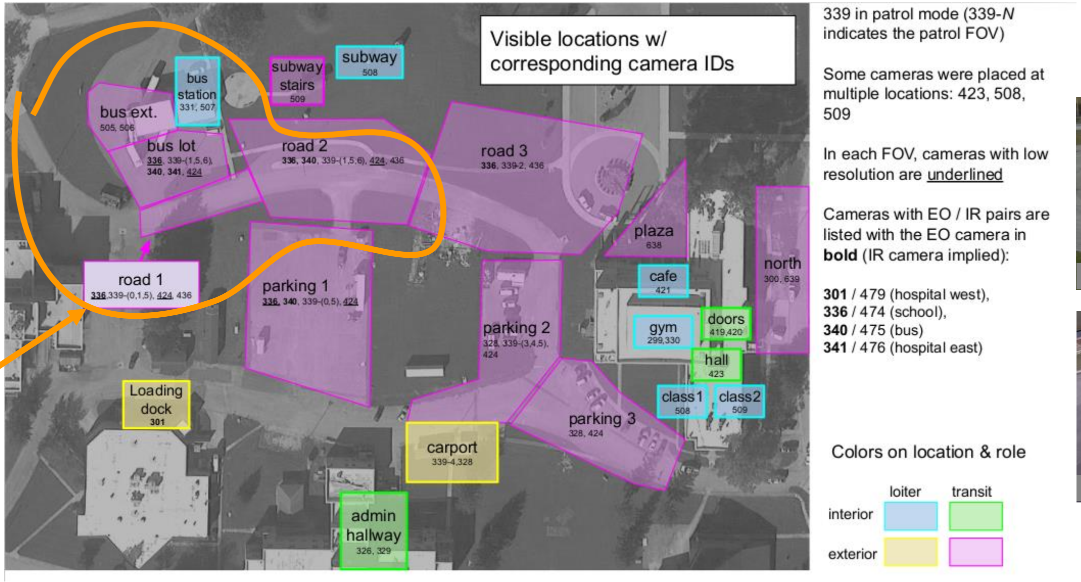


Figure 7. The actor in the purple jacket, actor 544 in Figure 6, is visible in multiple cameras during the scenario. Her height is (a) 301 pixels, (b) 118 pixels, (c) 176 pixels, and (d) 89 pixels in each of the respective fields of view.

* Images are brightened for better visibility.

[4] Images are taken from Corona, Kellie, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. "Meva: A large-scale multiview, multimodal video dataset for activity detection." In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 1060-1068. 2021.

MEVA Dataset Site Map



Bus stop ext.

Bus stop int.

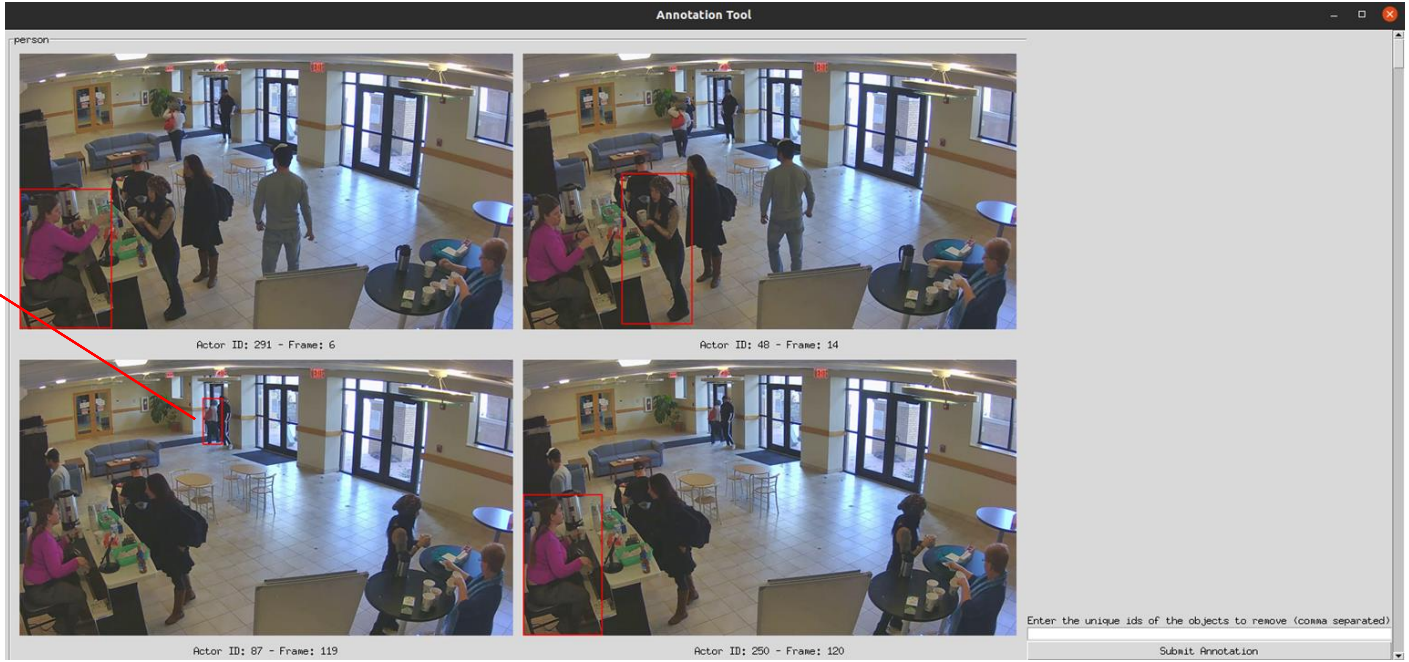
[4] Images are taken from Corona, Kellie, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. "Meva: A large-scale multiview, multimodal video dataset for activity detection." In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 1060-1068. 2021.

A Problem with MEVA Dataset

- Used an automated tool to annotate people
- Each video has many disconnected tracklets, resulting in an **inconsistent ID** associated with the same person
- IDs are inconsistent across cameras, preventing from evaluating the framework

Data Annotator Tool

- Use a developed mava labeler tool to assign unique IDs to individuals across cameras



Manual Annotation Results

Before

After

Evaluation Metrics

- IDF1
  - assesses how well the **tracking** system maintains consistent **identities** over time.
- IDP
  - measures the proportion of correct re-identifications out of all re-identifications made.
- IDR
  - measures the proportion of correct re-identifications out of all actual re-identifications.
- IDS
  - counts number of **identity switches** across tracklets.

$$\mathrm{IDF1} = \frac{2\times \mathrm{IDTP}}{2\times \mathrm{IDTP} + \mathrm{IDFP} + \mathrm{IDFN}} \qquad \mathrm{IDP} = \frac{\mathrm{IDTP}}{\mathrm{IDTP} + \mathrm{IDFP}} \qquad \mathrm{IDR} = \frac{\mathrm{IDTP}}{\mathrm{IDTP} + \mathrm{IDFN}}$$

IDTP: True Positives (correct re-identifications)
IDFP: False Positives (incorrect re-identifications)
IDFN: False Negatives (missed identifications)

Quantitative Results

method	IDF1	IDP	IDR	IDs
Pipeline w/o refinement	27.3%	22.2%	34.7%	375
Pipeline + pre-clustering refinement	32.2%	26.9%	37.9%	386
Pipeline + pre/post-clustering refinement	34.8%	28.8%	44.3%	353

Pre-clustering Tracklet Refinement Effects

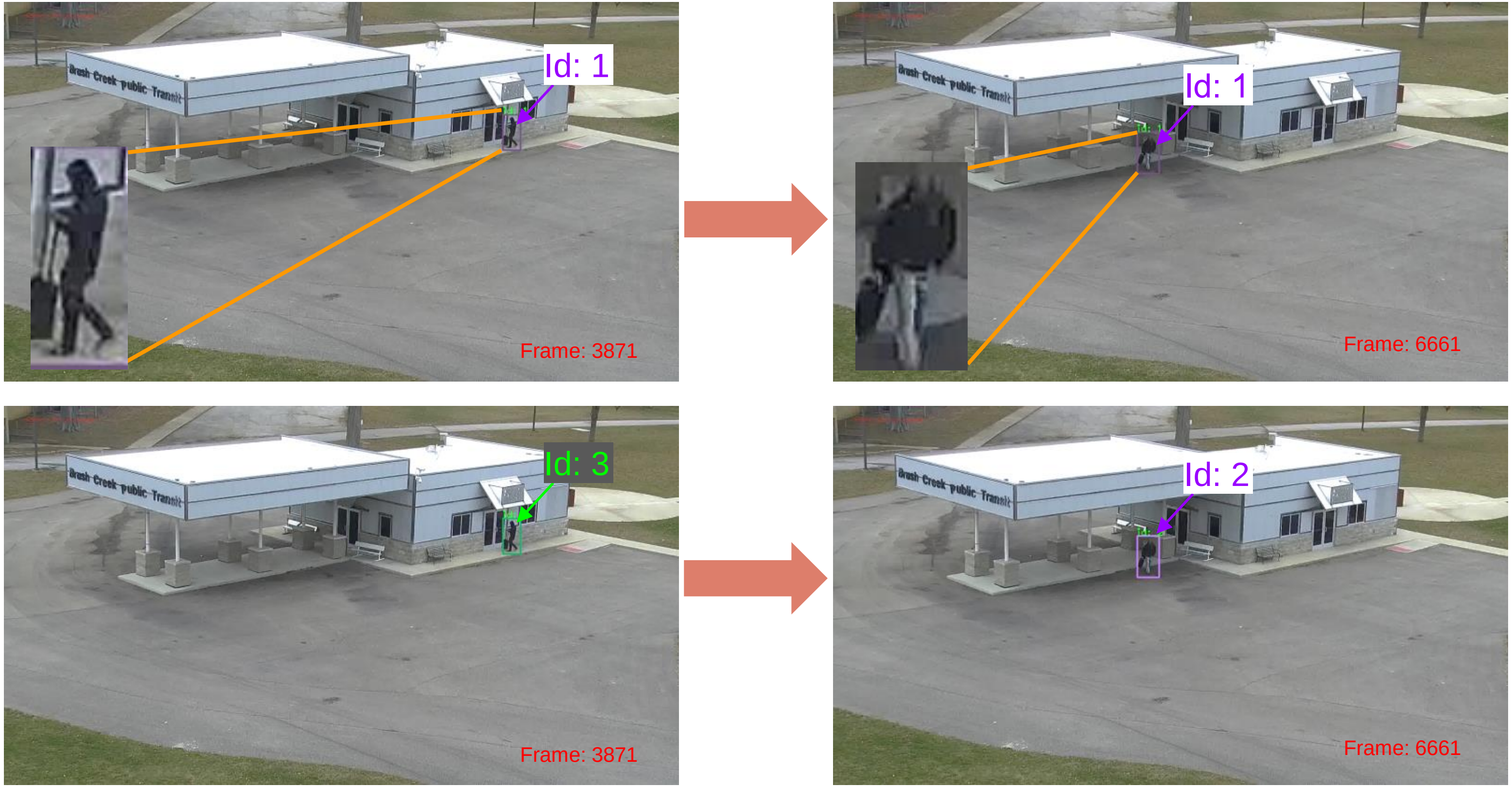
* Results are brightened for better visibility.

2018-03-05.13-10-01.13-15-01.bus.G331

Post-clustering Refinement Effects

Before

After



2018-03-05.13-15-00.13-20-00.hospital.G341

* Results are brightened for better visibility.

Results – Single-camera Re-appearance

Initial SCT tracklets: 27
Refined tracklets: 29
Clustered tracklets: 8
Refined clusters: 9

2018-03-05.13-15-01.13-20-01.bus.G331

* Different box colors represent different identities.

Results – Different Time

2018-03-05.13-10-01.13-15-01.bus.G331

2018-03-05.13-15-01.13-20-01.bus.G331

* Different box colors represent different identities.

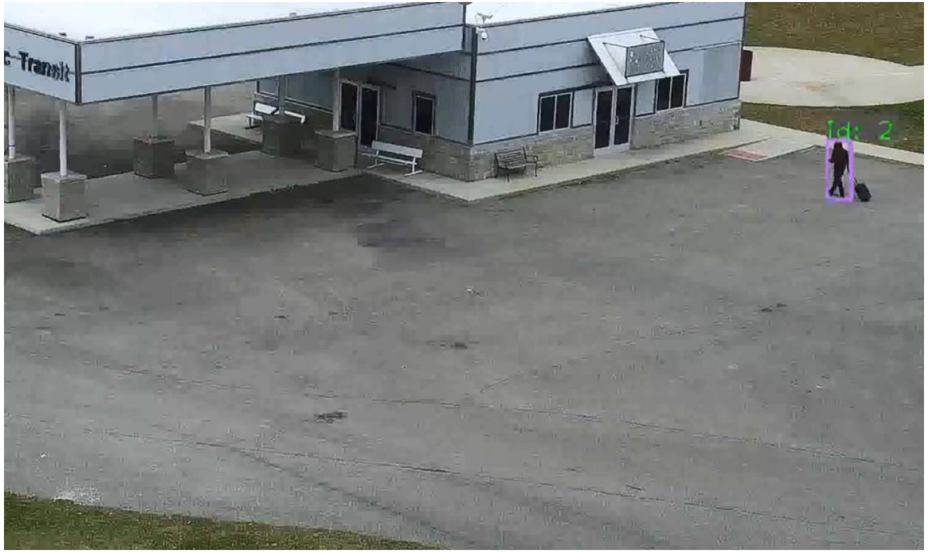
Results – Different Cameras

2018-03-05.13-15-00.13-20-00.bus.G506

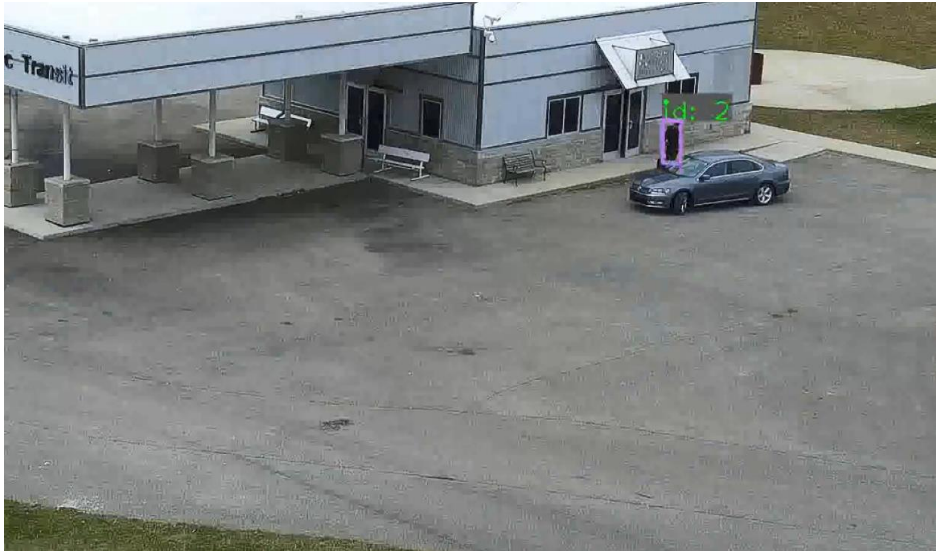
2018-03-05.13-15-00.13-20-00.hospital.G341

* Different box colors represent different identities.
* Videos are cropped to make detections appear larger for better visibility.

Results – Different Time



2018-03-05.13-15-00.13-20-00.hospital.G341



2018-03-05.13-20-00.13-25-00.hospital.G341

* Different box colors represent different identities.
* Videos are cropped to make detections appear larger for better visibility.

Failure (ID Switch)

2018-03-05.13-15-00.13-20-00.bus.G506

* Different box colors represent different identities.

Failure (Incorrect Matching)

2018-03-05.13-20-00.13-25-00.hospital.G341

* Different box colors represent different identities.

Conclusion

- MTMCT is a very challenging task due to variation in illuminations, lighting conditions, view angles, etc.
- Person appearance feature extractor should be more background-agnostic and be stronger to distinguish small-scale detections.
- SCT needs further enhancement for occlusion scenarios to prevent ID switches.

Future Work

- Incorporating gait information into the pipeline to refine identity assignment.
- Improve appearance feature extractor.
- Employ geometry knowledge of the cameras for more accurate association.
- Experiment fusing human-object-interaction to help enhance intra-camera tracking.
- Expand this to multi-category object tracking and re-identification across cameras.

Gratitude

- Prof. Manjunath
- Prof. Hollerer
- VRL members
  - Raphael, Satish, Conner, Bowen, Iftekhar
  - All other lab members (Amil, Chandrakanth, Devendra, Joaquin, Umang)
- Prof. Majedi
- Family
- My friends and everyone who attended