GabriellaV2: UCF DIVA System

Center for Research in Computer Vision (CRCV) at the University of Central Florida (UCF)

Zacchaeus Scheffer, Ishan Dave, Mubarak Shah, Yogesh Rawat

December 8, 2021

Outline

Outline

- Problem Statement

Outline

- Problem Statement
- New in GabriellaV2

Outline

- Problem Statement
- New in GabriellaV2
  - Localization

Outline

- Problem Statement
- New in GabriellaV2
  - Localization
  - Action Classification

Outline

- Problem Statement
- New in GabriellaV2
  - Localization
  - Action Classification
  - Post Processing

Outline

- Problem Statement
- New in GabriellaV2
  - Localization
  - Action Classification
  - Post Processing
- Scores

Outline

- Problem Statement
- New in GabriellaV2
  - Localization
  - Action Classification
  - Post Processing
- Scores
- Qualitative Results

Outline

- Problem Statement
- New in GabriellaV2
  - Localization
  - Action Classification
  - Post Processing
- Scores
- Qualitative Results
- Questions

Problem Statement

Problem Statement

Untrimmed Video

Problem Statement

Untrimmed Video

- No special processing

Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"

Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"

Detect activities

Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"

Detect activities

- Human and vehicle

Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"

Detect activities

- Human and vehicle
- Indoor and outdoor

Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"

Detect activities

- Human and vehicle
- Indoor and outdoor

Types

Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"

Detect activities

- Human and vehicle
- Indoor and outdoor

Types

- Single-actor

Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"

Detect activities

- Human and vehicle
- Indoor and outdoor

Types

- Single-actor
- Multi-actor

Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"

Detect activities

- Human and vehicle
- Indoor and outdoor

Types

- Single-actor
- Multi-actor
- Actor-object

Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"

Detect activities

- Human and vehicle
- Indoor and outdoor

Types

- Single-actor
- Multi-actor
- Actor-object

Output

Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"

Detect activities

- Human and vehicle
- Indoor and outdoor

Types

- Single-actor
- Multi-actor
- Actor-object

Output

- Start/end times

Problem Statement

Untrimmed Video

- No special processing
- Should be comparable to "real-world"

Detect activities

- Human and vehicle
- Indoor and outdoor

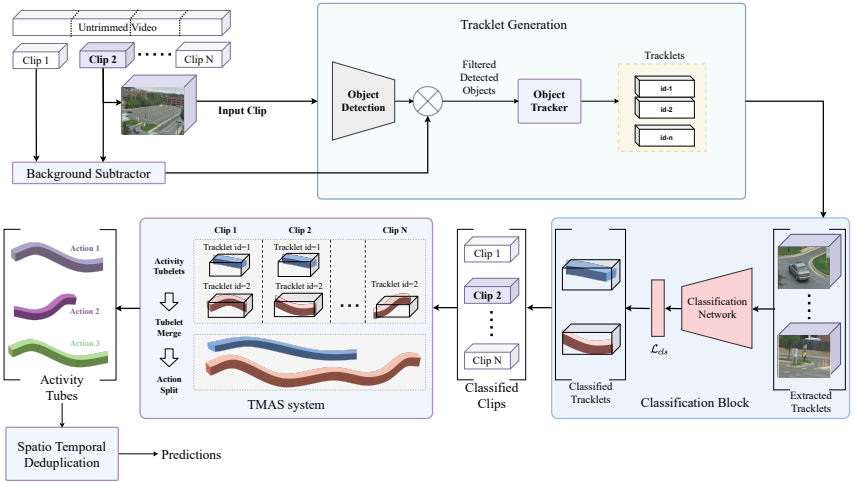
Types

- Single-actor
- Multi-actor
- Actor-object

Output

- Start/end times
- Spacial location

GabriellaV2 System



PreProcessing

PreProcessing

- First, we split the video using a sliding window approach

PreProcessing

- First, we split the video using a sliding window approach
  - 0-16

PreProcessing

- First, we split the video using a sliding window approach
  - 0-16
  - 8-24

PreProcessing

- First, we split the video using a sliding window approach
  - 0-16
  - 8-24
  - 16-32

PreProcessing

- First, we split the video using a sliding window approach
  - 0-16
  - 8-24
  - 16-32
- Every 4th frame used for tracklet generation

PreProcessing

- First, we split the video using a sliding window approach
  - 0-16
  - 8-24
  - 16-32
- Every 4th frame used for tracklet generation
- Full clip is used in tracklet extraction/classification

Tracklet Generation

Tracklet Generation

Tracklet Generation happens in 3 steps

Tracklet Generation

Tracklet Generation happens in 3 steps

- Object Detector (YOLOv5)

Tracklet Generation

Tracklet Generation happens in 3 steps

- Object Detector (YOLOv5)
  - spacially localizes objects

Tracklet Generation

Tracklet Generation happens in 3 steps

- Object Detector (YOLOv5)
  - spacially localizes objects
- Background Subtractor (MOG)

Tracklet Generation

Tracklet Generation happens in 3 steps

- Object Detector (YOLOv5)
  - spacially localizes objects
- Background Subtractor (MOG)
  - Removes Stationary Objects

Tracklet Generation

Tracklet Generation happens in 3 steps

- Object Detector (YOLOv5)
  - spacially localizes objects
- Background Subtractor (MOG)
  - Removes Stationary Objects
- Object Tracker (SORT)

Tracklet Generation

Tracklet Generation happens in 3 steps

- Object Detector (YOLOv5)
  - spacially localizes objects
- Background Subtractor (MOG)
  - Removes Stationary Objects
- Object Tracker (SORT)
  - groups detections of the same object

Object Detection

For each Clip

Object Detection

For each Clip

- send 4 frames to Object Detector

Object Detection

For each Clip

- send 4 frames to Object Detector
- Place bounding box around potential actor

Object Detection

For each Clip

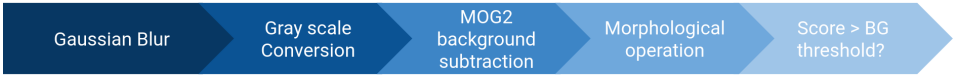
- send 4 frames to Object Detector
- Place bounding box around potential actor
  - person

Object Detection

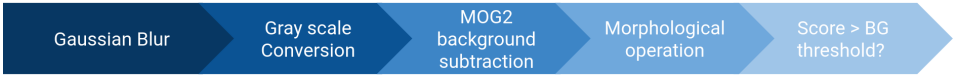
For each Clip

- send 4 frames to Object Detector
- Place bounding box around potential actor
  - person
  - vehicle

Background Subtraction

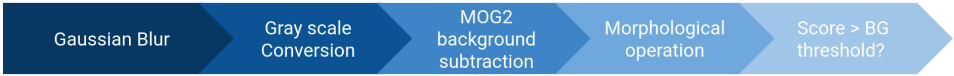


Background Subtraction



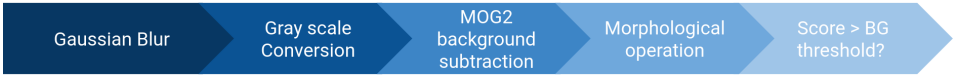
- We use the MOG2 background subtractor

Background Subtraction



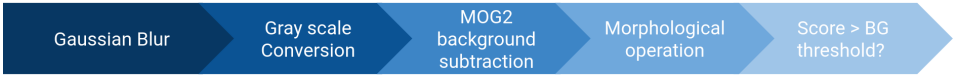
- We use the MOG2 background subtractor
- simple to use

Background Subtraction



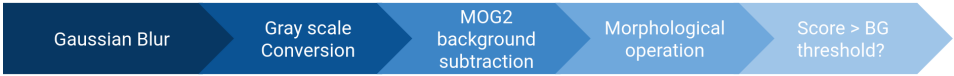
- We use the MOG2 background subtractor
- simple to use
- drastically reduce localization false alarm

Background Subtraction



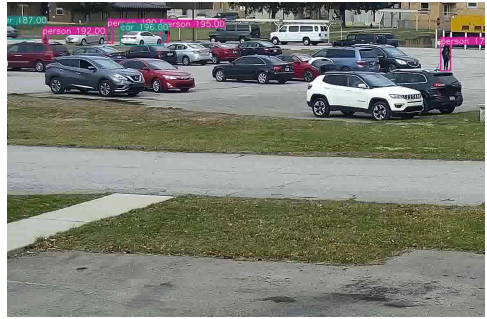
- We use the MOG2 background subtractor
- simple to use
- drastically reduce localization false alarm
  - 40% FA reduction

Background Subtraction



- We use the MOG2 background subtractor
- simple to use
- drastically reduce localization false alarm
  - 40% FA reduction
  - Runtime reduced by half

Background Subtractor Results



Object Tracker

Object Tracker

- We use the SORT tracker

Object Tracker

- We use the SORT tracker
- IOU-based matching

Object Tracker

- We use the SORT tracker
- IOU-based matching
- Same object gets same ID in subsequent frames

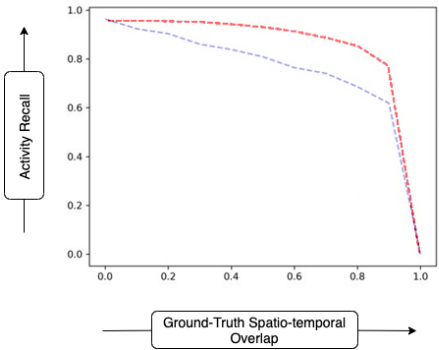
Object Tracker

- We use the SORT tracker
- IOU-based matching
- Same object gets same ID in subsequent frames
- Can "remember" 2 frames prior

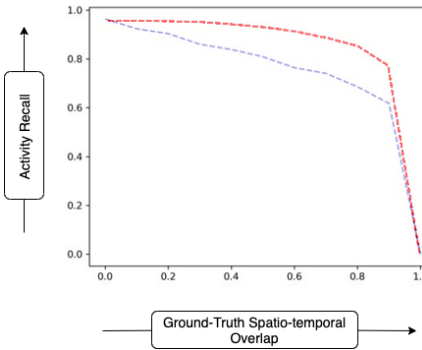
Object Tracker

- We use the SORT tracker
- IOU-based matching
- Same object gets same ID in subsequent frames
- Can "remember" 2 frames prior
  - Track continues if one bounding box is missing

Recall Analysis of Tracklet Generation

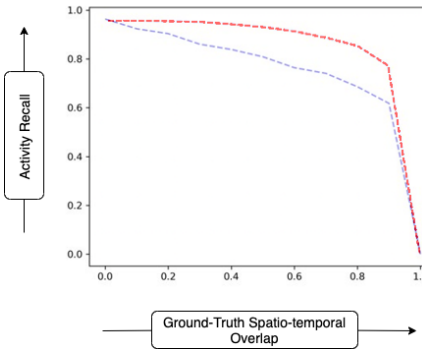


Recall Analysis of Tracklet Generation



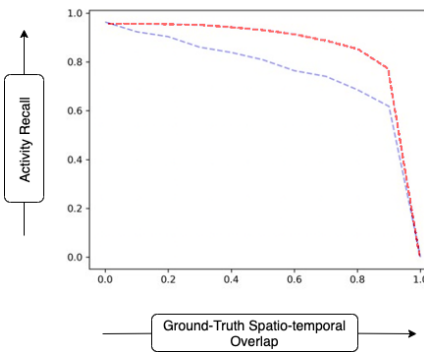
- Legacy method used action localization instead of object detection/tracking

Recall Analysis of Tracklet Generation



- Legacy method used action localization instead of object detection/tracking
- New method attains higher activity recall than legacy method

Recall Analysis of Tracklet Generation



- Legacy method used action localization instead of object detection/tracking
- New method attains higher activity recall than legacy method
  - 22% higher recall using 0.8 IOU threshold

Tracklet Extraction

Tracklet Extraction

For each object ID produced by the SORT tracker in a given clip:

Tracklet Extraction

For each object ID produced by the SORT tracker in a given clip:

- collect all bounding boxes with that object ID

Tracklet Extraction

For each object ID produced by the SORT tracker in a given clip:

- collect all bounding boxes with that object ID
- Combine them by taking the smallest-bounding bounding box

Tracklet Extraction

For each object ID produced by the SORT tracker in a given clip:

- collect all bounding boxes with that object ID
- Combine them by taking the smallest-bounding bounding box
- Extend the resulting bounding box in the shorter dimension to obtain a square

Tracklet Extraction

For each object ID produced by the SORT tracker in a given clip:

- collect all bounding boxes with that object ID
- Combine them by taking the smallest-bounding bounding box
- Extend the resulting bounding box in the shorter dimension to obtain a square
  - classifier gets consistent aspect ratio

Tracklet Extraction

For each object ID produced by the SORT tracker in a given clip:

- collect all bounding boxes with that object ID
- Combine them by taking the smallest-bounding bounding box
- Extend the resulting bounding box in the shorter dimension to obtain a square
  - classifier gets consistent aspect ratio
- Take spatial crop of clip. <- Extracted Tracklet

Action Classification

Action Classification

Surveillance video activities are multi-label

Action Classification

Surveillance video activities are multi-label

Ground truth tracks to train 3D-CNN backbone + sigmoid activation with standard BCE loss

Action Classification

Surveillance video activities are multi-label
Ground truth tracks to train 3D-CNN backbone + sigmoid activation with standard BCE loss

$$\mathcal{L}_{\text{BCE}}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) - (1 - y_i)\log(1 - \hat{y}_i)]$$

Action Classification

Surveillance video activities are multi-label

Ground truth tracks to train 3D-CNN backbone + sigmoid activation with standard BCE loss

$$\mathcal{L}_{\text{BCE}}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) - (1 - y_i)\log(1 - \hat{y}_i)]$$

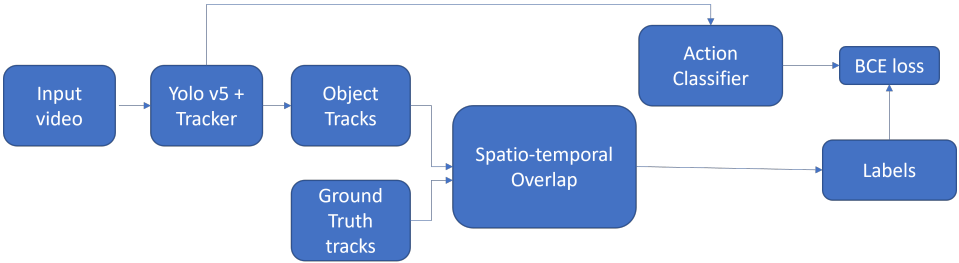
Each action class is independent of each other

Adapting to Yolo Tracklets

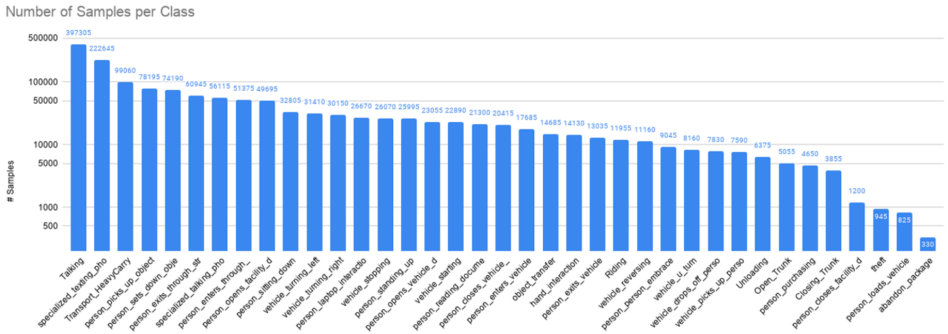
Need to train classifier to take YOLO tracklets

Adapting to Yolo Tracklets

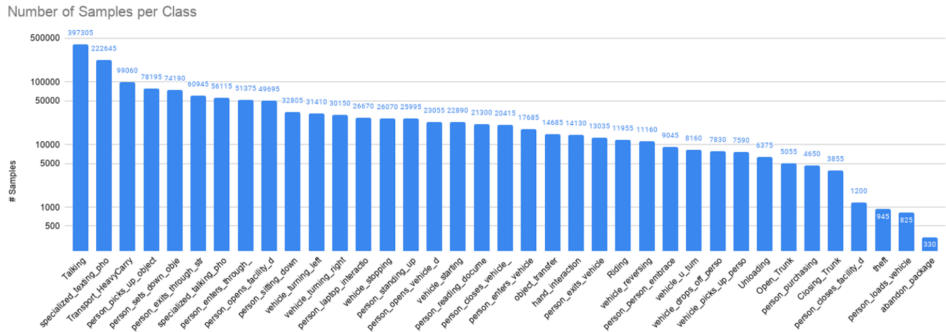
Need to train classifier to take YOLO tracklets



Class Balanced Training

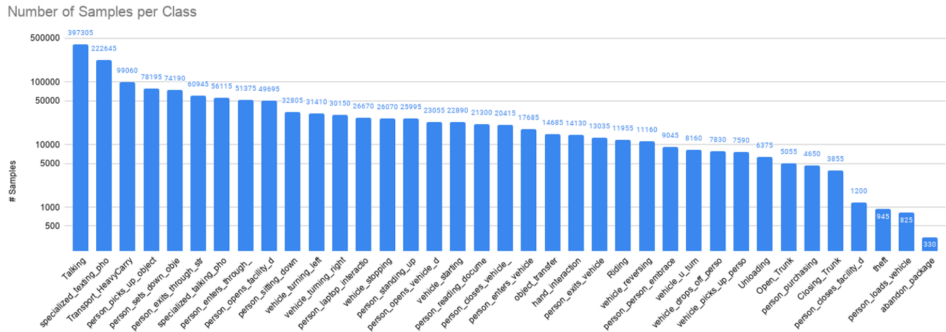


Class Balanced Training



Some classes are very frequent, some very rare

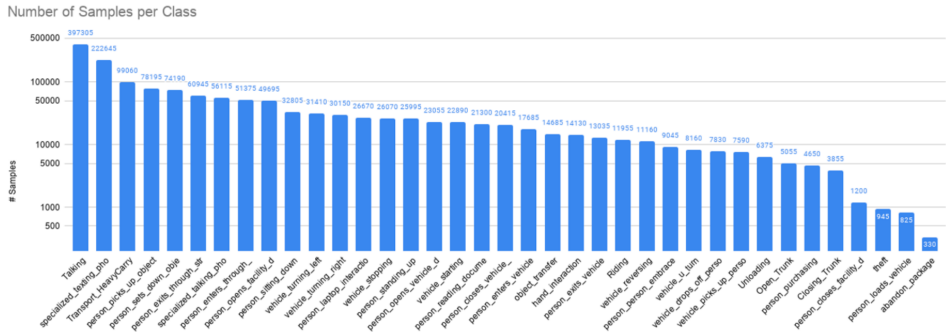
Class Balanced Training



Some classes are very frequent, some very rare

In the most extreme, 1:1000 difference

Class Balanced Training



Some classes are very frequent, some very rare

In the most extreme, 1:1000 difference

Used Partial Label Masking (PLM) for class imbalance

Learning Multi-Label Class Correlations

Learning Multi-Label Class Correlations

Log-Sum-Exponential Pairwise (LSEP) loss for multilabel action recognition.

Learning Multi-Label Class Correlations

Log-Sum-Exponential Pairwise (LSEP) loss for multilabel action recognition.

$$\mathcal{L}_{\text{LSEP}} = \log\left(1 + \sum_{i \in y} \sum_{j \notin y} e^{x_j - x_i}\right)$$

Learning Multi-Label Class Correlations

Log-Sum-Exponential Pairwise (LSEP) loss for multilabel action recognition.

$$\mathcal{L}_{\text{LSEP}} = \log\left(1 + \sum_{i \in y} \sum_{j \notin y} e^{x_j - x_i}\right)$$

Standard LSEP is based on BCE loss

Learning Multi-Label Class Correlations

Log-Sum-Exponential Pairwise (LSEP) loss for multilabel action recognition.

$$\mathcal{L}_{\text{LSEP}} = \log\left(1 + \sum_{i \in y} \sum_{j \notin y} e^{x_j - x_i}\right)$$

Standard LSEP is based on BCE loss

Use LSEP loss with PLM to reweight samples to class balance and learn class correlations together

Two level Knowledge Distillation

We perform Knowledge Distillation in two stages:

Two level Knowledge Distillation

We perform Knowledge Distillation in two stages:

- standard phase

Two level Knowledge Distillation

We perform Knowledge Distillation in two stages:

- standard phase
  - reduce ensemble size

Two level Knowledge Distillation

We perform Knowledge Distillation in two stages:

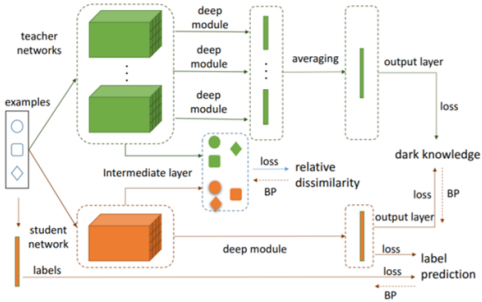
- standard phase
  - reduce ensemble size
- compression phase

Two level Knowledge Distillation

We perform Knowledge Distillation in two stages:

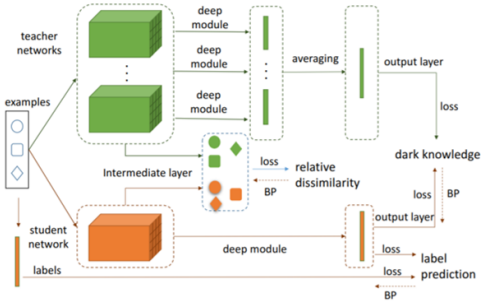
- standard phase
  - reduce ensemble size
- compression phase
  - reduce model size

Knowledge Distillation (Standard)

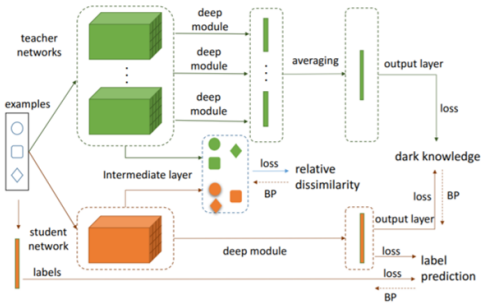


Knowledge Distillation (Standard)

- Train student model with ensemble

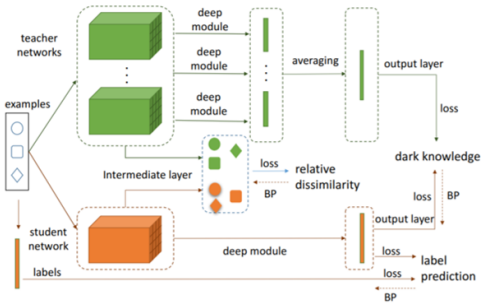


Knowledge Distillation (Standard)



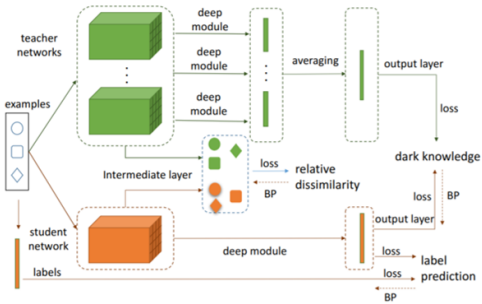
- Train student model with ensemble
- loss based on hidden layers

Knowledge Distillation (Standard)



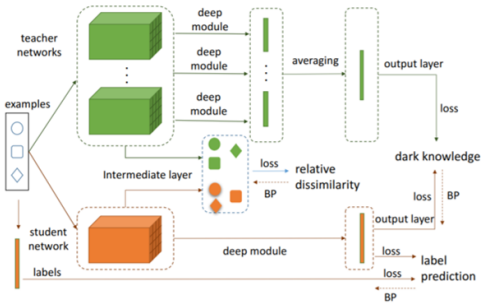
- Train student model with ensemble
- loss based on hidden layers
- loss on raw outputs

Knowledge Distillation (Standard)



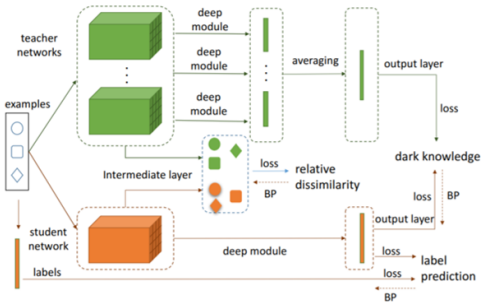
- Train student model with ensemble
- loss based on hidden layers
- loss on raw outputs
- loss on label prediction

Knowledge Distillation (Standard)



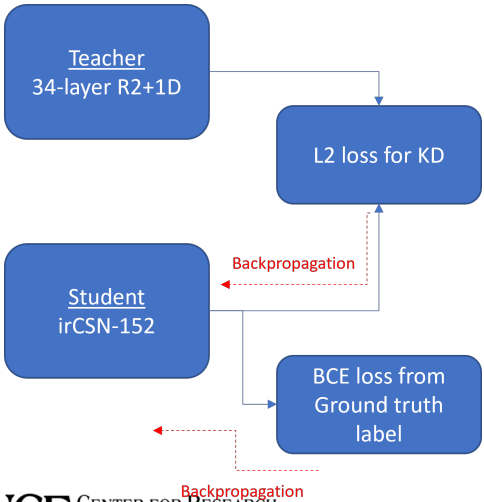
- Train student model with ensemble
- loss based on hidden layers
- loss on raw outputs
- loss on label prediction
- val set mAP >4% improvment

Knowledge Distillation (Standard)

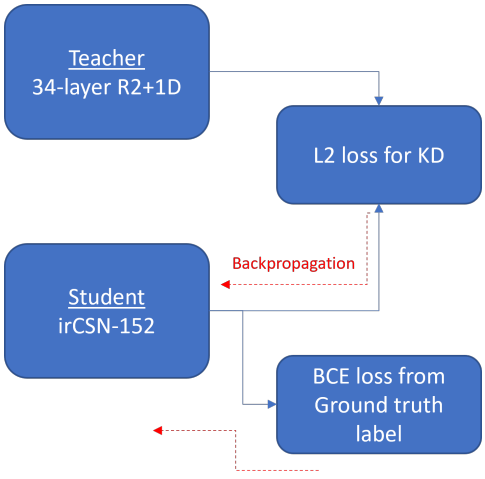


- Train student model with ensemble
- loss based on hidden layers
- loss on raw outputs
- loss on label prediction
- val set mAP >4% improvment
  - same mAP as fulla ensemble

Knowledge Distillation (Model Compression)

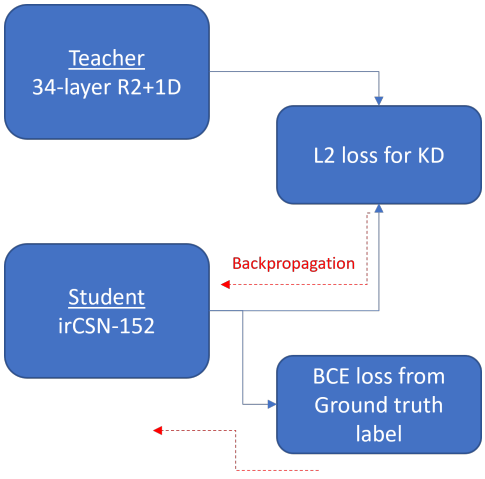


Knowledge Distillation (Model Compression)



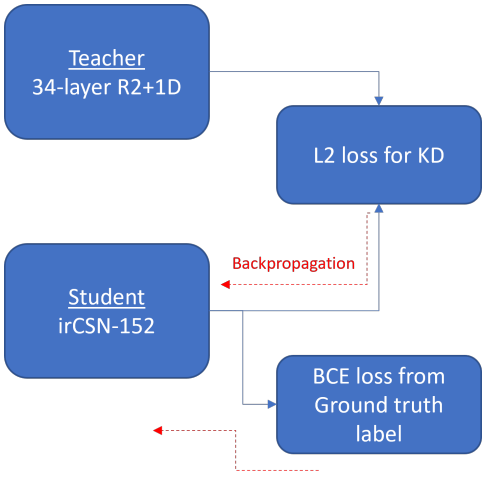
- Student model from before as Teacher

Knowledge Distillation (Model Compression)



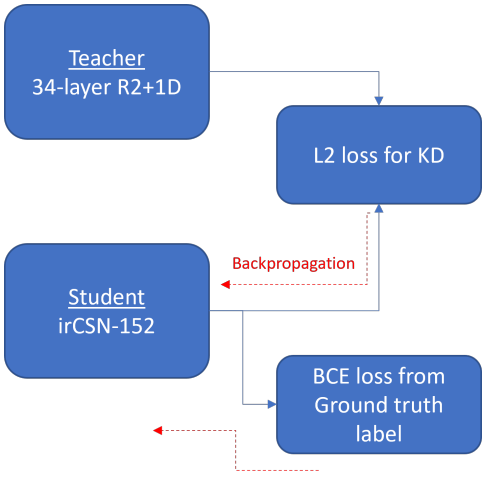
- Student model from before as Teacher
- New student is slimmer

Knowledge Distillation (Model Compression)



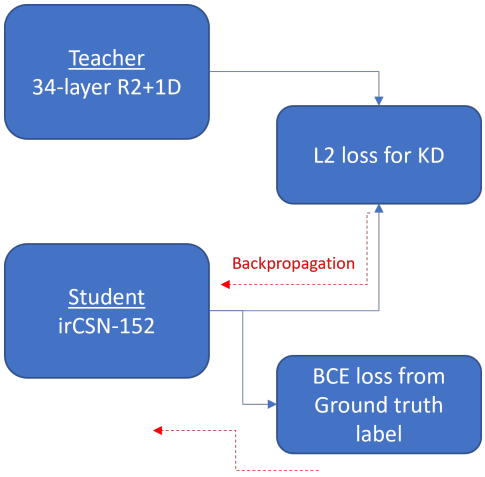
- Student model from before as Teacher
- New student is slimmer
- loss on raw outputs

Knowledge Distillation (Model Compression)



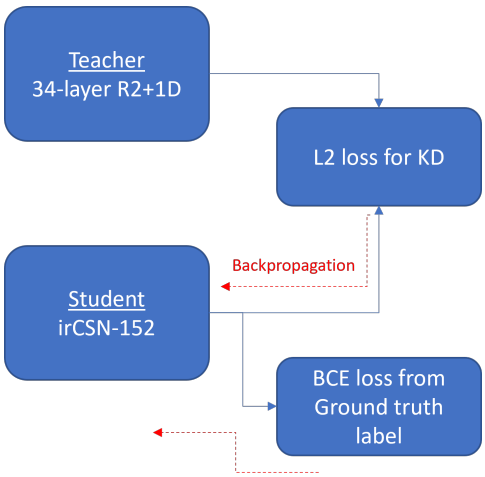
- Student model from before as Teacher
- New student is slimmer
- loss on raw outputs
- loss on label prediciton

Knowledge Distillation (Model Compression)



- Student model from before as Teacher
- New student is slimmer
- loss on raw outputs
- loss on label prediciton
- val set mAP >7% improvement

Knowledge Distillation (Model Compression)



- Student model from before as Teacher
- New student is slimmer
- loss on raw outputs
- loss on label prediciton
- val set mAP >7% improvement
- same mAP as full ensemble

Post-Processing

Post-Processing has two steps:

Post-Processing

Post-Processing has two steps:

- TMAS

Post-Processing

Post-Processing has two steps:

- TMAS
  - smooth detection temporally

Post-Processing

Post-Processing has two steps:

- TMAS
  - smooth detection temporally
  - convert actor tracks into action tubes

Post-Processing

Post-Processing has two steps:

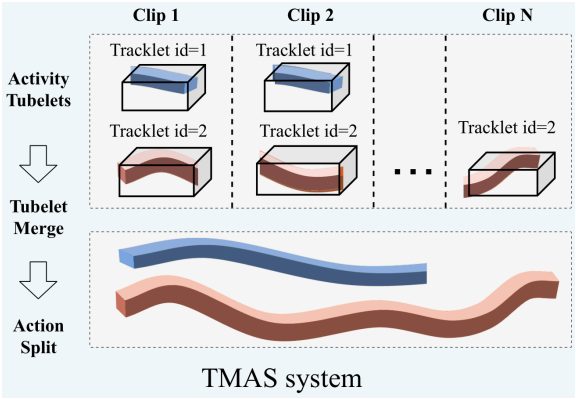
- TMAS
  - smooth detection temporally
  - convert actor tracks into action tubes
- NMS

Post-Processing

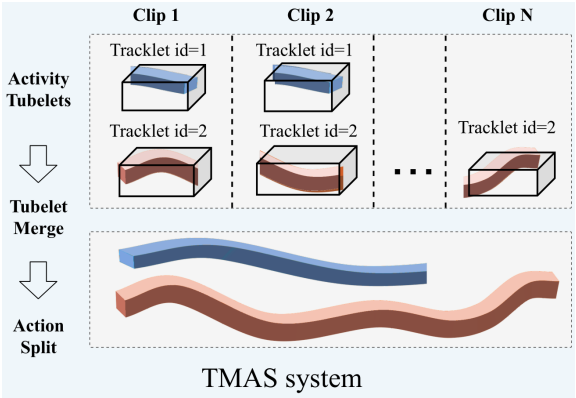
Post-Processing has two steps:

- TMAS
  - smooth detection temporally
  - convert actor tracks into action tubes
- NMS
  - Remove duplicate predictions

TMAS

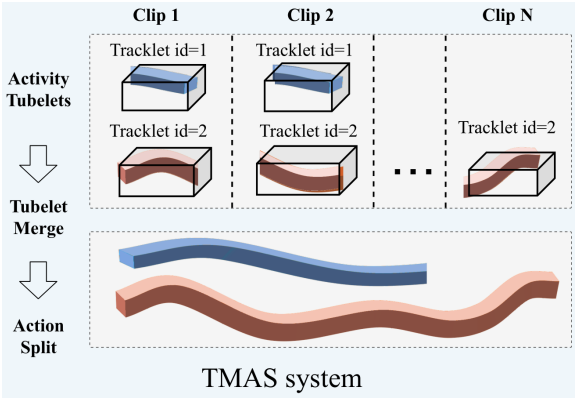

Clip 1
Clip 2
Clip N
Activity Tubelets
Tracklet id=1
Tracklet id=1
Tracklet id=2
Tracklet id=2
Tracklet id=2
Tubelet Merge
Action Split
TMAS system

TMAS



Chain tracklets with same ID into one actor track

TMAS

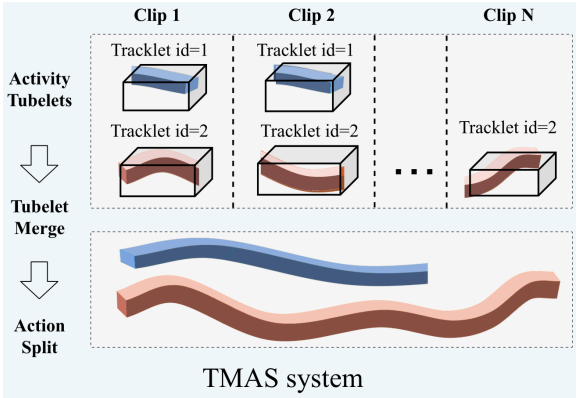


TMAS system

Chain tracklets with same ID into one actor track

Smooth classwise detection temporally

TMAS



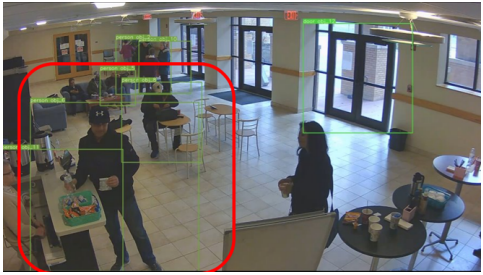
TMAS system

Chain tracklets with same ID into one actor track

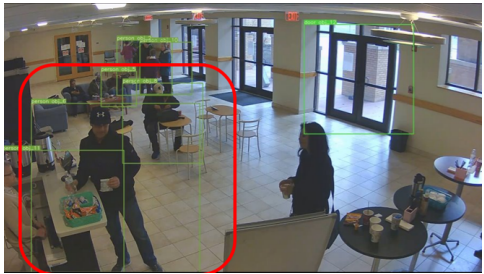
Smooth classwise detection temporally

Create action tubes for each class using (connected) regions of actor tracks where class-scores is high

NMS

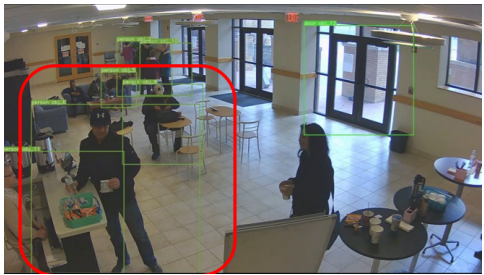


NMS



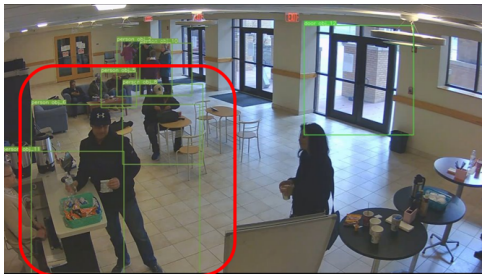
- We square our object detections

NMS



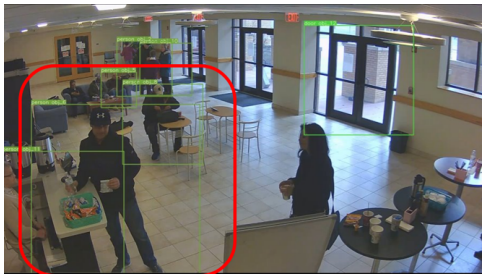
- We square our object detections
- Very large overlaps

NMS



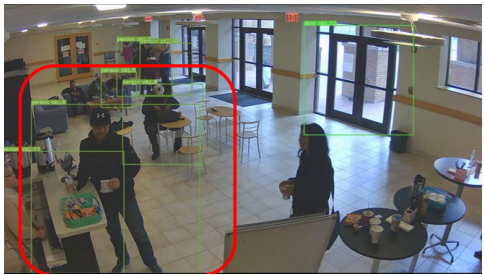
- We square our object detections
- Very large overlaps
- Actions contained in multiple bounding boxes

NMS



- We square our object detections
- Very large overlaps
- Actions contained in multiple bounding boxes
- multi-actor actions

NMS



- We square our object detections
- Very large overlaps
- Actions contained in multiple bounding boxes
- multi-actor actions
- Need to remove duplicate predictions

NMS

Need method to work regardless of actor size

NMS

Need method to work regardless of actor size
use dimensionless distance metric

NMS

Need method to work regardless of actor size
use dimensionless distance metric

$$d(A,B) = \frac{d_e(A,B)}{\sqrt[4]{AreaA * Area(B)}}$$

NMS

Need method to work regardless of actor size
use dimensionless distance metric

$$d(A,B) = \frac{d_e(A,B)}{\sqrt[4]{AreaA * Area(B)}}$$

Where d_e is the euclidian distance.

NMS

Need method to work regardless of actor size
use dimensionless distance metric

$$d(A,B) = \frac{d_e(A,B)}{\sqrt[4]{AreaA * Area(B)}}$$

Where d_e is the euclidian distance. If actors are close, treat as same action, if VERY close, same actor

NMS

Need method to work regardless of actor size
use dimensionless distance metric

$$d(A,B) = \frac{d_e(A,B)}{\sqrt[4]{AreaA * Area(B)}}$$

Where d_e is the euclidian distance. If actors are close, treat as same action, if VERY close, same actor
False Negatives very rare

TrecVID21 Results

Rank	team_name	team_abbrev	nAUDC@0.2tfa	p_miss@0.15tfa
1	BUPT-MCPRL	BUPT-MC_26542	0.4085	0.3249
2	UCF	UCF_26546	0.4306	**0.3408**
3	INF	INF_26532	0.4444	0.3508
4	M4D_2021	M4D_202_26467	0.8466	0.7941
5	TokyoTech_AIST	TOKYOTE_26508	0.8516	0.8197
6	Team UEC	TEAMUE_26530	0.964	0.9503

TrecVID21 Results

Rank	team_name	team_abbrev	nAUDC@0.2tfa	p_miss@0.15tfa
1	BUPT-MCPRL	BUPT-MC_26542	0.4085	0.3249
2	UCF	UCF_26546	0.4306	**0.3408**
3	INF	INF_26532	0.4444	0.3508
4	M4D_2021	M4D_202_26467	0.8466	0.7941
5	TokyoTech_AIST	TOKYOTE_26508	0.8516	0.8197
6	Team UEC	TEAMUE_26530	0.964	0.9503

- based on VIRAT dataset

TrecVID21 Results

Rank	team_name	team_abbrev	nAUDC@0.2tfa	p_miss@0.15tfa
1	BUPT-MCPRL	BUPT-MC_26542	0.4085	0.3249
2	UCF	UCF_26546	0.4306	**0.3408**
3	INF	INF_26532	0.4444	0.3508
4	M4D_2021	M4D_202_26467	0.8466	0.7941
5	TokyoTech_AIST	TOKYOTE_26508	0.8516	0.8197
6	Team UEC	TEAMUE_26530	0.964	0.9503

- based on VIRAT dataset
- has known cameras

TrecVID21 Results

Rank	team_name	team_abbrev	nAUDC@0.2tfa	p_miss@0.15tfa
1	BUPT-MCPRL	BUPT-MC_26542	0.4085	0.3249
2	UCF	UCF_26546	0.4306	**0.3408**
3	INF	INF_26532	0.4444	0.3508
4	M4D_2021	M4D_202_26467	0.8466	0.7941
5	TokyoTech_AIST	TOKYOTE_26508	0.8516	0.8197
6	Team UEC	TEAMUE_26530	0.964	0.9503

- based on VIRAT dataset
- has known cameras
- 2nd place

SDL21 Results

Rank	Team Name	sub_id	mean p_miss@0.01tfa	mean p_miss@0.02tfa	mean nAUDC@0.2tfa	relative_processing_time
1	UCF	25908	**0.62**	**0.5372**	0.3518	0.684
2	CMU-DIVA	26095	0.65	0.5438	0.333	0.776
3	IBM-Purdue	26113	0.65	0.5531	0.3533	0.575
4	UMD	26619	0.68	0.5938	0.3898	0.515
5	UMD-Columbia	25031	0.68	0.5975	0.4002	0.52
6	UMCMU	25576	0.75	0.6861	0.4922	0.614
7	Purdue	25782	0.8	0.7294	0.4942	0.239
8	MINDS_JHU	24666	0.84	0.7791	0.6343	0.898

SDL21 Results

Rank	Team Name	sub_id	mean p_miss@0.01tfa	mean p_miss@0.02tfa	mean nAUDC@0.2tfa	relative_processing_time
1	UCF	25908	**0.62**	**0.5372**	0.3518	0.684
2	CMU-DIVA	26095	0.65	0.5438	0.333	0.776
3	IBM-Purdue	26113	0.65	0.5531	0.3533	0.575
4	UMD	26619	0.68	0.5938	0.3898	0.515
5	UMD-Columbia	25031	0.68	0.5975	0.4002	0.52
6	UMCMU	25576	0.75	0.6861	0.4922	0.614
7	Purdue	25782	0.8	0.7294	0.4942	0.239
8	MINDS_JHU	24666	0.84	0.7791	0.6343	0.898

- based on MEVA dataset

SDL21 Results

Rank	Team Name	sub_id	mean p_miss@0.01tfa	mean p_miss@0.02tfa	mean nAUDC@0.2tfa	relative_processing_time
1	UCF	25908	**0.62**	**0.5372**	0.3518	0.684
2	CMU-DIVA	26095	0.65	0.5438	0.333	0.776
3	IBM-Purdue	26113	0.65	0.5531	0.3533	0.575
4	UMD	26619	0.68	0.5938	0.3898	0.515
5	UMD-Columbia	25031	0.68	0.5975	0.4002	0.52
6	UMCMU	25576	0.75	0.6861	0.4922	0.614
7	Purdue	25782	0.8	0.7294	0.4942	0.239
8	MINDS_JHU	24666	0.84	0.7791	0.6343	0.898

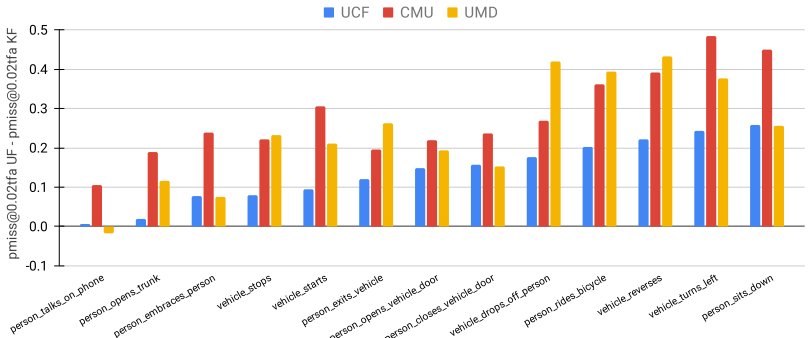
- based on MEVA dataset
- has unknown cameras

SDL21 Results

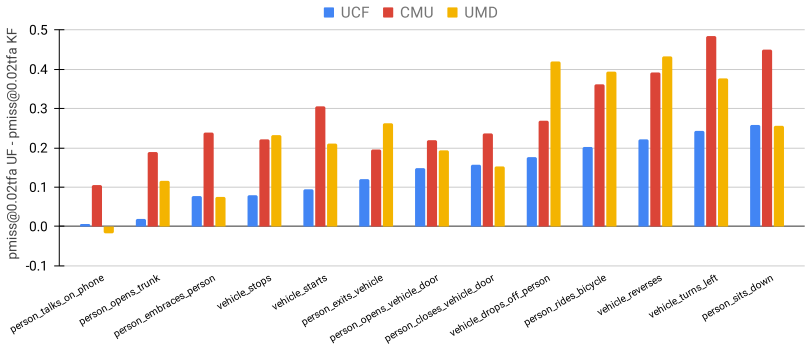
Rank	Team Name	sub_id	mean p_miss@0.01tfa	mean p_miss@0.02tfa	mean nAUDC@0.2tfa	relative_processing_time
1	UCF	25908	**0.62**	**0.5372**	0.3518	0.684
2	CMU-DIVA	26095	0.65	0.5438	0.333	0.776
3	IBM-Purdue	26113	0.65	0.5531	0.3533	0.575
4	UMD	26619	0.68	0.5938	0.3898	0.515
5	UMD-Columbia	25031	0.68	0.5975	0.4002	0.52
6	UMCMU	25576	0.75	0.6861	0.4922	0.614
7	Purdue	25782	0.8	0.7294	0.4942	0.239
8	MINDS_JHU	24666	0.84	0.7791	0.6343	0.898

- based on MEVA dataset
- has unknown cameras
- 1st place in UF

Generalization

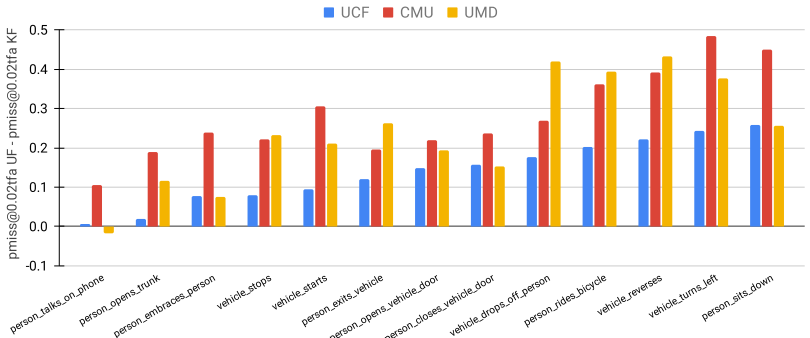


Generalization



- Drop in performance going from Known Facility to Unknown

Generalization



- Drop in performance going from Known Facility to Unknown
- Our system gets best generalization

Qualitative Results

Many common situations that hurt performance

Qualitative Results

Many common situations that hurt performance

- Distant actors

Qualitative Results

Many common situations that hurt performance

- Distant actors
- Actor changes distance over time

Qualitative Results

Many common situations that hurt performance

- Distant actors
- Actor changes distance over time
- Temporal variability

Qualitative Results

Many common situations that hurt performance

- Distant actors
- Actor changes distance over time
- Temporal variability
- Crowded Scenes

Conclusion

GabriellaV2 is a real-time action detection system which can generalize very well to the unknown facility cameras.

Conclusion

GabriellaV2 is a real-time action detection system which can generalize very well to the unknown facility cameras.Built upon our Gabriella system by

Conclusion

GabriellaV2 is a real-time action detection system which can generalize very well to the unknown facility cameras.Built upon our Gabriella system by

- Replacing localization by tracklet generation,

Conclusion

GabriellaV2 is a real-time action detection system which can generalize very well to the unknown facility cameras.Built upon our Gabriella system by

- Replacing localization by tracklet generation,
- Better classifiers: class imbalance, multi-label class correlation, Knowledge distillation

Conclusion

GabriellaV2 is a real-time action detection system which can generalize very well to the unknown facility cameras.Built upon our Gabriella system by

- Replacing localization by tracklet generation,
- Better classifiers: class imbalance, multi-label class correlation, Knowledge distillation
- Improved Post processing using Spatio-temporal deduplication

Conclusion

GabriellaV2 is a real-time action detection system which can generalize very well to the unknown facility cameras.Built upon our Gabriella system by

- Replacing localization by tracklet generation,
- Better classifiers: class imbalance, multi-label class correlation, Knowledge distillation
- Improved Post processing using Spatio-temporal deduplication
- Top place for ActEV-SDL21 UF leaderboard and Runners-up for TRECVID21

References

Monfort, Mathew, et al. "Multi-moments in time: Learning and interpreting models for multi-action video understanding." arXiv preprint arXiv:1911.00232 (2019).
You, Shan, et al. "Learning from multiple teacher networks." Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017.
Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).
Duarte, Kevin, Yogesh Rawat, and Mubarak Shah. "PLM: Partial Label Masking for Imbalanced Multi-label Classification." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.