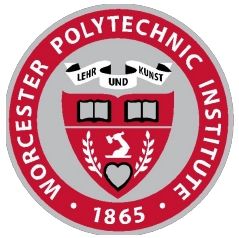


WPI

Learning Deep Social Interactions to Identify Positive Classroom Climate

Przemek Gardias

Classroom Observations

"Morning Circle at Preschool" public listing: https://www.youtube.com/watch?v=PZY-hB2C_Iw

Classroom Observation Protocols

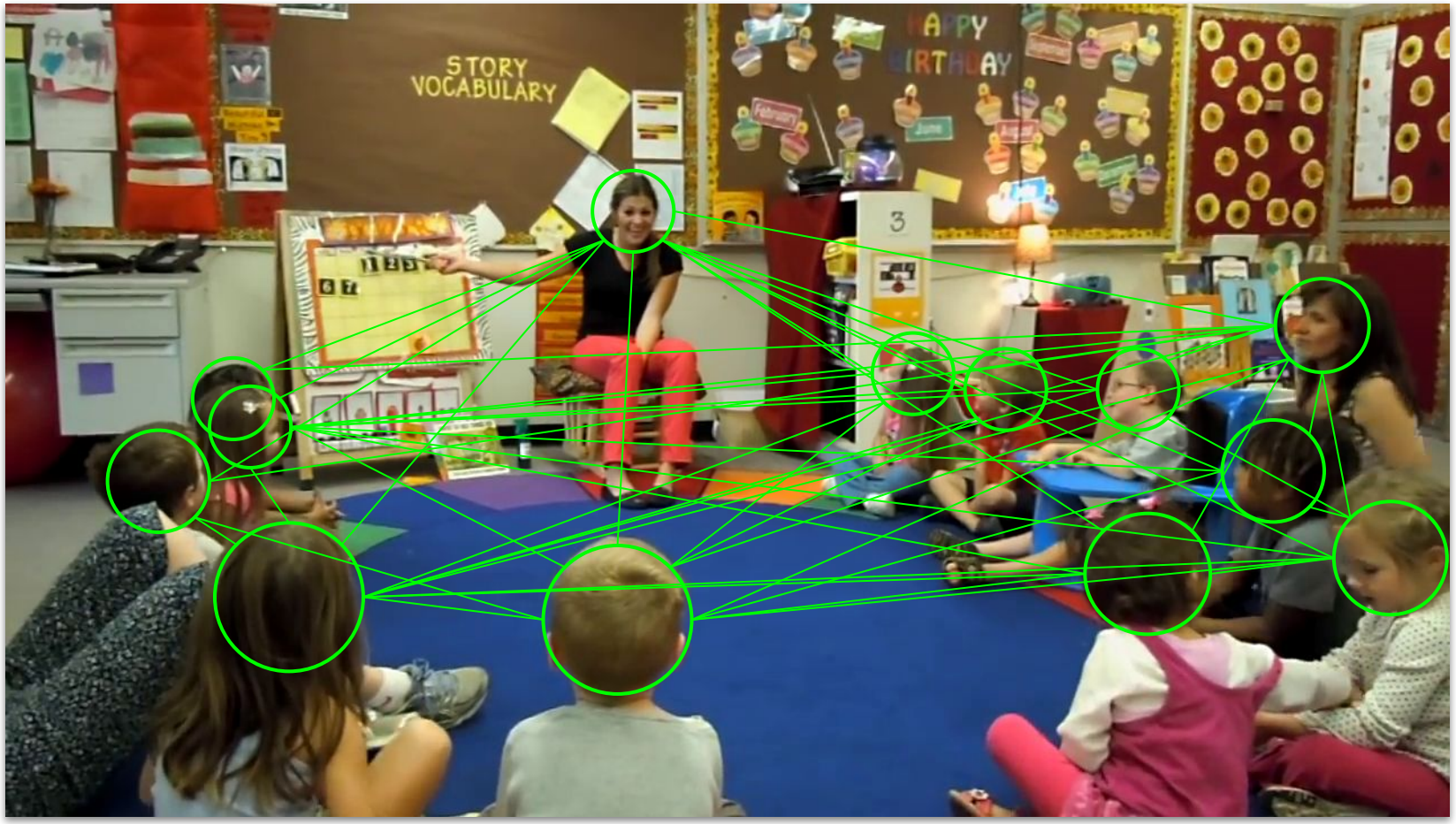
- Classroom Assessment Scoring System (CLASS) (Pianta et al., 2008)
  - Measure quality of teacher-student interactions in PK-12 classrooms
  - Coded manually by trained CLASS observers
  - Ten dimensions of classroom quality

Positive climate — "warmth, respect, and enjoyment communicated by verbal and nonverbal interactions"

Classroom Climate

"Morning Circle at Preschool" public listing: https://www.youtube.com/watch?v=PZY-hB2C_Iw

Modeling a Classroom as a Graph



"Morning Circle at Preschool" public listing: https://www.youtube.com/watch?v=PZY-hB2C_Iw

ACORN (Ramakrishnan et al., 2019-2021)

- Ensemble perceptual + auditory deep networks to estimate climate
- Success identifying particular moments in classroom sessions with graph convolution networks
  - Ignores who is where

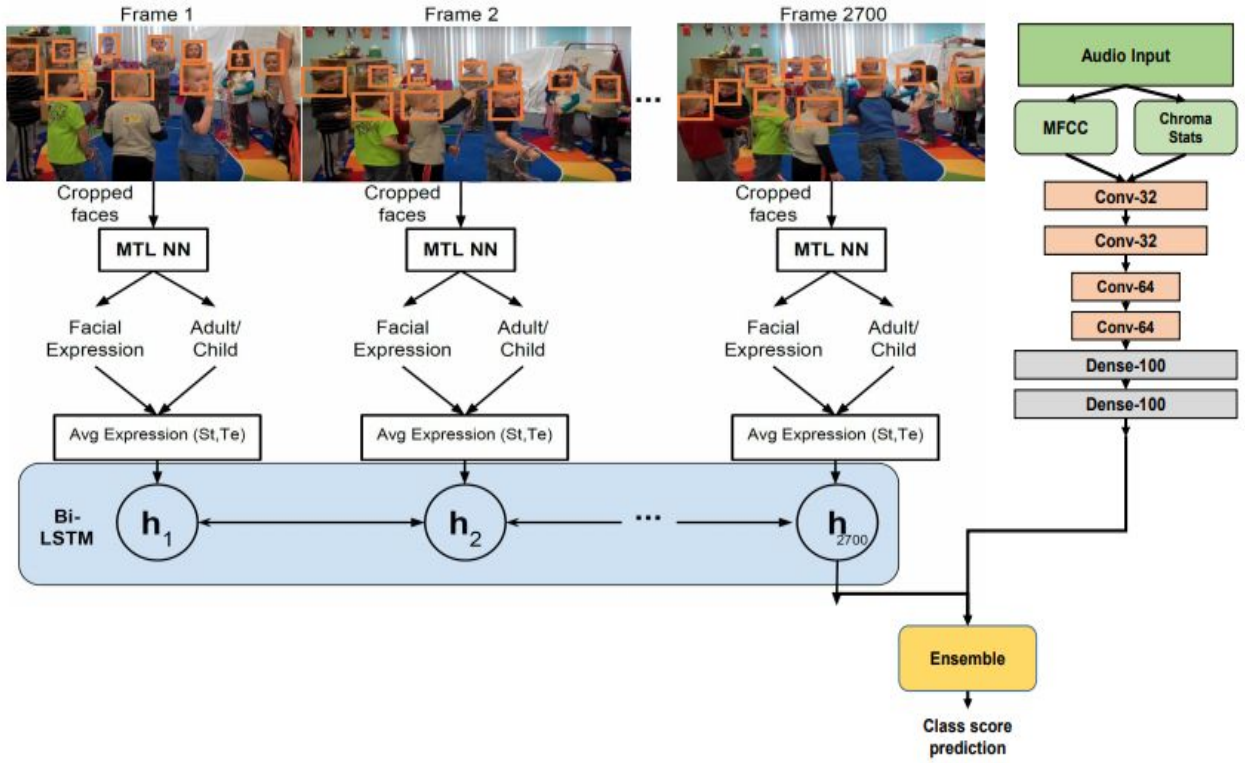
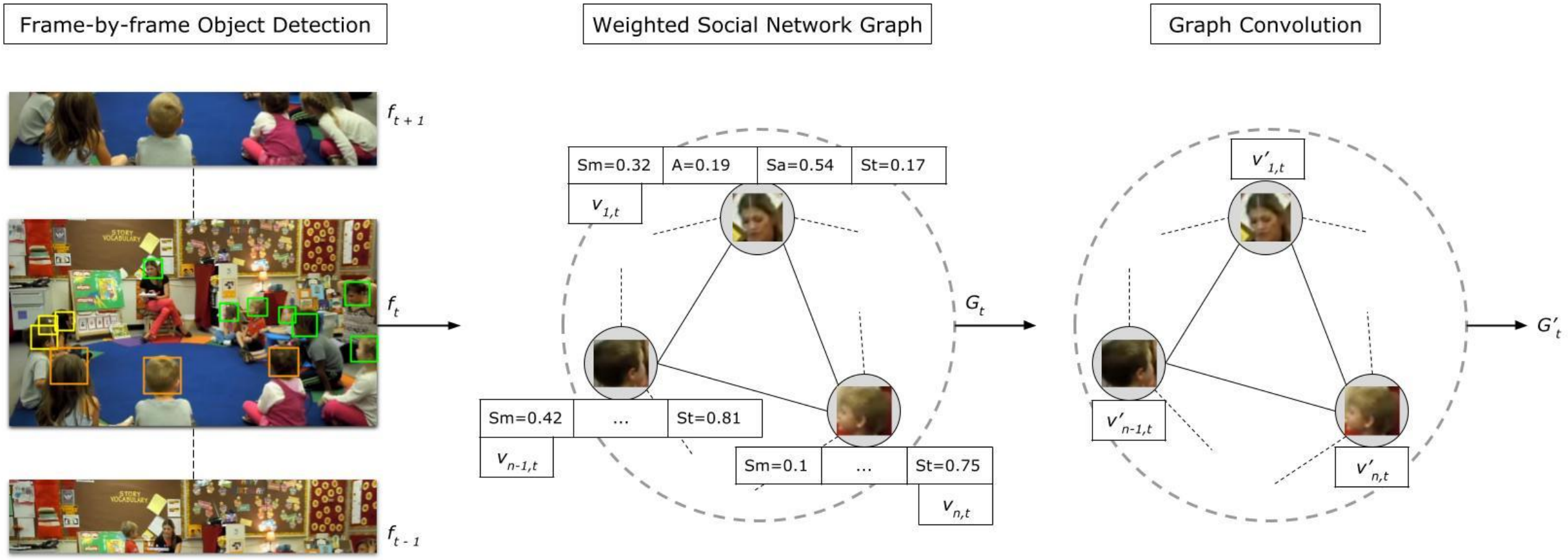


Fig. 1: Ensemble of visual and audio-based models to predict CLASS climate. (*Ramakrishnan, et al., 2019*)

Dynamic Social Network Graph



Graph Convolution Networks (Kipf & Welling, 2017)

- Aggregate neighboring features with weights dependent on adjacency matrix

$$H_t^{(0)} = V_t$$

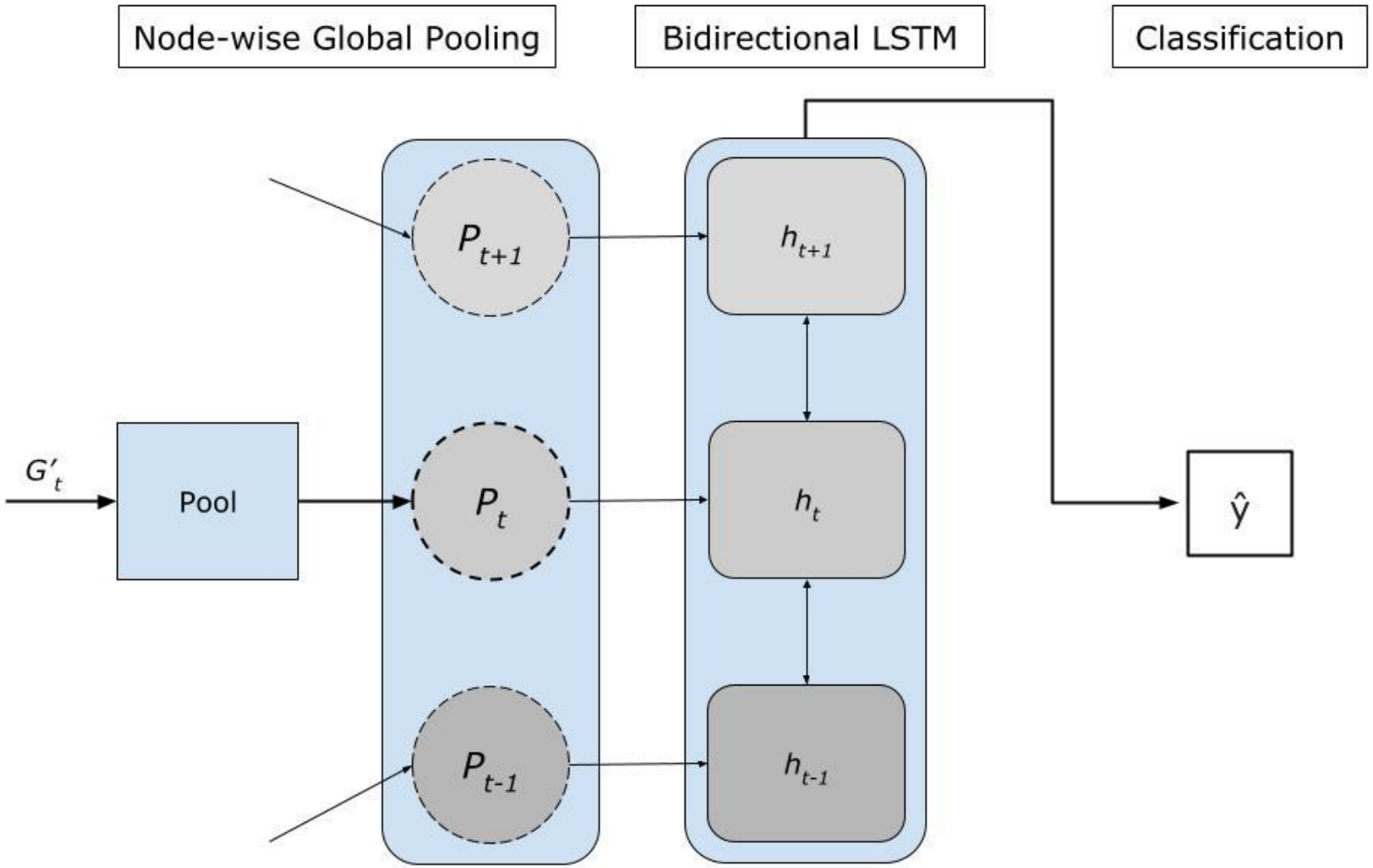
$$H_t^{(l+1)} = \sigma(L_t H_t^{(l)} W^{(l)})$$

Symmetric, normalized Laplacian

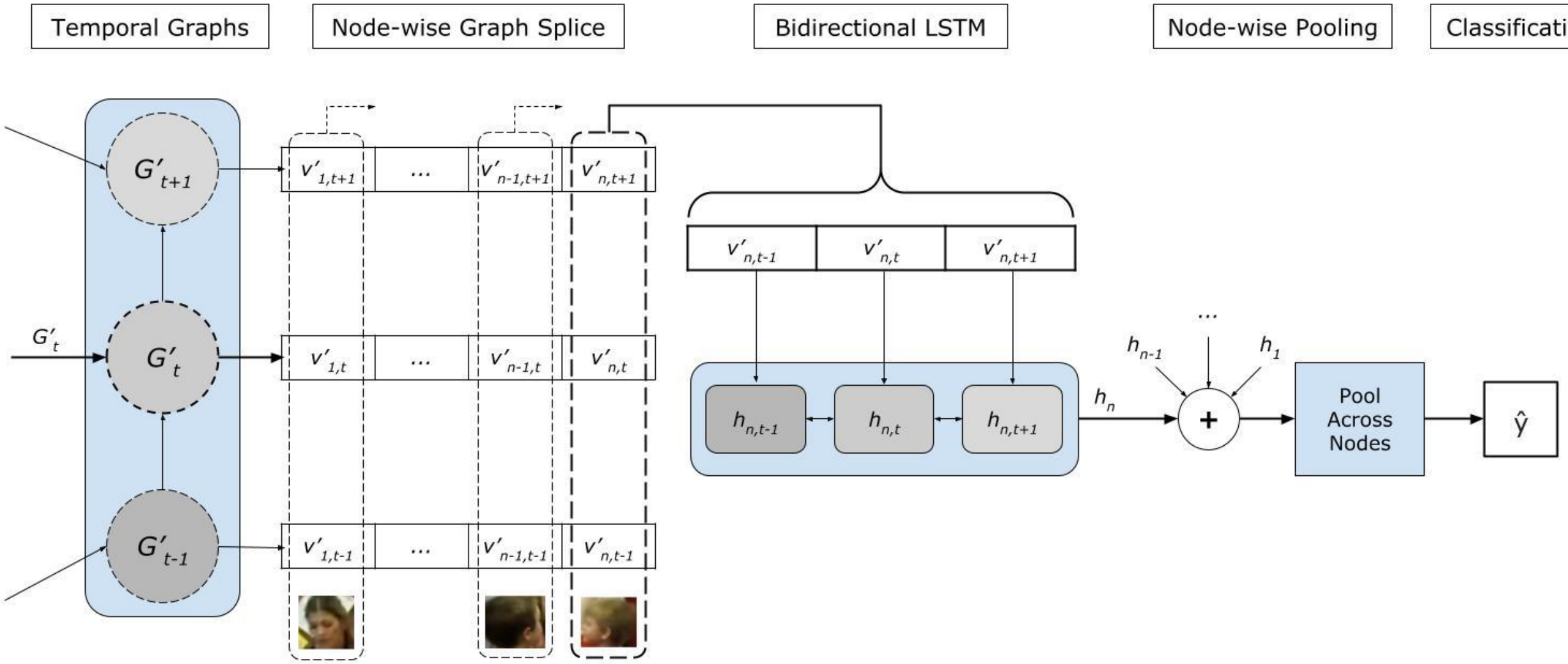
Convolution weights

Tracking vs Non-tracking

Non-tracking GNN



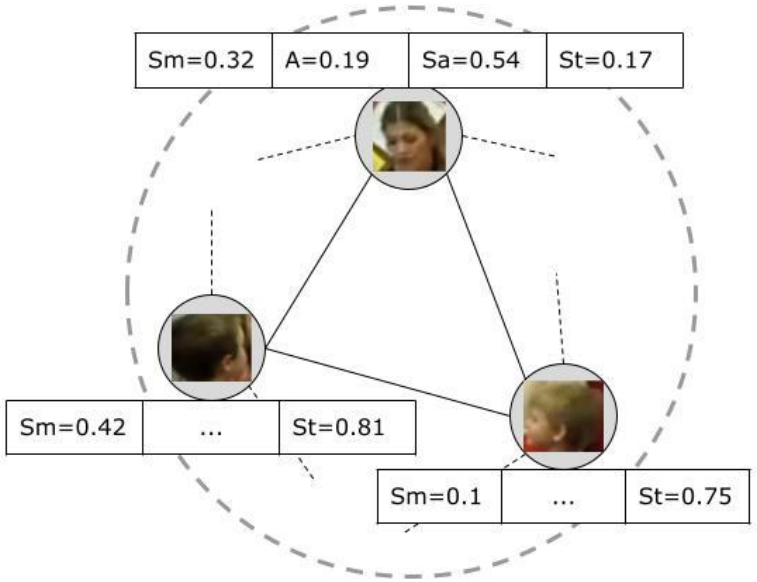
Tracking GNN



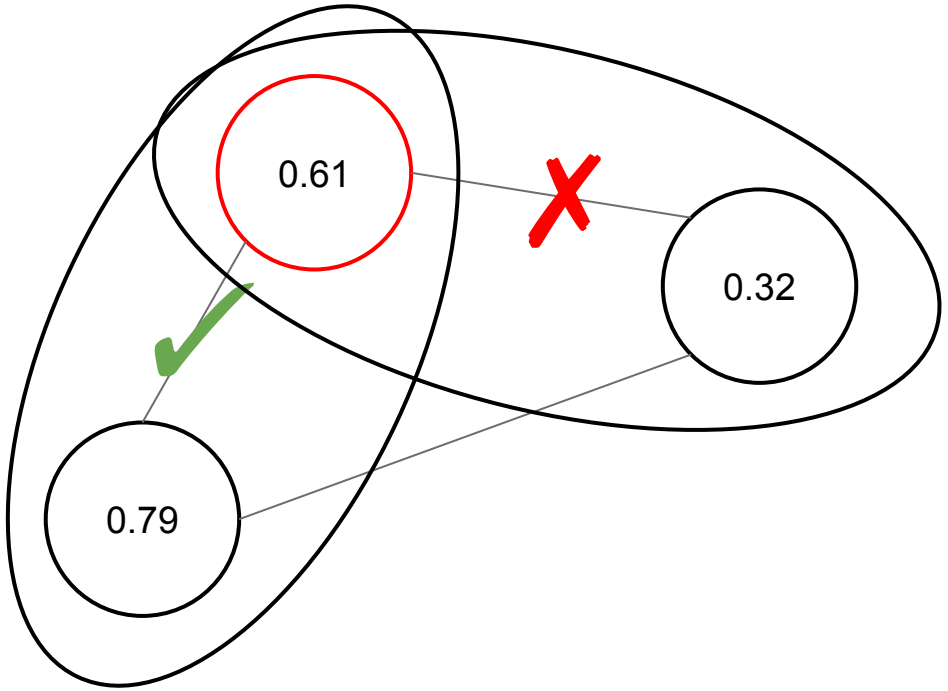
Simulation #1

Classroom Graphs

- Graph data structure with vertex set **V** and adjacency matrix **A**
- Each node in **V** contains features indicating:
  - Smiling
  - Angry
  - Sad
  - Is the individual a student
- Each adjacency $\mathbf{a_{ij}}$ in **A** is $1 - \frac{d(v_i, v_j)}{\sqrt{w^2 + h^2}}$



Simulating Classroom Graphs



Simulating Classroom Graphs Pt. 2

- Steps to construct graph:
  - Establish one individual in simulated graph as the Teacher
  - For each person in each timestep:
    i. Randomly sample smiling feature
    ii. Randomly sample movement vector
  - Evaluate ground truth **y** label as positive if all individuals had positive interactions with high smiling feature values
- Repeat until we create a balanced dataset

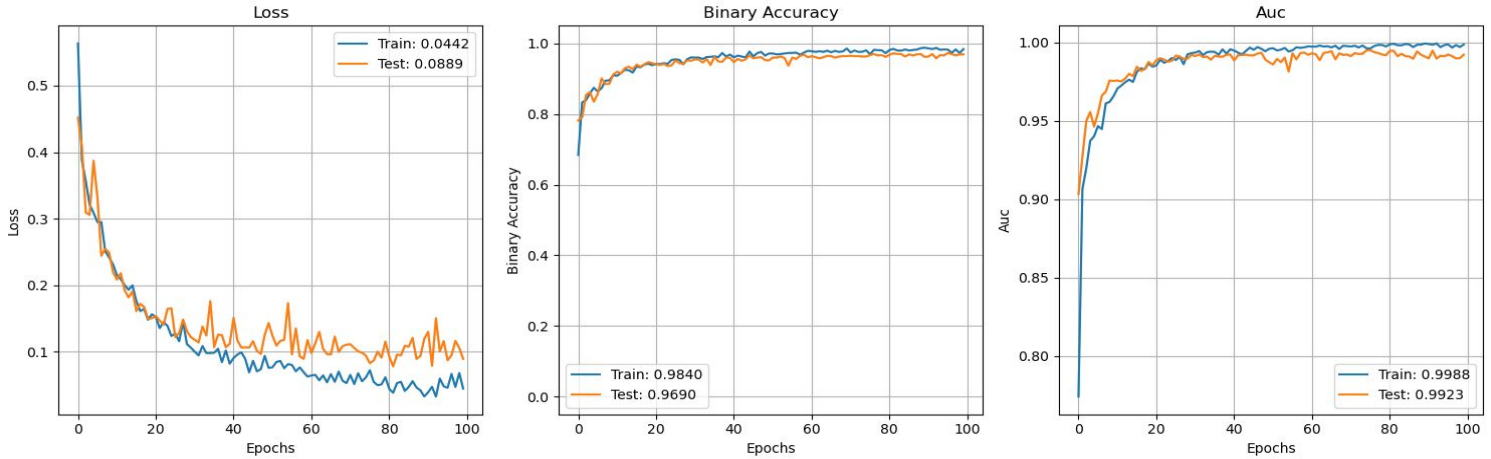
Comparing Networks

Both trained for 100 epochs at 0.01 learning rate on a dataset simulated with n = 4 nodes and t = 10 timesteps.

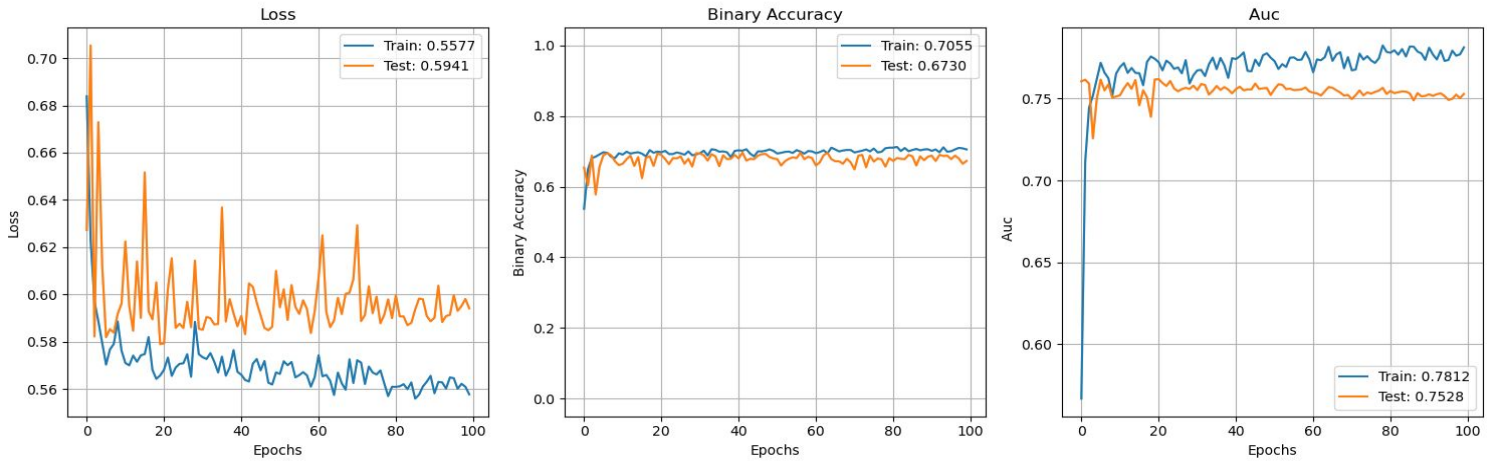
Tracking: **96.9%** test accuracy

Non-tracking: 67.3% test accuracy

Tracking



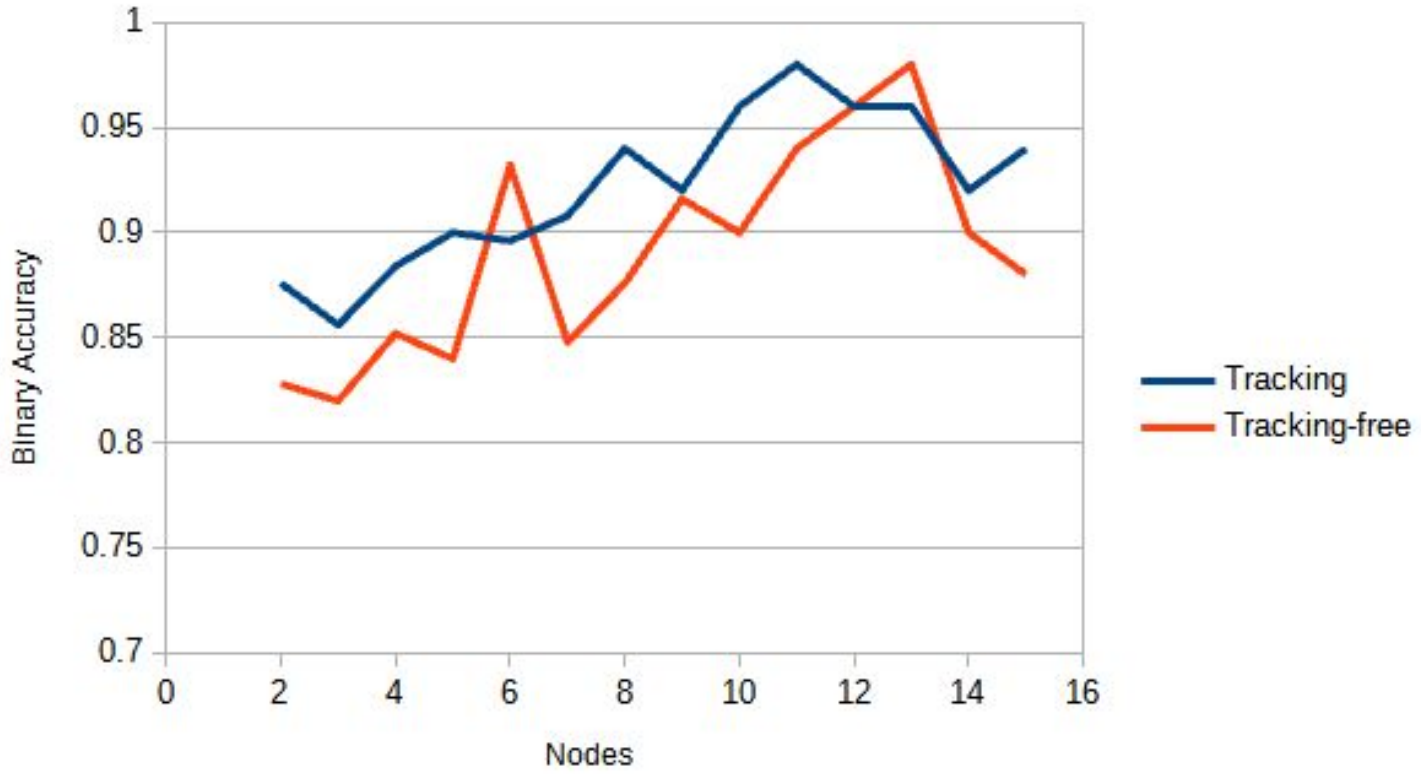
Non-tracking



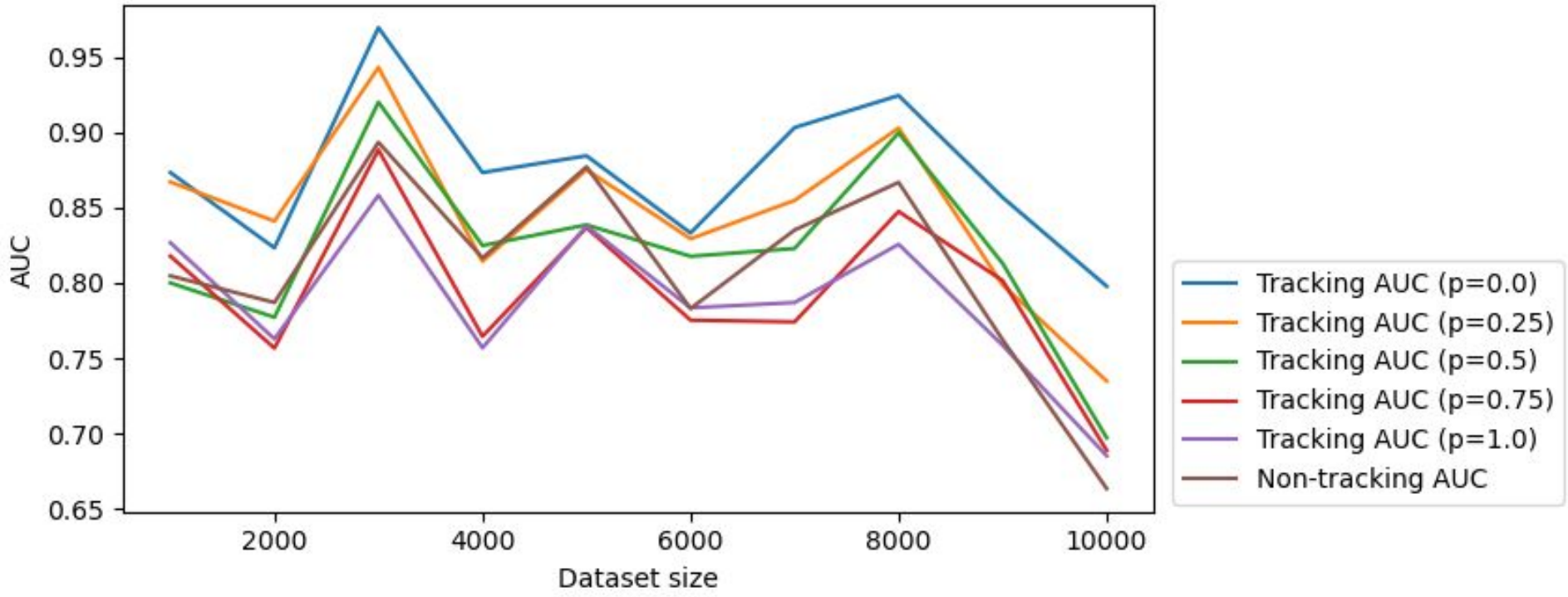
Larger Graphs

Previous experimentation showed lower accuracy when trained on simulated data with more nodes.

We repeat these experiments for an increasing number of nodes, stopping training early to optimize validation loss.



Tracking Error



1) Longer simulated datasets have more swaps, networks fit to lower AUC.
2) Non-tracking is worse than tracking with no swaps (p=0).

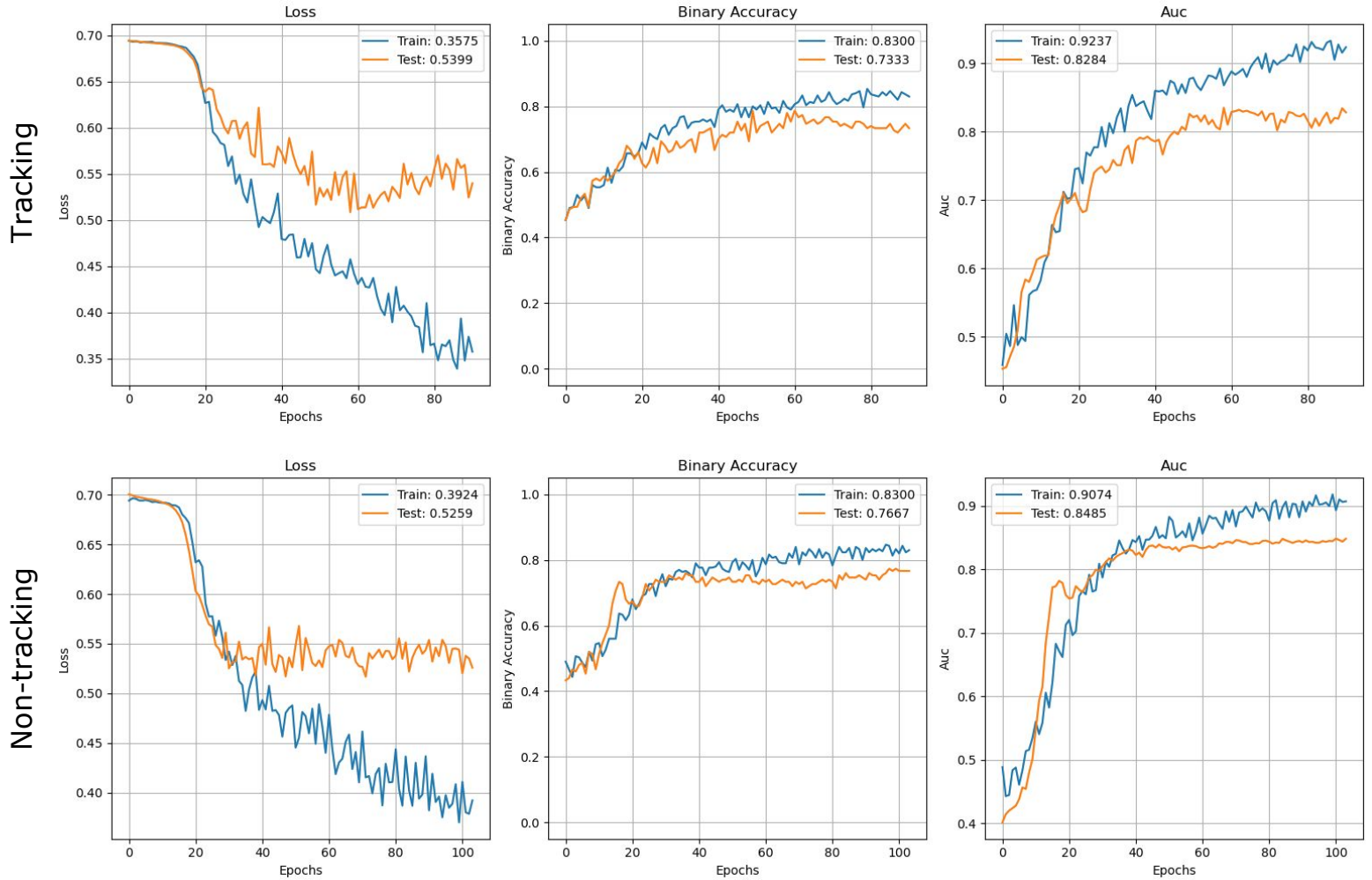
Simulation #2

Comparing Networks

Both trained for 100 epochs at 0.001 learning rate on a dataset simulated with n = 22 nodes and t = 10 timesteps, in which participants *must be close in proximity to have a positive interaction*.

Tracking: 73% test accuracy

Non-tracking: 76.7% test accuracy.



Tracking

Tracking for an Ordered Graph

- To construct our ordered social network graph, we have to know who is who in sequential frames of video
- We evaluate a number of facial detection tools and choose YOLOv3 (Redmon & Farhadi, 2018) due to the amount of detected faces and consistency
- FaceNet (Schroff et al., 2017) is the embedding network we choose to generate a discriminable representation from the detected faces

Methods of Matching

Given our embedding network FaceNet **E(x)** and parameter **δ**, we evaluate the similarity between two embeddings **A** and **B**, with coordinates $\mathbf{A_{xy}}$ and $\mathbf{B_{xy}}$ and cropped face pixels $\mathbf{A_p}$ and $\mathbf{B_p}$, respectively:

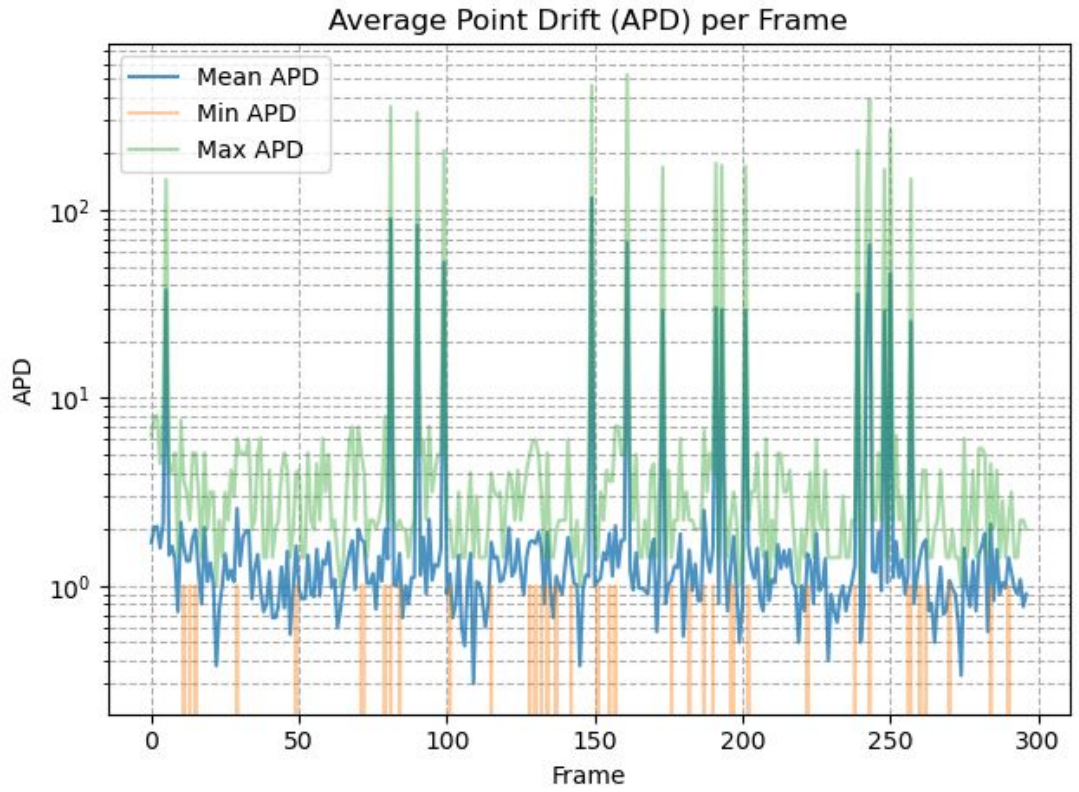
$$\delta * \frac{E(A_p) \cdot E(B_p)}{\|E(A_p)\|\|E(B_p)\|} + (1 - \delta) * \|A_{xy} - B_{xy}\|$$

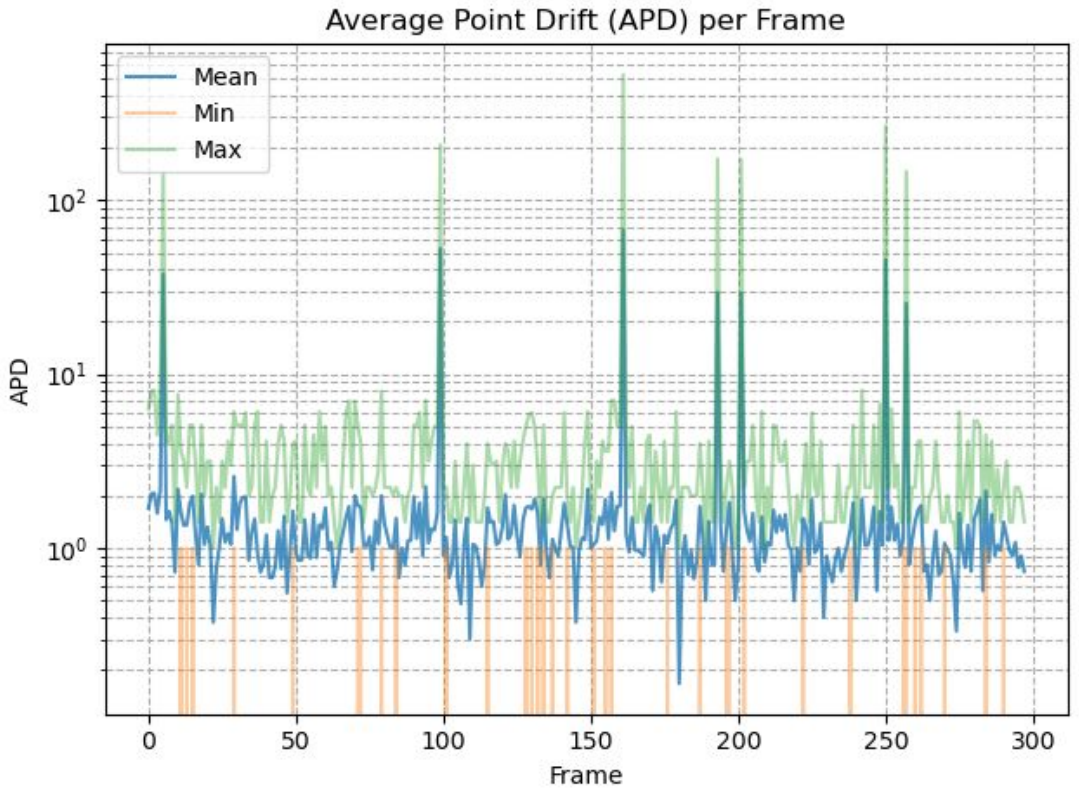
Which is a combination of the cosine similarity of the embeddings (left) and the Euclidean distance of the faces (right).

Matching Results

- On a small dataset of faces from first and last frames of video, both $\delta = 0$ and $\delta = 1$ perform inconsistently
- Using Intersection over Union (IoU) as indication of same faces in sequential frames:
  - $\delta = 0$ is not a fair method of matching because IoU accounts for Euclidean distance
  - $\delta = 1$ results in an **89%** matching accuracy

Average Point Drift

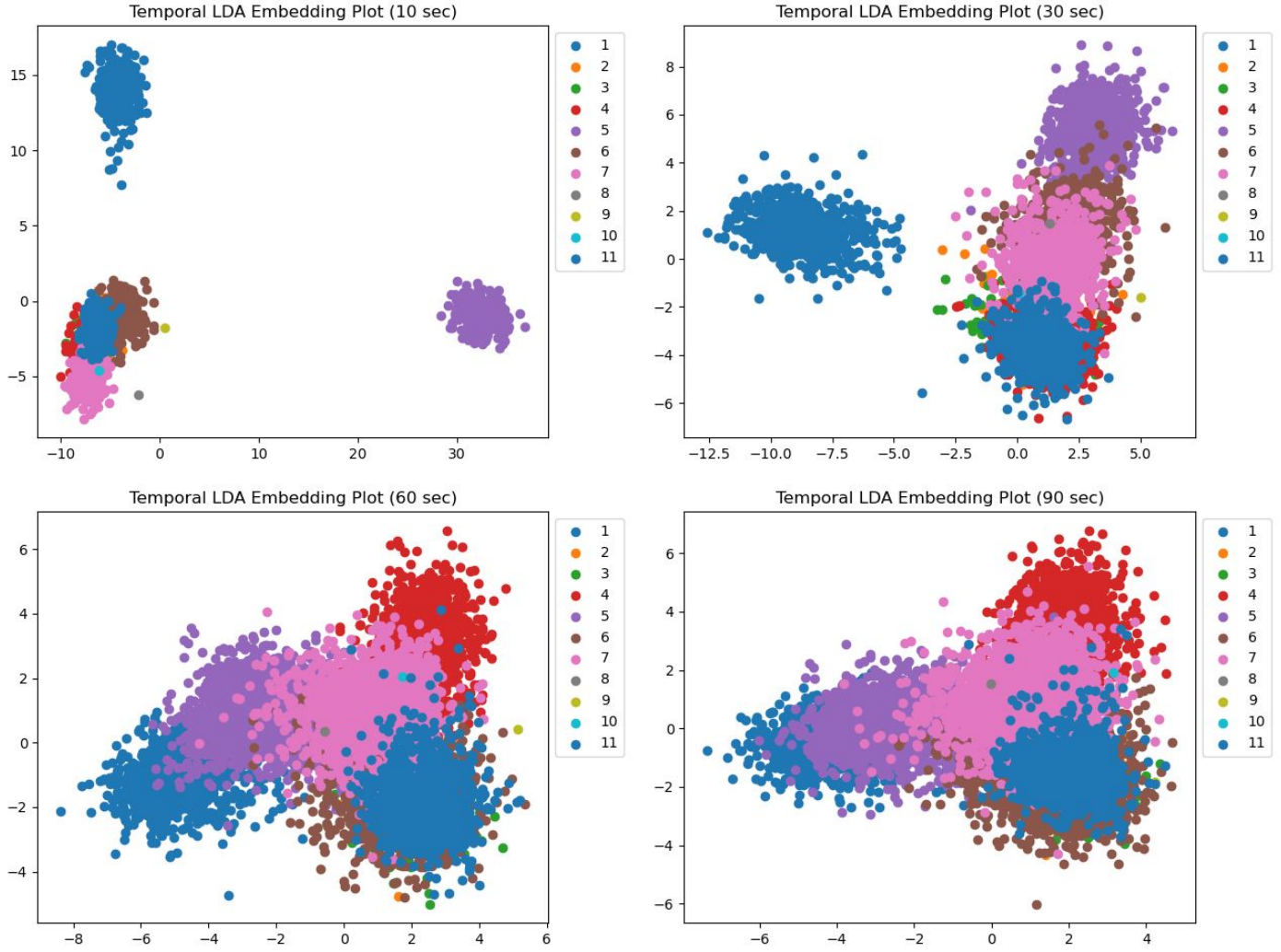




With δ = 1 (left) and δ = 0.5 (right), we observe a decrease in likely erroneous swaps (indicated by spikes) using a combination of our two heuristics.

Latent Space Analysis

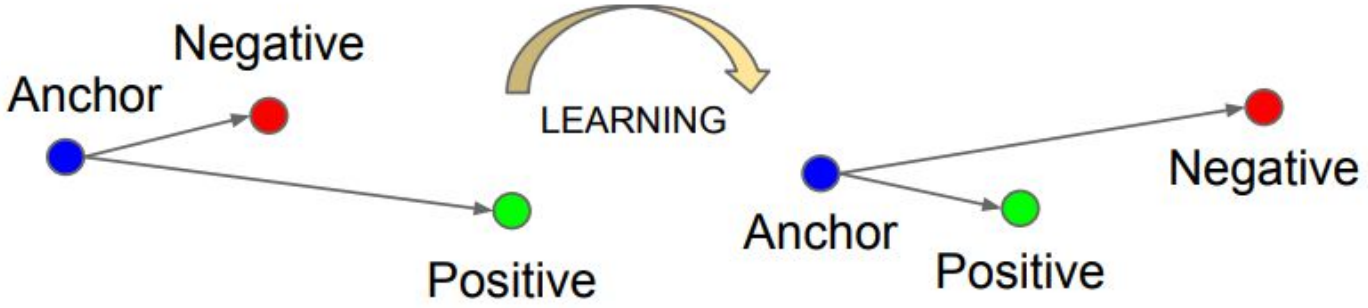
Progression of linear discriminant analysis (LDA) of embedding space over the first 90 seconds of a video clip (start: top left, end: bottom right).



Fine-tuning FaceNet

Triplet Loss

- Triplet: anchor **a**, positive **p**, and negative **n**
  - The goal of triplet loss to maximize the distance between the embeddings of the same face (**a** and **p**) is small and the distance between the embeddings of different faces (**a** and **n**) is large



Schroff et al., 2015

FaceNet (Schroff et al., 2015)

- FaceNet is an embedding network that maps face images to a Euclidean space where distances correspond to a measure of similarity
- Deep convolutional network that optimizes triplet loss
- Essential to efficiently fitting the network is the process of mining hard triplets, which are chosen because of their initially high triplet loss

Classroom Observation Faces

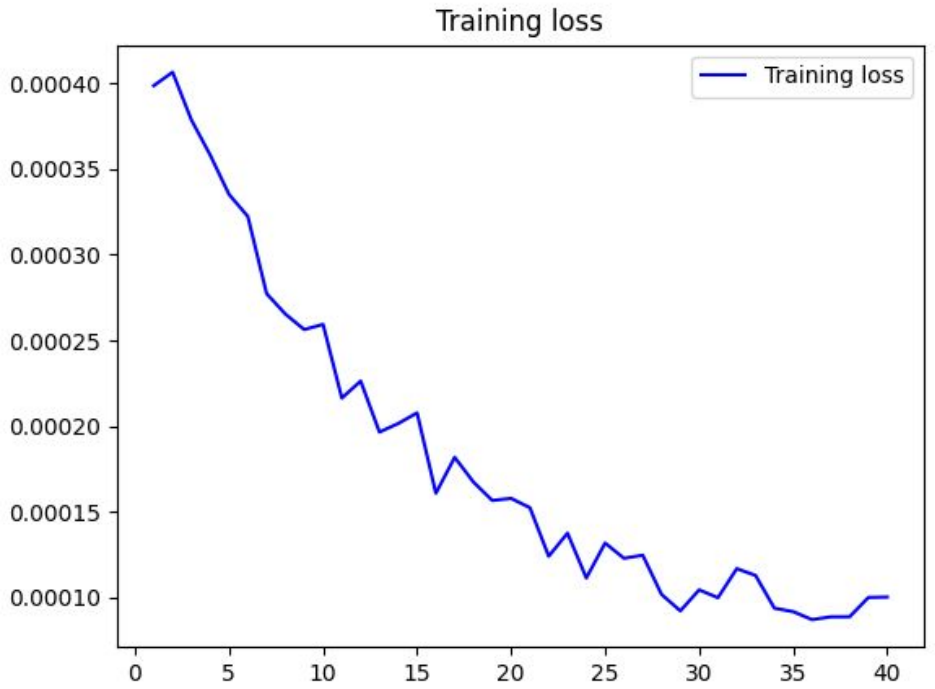
- Sample matching faces from frame pairs (3 frames apart), and use these combinations to generate triplets
- Successful matching frames contain at least 2 or more faces which we can use to generate triplets from

20 YouTube Videos
33,891 frame pairs

UVA Toddler
168,254 frame pairs

Results

- We restore weights from training on VGGFace2 and train with 0.0001 learning rate, annealed further over 40 epochs
- Improve ROC AUC of distinguishing same vs different face from 0.95 to **0.98** on unseen classroom observations



Thank you!