ECCV

EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO

2024

VideoClusterNet: Self-Supervised and Adaptive Face Clustering For Videos

By
Devesh Walawalkar, Pablo Garrido
Flawless AI Inc.

E: {devesh.walawalkar, pablo.garrido}@flawlessai.com
Arxiv link: https://arxiv.org/abs/2407.12214

EUROPEAN CONFERENCE ON COMPUTER VISION

Overview

1. We present a fully self-supervised video face clustering framework that efficiently auto-adapts to specific variations observed in set of faces in a given video.

2. Major highlights of our work include:

    1. A self-supervised model finetuning method that depends on only positive face match pairs to improve the face embeddings.
    2. A deep learning-based similarity metric for face clustering, which automatically adapts to a given model's learned embedding space.
    3. A fully automated video face clustering algorithm that does not require any user input parameters.
    4. Release of a movie face clustering benchmark dataset called MovieFaceCluster which provides extreme challenging face clustering scenarios present in the movie domain.

EUROPEAN CONFERENCE ON COMPUTER VISION

Objective

Clustering face motion tracks in a given video across common facial identities into a single group.

Specific challenges for movie domain

1. Movie/TV series domain has higher than usual count of hard to identify faces (i.e. extreme pose, dark lighting, heavy occlusion, blurriness etc.)
2. A character's appearance can change drastically through the progression of the movie.
3. For certain movie settings, large count of background/secondary characters are present. (i.e. crowd scenes etc.)

EUROPEAN CONFERENCE ON COMPUTER VISION

Current limitations

1. Use of ground truth cluster labels to train DL model to *pull close/push away embeddings*.
2. Requirement of either *number of clusters* as algorithm input or
3. Requirement of *one or more user defined thresholds* to define cluster boundaries.
4. Use of *euclidean or cosine distance metric* for comparing embedding in model feature space.
5. Self-supervised methods use complex methods based on temporal constraints (such as co-occurring tracks) to *mine negative samples to contrast with*.

Proposed framework tackles all these limitations ….

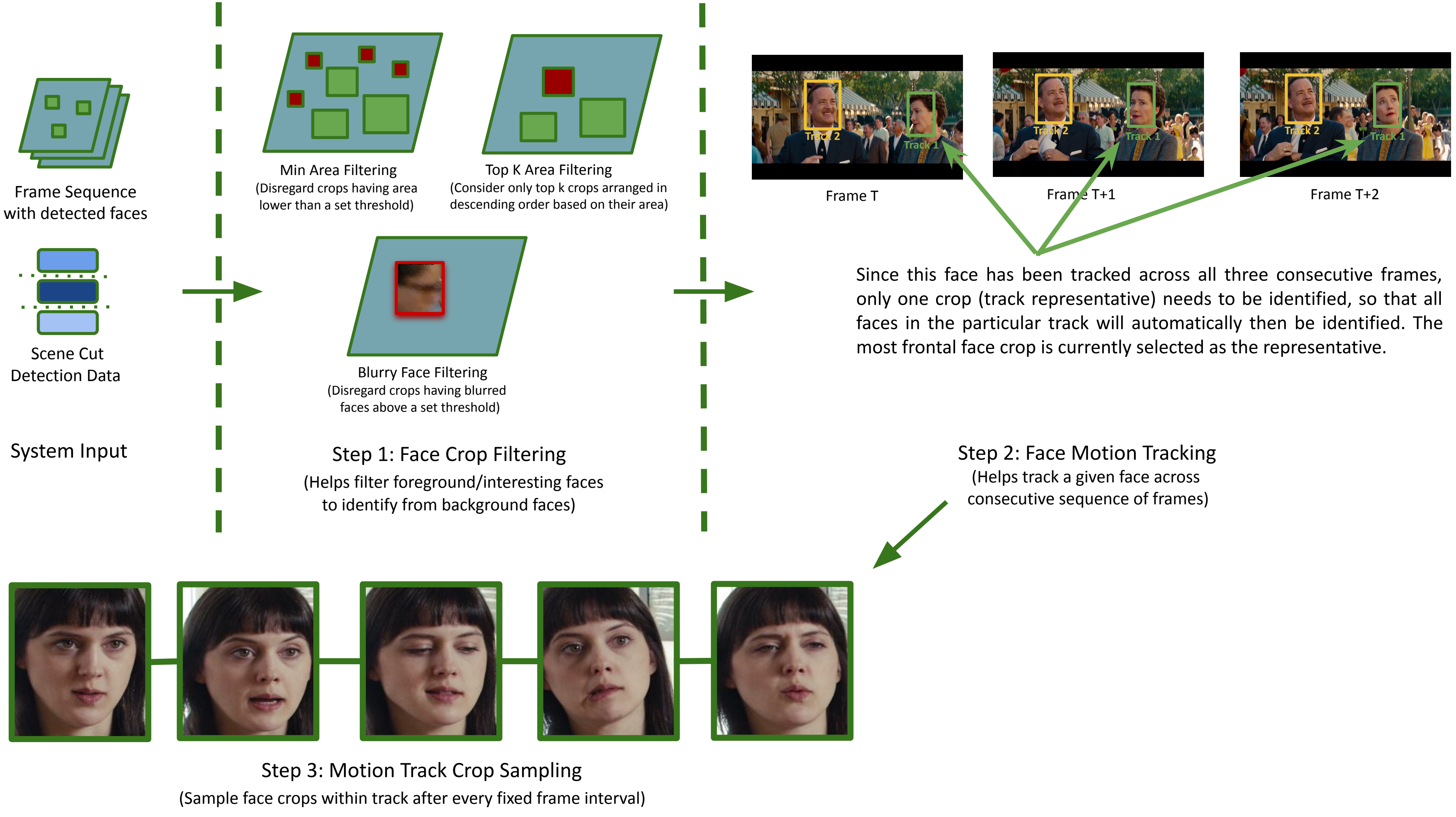
Central Concept

Given a large scale pretrained face identification model,

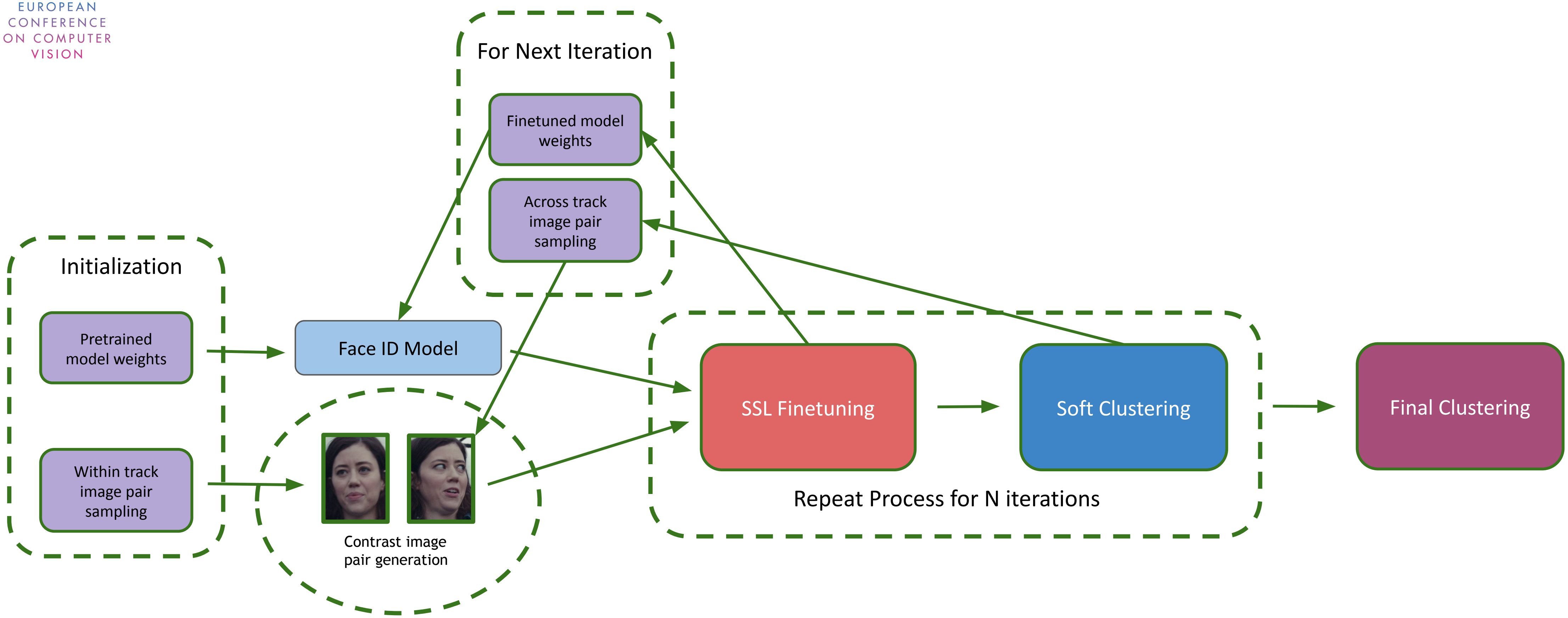
1. *Motion track faces* along various shots in a given frame sequence (movies/TV series).
2. *Finetune the model* on specific faces present in the motion tracks, using *temporal self-supervision*.
3. *Soft match tracks for common identities*, which subsequently enhances the self-supervised model finetuning.
4. *Iteratively* perform soft matching and finetuning, which *progressively helps model adapt to specific faces*.
5. Use *model learnt similarity metric* for final clustering of tracks.

ECCV

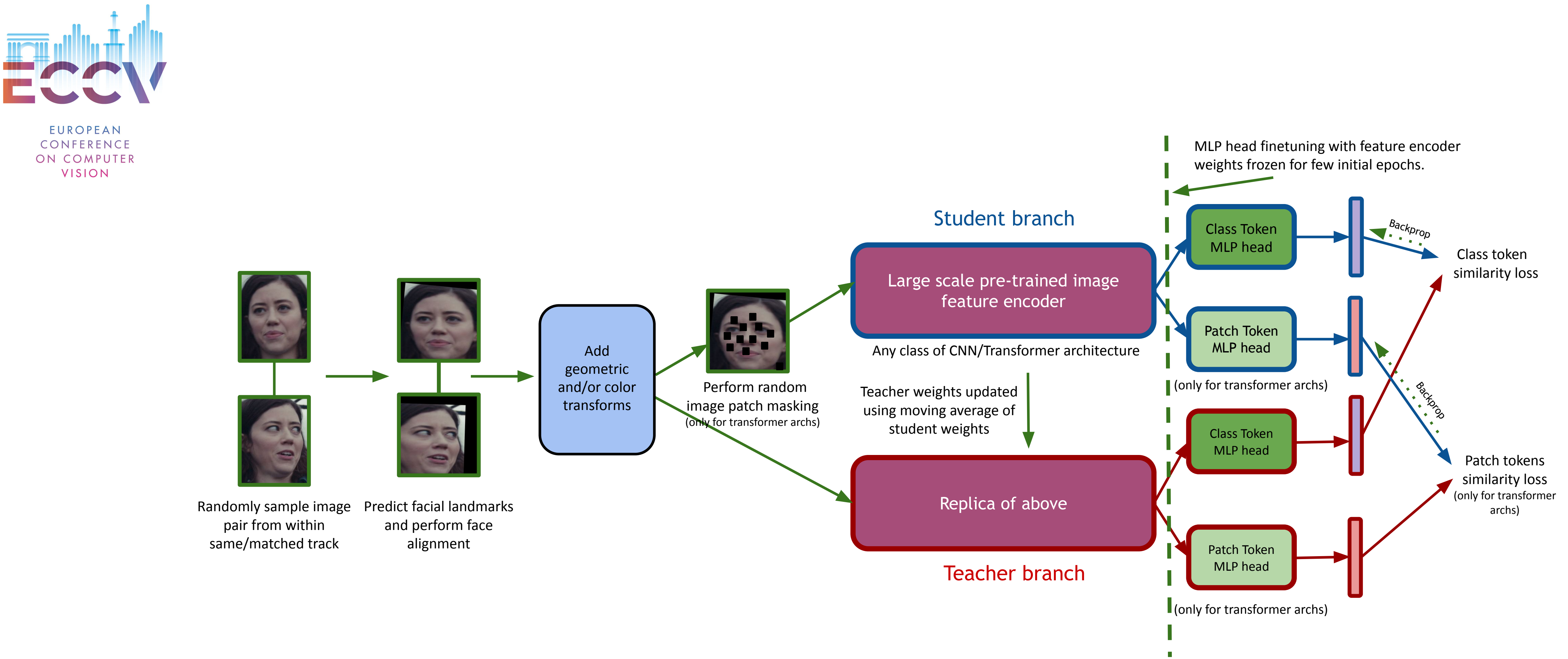
EUROPEAN CONFERENCE ON COMPUTER VISION



Stage 1: Face Motion Track Pre-Processing



Self Supervised Video Face Clustering - Central Idea

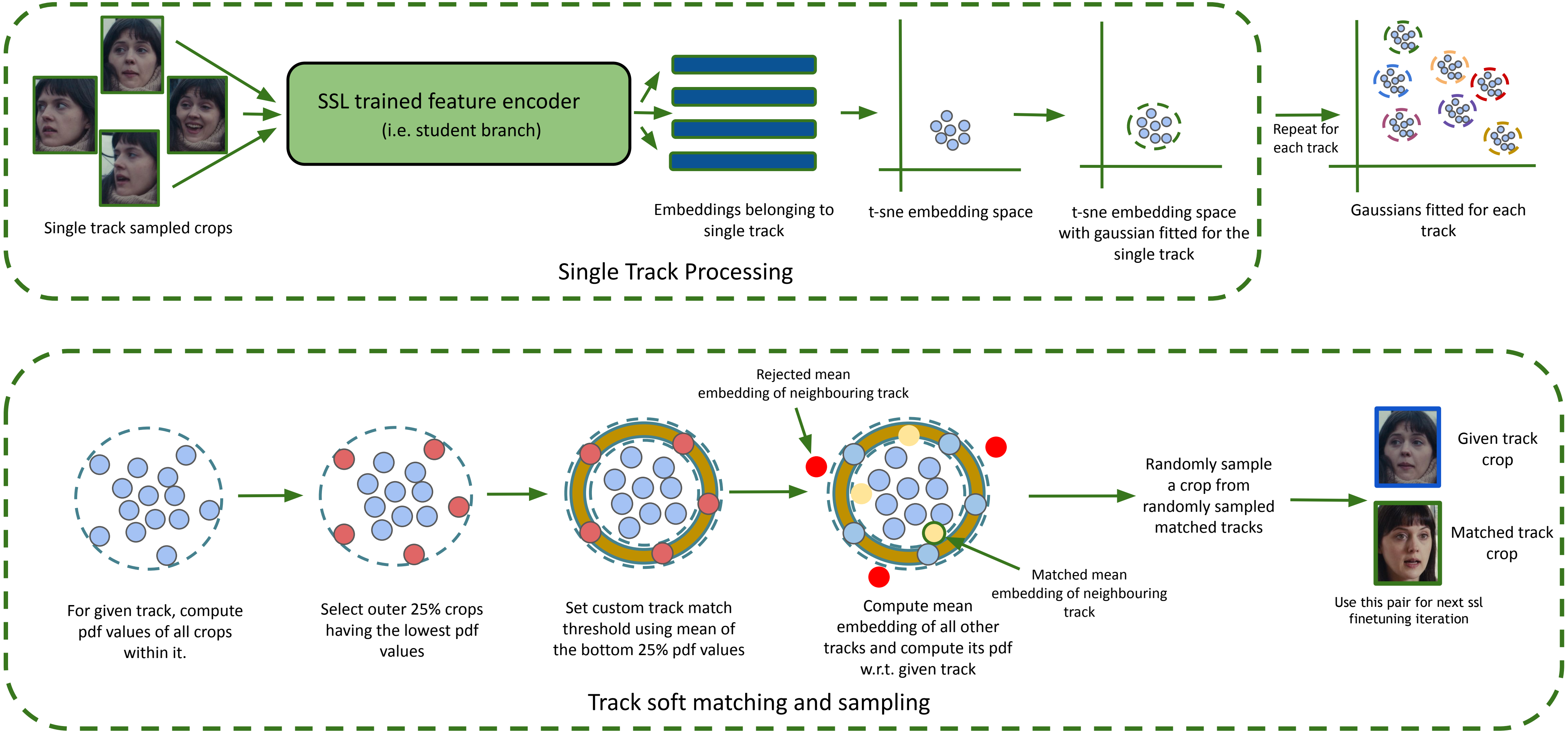


loss: -1*softmax(vec_1/temp_1)*log_softmax(vec_2/temp_2)

temp_* : softening temperature

Stage 2: Self Supervised Model Finetuning

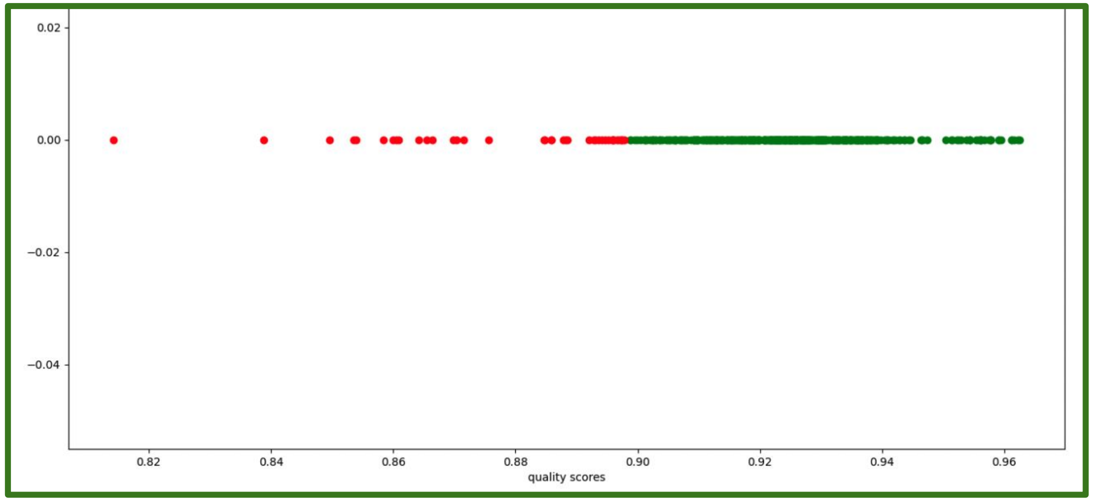
EUROPEAN
CONFERENCE
ON COMPUTER
VISION



Stage 3: Soft Clustering using fitted Track Normal Distribution matching

Final Clustering

Step 1: Compute the facial crop quality of a track using the dropout based embedding variance method[1] and filter outliers based on the distribution of computed track quality values.



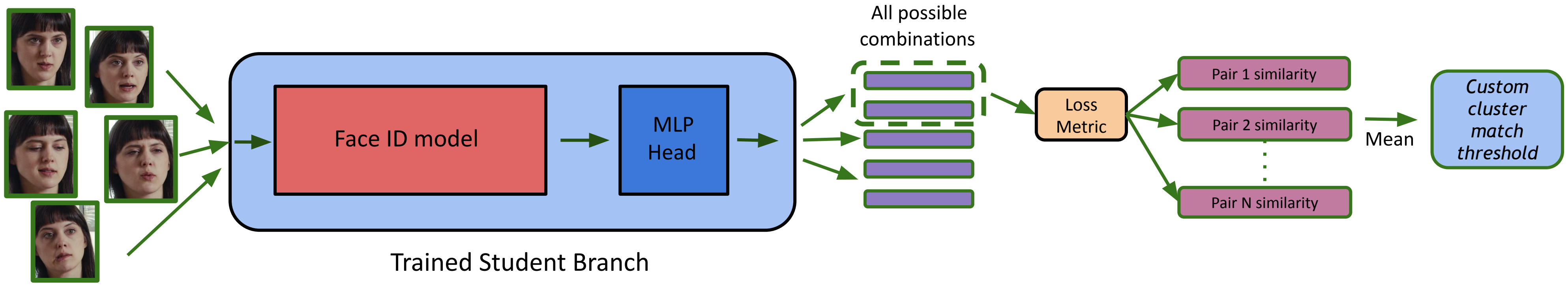
Stage 4: Final Clustering algorithm using learnt embedding similarity metric

[1] Terhorst, Philipp, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. "SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5651-5660. 2020.

Final Clustering

Step 2: Initialize a cluster instance for each valid track.

Step 3: Compute a custom match threshold for each cluster by passing all crops within the cluster through the trained student branch. Compute the similarity between all possible combinations of cluster crops through the loss function (i.e. learnt similarity metric). Assign the loss value mean as custom threshold for given cluster.



Stage 4: Final Clustering algorithm using learnt embedding similarity metric

Final Clustering

Step 4:

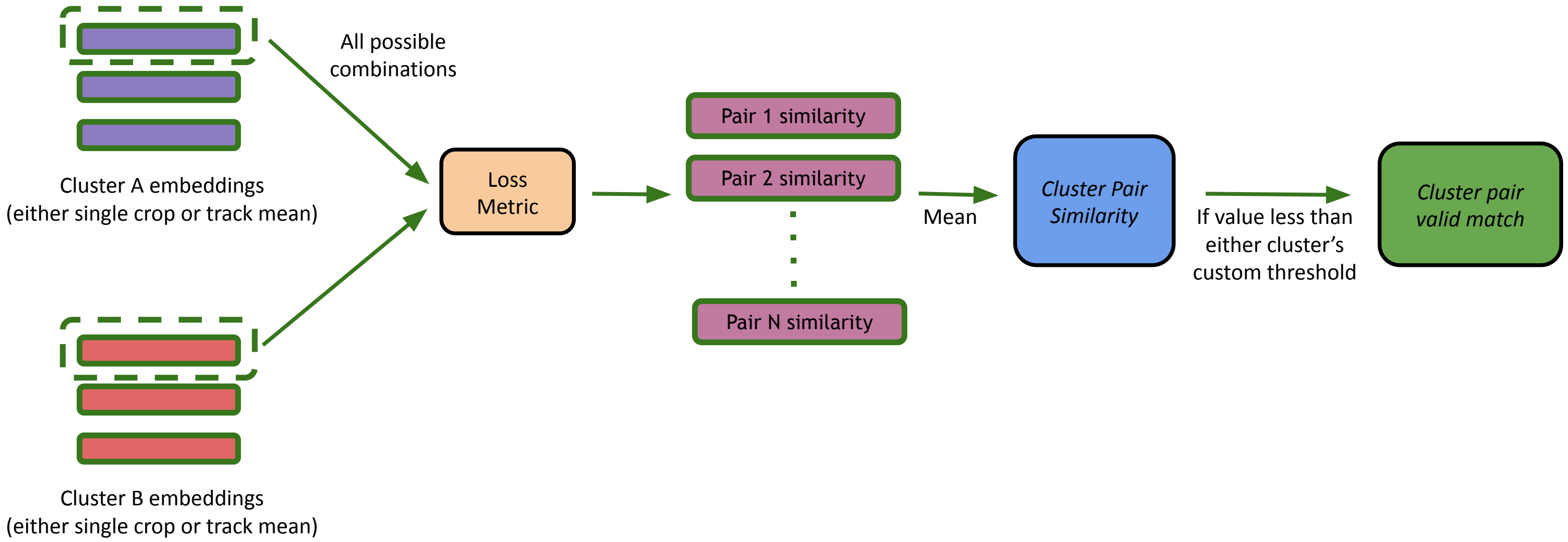
For a given pair of clusters, compute the similarity metric between all possible combination of tracks between them.

If a cluster has only one track, create query match pairs using each individual crop.

If a cluster has multiple tracks, create query match pairs using mean embedding of each track.

The overall cluster pair similarity metric is a mean of all individual computed match values.

The pair is considered a valid match if overall similarity metric is lower than either cluster’s custom match threshold.



Stage 4: Final Clustering algorithm using learnt embedding similarity metric

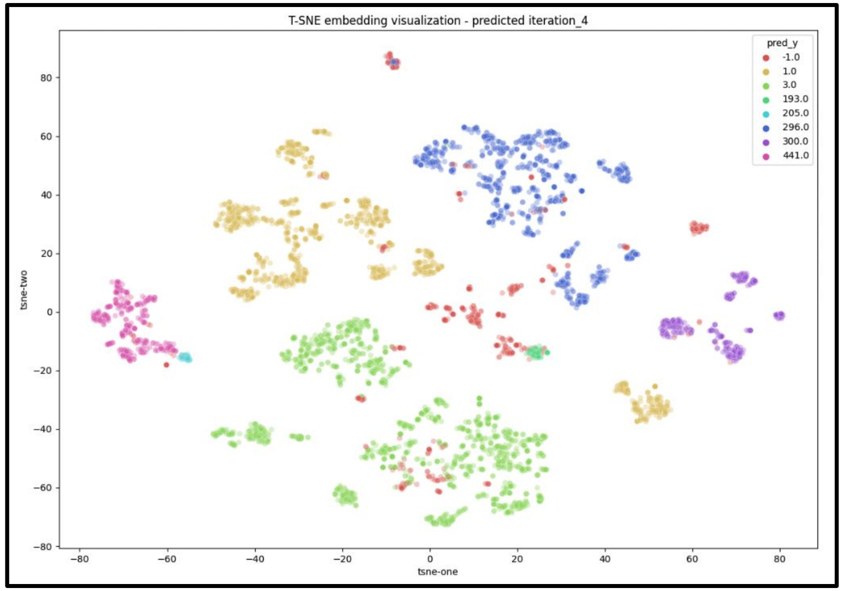
Final Clustering

Step 5:

Merge clusters with valid matches and daisy chain merges through all the cluster pairs. (e.x. If cluster pairs 1,2 and pairs 2,3 are matched, then link them up together)

Step 6:

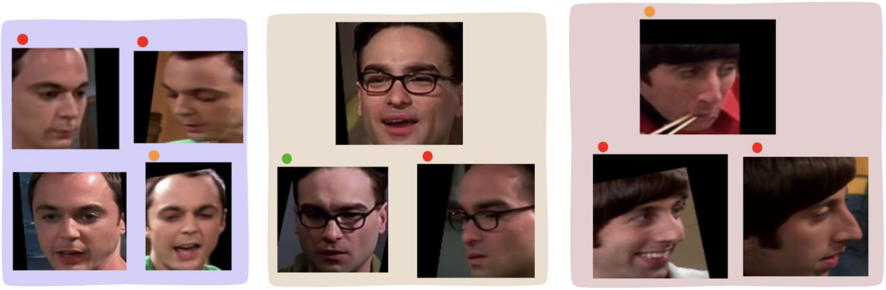
Repeat steps 3 to 5 until there is no change in cluster definitions over previous iterations.



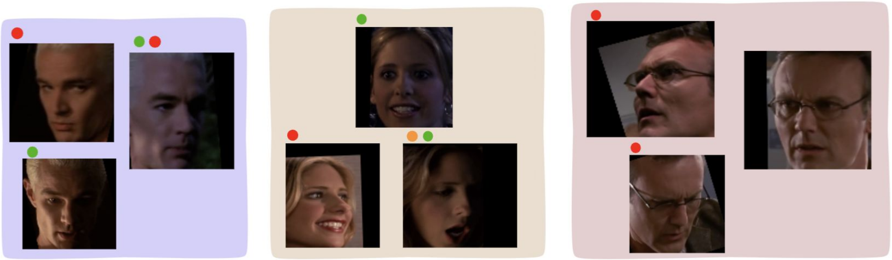
Stage 4: Final Clustering algorithm using learnt embedding similarity metric

Experimental Analysis

TV Series Datasets - Big Bang Theory (BBT) S01 and Bluffy, The Vampire Slayer (BVS) S05



The Big Bang Theory (BBT)



Buffy, The Vampire Slayer (BVS)

Method	BBT S01 Episode						
	S1E1	S1E2	S1E3	S1E4	S1E5	S1E6	Combined
SCTL [54]	66.48	-	-	-	-	-	-
TSiam [41]	96.4	-	-	-	-	-	-
SSiam [41]	96.2	-	-	-	-	-	-
MLR [4]	95.18	94.16	77.81	79.35	79.93	75.85	83.71
BCL [47]	98.63	98.54	90.61	86.95	89.12	81.07	89.63
CCL [42]	98.2	-	-	-	-	-	-
VCTRSF [53]	99.39	**99.84**	97.58	96.41	98.47	93.33	94.20
Ours⋆†	**99.70**	99.67	**98.60**	**98.80**	**99.10**	**97.10**	**98.70**

Method	BVS S05 Episode						
	S5E1	S5E2	S5E3	S5E4	S5E5	S5E6	Combined
HMRF [55]	-	50.3	-	-	-	-	-
WBSLRR [56]	-	62.7	-	-	-	-	-
TSiam [41]	-	92.46	-	-	-	-	-
SSiam [41]	-	90.87	-	-	-	-	-
CP-SSC [44]	-	65.2	-	-	-	-	-
MvCorr [43]	-	97.7	-	-	-	-	-
MLR [4]	71.99	61.27	66.60	67.07	69.59	61.72	66.37
BCL [47]	92.08	79.76	84.00	84.97	89.05	80.58	83.62
CCL [42]	-	92.1	-	-	-	-	-
Ours⋆†	**96.30**	**99.10**	**98.70**	**97.43**	**99.00**	**96.78**	**96.10**

Table 1: WCP/Clustering Accuracy on BBT-S01 and BVS-S05. ⋆We use ArcFace-R100 [15] as our pre-trained base model. Combined results indicate clustering performance on set of face tracks from all six episodes combined together. † For fair literature comparison, we use the same face detection, tracking, and clustering labels as provided in [41, 47], thereby not utilizing our proposed advanced pre-processing modules in order to effectively compare pure track clustering performance against literature methods.

[15] Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699. IEEE (2019)
[4] Bäuml, M., Tapaswi, M., Stiefelhagen, R.: Semi-supervised learning with constraints for person identification in multimedia data. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3602–3609. IEEE (2013)
[41] Sharma, V., Tapaswi, M., Sarfraz, M.S., Stiefelhagen, R.: Self-supervised learning of face representations for video face clustering. In: International Conference on Automatic Face & Gesture Recognition. pp. 1–8. IEEE (2019)
[42] Sharma, V., Tapaswi, M., Sarfraz, M.S., Stiefelhagen, R.: Clustering based contrastive learning for improving face representations. In: IEEE International Conference on Automatic Face and Gesture Recognition. pp. 109–116. IEEE (2020)
[43] Somandepalli, K., Hebbar, R., Narayanan, S.: Robust character labeling in movie videos: Data resources and self-supervised feature adaptation. IEEE Transactions on Multimedia 24, 3355–3368 (2021)
[44] Somandepalli, K., Narayanan, S.S.: Reinforcing self-expressive representation with constraint propagation for face clustering in movies. In: International Conference on Acoustics,Speech and Signal Processing (ICASSP). pp. 4065–4069. IEEE (2019)
[47] Tapaswi, M., Law, M.T., Fidler, S.: Video face clustering with unknown number of clusters. In: International Conference on Computer Vision (ICCV). pp. 5027–5036. IEEE (2019)
[53] Wang, Y., Dong, M., Shen, J., Luo, Y., Lin, Y., Ma, P., Petridis, S., Pantic, M.: Self-supervised video-centralised transformer for video face clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 45(11), 12944–12959 (2023)
[54] Wu, B., Lyu, S., Hu, B., Ji, Q.: Simultaneous clustering and tracklet linking for multi-face tracking in videos. In: International Conference on Computer Vision (ICCV). pp. 2856–2863. IEEE (2013)
[55] Wu, B., Zhang, Y., Hu, B., Ji, Q.: Constrained clustering and its application to face clustering in videos. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3507–3514. IEEE (2013)
[56] Xiao, S., Tan, M., Xu, D.: Weighted block-sparse low rank representation for face clustering in videos. In: European Conference on Computer Vision (ECCV). pp. 123–138. Springer (2014)

Experimental Analysis

Release of MovieFaceCluster Dataset[1]

1. Given the unique in-the-wild challenges in video production domain, we present a novel video face clustering dataset, which incorporates challenging movies hand-selected by experienced film post-production specialists.

2. It is a collection of nine movies, with facial identity labels provided for each movie face motion track.

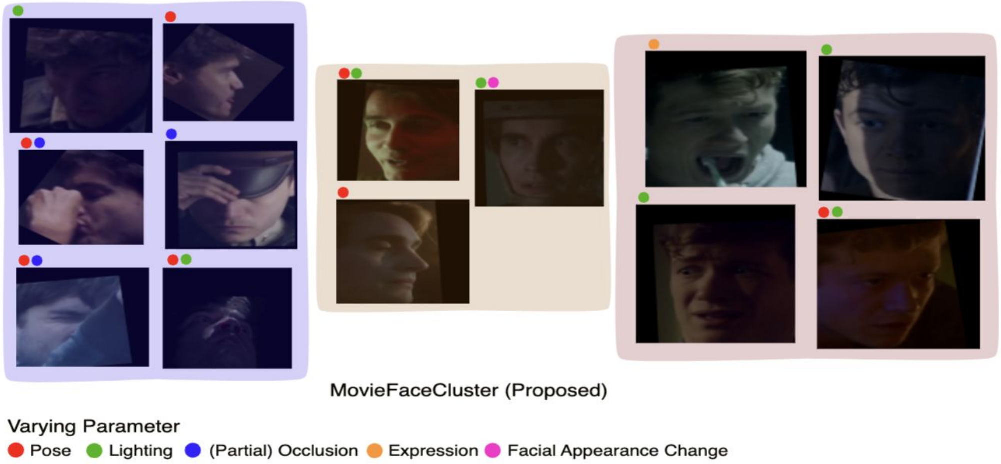


Fig. 1: Select hard case clusters predicted using our algorithm from within **MovieFaceCluster** dataset. Trivial face represents an easy ID sample for each cluster. The term "varying parameter" depicts the dominant image attributes that are particularly challenging for a given face crop. It is not part of the dataset annotations but is depicted for enhanced reader understanding.

[1] https://www.flawlessai.com/dataset/

Experimental Analysis - MovieFaceCluster

Method	Movie								
	An Elephant's Journey (2019)	Armed Response	Angel Of The Skies	Death Do Us Part (2019)	American Fright Fest	The Fortress	Under The Shadow	The Hidden Soldier	S.M.A.R.T. Chase
	Weighted Cluster Accuracy (%) & Pred Cluster Ratio (Pred / GT)								
TSiam [41]	90.7 & 1.44	84.9 & 1.36	77.1 & 0.62	92.9 & 1.57	89.3 & 0.83	68.6 & 0.69	71.8 & 2.11	90.7 & 1.33	79.6 & 1.70
SSiam [41]	88.1 & 1.61	86.6 & 1.21	75.5 & 0.59	94.4 & 1.28	86.2 & 0.78	71.1 & 0.73	68.3 & 2.33	88.7 & 1.24	82.3 & 1.80
JFRAC [61]	91.4 & 1.33	85.2 & 1.50	73.4 & 0.62	90.8 & 0.71	91.5 & 0.86	65.3 & 0.77	73.1 & 2.00	92.6 & 1.19	85.8 & 1.70
CCL [42]	89.5 & N.A.†	89.7 & N.A.†	75.0 & N.A.†	95.4 & N.A.†	87.2 & N.A.†	62.7 & N.A.†	77.4 & N.A.†	84.0 & N.A.†	89.9 & N.A.†
VCTRSF [53]	96.3 & N.A.†	92.2 & N.A.†	77.7 & N.A.†	96.5 & N.A.†	91.3 & N.A.†	78.8 & N.A.†	78.7 & N.A.†	94.4 & N.A.†	88.4 & N.A.†
Ours	**97.2 & 1.11**	**94.1 & 0.93**	**85.9 & 0.72**	**98.0 & 1.14**	**97.6 & 0.92**	**89.3 & 1.02**	**82.5 & 1.88**	**98.5 & 1.04**	**93.8 & 1.50**

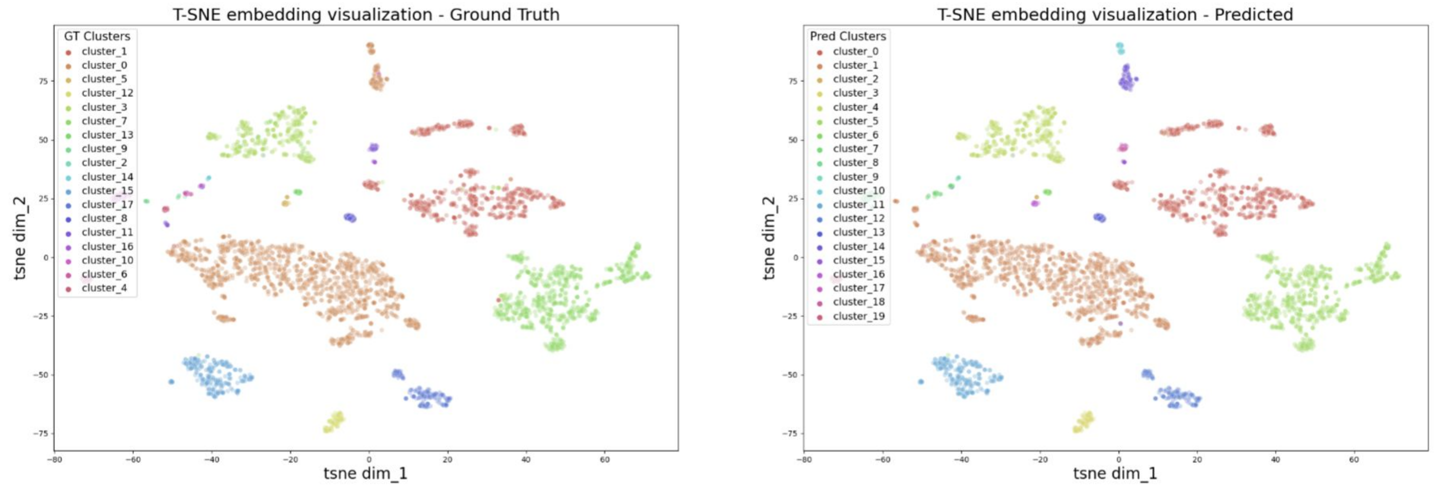
Table 3: Quantitative comparisons on each MovieFaceCluster dataset movie. For a fair comparison, we incorporate ArcFace-R100 [15] as the pre-trained feature extractor for all reported methods, including ours. We outperform SoTA methods w.r.t. cluster accuracy and predicted cluster ratio. Details on our implementation of all comparative methods can be found in the supplementary material. †Number of ground truth clusters is required as input for these methods, so PCR isn't a valid performance metric while also being a major limitation for these methods.

Fig. 4: Comparative t-SNE embedding visualizations on select **MovieFaceCluster** movie datasets. *Left:* Ground truth, *Right:* Our method. Each dot in the diagram above represents the finetuned model's extracted embedding for a face crop I_{t_n} in a given track's sampled crop set t. Face embeddings assigned to a given color constitute a single cluster.

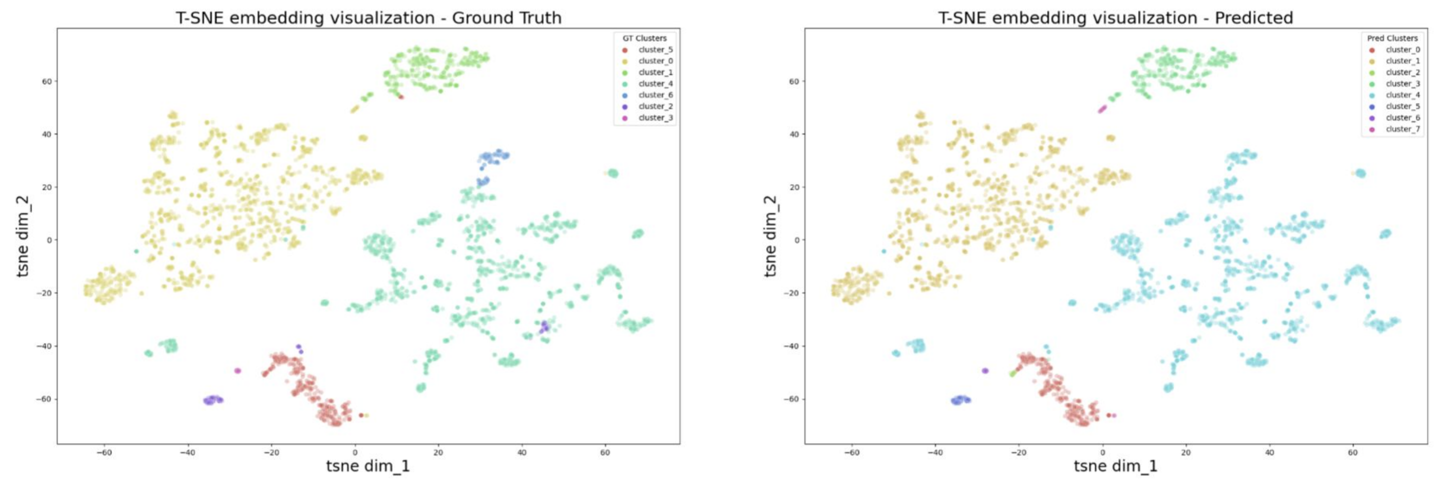
[41] Sharma, V., Tapaswi, M., Sarfraz, M.S., Stiefelhagen, R.: Self-supervised learning of face representations for video face clustering. In: International Conference on Automatic Face & Gesture Recognition. pp. 1–8. IEEE (2019)
[42] Sharma, V., Tapaswi, M., Sarfraz, M.S., Stiefelhagen, R.: Clustering based contrastive learning for improving face representations. In: IEEE International Conference on Automatic Face and Gesture Recognition. pp. 109–116. IEEE (2020)
[53] Wang, Y., Dong, M., Shen, J., Luo, Y., Lin, Y., Ma, P., Petridis, S., Pantic, M.: Self-supervised video-centralised transformer for video face clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 45(11)
[61] Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Joint face representation adaptation and clustering in videos. In: European Conference on Computer Vision (ECCV). pp. 236–251. Springer (2016)
[15] Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699. IEEE (2019)

EUROPEAN CONFERENCE ON COMPUTER VISION

Experimental Analysis - MovieFaceCluster



MovieFaceCluster : An Elephant's Journey



MovieFaceCluster : Death Do Us Part

Summary:

1. We present a novel *video face clustering algorithm* that specifically adapts to a given set of face tracks through a *fully self-supervised mechanism*.
2. Our fully automated approach to video face clustering specifically helps avoid any sub-optimal solutions that may be induced from *non-intuitive user-defined parameters*.
3. In addition, using a *model-learned similarity metric* over generic distance functions helps provide SoTA video face clustering performance over other competing methods.
4. *Extensive experiments and ablation studies* on our presented comprehensive movie dataset and traditional benchmarks underline our method's effectiveness under extremely challenging real-world scenarios.

Thank You