



FLAWLESS

VideoClusterNet: Self-Supervised and Adaptive Clustering For Videos

Devesh Walawalkar, Pablo Garrido
Flawless AI Inc.



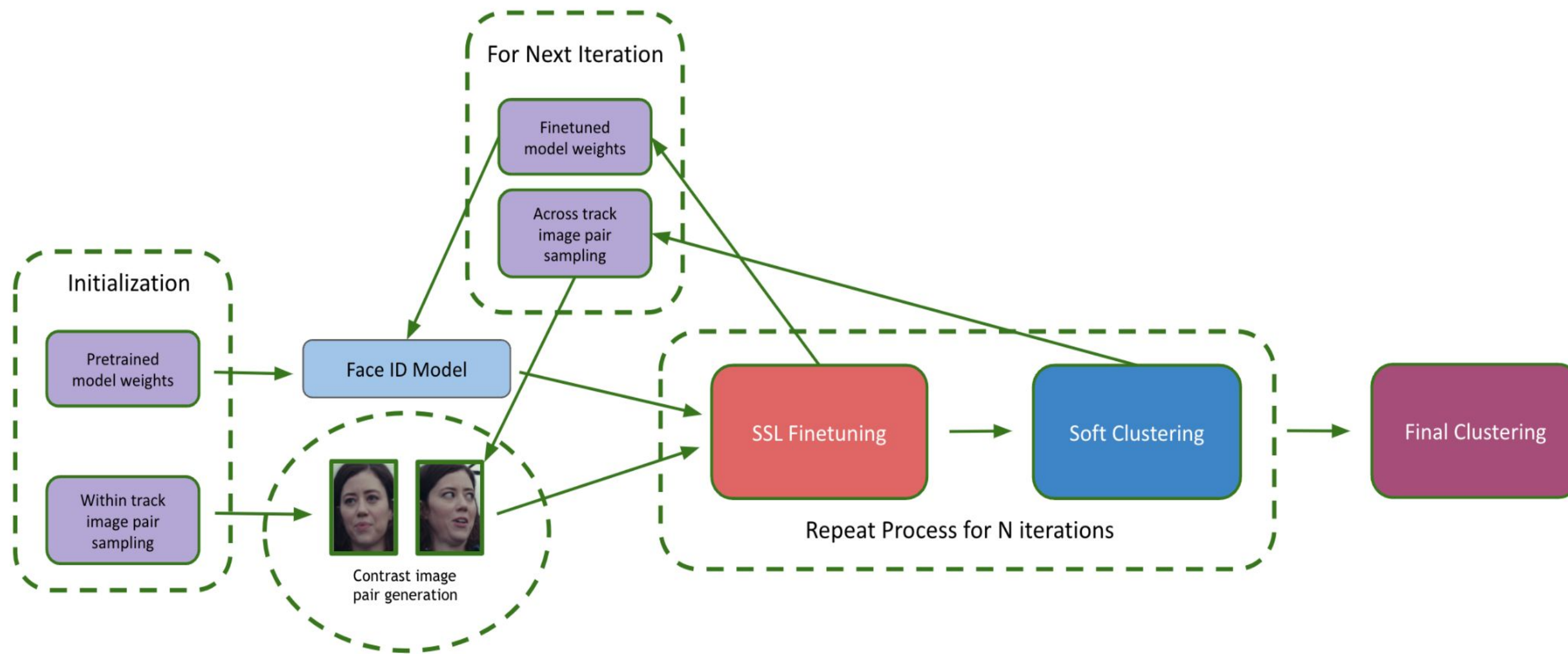
EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO
2024

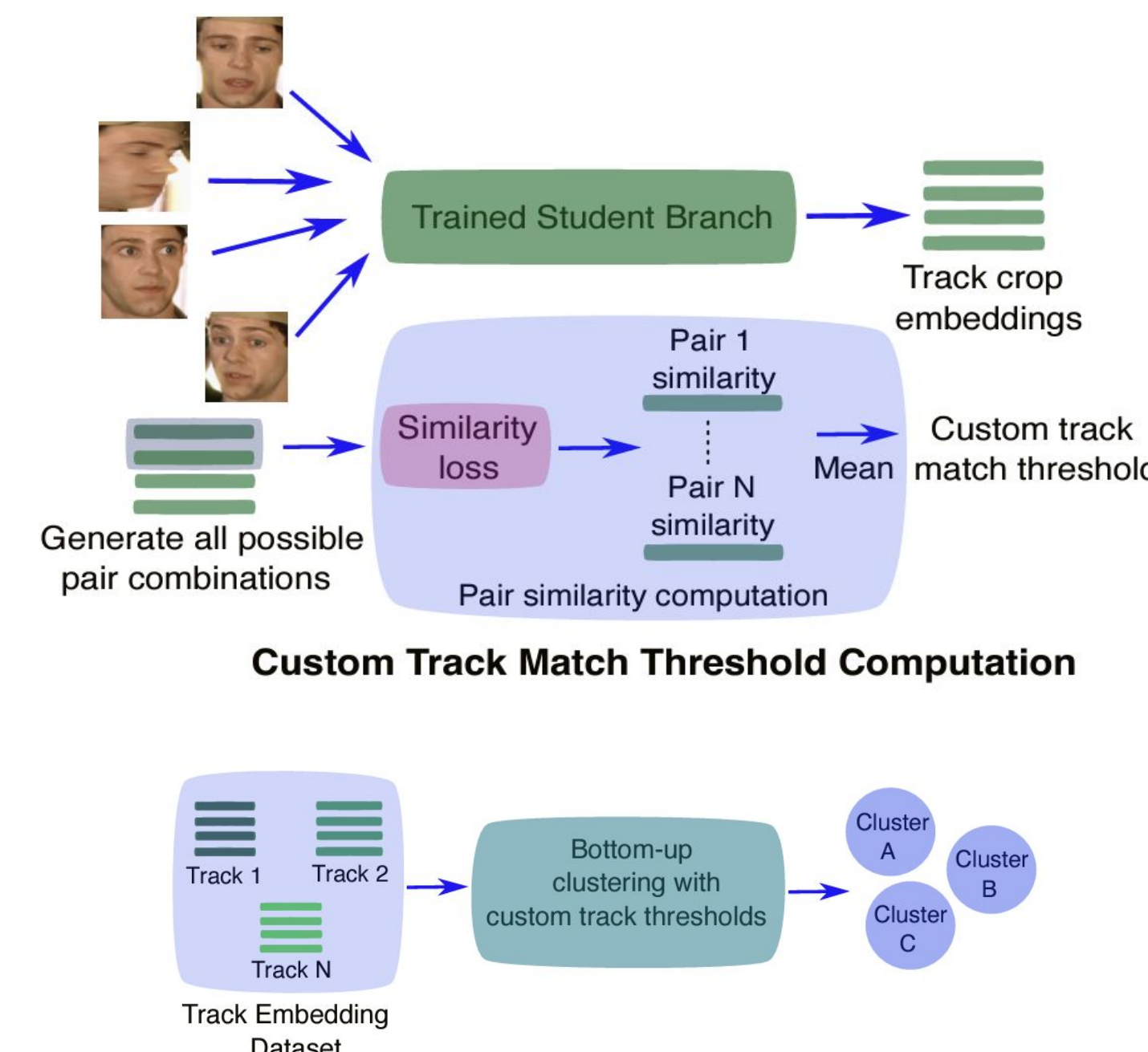
Contributions

- A **fully self-supervised video face clustering algorithm**, which progressively learns robust identity embeddings for all faces within a given video face dataset.
- A **self-supervised model finetuning approach** that removes any dependence on manual ground truth cluster labels.
- A **deep learning-based similarity metric for face clustering**, which automatically adapts to a given model's learned embedding space.
- A **novel video face clustering algorithm** that does not depend on any user-input parameters.
- **Release of a novel video face clustering benchmark dataset** with extreme challenging face clustering scenarios in movie domain.

Central Idea



Autonomous Face Clustering



Algorithm 1: Face Track Clustering

Input: Filtered Face Tracks
 $T = \{t_j | j = 1, 2, \dots, N\}$
 $\exists t_j = \{I_{t_1}, I_{t_2}, \dots, I_{t_n} | t_n = 12 * n + f_1\}$,
 finetuned model θ_{ft} , Similarity metric S

Stage 3.1 (Compute track crop embedding set):
 for t_j in T do
 for I_{t_n} in t_j do
 $E_{t_n} \leftarrow \theta_{ft}(I_{t_n})$
 end
 $t_{jE} = \{E_{t_1}, E_{t_2}, \dots, E_{t_n}\}$
 end
 $T_E = \{t_{1E}, t_{2E}, \dots, t_{NE}\}$

Stage 3.2 (Compute custom track threshold):
 for t_{jE} in T_E do
 $Sim_{t_{jE}} = \{S(E_{t_1}, E_{t_m}) | (l, m) \in {}^n C_2\}$
 $Thresh_{t_j} \leftarrow \text{mean}(Sim_{t_{jE}})$
 end
 $Threshes = \{Thresh_{t_1}, \dots, Thresh_{t_N}\}$

Stage 3.3 (Perform track clustering):
 Initialize $i = 0$, $C = \{C_j = \{t_{jE} | j = 1, 2, \dots, N\}\}$
 repeat
 for C_j in C do
 for C_k in C if $k \neq j$ do
 $Sim_{jk} \leftarrow \text{mean}(\{S(t_{jE_n}, t_{kE_k}) | \forall (a, b) \in {}^n C_2\})$
 if $Sim_{jk} < Thresh_{t_j} / Thresh_{t_k}$ then
 $C_j \leftarrow \text{merge}(C_j, C_k)$
 end
 end
 end
 $C = \text{link_merges}(C)$
 $N_{C_i} = \text{cluster_count}(C)$
 Repeat Stage 2 for new merged cluster set C
 $i = i + 1$
 until $N_{C_i} - N_{C_{i-1}} = 0$;
Output: Clustered track IDs C

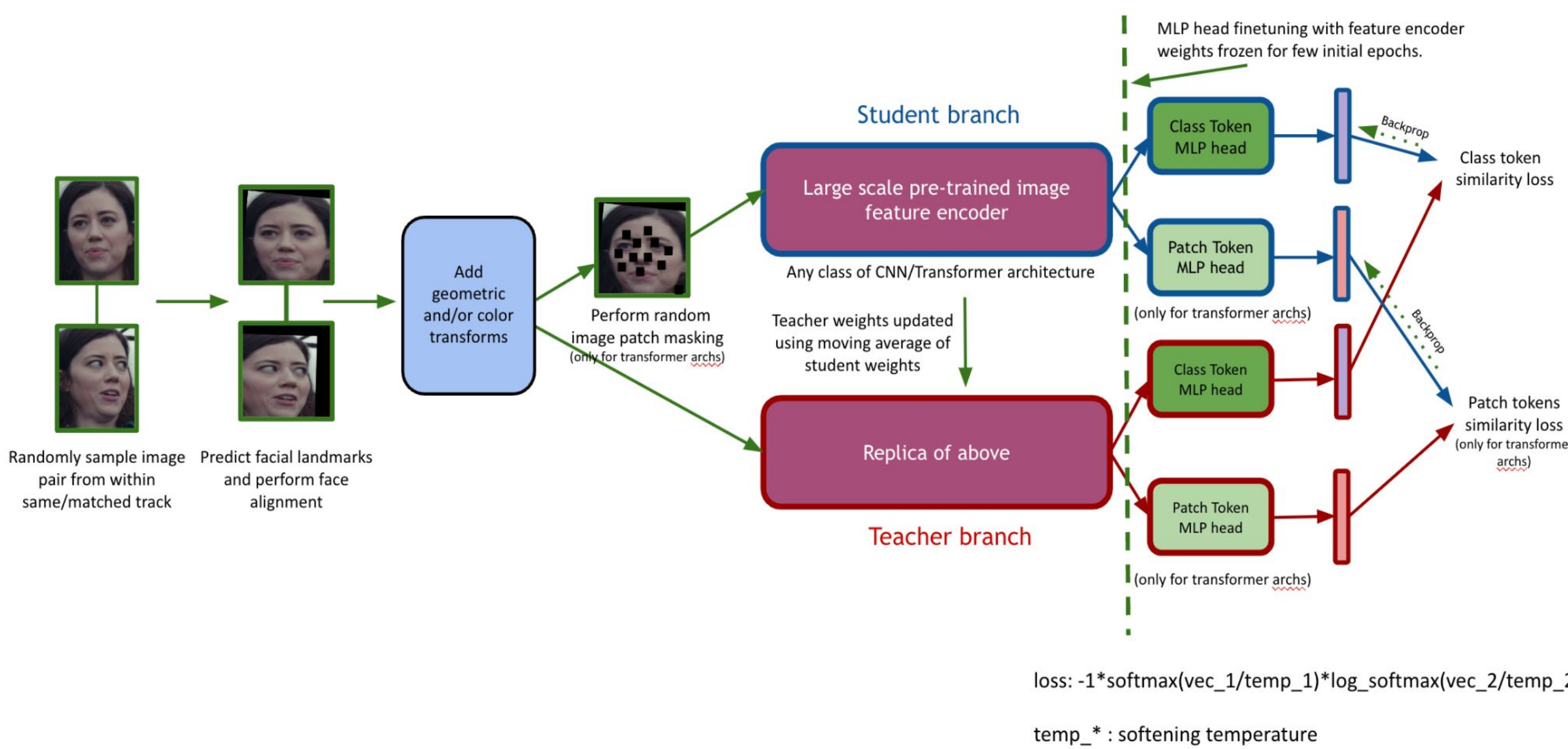
Character Face Clustering in Movies

Objective: Given an entire movie sequence, cluster main character face tracks across common facial identities.

Movie domain specific challenges: Extreme variations across character face pose, lighting conditions, heavy occlusion, blur and facial appearance.



Self-Supervised Face ID Model Finetuning

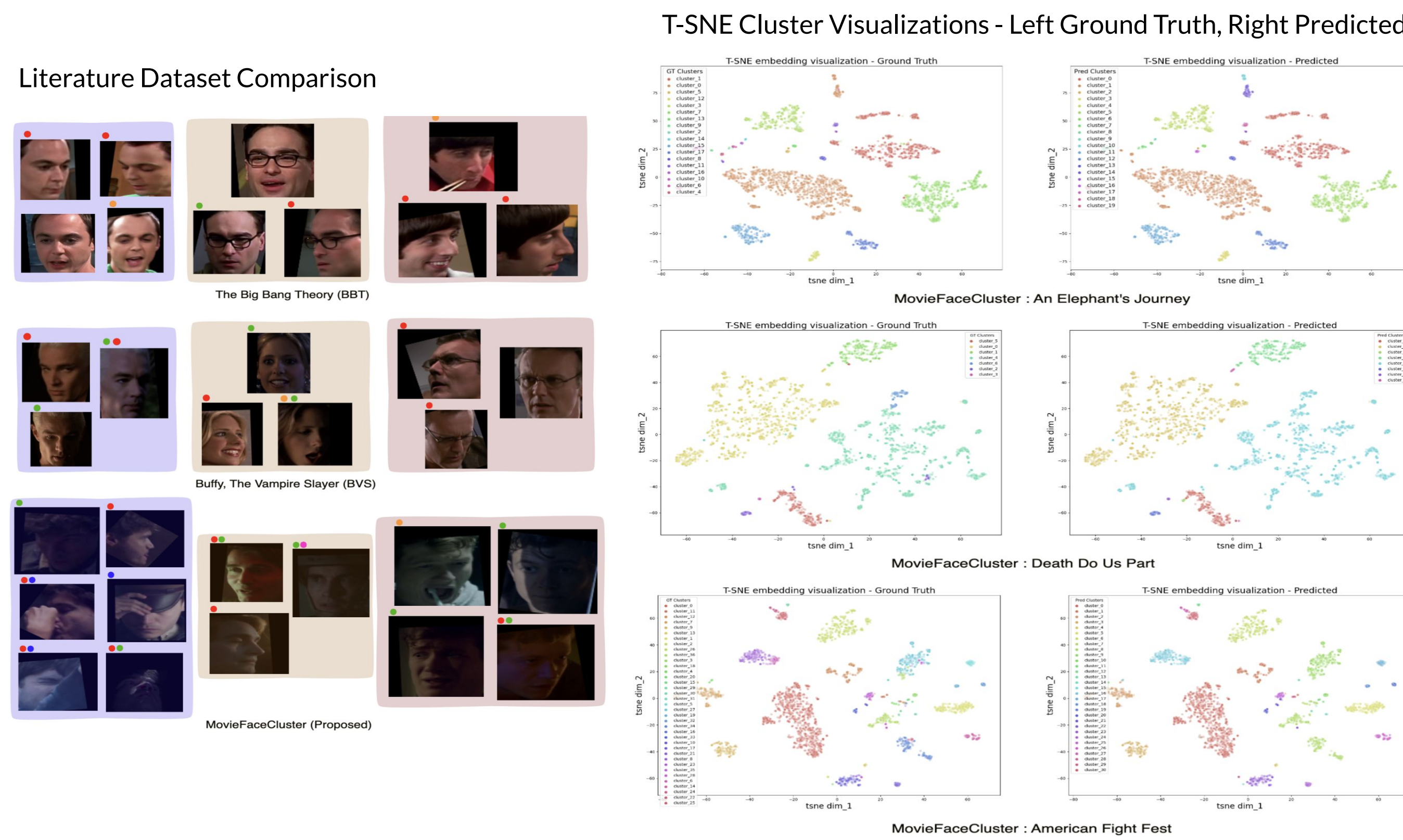


Results

Literature Datasets: Big Bang Theory S01 & Buffy The Vampire Slayer S05

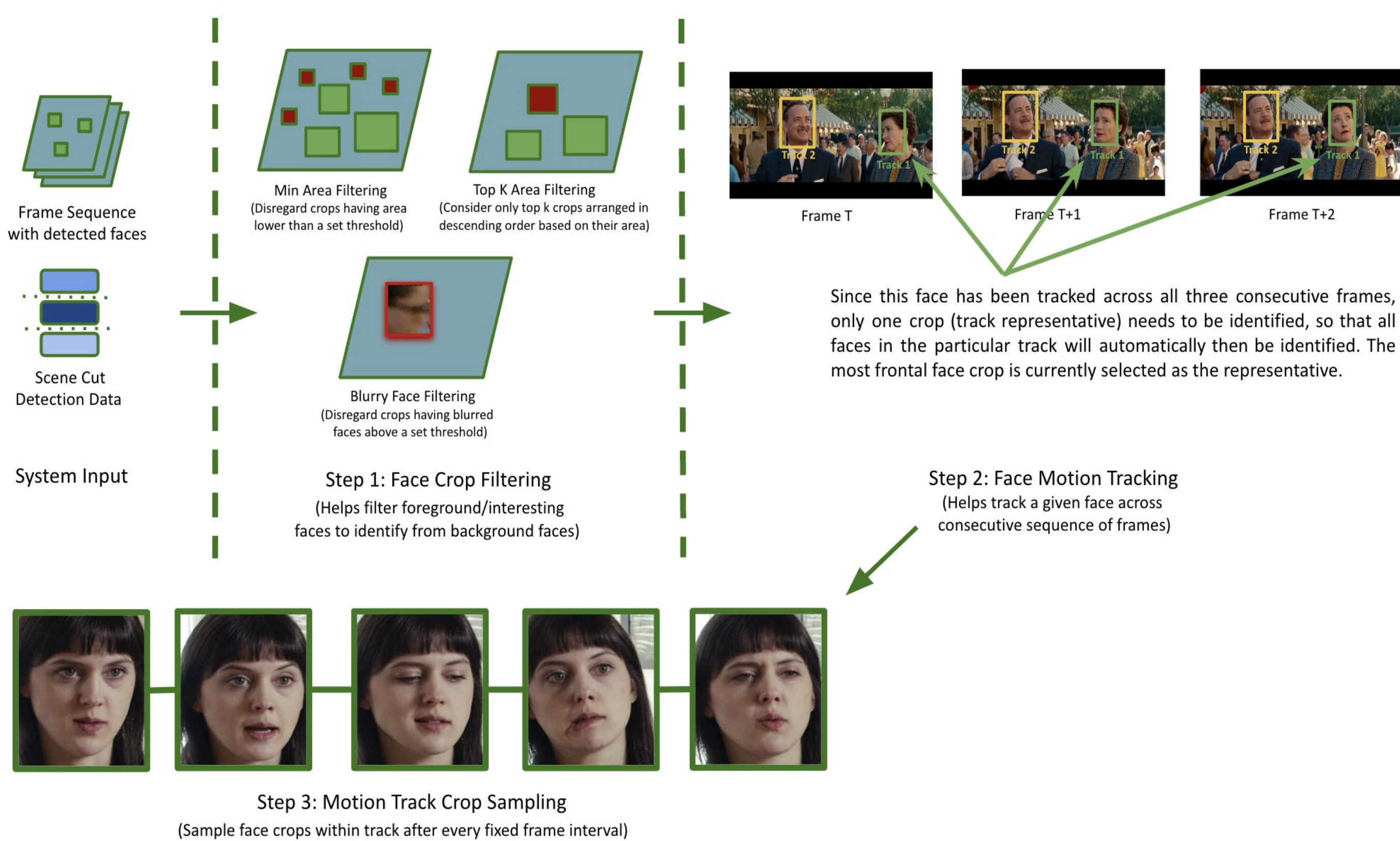
Method	BBT S01 Episode						Combined		BVS S05 Episode						Combined
	S1E1	S1E2	S1E3	S1E4	S1E5	S1E6			S5E1	S5E2	S5E3	S5E4	S5E5	S5E6	
SCTL [54]	66.48	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TSiam [41]	96.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TSiam [41]	96.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SSiam [41]	96.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-
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	-	96.30	99.10	98.70	97.43	99.00	96.78	96.10

Release of MovieFaceCluster Dataset



Method	An Elephant's Journey (2019)	Armed Response	Angel Of The Skies	Death Do Us Part (2019)	Movie American Fight Fest	The Fortress	Under The Shadow	The Hidden Soldier	S.M.A.R.T. Chase
	Weighted Cluster Accuracy (%)	Weighted Cluster Accuracy (%)	Weighted Cluster Accuracy (%)	Weighted Cluster Accuracy (%)	Weighted Cluster Accuracy (%)	Weighted Cluster Accuracy (%)	Weighted Cluster Accuracy (%)	Weighted Cluster Accuracy (%)	Weighted Cluster Accuracy (%)
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.†	94.4 & 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

Face Track Preprocessing



Coarse Track Matching

