Audio-Visual Representation Learning For Lip-Sync Estimation Through Ranking Augmented Contrastive Training

Bhavin Jawade, Ravi Teja Gadde, Christophe Bejjani , Yinghong Lan

NETFLIX Research

2025 icassp HYDERABAD, INDIA Celebrating Signal Processing

Problem Statement

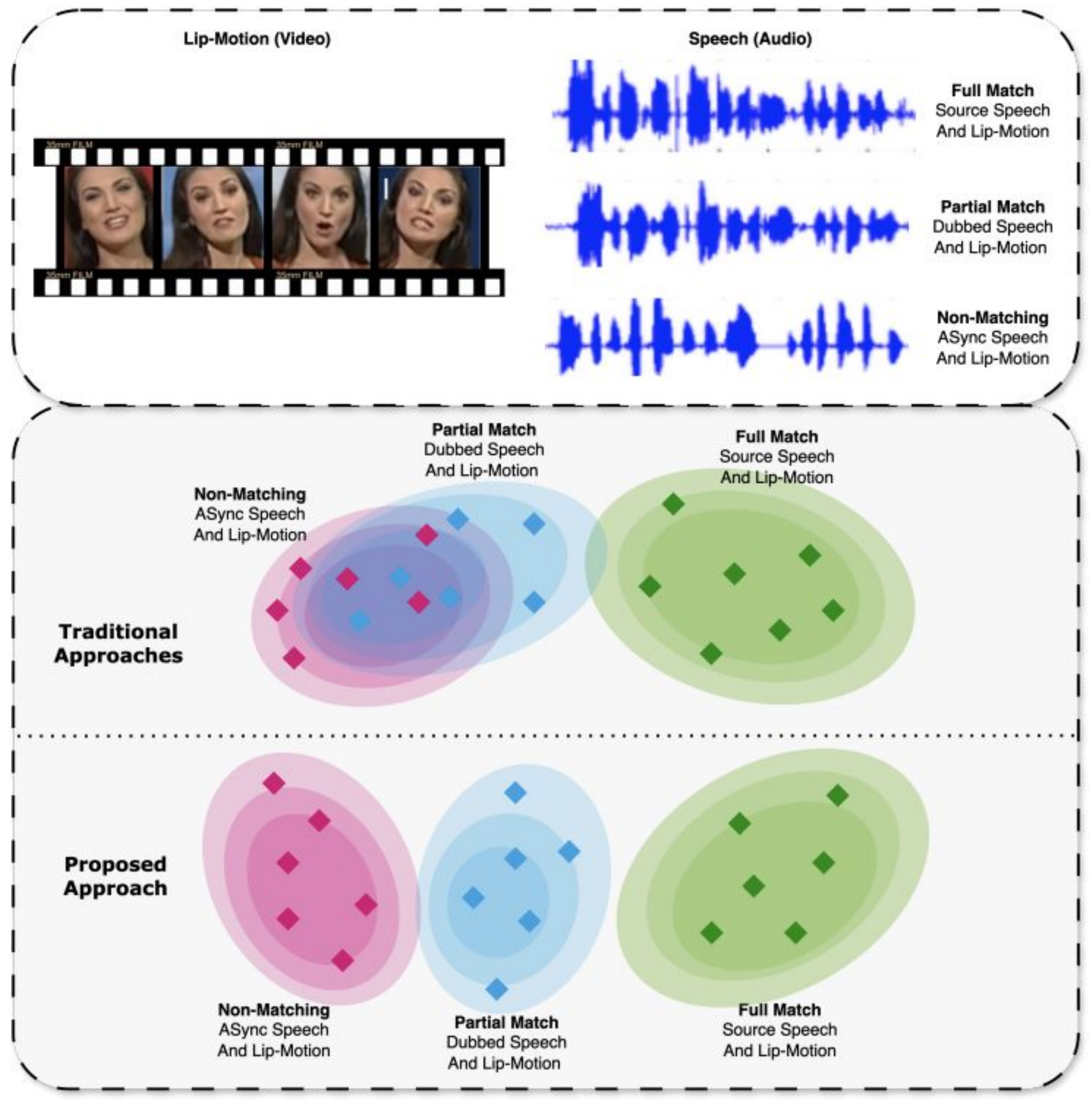
- ★ Estimate quality of lip-sync between a Video (Lip-Motion) and Audio - (Speech)

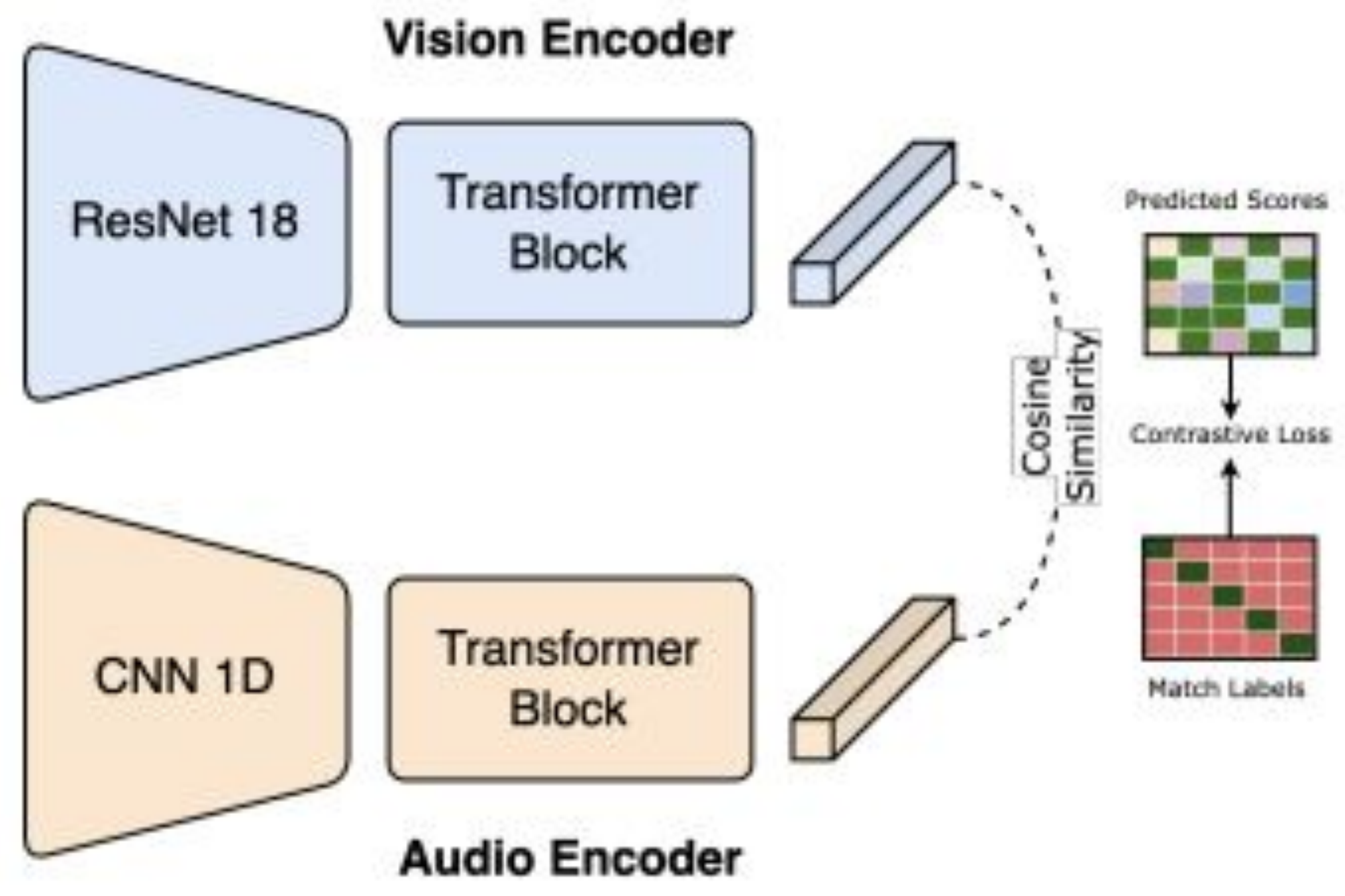
Traditional Lip-Sync Models

- ★ Trained contrastively to align perfectly synced audio-videos.
- ★ Push apart all other forms of audios.
- ★ Effective at distinguishing perfect-sync from imperfect sync.

Limitations

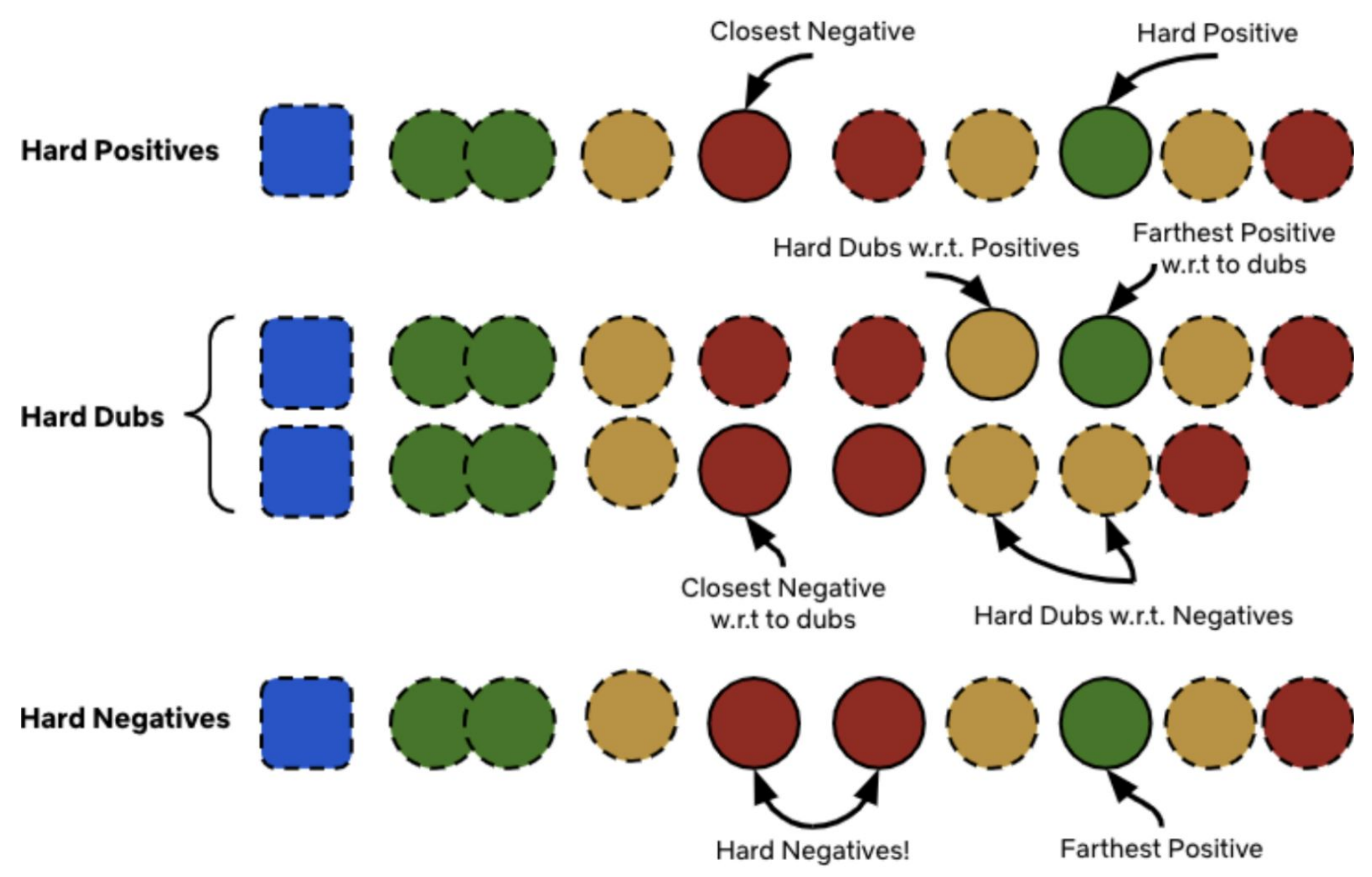
- ★ Ineffective at detecting degrees of sync
- ★ Unable to effectively rank dubbed content based on lip-sync





Model Architecture

Method


Closest Negative
Hard Positive
Hard Positives
Hard Dubs w.r.t. Positives
Farthest Positive w.r.t to dubs
Hard Dubs
Closest Negative w.r.t to dubs
Hard Dubs w.r.t. Negatives
Hard Negatives
Hard Negatives!
Farthest Positive

Ranking Supervised Multi-Similarity (RSMS)

Hard positive audios: Positive audio samples that are closer to the nearest negative or nearest dubbed audio within a margin for the given video

$$\hat{S}_i^p = \{p \,|\, S(v_i, p) - \lambda < \max(S(v_i, n), S(v_i, d))\}$$

Hard negative audios: Negative audio samples that are closer than the farthest positive or farthest dubbed audio within a margin for the given video.

$$\hat{S}_i^n = \{n \,|\, S(v_i, n) + \lambda > \min(S(v_i, p), S(v_i, d))\}$$

Hard dubbed audios with respect to positives: Dubbed audio samples that are closer than the farthest positive audios

$$\hat{S}_i^{nr} = \{d \,|\, S(v_i, d) + \lambda_d > \min(S(v_i, p))\}$$

Hard dubbed audios with respect to negatives: Dubbed audio samples that are farther away from the closest negative audio

$$\hat{S}_i^{pr} = \{d \,|\, S(v_i, d) - \lambda_d < \max(S(v_i, n))\}$$

$$L_i^p = \log\left(1 + \sum e^{-\alpha(\hat{S}_i^p - \sigma)}\right) L_i^n = \log\left(1 + \sum e^{\beta(\hat{S}_i^n - \sigma)}\right)$$

$$L_i^{pr} = \log\left(1 + \sum e^{-\gamma(\hat{S}_i^{pr} - \sigma)}\right) L_i^{np} = \log\left(1 + \sum e^{\delta(\hat{S}_i^{nr} - \sigma)}\right)$$

$$L_{RSMS} = \frac{1}{B} \sum_{i=1}^{B} \left\{ \frac{1}{\alpha} \cdot L_i^p + \frac{1}{\beta} \cdot L_i^n + \frac{1}{\gamma} \cdot L_i^{pr} + \frac{1}{\delta} \cdot L_i^{np} \right\}$$

Datasets

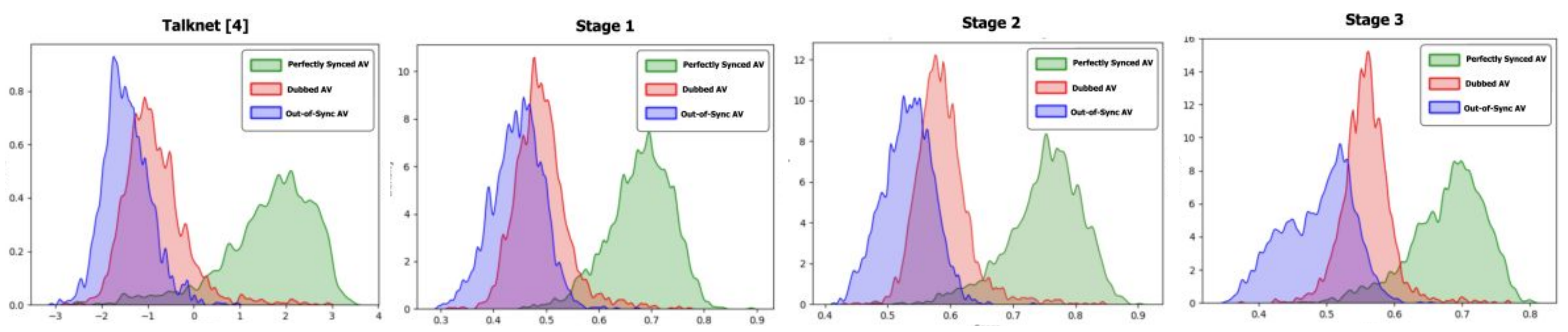
- ★ **Training Dataset:** For **pretraining** - VoxCeleb dataset, consisting of 862,885 videos (1,868 hours) widely employed in speaker and face recognition tasks. For **fine-tuning**, we used a 37-hour internal Netflix dataset with 66,847 videos featuring real-world partial-syncs from dubbed videos in multiple languages.
- ★ **Evaluation Dataset:** Curated disjoint set of 5,742 Netflix videos
  (i) 1,900 videos with non-matching audios (outof-syncs)
  (ii) 1,900 videos with **dubbed audios** (partial-syncs)
  (iii) 1,900 videos with original audios (perfect-syncs).
  The dataset spans 10 languages, including *English, Brazilian Portuguese, German, Spanish, French, Hindi, Italian, Japanese, Russian, and Turkish*

Comparison To SoTA

Zero-shot performance on Netflix Dubbed Content

Method	Source Vs Dubs (S/D)		Source Vs Out-of-syncs (S/O)		Dubs Vs Out-of-syncs (D/O)	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
SyncNet	70.48%	75.52%	71.38%	76.83%	58.16%	60.81%
VocaLiST	75.29%	79.70%	91.32%	94.20%	60.29%	72.28%
MTDVocaLiST	77.38%	83.48%	93.55%	96.72%	64.55%	74.76%
TalkNet	92.78%	95.61%	96.20%	98.43%	70.89%	77.69%
Ours	93.85%	97.68%	98.28%	99.73%	74.92%	82.73%

Density plot of score distribution for positives, dubs and negatives on our evaluation dataset



Comparison (AUC) of contrastive losses for lip-sync ranking

Method	Rank Priors	S / D	S / O	D / O
InfoNCE	✗	93.22	95.48	67.4
MultiSimilarity	✗	95.38	98.29	74.64
RINCE	✓	95.47	98.35	76.95
RSMS (Ours)	✓	97.68	99.73	82.73

Does Fine-tuning on Real-Dubs Help?

Method	Dataset	D/O	
		Acc	AUC
Zero-Shot	Synthetic Shifted-Sync	74.92%	82.73%
Fine-Tuned	Real Dubbed Audio	81.31%	88.60%

Selected References

[TalkNet] - R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, “Is someone speaking? exploring long-term temporal features for audiovisual active speaker detection,” in Proceedings of the 29th ACM international conference on multimedia
[RINCE] - D. T. Hoffmann, N. Behrmann, J. Gall, T. Brox, and M. Noroozi, “Ranking info noise contrastive estimation: Boosting contrastive learning via ranked positives,” in Proceedings of the AAAI Conference on Artificial Intelligence
[SyncNet] - J. S. Chung and A. Zisserman, “Out of time: Automated lip sync in the wild,” in Computer Vision – ACCV 2016 Workshops (C.-S. Chen, J. Lu, and K.-K. Ma, eds.), (Cham)