Disjoint Mapping Network for Cross-Modal Matching of Voices and Faces

Yandong Wen[1], Mahmoud Al Ismail[1], Weiyang Liu[2], Bhiksha Raj[1], Rita Singh[1]
[1]Carnegie Mellon University, [2]Georgia Institute of Technology

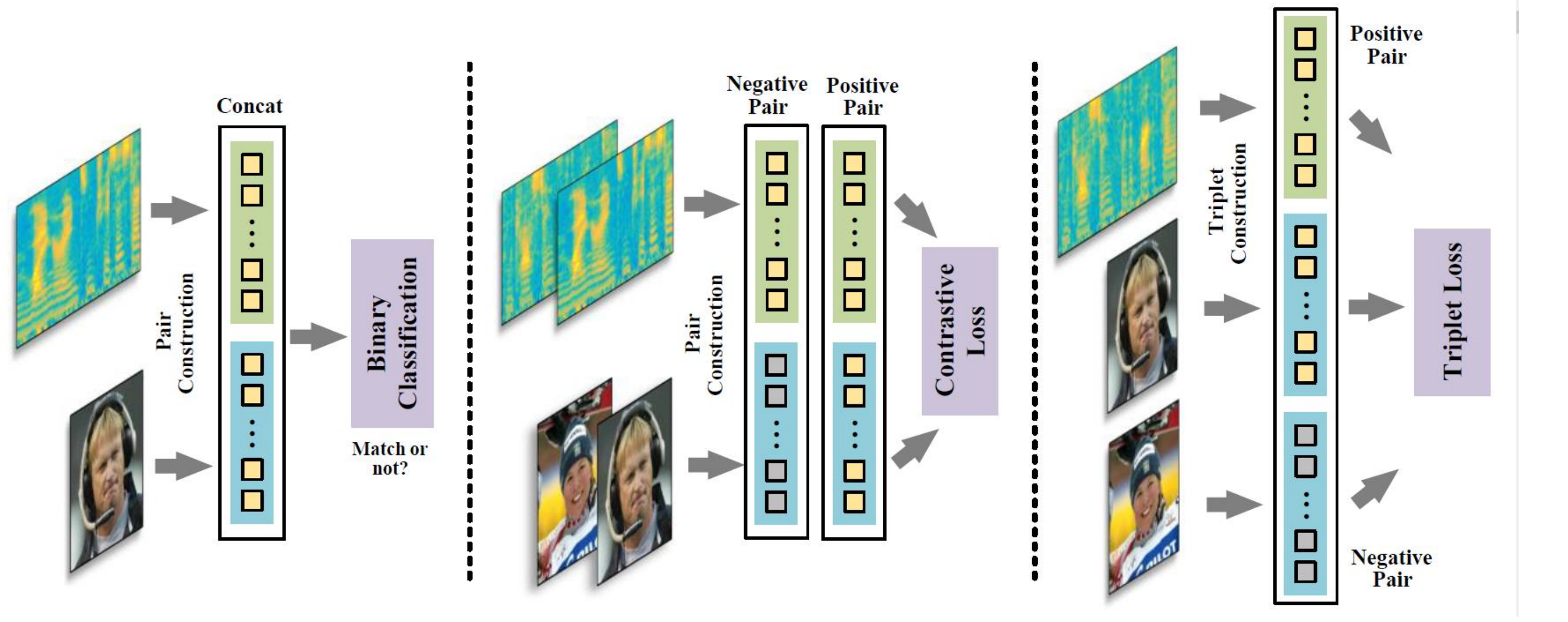
Task: voices and faces matching

voice to face   face to voice

Motivation

- Biological influences (genetic, physical, and/or environment) affect the face as well as the voice
- Humans associate voices of unknown individuals to pictures of their faces
- Explore the possibility of finding the association between voices and faces algorithmically through covariates
- Enable cross-modal profiling

Related Works



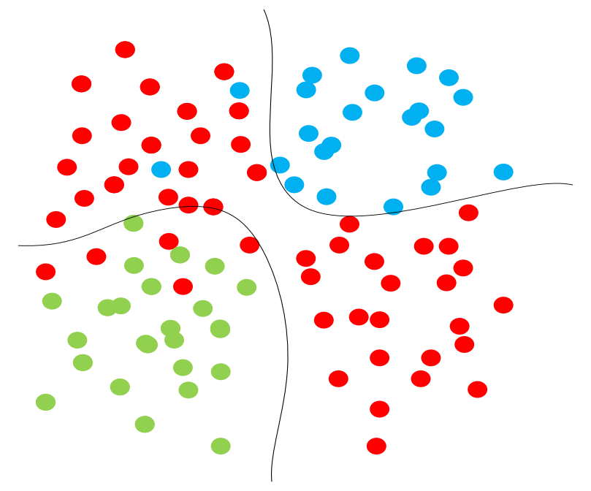
(a) Seeing Faces and Hearing Voices (b) Learnable PINs (c) Learning Face-Voice Association

- Covariates are implicitly used. Only one covariate can be used for data construction
- Pair or triplet construction is needed

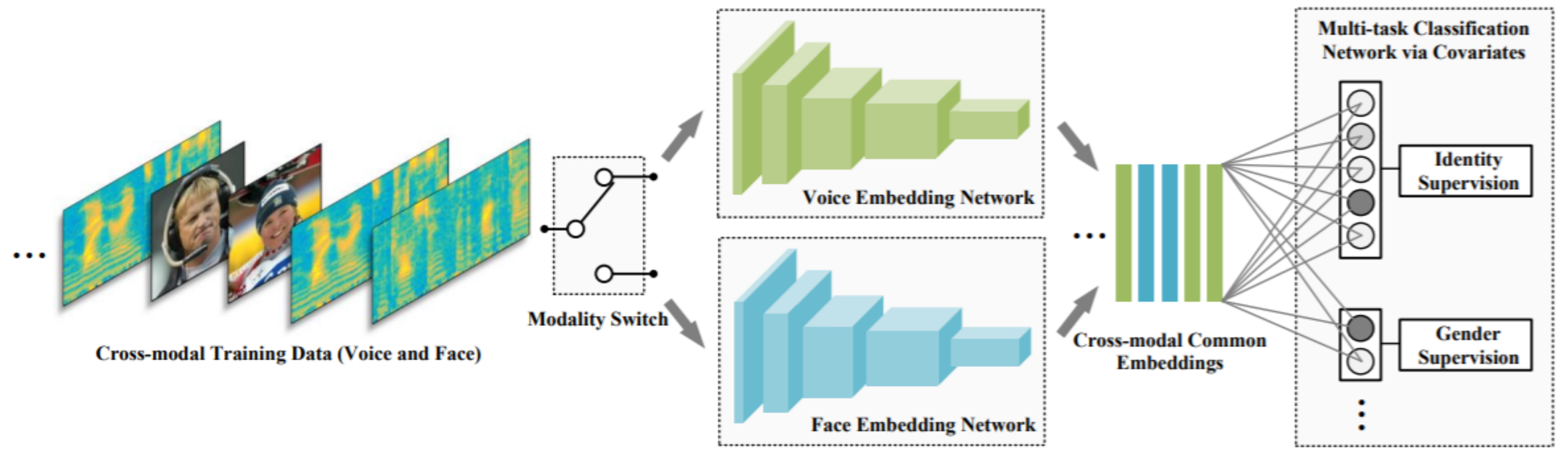
Disjoint Mapping Networks (DIMNets)

The key ideas:

A joint classifier produces common decision regions for both voice and face embeddings.



The proposed framework:


Cross-modal Training Data (Voice and Face)
Modality Switch
Voice Embedding Network
Face Embedding Network
Cross-modal Common Embeddings
Multi-task Classification Network via Covariates
Identity Supervision
Gender Supervision

Individual supervision: DIMNet formulates the problem as learning common embeddings for the two through individual supervision.
Multiple covariates: multiple classifiers can be learned by different covariates.
Ablation study: the effect of the individual covariate on the performance can be easily isolated and analyzed.

Dataset

# of samples	train	validation	test	total
speech segments	112,697	14,160	21,799	148,656
face images	313,593	36,716	58,420	408,729
IDs	924	112	189	1,225
genders	2	2	2	2
nationalities	32	11	18	36
testing instances	-	4,678,897	6,780,750	11,459,647

Statistics for the data appearing in `VoxCeleb` and `VGGFace`

Evaluation

- **1:N Matching**. We are given a probe input from one modality, and a gallery of N inputs from the other modality, including one that belongs to the same subject as the probe, and N-1 "imposter". The task is to identify which entry in the gallery matches the probe
- **Verification**. We are given two inputs, one a face, and another a voice. The task is to determine if they belong to the same subject.
- **Retrieval**. The gallery comprises a large number of instances, one or more of which might match the probe. The task is to order the gallery such that the entries in the gallery that match the probe lie at the top of the ordering.

Results

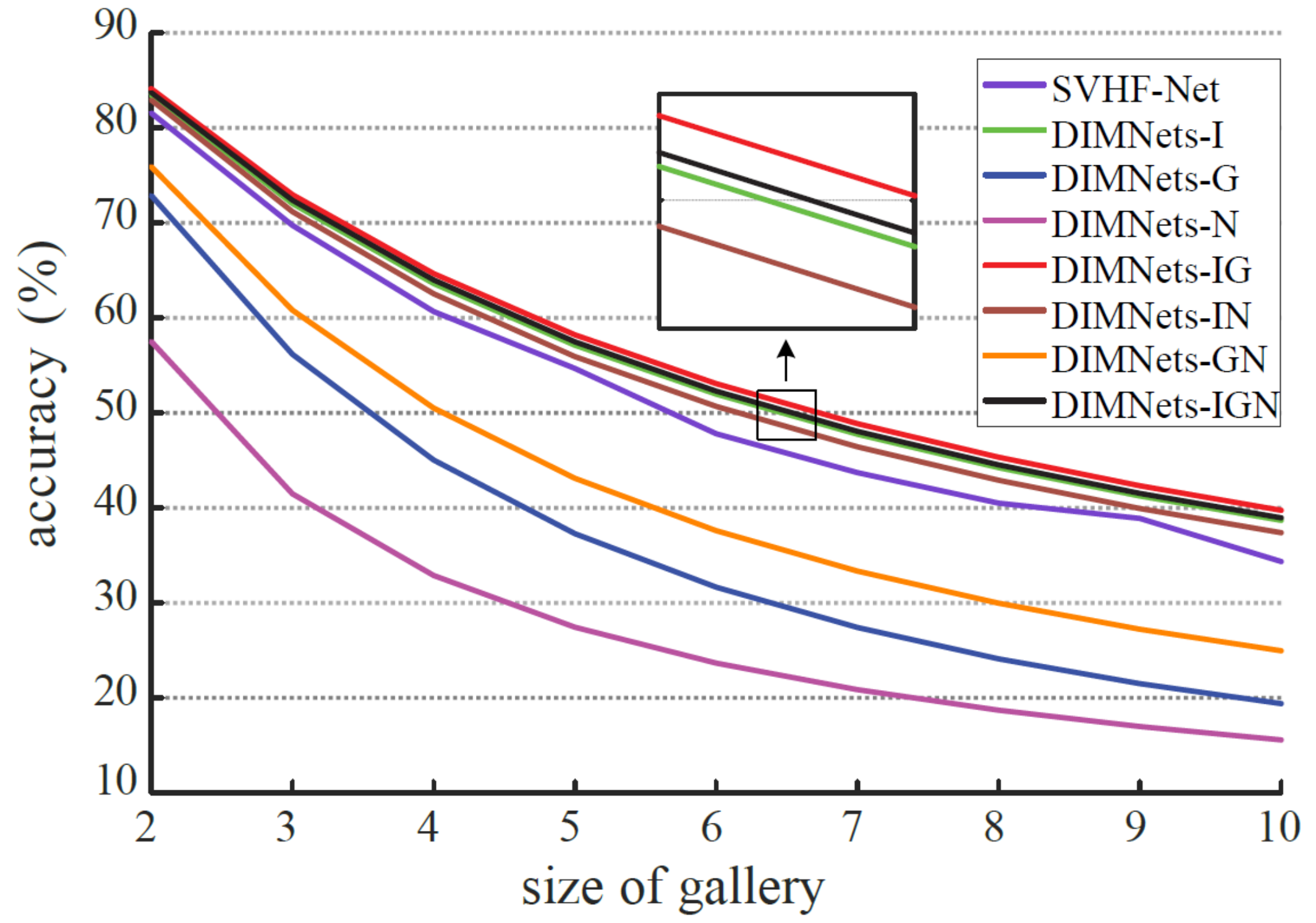
1:2 matching

method	voice → face (ACC %) U	G	N	G, N
SVHF-Net	81.00	63.90	-	-
DIMNet-I	83.45± 0.42	70.91±0.56	81.97±0.51	69.89±0.78
DIMNet-G	72.90±0.55	50.32±0.70	71.92±0.51	50.21±0.65
DIMNet-N	57.53±0.45	55.33±0.67	53.04±0.43	51.96±0.59
DIMNet-IG	**84.12**±0.44	**71.32**±0.60	**82.65**±0.57	**70.39**±0.80
DIMNet-IN	82.95±0.40	70.04±0.67	81.04±0.55	68.59±0.76
DIMNet-GN	75.92±0.42	56.66±0.55	72.94±0.48	53.48±0.73
DIMNet-IGN	83.73±0.53	70.76±0.34	81.75±0.48	69.17±0.71

method	face → voice (ACC %) U	G	N	G, N
SVHF-Net	79.50	63.40	-	-
DIMNet-I	83.52±0.45	71.78±0.55	82.41±0.48	**70.90**±0.81
DIMNet-G	72.47±0.54	50.48±0.71	72.15±0.54	50.61±0.68
DIMNet-N	56.20±0.43	54.34±0.61	53.90±0.44	51.97±0.57
DIMNet-IG	**84.03**±0.39	**71.65**±0.60	**82.96**±0.49	70.78±0.47
DIMNet-IN	82.86±0.35	70.91±0.59	81.91±0.52	70.22±0.77
DIMNet-GN	73.78±0.69	54.90±0.54	72.63±0.48	53.45±0.85
DIMNet-IGN	83.63±0.66	71.42±0.49	82.50±0.43	70.46±0.62

Results (continues.)

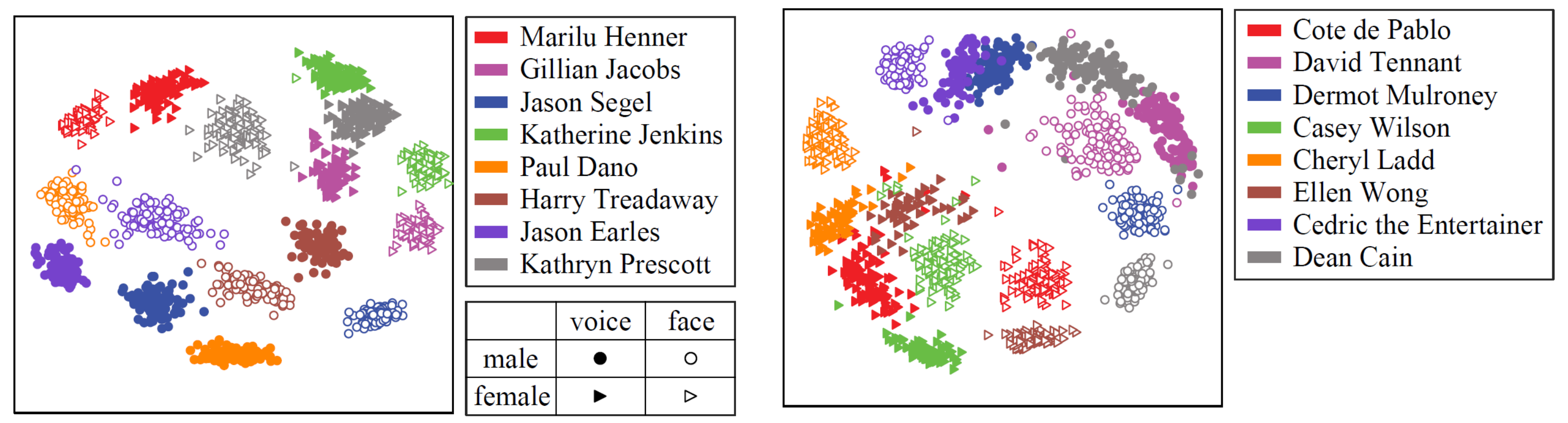
1:N matching



Verification

method	verification (EER %) U	G	N	G, N
DIMNet-I	24.95±0.20	34.95±0.45	25.92±0.68	35.74±0.87
DIMNet-G	34.86±0.11	49.69±0.24	35.13±0.36	49.67±0.51
DIMNet-N	45.89±0.39	46.97±0.55	47.89±0.82	48.87±1.14
DIMNet-IG	**24.56**±0.23	**34.84**±0.41	**25.54**±0.65	**35.73**±0.79
DIMNet-IN	25.54±0.18	36.22±0.40	27.25±0.72	37.39±0.79
DIMNet-GN	33.28±0.52	46.65±0.16	34.77±0.26	48.08±0.52
DIMNet-IGN	25.00±0.19	35.76±0.36	26.80±0.69	37.30±0.74

Embedding visualization



Multi-dimensional scaling (MDS) is adopted for the visualization of voice and face embeddings, since it tends to preserve distances and global structure.