BOSTON UNIVERSITY

Dartmouth

Unimodal Face Classification with Multimodal Training

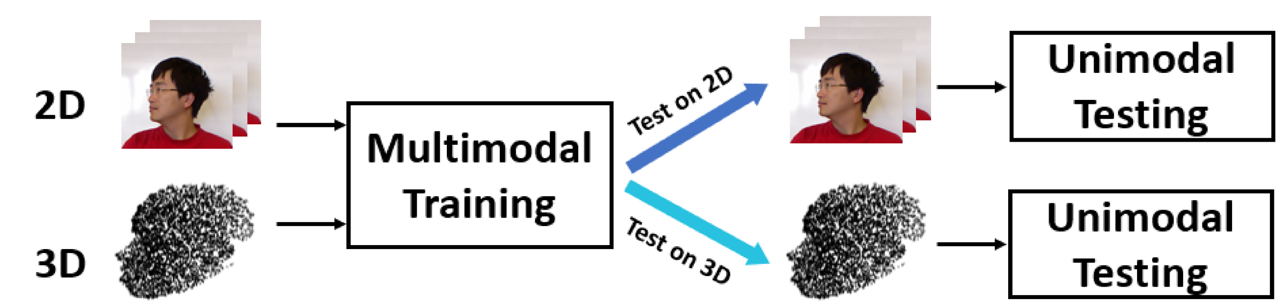
Wenbin Teng[1] Chongyang Bai[2]

[1] Boston University [2] Dartmouth College

FG 2021
Jodhpur, India
December 15 - 18, 2021

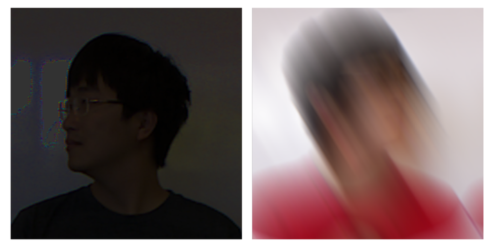
Motivation

Goal: Train classification model with both 2D and 3D face data and test with single modality.

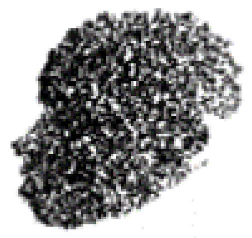


Why Multimodal Training Unimodal Testing (MTUT)?

(a) Original RGB



(b) Low quality RGB



(c) Point cloud

- **MT**: Low quality of RGB image/point cloud lack of texture features.
- **UT**: Not both modalities are available in practice.

Contributions

- Propose MTUT framework for face classification.
- Establish cross-modal autoencoders to learn embeddings containing information of both the available and missing modalities during test.
- Develop adaptive embedding divergence (AED) loss to avoid interference from any potential noisy modality.

Cross-modal Autoencoder

Encoder. We encode 2D RGB images $\mathbf{I}$ into $\mathbf{x}^I \in \mathbb{R}$ with ResNet-18 [1] and encode 3D point clouds $\mathbf{P}$ into $\mathbf{x}^P \in \mathbb{R}$ with PointNet [2]. Encoded features are concatenated with face attribute vector $\mathbf{a} \in \mathbb{R}$:

$$\mathbf{x}'^I = f(enc^I(\mathbf{I}), \mathbf{a}), \ \mathbf{x}'^P = f(enc^P(\mathbf{P}), \mathbf{a}) \tag{1}$$

where f is a fully-connected neural network.

Decoder. There are two cases based on the available modality during test and we perform optimization of reconstruction loss:

- **Case 1: 3D available and 2D missing.** Both $\mathbf{x}'^I$ and $\mathbf{x}'^P$ are decoded into $\hat{\mathbf{I}}$ and $\hat{\mathbf{P}}$ to reconstruct $\mathbf{I}$:

$$\hat{\mathbf{I}} = dec^I(\mathbf{x}'^I), \ \hat{\mathbf{P}} = dec^I(\mathbf{x}'^P), \ \mathcal{L}^I_{RE} = \left\|\hat{\mathbf{I}} - \mathbf{I}\right\|_2 + \left\|\hat{\mathbf{P}} - \mathbf{I}\right\|_2 \tag{2}$$

- **Case 2: 2D available and 3D missing.** Both $\mathbf{x}'^I$ and $\mathbf{x}'^P$ are decoded into $\hat{\mathbf{I}}$ and $\hat{\mathbf{P}}$ to reconstruct $\mathbf{P}$:

$$\hat{\mathbf{I}} = dec^P(\mathbf{x}'^I), \ \hat{\mathbf{P}} = dec^P(\mathbf{x}'^P), \ \mathcal{L}^P_{RE} = \left\|\hat{\mathbf{I}} - \mathbf{P}\right\|_2 + \left\|\hat{\mathbf{P}} - \mathbf{P}\right\|_2 \tag{3}$$

Cross-modal Autoencoder (cont'd)

3D Autoencoder. According to [2], PointNet extracts a 1024-dimension node embedding for each vertex $v \in \mathbf{P}$. The global point features are aggregated by a max pooling operator:

$$\mathbf{x}^P = enc^P(\mathbf{P}) = \max_{i=1...n} h(\mathbf{x}_i) \tag{4}$$

where h is a set of graph convolutional neural networks. The max pooling operation is reversed by applying a Gaussian sampling with the maximum value is set as $\mathbf{x}^P$:

$$\hat{\mathbf{P}}^0_{i,k} = \min\left\{\mathcal{N}(0,1), \mathbf{x}^P_k\right\}, \ \text{where } k = 1, 2, ..., 1024 \tag{5}$$

Adaptive Embedding Divergence Loss

Assume $\mathcal{M}$ is the missing modality and $\mathcal{A}$ is the available modality during test mode, where $\mathcal{M}, \mathcal{A} \in \{\mathbf{I}, \mathbf{P}\}$. The AED loss is defined as:

$$\mathcal{L}_{AED} = \rho\left\|\mathbf{x}^\mathcal{M} - \mathbf{x}^\mathcal{A}\right\|_2, \ \text{where } \rho^\mathcal{M} = \begin{cases} e^{\beta\Delta_\mathcal{M}\mathcal{L}} - 1, & \Delta_\mathcal{M}\mathcal{L} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\Delta_\mathcal{M} = \mathcal{L}^\mathcal{M}_{cls} - \mathcal{L}^\mathcal{A}_{cls}$ is the difference between loss of classification and $\beta > 0$ is a hyper-parameter controlling the impact from the loss difference.

Objective Function

The objective function is similarly separated by two cases based on the availability of testing modality:

- **Case 1: 3D available and 2D missing.** $\mathbf{x}^P$ is used for classification.

$$\mathcal{L} = \mathcal{L}^P_{cls} + \lambda_1\mathcal{L}^I_{RE} + \lambda_2\mathcal{L}_{AED}, \ \text{where } \mathcal{L}^P_{cls} = -\mathbb{E}_{\mathbf{x}^P}\log\mathbb{P}(\mathbf{y}|\mathbf{x}^P) \tag{7}$$

- **Case 2: 2D available and 3D missing.** $\mathbf{x}^I$ is used for classification.

$$\mathcal{L} = \mathcal{L}^I_{cls} + \lambda_1\mathcal{L}^P_{RE} + \lambda_2\mathcal{L}_{AED}, \ \text{where } \mathcal{L}^I_{cls} = -\mathbb{E}_{\mathbf{x}^I}\log\mathbb{P}(\mathbf{y}|\mathbf{x}^I) \tag{8}$$

End-to-End Architecture

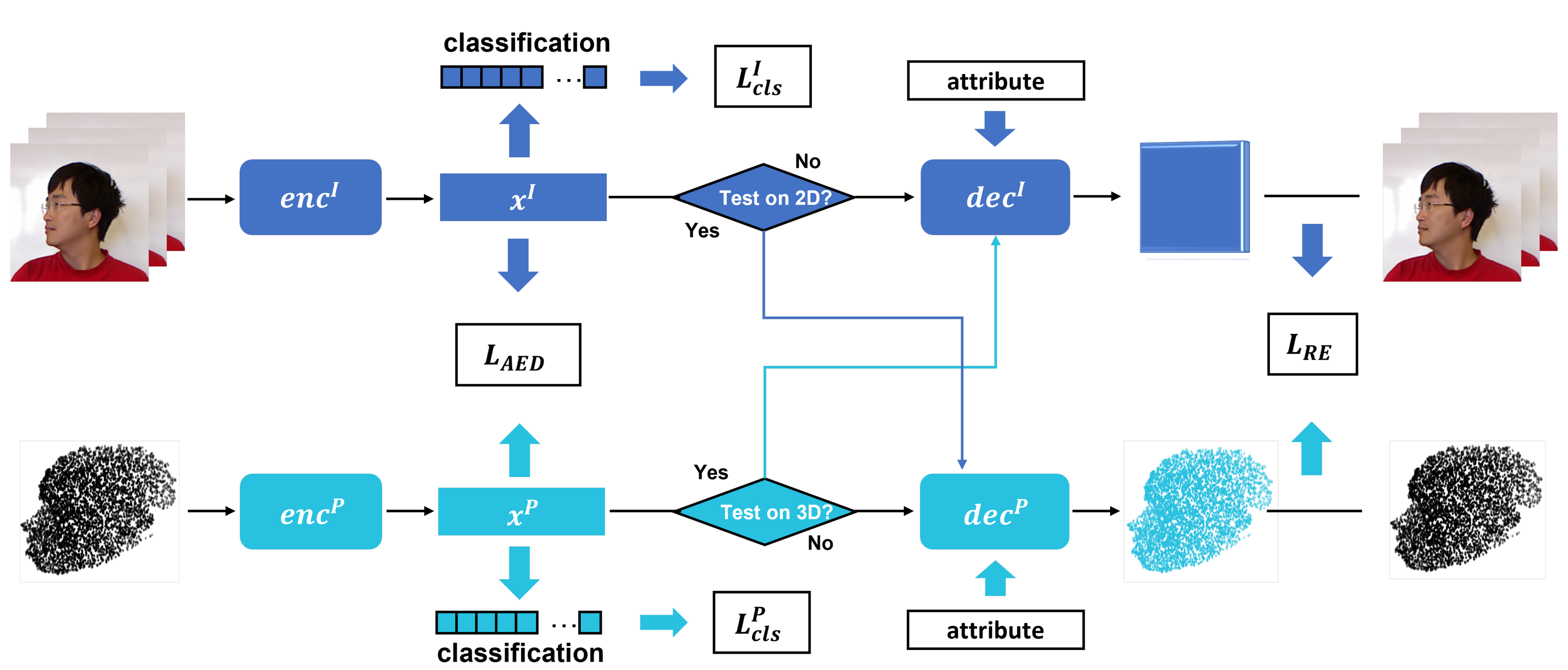


Figure: An overview of proposed MTUT face recognition framework.

Results

Test Modality	Method	Kinect		CASIA	
		Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)
2D	VGG-11	81.04	80.84	90.87	91.05
	VGG-11+MTUT	**84.29**	**82.92**	**92.00**	**92.28**
	ResNet-18	94.87	94.65	94.70	94.55
	ResNet-18+MTUT	**97.43**	**97.17**	**95.24**	**94.87**
	FaceNet	90.66	90.56	91.57	91.43
	FaceNet+MTUT	**93.12**	**92.97**	**93.42**	**93.28**
	DeepID	63.19	63.19	76.54	76.12
	DeepID+MTUT	**67.50**	**67.49**	**78.24**	**77.82**
	DeepFace	75.00	74.55	74.38	73.49
	DeepFace+MTUT	**80.19**	**79.21**	**75.45**	**74.18**
3D	PointNet	79.49	79.17	82.49	81.31
	PointNet+MTUT	**86.58**	**86.34**	**89.84**	**89.41**

Table: Face classification accuracy and F1-score on Kinect and CASIA datasets. We compare the results of our MTUT methods and model trained with single modality.

Method	Test Modality	Kinect	CASIA
DCC-CAE	2D	91.67	92.64
SSA	2D	93.59	95.03
MTUT (ours)	2D	**97.43**	**95.24**
DCC-CAE	3D	73.72	82.81
SSA	3D	85.89	89.08
MTUT (ours)	3D	**85.90**	**89.84**

Table: Comparison with other state-of-the-art multimodal learnng methods on Kinect and CASIA. The backbone method for 2D and 3D modality are ResNet-18 [1] and PointNet [2]. The scores are reported as accuracy.

Ablation Study

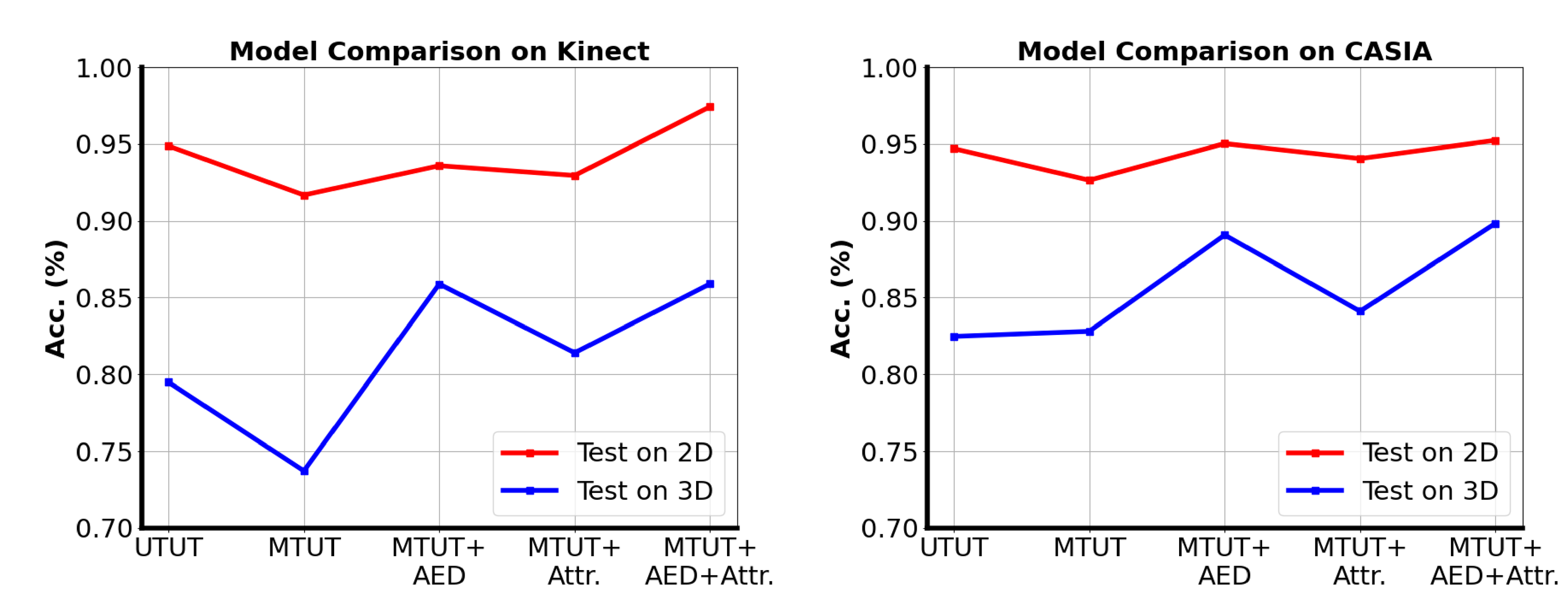


Figure: Ablation study. Our methods are *MTUT+AED+Attr.* UTUT stands for unimodal training unimodal testing. Note AED loss contribute significantly to the overall performance of our architecture.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[3] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.