Disentangled Variational Representation for Heterogeneous Face Recognition

Xiang Wu, Huaibo Huang, Vishal M. Patel, Ran He, Zhenan Sun
[1] Center for Research on Intelligent Perception and Computing, CASIA
[2] National Laboratory of Pattern Recognition, CASIA
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
[4] Johns Hopkins University, 3400 N. Charles St, Baltimore, MD 21218, USA

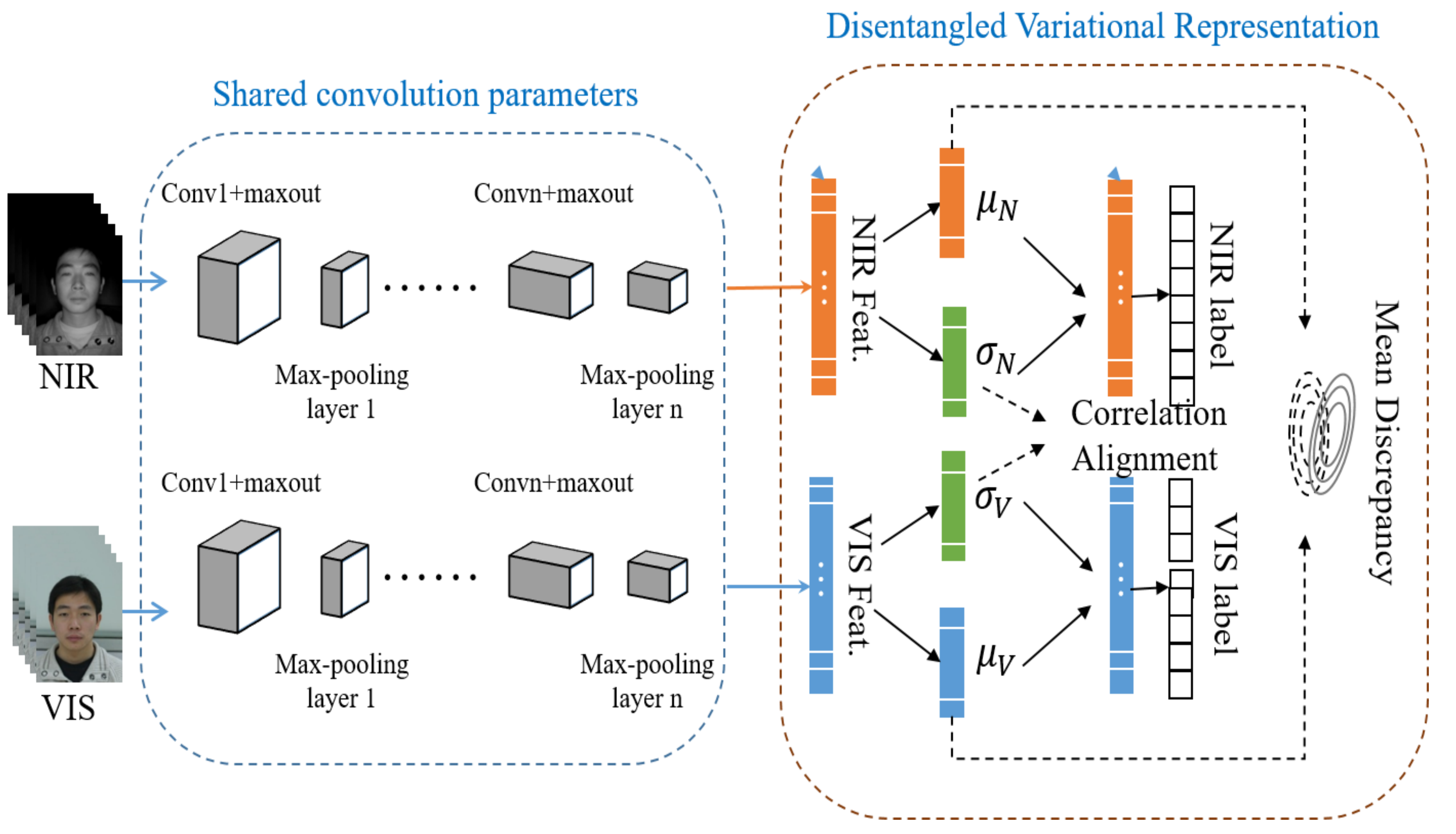
Problems

- The performance of face recognition methods, trained on large-scale VIS face images, degrades significantly when confronted by the NIR face images due to the modality discrepancy.
- The lack of sufficient training samples for heterogeneous face recognition may lead to overfitting during training.

Contributions

- An end-to-end Disentangled Variational Representation (DVR) framework is developed for cross-modal NIR-VIS face matching, aiming at disentangling the NIR and VIS face representations.
- We propose to minimize the identity information for the same subject and the relaxed correlation alignment constraint on modality variations that facilitate modeling the compact and discriminative disentangled latent variable spaces for heterogeneous modalities.
- An alternative optimization is proposed to provide mutual promotion between HFR network and disentangled variational representation part.

Framework



An overview of the proposed DVR approach. The NIR and VIS representations x_N and x_V are disentangled into (μ_N, σ_N) and (μ_V, σ_V), respectively. We assume that there is a linear relationship, P, between lighting variations, i.e. $\sigma_V = P\sigma_N$. The mean discrepancy is used to measure the difference between NIR and VIS distributions in the latent space. The reconstructions $\hat{x}_N$ and $\hat{x}_V$ are obtained from the likelihood $p(x_N|z_N)$ and $p(x_V|z_V)$, respectively and are constrained by the cross-entropy loss.

Objective

- Disentangled Varitional Representation

VAE

$$\mathcal{J}_{\text{DVR}} = \boxed{\begin{aligned} &\underbrace{-\frac{1}{2}\sum_j\left(1+\log\sigma_{Nj}^{2(i)}-\mu_{Nj}^{2(i)}-\sigma_{Nj}^{2(i)}\right)}_{\text{NIR approximate posterior estimator}} \\ &\underbrace{-\frac{1}{2}\sum_j\left(1+\log\sigma_{Vj}^{2(i)}-\mu_{Vj}^{2(i)}-\sigma_{Vj}^{2(i)}\right)}_{\text{VIS approximate posterior estimator}} \\ &\underbrace{+\mathbb{E}\left[\log p(x_N^{(i)}|z_N)\right]+\mathbb{E}\left[\log p(x_V^{(i)}|z_V)\right]}_{\text{reconstruction parts}} \end{aligned}}$$

$$\underbrace{+\lambda_1\|\mu_N^{(i)}-\mu_V^{(i)}\|_2^2}_{\text{mean discrepancy part}}$$

$$\underbrace{+\lambda_2\|\sigma_V-P\sigma_N\|_2^2+\lambda_3\|P^\top P-I\|_F^2,}_{\text{correlation alignment constraint}}$$

- Heterogeneous Recognition Network

$$\mathcal{J}_{\text{cls}} = \underline{\underline{\text{softmax}(x_i, y; W, \Theta)}} + \underline{\underline{\text{softmax}(\hat{x}_i, y; W, \Theta)}}, i \in \{N, V\}\ .$$

original features from true images

synthesis features sampling from DVR

Ablation Study

Table 1: The ablation study for DVR. Both LightCNN-9 and LightCNN-29 are used as the backbones.

Backbone	Disentangled Variational Part	Mean Discrepancy	Correlation Alignment	CASIA NIR-VIS 2.0		Oulu-CASIA NIR-VIS			BUAA-VisNir		
				Rank-1	FAR=0.1%	Rank-1	FAR=1%	FAR=0.1%	Rank-1	FAR=1%	FAR=0.1%
LightCNN-9	-	-	-	97.1	93.7	93.8	80.4	43.8	94.8	94.3	83.5
	√	-	-	98.0	97.3	96.3	85.9	50.7	96.5	95.8	88.3
	√	√	-	98.2	98.1	98.0	88.6	61.3	97.3	96.6	91.0
	√	√	√	**99.1**	**98.6**	**99.3**	**89.7**	**65.8**	**97.9**	**97.0**	**92.8**
LightCNN-29	-	-	-	98.1	97.4	99.0	93.1	68.3	96.8	97.0	89.4
	√	-	-	99.0	99.1	**100.0**	95.2	79.8	98.0	97.9	93.0
	√	√	-	99.5	99.3	**100.0**	96.5	83.0	98.9	98.4	95.6
	√	√	√	**99.7**	**99.6**	**100.0**	**97.2**	**84.9**	**99.2**	**98.5**	**96.9**

Comparisons

- **Database**

Three publicly available VIS-to-NIR face recognition datasets including the CASIA NIR-VIS 2.0 Face Database, the Oulu-CASIA NIR-VIS Database, and BUAA-VisNir Face Database are used for evaluations.

- **Implementation Details**

(1) We employ LightCNN as a basic network architecture. Both LightCNN-9 and LightCNN-29, which are pretrained on MS-Celeb-1M, are used to initialize parameters for HFR.

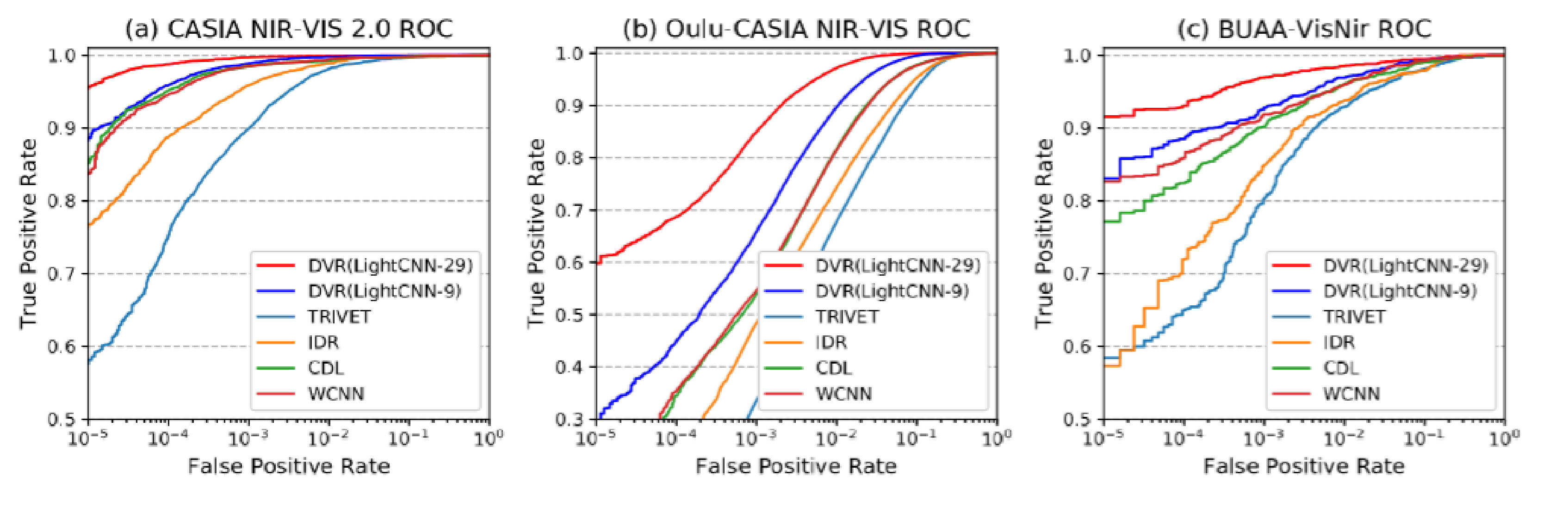
(2) The multilayer perceptron is used to model the DVR parts. The hidden layer dimension is set to 64.

(3) SGD is employed for recognition parts and Adam is used for DVR. The initial learning rate is set to 1e-3 and gradually reduced to 1e-5. The batch size is set to 128.

(4) The trade-off parameters $\lambda_1, \lambda_2, \lambda_3$ is set to 1.0, 0.1 and 0.001, respectively.

Table 2: Comparisons with other state-of-the-art HFR methods on the CASIA NIR-VIS 2.0 database, the Oulu-CASIA NIR-VIS database and the BUAA-VisNir database.

Method	CASIA NIR-VIS 2.0		Oulu-CASIA NIR-VIS			BUAA-VisNir		
	Rank-1	FAR=0.1%	Rank-1	FAR=1%	FAR=0.1%	Rank-1	FAR=1%	FAR=0.1%
KDSR (Huang et al. 2013)	37.5	9.3	66.9	56.1	31.9	83.0	86.8	69.5
H2(LBP3) (Shao and Fu 2017)	43.8	10.1	70.8	62.0	33.6	88.8	88.8	73.4
Gabor+RBM (Yi et al. 2015)	86.2 ± 1.0	81.3 ± 1.8	-	-	-	-	-	-
Recon.+UDP (Juefei-Xu, Pal, and Savvides 2015)	78.5 ± 1.7	85.8	-	-	-	-	-	-
Gabor+JB (Chen et al. 2012)	89.5 ± 0.8	83.2 ± 1.0	-	-	-	-	-	-
Gabor+HJB (Shi et al. 2017)	91.6 ± 0.8	89.9 ± 0.9	-	-	-	-	-	-
IDNet (Reale et al. 2016)	87.1 ± 0.9	74.5	-	-	-	-	-	-
HFR-CNN (Saxena and Verbeek 2016)	85.9 ± 0.9	78.0	-	-	-	-	-	-
Hallucination (Lezama, Qiu, and Sapiro 2017)	89.6 ± 0.9	-	-	-	-	-	-	-
TRIVET (Liu et al. 2016)	95.7 ± 0.5	91.0 ± 1.3	92.2	67.9	33.6	93.9	93.0	80.9
IDR (He et al. 2017)	97.3 ± 0.4	95.7 ± 0.7	94.3	73.4	46.2	94.3	93.4	84.7
ADFL (Song et al. 2018)	98.2 ± 0.3	97.2 ± 0.3	95.5	83.0	60.7	95.2	95.3	88.0
CDL (Wu et al. 2018b)	98.6 ± 0.2	98.3 ± 0.1	94.3	81.6	53.9	96.9	95.9	90.1
W-CNN (He et al. 2018)	98.7 ± 0.3	98.4 ± 0.4	98.0	81.5	54.6	97.4	96.0	91.9
DVR (LightCNN-9)	99.1 ± 0.2	98.6 ± 0.2	99.3	89.7	65.8	97.9	97.0	92.8
DVR (LightCNN-29)	$\mathbf{99.7 \pm 0.1}$	$\mathbf{99.6 \pm 0.3}$	**100.0**	**97.2**	**84.9**	**99.2**	**98.5**	**96.9**


(a) CASIA NIR-VIS 2.0 ROC
(b) Oulu-CASIA NIR-VIS ROC
(c) BUAA-VisNir ROC
True Positive Rate
False Positive Rate
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
DVR(LightCNN-29)
DVR(LightCNN-9)
TRIVET
IDR
CDL
WCNN