Learning Unified Representations for Multi-Resolution Face Recognition
- Supplementary Materials -

Anonymous Author(s)
Affiliation
Address
email

A Appendix

A.1 Theoretical Derivation of Up-sampling Error

Here, we take bilinear interpolation, a typical image interpolation method, as an example to analyze the relationship between the interpolation error and the resolution of a face image. Bilinear interpolation can be considered as a bivariate Lagrange interpolation problem containing two interpolation nodes in each of the two dimensions.

Let D be a unit-bounded closed region in a two-dimensional image space, and $Q_1\left(x_0, y_0\right),\ Q_2\left(x_1, y_0\right),\ Q_3\left(x_0, y_1\right),\ Q_4(x_1, y_1) \in D$ be four adjacent pixel points in this region. We use an interpolation polynomial $P(x, y)$ for the interpolation approximation of the bivariate continuous function $f(x, y)$ defined on D , and the interpolation error $E(x, y)$ can be expressed as

$$E(x, y)=f(x, y)-P(x, y) \tag{1}$$

which indicates the potential error information introduced to the recognition of different identities. According to the the Rolle's theorem, we can obtain

$$E(x, y)=\frac{\frac{\partial^4 f(\xi, \eta)}{\partial x^2 \partial y^2}}{4} \omega_2(x) \mu_2(y) \tag{2}$$

where ξ, η is an interior point of D and

$$\omega_2(x)=(x-x_0)(x-x_1) \tag{3}$$

$$\mu_2(y)=(y-y_0)(y-y_1) \tag{4}$$

As $x_1-x_0=y_1-y_0=1$ for adjacent pixel points, we can get the upper bound of $|\omega_2(x)|$ and $\left|\mu_2\left(y\right)\right|$

$$|\omega_2(x)|<\frac{1}{4}, |\mu_2(y)|<\frac{1}{4} \tag{5}$$

Thus, the error estimation can be expressed as

Table 1: Comparison of different training methods for our BTNet. “Acc.” denotes average 1:1 verification accuracy. “# Params.” indicates the amount of parameter storage for the branch network B_{14}.

Training method	Acc. (%)		# Params. (M)
	112&14	14&14	
Scratch	49.90	78.00	43.59
Pretraining	78.05	76.87	43.59
Pretraining + BCT	85.90	78.04	43.59
Pretraining + BCT + Fix Trunk	85.07	77.22	2.29
Pretraining + BCT + Fix Trunk + Branch Distillation	94.08	90.90	2.29

$$E(x,y) \leq \frac{\left|\frac{\partial^4 f(\xi,\eta)}{\partial x^2 \partial y^2}\right|}{64} \tag{6}$$

where $\frac{\partial^4 f(\xi,\eta)}{\partial x^2 \partial y^2}$ can be approximated using the difference operator

$$\begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix} \tag{7}$$

Based on the above theoretical analysis, we can experimentally study the relationship between the estimated up-sampling error and the image resolution.

A.2 Instantiation of BTNet-res50

We provide the detailed architecture of BTNet-res50 (φ_{bt}), an instantiation of BTNet framework based on ResNet50 [1]. Our method can be easily implemented by refining a network with the top-down hierarchical representation structure.

A.3 Ablation Study

In all these experiments, we report the average verification results on six benchmarks in 112&14 and 14&14 matching, representing cross-resolution and same-resolution performance respectively.

Training Method Alternatives. Here, we experimentally compare different training methods: (1) Scratch: train without pretrained trunk parameters. (2) Pretraining: initialize the backbone and classifier with the pretrained trunk network. (3) Backward-compatible training (BCT, [2]): fix parameters of the old classifier. (4) Fix-trunk: fix parameters of the trunk subnet T_r. (5) Branch distillation: use L2-distance to obtain the loss between the intermediate feature maps at the coupling layer of the pretrained trunk T and the branch B_r.

We compare different training method combinations in Table 1 and find that both pretraining and BCT succeeded in ensuring representation compatibility. Among these two, BCT performs better since it imposes a stricter constraint during training. Furthermore, we are able to observe that branch distillation is crucial for improving the discriminative power by transferring high-resolution information to low-resolution branches.

Where should we have resolution-specific layers? We conducted an ablation to see the effects of different specific-shared layer allocation strategies. The experiment was done with different trunk layers (i.e., the parameters of these layers are inherited from the pretrained trunk without updating). Figure 2 shows the results. We find that increasing the number of branch layers (i.e., specific layers for different resolutions) will lead to better performance due to increased flexibility. Our specific-shared layer allocation of BTNet can achieve better parameter/accuracy tradeoffs. Since further increasing the number of trunk layers based on BTNet cannot lead to significantly better performance but

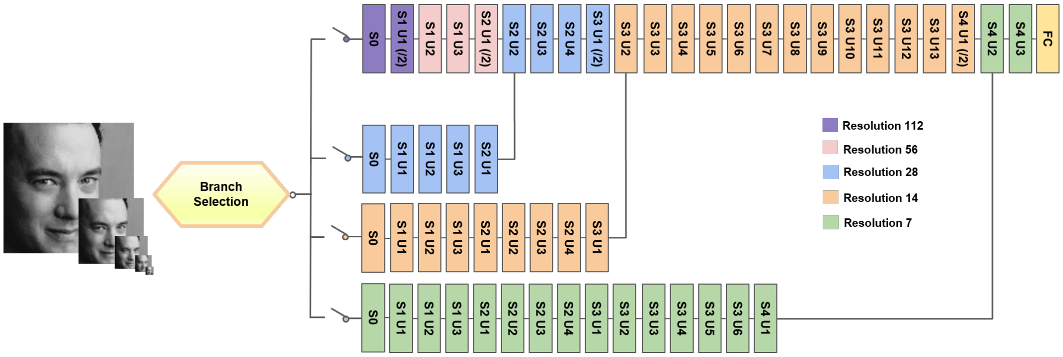


Figure 1: Detailed architecture of BTNet-res50 (φ_{bt}). Note that 'S' and 'U' represent stage and unit respectively, and '/2' means down-sampling by convolution with stride 2.

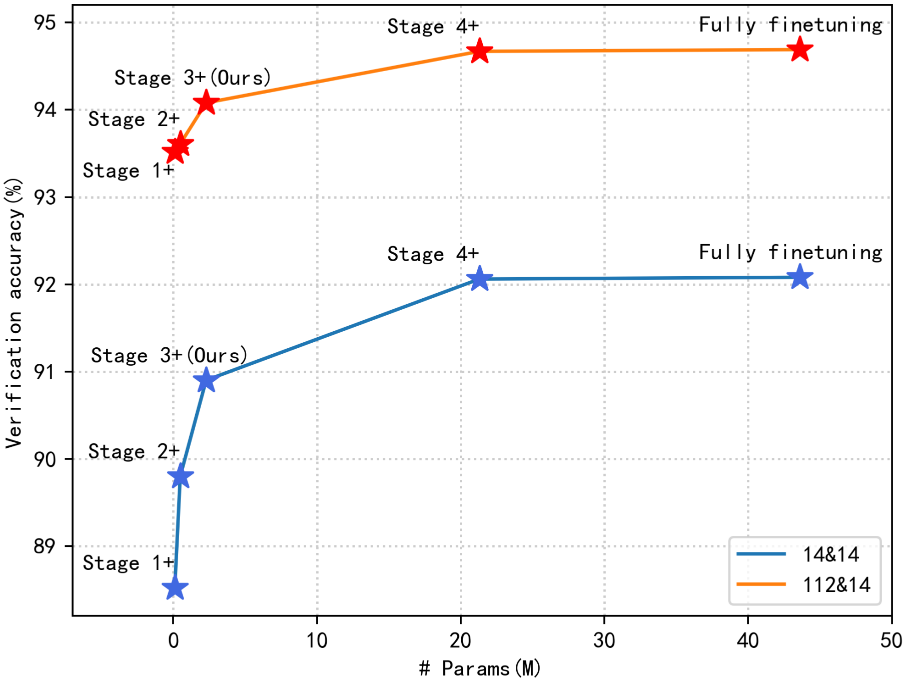


Figure 2: Comparison of verification accuracy and the amount of stored parameters for different specific-shared layer allocation strategies. Note that "Stage x+" indicates that layers deeper than "Stage x, Unit 1" are inherited from the pretrained trunk without updating.

increases parameter storage cost by a large margin, we use resolution-specific layers as shown in Figure 1.

A.4 Visualization

To interpret the behavior of learning compatible and discriminative representations, we visualize the intermediate feature maps in Figure 3. We find that φ_{hr} introduces the noise information while φ_{mm} has more discriminative but resolution-variant feature maps. The feature maps of φ_{mr} tend to be smoother, diminishing the error information, but the discriminability could be limited as high-frequency details benefit recognition [3].

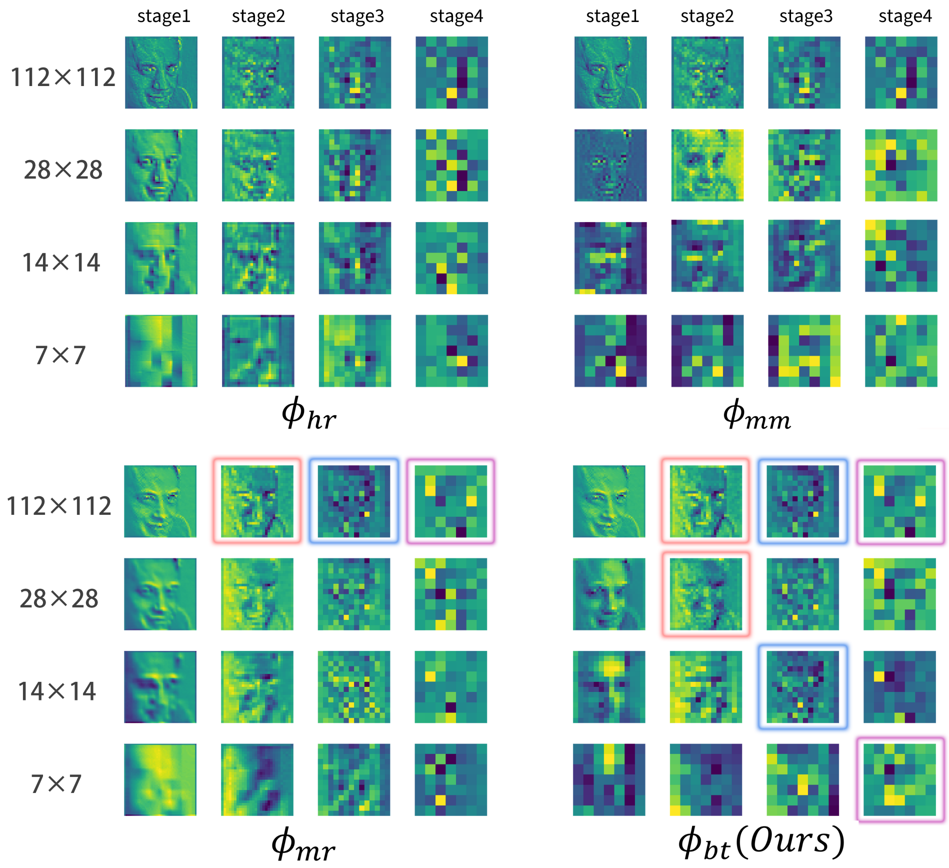


Figure 3: Visualization of intermediate feature maps for inputs with different resolutions. We show the feature maps located at output layers of BNets, denoted as stage1/2/3/4 respectively. We see our method can transfer multi-resolution visual inputs to intermediate feature maps at corresponding layers (indicated by bounding boxes of the same color) of TNet.

We also show that through the resolution-specific feature transfer of multiple branches, φ_{bt} can encourage the transferred features to be aligned before fed into the trunk network in corresponding layers. For instance, at stage 2, the feature maps of φ_{bt} with input resolution 112 and 28 are more similar than those of φ_{hr}, φ_{mm}, φ_{mr}. Furthermore, more detailed information can be found in the feature maps of φ_{bt} with input resolution 28 compared to φ_{mr}. This inspiring phenomenon suggests that BTNet can learn compatible representations while improving the discriminability in low-resolution domain through the knowledge transferred from high-resolution visual signals.

A.5 Additional Experimental Results

Multi-Resolution Identity Matching. We report the detailed results for 1:1 verification on each dataset (i.e., LFW, CFP-FF, CFP-FP, AgeDB-30, CALFW and CPLFW). The relative drop of φ_{bt} in high-resolution setting (i.e., 112&112) becomes almost negligible compared to the improvement for all the other settings which incorporate low-resolution inputs.

Multi-Resolution Feature Aggregation. We report the detailed results on the IJB-C dataset, including TAR at different FAR (see Table 3, 4), ROC Curve (see Figure 4, 5) for 1:1 verification, and TPIR at FPIR=0.01, Top-1, Top-5, Top-10 accuracy (see Table 6, 7) for 1:N identification. We are able to observe that φ_{bt} can be comparable to or serve as the paradigm model (i.e., model with the best performance) in each resolution setting, both for identity matching and feature aggregation.

Table 2: Detailed cross-resolution 1:1 verification accuracy (mean±std over 5 trails) per-benchmark.

		φ_{hr}	φ_{mm}	φ_{mr}	φ_{bt} (Ours)
112&7	LFW	63.0±2.0	51.5±2.6	77.4±1.4	**96.1±0.7**
	CFP-FF	56.2±1.2	51.2±2.0	64.7±2.0	**90.9±1.3**
	CFP-FP	54.9±1.3	49.8±1.6	60.8±2.5	**80.2±2.3**
	AgeDB-30	57.3±1.6	50.0±1.4	60.5±2.0	**79.8±2.3**
	CALFW	58.5±1.7	51.2±1.6	66.1±1.8	**87.8±1.7**
	CPLFW	56.6±1.6	49.8±1.5	65.6±1.3	**81.8±1.3**
112&14	LFW	91.0±1.2	49.1±1.3	96.9±0.6	**99.4±0.3**
	CFP-FF	81.7±1.7	50.0±1.5	90.4±1.0	**98.2±0.4**
	CFP-FP	75.5±1.4	50.2±2.0	82.3±1.8	**92.6±1.2**
	AgeDB-30	76.9±1.5	51.3±1.8	82.7±1.1	**91.3±1.2**
	CALFW	81.8±1.0	49.7±1.5	88.0±0.8	**93.9±1.1**
	CPLFW	79.2±1.3	49.1±1.4	84.5±1.4	**89.1±1.7**
112&28	LFW	99.5±0.3	48.9±1.1	**99.7±0.2**	**99.7±0.2**
	CFP-FF	99.0±0.3	51.5±1.8	99.5±0.3	**99.7±0.2**
	CFP-FP	94.9±1.1	49.4±2.0	95.4±0.8	**97.0±0.5**
	AgeDB-30	95.7±1.0	49.5±0.7	95.5±1.1	**96.3±1.1**
	CALFW	95.0±1.0	50.6±0.7	94.9±1.0	**95.5±1.0**
	CPLFW	91.3±1.2	50.3±1.2	91.3±1.2	**91.7±1.0**

Table 3: Detailed same-resolution 1:1 verification accuracy (mean±std over 5 trails) per-benchmark.

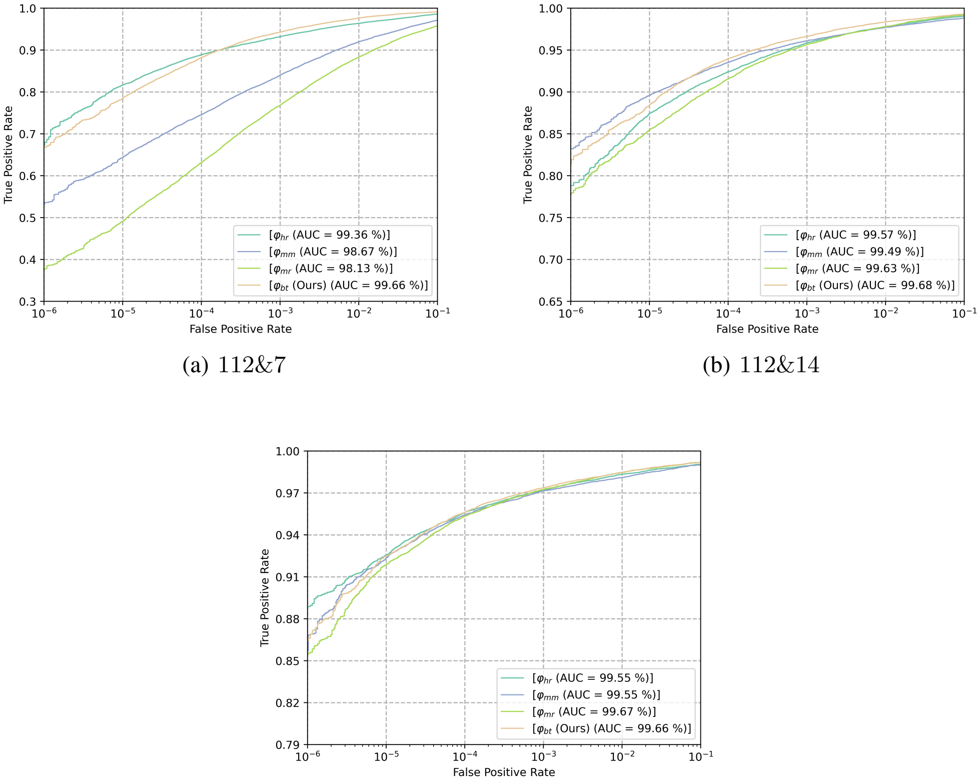
		φ_{hr}	φ_{mm}	φ_{mr}	φ_{bt} (Ours)
7&7	LFW	70.8±2.9	74.0±1.5	72.5±1.8	**92.7±1.2**
	CFP-FF	67.4±2.3	69.4±1.7	67.7±1.4	**86.1±1.5**
	CFP-FP	57.1±2.3	59.1±2.3	56.5±1.3	**73.8±1.4**
	AgeDB-30	54.3±1.9	54.2±2.2	53.5±1.8	**62.5±2.2**
	CALFW	58.2±1.0	60.1±1.3	59.2±2.1	**76.0±1.5**
	CPLFW	56.4±1.7	58.6±1.2	56.7±1.1	**75.6±1.5**
14&14	LFW	87.9±0.9	93.4±1.2	94.8±0.9	**98.5±0.5**
	CFP-FF	79.0±2.2	84.7±1.6	86.7±1.6	**96.2±0.7**
	CFP-FP	68.2±1.5	73.7±2.0	78.0±1.5	**89.0±1.0**
	AgeDB-30	64.1±1.7	64.2±2.4	65.9±2.3	**84.2±1.6**
	CALFW	71.8±0.9	75.6±1.3	77.5±1.4	**89.9±0.7**
	CPLFW	72.3±1.6	76.4±1.8	79.0±1.7	**87.6±1.9**
28&28	LFW	99.1±0.4	99.6±0.6	99.6±0.3	**99.8±0.3**
	CFP-FF	97.2±0.7	98.4±0.7	99.1±0.4	**99.4±0.3**
	CFP-FP	91.9±1.3	93.5±1.3	95.0±1.0	**96.8±0.9**
	AgeDB-30	90.9±1.2	92.6±0.8	92.4±1.0	**94.9±1.1**
	CALFW	92.9±1.3	93.4±0.9	93.9±1.3	**95.0±0.9**
	CPLFW	89.5±1.3	90.6±1.2	90.7±1.3	**91.7±0.9**
112&112	LFW	**99.8±0.2**	**99.8±0.2**	**99.8±0.2**	**99.8±0.2**
	CFP-FF	**99.9±0.1**	**99.9±0.1**	99.8±0.2	99.8±0.2
	CFP-FP	**98.9±0.3**	**98.9±0.3**	98.1±0.4	98.1±0.4
	AgeDB-30	**98.4±0.7**	**98.4±0.7**	97.2±0.8	97.2±0.8
	CALFW	**96.0±1.2**	**96.0±1.2**	95.9±1.0	95.9±1.0
	CPLFW	**93.1±1.3**	**93.1±1.3**	92.7±1.0	92.7±1.0

Table 4: 1:1 verification TAR at different FAR on the IJB-C dataset for cross-resolution feature aggregation.

	112&7					112&14					112&28				
FAR	10^{-6}	10^{-5}	10^{-3}	10^{-2}	10^{-1}	10^{-6}	10^{-5}	10^{-3}	10^{-2}	10^{-1}	10^{-6}	10^{-5}	10^{-3}	10^{-2}	10^{-1}
φ_{hr}	**67.99**	**81.65**	93.18	96.38	98.65	78.83	87.44	95.86	97.79	99.05	**88.87**	**92.56**	97.19	98.33	99.06
φ_{mm}	53.57	64.34	84.01	91.96	97.12	**83.22**	**89.56**	96.10	97.71	98.82	86.84	92.33	97.16	98.10	99.01
φ_{mr}	37.83	49.12	76.80	88.32	95.79	77.97	85.46	95.64	97.79	99.21	85.55	91.86	97.25	98.46	99.19
φ_{bt} (Ours)	66.84	78.40	**94.27**	**97.63**	**99.16**	81.92	88.38	**96.64**	**98.34**	**99.28**	86.61	92.48	**97.38**	**98.47**	**99.20**

Table 5: 1:1 verification TAR at different FAR on the IJB-C dataset for same-resolution feature aggregation.

	7&7					14&14					28&28					112&112				
FAR	10^{-6}	10^{-5}	10^{-3}	10^{-2}	10^{-1}	10^{-6}	10^{-5}	10^{-3}	10^{-2}	10^{-1}	10^{-6}	10^{-5}	10^{-3}	10^{-2}	10^{-1}	10^{-6}	10^{-5}	10^{-3}	10^{-2}	10^{-1}
φ_{hr}	0.69	1.73	12.58	27.63	56.81	9.82	20.38	52.57	72.61	90.30	75.67	83.24	94.21	97.15	98.74	**89.58**	**94.51**	**97.57**	98.40	99.06
φ_{mm}	0.68	1.73	11.93	27.48	56.84	7.59	15.61	48.28	71.13	91.04	73.68	85.14	95.82	97.65	98.89	**89.58**	**94.51**	**97.57**	98.40	99.06
φ_{mr}	0.74	1.76	11.11	25.98	54.26	14.21	24.72	60.39	79.84	94.35	78.91	86.42	96.04	98.07	99.09	88.48	93.37	97.50	**98.51**	**99.23**
φ_{bt} (Ours)	**12.09**	**20.70**	**57.17**	**79.02**	**93.90**	**57.75**	**70.63**	**90.85**	**96.06**	**98.68**	**82.85**	**90.32**	**96.94**	**98.31**	**99.15**	88.48	93.37	97.50	**98.51**	**99.23**



(a) 112&7

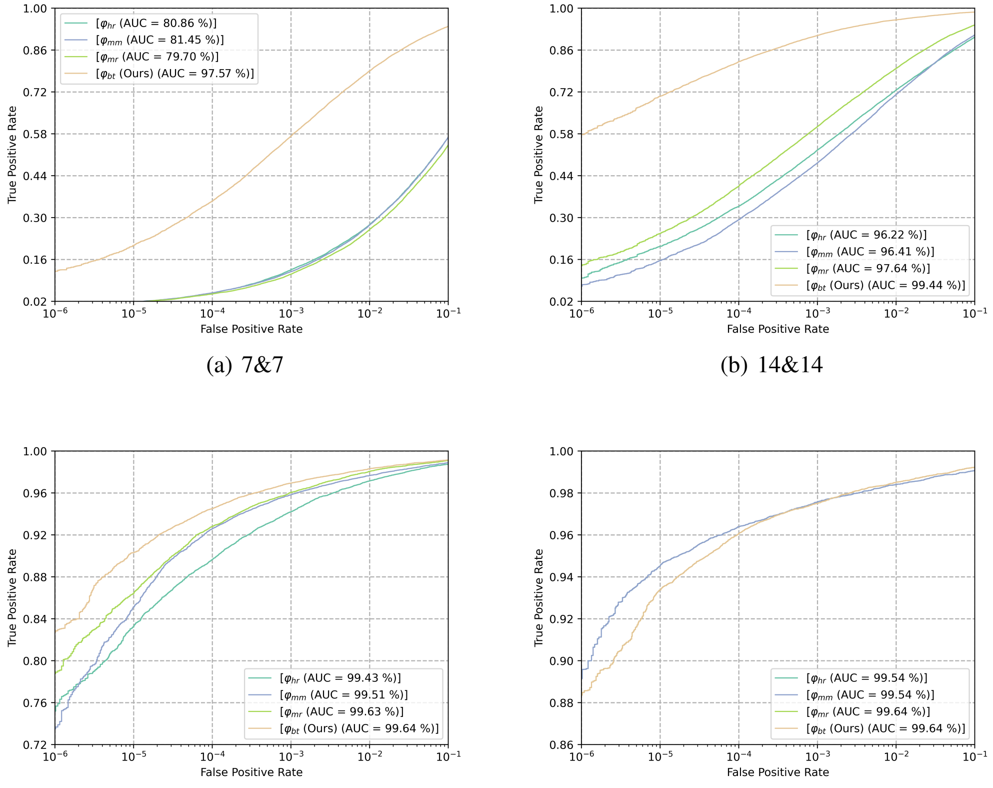
(b) 112&14

(c) 112&28

Figure 4: 1:1 verification ROC Curve on the IJB-C dataset for cross-resolution feature aggregation.

Table 6: 1: N identification TPIR(%@FPIR=0.01), Top-1, Top-5, Top-10 accuracy on the IJB-C dataset for cross-resolution feature aggregation.

	112&7				112&14				112&28			
	TPIR	Top-1	Top-5	Top-10	TPIR	Top-1	Top-5	Top-10	TPIR	Top-1	Top-5	Top-10
φ_{hr}	**75.35**	**92.76**	**95.14**	**95.92**	81.98	93.89	96.25	96.98	**90.42**	96.05	**97.47**	97.80
φ_{mm}	59.07	88.89	92.33	93.35	**86.39**	**95.15**	**96.86**	97.31	90.04	96.00	97.31	97.72
φ_{mr}	43.89	82.29	87.74	89.42	82.18	93.87	96.20	96.89	88.90	95.93	97.36	97.84
φ_{bt}(Ours)	73.40	91.30	94.86	95.88	84.78	94.78	96.84	**97.41**	89.84	**96.16**	97.46	**97.90**



(a) 7&7

(b) 14&14

(c) 28&28

(d) 112&112

Figure 5: 1:1 verification ROC Curve on the IJB-C dataset for same-resolution feature aggregation.

Table 7: 1: N identification TPIR(%@FPIR=0.01), Top-1, Top-5, Top-10 accuracy on the IJB-C dataset for same-resolution feature aggregation.

	7&7				14&14				28&28				112&112			
	TPIR	Top-1	Top-5	Top-10	TPIR	Top-1	Top-5	Top-10	TPIR	Top-1	Top-5	Top-10	TPIR	Top-1	Top-5	Top-10
φ_{hr}	1.20	11.77	19.95	24.28	15.16	50.96	63.62	68.68	77.52	91.62	94.95	95.99	**92.66**	**96.58**	**97.71**	97.94
φ_{mm}	1.24	20.38	30.23	34.83	11.62	62.08	72.33	76.33	79.31	93.87	96.09	96.81	**92.66**	**96.58**	**97.71**	97.94
φ_{mr}	1.36	17.41	26.53	31.03	23.72	68.64	78.38	81.99	83.82	94.53	96.67	97.33	90.89	96.44	97.65	**98.00**
φ_{bt}(Ours)	**15.55**	**55.49**	**67.98**	**73.05**	**63.69**	**86.35**	**92.14**	**94.01**	**86.87**	**95.42**	**97.06**	**97.62**	90.89	96.44	97.65	**98.00**

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[2] Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6367–6376. Computer Vision Foundation / IEEE, 2020.

[3] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8681–8691. Computer Vision Foundation / IEEE, 2020.