Domain Generalization for Object Recognition with Multi-task Autoencoders

VICTORIA UNIVERSITY OF WELLINGTON

Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, David Balduzzi

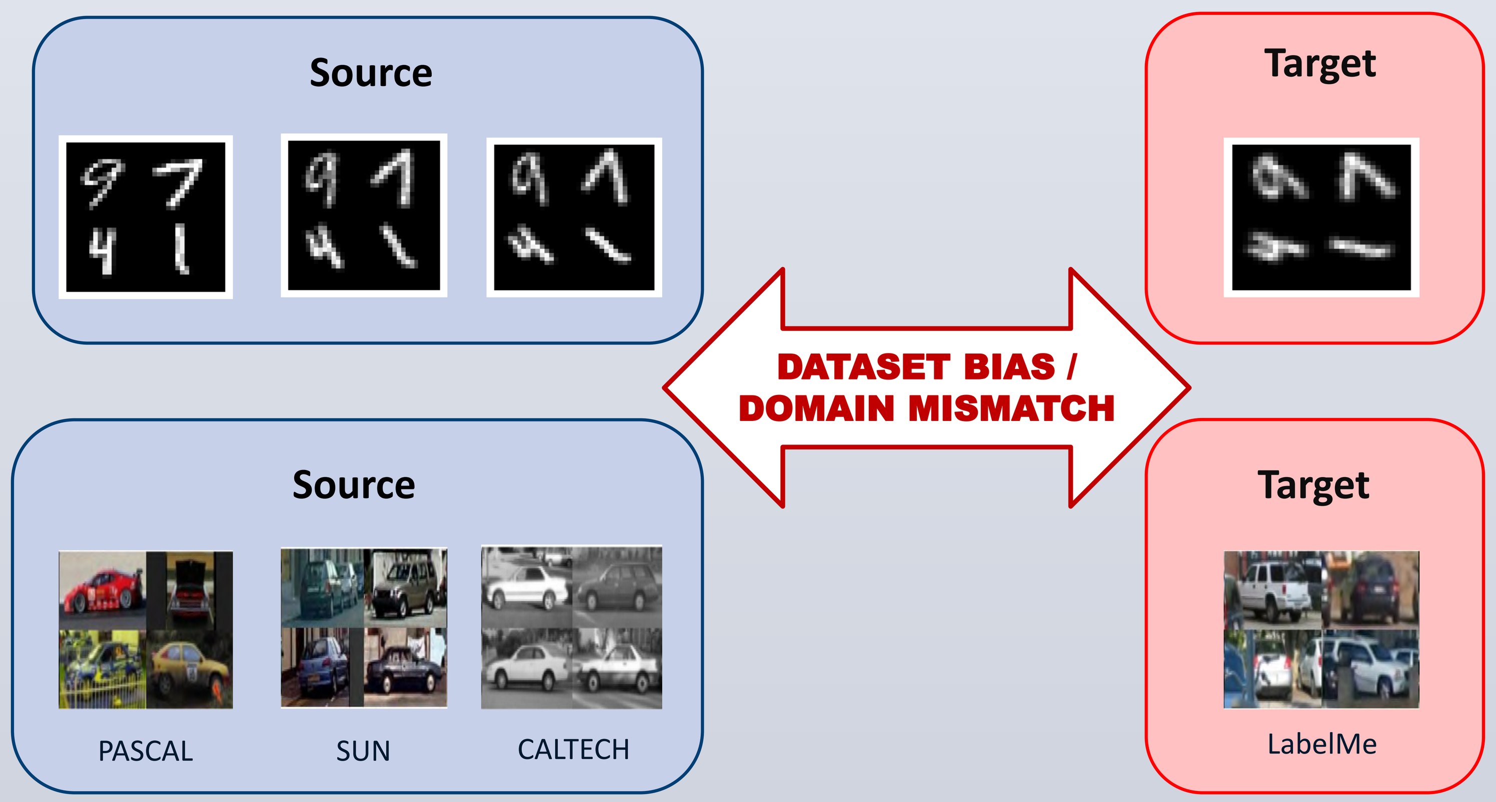
{muhammad.ghifary, bastiaan.kleijn, mengjie.zhang}@ecs.vuw.ac.nz, david.balduzzi@vuw.ac.nz

International Conference on Computer Vision (ICCV2015)

1. Summary

- Propose novel feature learning algorithms based on multi-task learning for domain generalization: ***Multi-task Autoencoder*** **(MTAE)** and ***denoising MTAE***.
- MTAE learns to transform an original image into analogs in multiple related domains; the learned features are robust to variations across domains.
- The learned weights of MTAE and D-MTAE appear to capture the *underlying transformation* among source domains.
- Evaluations on benchmark datasets: MNIST, ETH-80, PASCAL VOC 2007, LabelMe, Caltech-101, SUN09, and Office show the effectiveness of MTAE and D-MTAE.

2. Domain Generalization



- In discriminative tasks, standard supervised learning algorithms may suffer from dataset bias [Torralba and Efros, 2011] or domain mismatch if the training data do not represent the target environment.
- Domain generalization is concerned with mitigating dataset bias, i.e., improving generalization ability of classifiers to any unseen target domain given other different but related domains as the training sources.

3. Autoencoder

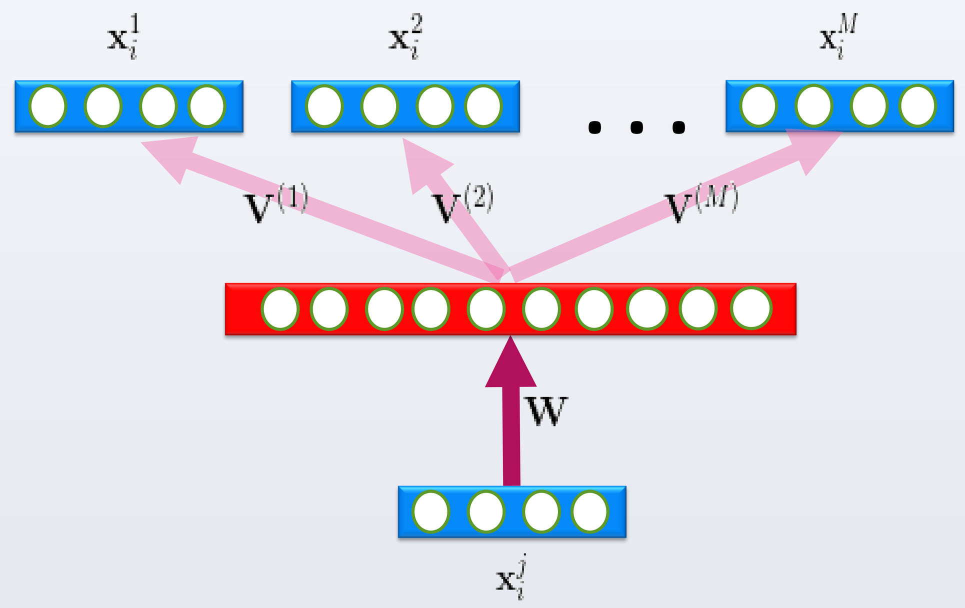
- An autoencoder (AE) $f_\Theta : \mathbb{R}^d \to \mathbb{R}^d$ is a single-layer neural net with the output signal is the input itself. It consists of two stages:
  1. Encoding: $\mathbf{h} = \sigma_{\mathrm{enc}}(\mathbf{W}^\top \mathbf{x})$
  2. Decoding: $\hat{\mathbf{x}} = \sigma_{\mathrm{dec}}(\mathbf{V}^\top \mathbf{h}) =: f_\Theta(\mathbf{x})$

  where $\sigma_{\mathrm{enc}}(\cdot)$ and $\sigma_{\mathrm{dec}}(\cdot)$ are non-linear, element-wise activation functions, and $\Theta = \{\mathbf{W}, \mathbf{V}\}$
- Define a *reconstruction* loss function $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, e.g., the mean-squared loss. Given a finite sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, the objective of autoencoders is

$$\hat{\Theta} := \arg\min_{\Theta} \sum_{i=1}^{N} \ell(\mathbf{x}_i, f_\Theta(\mathbf{x}_i)) + \eta \mathcal{R}(\Theta)$$

  where $\mathcal{R}(\Theta)$ is any useful regularization term, e.g., $\mathcal{R}(\Theta) = \|\mathbf{W}\|_2^2 + \|\mathbf{V}\|_2^2$
- Feature extraction: given a datapoint $\mathbf{x}'$, $\mathbf{h}' = \sigma_{\mathrm{enc}}(\hat{\mathbf{V}}^\top \mathbf{x}')$
- Denoising autoencoder (DAE) [Vincent et al, 2010]: the original input to the autoencoder $f_\Theta(\cdot)$ is corrupted $\tilde{\mathbf{x}} \sim \mathcal{Q}(\tilde{X}|X)$ with, e.g., zero-masking or Gaussian noise.

4. Multi-task Autoencoder (MTAE)

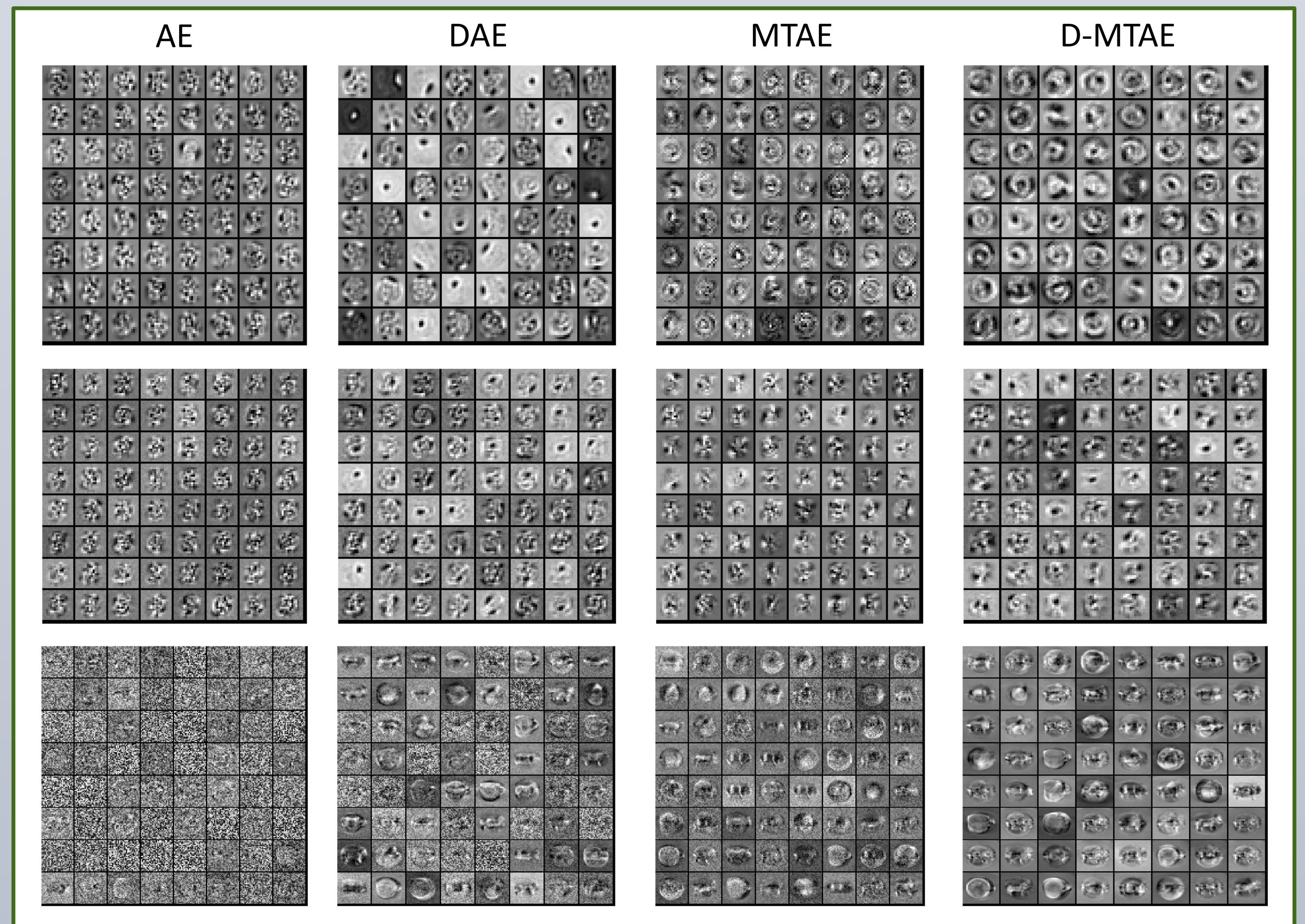


- MTAE has multiple outputs with a shared encoding layer [Caruana 1997], where the target outputs are the data points from multiple domains.
- *Category-level reconstruction:* the class of output data points is the same as that of the input.
- Let $S^l = \{\mathbf{x}_1^l, \ldots, \mathbf{x}_N^l\} \sim \mathcal{D}^l$ be a sample from the *l*-th domain, where $l = 1, \ldots, M$. Suppose that $\{\mathbf{x}_i^l\}_{l=1}^{M}$ have the same label. The objective of MTAE:

$$\min_{\Theta^1,\ldots,\Theta^M} \sum_{l,j=1}^{M} \sum_{i=1}^{N} \ell(f_{\Theta^l}(\mathbf{x}_i^j), \mathbf{x}_i^l) + \eta \mathcal{R}(\Theta^l)$$

  where $\Theta^l = \{\mathbf{V}, \mathbf{W}^l\}$ is the set of parameters with $\mathbf{V} \in \mathbb{R}^{d \times k}$ are the shared weights.
- The denoising strategy can be easily incorporated in MTAE by simply replacing $\mathbf{x}_i^j$ in the above objective with its corrupted pair $\tilde{\mathbf{x}}_i^j \sim \mathcal{Q}(\tilde{X}|X)$. We refer to the model as D-MTAE.

5. Weight Visualization



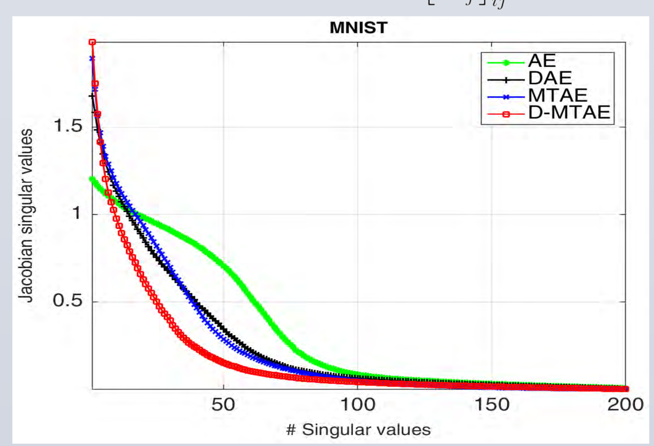
- AE and DAE only captures the *content* of the objects, i.e., local blob and stroke detectors, while MTAE also learns the *transformation* among domains.
- The D-MTAE's weights appear to be a mixture of DAE and MTAE's weights.

6. Experiment I: MNIST and ETH-80

- We created several *object views / domains* from MNIST and ETH-80 images and evaluated the cross-domain recognition tasks: feature learning + L-SVM classification.
- We compare MTAE and D-MTAE with autoencoder-based algorithms: AE, DAE, and CAE, and uDICA [Muandet et al. 2013].
- The inputs to the algorithms are the raw pixels.

Source	Target	Raw	AE	DAE	CAE	uDICA	**MTAE**	**D-MTAE**
MNIST-r leave-one-**roll-rotation**-out								
$M_{15°}, M_{30°}, M_{45°}, M_{60°}, M_{75°}$	M	52.40	74.20	76.90	72.10	67.20	**77.90**	**82.50**
$M, M_{30°}, M_{45°}, M_{60°}, M_{75°}$	$M_{15°}$	74.10	93.20	93.20	95.30	87.80	**95.70**	**96.30**
$M, M_{15°}, M_{45°}, M_{60°}, M_{75°}$	$M_{30°}$	71.40	89.90	91.30	**92.60**	88.80	91.20	**93.40**
$M, M_{15°}, M_{30°}, M_{60°}, M_{75°}$	$M_{45°}$	61.40	**82.20**	81.10	**81.50**	77.80	77.30	78.60
$M, M_{15°}, M_{30°}, M_{45°}, M_{75°}$	$M_{60°}$	67.40	90.00	**92.80**	92.70	84.20	92.40	**94.20**
$M, M_{15°}, M_{30°}, M_{45°}, M_{60°}$	$M_{75°}$	55.40	73.80	76.50	79.30	69.50	**79.90**	**80.50**
Average		63.68	83.88	85.30	85.58	79.22	**85.73**	**87.58**
MNIST-s leave-one-**dilation**-out								
$M_{*0.9}, M_{*0.8}, M_{*0.7}, M_{*0.6}$	M	54.00	67.50	71.80	**75.80**	**75.80**	74.50	**76.00**
$M, M_{*0.8}, M_{*0.7}, M_{*0.6}$	$M_{*0.9}$	80.40	95.10	94.00	94.90	88.60	**97.80**	**98.00**
$M, M_{*0.9}, M_{*0.7}, M_{*0.6}$	$M_{*0.8}$	82.60	94.60	92.90	94.90	86.60	**96.30**	**96.40**
$M, M_{*0.9}, M_{*0.8}, M_{*0.6}$	$M_{*0.7}$	78.20	93.70	91.60	92.50	87.40	**95.80**	**94.90**
$M, M_{*0.9}, M_{*0.8}, M_{*0.7}$	$M_{*0.6}$	64.70	74.80	76.10	77.50	75.30	**78.00**	**78.30**
Average		71.98	85.14	85.28	87.12	82.74	**88.48**	**88.72**

The singular value spectrum of the Jacobian matrix $J_x(\mathbf{h}) = \left[\frac{\partial x_i}{\partial x_j}\right]_{ij}$



Source	Target	AE	DAE	CAE	**MTAE**	**D-MTAE**
ETH80-p leave-one-**pitch**-out						
$E_{p22}, E_{p45}, E_{p68}, E_{p90}$	E_{p0}	70.00	73.73	**74.50**	73.76	**77.50**
$E_{p0}, E_{p45}, E_{p68}, E_{p90}$	E_{p22}	86.25	88.74	88.50	**92.50**	**92.50**
$E_{p0}, E_{p22}, E_{p68}, E_{p90}$	E_{p45}	92.51	93.77	93.49	**97.51**	**97.53**
$E_{p0}, E_{p22}, E_{p45}, E_{p90}$	E_{p68}	95.01	**98.74**	**99.00**	**98.78**	**98.78**
$E_{p0}, E_{p22}, E_{p45}, E_{p68}$	E_{p90}	75.00	**75.02**	**74.49**	**75.03**	72.78
Average		83.75	86.00	86.00	**87.51**	**87.82**
ETH80-y leave-one-**yaw**-out						
$E_{+y45}, E_{y0}, E_{-y45}, E_{-y90}$	E_{+y90}	84.98	**94.97**	91.20	91.26	**92.50**
$E_{+y90}, E_{y0}, E_{-y45}, E_{-y90}$	E_{+y45}	98.75	98.75	98.75	98.75	98.75
$E_{+y90}, E_{+y45}, E_{-y45}, E_{-y90}$	E_{y0}	92.48	93.73	94.72	**96.25**	**97.50**
$E_{+y90}, E_{+y45}, E_{y0}, E_{-y90}$	E_{-y45}	97.49	98.75	98.75	**100.00**	**100.00**
$E_{+y90}, E_{+y45}, E_{y0}, E_{-y45}$	E_{-y90}	91.23	94.96	93.80	**96.25**	**96.25**
Average		92.99	96.23	95.44	**96.50**	**97.00**

7. Experiment II: PASCAL VOC, Caltech, SUN09, LabelMe, and Office

- The inputs to the algorithms are the DeCAF$_6$ features [Donahue et al. 2014].
- MTAE and D-MTAE are used as the pretraining algorithm for a feed-forward neural net.
- We compared MTAE and D-MTAE with state-of-the-art algorithms: Undo-Bias [Khosla et al. 2012], UML [Fang et al. 2013], LRE-SVM [Xu et al. 2014].

Algorithms	L,C,S → V	V,C,S → L	V,L,S → C	V,L,C → S	Avg.
L-SVM	58.86 ± 0.00	52.49 ± 0.00	77.67 ± 0.00	49.09 ± 0.00	59.93 ± 0.00
1HNN	59.10 ± 0.09	58.20 ± 0.04	86.87 ± 0.12	57.86 ± 0.15	65.46 ± 0.10
DAE+1HNN	$\mathbf{62.00 \pm 0.11}$	59.23 ± 0.00	87.50 ± 0.09	54.12 ± 0.08	65.75 ± 0.07
Undo-Bias	54.29 ± 0.00	58.09 ± 0.00	87.50 ± 0.00	54.21 ± 0.00	63.52 ± 0.00
UML	56.26 ± 0.00	58.50 ± 0.00	$\mathbf{91.13 \pm 0.00}$	58.49 ± 0.00	65.85 ± 0.00
LRE-SVM	60.58 ± 0.00	$\mathbf{59.74 \pm 0.00}$	88.11 ± 0.00	54.88 ± 0.00	65.83 ± 0.00
MTAE+1HNN	61.09 ± 0.02	$\mathbf{59.24 \pm 0.00}$	$\mathbf{90.17 \pm 0.08}$	$\mathbf{60.20 \pm 0.06}$	$\mathbf{67.81 \pm 0.04}$
D-MTAE+1HNN	$\mathbf{63.90 \pm 0.07}$	$\mathbf{60.13 \pm 0.00}$	$\mathbf{89.05 \pm 0.06}$	$\mathbf{61.33 \pm 0.08}$	$\mathbf{68.60 \pm 0.05}$

Algorithms	A,C → D,W	D,W → A,C	C,D,W → A	A,W,D → C	Avg.
L-SVM	82.08 ± 0.00	76.12 ± 0.00	90.61 ± 0.00	84.51 ± 0.00	83.33 ± 0.00
1HNN	83.41 ± 0.15	76.49 ± 0.11	92.13 ± 0.08	85.89 ± 0.03	84.48 ± 0.09
DAE+1HNN	82.05 ± 0.19	79.04 ± 0.09	92.02 ± 0.07	85.17 ± 0.04	84.57 ± 0.10
Undo-Bias	80.49 ± 0.00	69.98 ± 0.00	90.98 ± 0.00	85.95 ± 0.00	81.85 ± 0.00
UML	82.29 ± 0.00	79.54 ± 0.00	91.02 ± 0.00	84.59 ± 0.00	84.36 ± 0.00
LRE-SVM	$\mathbf{84.59 \pm 0.00}$	$\mathbf{81.17 \pm 0.00}$	91.87 ± 0.00	$\mathbf{86.38 \pm 0.00}$	$\mathbf{86.00 \pm 0.00}$
MTAE+1HNN	84.23 ± 0.11	79.30 ± 0.10	$\mathbf{92.20 \pm 0.04}$	85.98 ± 0.09	85.43 ± 0.09
D-MTAE+1HNN	$\mathbf{85.35 \pm 0.20}$	$\mathbf{80.52 \pm 0.15}$	$\mathbf{93.13 \pm 0.05}$	$\mathbf{86.15 \pm 0.08}$	$\mathbf{86.29 \pm 0.12}$