Domain Invariant Masked Autoencoders for Self-supervised Learning from Multi-domains

Haiyang Yang[1,4], Shixiang Tang[3,4], Meilin Chen[2,4], Yizhou Wang[2,4], Feng Zhu[4], Lei Bai[3], Rui Zhao[4,5], and Wanli Ouyang[3]

[1]Nanjing University, China [2]Zhejiang University, China
[3]The University of Sydney, Austrilia [4]SenseTime Research, [5]Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, China
hyyang@smail.nju.edu.cn, baisanshi@gmail.com, {yizhouwang, merlinis}@zju.edu.cn, stan3906@uni.sydney.edu.au, {zhufeng, zhaorui}@sensetime.com, wanli.ouyang@sydney.edu.au

1 More Experiments on DomainNet in Cross Domain Setting

In Table 3 (in the main text), we train our model on Painting, Real and Sketch, and evaluate the model's generalization ability on Clipart, Infograph, and Quickdraw. In this section, we evaluate our proposed DiMAE in the opposite setting in Table 3 in the main text. Specifically, we train our model on Clipart, Infograph, and Quickdraw, and then evaluate our model on Painting, Real and Sketch. Exactly following the evaluating setting in DIUL [8], we leverage the linear evaluation on 1% and 5% label fraction setting, and leverage the full network finetuning on 10% and 100% label fraction setting. We report the averaged results on 10 runs.

We present the experimental results in S-Table 1 (DomainNet) . Our DiMAE gets significant gains over DIUL [8] and other SSL methods [2, 1, 3, 6, 4] on overall and average accuracy[1]. Specifically, compared with contrastive learning based methods on Average Metrics, such as MoCo V2 [2], SimCLR V2 [1], BYOL [3], AdCo [6], our generative based method, *i.e.,* DiMAE, improves the cross-domain generalization tasks by **+4.52%** and **+2.96%** for DomainNet [7] on 1% and 5% fraction setting, respectively. Our DiMAE also improves the contrastive learning based methods on 10% and 100% label fraction setting by **+27.02%** and **+19.21%** , respectively. When we compare our DiMAE with the generative baseline method, *i.e.,* MAE [4], our DiMAE also improves **+3.07%** and **+4.13%** for DomainNet on 1% and 5% fraction setting respectively, where our model is tested by linear evaluation. Our DiMAE also improves the MAE baseline by **+11.9%** and **+18.29%** for DomainNet on 10% and 100% fraction setting, respectively, where the whole model is finetuned. The significant improvement to other states-of-the-art methods illustrates that our proposed

[1] Overall and Avg. indicate the overall accuracy of all the test data and the arithmetic mean of the accuracy of 3 domains, respectively. Note that they are different because the capacities of different domains are not equal.

S-Table 1: Results of the cross-domain generalization on DomainNet. All of the models are trained on Clipart, Infograph, Quickdraw domains of DomainNet and tested on the other three domains. The title of each column indicates the name of the domain used as target. All the models are pretrained for 1000 epoches before finetuned on the labeled data. Results style: **best**, second best.

	Label Fraction 1%					Label Fraction 5%				
method	Painting	Real	Sketch	Overall	Avg.	Painting	Real	Sketch	Overall	Avg.
ERM	6.68	6.97	7.25	6.94	6.96	7.45	6.08	5.00	6.24	6.18
MoCo V2 [2]	11.38	14.97	15.28	14.04	13.88	20.80	24.91	21.44	23.06	22.39
SimCLR V2 [1]	16.97	20.25	17.84	18.85	18.36	21.35	24.34	27.46	24.15	24.38
BYOL [3]	5.00	8.47	4.42	6.68	5.96	9.78	10.73	3.97	9.09	8.16
AdCo [6]	11.13	16.53	17.19	15.16	14.95	19.97	24.31	24.19	23.08	22.82
MAE(ViT_small) [4]	17.86	24.57	19.33	21.63	20.59	24.55	30.43	26.07	27.90	27.02
DIUL [8]	14.45	21.68	**21.30**	19.59	19.14	21.09	30.51	**28.49**	27.48	28.19
DiMAE(ViT_tiny)	15.36	23.73	13.8	19.37	17.63	18.11	28.17	14.48	22.57	20.25
DiMAE(ViT_small)	**20.18**	**30.77**	20.03	**25.63**	**23.66**	**27.02**	**39.92**	26.50	**33.59**	**31.15**
	Label Fraction 10%					Label Fraction 100%				
method	Painting	Real	Sketch	Overall	Avg.	Painting	Real	Sketch	Overall	Avg.
ERM	9.90	9.19	5.12	8.56	8.07	31.50	40.21	24.01	34.48	31.91
MoCo V2 [2]	25.35	29.91	23.71	27.37	26.32	43.42	58.61	40.38	50.66	47.47
SimCLR V2 [1]	24.01	30.17	31.58	28.75	28.59	46.79	62.32	51.05	55.71	53.39
BYOL [3]	9.50	10.38	4.45	8.92	8.11	34.02	46.48	24.82	38.59	35.11
AdCo [6]	23.35	29.98	27.57	27.65	26.97	43.55	61.42	51.23	54.37	52.07
MAE (ViT_small) [4]	41.24	54.68	39.41	47.82	45.11	53.13	68.51	48.86	60.21	56.83
DIUL [8]	25.90	33.29	30.77	30.72	29.99	49.64	63.77	54.31	57.91	55.91
DiMAE (ViT_tiny)	44.59	58.24	44.05	51.54	48.96	55.41	70.26	51.52	62.30	59.06
DiMAE (ViT_small)	**50.73**	**64.89**	**55.41**	**59.01**	**57.01**	**70.48**	**82.79**	**72.10**	**77.18**	**75.12**

DiMAE can learn more domain-invariant features from multiple domain data in the self-supervised learning.

2 Experiments with ViT_tiny backbone

ViT_tiny is a smaller backbone than ViT-small used in our paper. To illustrate the effectiveness of our proposed DiMAE, we also provide the results when we use ViT_tiny as the backbone, because ViT_tiny[2] is much smaller than current popular CNN backbone, *i.e.,* ResNet18 [5]. We evaluate our model by exactly following the setting in DIUL [8]. Specifically, we evaluate our model in three different settings on two different datasets.

[2] The number of parameters for the following backbone network: ViT_tiny 5.49M, ViT_small 21.59M, Resnet18 11.69M. We can see that ViT_tiny is much smaller than ResNet18.

S-Table 2: Results of the cross-domain generalization on DomainNet. All of the models are trained on Painting, Real, Sketch domains of DomainNet and tested on the other three domains. The title of each column indicates the name of the domain used as target. All the models are pretrained for 1000 epoches before finetuned on the labeled data. Results style: **best**, second best.

	Label Fraction 1%					Label Fraction 5%				
method	Clipart	Infograph	Quickdraw	Overall	Avg.	Clipart	Infograph	Quickdraw	Overall	Avg.
ERM	6.54	2.96	5.00	4.75	4.83	10.21	7.08	5.34	6.81	7.54
MoCo V2 [2]	18.85	10.57	6.32	10.05	11.92	28.13	13.79	9.67	14.56	17.20
SimCLR V2 [1]	23.51	15.42	5.29	11.80	14.74	34.03	17.17	10.88	17.32	20.69
BYOL [3]	6.21	3.48	4.27	4.45	4.65	9.60	5.09	6.02	6.49	6.90
AdCo [6]	16.16	12.26	5.65	9.57	11.36	30.77	18.65	7.75	15.44	19.06
MAE (ViT_small) [4]	22.38	12.62	10.50	13.51	15.17	32.60	15.28	13.43	17.85	20.44
DIUL [8]	18.53	10.62	12.65	13.29	13.93	39.32	**19.09**	10.50	18.73	22.97
DiMAE (ViT_tiny)	26.27	15.10	15.43	17.55	18.93	39.95	16.46	11.34	18.47	22.50
DiMAE (ViT_small)	**26.52**	**15.47**	**15.47**	**17.72**	**19.15**	**42.31**	18.87	**15.00**	**21.68**	**25.39**
	Label Fraction 10%					Label Fraction 100%				
method	Clipart	Infograph	Quickdraw	Overall	Avg.	Clipart	Infograph	Quickdraw	Overall	Avg.
ERM	15.10	9.39	7.11	9.36	10.53	52.79	23.72	19.05	27.19	31.85
MoCo V2 [2]	32.46	18.54	8.05	15.92	19.69	64.18	27.44	25.26	33.76	38.96
SimCLR V2 [1]	37.11	19.87	12.33	19.45	23.10	68.72	27.60	30.56	37.47	42.29
BYOL [3]	14.55	8.71	5.95	8.46	9.74	54.44	23.70	20.42	28.23	32.86
AdCo [6]	32.25	17.96	11.56	17.53	20.59	62.84	26.69	26.26	33.80	38.60
MAE (ViT_small) [4]	51.86	24.81	23.94	29.87	33.54	59.21	28.53	23.27	32.06	37.00
DIUL [8]	35.15	20.88	15.69	21.08	23.91	72.79	32.01	33.75	41.19	46.18
DiMAE (ViT_tiny)	68.58	36.14	21.08	34.95	41.93	81.42	42.57	27.89	42.88	50.63
DiMAE (ViT_small)	**70.78**	**38.06**	**27.39**	**39.20**	**45.41**	**83.87**	**44.99**	**39.30**	**49.96**	**56.05**

2.1 Evaluation on DomainNet

In this section, we evaluate our DiMAE with ViT_tiny in two settings on the DomainNet dataset, *i.e.,* pretraining on Clipart, Infograph, Quickdraw, evaluating on Painting, Real, Sketch (Clipart, Infograph, Quickdraw → Painting, Real, Sketch), and pretraining on Painting, Real, Sketch, evaluating on Clipart, Infograph, Quickdraw (Painting, Real, Sketch → Clipart, Infograph, Quickdraw).

Clipart, Infograph, Quickdraw → Painting, Real, Sketch. We exactly follow the cross-domain generalization evaluation process in DIUL [8], which is divided into three steps. First, we train our model on Clipart, Infograph, Quickdraw in the unsupervised manner. Then, we will use a small number of labeled training examples of the validation subset in Clipart, Infograph, Quickdraw to finetune the classifier or the whole network. In detail, when the fraction of labeled finetuning data is lower than 10% of the whole validation subset in the source domains, we only finetune the linear classifier for all the methods. When the fraction of labeled finetuning data is larger than 10% of the whole validation subset in the source domains, we finetune the whole network, including the backbone and the classifier. Last, we can evaluate the model on Painting, Real, Sketch.

We report our results with ViT-tiny in S-Table 1. The results in S-Table 1 show our DiMAE with ViT_tiny backbone has competitive results with other states-of-the-art methods that usually use ResNet18 or ViT_small backbones that are larger than ViT_tiny. Specifically, our DiMAE with ViT_tiny backbone improves other states-of-the-art methods, including contrastive learning based methods and generative methods, by **+3.72%** and **+3.85%** on Overall and Average Metrics for 10% label fraction, respectively. For 100% Label Fraction setting, our method using ViT_tiny backbone improves other states-of-the-art methods by **+2.09%** and **+2.23%** on Overall and Average Metrics, respectively.

Painting, Real, Sketch → Clipart, Infograph, Quickdraw. We exactly follow the cross-domain generalization evaluation process in DIUL [8], which is divided into three steps. First, we train our model on Painting, Real, Sketch in the unsupervised manner. Then, we will use a small number of labeled training examples of the validation subset in Painting, Real, Sketch to finetune the classifier or the whole network. In detail, when the fraction of labeled finetuning data is lower than 10% of the whole validation subset in the source domains, we only finetune the linear classifier for all the methods. When the fraction of labeled finetuning data is larger than 10% of the whole validation subset in the source domains, we finetune the whole network, including the backbone and the classifier. Last, we can evaluate the model on Clipart, Infograph, Quickdraw.

We report our results with ViT-tiny in S-Table 2. The results in S-Table 2 show our DiMAE with vit_tiny backbone has competitive results with other states-of-the-art methods that usually use ResNet18 or ViT_small backbones that are larger than ViT_tiny. Specifically, our DiMAE with ViT_tiny backbone improves other states-of-the-art methods, including contrastive learning based methods and generative methods, by **+4.04%** and **+3.76%** on Overall and Average Metrics for 1% label fraction, respectively. For Label Fraction 10% and 100% setting, our DiMAE with ViT_tiny backbone improves other states-of-the-art methods by **+8.39%** and **+4.45%** on Average Metric, respectively.

2.2 Evaluation on PACS

We exactly follow the protocol of [8]. Specifically, when evaluating on the "Photo" domain, we learn the backbone on the training subset of Art, Cartoon and Sketch on PACS in a self-supervised manner, and then linearly train a classifier with the backbone fixed on 1% and 5% label fraction setting, and finetune a classifier with the backbone trained on 10% and 100% label fraction setting. We evaluate our model on the validation subset in Photo, and report the averaged results by 10 runs. We leverage the similar strategy when evaluating our method on "Art", "Cartoon" and "Sketch".

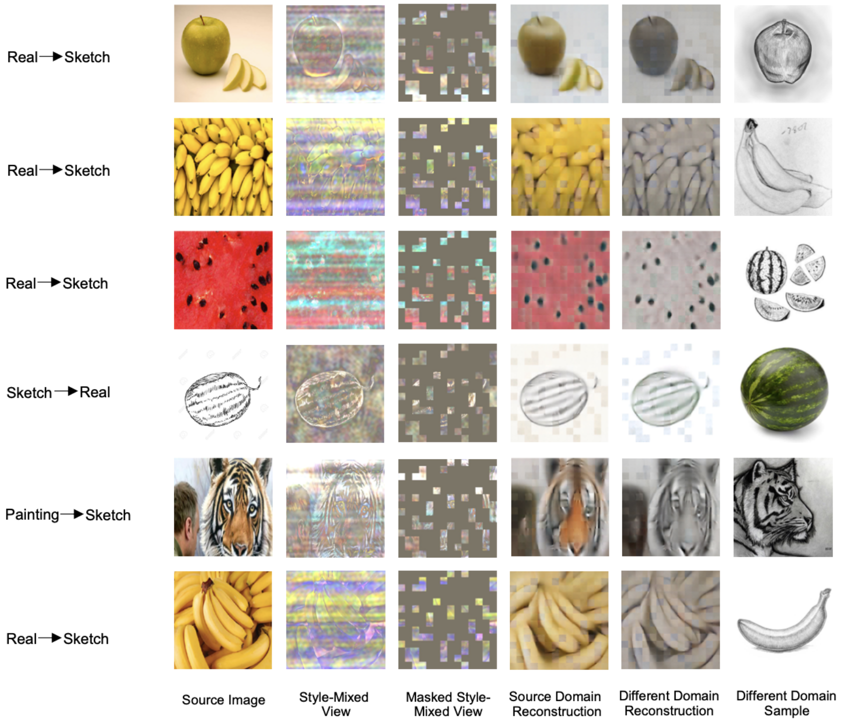
We present our results in S-Table 3. In this setting, our DiMAE achieves a better performance than previous works on most tasks and gets significant gains over DIUL and other SSL methods on average accuracy. Compared with state-of-the-art method DIUL and other SSL methods, our DiMAE improves the accuracy by **+10.88%**, **+8.78%** and **+2.24%** on average on 1%, 10% and 100% Label Fraction setting, respectively.

S-Table 3: Results of the cross-domain generalization setting on PACS. Given the experiment for each target domain runs respectively, there is no overall accuracy across domains. Thus we report the average accuracy and the accuracy for each domain. The title of each column indicates the name of the domain used as target. All the models are pretrained for 1000 epochs before finetuned on the labeled data. Results style: **best**, second best.

	Label Fraction 1%					Label Fraction 5%				
method	Photo	Art.	Cartoon	Sketch	Avg.	Photo	Art.	Cartoon	Sketch	Avg.
MoCo V2 [2]	22.97	15.58	23.65	25.27	21.87	37.39	25.57	28.11	31.16	30.56
SimCLR V2 [1]	30.94	17.43	**30.16**	25.20	25.93	**54.67**	35.92	35.31	36.84	40.68
BYOL [3]	11.20	14.53	16.21	10.01	12.99	26.55	17.79	21.87	19.65	21.47
AdCo [6]	26.13	17.11	22.96	23.37	22.39	37.65	28.21	28.52	30.35	31.18
MAE(ViT_small) [4]	30.72	23.54	20.78	24.52	24.89	32.69	24.61	27.35	30.44	28.77
DIUL [8]	27.78	19.82	27.51	29.54	26.16	44.61	39.25	**36.41**	36.53	39.20
DiMAE (ViT_tiny)	**50.48**	**41.35**	29.39	26.93	**37.04**	49.10	**44.71**	32.25	29.80	38.97
DiMAE (ViT_small)	48.86	31.73	25.83	**32.50**	34.23	50.00	41.25	34.40	**38.00**	**40.91**
	Label Fraction 10%					Label Fraction 100%				
method	Photo	Art.	Cartoon	Sketch	Avg.	Photo	Art.	Cartoon	Sketch	Avg.
MoCo V2 [2]	44.19	25.85	33.53	24.97	32.14	59.86	28.58	48.89	34.79	43.03
SimCLR V2 [1]	54.65	37.65	46.00	28.25	41.64	67.45	43.60	54.48	34.73	50.06
BYOL [3]	27.01	25.94	20.98	19.69	23.40	41.42	23.73	30.02	18.78	28.49
AdCo [6]	46.51	30.21	31.45	22.96	32.78	58.59	29.81	50.19	30.45	42.26
MAE(ViT_small) [4]	35.89	25.59	33.28	32.39	31.79	36.84	25.24	32.25	34.45	32.20
DIUL [8]	53.37	39.91	46.41	30.17	42.47	68.66	41.53	56.89	37.51	51.15
DiMAE (ViT_tiny)	61.85	**65.09**	49.79	28.28	51.25	63.75	62.74	56.51	30.56	53.39
DiMAE (ViT_small)	**77.87**	59.77	**57.72**	**39.25**	**58.65**	**78.99**	**63.23**	**59.44**	**55.89**	**64.39**

3 Visualization

We visualize more reconstruction results of DiMAE using ViT-base in S-Figure 1. The results show that the encoder of our DiMAE removes the domain style and the domain-specific decoders learn specific style information. That means DiMAE could eliminate the style noise on visible patches as no messy style information appears in reconstructions and provide complete reconstructions with specific domain styles.



S-Figure 1: Reconstruction visualization of different decoders. Sketch→Real denotes using Sketch as source domain and Real as a different domain to reconstruct.

References

1. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029 (2020)
2. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
3. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)
4. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Hu, Q., Wang, X., Hu, W., Qi, G.J.: Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1074–1083 (2021)
7. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1406–1415 (2019)
8. Zhang, X., Zhou, L., Xu, R., Cui, P., Shen, Z., Liu, H.: Domain-irrelevant representation learning for unsupervised domain generalization. arXiv preprint arXiv:2107.06219 (2021)