Self-Distilled Vision Transformer for Domain Generalization

Maryam Sultana[1,2], Muzammal Naseer[1,3], Muhammad Haris Khan[1], Salman Khan[1,3], and Fahad Shahbaz Khan[1,4]

[1] Mohamed Bin Zayed University of AI, UAE,
[2] VAIL, Oxford Brookes University, UK,
[3] Australian National University, AU,
[4] Linköping University, Sweden,
maryam.sultana, muzammal.naseer, muhammad.haris, salman.khan, fahad.khan, @mbzuai.ac.ae

Supplementary Material

More t-SNE feature visualizations: Fig. 1 (left) visualizes the class-wise feature representations of different blocks using t-SNE in baseline (ERM-ViT) and our model (ERM-SDViT) for Caltech101 target domain in the VLCS dataset. In comparison to baseline, our method facilitates improved learning of discriminative features and hence reduces the intra-class variance while increasing the inter-class variance in the feature space. Similarly, Fig. 1 (right) visualizes the same features, however, on the basis of source and target domain labels. Compared to baseline, our method promotes a greater overlap between the features of source and target domain features.

Hyperparameters analysis: We show test performance as a function of temperature (τ) and weight λ (Tab. 1). Note that, we consider the final output (in our all experimental results) which is obtained as the highest validation accuracy via grid search. The individual results of hyper-parameters mentioned in the Tab. 1 could not be considered best as they show output on the test data, hence violating the DG protocols of model performance on unseen target data.

Table 1: Analysis of temperature (τ) and weight λ with CvT-21 on PACS.

τ, λ=3.0, 0.1	τ, λ=3.0, 0.2	τ, λ=3.0, 0.5	τ, λ=5.0, 0.1	τ, λ=5.0, 0.2	τ, λ=5.0, 0.5
88.2 ± 0.3	88.2 ± 0.4	87.2 ± 0.5	88.4 ± 0.1	**89.7 ± 0.7**	88.5 ± 0.4

Confusion matrices on other DG dataset: Fig. 2 visualizes the confusion matrices for the baseline and our method on VLCS dataset. In comparison to the baseline, our method is capable of reducing false positives in all four target domains.

Attention visualizations on other DG datasets: We also visualize attention maps from different images of four datasets, including VLCS, OfficeHome,

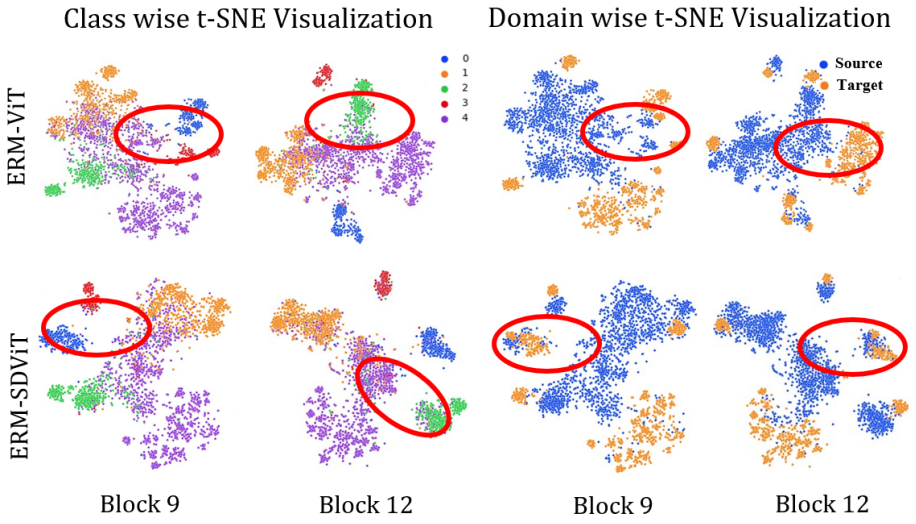


Fig. 1: t-SNE visualization of features from different blocks (9 & 12) in baseline (ERM-ViT) and our approach for Caltech101 target domain in `VLCS` dataset. Left: Features are colored corresponding to their class labels (classes: 5). Right: Features are colored corresponding to their domain labels. Our approach (ERM-SDViT) improves class-wise discrimination. For instance, in class-wise t-SNE in block 9, the features of class 0 and 3 (highlighted in red circle) are well-separated as compared to the baseline (ERM-ViT). Similarly, in block 12, the features of class 2 and 4 are clearly distinguishable. In domain-wise t-SNE, for our approach, source and target domain features show greater overlap with each other.

`TerraIncognita` and `DomainNet` in Figs. 3, 4 and 5. It can be observed that in all target domains of the four datasets, our method mostly relies on features corresponding to the foreground object's semantics rather than the background information. However, the baseline approach (ERM-ViT) mostly capitalizes more on the background features and pays less attention to the features belonging to the foreground object. For instance, in Fig. 3, target domain: Location_46 of the `TerraIncognita` dataset, our method is capable of focussing on the foreground object (a dog), which occupies a small fraction of the overall image. However, the baseline model is prone to attending more to the background features, which are prevalent in the image. Note that the attention maps are computed at the final block of ViT models.

Recognition accuracy on target domains of other DG datasets: Tables 2 and 3 compares target domain-wise recognition accuracy on `VLCS`, `OfficeHome`, `TerraIncognita`, and `DomainNet` datasets of our method with the baseline utilizing three ViT backbones and a DG baseline (T3A [1]).

Training overhead on target domains of other DG datasets: Table 4 and 5 reports training overhead, computed as relative % increase in training time (hrs.) on `TerraIncognita` and `DomainNet` datasets. The numbers report

Table 2: Comparison of target domain-wise classification accuracy on `VLCS`, `OfficeHome`, and `TerraIncognita` datasets. Results are reported of our method with the baseline using three different ViT backbones, including DeiT-Small [2], CvT-21 [3], and T2T-ViT-14 [4], and a DG baseline (T3A [1]).

Dataset			`VLCS`				
Model	Backbone	# of Params	Caltech101	LableMe	SUN09	VOC2007	Average
ERM	ResNet-50	23.5M	98.1 ± 0.4	64.1 ± 0.5	70.7 ± 0.9	74.8 ± 2.4	76.9 ± 0.6
ERM-ViT	DeiT-Small	22M	96.7 ± 0.8	65.2 ± 1.0	73.9 ± 0.3	77.4 ± 0.3	78.3 ± 0.5
ERM-SDViT	DeiT-Small	22M	96.8 ± 0.5	64.2 ± 0.8	76.2 ± 0.4	78.5 ± 0.4	78.9 ± 0.4
ERM-SDViT + T3A	DeiT-Small	22M	98.9 ± 0.2	65.9 ± 0.3	79.8 ± 0.4	81.9 ± 0.4	81.6 ± 0.1
ERM-ViT	CvT-21	32M	97.3 ± 0.5	65.2 ± 0.9	76.6 ± 1.1	76.9 ± 0.3	79.0 ± 0.3
ERM-SDViT	CvT-21	32M	96.5 ± 0.7	63.3 ± 0.4	78.1 ± 0.2	78.9 ± 0.8	79.2 ± 0.4
ERM-SDViT + T3A	CvT-21	32M	98.4 ± 0.3	66.8 ± 0.5	80.1 ± 1.0	80.6 ± 0.7	**81.9 ± 0.4**
ERM-ViT	T2T-ViT-14	21.5M	96.5 ± 0.5	64.5 ± 0.1	76.4 ± 0.4	78.2 ± 1.0	78.9 ± 0.3
ERM-SDViT	T2T-ViT-14	21.5M	96.9 ± 0.4	64.0 ± 0.5	76.7 ± 1.4	80.4 ± 1.3	79.5 ± 0.8
ERM-SDViT + T3A	T2T-ViT-14	21.5M	98.6 ± 0.3	66.5 ± 0.7	78.2 ± 0.5	81.7 ± 0.9	81.2 ± 0.3
Dataset			`OfficeHome`				
Model	Backbone	# of Params	Art	Clipart	Product	Real World	Average
ERM	ResNet-50	23.5M	58.8 ± 1.0	51.3 ± 0.4	73.7 ± 0.4	74.7 ± 0.6	64.6 ± 0.2
ERM-ViT	DeiT-Small	22M	67.6 ± 0.3	57.0 ± 0.6	79.4 ± 0.1	81.6 ± 0.4	71.4 ± 0.1
ERM-SDViT	DeiT-Small	22M	68.3 ± 0.8	56.3 ± 0.2	79.5 ± 0.3	81.8 ± 0.1	71.5 ± 0.2
ERM-SDViT + T3A	DeiT-Small	22M	69.1 ± 1.0	57.9 ± 0.4	80.7 ± 0.0	82.3 ± 0.1	72.5 ± 0.3
ERM-ViT	CvT-21	32M	74.4 ± 0.2	59.8 ± 0.5	83.5 ± 0.4	84.1 ± 0.2	75.5 ± 0.0
ERM-SDViT	CvT-21	32M	73.8 ± 0.6	60.7 ± 0.9	83.0 ± 0.3	85.0 ± 0.3	75.6 ± 0.2
ERM-SDViT + T3A	CvT-21	32M	75.2 ± 0.7	62.7 ± 0.8	84.2 ± 0.6	86.1 ± 0.0	**77.0 ± 0.2**
ERM-ViT	T2T-ViT-14	21.5M	70.2 ± 0.5	59.0 ± 0.6	81.9 ± 0.3	83.6 ± 0.6	73.7 ± 0.2
ERM-SDViT	T2T-ViT-14	21.5M	71.1 ± 0.5	59.2 ± 0.3	82.8 ± 0.4	83.5 ± 0.3	74.2 ± 0.3
ERM-SDViT + T3A	T2T-ViT-14	21.5M	70.8 ± 0.4	61.9 ± 0.7	84.1 ± 0.2	85.0 ± 0.3	75.5 ± 0.2
Dataset			`TerraIncognita`				
Model	Backbone	# of Params	location_38	location_43	location_46	location_100	Average
ERM	ResNet-50	23.5M	56.3 ± 1.1	36.8 ± 4.6	52.6 ± 0.4	35.2 ± 1.7	45.2 ± 1.2
ERM-ViT	DeiT-Small	22M	50.2 ± 1.4	30.6 ± 0.9	53.2 ± 0.2	39.6 ± 1.0	43.4 ± 0.5
ERM-SDViT	DeiT-Small	22M	55.9 ± 1.7	31.7 ± 2.6	52.2 ± 0.3	37.4 ± 0.6	44.3 ± 1.0
ERM-SDViT + T3A	DeiT-Small	22M	53.8 ± 1.2	36.2 ± 1.0	51.1 ± 1.0	38.5 ± 1.3	44.9 ± 0.4
ERM-ViT	CvT-21	32M	51.4 ± 1.8	40.1 ± 1.7	57.6 ± 1.0	45.7 ± 0.6	48.7 ± 0.4
ERM-SDViT	CvT-21	32M	53.6 ± 3.3	42.7 ± 1.6	58.2 ± 1.0	44.5 ± 1.8	49.7 ± 1.4
ERM-SDViT + T3A	CvT-21	32M	58.1 ± 0.7	46.2 ± 0.3	57.0 ± 1.0	44.1 ± 2.2	**51.4 ± 0.7**
ERM-ViT	T2T-ViT-14	21.5M	52.5 ± 1.7	43.0 ± 1.3	53.7 ± 1.1	43.0 ± 1.6	48.1 ± 0.2
ERM-SDViT	T2T-ViT-14	21.5M	57.2 ± 2.9	45.4 ± 2.4	57.7 ± 0.8	41.9 ± 0.4	50.6 ± 0.8
ERM-SDViT + T3A	T2T-ViT-14	21.5M	59.3 ± 1.2	48.2 ± 1.0	53.1 ± 0.9	41.5 ± 0.2	50.5 ± 0.6

Table 3: Comparison of target domain-wise classification accuracy on `DomainNet` dataset. Results are reported of our method with the baseline using three different ViT backbones, including DeiT-Small [2], CvT-21 [3], and T2T-ViT-14 [4], and a DG baseline (T3A [1]).

Dataset			`DomainNet`					
Model	Backbone	# of Params	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Average
ERM	ResNet-50	23.5M	57.6 ± 0.6	18.5 ± 0.3	45.9 ± 0.7	11.6 ± 0.1	59.5 ± 0.3	48.6 ± 0.3	40.3 ± 0.1
ERM-ViT	DeiT-Small	22M	62.9 ± 0.2	23.3 ± 0.1	53.1 ± 0.2	15.7 ± 0.1	65.7 ± 0.1	52.4 ± 0.2	45.5 ± 0.0
ERM-SDViT	DeiT-Small	22M	63.4 ± 0.1	22.9 ± 0.0	53.7 ± 0.1	15.0 ± 0.4	67.4 ± 0.1	52.6 ± 0.2	45.8 ± 0.0
ERM-SDViT + T3A	DeiT-Small	22M	64.3 ± 0.2	23.7 ± 0.0	54.2 ± 0.3	19.7 ± 0.4	69.6 ± 0.1	53.2 ± 0.2	47.4 ± 0.1
ERM-ViT	CvT-21	32M	69.0 ± 0.2	27.2 ± 0.2	58.4 ± 0.2	17.1 ± 0.3	71.6 ± 0.1	59.2 ± 0.3	50.4 ± 0.1
ERM-SDViT	CvT-21	32M	68.9 ± 0.1	26.7 ± 0.3	58.0 ± 0.1	17.3 ± 0.1	71.9 ± 0.0	59.1 ± 0.3	50.4 ± 0.0
ERM-SDViT + T3A	CvT-21	32M	69.7 ± 0.1	27.6 ± 0.2	58.7 ± 0.1	23.0 ± 0.1	73.6 ± 0.2	59.6 ± 0.1	**52.0 ± 0.0**
ERM-ViT	T2T-ViT-14	21.5M	67.0 ± 0.3	25.2 ± 0.2	55.3 ± 0.3	15.3 ± 0.2	70.3 ± 0.1	55.9 ± 0.2	48.1 ± 0.1
ERM-SDViT	T2T-ViT-14	21.5M	67.6 ± 0.2	25.0 ± 0.2	55.8 ± 0.4	15.2 ± 0.3	70.0 ± 0.1	55.9 ± 0.1	48.2 ± 0.2
ERM-SDViT + T3A	T2T-ViT-14	21.5M	68.2 ± 0.1	25.8 ± 0.2	56.7 ± 0.3	20.7 ± 0.2	72.4 ± 0.1	57.0 ± 0.2	50.2 ± 0.1

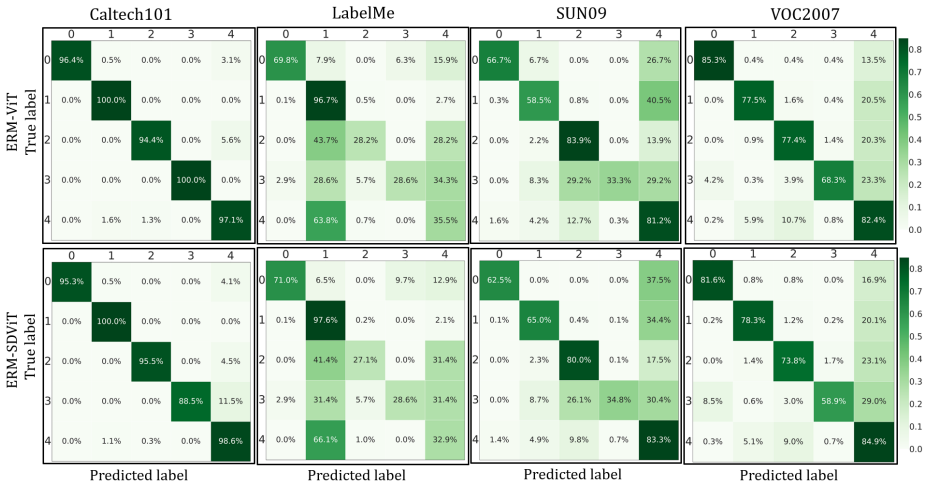
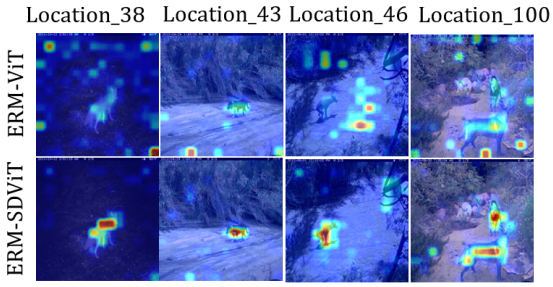


Fig. 2: Confusion matrices of the baseline and our method on `VLCS` dataset. The classes in the figure are '0':Bird, '1':Car, '2':Chair, '3':Dog, and '4':Person.

Fig. 3: Comparison of attention maps between the baseline (ERM-ViT) and our proposed method (ERM-SDViT) on four target domains of `TerraIncognita` dataset. The ViT backbone is DeiT-Small.



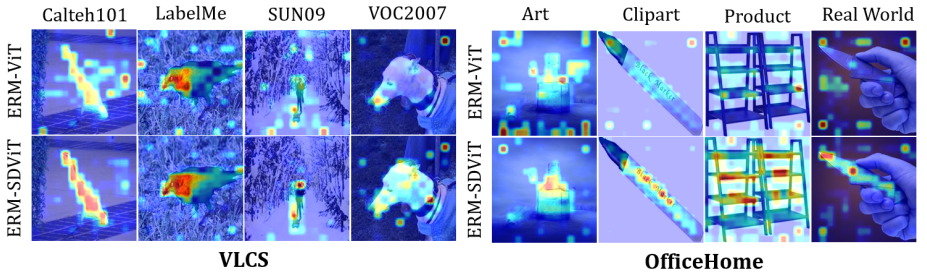


Fig. 4: Comparison of attention maps between the baseline (ERM-ViT) and our proposed method (ERM-SDViT) on four target domains of `VLCS` and `OfficeHome` datasets. The ViT backbone is DeiT-Small.

the training time increase introduced by our method on top of the baseline. The results show that in both large-scale DG benchmark datasets i.e. `TerraIncognita` (24K images) and `DomainNet` (500K images), our model (ERM-SDViT) is not

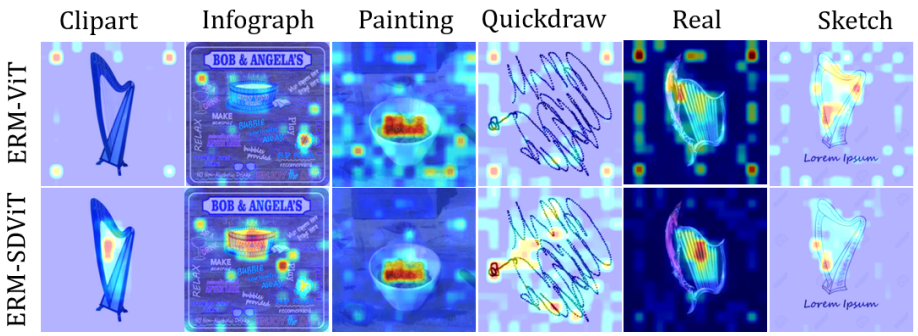


Fig. 5: Comparison of attention maps between the baseline (ERM-ViT) and our proposed method (ERM-SDViT) on six target domains of `DomainNet` datasets. The ViT backbone is DeiT-Small.

Table 4: Training overhead, computed as relative % increase in training time (hrs.), introduced by our method on top of the baseline.

Dataset:	`TerraIncognita`			
Model	Location_38	Location_43	Location_46	Location_100
ERM-ViT	0.268	0.268	0.270	0.268
ERM-SDViT	0.276	0.282	0.282	0.302
Rel.overhead	2.975	5.068	4.447	12.620

Table 5: Training overhead, computed as relative % increase in training time (hrs.), introduced by our method on top of the baseline.

Dataset:	`DomainNet`					
Model	Clipart	Infograph	Painting	Quickdraw	Real	Sketch
ERM-ViT	0.418	0.423	0.422	0.430	0.436	0.430
ERM-SDViT	0.482	0.444	0.510	0.469	0.446	0.460
Rel.overhead	15.376	5.124	20.928	9.079	2.463	7.089

exceeding more than 20% relative overhead training time. Note that this training time could differ with GPU utilization. Results are reported with DeiT-Small (22M params.) backbone on Nvidia RTX A6000 GPU.

References

1. Iwasawa, Y., Matsuo, Y.: Test-time classifier adjustment module for model-agnostic domain generalization. Advances in Neural Information Processing Systems **34** (2021)
2. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, PMLR (2021) 10347–10357

3. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 22–31
4. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 558–567