TVT: Transferable Vision Transformer for Unsupervised Domain Adaptation

Jinyu Yang[1], Jingjing Liu[2], Ning Xu[2], and Junzhou Huang[1]
[1]University Of Texas at Arlington, [2]Kuaishou Technology
jinyu.yang@mavs.uta.edu, jjliu08cas@gmail.com, ningxu01@gmail.com, jzhuang@uta.edu

1. Supplementary

Since there does not exist a golden-standard vision transformer backbone in unsupervised domain adaptation (UDA), several contemporary work [14, 12, 9] use DeiT [11] and Swin [7] in their experiments. Specifically, CDTrans [14] and WinTR [9] use DeiT-S and DeiT-B, while BCAT [12] uses Swin-B. Table 1 shows the detailed comparison between different vision transformers. It is noteworthy that previous CNN-based methods mainly use ResNet-50 (23M parameters) as the backbone in the Office-Home and Office-31 datasets, and use ResNet-101 (45M parameters) in the VisDA-2017 dataset. For a fair comparison with CNN-based counterparts, it is critical to guarantee a comparable backbone size. According to Table 1, DeiT-S or Swin-T can be used as the backbone in the Office-Home and Office-31 datasets, while Swin-S can be used in the VisDA-2017 dataset. Therefore, we replace ViT [3] in our proposed framework with Swin to check its generalization ability.

1.1. Swin-based TVT

Different from ViT which uses an additional class token for classification, Swin applies a global average pooling (GAP) layer on the output patches f_{ir} of the last stage, where i and r indicate the index of the input image and image patches, respectively. To adopt Swin in our TVT framework, we apply the patch-level domain discriminator D_l to f_{ir} and obtain the transferability of each output patch by $t_{ir} = H(D_l(f_{ir})) \in [0, 1]$, where $H(\cdot)$ is the standard entropy function. After that, we apply the element-wise multiplication to f_{ir} and t_{ir}, followed by a GAP layer and a linear classifier.

1.2. Implementation Details

The Swin-T and Swin-S with 4×4 input patch size and 7×7 window size [7] pre-trained on ImageNet-1K [2] are used as our backbone. The architecture hyper-parameters of these two backbones are as follows:

- Swin-T: $C = 96$, layer numbers = {2, 2, 6, 2}
- Swin-S: $C = 96$, layer numbers ={2, 2, 18, 2},

Methods	Image Size	#parameters	FLOPs	throughput	acc.
ViT-B/16 [3]	384^2	86M	55.4G	85.9	77.9
DeiT-S [11]	224^2	22M	4.6G	940.4	79.8
DeiT-B [11]	224^2	86M	17.5G	292.3	81.8
Swin-T [7]	224^2	29M	4.5G	755.2	81.3
Swin-S [7]	224^2	50M	8.7G	436.9	83.0
Swin-B [7]	224^2	88M	15.4G	278.1	83.5

Table 1. Comparison of different vision transformers pretrained on ImageNet-1K, where acc. indicates the ImageNet top-1 accuracy

where C is the embedding dimension. We train Swin-based TVT using mini-batch Stochastic Gradient Descent (SGD) optimizer with the momentum of 0.9. We initialize the learning rate as 0 and linearly increase it to $lr = 0.003$ after 500 training steps. We then decrease it by the cosine decay strategy. The total training step is 5,000. For the Office-Home and Office-31 datasets, we set $\alpha = \beta = \gamma = 0.1$. For the Visda-2017 dataset, we set $\alpha = 1.0$, $\beta = 0.1$, and $\gamma = 1.0$.

1.3. Results

As shown in Table 2 3, 4, our Swin-based TVT outperforms CNN-based counterparts by a large margin. Note that the complexity of Swin-T and Swin-S are similar to ResNet-50 (DeiT-S) and ResNet-101, respectively. Therefore, the comparison is guaranteed to be fair in terms of the backbone complexity. Furthermore, Swin-based TVT achieves very competitive performance compared with existing vision transformer-based methods, indicating the generalization ability of TVT to various vision transformer backbones.

1.4. Attention Visualization

We visualize the attention map of the class token in TAM to verify that our model can attend to local features that are both transferable and discriminative. Without loss of generality, we randomly sample target-domain images in VisDA-2017 dataset for comparison. As shown in Figure 1, our method captures more accurate regions than Source Only and Baseline. For instance, to recognize the person in the top-left image, Source Only mainly focus on women's

Algorithm		A→ W	D→ W	W→ D	A→ D	D→ A	W→ A	Avg
Source Only	ResNet-50	68.4	96.7	99.3	68.9	62.5	60.7	76.1
TADA [13]		94.3	98.7	99.8	91.6	72.9	73.0	88.4
TAT [6]		92.5	99.3	100.0	93.2	73.1	72.1	88.4
SHOT [5]		90.1	98.4	99.9	94.0	74.7	74.3	88.6
ALDA [1]		95.6	97.7	100.0	94.0	72.2	72.5	88.7
Source Only-S	DeiT	86.9	97.7	99.6	87.6	74.9	73.5	86.7
CDTrans-S [14]		93.5	98.2	99.6	94.6	78.4	78.0	90.4
Source Only-T	Swin	85.5	99.2	100.0	87.6	73.9	72.3	86.4
TVT-T		96.9	99.2	100.0	96.6	79.1	78.9	91.8

Table 2. Performance comparison on the Office-31 dataset

Algorithm		A→CA	→PA	→RC	→AC	→PC	→RP	→AP	→CP	→RR	→AR	→CR	→P	Avg
Source Only	ResNet-50	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [8]		43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
ALDA [1]		53.7	70.1	76.4	60.2	72.6	71.5	56.8	51.9	77.1	70.2	56.3	82.1	66.6
TADA [13]		53.1	72.3	77.2	59.1	71.2	72.1	59.7	53.1	78.4	72.4	60.0	82.9	67.6
SHOT [5]		57.1	78.1	81.5	68.0	78.2	78.1	67.4	54.9	82.2	73.3	58.8	84.3	71.8
Source Only-S	DeiT	55.6	73.0	79.4	70.6	72.9	76.3	67.5	51.0	81.0	74.5	53.2	82.7	69.8
CDTrans-S [14]		60.6	79.5	82.4	75.6	81.0	82.3	72.5	56.7	84.4	77.0	59.1	85.5	74.7
WinTR-S [9]		65.3	84.1	85.0	76.8	84.5	84.4	73.4	60.0	85.7	77.2	63.1	86.8	77.2
Source Only-T	Swin	54.8	73.6	80.9	67.6	74.7	76.8	66.2	49.7	81.5	72.4	52.8	82.3	69.4
TVT-T		63.1	82.5	86.2	76.7	82.9	83.9	75.1	60.4	86.3	77.3	63.8	87.3	77.1

Table 3. Performance comparison on the Office-Home dataset

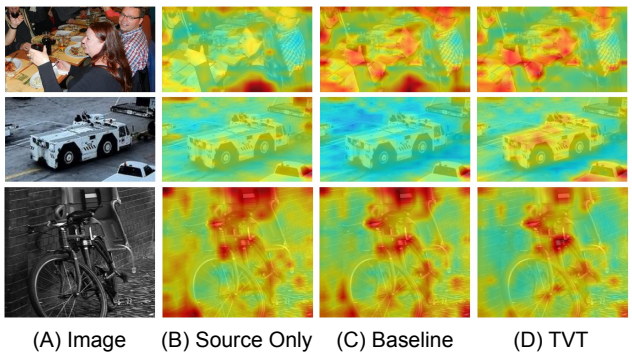

Figure 1. Attention map visualization of person, truck, and bicycle in VisDA-2017 dataset. The hotter the color, the higher the attention

shoulder which is discriminative yet not highly transferable. Moving beyond the shoulder region, the baseline also attends to faces and hands that can generalize well across domains. Our method, instead, ignores the shoulder and only highlight those regions that are important for classification and transferable. Certainly, by leveraging the intrinsic attention mechanism and fine-grained features captured by sequential patches, our method promotes the capability of ViT in transferring domain knowledge.

References

[1] Minghao Chen, Shuai Zhao, Haifeng Liu, and Deng Cai. Adversarial-learned loss for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Seungmin Lee, Dongwan Kim, Namil Kim, and Seong-Gyun Jeong. Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 91–100, 2019.

[5] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *In-*

Algorithm		plane	bcycl	bus	car	house	knife	mcycl	person	plant	sktbrd	train	truck	Avg
Source Only	ResNet-101	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
MCD [10]		87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
ALDA [1]		93.8	74.1	82.4	69.4	90.6	87.2	89.0	67.6	93.4	76.1	87.7	22.2	77.8
DTA [4]		93.7	82.2	85.6	83.8	93.0	81.0	90.7	82.1	95.1	78.1	86.4	32.1	81.5
SHOT [5]		94.3	88.5	80.1	57.3	93.1	94.9	80.7	80.3	91.5	89.1	86.3	58.2	82.9
Source Only-S	Swin	97.5	52.4	85.9	65.2	60.0	54.5	93.2	20.1	82.3	54.3	94.1	31.8	65.9
TVT-S		97.6	88.0	86.0	64.4	95.0	96.6	89.7	76.6	94.9	79.4	91.2	64.1	85.3

Table 4. Performance comparison on the VisDA-2017 dataset

ternational Conference on Machine Learning, pages 6028–6039. PMLR, 2020.

[6] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, pages 4013–4022. PMLR, 2019.

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[8] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[9] Wenxuan Ma, Jinming Zhang, Shuang Li, Chi Harold Liu, Yulin Wang, and Wei Li. Exploiting both domain-specific and invariant knowledge via a win-win transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2111.12941*, 2021.

[10] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018.

[11] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[12] Xiyu Wang, Pengxin Guo, and Yu Zhang. Domain adaptation via bidirectional cross-attention transformer. *arXiv preprint arXiv:2201.05887*, 2022.

[13] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[14] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021.