A2XP: Towards Private Domain Generalization

Supplementary Material

1. Implementation of Generalization

In this section, we present detailed implementation of the *Attention-based Generalization* module in a pseudo-code form from initialization to forwarding Algorithm 1.

Algorithm 1 Generalization Implementation

```
procedure INIT(self, $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_i, \cdots, \mathbf{p}_N$)
    self.$\mathcal{E}_{\text{shared}} \leftarrow$ resnet18_1k()    ▷ Initialize embedders.
    self.$\mathcal{E}_{\text{T}}$, self.$\mathcal{E}_{\text{E}} \leftarrow$ linear(), linear()
    self.$\mathbf{p}_i \leftarrow \mathbf{p}_i / \|\mathbf{p}_i\|_2 \quad \forall i \in [1, N]$    ▷ Normalize experts.
end procedure

procedure FORWARD(self, $\mathbf{x}_{N+1,j}$)
    $\mathbf{z}_{\mathbf{x}} \leftarrow$ self.$\mathcal{E}_{\text{T}}$(self.$\mathcal{E}_{\text{shared}}(\mathbf{x}_{N+1,j})$)
    $\mathbf{z}_{\mathbf{p}_i} \leftarrow$ self.$\mathcal{E}_{\text{E}}$(self.$\mathcal{E}_{\text{shared}}$(self.$\mathbf{p}_i$))    $\forall i \in [1, N]$
    $\lambda_i \leftarrow \mathbf{z}_{\mathbf{x}} \mathbf{z}_{\mathbf{p}_i}^{\top} \quad \forall i \in [1, N]$    ▷ Calculate attention scores.
    $\mathbf{p}_{N+1,j} \leftarrow \sum_{i=1}^{N} \lambda_i$self.$\mathbf{p}_i$
    return $\mathbf{x}_{N+1,j} + \mathbf{p}_{N+1,j}$
end procedure
```

2. Further Analysis on the Experts

In this section, we conduct further analysis on the expert prompts of A2XP. We analyzed the prompts by changing various components: the size of the experts, the number of experts, the type of prompts, the way to mix the experts.

2.1. Size of the Experts

We analyzed the prompt size in the performance and the memory requirement perspectives (see Figure 1). We empirically found that 30 is the best prompt size among the five sizes and applied it to our method.

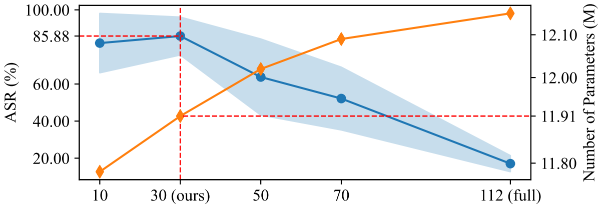


Figure 1. Expert adaptation performance of A2XP with Gaussian initialization. The blue transparent range shows $\mu \pm \sigma$ of ASR.

2.2. Ablation Study on Various Experts

We compared domain generalization performance among various experts.

2.2.1 Various Prompts

Generalization *without* prompts, which is equivalent to linear probing, loses the benefits of linear combination; therefore, it has a lower generalization performance (see Table 1). Utilizing *random* prompts show performance improvement, indicating that prompts can contribute to better generalization performance. Furthermore, we show the effectiveness of our method that enhances generalization performance more efficiently by leveraging *experts* trained from each domain.

	Picture	Art	Cartoon	Sketch	Avg.
Without	86.95	83.11	94.04	86.79	87.72 (-7.35)
Random	98.98	93.85	90.19	88.09	92.78 (-2.29)
Experts	**99.07**	**95.27**	**98.07**	**87.85**	**95.07** (-0.00)

Table 1. Comparison by various prompts.

2.2.2 Number of Experts

As shown in Table 2, using either no experts or just a single expert offers limited generalization potential. In contrast, employing multiple experts, particularly in numbers matching the domain count, broadens the scope to identify the optimal direction for generalization.

# experts	Picture	Art	Cartoon	Sketch	Avg.
0	86.95	83.11	94.04	86.79	87.72 (-7.35)
1	83.75	95.69	86.82	86.49	88.19 (-6.88)
2	97.35	**99.34**	93.41	87.27	94.34 (-0.73)
3 (all)	**99.07**	95.27	**98.07**	**87.85**	**95.07** (-0.00)

Table 2. Comparison by # experts.

2.3. Various Ways to Mix Experts

In Table 3, we compared various methods to mix pre-trained experts for unseen datasets. We demonstrate that our *attention*-based approach outperforms methods that mix experts in *constant* or *random* weights.

	Picture	Art	Cartoon	Sketch	Avg.
Constant	97.82	99.40	94.24	86.99	94.61 (-0.46)
Random	97.65	99.16	93.85	87.05	94.43 (-0.64)
Attention	**99.07**	**95.27**	**98.07**	**87.85**	**95.07** (-0.00)

Table 3. Various ways to mix experts.

3. Further Analysis on the Framework

We analyzed more about the A2XP framework itself in the perspective of how evenly the experts are mixed, how does the objective network architecture affects, and the scalability of A2XP.

3.1. Attention Distribution

When A2XP is applied on the source domain, we expected the attention weights of A2XP emphasize the experts of the source domain. This study analyzes how A2XP attends to different experts depending on the domain of the input images. The violin plots in Figure 2 show the distribution of normalized attention weights in PACS [7] dataset. Each cell shows the distribution of attention weights on each domain. Across all combinations of target and source domains, a significant standard deviation was observed, indicating a wide range of variation in the attention weights. This suggests that the attention weights have a very large range.

	P	A	C	S
P	1.729E-1	**1.330E-2**	3.424E-1	**2.377E-4**
A	4.966E-1	5.752E-2	**4.210E-2**	5.739E-2
C	**2.127E-2**	**1.641E-3**	1.759E-1	**1.797E-2**
S	2.556E-1	2.526E-1	5.566E-1	**2.460E-9**

Table 4. p-values of RM-ANOVA [9] with the normalized attention weights on PACS [7] dataset. Bold styled cells are significant with $p \leq 0.05$.

To be analytic, we performed Repeated Measures-ANalysis Of VAriance (RM-ANOVA) [9] on the normalized attention weights, and the result is in Table 4. Each cell contains the p-value of a combination of the target domain and tested domain. For example, p-value of weights when trained on 'P' and tested on 'A' is 1.330E-2. In this case, the experts are from the 'A,' 'C,' and 'S' domains. The smaller a p-value is, the more the combination showed a significant correlation among weights for experts. The p-values are significant with $p \leq 0.05$ in some cases but not dominant. As a result, A2XP mixes the experts differently depending to the input images, and the mixing ratios are not always similar even if the target and testing domain is the same.

3.2. Various Objective Networks

We are concerned only about CLIP [8]-pretrained Vision Transform (ViT) [3] for the objective network in the main paper. We present another result on a convolutional neural network ResNet50 [4] and ImageNet [2] supervised pretraining to reveal another characteristic of A2XP. The leave-one-domain-out evaluation result is compared in Table 5. The number of updates was limited to 3K for ImageNet and 1K for CLIP pretrained models in the adaptation step. And we initialized the experts by zero before adaptation.

We observed that the experts must be well adapted for all domain from ResNet50 with both ImageNet and CLIP pretraining. Moreover, even if the adaptation was successful, the model itself have to be generalized at the pretext task. Both the average accuracy of the both ResNet50 was lower compared to other existing methods [1, 5]. As a result, A2XP is sensitive to the adaptation method, the objective network architecture, and the pretext task.

Architecture	Pretraining	Expert Adaptation P	A	C	S	Avg.
ResNet50 [4]	ImageNet [2]	92.40	72.36	85.24	66.28	79.07
ResNet50 [4]	CLIP [8]	67.25	52.83	59.98	56.73	59.20
ViT-base [3]	ImageNet [2]	96.95	79.30	92.41	87.94	89.15
ViT-base [3]	CLIP [8]	97.54	73.88	95.52	94.55	90.37
Architecture	**Pretraining**	**Attention-based Generalization P**	**A**	**C**	**S**	**Avg.**
ResNet50 [4]	ImageNet [2]	51.56	49.12	46.25	36.12	45.76
ResNet50 [4]	CLIP [8]	74.31	44.38	42.62	16.34	44.41
ViT-base [3]	ImageNet [2]	81.02	69.53	49.23	31.38	57.79
ViT-base [3]	CLIP [8]	99.07	95.07	98.12	88.22	95.12

Table 5. The result of leave-one-domain-out evaluation using ViT [3] and ResNet50 [4].

3.3. Scalability

We applied our method to larger datasets: Office-Home (Table 6) and DomainNet (Table 7). The results show that A2XP outperforms current methods, validating its applicability across datasets of varying sizes.

	Art	Clipart	Product	Real	Avg.
ERM	48.04	42.27	48.25	47.63	46.55
MIRO	56.49	58.56	43.30	54.43	53.20
A2XP	**77.42**	**65.73**	**81.93**	**83.15**	**77.06**

Table 6. Office-Home evaluation.

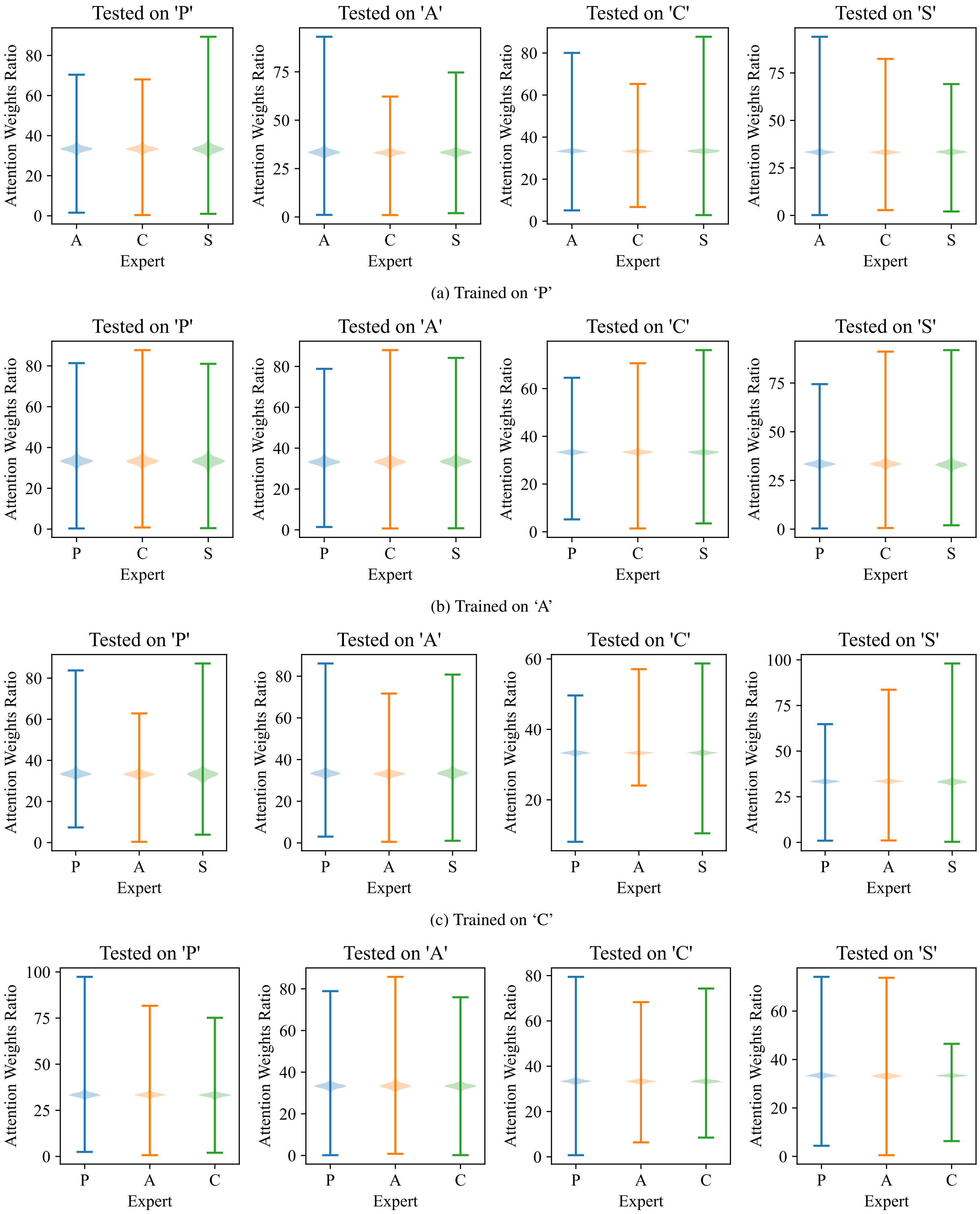
	Clip	Info	Paint	Quick	Real	Sketch	Avg.
ERM	0.32	0.35	0.45	0.39	0.41	0.57	0.41
MIRO	39.31	39.48	40.10	**39.77**	40.59	42.18	40.24
A2XP	**62.88**	**43.58**	**58.99**	13.72	**55.96**	**58.45**	**48.93**

Table 7. DomainNet evaluation.

We further investigated the scalability of A2XP for datasets of various sizes by measuring the number of parameters, memory requirements, computational resources (GFLOPs), and the training time (see Table 8). The number of parameters and the memory requirement of A2XP only depends on the number of experts. Training time primarily depends on the number of training samples. From this perspective, we show the practical applicability of A2XP for larger datasets.

Dataset	# classes	# domains	# samples	# params	Mem. load	GFLOPs	Time (s)	Avg. Acc.
PACS	7	4	9,991	11.91M	17817MiB	1.814	2.12	95.07
VLCS	5	4	10,729	11.91M	17777MiB	1.814	2.51	83.15
Office-Home	65	4	15,588	11.91M	17779MiB	1.814	2.87	77.06
DomainNet	345	6	586,575	12.05M	18185MiB	2.539	136.46	48.93

Table 8. Scalability analysis.



(a) Trained on ‘P’

(b) Trained on ‘A’

(c) Trained on ‘C’

(d) Trained on ‘S’

Figure 2. Visualization of normalized attention weights of correctly classified samples from A2XP on PACS [7] dataset.

4. Discussion

4.1. Failure Cases

We found that A2XP struggles to generalize for specific domains such as *Quick* in the DomainNet dataset and *Sketch* domain in the PACS dataset; both have more significant domain shifts from another dataset. Despite limitations in generalizing distinct domains, the performance of our approach still achieved state-of-the-art average accuracy.

4.2. Interpretability

As noted in [6], a visual prompt facilitates domain adaptation by aligning features between the source and target domains. This can be interpreted as our experts are responsible for shifting features towards the target domain's manifolds. In our privacy setting, the alignment target is the manifold characterized by features from the data used to pre-train the model. Consequently, generating an expert for an unseen domain by mixing the experts from other domains can be considered crafting a mapping function to the pre-trained manifold, which we interpret as contributing to enhancing decision-making when using a pre-trained model while keeping it private.

References

[1] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. In *European Conference on Computer Vision*, pages 440–457. Springer, 2022. 2

[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, 2009. 2

[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. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] Samyak Jain, Sravanti Addepalli, Pawan Kumar Sahu, Priyam Dey, and R. Venkatesh Babu. Dart: Diversify-aggregate-repeat training improves generalization of neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16048–16059, 2023. 2

[6] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 4

[7] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 2, 3

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[9] Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951. 2