FDS: Feedback-guided Domain Synthesis with Multi-Source Conditional Diffusion Models for Domain Generalization - Supplementary Material

Mehrdad Noori* Milad Cheraghalikhani
Ali Bahri Gustavo A. Vargas Hakim David Osowiechi Moslem Yazdanpanah
Ismail Ben Ayed Christian Desrosiers

ÉTS Montreal, Canada
International Laboratory on Learning Systems (ILLS)

1. Implementation

Our proposed FDS method is built using the Python language and the PyTorch framework. We utlized four NVIDIA A100 GPUs for all our experiments. For initializing our models, we utilize the original Stable Diffusion version 1.5 as our initial weight [20]. The key hyperparameter configurations employed for training these diffusion models and generating new domains are detailed in Tables 1 and 2, respectively.

Furthermore, for classifier training, we adhere to the methodologies and parameter settings described by Cha et al. [4], ensuring consistency and reproducibility in our experimental setup. The original implementation and instructions for reproducing our results are accessible via https://github.com/Mehrdad-Noori/FDS.

2. Additional Ablation

Selection/Filtering. In this section, we provide visual examples to show the efficacy of our synthetic sample selection and filtering mechanism. As mentioned in the method section, this mechanism is intricately designed to scrutinize the generated images through two lenses: the alignment of the predicted class with the intended label, and the entropy indicating the prediction's uncertainty.

The Figures 2, 3, 4 showcase a set of images generated from interpolations between two domains. Specifically, the diffusion model is trained on *"art"*, *"sketch"*, and *"photo"* of the PACS dataset, and the selected images, demonstrated in the first two rows, exemplify successful blends of domain characteristics, embodying a balanced mixture that enriches the training data with novel, domain-bridging examples. These images were chosen based on their ability to meet our criteria: correct class prediction aligned with high entropy scores. The third and fourth rows highlight the filtering aspect of our mechanism, displaying images not selected due to class mismatches and low entropy, respectively. This visual demonstration underlines the pivotal role of our selection/filtering process in refining the synthetic dataset, ensuring only the most challenging and domain-representative samples are utilized for model training. Through this approach, we aim to significantly bolster the model's capacity to generalize across diverse visual domains.

Inter-domain Transition. In this section, we demonstrate the model's ability to navigate between distinct visual domains, a capability enabled by adjusting the mix coefficient α. Trained on multiple source domains, our model can generate images that blend the unique attributes of each source domain. By varying α from 0.0 to 1.0, we enable smooth transitions between two source domains, where $\alpha = 0.0$ and $\alpha = 1.0$ correspond to generating pure images of the first and second domain, respectively. As an example, we illustrated this ability for our model trained on the PACS sources' *"art"*, *"sketch"*, and *"photo"*. These domain transitions are illustrated in the figures, showcasing transitions from *"photo"* to *"art"* domain in Figure 5, *"sketch"* to *"art"* domain in Figure 6, and *"sketch"* to *"photo"* domain in Figure 7, respectively. The examples provided highlight the effectiveness of our interpolation method in producing images that incorporate the distinctive features of the mixed domains, thus affirming the model's capability to generate novel and coherent visual content that bridges the attributes of its training domains. Note that in all of our generation experiments, we constrained α to the range of 0.3 to 0.7 to ensure the generated images optimally embody the characteristics of the two mixing domains, as detailed in Table 2.

Number of Generated Domains The impact of varying the number of generated domains on model performance was rigorously evaluated, as summarized in Table 3. This analysis aimed to understand how different combinations of augmented domains influence the overall accuracy across various dataset domains such as Art, Cartoon, Photo, and Sketch. By integrating diverse domain combinations, identi-

*Correspondence to mehrdad.noori.1@ens.etsmtl.ca

Config	Value
Number of GPUs	4
Learning rate	1e-4
Learning rate scheduler	LambdaLinear
Batch size	96 (24 per GPU)
Precision	FP16
Max training steps	10000
Denosing timesteps	1000
Sampler	DDPM [9]
Autoencoder input size	256 x 256 x 3
Latent diffusion input size	32 x 32 x 4

Table 1. Hyperparameter Configuration for Training Diffusion Models.

Config	Value
Sampler	DDIM [22]
Denosing timesteps	50
Classifier-free guidance (CFG)	Randomly from [5, 6]
Mix coefficient α	Randomly from [0.3, 0.7]
Mix timestep T	Randomly from [20, 45]
Generated images (PACS)	32k per class
Generated images (VLCS)	32k per class
Generated images (OfficeHome)	16k per class

Table 2. Hyperparameter Configuration for Generating New Domains.

fied by IDs (as defined in Table 4), we observed improvement gain when we add more generated domain of different combinations. Notably, all possible combinations of augmented domains (3 new domains for PACS, VLCS and OfficeHome) were utilized as the final method, leveraging the full spectrum of available data domains.

Stability Analysis. In this section, we demonstrate the performance of our model across different stages of training within two domains of the PACS dataset, depicted in Figure 1. It is important to note that these test accuracies *were not used* in the selection of the best-performing model mentioned in earlier sections and all of our experiments follow leave-one-out settings suggested by DomainBed. The results indicate that our model achieves higher stability and better mean accuracy with lower standard deviation compared to the ERM trained on original data. Note that we cannot plot the figures for SWAD since it is a WA of ERM and does not have individual training curves. These results demonstrate the robustness and stability of our model during training, which is crucial for domain generalization algorithms.

t-SNE Visualizations. This section presents a comprehensive t-SNE analysis for all classes in the PACS dataset, demonstrating the effectiveness of the FDS method in generating diverse, high-quality samples. The plots are provided in Figure 8. Each t-SNE plot illustrates the distribution of both original and FDS-generated samples across different domains. The results shown here are based on our diffusion model trained on the “Art,” “Photo,” and “Sketch” source domains from the PACS dataset. To create these visualizations, we extracted features using the CLIP vision encoder [17]. Each class in the PACS dataset is represented as distinct clusters, with “x” markers indicating the location of the average representation of each domains. These averages serve as a reference to assess how well the FDS-generated samples are compared with the original domains. These plots demonstrate how FDS enables smooth transitions between domains by interpolating between domain characteristics. This ability to generate synthetic data across a broad spectrum of domain representations improves the diversity of training data and enhances the model’s generalization ability. By covering a wider range of the domain space, FDS helps the model better handle unseen domains, making it more robust in real-world applications. These visualizations also suggest that the generated domains can be viewed as new pseudo-domains, as the FDS samples exhibit distributions distinct from their original sources domains. This additional diversity is critical for training models capable of generalizing beyond the source domains.

Visual Comparisons. This section visually compares the original images from the PACS dataset with the synthetic images generated by our FDS method, highlighting the ability of FDS to interpolate between domains. We provide examples for each pair of source domains used in training: "Art," "Photo," and "Sketch.". The visual comparisions are illustrated in Figures 9, 10, and 11. Each figure contains three sections: the first section shows samples from one original PACS domain, the middle section contains FDS-generated images combining the two selected domains, and the final section shows samples from the other original domain. These visual comparisons show that the FDS-generated images effectively blend domain-specific features, offering new pseudo-domain that can enrich the training set and enhance model generalization.

3. Oracle Results

In addition to leave-one-out setting, where the validation set is selected from the training domains, some studies also report the results of oracle (test-domain validation set). This can be particularly useful for understanding the potential of a method when domain knowledge is available. In this section, we compare our method (FDS+ERM) with the state-of-the-art results, as shown in Table 5. It is important to note that no Weight Averaging (WA) methods reported their oracle results within the DomainBed framework for a fair comparison. Therefore, we only train and report our ERM results here. Our proposed method, FDS+ERM, demonstrates superior performance across multiple benchmarks. Specifically,

Method	Augmented Domains	Accuracy (%)				
		Art	Cartoon	Photo	Sketch	Avg.
SWAD (reproduced)	—	89.49 ±0.2	83.65 ±0.4	97.25 ±0.2	82.06 ±1.0	88.11 ±0.45
SWAD + FDS	ID0	91.03 ±0.5	83.87 ±0.6	97.75 ±0.3	85.77 ±0.4	89.61 ±0.30
SWAD + FDS	ID1	91.01 ±0.6	85.06 ±1.3	97.90 ±0.3	83.64 ±0.4	89.40 ±0.65
SWAD + FDS	ID2	91.46 ±0.3	85.22 ±0.8	97.88 ±0.2	84.27 ±0.3	89.71 ±0.40
SWAD + FDS	ID0 + ID1	91.52 ±0.0	85.87 ±0.7	98.03 ±0.3	85.70 ±1.0	90.28 ±0.50
SWAD + FDS	ID1 + ID2	91.62 ±0.8	85.57 ±0.4	98.20 ±0.3	83.88 ±0.6	89.82 ±0.53
SWAD + FDS	ID0 + ID2	91.52 ±0.1	84.54 ±0.5	**98.28** ±0.1	**86.45** ±0.8	90.20 ±0.38
SWAD + FDS	ID0 + ID1 + ID2	**91.80** ±0.3	**86.03** ±0.8	98.05 ±0.2	86.11 ±0.1	**90.50** ±0.35

Table 3. Analysis of the impact of utilizing different numbers/combinations of generated domains on final model performance across the PACS dataset domains (Leave-one-out accuracy). For definitions of each augmented domain (ID0, ID1, ID2), see Table 4.

Augmented Domains	Art	Cartoon	Photo	Sketch
ID0	Cartoon + Photo	Art + Photo	Art + Cartoon	Art + Cartoon
ID1	Cartoon + Sketch	Art + Sketch	Art + Sketch	Art + Photo
ID2	Photo + Sketch	Photo + Sketch	Cartoon + Sketch	Cartoon + Photo

Table 4. Explanation of augmented domains ID definitions for each target domain of PACS dataset.

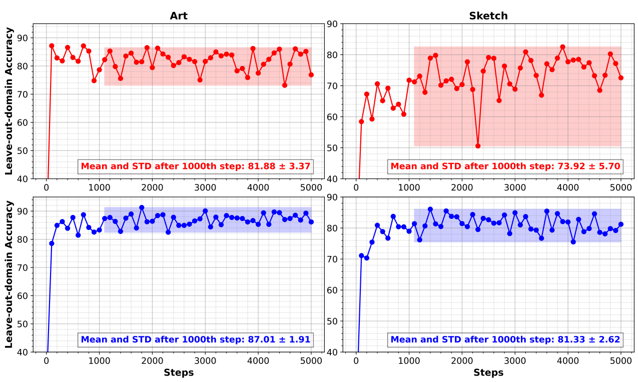

Figure 1. Accuracy (%) across training steps: Comparison between ERM (top row) vs. FDS (bottom row) in "Art" and "Sketch" domains of PACS dataset.

it achieves an average accuracy of 81.2%, outperforming all other methods. On the PACS dataset, FDS+ERM attains the highest accuracy of 89.7%, with significant improvements in the VLCS and OfficeHome datasets as well, achieving accuracies of 82.0% and 71.8% respectively. In addition to leave-one-out setting, these results also highlight the effectiveness of our approach in enhancing the performance under the oracle setting.

4. Detailed Results

Here we present the comprehensive tables containing all the detailed information that was summarized in the main paper. The leave-one-out performance (train-domain validation set) across different domains of PACS, VLCS, and OfficeHome datasets are detailed in the tables 6, 7, and 8, respectively. Additionally, the oracle (test-domain validation set) accuracy results for the PACS, VLCS, and OfficeHome benchmarks are detailed in Table 9, 10, and 11, respectively.

	Method	Aug.	PACS	VLCS	OfficeHome	Avg.
Standard Methods	ERM (*baseline*) [7]	✗	86.7 ±0.3	77.6 ±0.3	66.4 ±0.5	76.9
	ERM (*reproduced*)	✗	86.6 ±0.8	79.8 ±0.4	68.4 ±0.3	78.3
	IRM [2]	✗	84.5 ±1.1	76.9 ±0.6	63.0 ±2.7	74.8
	GroupDRO [21]	✗	87.1 ±0.1	77.4 ±0.5	66.2 ±0.6	76.9
	Mixup [24]	✓	86.8 ±0.3	78.1 ±0.3	68.0 ±0.2	77.6
	CORAL [23]	✗	87.1 ±0.5	77.7 ±0.2	68.4 ±0.2	77.7
	MMD [14]	✗	87.2 ±0.1	77.9 ±0.1	66.2 ±0.3	77.1
	DANN [6]	✗	85.2 ±0.2	79.7 ±0.5	65.3 ±0.8	76.7
	SagNet [16]	✓	86.4 ±0.4	77.6 ±0.1	67.5 ±0.2	77.2
	RSC [10]	✓	86.2 ±0.5	—	66.5 ±0.6	—
	SelfReg [11]	✓	86.7 ±0.8	78.2 ±0.1	68.1 ±0.3	77.7
	Fishr [18]	✗	85.8 ±0.6	78.2 ±0.2	66.0 ±2.9	76.7
	CDGA [8]	✓	<u>89.6 ±0.3</u>	<u>80.9 ±0.1</u>	<u>68.8 ±0.3</u>	<u>79.3</u>
	ERM + FDS (ours)	✓	**89.7** ±0.8	**82.0** ±0.1	**71.8** ±0.9	**81.2**

Table 5. Oracle (test-domain validation set) accuracy (%) results on the PACS, VLCS, and OfficeHome benchmarks. "**Aug.**" indicates whether advanced augmentation or domain mixing techniques are used. The **best results** and <u>second-best results</u> are highlighted.

	Method	Aug.	Target Domains				
			Art	Cartoon	Photo	Sketch	Avg.
Standard Methods	ERM (baseline) [7]	✗	84.7 ±0.4	80.8 ±0.6	97.2 ±0.3	79.3 ±1.0	85.5 ±0.2
	ERM (reproduced)	✗	86.9 ±0.6	80.2 ±0.7	96.6 ±0.4	74.5 ±2.9	84.3 ±1.1
	IRM [2]	✗	84.8 ±1.3	76.4 ±1.1	96.7 ±0.6	76.1 ±1.0	83.5 ±0.8
	GroupDRO [21]	✗	83.5 ±0.9	79.1 ±0.6	96.7 ±0.3	78.3 ±2.0	84.4 ±0.8
	Mixup [24]	✓	86.1 ±0.5	78.9 ±0.8	**97.6 ±0.1**	75.8 ±1.8	84.6 ±0.6
	CORAL [23]	✗	88.3 ±0.2	80.0 ±0.5	97.5 ±0.3	78.8 ±1.3	86.2 ±0.3
	MMD [14]	✗	86.1 ±1.4	79.4 ±0.9	96.6 ±0.2	76.5 ±0.5	84.6 ±0.5
	DANN [6]	✗	86.4 ±0.8	77.4 ±0.8	97.3 ±0.4	73.5 ±2.3	83.6 ±0.4
	MLDG [13]	✗	85.5 ±1.4	80.1 ±1.7	97.4 ±0.3	76.6 ±1.1	84.9 ±1.1
	VREx [12]	✗	86.0 ±1.6	79.1 ±0.6	96.9 ±0.5	77.7 ±1.7	84.9 ±1.1
	ARM [26]	✗	86.8 ±0.6	76.8 ±0.5	97.4 ±0.3	79.3 ±1.2	85.1 ±0.6
	SagNet [16]	✓	87.4 ±1.0	80.7 ±0.6	97.1 ±0.1	80.0 ±0.4	86.3 ±0.2
	RSC [10]	✓	85.4 ±0.8	79.7 ±1.8	<u>97.6 ±0.3</u>	78.2 ±1.2	85.2 ±0.9
	Mixstyle [27]	✓	86.8 ±0.5	79.0 ±1.4	96.6 ±0.1	78.5 ±2.3	85.2 ±0.3
	mDSDI [3]	✗	87.7 ±0.4	80.4 ±0.7	98.1 ±0.3	78.4 ±1.2	86.2 ±0.2
	SelfReg [11]	✓	87.9 ±1.0	79.4 ±1.4	96.8 ±0.7	78.3 ±1.2	85.6 ±0.4
	Fishr [18]	✗	88.4 ±0.2	78.7 ±0.7	97.0 ±0.1	77.8 ±2.0	85.5 ±0.5
	DCAug [1]	✓	88.5 ±0.8	78.8 ±1.5	96.3 ±0.1	80.8 ±0.5	86.1 ±0.7
	DomainDiff [15]	✓	84.9 ±1.6	<u>82.9 ±0.0</u>	95.5 ±0.0	79.0 ±0.9	85.6 ±0.6
	DSI [25]	✓	84.6 ±2.4	81.4 ±1.6	96.8 ±0.5	82.5 ±1.0	86.9 ±1.4
	CDGA [8]	✓	<u>89.1 ±1.0</u>	82.5 ±0.5	97.4 ±0.2	**84.8 ±0.9**	<u>88.5 ±0.5</u>
	ERM + FDS (ours)	✓	**90.7 ±0.9**	**84.2 ±0.6**	97.2 ±0.1	<u>83.0 ±0.4</u>	**88.8 ±0.1**
WA Methods	SWAD (baseline) [4]	✗	89.3 ±0.2	83.4 ±0.6	97.3 ±0.3	82.5 ±0.5	88.1 ±0.1
	SWAD (reproduced)	✗	89.5 ±0.2	<u>83.7 ±0.4</u>	97.3 ±0.2	82.1 ±0.1	88.1 ±0.4
	SelfReg SWA [11]	✓	85.9 ±0.6	81.9 ±0.4	96.8 ±0.1	81.4 ±0.6	86.5 ±0.3
	DNA [5]	✗	89.8 ±0.2	83.4 ±0.4	97.7 ±0.1	82.6 ±0.2	88.4 ±0.1
	DiWA [19]	✓	<u>90.1 ±0.6</u>	83.3 ±0.6	**98.2 ±0.1**	83.4 ±0.4	<u>88.8 ±0.4</u>
	TeachDCAug [1]	✓	89.6 ±0.0	81.8 ±0.5	97.7 ±0.0	<u>84.5 ±0.2</u>	88.4 ±0.2
	SWAD + FDS (ours)	✓	**91.8 ±0.3**	**86.0 ±0.8**	<u>98.1 ±0.2</u>	**86.1 ±0.1**	**90.5 ±0.3**

Table 6. Leave-one-out accuracy (%) results on the PACS dataset. "**Aug.**" indicates whether advanced augmentation or domain mixing techniques are used. The **best results** and <u>second-best results</u> are highlighted.

	Method	Aug.	Target Domains				
			Caltech101	LabelMe	SUN09	VOC2007	Avg.
Standard Methods	ERM (baseline) [7]	✗	97.7 ±0.4	64.3 ±0.9	73.4 ±0.5	74.6 ±1.3	77.5 ±0.4
	ERM (reproduced)	✗	96.9 ±1.4	64.1 ±1.4	71.1 ±1.5	72.8 ±0.9	76.2 ±1.1
	IRM [2]	✗	98.6 ±0.1	64.9 ±0.9	73.4 ±0.6	77.3 ±0.9	78.5 ±0.5
	GroupDRO [21]	✗	97.3 ±0.3	63.4 ±0.9	69.5 ±0.8	76.7 ±0.7	76.7 ±0.6
	Mixup [24]	✓	98.3 ±0.6	64.8 ±1.0	72.1 ±0.5	74.3 ±0.8	77.4 ±0.6
	CORAL [23]	✗	98.3 ±0.1	66.1 ±1.2	73.4 ±0.3	77.5 ±1.2	78.8 ±0.6
	MMD [14]	✗	97.7 ±0.1	64.0 ±1.1	72.8 ±0.2	75.3 ±3.3	77.5 ±0.9
	DANN [6]	✗	**99.0 ±0.3**	65.1 ±1.4	73.1 ±0.3	77.2 ±0.6	78.6 ±0.4
	MLDG [13]	✗	97.4 ±0.2	65.2 ±0.7	71.0 ±1.4	75.3 ±1.0	77.2 ±0.8
	VREx [12]	✗	98.4 ±0.3	64.4 ±1.4	74.1 ±0.4	76.2 ±1.3	78.3 ±0.8
	ARM [26]	✗	98.7 ±0.2	63.6 ±0.7	71.3 ±1.2	76.7 ±0.6	77.6 ±0.6
	SagNet [16]	✓	97.9 ±0.4	64.5 ±0.5	71.4 ±1.3	77.5 ±0.5	77.8 ±0.5
	RSC [10]	✓	97.9 ±0.1	62.5 ±0.7	72.3 ±1.2	75.6 ±0.8	77.1 ±0.5
	Mixstyle [27]	✓	98.6 ±0.3	64.5 ±1.1	72.6 ±0.5	75.7 ±1.7	77.9 ±0.5
	mDSDI [3]	✗	97.6 ±0.1	<u>66.4 ±0.4</u>	74.0 ±0.6	<u>77.8 ±0.7</u>	79.0 ±0.3
	SelfReg [11]	✓	96.7 ±0.4	65.2 ±1.2	73.1 ±1.3	76.2 ±0.7	77.8 ±0.9
	Fishr [18]	✗	<u>98.9 ±0.3</u>	64.0 ±0.5	71.5 ±0.2	76.8 ±0.7	77.8 ±0.5
	DCAug [1]	✓	98.3 ±0.1	64.2 ±0.4	<u>74.4 ±0.6</u>	77.5 ±0.3	78.6 ±0.5
	CDGA [8]	✓	96.3 ±0.7	**75.7 ±1.0**	72.8 ±1.3	73.7 ±1.3	<u>79.6 ±0.9</u>
	ERM + FDS (ours)	✓	98.8 ±0.3	65.6 ±0.9	**75.5 ±0.9**	**79.3 ±1.8**	**79.8 ±0.5**
WA Methods	SWAD (baseline) [4]	✗	<u>98.8 ±0.1</u>	63.3 ±0.3	75.3 ±0.5	79.2 ±0.6	<u>79.1 ±0.1</u>
	SWAD (reproduced)	✗	98.7 ±0.2	**63.9 ±0.3**	74.3 ±1.1	78.6 ±0.6	78.9 ±0.5
	SelfReg SWA [11]	✓	97.4 ±0.4	63.5 ±0.3	72.6 ±0.1	76.7 ±0.7	77.5 ±0.0
	DNA [5]	✗	<u>98.8 ±0.1</u>	63.6 ±0.2	74.1 ±0.1	<u>79.5 ±0.4</u>	79.0 ±0.5
	DiWA [19]	✓	98.4 ±0.1	63.4 ±0.1	75.5 ±0.3	78.9 ±0.6	79.1 ±0.2
	TeachDCAug [1]	✓	98.5 ±0.1	<u>63.7 ±0.3</u>	<u>75.6 ±0.5</u>	77.0 ±0.7	78.7 ±0.5
	SWAD + FDS (ours)	✓	**99.5 ±0.2**	62.9 ±0.2	**76.9 ±0.4**	**79.6 ±1.3**	**79.7 ±0.5**

Table 7. Leave-one-out accuracy (%) results on the VLCS dataset. "**Aug.**" indicates whether advanced augmentation or domain mixing techniques are used. The **best results** and <u>second-best results</u> are highlighted.

	Method	Aug.	Art	Clipart	Product	Real World	Avg.
			Target Domains				
Standard Methods	ERM (baseline) [7]	✗	61.3 ±0.7	52.4 ±0.3	75.8 ±0.1	76.6 ±0.3	66.5 ±0.3
	ERM (reproduced)	✗	59.5 ±2.1	51.3 ±1.3	73.8 ±0.8	73.8 ±0.2	64.6 ±1.1
	IRM [2]	✗	58.9 ±2.3	52.2 ±1.6	72.1 ±2.9	74.0 ±2.5	64.3 ±2.2
	GroupDRO [21]	✗	60.4 ±0.7	52.7 ±1.0	75.0 ±0.7	76.0 ±0.7	66.0 ±0.7
	Mixup [24]	✓	62.4 ±0.8	54.8 ±0.6	76.9 ±0.3	78.3 ±0.2	68.1 ±0.3
	CORAL [23]	✗	<u>65.3 ±0.4</u>	54.4 ±0.5	76.5 ±0.1	78.4 ±0.5	68.7 ±0.3
	MMD [14]	✗	60.4 ±0.2	53.3 ±0.3	74.3 ±0.1	77.4 ±0.6	66.3 ±0.1
	DANN [6]	✗	59.9 ±1.3	53.0 ±0.3	73.6 ±0.7	76.9 ±0.5	65.9 ±0.6
	MLDG [13]	✗	61.5 ±0.9	53.2 ±0.6	75.0 ±1.2	77.5 ±0.4	66.8 ±0.7
	VREx [12]	✗	60.7 ±0.9	53.0 ±0.9	75.3 ±0.1	76.6 ±0.5	66.4 ±0.6
	ARM [26]	✗	58.9 ±0.8	51.0 ±0.5	74.1 ±0.1	75.2 ±0.3	64.8 ±0.4
	SagNet [16]	✓	63.4 ±0.2	54.8 ±0.4	75.8 ±0.4	78.3 ±0.3	68.1 ±0.1
	RSC [10]	✓	60.7 ±1.4	51.4 ±0.3	74.8 ±1.1	75.1 ±1.3	65.5 ±0.9
	Mixstyle [27]	✓	51.1 ±0.3	53.2 ±0.4	68.2 ±0.7	69.2 ±0.6	60.4 ±0.3
	mDSDI [3]	✗	**68.1 ±0.3**	52.1 ±0.4	76.0 ±0.2	80.4 ±0.2	<u>69.2 ±0.4</u>
	SelfReg [11]	✓	63.6 ±1.4	53.1 ±1.0	76.9 ±0.4	78.1 ±0.4	67.9 ±0.7
	Fishr [18]	✗	62.4 ±0.5	54.4 ±0.4	76.2 ±0.5	78.3 ±0.1	67.8 ±0.5
	DCAug [1]	✓	61.8 ±0.6	<u>55.4 ±0.6</u>	77.1 ±0.3	78.9 ±0.3	68.3 ±0.4
	DomainDiff [15]	✓	57.6 ±0.4	49.2 ±0.6	73.0 ±0.6	75.2 ±0.9	63.7 ±0.6
	CDGA [8]	✓	60.1 ±1.4	54.2 ±0.5	<u>78.2 ±0.6</u>	<u>80.4 ±0.1</u>	68.2 ±0.6
	ERM + FDS (ours)	✓	64.6 ±0.2	**57.7 ±0.1**	**80.2 ±0.5**	**82.0 ±0.4**	**71.1 ±0.1**
WA Methods	SWAD (baseline) [4]	✗	66.1 ±0.4	57.7 ±0.4	78.4 ±0.1	80.2 ±0.2	70.6 ±0.2
	SWAD (reproduced)	✗	65.9 ±0.9	56.8 ±0.4	78.8 ±0.3	80.0 ±0.2	70.3 ±0.4
	SelfReg SWA [11]	✓	64.9 ±0.8	55.4 ±0.6	78.4 ±0.2	78.8 ±0.1	69.4 ±0.2
	DNA [5]	✗	**67.7 ±0.2**	57.7 ±0.3	78.9 ±0.2	<u>80.5 ±0.2</u>	<u>71.2 ±0.1</u>
	DiWA [19]	✓	<u>67.3 ±0.2</u>	<u>57.9 ±0.2</u>	<u>79.0 ±0.2</u>	79.9 ±0.1	71.0 ±0.1
	TeachDCAug [1]	✓	66.2 ±0.2	57.0 ±0.3	78.3 ±0.1	80.1 ±0.0	70.4 ±0.2
	SWAD + FDS (ours)	✓	67.3 ±0.8	**60.5 ±0.5**	**82.6 ±0.1**	**83.6 ±0.3**	**73.5 ±0.4**

Table 8. Leave-one-out accuracy (%) results on the OfficeHome dataset. "**Aug.**" indicates whether advanced augmentation or domain mixing techniques are used. The **best results** and <u>second-best results</u> are highlighted.

	Method	Aug.	Art	Cartoon	Photo	Sketch	Avg.
			Target Domains				
Standard Methods	ERM (baseline) [7]	✗	86.5 ±1.0	81.3 ±0.6	96.2 ±0.3	82.7 ±1.1	86.7 ±0.8
	ERM (reproduced)	✗	88.6 ±0.9	80.9 ±1.9	**98.4 ±0.4**	78.4 ±1.2	86.6 ±1.0
	IRM	✗	84.2 ±0.9	79.7 ±1.5	95.9 ±0.4	78.3 ±2.1	84.5 ±1.2
	GroupDRO	✗	87.5 ±0.5	82.9 ±0.6	97.1 ±0.3	81.1 ±1.2	87.2 ±0.7
	Mixup	✓	87.5 ±0.4	81.6 ±0.7	97.4 ±0.2	80.8 ±0.9	86.8 ±0.6
	CORAL	✗	86.6 ±0.8	81.8 ±0.9	97.1 ±0.5	82.7 ±0.6	87.1 ±0.7
	MMD	✗	88.1 ±0.8	82.6 ±0.7	97.1 ±0.5	81.2 ±1.2	87.3 ±0.8
	DANN	✗	87.0 ±0.4	80.3 ±0.6	96.8 ±0.3	76.9 ±1.1	85.3 ±0.6
	SagNet	✓	87.4 ±0.5	81.2 ±1.2	96.3 ±0.8	80.7 ±1.1	86.4 ±0.9
	RSC	✓	86.0 ±0.7	81.8 ±0.9	96.8 ±0.7	80.4 ±0.5	86.3 ±0.7
	Fishr	✗	87.9 ±0.6	80.8 ±0.5	97.9 ±0.4	81.1 ±0.8	86.9 ±0.6
	SelfReg	✓	87.9 ±0.5	80.6 ±1.1	97.1 ±0.4	81.1 ±1.3	86.7 ±0.8
	CDGA	✓	89.6 ±0.8	**85.3 ±0.7**	97.3 ±0.3	**86.2 ±0.5**	89.6 ±0.6
	ERM + FDS (ours)	✓	**91.1 ±0.3**	84.9 ±0.7	97.3 ±0.5	85.6 ±2.3	**89.7 ±0.8**

Table 9. Oracle (test-domain validation set) accuracy (%) results on the PACS dataset. "**Aug.**" indicates whether advanced augmentation or domain mixing techniques are used. The **best results** and second-best results are highlighted.

	Method	Aug.	Caltech101	LabelMe	SUN09	VOC2007	Avg.
			Target Domains				
Standard Methods	ERM (baseline) [7]	✗	97.6 ±0.3	67.9 ±0.7	70.9 ±0.2	74.0 ±0.6	77.6 ±0.5
	ERM (reproduced)	✗	98.6 ±0.2	68.6 ±0.7	73.6 ±1.7	78.6 ±1.2	79.8 ±0.4
	IRM	✗	97.3 ±0.2	66.7 ±0.1	71.0 ±2.3	72.8 ±0.4	77.0 ±0.8
	GroupDRO	✗	97.7 ±0.2	65.9 ±0.2	72.8 ±0.8	73.4 ±1.3	77.5 ±0.6
	Mixup	✓	97.8 ±0.4	67.2 ±0.4	71.5 ±0.2	75.7 ±0.6	78.1 ±0.4
	CORAL	✗	97.3 ±0.2	67.5 ±0.6	71.6 ±0.6	74.5 ±0.0	77.7 ±0.4
	MMD	✗	98.8 ±0.0	66.4 ±0.4	70.8 ±0.5	75.6 ±0.4	77.9 ±0.3
	DANN	✗	9.0 ±0.2	66.3 ±1.2	73.4 ±1.4	80.1 ±0.5	79.7 ±0.8
	SagNet	✓	97.4 ±0.3	66.4 ±0.4	71.6 ±0.1	75.0 ±0.8	77.6 ±0.4
	RSC	✓	98.0 ±0.4	67.2 ±0.3	70.3 ±1.3	75.6 ±0.4	77.8 ±0.6
	Fishr	✗	97.6 ±0.7	67.3 ±0.5	72.2 ±0.9	75.7 ±0.3	78.2 ±0.6
	SelfReg	✓	98.2 ±0.3	63.9 ±0.8	72.2 ±0.1	75.5 ±0.4	77.5 ±0.2
	CDGA	✓	96.6 ±0.7	**75.5 ±1.9**	73.6 ±1.1	77.8 ±1.0	80.9 ±1.2
	ERM + FDS (ours)	✓	**99.5 ±0.1**	68.7 ±0.3	**77.4 ±0.7**	**82.6 ±0.1**	**82.0 ±0.1**

Table 10. Oracle (test-domain validation set) accuracy (%) results on the VLCS dataset. "**Aug.**" indicates whether advanced augmentation or domain mixing techniques are used. The **best results** and second-best results are highlighted.

	Method	Aug.	Target Domains				
			Art	**Clipart**	**Product**	**Real World**	**Avg.**
Standard Methods	ERM (baseline) [7]	✗	61.7 ±0.7	53.4 ±0.3	74.1 ±0.4	76.2 ±0.6	66.4 ±0.5
	ERM (reproduced)	✗	64.0 ±0.9	53.7 ±1.1	77.1 ±0.3	78.8 ±0.4	68.4 ±0.3
	IRM	✗	56.4 ±3.2	51.2 ±2.3	71.7 ±2.7	72.7 ±2.7	63.0 ±2.7
	GroupDRO	✗	60.5 ±1.6	53.1 ±0.3	75.5 ±0.3	75.9 ±0.7	66.3 ±0.7
	Mixup	✓	63.5 ±0.2	54.6 ±0.4	76.0 ±0.3	78.0 ±0.7	68.0 ±0.4
	CORAL	✗	<u>64.8 ±0.8</u>	54.1 ±0.9	76.5 ±0.4	78.2 ±0.4	68.4 ±0.6
	MMD	✗	60.4 ±1.0	53.4 ±0.5	74.9 ±0.1	76.1 ±0.7	66.2 ±0.6
	DANN	✗	60.6 ±1.4	51.8 ±0.7	73.4 ±0.5	75.5 ±0.9	65.3 ±0.9
	SagNet	✓	62.7 ±0.5	53.6 ±0.5	76.0 ±0.3	77.8 ±0.1	67.5 ±0.4
	RSC	✓	61.7 ±0.8	53.0 ±0.9	74.8 ±0.8	76.3 ±0.5	66.5 ±0.8
	Fishr	✗	63.4 ±0.8	54.2 ±0.3	76.4 ±0.3	78.5 ±0.2	68.1 ±0.4
	SelfReg	✓	64.2 ±0.6	53.6 ±0.7	76.7 ±0.3	77.9 ±0.5	68.1 ±0.3
	CDGA	✓	61.1 ±1.1	<u>55.9 ±1.0</u>	<u>78.2 ±0.8</u>	<u>79.8 ±0.2</u>	<u>68.8 ±0.8</u>
	ERM + FDS (ours)	✓	**65.3** ±**0.8**	**58.4** ±**0.8**	**81.2** ±**0.2**	**82.4** ±**0.6**	**71.8** ±**0.9**

Table 11. Oracle (test-domain validation set) accuracy (%) results on the OfficeHome dataset. "**Aug.**" indicates whether advanced augmentation or domain mixing techniques are used. The **best results** and <u>second-best results</u> are highlighted.

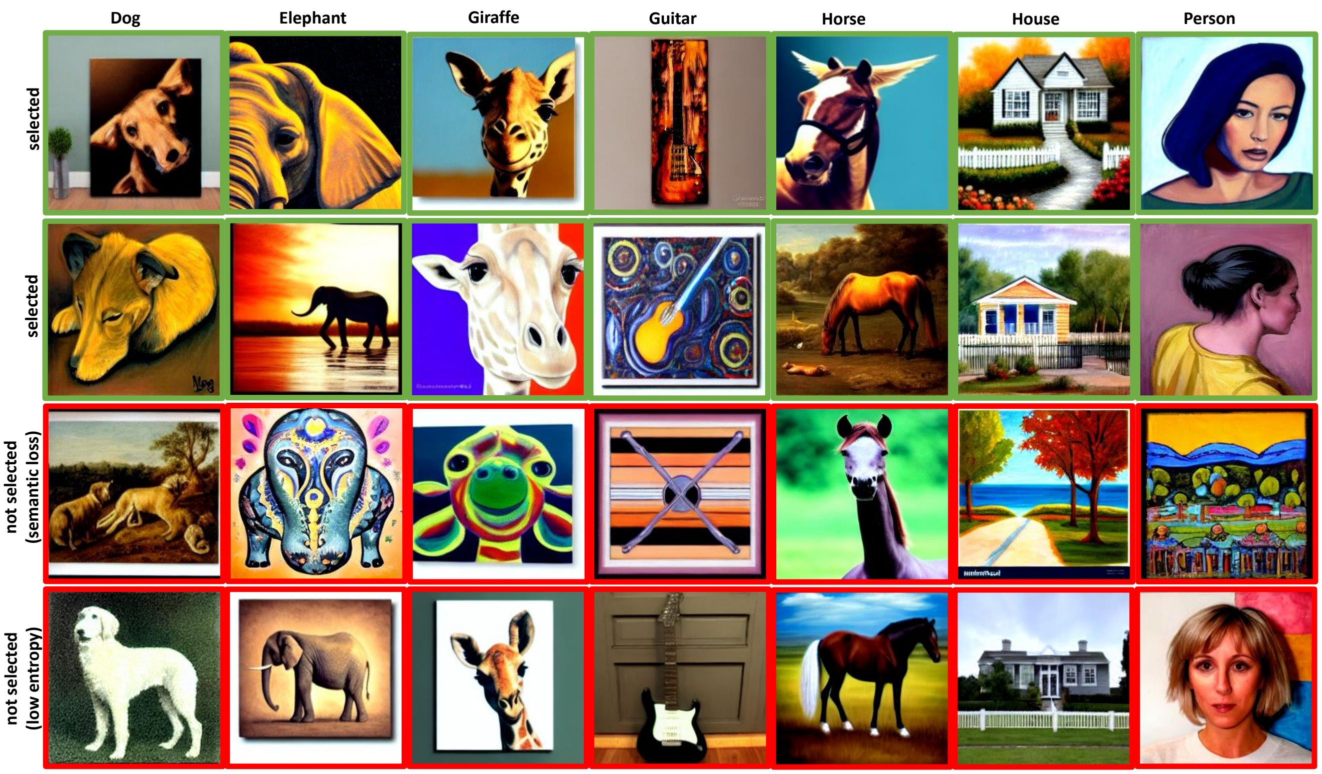

Figure 2. Synthetic images from interpolating between *"art"* and *"photo"* domains of PACS, with selected images showcasing a blend of artistic and realistic features (top two rows) and non-selected images (bottom rows) due to class mismatches and low entropy.

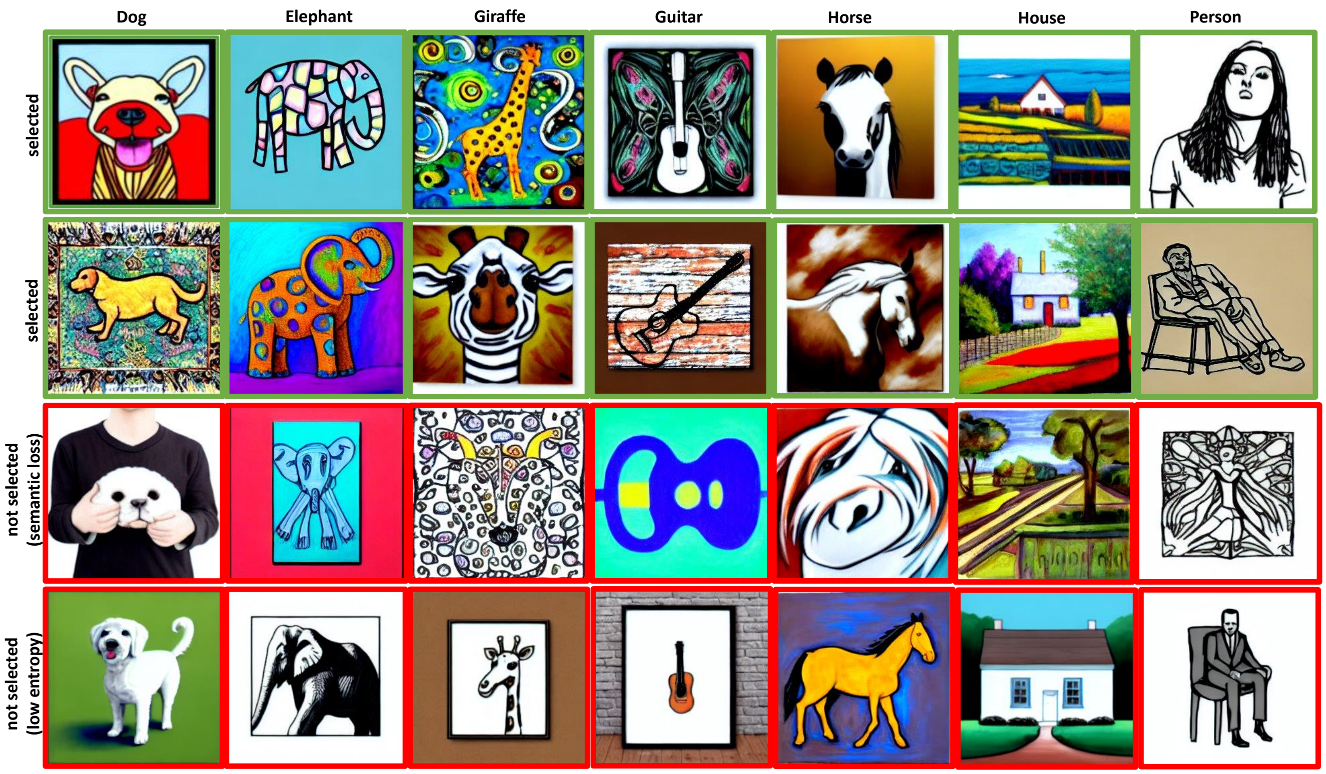

Figure 3. Interpolation between *"art"* and *"sketch"* in PACS highlights selected images (top rows) merging textures and outlines, and non-selected images (bottom rows) for failing selection criteria.

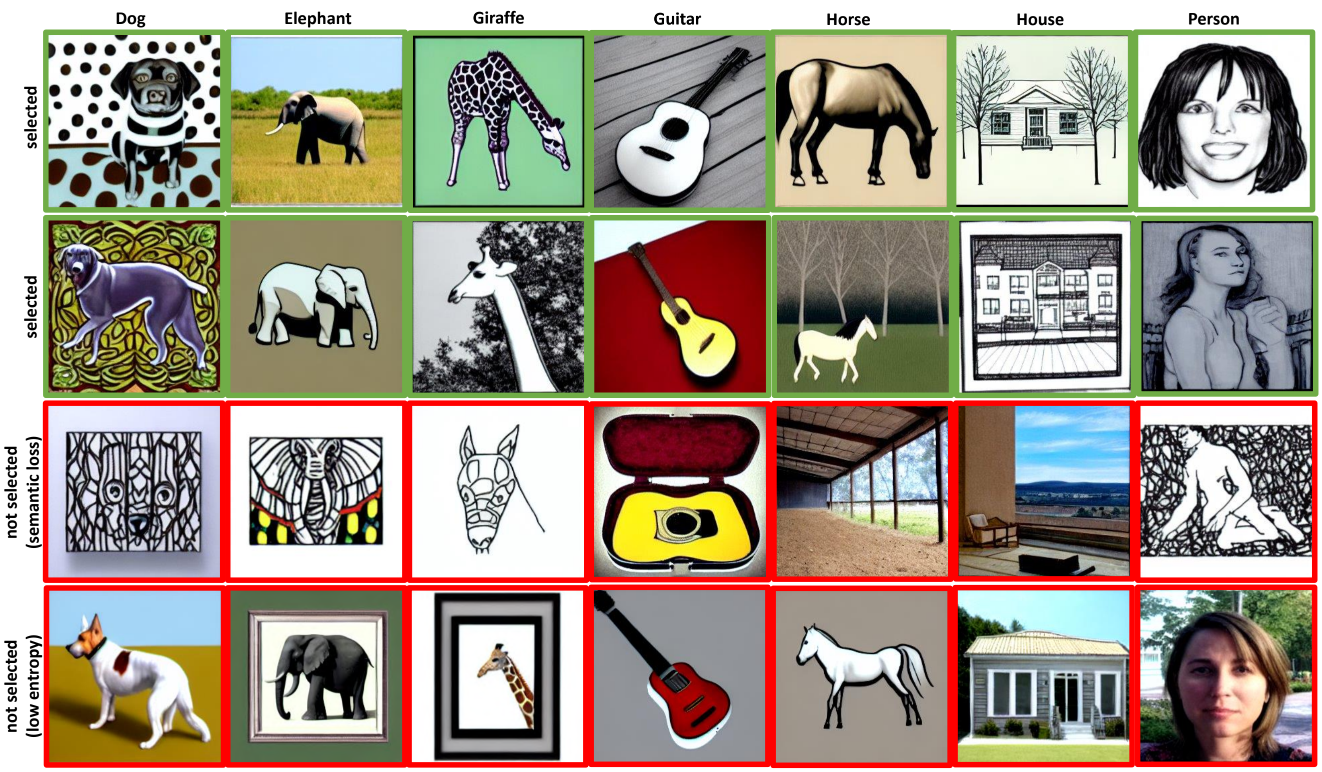

Figure 4. Results from *"photo"* and *"sketch"* domain interpolation in PACS, with selected synthetic images (top rows) and non-selected due to predictability and class misalignment (bottom rows).

Figure 5. Inter-domain Transition from *"photo"* to *"art"*. This sequence illustrates how varying α from 0.0 (purely photorealistic images) to 1.0 (purely artistic representations) enables the model to seamlessly blend photographic realism with artistic expression, demonstrating a smooth progression from real-world imagery to stylized art.

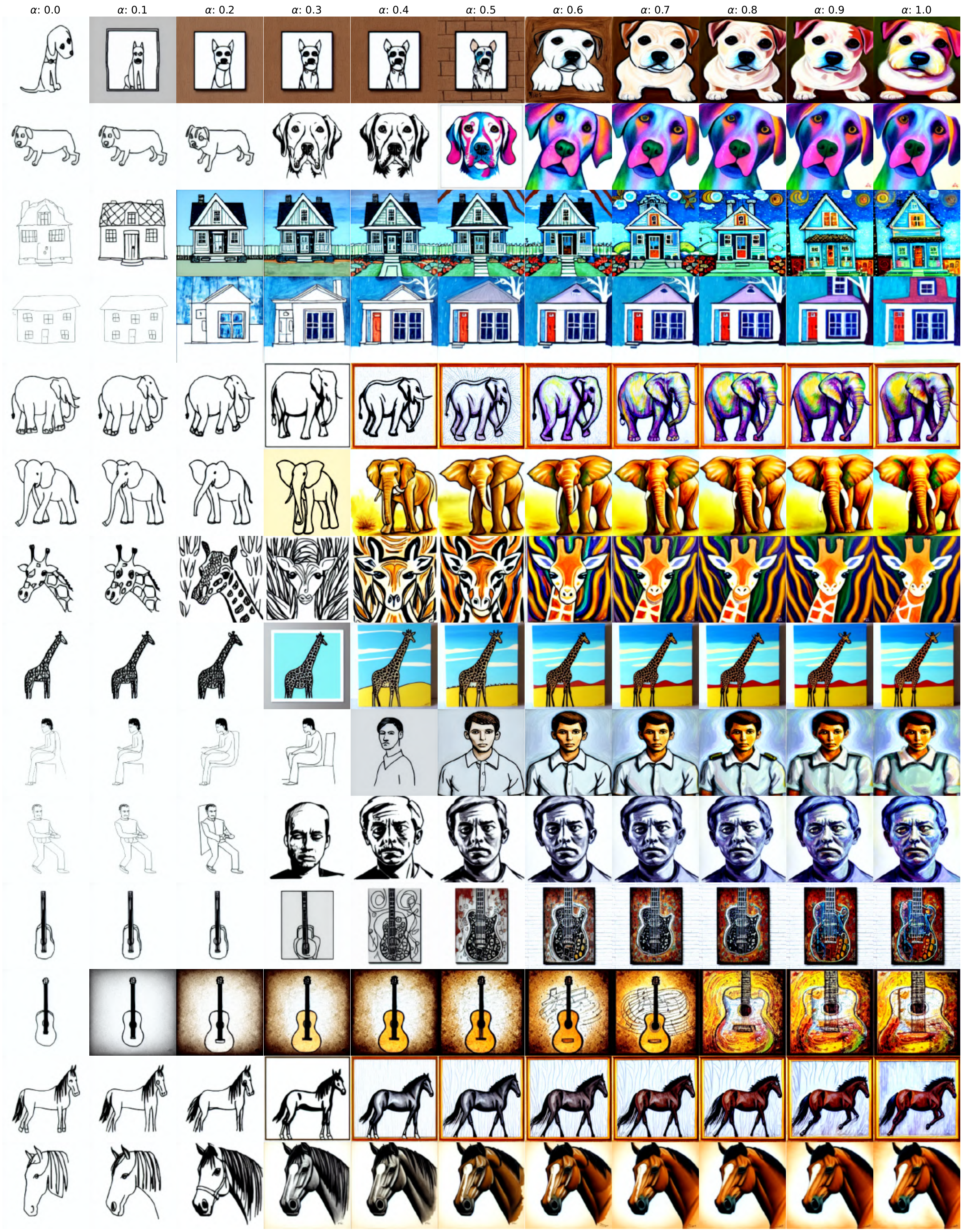

Figure 6. Inter-domain Transition from *"sketch"* to *"art"*. Displayed here is the transformation that occurs as α is adjusted, beginning with 0.0 (pure sketches) and moving towards 1.0 (fully art-inspired images). The model effectively infuses basic sketches with complex textures and colors, transitioning from minimalistic line art to detailed and vibrant artistic images.

Figure 7. Inter-domain Transition from *"sketch"* to *"photo"*. This figure demonstrates the capability of the model to morph sketches into photorealistic images by altering α from 0.0 (entirely sketch-based) to 1.0 (completely photorealistic). The transition highlights the model's proficiency in enriching simple outlines with lifelike details and textures, bridging the gap between abstract sketches and reality.

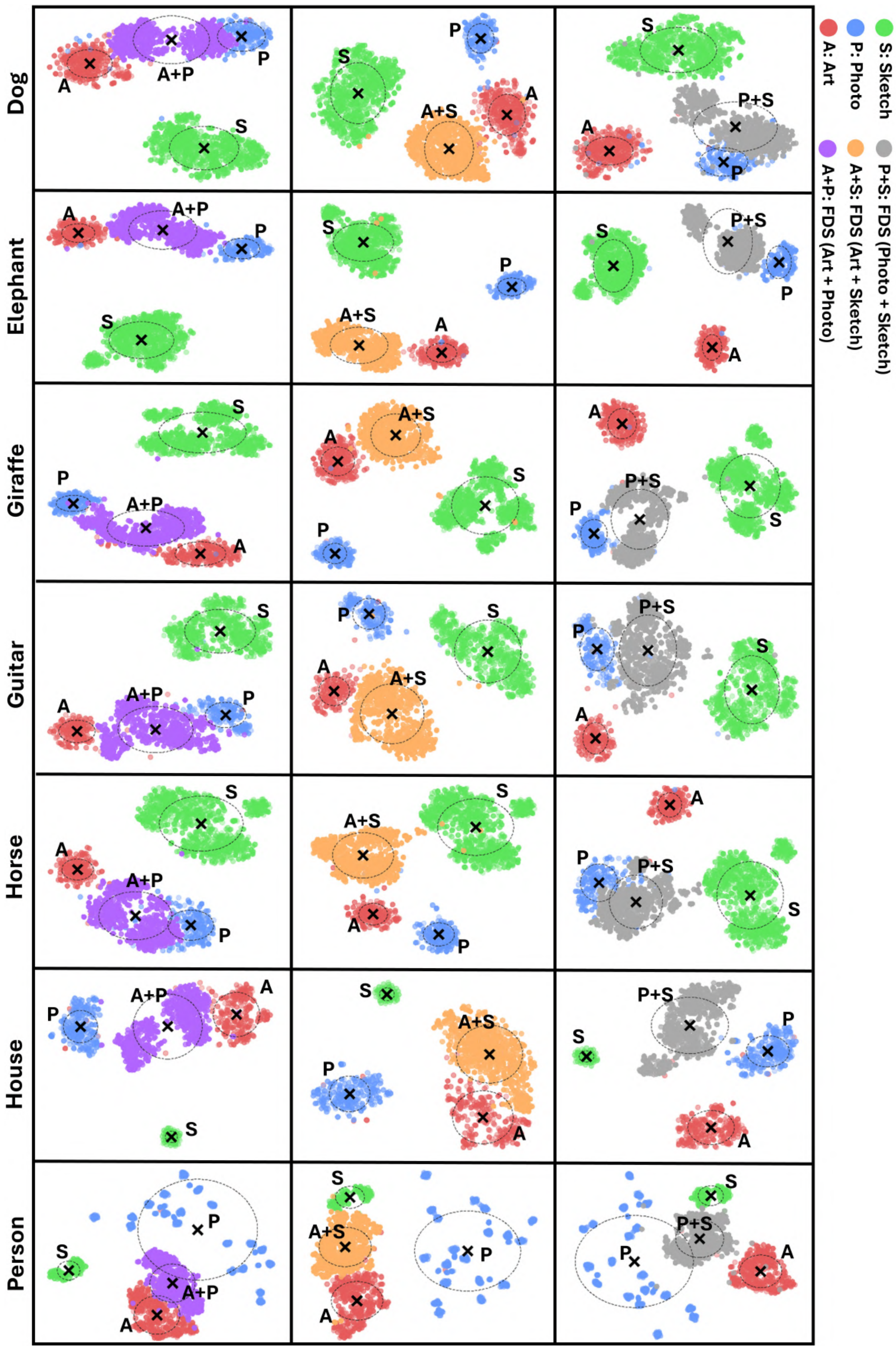

Figure 8. t-SNE visualization of all classes from the PACS dataset, showing the distribution of original source domains (Art, Photo, Sketch) and FDS ones.

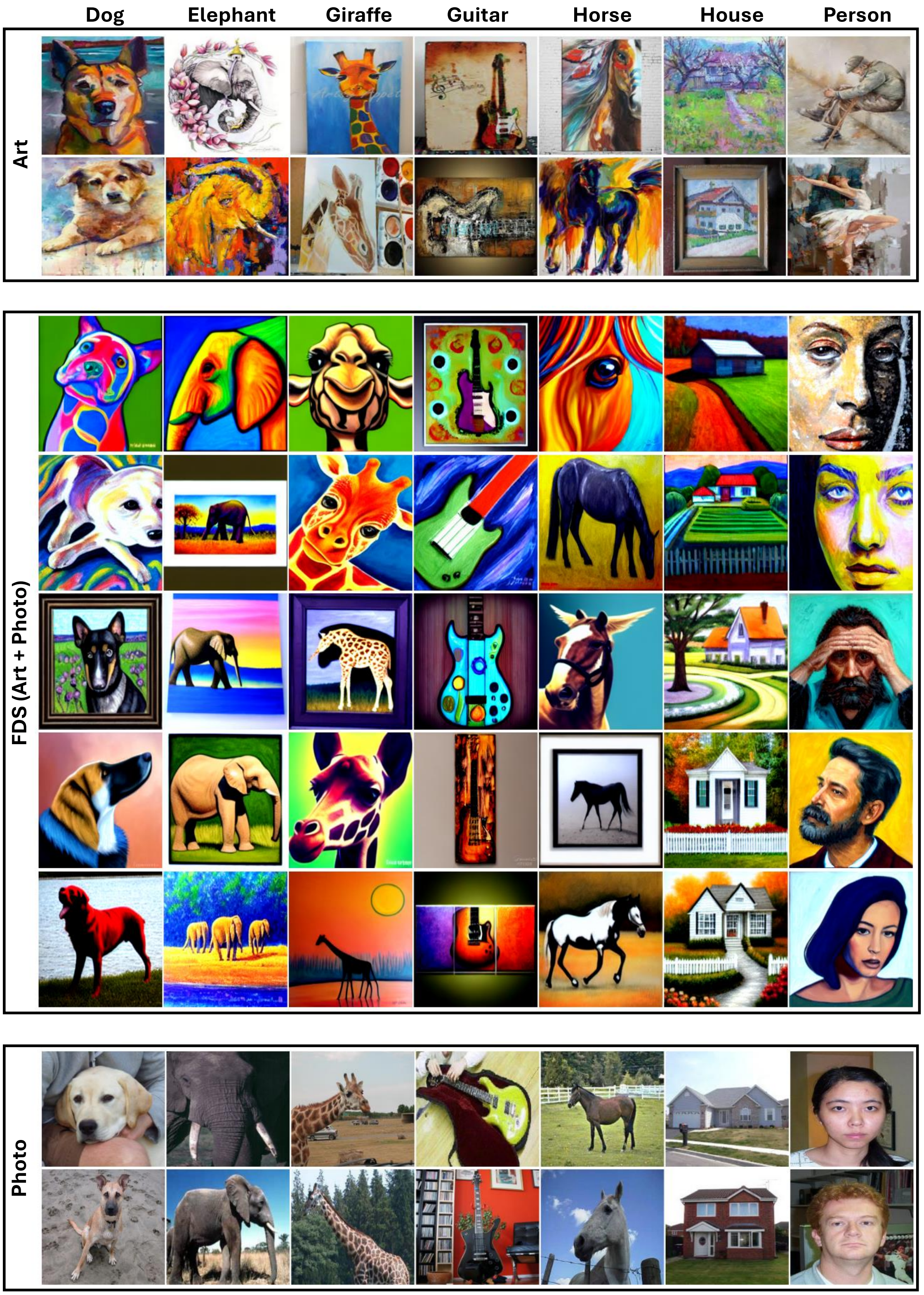

Figure 9. Visual comparison of original "Art" and "Photo" samples from PACS with synthetic images generated by FDS (Art + Photo). The middle section illustrates how FDS combines visual elements from both domains, producing diverse, domain-bridging images.

Figure 10. Visual comparison of original "Sketch" and "Art" samples from PACS with synthetic images generated by FDS (Sketch + Art). The generated images in the middle section showcase a blend of artistic textures and sketched outlines.

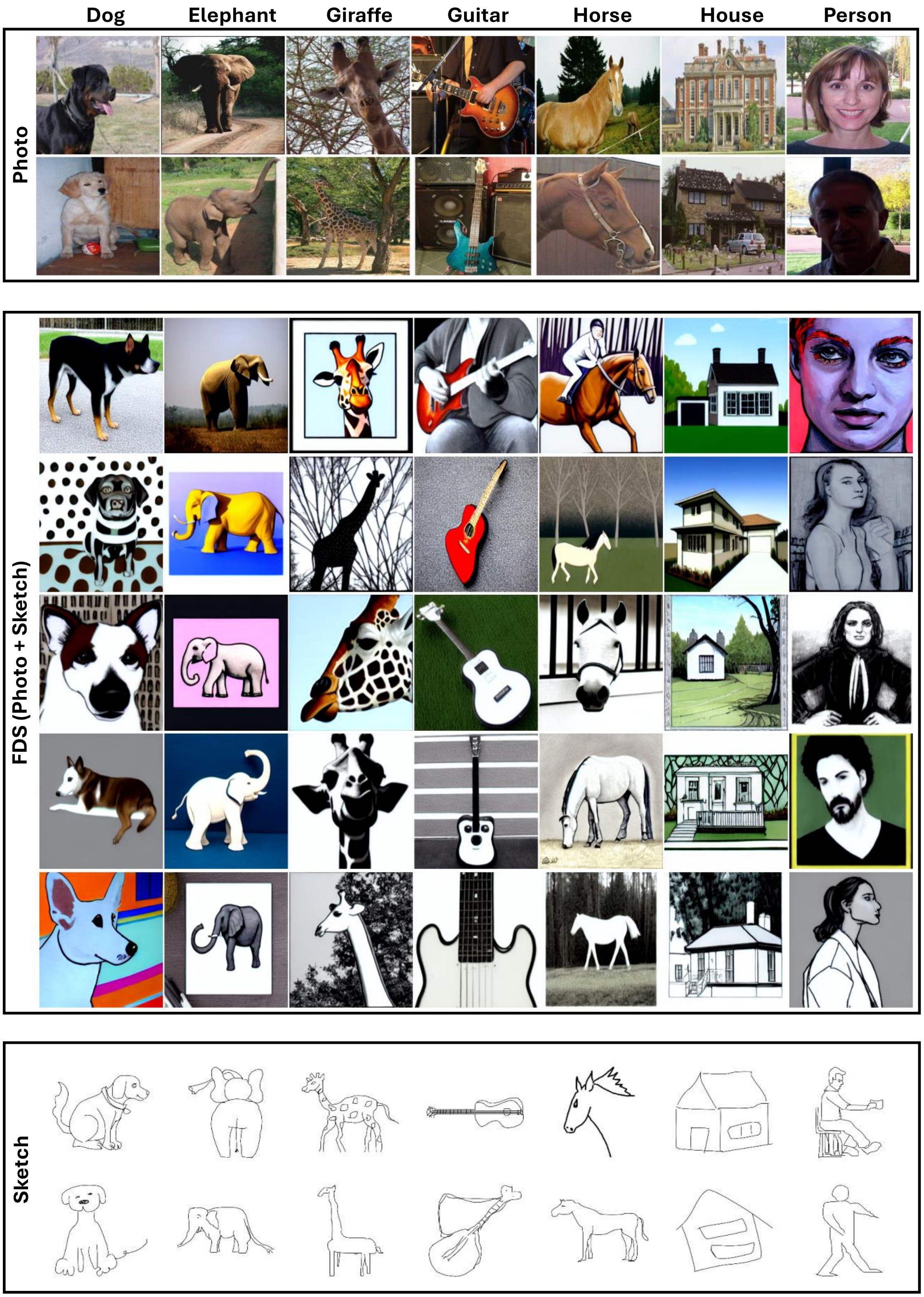

Figure 11. Visual comparison of original "Photo" and "Sketch" samples from PACS with synthetic images generated by FDS (Photo + Sketch). The middle section demonstrates how FDS integrates the photorealistic details of the "Photo" domain with the elements of the "Sketch" domain.

References

[1] Masih Aminbeidokhti, Fidel A Guerrero Pena, Heitor Rapela Medeiros, Thomas Dubail, Eric Granger, and Marco Pedersoli. Domain generalization by rejecting extreme augmentations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2215–2225, 2024. 5, 6, 7

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 4, 5, 6, 7

[3] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34, 2021. 5, 6, 7

[4] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 5, 6, 7

[5] Xu Chu, Yujie Jin, Wenwu Zhu, Yasha Wang, Xin Wang, Shanghang Zhang, and Hong Mei. DNA: Domain generalization with diversified neural averaging. In *International Conference on Machine Learning*, pages 4010–4034. PMLR, 2022. 5, 6, 7

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 4, 5, 6, 7

[7] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *ArXiv*, abs/2007.01434, 2021. 4, 5, 6, 7, 8, 9

[8] Sobhan Hemati, Mahdi Beitollahi, Amir Hossein Estiri, Bassel Al Omari, Xi Chen, and Guojun Zhang. Cross domain generative augmentation: Domain generalization with latent diffusion models. *arXiv preprint arXiv:2312.05387*, 2023. 4, 5, 6, 7

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[10] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision*, pages 124–140. Springer, 2020. 4, 5, 6, 7

[11] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9619–9628, 2021. 4, 5, 6, 7

[12] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 5, 6, 7

[13] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 5, 6, 7

[14] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 4, 5, 6, 7

[15] Qiaowei Miao, Junkun Yuan, Shengyu Zhang, Fei Wu, and Kun Kuang. Domaindiff: Boost out-of-distribution generalization with synthetic data. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5640–5644. IEEE, 2024. 5, 7

[16] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8690–8699, June 2021. 4, 5, 6, 7

[17] 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

[18] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022. 4, 5, 6, 7

[19] Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022. 5, 6, 7

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[21] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 4, 5, 6, 7

[22] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[23] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 4, 5, 6, 7

[24] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020. 4, 5, 6, 7

[25] Runpeng Yu, Songhua Liu, Xingyi Yang, and Xinchao Wang. Distribution shift inversion for out-of-distribution prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3592–3602, 2023. 5

[26] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34, 2021. 5, 6, 7

[27] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021. 5, 6, 7