ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446788

DOMAINDIFF: BOOST OUT-OF-DISTRIBUTION GENERALIZATION WITH SYNTHETIC DATA

Qiaowei Miao[1] *Junkun Yuan*[1] *Shengyu Zhang*[1] *Fei Wu*[1] *Kun Kuang*[1*]

[1] Zhejiang University, Hangzhou, China

ABSTRACT

In contemporary machine learning, enhancing model generalization through diversified datasets is essential. Yet, collecting additional data often faces prohibitive costs and privacy constraints, with no guarantee of improved diversity. In this paper, we propose DomainDiff, featuring a pivotal Word-to-Image Mapping (WIM) mechanism. WIM constructs precise mapping between prompts and images, where the prompts only comprise style and class words. It generates intra-domain data by employing identical prompts to produce source-style images, preserving style and class consistency, thereby diversifying the dataset. Expanding on this innovation, we fuse multiple WIMs and use the prompts with multiple style words to create inter-domain data, which captures a fusion style of multiple source domains. Inter-domain data significantly widens the training data distribution, amplifying diversity. Experimental results demonstrate DomainDiff's transformative potential, improving model performance on real-world data compared to using only real data. These findings highlight DomainDiff's utility in enhancing generalization across diverse real-world scenarios.

Index Terms— Domain generalization, image generation, data distribution shift, model robustness

1. INTRODUCTION

One of the foremost challenges in deep learning pertains to the deterioration of model performance when confronted with out-of-distribution (OOD) data. Consequently, there has been a heightened focus on research endeavors [1, 2] aimed at elevating generalization performance and bolstering reliability. Recent investigations [3, 4] have underscored the pivotal role of comprehensive and diversified datasets in augmenting the generalization capabilities of machine learning models. Nonetheless, aggregating and curating large-scale datasets from diverse real-world sources entail substantial human and financial resources and paramount data privacy considerations. These constraints invariably confine the scope of training data available for real-world applications, presenting a formidable challenge in harnessing data diversity from authentic sources to cultivate a more adaptable model.

Recently, text-to-image diffusion models, such as Stable Diffusion [5] and Imagen [6], have undergone training on billion-scale datasets [7], exhibiting the remarkable ability to generate lifelike images from textual inputs. Leveraging the impressive capabilities of these generative models, a compelling question arises: Could we use generated data to augment dataset diversity to boost generalization performance? In our evaluation of text-to-image diffusion models, with a particular focus on Stable Diffusion, we have unveiled several noteworthy limitations: (1) The diversity of generated images heavily hinges on using textual prompts. Using fixed and simplistic text prompts can significantly restrict diversity within the same image category. (2) Owing to the polysemy of certain words, generated images may inadvertently fall into different categories than the original data. (3) The stylistic variations in generated images remain limited by pre-training constraints.

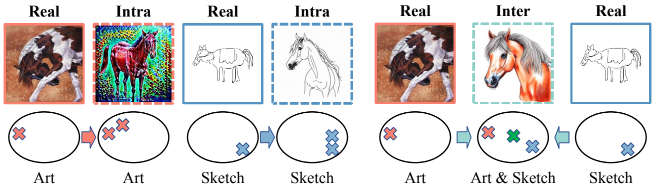


Fig. 1: DomainDiff generates intra-domain data and inter-domain data to boost generalization. The former has the same style as one source domain, and the latter has a mixed style of multiple source domains.

To tackle these challenges, we introduce DomainDiff, a novel data synthesis method that incorporates the Word-to-Image Mapping (WIM) module, meticulously crafted to ensure consistency in style and class, all conditioned on corresponding words. DomainDiff leverages style words and class names as textual prompts in the generation process, alleviating the need for extensive manual text design. The WIM module excels at reconstructing the mapping between style words, class names, and images within each domain, effectively resolving issues related to polysemy. This enhancement enables us to convey a style with a single word precisely. With this accurate word-to-image mapping, DomainDiff effectively generates data that faithfully captures the stylistic essence of the source domains, which we refer to as intra-domain data. Moreover, by amalgamating WIMs trained across multiple domains, DomainDiff gains the capability to create inter-domain data characterized by a fusion of unique styles. This approach significantly enhances dataset diversity, broadening its scope. These data exemplify the distinctive styles as illustrated in Fig. 1.

To prove that the data synthesized by DomainDiff can boost the generalization performance, we assess models trained on this data across three tasks: multi-source OOD, single-source OOD, and transfer learning, including 13 datasets. Our experiments confirm that DomainDiff-generated data effectively reduces the distribution gap between training and testing data, enhancing model generalization within and across domains. In summary, our contributions can be summarized as follows:

- We highlight three limitations of the data generated by Sta-

This work was supported in part by National Natural Science Foundation of China (No. 62376243, 62037001, U20A20387), the StarryNight Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010), Project by Shanghai AI Laboratory (P22KS00111) and Program of Zhejiang Province Science and Technology (2022C01044). (* Corresponding Author: Kun Kuang).

bleDiff: misinterpretation, inaccurate word-to-image mapping, and limited diversity.

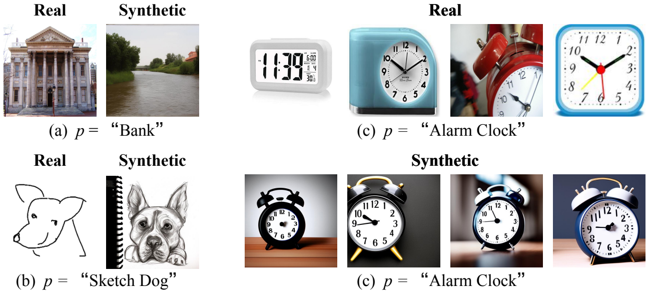
- We propose a DomainDiff data generation method, which reconstructs the word-to-image mapping with simple and fixed prompts. DomainDiff accurately generates intra-domain data that maintains consistency with the style of the source domain. Furthermore, DomainDiff employs weight fusion to produce innovative styles and generates inter-domain data that bridge the divide between disparate source domains.
- In multi-source OOD, single-source OOD, and transfer learning three tasks, DomainDiff synthesized data significantly boost the performance of models across 13 datasets.

2. RELATED WORKS

Out-of-domain generalization. In the realm of out-of-domain generalization [8, 9, 10, 11], the primary objective is to train a model capable of robust performance on unseen data. Traditionally, data generation methods [12, 13, 14, 15] employed in the OOD task have largely focused on enhancing dataset diversity through modifications to the original data. These modifications often involve techniques such as interpolation, cropping, and other transformations. Most of these methods have relied on two prevalent model architectures: generative models like Variational Auto-encoders (VAE) and Generative Adversarial Networks (GAN). However, these approaches are primarily constrained to manipulating existing input data. Consequently, they cannot generate a substantial volume of diverse, entirely novel data. This limitation poses challenges when attempting to scale datasets effectively.

Text-to-image diffusion models. These generative models are exemplified by Stable Diffusion [5] (StableDiff), known for their remarkable ability to produce high-quality images. These models operate by gradually introducing Gaussian noise to data and then recovering the original data through a series of intricate processes. In the realm of conditional diffusion models, this reverse process is conditioned on specific signals, such as class names and style words, enabling the generation of images tailored to particular conditions. During the generation phase, employing different textual descriptions yields a rich spectrum of images, showcasing the diversity-promoting potential of these models. Recent research efforts [6, 16] delve into the effectiveness of controlling image attributes using text inputs within diffusion models, further enriching the diversity of generated images. However, it's worth noting that manually specifying appropriate prompts for each class becomes impractical as the number of class names increases. Furthermore, inadequate prompt design during the training phase can generate images that lack diversity or even misrepresent their intended class, underscoring the limitations of text-to-image diffusion models.

Learning from synthetic data. Employing synthetic data to train models and boost their performance is common in various applications. In exploring the diffusion model for OOD tasks, we identify two related works that share certain aspects with our approach. Sariyildiz et al. [16] introduced ImageNet-clone, where they leveraged the Stable Diffusion model to generate a dataset of equal size. They employed complex prompts to generate different data for boosting classification models' generalization and transfer learning performance. However, the complex prompts necessitate extensive manual testing to select suitable prompts to synthesize data, showcasing the advantages over real data. Azizi et al.[17] focus on generating realistic data with complex prompts by fine-tuning the diffusion model. In contrast, DomainDiff stands out by specializing in creating highly diverse images while minimizing text dependencies. It offers the unique capability to generate both intra-domain and inter-domain data. Intra-domain data maintains the style of the source domain, while inter-domain data introduces entirely new styles. Moreover, DomainDiff achieves this with minimal reliance on textual prompts, making the process less labor-intensive.



(a) $p =$ "Bank"
(c) $p =$ "Alarm Clock"
(b) $p =$ "Sketch Dog"
(c) $p =$ "Alarm Clock"

Fig. 2: Qualitative examples. Under the condition of minimal text dependence, we present three main limitations of the current text-to-image models. (a) Misinterpretation due to word ambiguity; (b) Limited understanding of adjectives; (c) Limited diversity of the same object with the fixed prompt.

3. DOMAINDIFF

3.1. Generation Capability Check

The StableDiff model, renowned for its image generation capabilities, has demonstrated remarkable potential in generating images from textual descriptions. However, our investigation into StableDiff's generation capabilities has unearthed three notable weaknesses that hinder its application in OOD tasks: (i) **Misinterpretation due to word ambiguity**. StableDiff struggles when confronted with words carrying multiple meanings or ambiguous contexts. For instance, the word "Bank" may be misinterpreted as a riverbank instead of a financial institution, adversely impacting both classification and generalization, as shown in Fig. 2 (a). (ii) **Limited understanding of adjectives**. Without complex prompts, StableDiff could only apply filters to images of real styles based on adjectives rather than achieving the simplicity and abstraction of real data, as illustrated in Fig. 2 (b). (iii) **Limited diversity of the same object with fixed prompt.** Even when the words describing the target have no ambiguity, the generated images are not as rich in variety as real-world samples. As shown in Fig. 2 (c), the upper images demonstrate the diversity of alarm clocks in the real world. However, The lower part of the image consists of generated images in which the alarm clocks have only a single design. These limitations indicate that using text-to-image models blindly could be very risky. Especially when generating data on a large scale, not every incorrect image can be detected and removed. DomainDiff constructs a correct word-to-image mapping and uses domain fusion to enhance the diversity of synthetic data, which effectively overcomes the limitations and boosts the generalization performance.

3.2. Word-to-Image Mapping

The three weaknesses discussed above stem from the lack of proper word-to-image mapping. Misinterpretation Problem: StableDiff struggles to map class names to the correct images in the current dataset. Limited Understanding of Adjectives: StableDiff lacks sufficient training to capture the correct mapping between adjec-

Table 1: Results of multi-source OOD task. (∗:StableDiff, •:Real, †:Intra-domain, ‡:Inter-domain, **bold**: best results).

Algorithm	PACS A	PACS C	PACS P	PACS S	PACS Avg	OfficeHome A	OfficeHome C	OfficeHome P	OfficeHome R	OfficeHome Avg
ERM•	82.7 ± 0.9	79.1 ± 0.9	95.3 ± 0.4	75.9 ± 0.5	83.2	57.0 ± 1.1	50.5 ± 0.9	70.9 ± 0.0	74.3 ± 0.6	63.2
Mixup•	83.9 ± 0.5	74.4 ± 1.9	94.8 ± 0.1	75.3 ± 0.9	82.1	55.7 ± 0.7	50.6 ± 0.3	73.4 ± 0.4	73.7 ± 0.2	63.3
CORAL•	84.0 ± 1.4	78.7 ± 1.4	94.9 ± 0.3	75.3 ± 2.2	83.2	54.4 ± 0.9	51.1 ± 0.6	71.6 ± 1.0	73.5 ± 0.5	62.7
ERM•∗	83.4 ± 0.7	80.1 ± 0.9	93.9 ± 0.6	74.8 ± 0.4	83.0	54.3 ± 0.4	50.8 ± 1.1	72.5 ± 0.5	73.1 ± 0.7	62.7
Mixup•∗	82.4 ± 0.6	81.6 ± 0.4	93.6 ± 0.0	75.3 ± 1.5	83.2	55.6 ± 0.5	50.2 ± 0.6	72.6 ± 0.1	74.9 ± 0.3	63.3
CORAL•∗	79.7 ± 0.5	84.4 ± 0.0	94.7 ± 0.1	75.5 ± 1.8	83.6	55.8 ± 0.6	51.0 ± 0.4	72.8 ± 0.4	74.2 ± 0.1	63.5
ERM•†	83.5 ± 1.6	79.6 ± 0.7	95.7 ± 0.3	79.7 ± 2.0	84.6	55.4 ± 0.0	48.7 ± 0.9	74.3 ± 0.1	76.3 ± 0.0	**63.7**
Mixup•†	85.7 ± 0.7	80.8 ± 1.0	95.8 ± 0.6	80.2 ± 0.5	**85.6**	57.1 ± 0.4	49.1 ± 1.3	74.5 ± 0.6	74.5 ± 0.5	63.8
CORAL•†	84.3 ± 0.7	80.3 ± 0.9	95.2 ± 0.3	80.5 ± 2.2	85.1	57.9 ± 0.3	49.8 ± 0.2	72.2 ± 0.0	75.0 ± 0.3	63.7
ERM•†‡	84.9 ± 1.6	82.9 ± 0.0	95.5 ± 0.0	79.0 ± 0.9	**85.6**	57.6 ± 0.4	49.2 ± 0.6	73.0 ± 0.6	75.2 ± 0.9	**63.7**
Mixup•†‡	87.2 ± 1.0	80.3 ± 1.5	96.5 ± 0.2	78.5 ± 0.8	**85.6**	60.0 ± 1.2	50.7 ± 1.1	74.7 ± 0.3	75.9 ± 0.5	**65.3**
CORAL•†‡	87.5 ± 1.3	83.6 ± 0.0	95.4 ± 0.6	81.5 ± 0.3	**87.0**	59.5 ± 1.7	51.9 ± 0.8	74.8 ± 0.5	76.0 ± 0.4	**65.6**

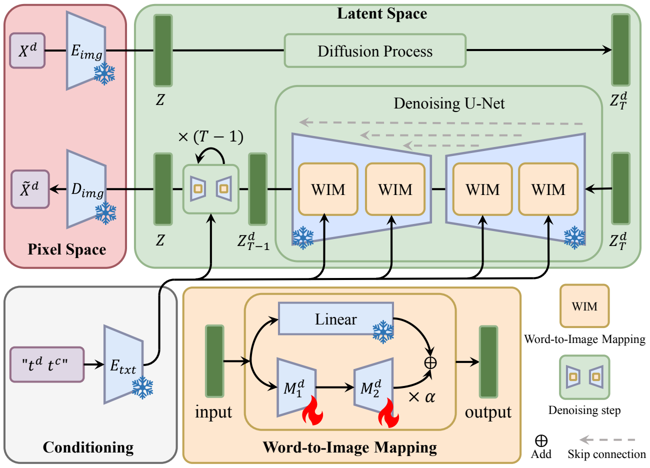


Fig. 3: Overview of DomainDiff. We revamp each linear layer in the UNet by replacing it with a Word-to-Image Mapping (WIM) module. The WIM module comprises two learnable linear layers, M_1^d and M_2^d, whose outputs as residuals are summed with the outputs of the original linear layer.

tives and the visual characteristics of the data. Limited Diversity: StableDiff imposes constraints that limit the diversity of shapes associated with class names. To address these issues, we propose the Word-to-Image Mapping (WIM) module, which aims to establish an accurate word-to-image mapping while minimizing manual intervention in the training process. For datasets like PACS [18] and OfficeHome [19], which are the multiple-source domain generalization datasets, we have access to only class names t^c and domain names t^d. So, we set a fixed prompt p^d = "t^d t^c" for the class c of images within the domain d. We send the images and the prompt into the diffusion model G^d with the Word-to-Image Mapping (WIM) module. Inspired by previous approaches [20], WIM applies two additional layers M_1^d and M_2^d, whose outputs as residuals are summed with the outputs of the original linear layer, as shown in Fig. 3. The output Z_{output} of each WIM can be calculated:

$$Z_{\text{output}} = L(Z_{\text{input}}) + \alpha M_2^d(M_1^d(Z_{\text{input}})). \tag{1}$$

where the L is the origin linear layer in Unet, and the α is the scaling hyperparameter. We freeze the weights of G^d except the weights of WIMs and employ a fixed text prompt p^d = "t^d t^c" to construct the word-to-image mapping. The DomainDiff with WIMs trained in domain d can generate images $\hat{X}^d$ that share the same style but are distinct from any data in source domain d. We donate these images as intra-domain data, which enhances the diversity of in-distribution data for boosting out-of-distribution generalization [21].

3.3. Domain Fusion

While intra-domain data effectively enhances the diversity within source domains, a style gap often persists among different source domains, leaving the data within these gaps untapped. This hidden data between distinct source domains holds the potential to enrich the overall training dataset further, making it imperative to explore and exploit this resource. To bridge the style gap and unlock the hidden diversity, we introduce the concept of domain fusion. Domain fusion involves the integration of multiple G^d models and their respective domain labels t^d. This integration empowers us to generate inter-domain data by sampling from the spaces between source domains. For instance, let's consider two DomainDiff models, G^A and G^S, representing source domains 'Art' and 'Sketch', respectively. To achieve domain fusion, we combine the Word-to-Image Mapping (WIM) modules from both models:

$$\begin{aligned} M_1^{A,S} &= \beta M_1^A + (1-\beta) M_1^S, \\ M_2^{A,S} &= \beta M_2^A + (1-\beta) M_2^S, \end{aligned} \tag{2}$$

Here, β represents the fusion hyperparameter. The output of the domain fusion WIM is computed as follows:

$$Z_{\text{output}} = L(Z_{\text{input}}) + \alpha M_2^{A,S}(M_1^{A,S}(Z_{\text{input}})). \tag{3}$$

After fusion, the DomainDiff $G_{inter}^{A,S}$ requires the use of textual prompts $p_{inter}^{A,S}$ = "t^A t^S t^c," such as "art sketch horse", to generate inter-domain images. In other words, $X^{A,S}inter = G_{inter}^{A,S}(p_{inter}^{A,S}, \epsilon)$, where ϵ represents random noise.

4. EXPERIMENTS

In this section, we conduct experiments to answer four main questions: **RQ1:** Does the synthetic data boost the multi-source OOD performance? **RQ2:** Does the synthetic data boost the single-source OOD performance? **RQ3:** Does the synthetic data boost the transfer learning performance? **RQ4:** How does synthetic data boost the generalization performance of the classification model?

Dataset. For multi-source OOD, we select the PACS [18] and OfficeHome [19] datasets. For single-source OOD, following the settings of ImageNet-clone [16], we consider ImageNet-100

Table 2: Results of single-source OOD task. Top-1 and Top-5 accuracy on several ImageNet datasets. IN-A is tested using only the categories that intersect with IN-100. (●:Real,∗:StableDiff, †:Intra-domain, **bold**: best results).

Data	Real:Syn	IN-Val		IN-V2		IN-Sketch		IN-A	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
IN-100●	1:0	87.5	93.8	78.1	92.2	37.5	58.9	26.7	62.5
IN-100●∗	1:0.5	87.5	95.3	79.7	90.6	35.9	**64.1**	31.3	67.2
IN-100●†	1:0.5	**89.1**	**96.9**	**81.2**	**93.8**	**39.1**	62.5	**32.8**	**73.4**

Table 3: Results of transfer learning. (●:Real,∗:StableDiff, †:Intra-domain, **bold**: best results).

Data	Real:Syn	Aircraft	Cars196	DTD	EuroSAT	Flowers	Pets	Food101
IN-100●	1:0	43.6	41.5	67.9	96.2	85.6	78.7	63.4
IN-100●∗	1:0.5	48.7	43.7	70.2	96.2	**89.6**	83.3	68.1
IN-100●†	1:0.5	**49.2**	**47.0**	**71.8**	**96.5**	**89.6**	**84.7**	**68.3**

(IN-100) as a single source domain and use ImageNet-val (IN-val), ImageNet-V2 (IN-V2), ImageNet-Sketch (IN-S), ImageNet-R (IN-R), and ImageNet-A (IN-A) as five target domains in our experiments. For transfer learning, We evaluate the transfer performance of our models on eight datasets: Aircraft [22], Cars196 [23], DTD [24], EuroSAT [25], Flowers [26], Pets [27], and Food101 [28].

Implementation details. ResNet50 is used as the backbone network for all models in all experiments reported in this paper. We use publicly available `DomainBed` [29] for multi-source OOD. We primarily test three algorithms: Empirical Risk Minimization (ERM), Inter-domain Mixup (Mixup) [30], and Deep CORrelation ALignment (CORAL) [31]. We use `TREX` [32] to evaluate models' single-source OOD and transfer learning performance. For these two tasks, we follow the settings of previous works [16], with text control parameters set to 2.0 and 50 iterations for inversion. Besides, we use ”*” as the domain label for training WIMs in IN-100. We use the publicly available StableDiff on `HuggingFace`, with the scaling hyperparameter α set to 1.0. In the domain fusion phase, the fusion hyperparameter β is set to 0.5 to balance the quality and diversity of the generated images.

RQ1: boosting multi-source OOD task performance. We use DomainDiff to synthesize equal amounts of intra-domain and inter-domain data. We train the classification model with synthetic and real data on PACS and OfficeHome. As shown in Table 1, when the model is trained using images generated by the StableDiff, there is no significant improvement compared to models trained exclusively on real data. In contrast, using inter-domain data leads to noticeable enhancements in model performance. What's even more significant is that when inter-domain images are introduced into the training data mix, the resulting models demonstrate a remarkable improvement in performance on both datasets. The accuracy of ERM, Mixup, and CORAL improved by an average of 1.8, 2.2 and 2.8 points on the two datasets, respectively.

RQ2: boosting single-source OOD task performance. Expanding our investigation further, we extend our study to the harder scenario: single-source domain task. ImageNet is a typical dataset where all training data share the same authentic style. We trained a classification model on ImageNet and assessed its generalization performance across several test sets with significantly distinct styles. As shown in Table 2, for test domains with similar styles (IN-val, IN-V2), DomainDiff notably boosts the model's generalization performance and gets 89.1% and 81.2% Top-1 accuracy. Furthermore, for test domains encompassing diverse styles (IN-Sketch, IN-A), DomainDiff outperforms StableDiff to generate more useful data to get 39.%1 and 32.8% Top-1 accuracy.

RQ3: boosting transfer learning performance. As a step forward, we evaluate the transfer learning performance of the model trained in the single-source domain task. As shown in Table 3, the model trained with DomainDiff generated data has a better ability to extract representations, which suits new classes, even new downstream tasks, and the model leads by an average of 1.0 points across seven datasets.

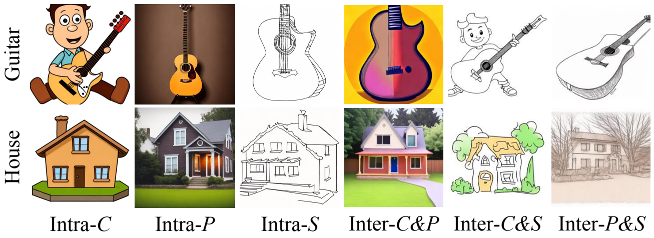
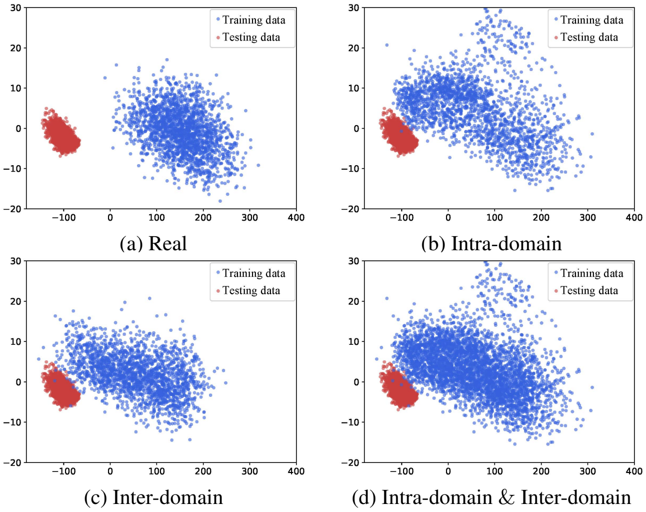


Fig. 4: Diverse data synthesized by DomainDiff.



(a) Real

(b) Intra-domain

(c) Inter-domain

(d) Intra-domain & Inter-domain

Fig. 5: Distribution comparison of testing and training data in feature space with the $Sketch$ as the target domain.

RQ4: DomainDiff reduces distribution gap. We present examples of both intra-domain and inter-domain data in Fig. 4. Intra-domain data maintains the original domain-specific style, while inter-domain data seamlessly blends multiple source domain styles. The complementarity of these styles is evident in the feature space's data distribution, as illustrated in Fig. 5. Intra-domain data exhibits a wide distribution that closely aligns with the target domain's data distribution. Furthermore, incorporating inter-domain data expands the training data distribution even further. Importantly, it's worth noting that inter-domain and intra-domain data distributions emphasize different aspects, underscoring the importance of generating inter-domain data within the context of multi-source domain generalization.

5. CONCLUSION

In this paper, we propose the DomainDiff, which reconstructs the word-to-image mapping in each domain to ensure consistency in styles and categories conditioned on corresponding words. DomainDiff relies on only one style word and one class name as textual prompts to minimize the need for manual text design. Experimental results demonstrate that both intra-domain and inter-domain data generated by DomainDiff can narrow the distribution gap between training and testing data, leading to improved generalization performance.

6. REFERENCES

[1] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen, "Deep stable learning for out-of-distribution generalization," in *CVPR*, June 2021, pp. 5372–5382.

[2] Qixun Wang, Yifei Wang, Hong Zhu, and Yisen Wang, "Improving out-of-distribution generalization by adversarial training with structured priors," in *NIPS*. 2022, vol. 35, pp. 27140–27152, Curran Associates, Inc.

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al., "Segment anything," *arXiv:2304.02643*, 2023.

[4] Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He, "Scaling vision-language models with sparse mixture of experts," *EMNLP(Findings)*, 2023.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, Jun 2022.

[6] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al., "Photorealistic text-to-image diffusion models with deep language understanding," *NIPS*, vol. 35, pp. 36479–36494, 2022.

[7] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al., "Laion-5b: An open large-scale dataset for training next generation image-text models," *arXiv:2210.08402*, 2022.

[8] Junkun Yuan, Xu Ma, Defang Chen, Kun Kuang, Fei Wu, and Lanfen Lin, "Domain-specific bias filtering for single labeled domain generalization," *IJCV*, vol. 131, no. 2, pp. 552–571, 2023.

[9] Junkun Yuan, Xu Ma, Ruoxuan Xiong, Mingming Gong, Xiangyu Liu, Fei Wu, Lanfen Lin, and Kun Kuang, "Instrumental variable-driven domain generalization with unobserved confounders," *TKDD*, 2023.

[10] Junkun Yuan, Xu Ma, Defang Chen, Fei Wu, Lanfen Lin, and Kun Kuang, "Collaborative semantic aggregation and calibration for federated domain generalization," *TKDE*, 2023.

[11] Anpeng Wu, Junkun Yuan, Kun Kuang, Bo Li, Runze Wu, Qiang Zhu, Yueting Zhuang, and Fei Wu, "Learning decomposed representations for treatment effect estimation," *IEEE TKDE*, vol. 35, no. 5, pp. 4989–5001, 2023.

[12] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot, "Domain generalization with adversarial feature learning," in *CVPR*, June 2018.

[13] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool, "Combogan: Unrestrained scalability for image domain translation," in *CVPR workshops*, 2018, pp. 783–790.

[14] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang, "Learning to generate novel domains for domain generalization," in *ECCV*. Springer, 2020, pp. 561–578.

[15] Nathan Somavarapu, Chih-Yao Ma, and Zsolt Kira, "Frustratingly simple domain generalization via image stylization," 2020.

[16] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis, "Fake it till you make it: Learning transferable representations from synthetic imagenet clones," in *CVPR*, 2023.

[17] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet, "Synthetic data from diffusion models improves imagenet classification," *arXiv:2304.08466*, 2023.

[18] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales, "Deeper, broader and artier domain generalization," in *ICCV*, 2017, pp. 5542–5550.

[19] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *CVPR*, 2017, pp. 5018–5027.

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," *arXiv:2106.09685*, 2021.

[21] John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt, "Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization," 2021.

[22] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi, "Fine-grained visual classification of aircraft," *arXiv:1306.5151*, 2013.

[23] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei, "Collecting a large-scale dataset of fine-grained cars," 2013.

[24] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi, "Describing textures in the wild," in *CVPR*, 2014, pp. 3606–3613.

[25] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J-STARS*, vol. 12, no. 7, pp. 2217–2226, 2019.

[26] Maria-Elena Nilsback and Andrew Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 2008, pp. 722–729.

[27] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar, "Cats and dogs," in *CVPR*. IEEE, 2012, pp. 3498–3505.

[28] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool, "Food-101–mining discriminative components with random forests," in *ECCV*. Springer, 2014, pp. 446–461.

[29] Ishaan Gulrajani and David Lopez-Paz, "In search of lost domain generalization," *arXiv:2007.01434*, 2020.

[30] Yufei Wang, Haoliang Li, and Alex C Kot, "Heterogeneous domain generalization via domain mixup," in *ICASSP*. IEEE, 2020, pp. 3622–3626.

[31] Baochen Sun and Kate Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Deep coral: Correlation align- ment for deep domain adaptation*. Springer, 2016, pp. 443–450.

[32] MertBulent Sariyildiz, Yannis Kalantidis, Karteek Alahari, and Diane Larlus, "No reason for no supervision: Improved generalization in supervised models," Jun 2022.