Adapting Without Seeing: Text-Aided Domain Adaptation for Adapting CLIP-like Models to Novel Domains

Louis Hémadou, Héléna Vorobieva, Ewa Kijak, Frédéric Jurie

Adapting Without Seeing: Text-Aided Domain Adaptation for Adapting CLIP-like Models to Novel Domains

Louis Hémadou[1,2,3] Héléna Vorobieva [1] Ewa Kijak[2] Frédéric Jurie[3]

[1] *Safran Tech, Digital Sciences & Technologies Department*
[2] *Université de Rennes, IRISA, INRIA, CNRS*
[3] *Université de Caen Normandie, ENSICAEN, CNRS*
email: louis.hemadou@safrangroup.com

***Abstract*—This paper addresses the challenge of adapting large vision models, such as CLIP, to domain shifts in image classification tasks. While these models, pre-trained on vast datasets like LAION 2B, offer powerful visual representations, they may struggle when applied to domains significantly different from their training data, such as industrial applications. We introduce TADA, a Text-Aided Domain Adaptation method that adapts the visual representations of these models to new domains without requiring target domain images. TADA leverages verbal descriptions of the domain shift to capture the differences between the pre-training and target domains. Our method integrates seamlessly with fine-tuning strategies, including prompt learning methods. We demonstrate TADA's effectiveness in improving the performance of large vision models on domain-shifted data, achieving state-of-the-art results on benchmarks like DomainNet.**

***Index Terms*—deep learning, domain generalization, vision-language models.**

I. Introduction

Recent advancements in machine learning have led to the development of large vision models, such as CLIP [1], BASIC [2], and ALIGN [3], which are pre-trained on vast datasets of image-text pairs. These models, comprising a visual encoder and a text encoder trained to minimize a contrastive loss, have demonstrated remarkable capabilities in various vision tasks. Their strength lies in the ability to extract useful information from images, aligned with text-based representations of their content.

However, the effectiveness of these models can be compromised when applied to domains that differ significantly from their pre-training data. This challenge is particularly evident in specialized fields like industrial applications, where the visual characteristics of the data may deviate substantially from the web-scraped images typically used in pre-training. To create a model that performs well on a specific problem, it is common to fine-tune the pre-trained model. However, in many cases, there is a shift between the domain of the available training data and that of the test data during the fine-tuning phase. For instance, in industrial applications, real-world data may be limited or unattainable, whereas synthetic data can be conveniently generated.

This mismatch between the training and target domains results in a domain shift, often leading to performance degradation. Traditional domain adaptation techniques [4] address such shifts using limited, potentially unlabeled samples from the target domain. Alternatively, domain generalization methods [5]–[8] aim to create models that are robust to unseen domains without any target domain information. Furthermore, the text/image alignment of multimodal models such as CLIP allows generalization guided by textual descriptions of the target images. Several methods focus on shifting the visual features towards the target domain using textual descriptions of the source and target images. PODA [9] and 'CLIP the gap' [10] propose learning a specific augmentation for each source image, operating within CLIP's visual encoder. These augmentations are simple functions, such as translation or instance normalization. The augmented features are subsequently forwarded to a semantic segmentation or detection algorithm. Additionally, LADS and conceptually similar methods [11]–[13] utilize StyleGAN-NADA's [14] directional loss to train an augmentation function per target domain directly in CLIP space. This directional loss lies on the empirical observation that a semantic shift in the text modality corresponds to the same shift in the visual modality in CLIP space.

In this context, we propose TADA (Text-Aided Domain Adaptation), a novel method designed specifically for adapting large vision models to new domains without requiring target domain images. TADA leverages high-level knowledge about the target domain, provided in the form of textual descriptions, to adapt the visual representations produced by the encoder of the large vision model. Our method is similar to [11]–[13] in that we learn an augmentation function per target domain, but does not rely on the directional loss of StyleGAN-NADA. TADA builds upon the inherent capabilities of text-image models to align visual and textual representations. By utilizing textual descriptions of the target domain, TADA aims to learn to transform the representation of source images from the training distribution to the target distribution. This approach allows for effective domain adaptation without the need for target domain images, making it particularly suitable for scenarios where such data is scarce, expensive, or impossible to obtain.

The main contributions of this paper are:

- The introduction of TADA, a novel text-based adaptation method for large multimodal vision models.

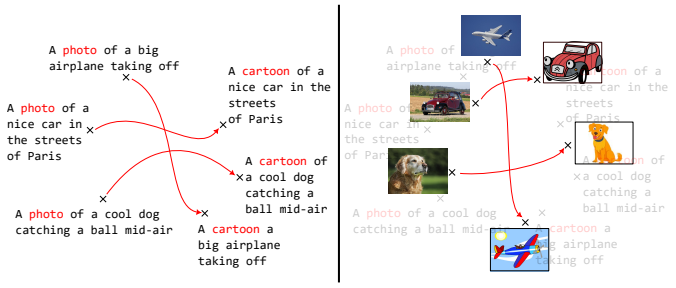

Fig. 1. TADA is a novel data augmentation method consisting of learning a function that maps source captions to target captions (left). It then augments the source images with this function to bring them closer to the target distribution (right). All these operations take place in CLIP space, where text and image are semantically aligned.

- A framework for learning to adapt visual representations to new domains using only textual descriptions.
- An empirical validation of TADA's effectiveness in improving the performance of CLIP on domain-shifted data in the standard benchmarks such as PACS [15], OfficeHome [16], and DomainNet [17].

II. Method

We propose TADA (**T**ext **A**ided **D**omain **A**daptation), a zero-shot domain adaptation method for image classification that exploits the text-image description space of a multimodal model. As illustrated in Fig. 1, the key idea is to learn an adaptation function that transforms textual representations from the source domain to the target domain. This adaptation function is then applied to the image features from the source domain, effectively transporting them into the target domain representation space. Finally, we fine-tune CLIP on these augmented image representations for the target domain, either via linear probing or prompt learning.

A. Problem Formulation and Notations

The paper focuses specifically on single-source zero-shot domain adaptation for N-way image classification, where N represents the number of classes.

In this context, labeled samples from one source domain are accessible at train time, represented as $\{(\mathbf{x}_{\mathrm{k}}^{\mathrm{s}}, y_k^s)\}_{k=1}^n \sim \mathcal{D}_s$. Here, $\mathbf{x}_{\mathrm{k}}$ refers to the k-th sample from the source domain and y_k is the one-hot encoding of its label. The test data, upon which the classifier is evaluated, originates from d target domains $\mathcal{D}_t^i$ and is denoted as $\{\mathbf{x}_{\mathrm{k}}^{\mathrm{i}}\}_{k=1}^{n_i} \sim \mathcal{D}_t^i$, $i \in \{1, \ldots, d\}$.

The approach builds on a multimodal model with text-image contrastive pretraining, such as CLIP [1]. We denote its visual encoder as E_V and its text encoder as E_T. The visual features encoded by E_V are represented as $\mathbf{z}_{\mathrm{k}}^{\mathrm{s}} = E_V(\mathbf{x}_{\mathrm{k}}^{\mathrm{s}})$. In addition, class names t_y and source domain textual description p_s are also available. Furthermore, it is assumed that textual descriptions p_i of the target domains $\mathcal{D}_t^i$ are provided during training. Such textual descriptions are used to transfer the image embeddings to the target domains.

TADA is designed to improve the generalization performance of fine-tuning strategies, ranging from zero-shot classification to more sophisticated prompt learning strategies [18], [19], through simple fine-tuning of the head (linear probing). The only constraint we impose is that the fine-tuning method must keep the visual encoder E_V frozen. There are two reasons for freezing the visual encoder. First, it is important to preserve the text-image alignment, a critical requirement for our method, by preventing its degradation. Secondly, maintaining a frozen encoder reduces the computational demands of the model, making it less resource intensive.

Our contribution is centered around the introduction of adaptation functions f_{θ_i}, which are distinct for each target domain. These functions aim to modify the CLIP features of the source images to make them more faithful to the target domains. As a result, the fine-tuning process can then be expressed as follows:

$$\min_{\phi} \sum_{k=1}^{n} \sum_{i=1}^{d} \ell(g_\phi(f_{\theta_i}(\mathbf{z}_{\mathrm{k}}^{\mathrm{s}})), y_k^s) \tag{1}$$

where ℓ is the cross-entropy loss and g_ϕ is the prediction function, parameterized by ϕ. For linear probing, $g_\phi(\mathbf{z}) = W^T\mathbf{z}$, where W is the matrix initialized with the zero shot classification head. For CoOp [18], a prompt learning method, ϕ represents the learnable context tokens $\{[V]_1, ..., [V]_m\}$ and $g_\phi(\mathbf{z}) = (W')^T\mathbf{z}$, where W' is a matrix whose i-th column is computed with $E_T([V]_1...[V]_m[CLASS]_i)$.

B. Learning Adaptation Functions

As previously mentioned, the core idea of TADA is to learn a domain adaptation function in the text modality, where domains can be conveniently described, and then apply this function to image representations. A key strength of our approach lies in the use of multiple prompts to represent the source and target domains, rather than relying on a single prompt. These prompts are generated by combining domain descriptions and diverse contents, providing a more comprehensive representation of the domains. The remainder of this section details the step-by-step procedure we follow.

A caption is a text describing a virtual image in the form `a {style} of {content}`, where `{style}` represents the image style (e.g., `painting`, `photo`) and `{content}` is the style-agnostic description of the image content. We learn augmentation functions using pairs of captions: (`a {source style} of {content}`, `a {target style} of {content}`). The style is the only part of the caption changing within a caption pair. To make the augmentation function robust on images, we want these captions to be as diverse as possible, while still faithfully representing the source and target distributions.

To this end, we first generate M generic captions using a large language model, with the following template: `a picture of {content}`, where `{content}` is a group of words describing the content of the virtual image depicted by the caption, for example `a picture of a cool dog catching a ball mid-air`. In order to cover the widest visual domain using text, we enforce diversity in generated

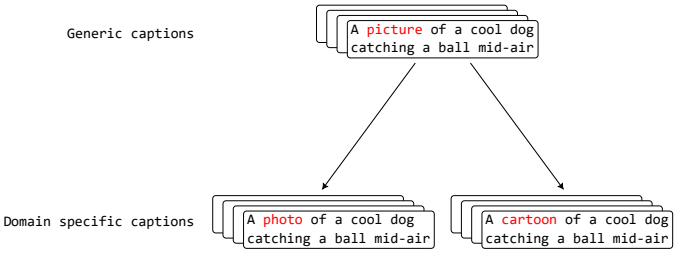

Fig. 2. Caption pairs stylization: the word `picture` is replaced by the textual description of the source and target domains (*e.g.* `photo` and `cartoon`) to generate source and target-specific caption sets.

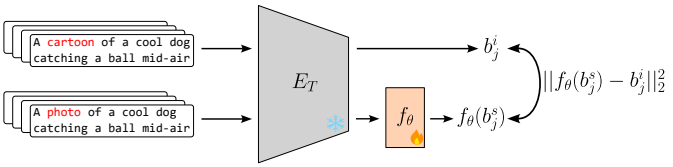

Fig. 3. Training an augmentation function f_θ to transform `photo` captions into `cartoon` captions in CLIP space.

captions by constraining the large language model to generate {`content`} related to diverse visual concepts.

We then add a style to the generic captions by replacing the word `picture` with a domain textual description (see Fig. 2). For every target domain $\mathcal{D}_t^i$, with textual description p_i, we replace the word `picture` by p_s and p_i to create source and target specific caption pairs $\{(c_j^s, c_j^i)\}_{j=1}^M$.

Similar to [11]–[13], we train an adaptation function f_{θ_i} for each target domain. Let $(b_j^s, b_j^i) = (E_T(c_j^s), E_T(c_j^i))$ denote the CLIP embedding of the source and target captions, respectively. We train the adaptation functions to transform each source caption embedding b_j^s into its corresponding target caption embedding b_j^i in the CLIP space. The parameters θ_i of the augmentation function for the i-th target domain are learned by minimizing the following $\mathcal{L}_2$ loss (see Fig. 3):

$$\forall i \;\; \theta_i^* = \arg\min_{\theta_i} \sum_{j=1}^{M} \| f_{\theta_i}(b_j^s) - b_j^i \|_2^2 \tag{2}$$

In our experiments, the adaptation functions f_{θ_i} are implemented as simple fully-connected neural networks (see Section III for architectural details).

C. Fine-tuning CLIP

While fine-tuning the linear zero shot head is a reasonable way to exploit the knowledge embedded in a pre-trained CLIP model, much work in recent years has focused on developing better CLIP-based classification models. Two categories of algorithm have emerged: robust fine-tuning [20]–[22] and prompt learning [18], [19], [23]–[25].

The adaptation functions are computed using text features of a frozen CLIP and should be effective on image features of a frozen CLIP. Thus, our augmentation strategy is compatible with any fine-tuning method that keeps the visual encoder frozen. While this precludes end-to-end fine-tuning, most prompt learning methods keep the visual encoder frozen, allowing us to use TADA in conjunction with state-of-the-art prompt learning methods. During the fine-tuning stage, we discard the original, unadapted image embeddings.

III. Experiments

A. Implementation Details

Generation of Generic Captions. We generated generic captions using Meta's openly accessible Llama3-8b [26]. To enforce diversity in generated captions, we leveraged a list of 7881 visual concepts from the OpenImages v3 dataset [27]. For each visual concept, we asked LLama3 to generate 10 sentences beginning with `a picture of`, related to the visual concept, leading to a generic caption dataset of 78,810 instances.

Modelling Adaptation Functions. We modeled the adaptation function f_{θ_i} with an isometry, represented by a matrix R subjected to the constraint $R^T R = I_d$, where I_d is the d-dimensional identity matrix. Eq. (2) can be solved using the Kabsch algorithm [28]. There are two reasons for modelling the adaptation function with an isometry. First, text and image live on the unit sphere in CLIP space. As a consequence, a function mapping text to text or image to image should be norm-preserving. Second, the structure of CLIP space, where semantically related images and texts have a small angle, leads us to use an angle preserving transformation.

B. Datasets and Textual Descriptions

We evaluate our approach on three datasets commonly used for domain adaptation. **DomainNet** [17] comprises 345 classes and six domains (quickdraw, infograph, sketch, painting, clipart, real). **OfficeHome** [16] comprises 65 classes and four domains (art, clipart, product, real world). **PACS** [15] comprises seven classes and four domains (photo, art painting, cartoon, sketch). Following previous single domain generalization works [13], [29], we train our classification model on the photo realistic domain of each dataset, and evaluate on the remaining domains.

Fig. 4 presents the textual descriptions we used to represent the different domains in our experiments. These descriptions aim to capture the salient characteristics and contexts of each domain, enabling the generation of relevant and domain-specific captions through the language model.

C. Fine-tuning algorithms

As stated before, our adaptation method is compatible with any CLIP-based classification algorithm that keeps the visual encoder frozen. We propose to test the effectiveness of TADA on four algorithms. **LP** (Linear Probing) is a simple fine-tuning of the zero shot linear classification head. We initialize the zero shot head using CLIP features of prompts with the template `A photo of a {classname}`. **WiSE-LP** [20] (Weight Space Ensembling - Linear Probing) effectively averages the predictions of the zero-shot and the fined-tuned linear heads during inference. We used a mixing coefficient of $\alpha = 0.5$. **CoOp** [18] is a prompt learning algorithm that learns the

DomainNet	
real	photo
painting	painting
clipart	clipart
sketch	sketch
quickdraw	quickdraw
infograph	infographic

PACS	
photo	photo
art painting	artistic painting
cartoon	clipart
sketch	sketch

OfficeHome	
real world	photo
art	creative artistic image
clipart	clipart
product	amazon product image

Fig. 4. Domain name (left) and corresponding textual description (right) for the datasets we used.

context, leading to the best accuracy. Finally, **CoCoOp** [19] adds an image-conditioned component into context generation, leading to a better generalization on novel classes and domains. In our implementation, we fixed the context length to 4 for both **CoOp** and **CoCoOp**.

Following the recommendations from DomainBed [30], we selected the stopping iteration and hyperparameters (learning rate, weight decay, batch size) that maximized the accuracy on a validation set comprising 20% of the training data. CLIP features of the images from the validation set are not augmented. This approach ensured that the model's parameters were optimally tuned for the given task.

Among the different available CLIP architectures, we have chosen to rely on the publicly available models from OpenAI's CLIP for our experiments, specifically the ViT-B/16.

D. Baselines

In addition to the four fine-tuning algorithms stated above, we compare TADA with several CLIP-based methods: **CLIP's zero-shot**, **PromptStyler** [31], and **LADS** [11]. **PromptStyler** is a source-free domain generalization algorithm that learns K pseudo-words to represent various styles while remaining close to the contents of the N classes. A linear classifier is then trained using $K \times N$ prompts and applied to the image features during inference. **LADS** [11] is a domain extension method that augments source features with a MLP trained to shift the image along a domain change direction while retaining useful information for classification. In our experiments, we report the accuracy of the unseen domain.

E. Domain Generalization Performance and Comparison with State-of-the-Art Methods

Table I gives the main results we obtained on DomainNet, OfficeHome and PACS. As stated before, we train the classifiers on the photorealistic domain of each dataset and evaluate on the remaining domains. We perform five runs and report the mean micro accuracy (i.e., all samples are given equal weight) over the target domains. Because of the prohibitive computing cost, we did not test **CoCoOp** on **DomainNet** dataset.

The zero shot methods, especially **PromptStyler** [31] demonstrate robust performance, often achieving results that

TABLE I

RESULTS USING THE CLIP VIT-B/16 MODEL. USED IN CONJUNCTION WITH FINE-TUNING ALGORITHMS, TADA SYSTEMATICALLY IMPROVES PERFORMANCE ON THE UNSEEN DOMAINS. THE GAIN OF TADA COMPARED TO NO ADAPTATION IS SHOWN IN RED.

	Method	Dom.Net	Off.Home	PACS
	CLIP ZS	52.70	79.96	95.25
	PromptStyler	54.16±0.00	81.20±0.00	96.33±0.00
	LADS	53.95±0.08	81.25±0.16	93.16±0.12
No adaptation	LP	51.78±0.09	81.08±0.14	92.99±0.05
	WiSE-LP	54.44±0.05	82.29±0.07	94.91±0.03
	CoOp	53.95±0.25	80.30±0.33	92.76±0.39
	CoCoOp	-	81.20±0.04	95.40±0.02
+ TADA	LP	54.20±0.15 +2.42	81.75±0.10 +0.67	93.54±0.05 +0.55
	WiSE-LP	55.50±0.05 +1.06	82.65±0.08 +0.36	95.65±0.06 +0.74
	CoOp	55.30±0.25 +1.35	81.16±0.19 +0.86	93.99±0.42 +1.23
	CoCoOp	-	81.82±0.05 +0.62	96.40±0.08 +1.00

are only slightly lower than robust fine-tuning methods. Due to domain shift, naive fine-tuning of the zero shot head (**LP**) often deteriorates the performance on the unseen domains.

The best results are achieved by combining TADA with strong fine-tuning methods. The improvement brought by TADA is higher for DomainNet than for OfficeHome and PACS. This is due to the less well-defined domains in OfficeHome. For instance, the **art** domain contains images with various styles, such as paintings, sketches and cartoons. Furthermore, there is minimal stylistic difference between the **product** and **real world** domains. This limits the effectiveness of TADA, as it heavily relies on specific domain descriptions. It is also noteworthy that PACS is approaching saturation, making significant accuracy improvements unexpected.

IV. CONCLUSIONS AND FUTURE WORK

This work introduces TADA, a novel adaptation method that effectively adapts images to unseen domains guided by textual descriptions. Used in conjunction with robust fine-tuning and prompt learning strategies, TADA consistently improves the accuracy on the unseen domains.

While current successes lie in aligning domains with clear verbal distinctions, future work will focus on extending this method to address subtler domain shifts. This could involve exploring alternative representations beyond text captions, or incorporating techniques from domain-invariant learning.

This work paves the way for further research on improving the accuracy and completeness of domain representations. This can be achieved by tackling limitations in caption generation and exploring techniques like utilizing source images to refine the caption generation process, drawing inspiration from methods like CoOp [18].

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[2] H. Pham, Z. Dai, G. Ghiasi, K. Kawaguchi, H. Liu, A. W. Yu, J. Yu, Y. Chen, M. Luong, Y. Wu, M. Tan, and Q. V. Le, "Combined scaling for zero-shot transfer learning," *Neurocomputing*, 2023.

[3] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021.

[4] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM*, 2020.

[5] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *CVPR*, 2018.

[6] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, 2016.

[7] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *AAAI*, 2016.

[8] H. Niu, H. Li, F. Zhao, and B. Li, "Domain-unified prompt representations for source-free domain generalization," *CoRR*, 2022.

[9] M. Fahes, T. Vu, A. Bursuc, P. Pérez, and R. de Charette, "Pøda: Prompt-driven zero-shot domain adaptation," in *ICCV*, 2023.

[10] V. Vidit, M. Engilberge, and M. Salzmann, "CLIP the gap: A single domain generalization approach for object detection," in *CVPR*, 2023.

[11] L. Dunlap, C. Mohri, D. Guillory, H. Zhang, T. Darrell, J. E. Gonzalez, A. Raghunathan, and A. Rohrbach, "Using language to extend to unseen domains," in *ICLR*, 2023.

[12] Z. Wang, L. Zhang, L. Wang, and M. Zhu, "Landa: Language-guided multi-source domain adaptation," *CoRR*, 2024.

[13] S. Yan, C. Luo, Z. Yu, and Z. Ge, "Generalizing clip to unseen domain via text-guided diverse novel feature synthesis," *CoRR*, 2024.

[14] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," *ACM*, 2022.

[15] D. Li, Y. Yang, Y. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *ICCV*, 2017.

[16] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *CVPR*, 2017.

[17] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *ICCV*, 2019.

[18] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *IJCV*, 2022.

[19] K. Zhou, J. Yang, C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," *CVPR*, 2022.

[20] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt, "Robust fine-tuning of zero-shot models," in *CVPR*, 2022.

[21] J. Silva-Rodríguez, S. Hajimiri, I. B. Ayed, and J. Dolz, "A closer look at the few-shot adaptation of large vision-language models," *CoRR*, 2023.

[22] X. Mao, Y. Chen, X. Jia, R. Zhang, H. Xue, and Z. Li, "Context-aware robust fine-tuning," *CoRR*, 2022.

[23] M. U. Khattak, H. A. Rasheed, M. Maaz, S. H. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *CVPR*, 2023.

[24] M. U. Khattak, S. T. Wasim, M. Naseer, S. Khan, M. Yang, and F. S. Khan, "Self-regulating prompts: Foundational model adaptation without forgetting," in *ICCV*, 2023.

[25] S. Bose, E. Fini, A. Jha, M. Singha, B. Banerjee, and E. Ricci, "Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization," *WACV*, 2023.

[26] AI@Meta, "Llama 3 model card," 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[27] K. et al., "Openimages: A public dataset for large-scale multi-label and multi-class image classification." *Dataset available from https://github.com/openimages*, 2017.

[28] J. Lawrence, J. Bernal, and C. Witzgall, "A purely algebraic justification of the kabsch-umeyama algorithm." *Journal of research of the National Institute of Standards and Technology*, 2019.

[29] Z. Wang, Y. Luo, R. Qiu, Z. Huang, and M. Baktashmotlagh, "Learning to diversify for single domain generalization," in *ICCV*, 2021.

[30] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *ICLR*, 2021.

[31] J. Cho, G. Nam, S. Kim, H. Yang, and S. Kwak, "Promptstyler: Prompt-driven style generation for source-free domain generalization," in *ICCV*, 2023.