Beyond Internet Images: Evaluating Vision-Language Models for Domain Generalization on Synthetic-to-Real Industrial Datasets

L. Hémadou[1,2,3] H.Vorobieva [1] E. Kijak[2] F. Jurie[3]

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

Abstract

Vision Language Foundation Models (VLFMs) have shown impressive generalization capabilities, making them suitable for Domain Generalization (DG) tasks, such as training on synthetic images and testing on real data. However, existing evaluations predominantly use academic benchmarks constructed from internet images, akin to the datasets used for training VLFMs. This paper assesses the performance of VLFM-based DG algorithms on two synthetic-to-real classification datasets, Rareplanes-tiles and Aerial Vehicles, designed to emulate industrial contexts. Our findings reveal that while VLFMs excel on academic benchmarks, outperforming randomly initialized networks, their advantage is significantly diminished on these industrial-like datasets. This study underscores the importance of evaluating models on diverse, representative data to understand their real-world applicability and limitations.

1. Introduction

The performance of deep learning algorithms heavily relies on the quantity and quality of training samples. In certain scenarios, obtaining realistic data can be costly, making the generation of synthetic data that mimics real data a more convenient alternative. Although synthetic data is becoming increasingly similar to real data, a distribution discrepancy persists. When dealing with images, this discrepancy can arise from various factors such as texture, lighting, and arrangement of elements. Consequently, a deep learning algorithm whose parameters have been learned using synthetic data may perform poorly on real data. This problem is addressed by domain adaptation [24] when a few target samples are available, and by domain generalization [7, 12, 20] when no target samples are available.

Vision Language Foundation Models (VLFMs) such as CLIP [15] or CoCa [29], trained on billions of (text, image) pairs, produce high-quality image representations. VLFMs consist of a visual encoder and a text encoder, trained to minimize a contrastive loss, bringing together semantically related text and images while separating unrelated ones. Their remarkable ability to generalize is a key driver for their successful application in domain generalization tasks.

The multimodal datasets used to train VLFMs [5, 17] are built by collecting images that are openly available on the internet. While internet images represent a vast domain, they are far from encompassing every conceivable task. Consequently, we can expect VLFMs to produce high-quality image representations for images that follow the distribution of "internet images," which may be the case for images from academic benchmarks, primarily constructed by collecting internet images [14, 16]. However, their effectiveness is far from obvious for industrial applications.

This work evaluates the performance of VLFM-based domain generalization algorithms on datasets that deviate from typical academic benchmarks, with a focus on the scenario where a classifier is trained on synthetic images and tested on real images. Specifically, we assess these algorithms on two internal datasets, Rareplanes-tiles and Aerial Vehicles, where we know the images are not available on the internet and thus not present in the multimodal datasets used to train VLFMs. By evaluating on these datasets that are more representative of industrial applications, we aim to gain insights into the real-world applicability and limitations of VLFM-based domain generalization approaches for synthetic-to-real transfer.

2. Related work

Several works have explored training deep models on synthetic data for various computer vision tasks, either generating synthetic datasets [9, 10, 28] or proposing techniques to leverage synthetic data for improving real-world performance [1, 18, 25]. [10] [28] generated large synthetic overhead imagery datasets and demonstrated their benefits for training segmentation and object detection models, respectively. [9] released a synthetic vehicle dataset, showing its advantages over ImageNet for pretraining vehicle-related tasks. [18] reviewed data augmentation and synthetic data generation methods for improving training without additional real data. [1] proposed a contrastive framework to enhance feature diversity and real-world generalization when training on synthetic data. [25] showed the possibility of performing face-related tasks using only highly realistic synthetic training data. [2] investigated domain adaptation techniques for augmenting rare classes with synthetic samples in imbalanced datasets.

More recently, due to the quality of the learned visual encoder, VLFMs have been widely used in domain generalization tasks. The simplest way to use VLFMs for domain generalization is to learn a linear classifier on top of the frozen visual encoder. While this yields good results at very low computational cost, methods have been developed to further improve generalization.

With a textual description of each class, the text-image alignment of VLFMs allows the construction of a simple zero shot algorithm by encoding a prompt representing each class in order to create a linear classification head. It has been shown that adding some context to the prompts enhances the zero-shot algorithm [15]. For example, if the test images have a particular style, e.g. cartoonish, then using prompts such as `a cartoon of a {classname}` will produce a better classifier than the generic prompts `{classname}`. Following this observation, [31] propose CoOp, a few shot classification algorithm using contextual optimisation. More precisely, they aim to learn the context words that lead to the best accuracy on a few set of images, turning CLIP into an effective few shot learner. CoCoOp [30] adds an image-conditioned part into context optimization, leading to better domain generalization performance.

Fine-tuning a model is usually the standard way to benefit from the knowledge embedded in the model. However, in the presence of a distribution shift, fine-tuning VLFMs on source data can actually degrade the performance of the fine-tuned algorithm on target data. Several methods have been developed to mitigate this problem. WiSE (Weight Space Ensembling) [27] proposes to ensemble weights from the zero shot and the fine-tuned model and ”model soups” [26] ensembles weights from fine-tuned models with different set of hyperparameters.

A final category of VLFMs-based DG algorithms relies on a textual description of the target domain to modify source images in order to reduce the domain gap with the target images. [22] and [4] propose to learn an augmentation function per source image that modifies its intermediate representation such that its CLIP representation is closer to the target domain description. The modified intermediate representations are then sent to a segmentation or object detection algorithm. Finally, LADS [3] and LANDA [23] use StyleGAN-NADA's [6] directional loss to learn a global augmentation function that operates in CLIP space.

3. Experimental Study

As mentioned in the introduction, this paper's primary contribution is to conduct a comprehensive experimental campaign aimed at evaluating the effectiveness of approaches based on Vision-Language Foundation Models (VLFMs) like CLIP in learning from synthetic data. The methodology employed for these experiments is detailed in this section.

3.1. Evaluated Methods

We evaluated the following six methods:

1. **SourceOnly** is a simple end-to-end learning approach on images from the source domains (i.e., using only synthetic images without any adaptation mechanism) of a randomly initialized ResNet50 with a linear classification head.
2. **CLIP ZS** is the zero-shot classification algorithm proposed in [15], where a linear classification head is computed with prompts representing each class, while the image encoder remains frozen.
3. **FT CLIP** refers to the end-to-end fine-tuning of the pretrained CLIP visual encoder ResNet50 and the zero shot classification head, aiming to leverage the generalization capability of the CLIP model.
4. **WiSE** [27] is a robust fine-tuning method that consists of averaging the weights of the zero-shot model and the fine-tuned model. In our experiments, we used a mixing coefficient of $\alpha = 0.5$.
5. **CoOp** [31] is a prompt learning method that learns the optimal context that leads to the best accuracy.
6. **CoCoOp** [30] (Conditioned Context Optimization) learns a context adapted to an unseen image, aiming to improve out-of-distribution performance. For **CoOp** and **CoCoOp**, we used a context length of $M = 4$ in our experiments.

Several VLFMs-based domain generalization image classification methods [3, 23] rely on a textual description of the target domain. As we will see in Section 5, these methods are not well-suited for the proposed industrial datasets, as there is no way to find relevant textual descriptions of the target domains.

3.2. Datasets from Scientific Literature

We evaluate VLFMs-based domain generalization algorithms on several datasets commonly used to benchmark domain generalization algorithms: **DomainNet** [14] (6 domains, 345 classes), **PACS** [11] (4 domains, 7 classes), and **OfficeHome** [21] (4 domains, 65 classes).

Each of these datasets has a domain depicting photorealistic images. We conduct our experiments utilizing the photorealistic domain as the target and the other domains as sources in a multi-source setup.

3.3. Industrial Datasets

We also run our experiments on two datasets that are more representative of the type of images encountered in industrial applications. These datasets are more challenging, sometimes featuring low-resolution images and representing objects rarely found in academic benchmarks.

For both datasets, the two domains are synthetic and real images. The first dataset addresses the recognition of **Rareplanes-tiles**, while the second focuses on the recognition of **Aerial Vehicles**. The former is created from the open-source detection dataset Rareplanes [19], and the latter is a dataset used in a company. The composition of the datasets is described in the following paragraphs.

Rareplanes-tiles: Rareplanes [19] is an object detection dataset containing synthetic and real satellite imagery of aircraft seen from the sky at multiple locations. It contains ∼630k synthetic aircraft annotations and ∼14k real aircraft annotations. Each aircraft is accompanied by a collection of characteristics, including the number of engines, role, wings, etc. In total, there are 7 different roles: civil large/medium/small transport, military fighter, military bomber, military transport, and military trainer. However, the synthetic images only contain three different aircraft roles: civil large/medium/small transport. We decided to train a classifier to distinguish between these three roles. We extracted individual aircraft images using the provided bounding boxes. This resulted in a three-way classification problem, with 14k realistic images and 37k synthetic images. For simplicity, we did not use all of the 630k synthetic aircraft annotations. See Figure 1 for image examples. In the following experiments, the classifiers are trained on synthetic images and evaluated on real images.

Aerial Vehicles consists of low-resolution synthetic and real infrared images of vehicles as seen from the sky. The classification problem is to predict the role of the vehicle (car, truck, military truck, armored vehicle, tank) or if there is no vehicle at all. This results in a six-way classification problem, with ∼24k synthetic images and ∼23k real images. See Figure 2 for image examples. In the following experiments, as for **Rareplanes-tiles**, the classifiers are trained on synthetic images and evaluated on real images.

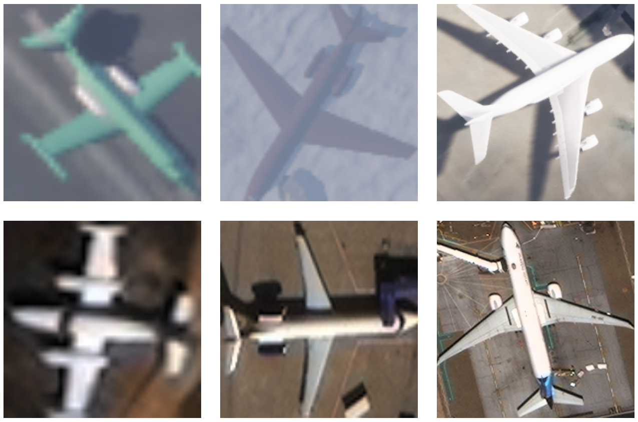

Figure 1. Examples of images from **Rareplanes-tiles**. Top row: synthetic images, bottom row: real images. Left column: small planes, middle: medium planes, right: large planes. Images have an average size of 80 pixels (std = 70 pixels).

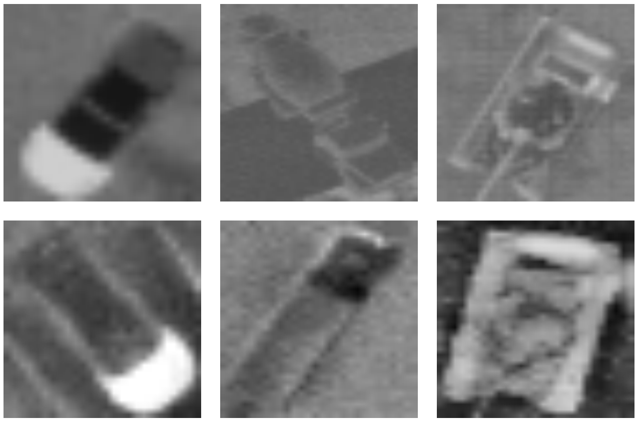

Figure 2. Examples of images from **Aerial Vehicles**. Top row: synthetic images, bottom row: real infrared images. Left column: car, middle: truck, right: tank. Images have an average size of 30 pixels (std = 12 pixels).

3.4. Visual-Language Model

We use the CLIP pre-trained ResNet50 from OpenAI[1]. Its architecture differs slightly from the original ResNet50 in that it replaces global pooling with multi-head attention. We use the same architecture when training the model in the SourceOnly experiments. For the experiments requiring a zero-shot algorithm, we build the linear classification head with prompts such as `a photo of a classname` for academic datasets and `an aerial photo of a classname` for industrial datasets.

3.5. Hyperparameter Tuning

Following the recommendations of DomainBed [8], we chose the hyperparameters (learning rate, weight decay,

[1] https://github.com/openai/CLIP

	OH	**PACS**	**DN**	**RP**	**AV**
SourceOnly	26.1	30.7	12.6	72.9	30.8
CLIP ZS [15]	80.2	**99.5**	76.6	37.4	19.5
FT CLIP	75.3	83.7	55.4	**74.2**	33.5
WiSE [27]	81.6	97.0	77.4	60.4	26.9
CoOp [31]	81.8	99.2	78.9	73.2	**37.0**
CoCoOp [30]	**82.3**	99.3	**79.6**	74.0	36.8

Table 1. Macro-accuracy on OH = OfficeHome, PACS, DN = DomainNet, RP = Rareplanes-tiles, AV = Aerial vehicles.

stopping iteration) that maximized accuracy on a validation set containing 20% of the training images. In all our experiments, we used the AdamW optimizer [13].

4. Results

In these experiments, the source domain comprises synthetic images, while the target domain consists of real images, evaluating the synthetic-to-real transfer capability. As academic datasets from the scientific community do not include synthetic data, we utilize the photorealistic domains as targets and the other domains as sources for a multi-source setup. Each experiment is repeated 5 times, and the average accuracy is reported. Due to a high class imbalance in the real domains of both **Rareplanes-tiles** and **Aerial Vehicles**, we compute the macro accuracy (treating all classes equally) on the test domain to evaluate the methods fairly. The results are presented in Table 1.

On academic datasets, CLIP-based algorithms significantly outperform randomly initialized end-to-end training. For example, accuracy is more than tripled on **OfficeHome**. Prompt learning strategies (CoOp, CoCoOp) seem to outperform fine-tuning the whole encoder, even with the addition of WiSE robust fine-tuning.

For industrial datasets, **CLIP ZS** performs only marginally better than random guessing, indicating poor text-image alignment for the images in **Rareplanes-tiles** and **Aerial Vehicles**. CLIP-based DG algorithms improve the performance of a randomly initialized network, but the gain is much lower than that observed on academic datasets. The former observation partially explains the latter, as each of the CLIP-based DG algorithms we tested relies heavily on the ZS model. **FT CLIP** involves fine-tuning the zero-shot model, **WiSE** ensembles the fine-tuned model and the ZS model, and prompt learning strategies build a classifier using textual descriptions of the classes, similar to ZS.

	photo	**art painting**	**cartoon**	**sketch**
`photo`	**18.15**	17.53	**17.47**	18.39
`artistic painting`	14.81	**20.49**	16.25	18.67
`cartoon`	12.46	14.83	16.89	19.17
`sketch`	12.98	15.80	14.51	**19.56**

Table 2. Mean cosine similarities (%) between images of PACS's domains and a textual description.

	real	**synthetic**
`aerial photo`	16.29	17.53
`aerial synthetic image`	**18.08**	**19.23**

Table 3. Mean cosine similarities (%) between images of Rareplanes-tiles and a textual description.

5. The challenge of describing atypical domains

Some domain generalization methods [3, 23] leverage a textual description of the target domain to modify the source images in order to bridge the domain gap. These methods require a textual description that specifically describes the target domain. It is relatively straightforward to find textual descriptions for the domains of academic benchmarks, where the name of the domain is often a sufficient description, possibly with a slight modification (**art painting** → **artistic painting**). However, crafting textual descriptions for **Rareplanes-tiles** and **Aerial Vehicles** is challenging.

We measure the capacity of a textual description $t_{\mathcal{D}}$ to represent a domain $\mathcal{D}$ by computing the mean cosine similarity in CLIP space between $t_{\mathcal{D}}$ and the images of $\mathcal{D}$:

$$Sim(t_{\mathcal{D}}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \frac{E_T(t_{\mathcal{D}}) \cdot E_V(x)}{\|E_T(t_{\mathcal{D}})\|_2 \|E_V(x)\|_2}$$

with E_V the visual encoder and E_T the textual encoder. Table 2 shows the similarity matrix between the domains of PACS and the textual descriptions corresponding to each PACS domain, computed using the textual and visual encoders of CLIP ResNet50. We observe that, in most cases, the textual description that best represents a domain is the one derived from the domain name. However, for **Rareplanes-tiles**, `aerial synthetic image` better represents images from both the real and synthetic domains (see Table 3). Such a textual description that describes both source and target domains is not a relevant information. We tried several variations of the word `synthetic` (`simulated`, `artificial`...) without finding a satisfactory textual description for the synthetic images. We make similar observations for **Aerial Vehicles**.

As a consequence, we cannot use the domain generalization methods that exploit a textual description of the target domain. This confirms that these type of images are in the blind spot of CLIP.

6. Conclusion and Future Work

This paper evaluated VLFMs-based Domain Generalization (DG) algorithms on two synthetic-to-real classification datasets, Rareplanes-tiles and Aerial Vehicles, designed to emulate industrial data. While VLFMs-based DG methods outperformed end-to-end training on academic benchmarks, their effectiveness was significantly limited on these industrial-like datasets. Further research is needed to leverage the knowledge in VLFMs for classifying atypical data not well-represented during their training.

References

[1] Wuyang Chen, Zhiding Yu, Shalini De Mello, Sifei Liu, Jose M. Alvarez, Zhangyang Wang, and Anima Anandkumar. Contrastive Syn-to-Real Generalization. *ICLR*, 2021. 2

[2] Tuhin Das, Robert-Jan Bruintjes, Attila Lengyel, Jan van Gemert, and Sara Beery. Domain Adaptation for Rare Classes Augmented with Synthetic Samples. *arXiv:2110.12216 [cs]*, 2021. 2

[3] Lisa Dunlap, Clara Mohri, Devin Guillory, Han Zhang, Trevor Darrell, Joseph E. Gonzalez, Aditi Raghunathan, and Anna Rohrbach. Using language to extend to unseen domains. In *ICLR*, 2023. 2, 4

[4] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. Pøda: Prompt-driven zero-shot domain adaptation. In *ICCV*, 2023. 2

[5] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M. Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. 2023. 1

[6] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM*, 2022. 2

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 2016. 1

[8] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021. 3

[9] Tae Soo Kim, Bohoon Shim, Michael Peven, Weichao Qiu, Alan Yuille, and Gregory D. Hager. Learning from Synthetic Vehicles. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 500–508, Waikoloa, HI, USA, 2022. IEEE. 1, 2

[10] Fanjie Kong, Bohao Huang, Kyle Bradbury, and Jordan M. Malof. The Synthinel-1 dataset: A collection of high resolution synthetic overhead imagery for building segmentation. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1803–1812, Snowmass Village, CO, USA, 2020. IEEE. 1, 2

[11] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 3

[12] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 1

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 4

[14] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 1, 2

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 4

[16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 2015. 1

[17] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. 2022. 1

[18] Viktor Seib, Benjamin Lange, and Stefan Wirtz. Mixing Real and Synthetic Data to Enhance Neural Network Training – A Review of Current Approaches. *arXiv:2007.08781 [cs]*, 2020. 2

[19] Jacob Shermeyer, Thomas Hossler, Adam Van Etten, Daniel Hogan, Ryan Lewis, and Daeil Kim. Rareplanes dataset, 2020. 3

[20] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016. 1

[21] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017. 3

[22] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. CLIP the gap: A single domain generalization approach for object detection. In *CVPR*, 2023. 2

[23] Zhenbin Wang, Lei Zhang, Lituan Wang, and Minjuan Zhu. Landa: Language-guided multi-source domain adaptation. *CoRR*, 2024. 2, 4

[24] Garrett Wilson and Diane J. Cook. A survey of unsupervised deep domain adaptation. *ACM*, 2020. 1

[25] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it

till you make it: Face analysis in the wild using synthetic data alone. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3661–3671, Montreal, QC, Canada, 2021. IEEE. 2

[26] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, 2022. 2

[27] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *CVPR*, 2022. 2, 4

[28] Yang Xu, Bohao Huang, Xiong Luo, Kyle Bradbury, and Jordan M. Malof. SIMPL: Generating synthetic overhead imagery to address zero-shot and few-shot detection problems. *CoRR*, abs/2106.15681, 2021. 1, 2

[29] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. 1

[30] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 2, 4

[31] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 2, 4