INDIGO: Intrinsic Multimodality for Domain Generalization

Puneet Mangla *† , Shivam Chandhok *‡ , Milan Aggarwal†, Vineeth N Balasubramanian◇, Balaji Krishnamurthy†
†Adobe Media and Data Science Research Lab, Noida ◇ Indian Institute of Technology Hyderabad ‡ INRIA, Universite Grenoble Alpes
{ai20resch11006, cs20mtech14006,vineethnb}@iith.ac.in, chandhokshivam@gmail.com

Abstract

For models to generalize under unseen domains (a.k.a domain generalization), it is crucial to learn feature representations that are domain-agnostic and capture the underlying semantics that makes up an object. Recent advances towards weakly supervised vision-language models (such as CLIP) have shown their ability on object understanding by capturing semantic characteristics that generalize under different domains. Hence, it becomes important to investigate how the semantic knowledge present in their representations can be effectively incorporated and utilized for domain generalization. Motivated from this, we study how semantic information from existing pre-trained multimodal networks can be leveraged in an "intrinsic" way to make systems generalize under unseen domains.We propose IntriNsic multimodality for DomaIn GeneralizatiOn (INDIGO), a simple and elegant framework that leverages the intrinsic modality present in pre-trained multimodal networks to enhance generalization to unseen domains at test-time. We experiment on different Domain Generalization benchmarks– DomainNet and Office Home and show state-of-the-art generalization performance on unseen domains. Further, we provide a thorough analysis to develop a holistic understanding of INDIGO.

1. Introduction

Domain Generalization (DG) [14, 37, 38] – a setting that aims to learn a model using data from source domains (for e.g. *clipart*, *painting*, *real world*) in order to generalize and predict effectively on an unseen domain (e.g. *sketch*) – has gained significant importance recently, to address this need. Besides the standard DG setting, the community has also seen concerted efforts towards defining new DG-related problem settings [21, 33] as well as leveraging any available data to generalize to unseen domains [5, 25]. Most previous approaches [9, 14, 16, 20, 22, 37, 38] that address/aim to tackle the DG problem use different learning paradigms and training strategies to learn domain-agnostic semantic features that represent an object category and can thus extend to unseen domain samples at test-time. Other methods [6, 30, 31] have also shown that leveraging domain-specific features along with domain-invariant information can further improve the model's generalization on unseen domains. More recently, vision transformers (ViTs) [7, 35] have demonstrated robustness to domain-shift [24, 39] which is desirable for making models generalize to unseen domains.

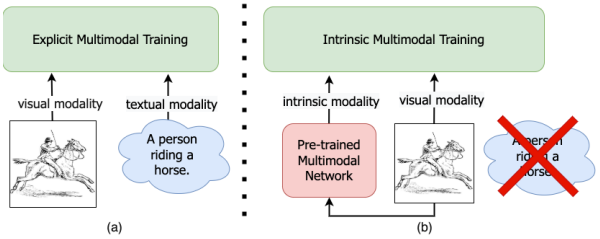

Figure 1. **Illustration of our broader idea.** In scenarios where we don't have access to explicit modalities like image captions for source domain data, we leverage the "intrinsic" modality present in pre-trained multimodal networks along with visual modality obtained from image.

An alternative strategy to address this task can be to look for other sources of information that can help disentangle domain-specific and domain-agnostic characteristics and thereby equip models with the ability to capture general domain-agnostic class-level cues. Recent progress towards weakly supervised vision-language models [17, 19, 27] have shown their abilities on semantic understanding and triggered the interest in using them for practical use in various settings. These methods are known to learning holistic object representations from cheap, weakly supervised noisy text annotations that capture class-level semantics of object categories such as shape/content [27]. Such representations can inherently capture object characteristics that generalize to unseen domains. We leverage this potential of vision-language models in this work.

Motivated from this, we study how the multimodal information in pre-trained multimodal networks [17, 19, 27] can be leveraged intrinsically to make systems robust to domain-shift and enhance generalization on unseen do-

*Equal contribution

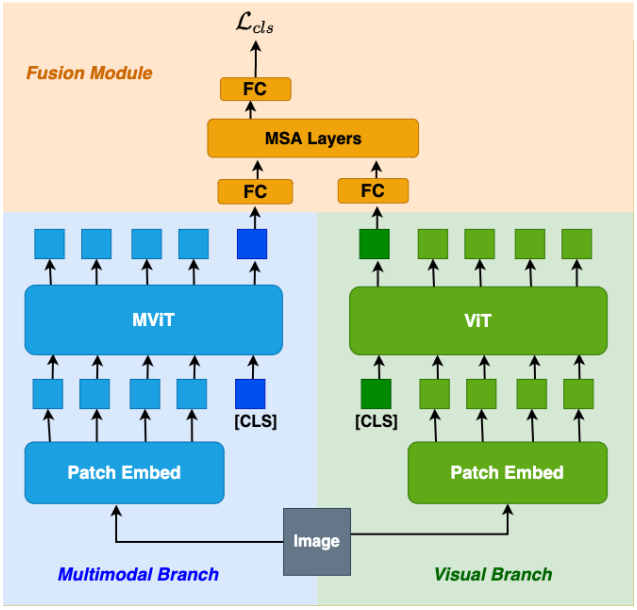

Figure 2. (Proposed approach) **INDIGO** consists of a multimodal branch comprised of pre-trained MViT to obtain intrinsic modality , a visual branch to extract visual modality and a fusion module to combine both.

mains. Specifically:

- We propose IntriNsic multimodality for DomaIn GeneralizatiOn (INDIGO), a simple and elegant way of leveraging the intrinsic modality present in pre-trained multimodal networks along with the visual modality in order to generalize better to unseen domains.
- We study the performance of recent popular vision-language models for domain generalization under different scenarios and show that fusing the visual modality from a trainable visual encoder (through our proposed attention based strategy) with intrinsic modality extracted from frozen pre-trained multimodal networks consistently offers better generalization performance.
- We perform comprehensive experiments on standard DG benchmarks - DomainNet and Office-Home and show that INDIGO achieves new state-of-the-art by outperforming prior SOTAs, conventional techniques like transfer learning, fine-tuning and zero-shot.

Figure 1 provides overview of the proposed idea. To the best of our knowledge, this is the first effort to study how multimodality can be leveraged intrinsically via pre-trained multimodal models to generalize better to unseen domains.

2. Proposed Methodology

Overall Framework. As depicted in Figure 2, there are three main components in our approach: (1) a *multimodal branch* which consists of a multimodal vision transformer (MViT) pre-trained on image-text pairs used to extract the intrinsic modality present in it; (2) a *visual branch*, which trains a vision transformer (ViT) to extract visual modality that will encode meaningful shape-biased concepts from the source domains, useful for generalization; and (3) a *fusion module* which combines best of both - intrinsic and visual modality through a multi-headed self-attention mechanism for final classification.

Multimodal branch. We leverage pre-trained large-scale vision-language networks like CLIP [27] that use a contrastive objective to push the embeddings of matched image-text pairs together and non-matched pairs apart. The pipeline generally consists of an image encoder $f^M(.)$ (in our case a ViT which we call MViT), a text encoder $g(.)$, and linear projection layers $h^I(.)$ and $h^T(.)$. The image and text features (obtained from their respective encoders) are projected to the same dimension, normalized, and then aligned using the following contrastive loss:

$$\mathbf{z}_i^I = \frac{h_I(f^M_{\texttt{CLS}}(\mathbf{x}_i))}{\|h_I(f^M_{\texttt{CLS}}(\mathbf{x}_i))\|_2}; \; \mathbf{z}_i^T = \frac{h_T(g(\mathbf{t}_i))}{\|h_T(g(\mathbf{t}_i))\|_2}$$

$$\mathcal{L}_I = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\mathbf{z}_i^I, \mathbf{z}_i^T)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{z}_i^I, \mathbf{z}_j^T)/\tau)}$$

$$\mathcal{L}_T = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\mathbf{z}_i^T, \mathbf{z}_i^I)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{z}_i^T, \mathbf{z}_j^I)/\tau)}$$

$$\mathcal{L}_{contrastive} = (\mathcal{L}_I + \mathcal{L}_T)/2$$

here $(\mathbf{x}_i, \mathbf{t}_i)$ denote the i^{th} image-text pair in a batch of size N. $f^M_{\texttt{CLS}}(.)$ represents the MViT's representation corresponding to the `CLS` token. The similarity function `sim(,)` is measured by dot product, and τ is a learnable temperature variable to scale the logits.

In scenarios where we do not have direct access to text annotations, we can assume that an image's unnormalized projected embedding $h^I(f^M_{\texttt{CLS}}(\mathbf{x}_i))$ would be weakly aligned with its hypothetical text description. This allows us to leverage the intrinsic modality present in a pre-trained multimodal vision transformer. Hence, we propose to use this unnormalized projected embedding $h^I(f^M_{\texttt{CLS}}(\mathbf{x}_i))$ as a "intrinsic" modality in our overall pipeline.

Visual branch. The visual branch is a sibling to the multimodal branch. We employ a trainable vision transformer, $f^V(.)$, to learn visual concepts from source domains that might be absent in MViT representations but are relevant to the task. These concepts can be dataset-, domain-, or even class-specific, which, when combined with the "intrinsic" modality, can help boost the overall performance on the given task. Moreover, by design, since ViTs are better than CNNs in recognizing object shapes [24,39], we believe their shape-biased representations $f^V_{\texttt{CLS}}(\mathbf{x})$ will further assist our overall pipeline in generalizing to unseen domains (as we show through our experiments).

Fusion module. The purpose of the fusion module is to fuse the "intrinsic" modality $h^I(f^M_{\texttt{CLS}}(\mathbf{x}))$ (obtained from the multimodal branch) and the visual modality $f^V_{\texttt{CLS}}(\mathbf{x})$ (obtained from the visual branch) to perform the final classification. We first project both of them to same space via linear projections $w^M(.)$ and $w^V(.)$ to obtain intrinsic modality $w^M(h^I(f^M_{\texttt{CLS}}(\mathbf{x})))$ and visual modality $w^V(f^V_{\texttt{CLS}}(\mathbf{x}))$ tokens. This is followed by a series of K multi-headed self-attention blocks ($\texttt{MSA}$) and feed-forward networks ($\texttt{FFN}$) to perform inter-modality attention on both tokens as follows

$$\begin{aligned} \mathbf{x}^M_0 = w^M(h^I(f^M_{\texttt{CLS}}(\mathbf{x}))); \ \mathbf{x}^V_0 = w^V(f^V_{\texttt{CLS}}(\mathbf{x})) \\ \mathbf{x}_0 = [\mathbf{x}^M_0 \ || \ \mathbf{x}^V_0] \\ \mathbf{o}_k = \mathbf{x}_{k-1} + \texttt{MSA}(\texttt{LN}(\mathbf{x}_{k-1})) \\ \mathbf{x}_k = \mathbf{o}_k + \texttt{FFN}(\texttt{LN}(\mathbf{o}_k)) \\ \mathbf{x}_K = [\mathbf{x}^M_K \ || \ \mathbf{x}^V_K] \end{aligned} \tag{1}$$

The attention mechanism allows the intrinsic modality token to attend with the visual modality token and incorporate any dataset, domain, or class-specific concepts present in it. Similarly, the visual modality token will interact with the intrinsic modality token to learn multimodal concepts present in it. This ensures that final representations leverage the best of both modalities. Finally, the transformed representation of intrinsic modality ($\mathbf{x}^M_K$) is passed through a linear layer $c^M(.)$ to get class predictions and minimize cross-entropy loss. In addition to this, we add a regularizer that also minimizes classification loss on the transformed representation of visual modality token (by passing $\mathbf{x}^V_K$ through another linear layer $c^V(.)$). Overall loss can be written as

$$\hat{y}_M = c^M(\mathbf{x}^M_K); \mathcal{L}_{cls} = \mathcal{L}_{CE}(\hat{y}_M, y)$$

3. Experiments and Analysis

We perform experiments on the following DG datasets - (1) DomainNet [26], a large scale dataset containing 586,575 examples from 345 classes and six domains (*clipart*, *infograph*, *painting*, *quickdraw*, *real*, *sketch*); and (2) Office-Home [36], containing 15,588 examples from 65 classes and four domains (*art*, *clipart*, *product*, *real*).

Baselines. We evaluate and compare four kinds of training pipelines - (1) CNNs, which include state-of-the-arts [2, 4, 28] that use a Resnet-50 backbone; (2) ViTs, which include DeiT-S [35] (considered equivalent to Resnet-50) backbone trained in AGG manner; (3) MViTs, which include conventional ways like zero-shot inference, transfer learning using linear layer (Linear Eval) and attention layers (Attention Eval) over frozen CLIP features; and (4) MViTs + ViTs, that include our proposed fusion, INDIGO (using a DeiT-S visual backbone).

Training and evaluation protocol. Following previous works [4, 10, 28], we consider each domain as the target domain and the rest domains as source domains for training.

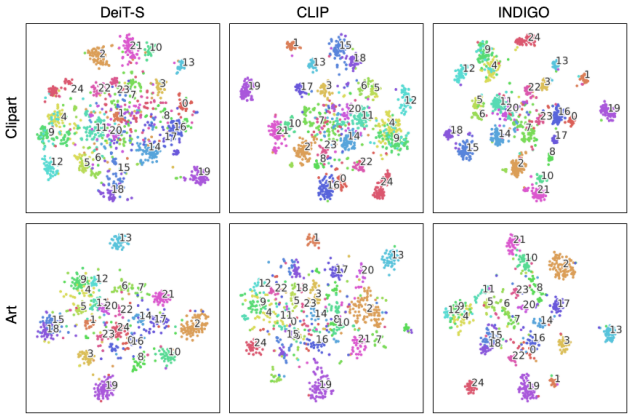

Figure 3. **t-SNE plots.** t-SNE visualization of learned feature representations by DeiT-S (standard AGG training), CLIP and our proposed INDIGO method when *clipart* and *art* are chosen as target domains for Office-Home dataset

We use test-domain validation (reporting best performance on test set) and training-domain validation model selection criteria (using a validation set) for DomainNet and Office-Home, respectively, as described in [10].

Results. Table 1 presents our results when CLIP-ViT-B/16 [27] is used as an MViT in the multimodal branch. As we can see, INDIGO achieves new state-of-the-art results by outperforming all the compared methods by good margins. In particular, on challenging domains like *quickdraw* where conventional ways of using MViTs perform worse than prior arts, INDIGO achieves the best performance by leveraging the best of both - intrinsic and the visual modality. Further, we can observe that ViTs trained with simple vanilla AGG loss easily beat state-of-the-art CNN-based approaches - SWAD [4], EoA [2]. This shows that their design offers shape-biased representations (compared to CNNs), which INDIGO leverages.

Choice of visual network and number of layers in fusion module. To highlight that INDIGO is leveraging the visual modality, we perform an ablation where we vary the strength of ViT used in the visual branch. Additionally, we also vary the number of layers used in the fusion module to show its effect on final performance. As shown in Table 2, by using more powerful (Hybrid ViTs) and large (ViT-B) vision transformers [7] in the visual branch, the domain generalization performance of INDIGO improves. This shows that INDIGO can attend to visual modality to learn additional shape-biased concepts, and the performance is not solely because of intrinsic modality. The gain in performance becomes prominent when more layers are used in the fusion module, implying a better inter-modality interaction between intrinsic and visual modality tokens.

Can fine-tuning MViT help further? In all our previous experiments, we used a frozen pre-trained multimodal

Type	Method	DomainNet						Office-Home				
		C	S	P	Q	I	Avg.	R	C	P	A	Avg.
CNNs	AGG	58.4	49.9	47.3	13.4	19.8	37.76	77.3	53.4	76.5	62.7	67.47
	IRM [1]	51.0	44.7	38.8	11.8	16.7	32.6	77.2	52.3	75.2	61.8	66.63
	DRO [29]	47.8	40.7	36.3	9.0	17.2	30.2	77.7	52.9	75.5	61.6	66.93
	Mixup [40]	55.8	49.2	46.2	12.8	19.2	36.64	79.2	54.7	77.3	64.7	68.98
	MLDG [15]	59.3	51.2	48.8	14.0	20.3	38.72	78.6	54.5	75.9	63.7	68.18
	CORAL [34]	58.8	50.8	47.5	13.6	20.8	38.3	77.9	55.3	76.7	64.4	68.58
	MMD [16]	54.6	47.5	44.9	12.6	19.6	35.84	78.1	53.7	76.1	63.0	67.73
	DANN [8]	53.8	46.7	43.5	11.8	17.5	34.66	76.6	51.7	74.1	59.3	65.43
	C-DANN [18]	53.4	46.5	44.7	12.9	18.4	35.18	76.0	51.1	74.1	61.0	65.55
	EoA [2]	65.9	57.1	55.3	16.5	23.4	43.64	81.5	59.8	79.5	69.1	72.48
	SelfReg [12]	62.4	53.7	51.7	14.7	22.5	41.0	78.8	55.4	78.4	64.9	69.37
	SagNet [23]	57.5	49.5	46.3	13.5	19.2	37.2	78.3	54.8	75.8	63.4	68.08
	ARM [41]	49.6	43.9	41.5	10.8	16.5	32.46	75.2	51.0	74.1	58.9	64.8
	V-REx [13]	43.3	37.7	32.5	9.8	14.1	27.48	76.6	53.0	75.3	60.7	66.4
	MTL [3]	58.0	49.0	46.2	12.7	19.2	37.02	76.8	52.4	74.9	61.5	66.4
	SAND [32]	43.8	39.9	38.2	9.0	15.2	29.22	76.2	53.3	73.5	60.3	65.82
	RSC [11]	55.5	47.8	44.4	12.5	18.3	35.7	75.1	51.4	74.8	60.7	65.50
	Fishr [28]	58.3	50.5	47.9	13.6	20.2	38.1	78.3	54.4	76.2	62.4	67.83
	SWAD [4]	66.0	55.5	53.5	16.1	22.4	42.7	80.2	57.7	78.4	66.1	70.6
ViTs	AGG	69.14	54.25	58.15	14.83	27.55	44.78	84.64	60.10	84.43	74.2	75.84
MViTs	Zero-Shot	67.8	61.79	64.13	13.9	45.7	50.66	84.7	60.8	83.37	78.9	76.94
	Linear Eval	63.2	59.37	57.36	10.34	41.7	46.39	82.51	66.66	81.22	72.86	75.81
	Attention Eval	75.3	64.68	64.33	16.30	44.23	52.97	88.14	69.00	88.99	77.53	80.92
MViTs + ViTs	INDIGO	**76.9**	**65.65**	**66.42**	**17.4**	**46.32**	**54.54**	**89.38**	**73.31**	**90.78**	**79.92**	**83.35**

Table 1. **ClosedDG results.** Performance of INDIGO on DomainNet (**C**: clipart, **S**: sketch, **P**: painting, **Q**: quickdraw, **I**: infograph) and Office-Home (**R**: real world, **C**: clipart, **P**: product, **A**: art) datasets under closed setting. We highlight the **best results** and the second best results. The results are averaged over five runs. INDIGO achieves new state-of-the-art by outperforming all compared methods by good margins.

Backbone	3 Layers					12 Layers				
	R	C	P	A	Avg.	R	C	P	A	Avg.
Resnet-50	88.8	72.91	90.1	79.34	82.78	89.0	72.48	90.2	78.2	82.47
DeiT-Ti	89.02	72.88	90.14	80.05	83.02	89.34	73.52	90.43	79.31	83.15
Hybrid-ViT-Ti	89.1	72.77	90.3	79.94	83.02	89.7	73.82	90.50	79.9	83.48
DeiT-S	89.38	73.31	90.78	79.92	83.35	90.1	74.32	90.99	80.2	83.90
Hybrid-ViT-S	89.73	73.71	91.05	81.16	83.91	90.88	75.45	91.22	81.62	84.80
ViT-B	91.4	74.23	91.84	82.33	84.95	91.76	75.85	92.13	83.51	85.81

Table 2. **Ablation on choice of visual network and number of layers in fusion module.** Performance of INDIGO when different networks are used in visual branch and layers of fusion module are increased on Office-Home (**R**: real world, **C**: clipart, **P**: product, **A**: art) under closed setting. The results are averaged over five runs. Stronger and Larger ViTs can be seen to further improve the generalization of INDIGO to unseen domains.

Backbone	Closed OfficeHome				
	R	C	P	A	Avg.
CLIP (FT B.N Layers)	84.00	68.92	84.13	76.47	78.38
INDIGO (FT B.N Layers)	89.50	73.50	91.02	80.2	83.55
CLIP (FT Last Layers)	89.6	74.5	89.70	83.43	84.30
INDIGO (FT Last Layer)	90.83	76.87	91.64	83.91	85.81
CLIP (All Layers)	88.71	73.15	88.78	81.37	83.00
INDIGO (All Layers)	89.51	75.42	90.34	82.73	84.5

Table 3. **Ablation on finetuning the MViT.** Performance of INDIGO when MViT is also finetuned on Office-Home (**R**: real world, **C**: clipart, **P**: product, **A**: art) under closed setting. We use training-domain validation set model selection criteria. The results are averaged over five runs.

network. As an additional experiment, along with training the visual branch and fusion module, we also finetune the multimodal network (i.e CLIP) . We finetune in two ways - (1) only normalization layers; and (2) last layer. Table 3 shows that the performance of INDIGO further improves while still outperforming standalone finetuning of CLIP.

t-SNE plots. We analyze and compare the representations learned by INDIGO with DeiT-S [35] and CLIP [27] on target domain (for 25 classes of Office-Home) via t-SNE plots in Figure 3. As can be seen, for INDIGO, the plot is less noisy and well segregated into class clusters as compared to DeiT-S and CLIP, resulting in state-of-the-art generalization on these target domains.

Conclusions and Future Work In this work, we study how multimodal information present in pre-trained vision-language models can be leveraged "intrinsically" to build systems that generalize to unseen domains. We propose INDIGO, a simple and elegant way to combine the intrinsic and visual modalities obtained from pre-trained multimodal network and vision transformer (ViT), respectively. Our future work will include the development of better methods to effectively fuse both modalities to improve generalization performance in unseen domains further. We also plan to extend same in other challenging settings like segmentation, and visual grounding.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. 2019. cite arxiv:1907.02893. 4

[2] Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *CoRR*, abs/2110.10832, 2021. 3, 4

[3] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22(2):1–55, 2021. 4

[4] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3, 4

[5] Shivam Chandhok, Sanath Narayan, Hisham Cholakkal, Rao Muhammad Anwer, Vineeth N. Balasubramanian, Fahad Shahbaz Khan, and Ling Shao. Structured latent embeddings for recognizing unseen classes in unseen domains. *British Machine Vision Conference*, abs/2107.05622, 2021. 1

[6] Prithvijit Chattopadhyay, Y. Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. volume abs/2008.12839, 2020. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 3

[8] 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

[9] Muhammad Ghifary, W. Kleijn, M. Zhang, and D. Balduzzi. Domain generalization for object recognition with multi-task autoencoders. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2551–2559, 2015. 1

[10] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. 3

[11] Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, 2020. 4

[12] 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

[13] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR, 18–24 Jul 2021. 4

[14] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5543–5551, 2017. 1

[15] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 4

[16] Haoliang Li, Sinno Jialin Pan, S. Wang, and A. Kot. Domain generalization with adversarial feature learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. 1, 4

[17] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 1

[18] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. 4

[19] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, 2021. 1

[20] Y. Li, X. Tian, Mingming Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018. 1

[21] Puneet Mangla, Shivam Chandhok, Vineeth N. Balasubramanian, and Fahad Shahbaz Khan. Cocoa: Context-conditional adaptation for recognizing unseen classes in unseen domains. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1618–1627, 2022. 1

[22] Krikamol Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. *ArXiv*, abs/1301.2115, 2013. 1

[23] 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*, 2021. 4

[24] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1, 2

[25] Prashant Pandey, Mrigank Raman, Sumanth Varambally, and A. P. Prathosh. Generalization on unseen domains via inference-time label-preserving target projections. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12919–12928, 2021. 1

[26] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019. 3

[27] 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, 3, 4

[28] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. *arXiv preprint arXiv:2109.02934*, 2021. 3, 4

[29] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 4

[30] Mattia Segu, A. Tonioni, and Federico Tombari. Batch normalization embeddings for deep domain generalization. *ArXiv*, abs/2011.12672, 2020. 1

[31] Seonguk Seo, Yumin Suh, D. Kim, Jongwoo Han, and B. Han. Learning to optimize domain specific normalization for domain generalization. 2020. 1

[32] Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. *CoRR*, abs/2106.02266, 2021. 4

[33] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9624–9633, 2021. 1

[34] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV 2016 Workshops*, 2016. 4

[35] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers and distillation through attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021. 1, 3, 4

[36] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 3

[37] Zheng Xu, W. Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *ECCV*, 2014. 1

[38] P. Yang and Wei Gao. Multi-view discriminant transfer learning. In *IJCAI*, 2013. 1

[39] Chongzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang feng Zhou, Zhongang Cai, Haiyu Zhao, Shuai Yi, Xianglong Liu, and Ziwei Liu. Delving deep into the generalization of vision transformers under distribution shifts. *ArXiv*, abs/2106.07617, 2021. 1, 2

[40] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 4

[41] M. Zhang, H. Marklund, N. Dhawan, A. Gupta, S. Levine, and C. Finn. Adaptive risk minimization: Learning to adapt to domain shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 4