Delving Deep into the Generalization of Vision Transformers under Distribution Shifts

Chongzhi Zhang[1,*], Mingyuan Zhang[2,*], Shanghang Zhang[3,*], Daisheng Jin[1], Qiang Zhou[4], Zhongang Cai[2,5], Haiyu Zhao[2,5], Xianglong Liu[1], Ziwei Liu[2,✉]
[1]Beihang University [2]S-Lab, Nanyang Technological University
[3]Peking University [4]AIR, Tsinghua University [5]Shanghai AI Laboratory

Abstract

Vision Transformers (ViTs) have achieved impressive performance on various vision tasks, yet their generalization under distribution shifts (DS) is rarely understood. In this work, we comprehensively study the out-of-distribution (OOD) generalization of ViTs. For systematic investigation, we first present a taxonomy of DS. We then perform extensive evaluations of ViT variants under different DS and compare their generalization with Convolutional Neural Network (CNN) models. Important observations are obtained: ***1)*** *ViTs learn weaker biases on backgrounds and textures, while they are equipped with stronger inductive biases towards shapes and structures, which is more consistent with human cognitive traits. Therefore, ViTs generalize better than CNNs under DS. With the same or less amount of parameters, ViTs are ahead of corresponding CNNs by more than 5% in top-1 accuracy under most types of DS.* ***2)*** *As the model scale increases, ViTs strengthen these biases and thus gradually narrow the in-distribution and OOD performance gap. To further improve the generalization of ViTs, we design the Generalization-Enhanced ViTs (GE-ViTs) from the perspectives of adversarial learning, information theory, and self-supervised learning. By comprehensively investigating these GE-ViTs and comparing with their corresponding CNN models, we observe:* ***1)*** *For the enhanced model, larger ViTs still benefit more for the OOD generalization.* ***2)*** *GE-ViTs are more sensitive to the hyper-parameters than their corresponding CNN models. We design a smoother learning strategy to achieve a stable training process and obtain performance improvements on OOD data by 4% from vanilla ViTs. We hope our comprehensive study could shed light on the design of more generalizable learning architectures. Codes and datasets are released in https://github.com/Phoenix1153/ViT_OOD_generalization.*

*These authors contributed equally to this work.
✉Corresponding author.

1. Introduction

Recently, transformer has made remarkable achievements in vision tasks, such as *e.g.* image classification [7, 8, 27], object detection [4, 36], and image processing [6]. Despite the encouraging performance achieved on standard benchmarks and several properties revealed in recent works [1, 5, 20, 21], the generalization ability of Vision Transformers (ViTs) is still less understood. While the traditional train-test scenario assumes the test data for model evaluation are independent identically distributed (IID) with sampled training data, this assumption does not always hold in real-world scenarios. Thus, out-of-distribution (OOD) generalization is a highly desirable capability of machine learning models. Recent works indicate current CNN architectures generalize poorly on various distribution shifts (DS) [11, 13, 14], whereas the investigation on ViTs remains scarce. Therefore, in this paper, we mainly focus on delving deep into the OOD generalization of ViTs under DS.

To comprehensively study the OOD generalization ability of ViTs, we first define a categorization of commonly appearing DS based on the modified semantic concepts in images. Generally, an image for classification contains a foreground object and background information. The foreground object consists of hierarchical semantic concepts including pixel-level elements, object textures, and shapes, object parts, and object itself [35]. A distribution shift usually causes variance on one or more semantics and we thus present a taxonomy of DS into four conceptual groups: background shifts, corruption shifts, texture shifts, and style shifts.

With the taxonomy of DS, we investigate the OOD generalization of ViTs by comparison with CNNs in each case. While models are desired to generalize to arbitrary OOD scenarios, the no-free-lunch theorem for machine learning [3, 12, 32] demonstrates that there is no entirely general-purpose learning algorithm, and that any learning algorithm implicitly or explicitly will generalize better on some distributions and worse on others. Thus some set of inductive biases are demanded to acquire generalization. Hence,

to achieve human-level generalization capability, machine learning models are supposed to have inductive biases that are most relevant to the human prior in the world. There have been many attempts to inject inductive biases into deep learning models that humans may exploit for the cognition operating at the level of conscious processing, *e.g.* the convolution [16] and self-attention mechanism [29]. Therefore, we examine whether transformers are equipped with inductive biases that are more related to human cognitive traits to better investigate the generalization properties of ViTs under DS. Extensive evaluations reveal the following observations on the OOD generalizations of ViTs: **1)** ViTs learn weaker biases on backgrounds and textures, while they are equipped with stronger inductive biases towards shapes and structures, which is more consistent with human cognitive traits. Therefore, ViTs generalize better than CNNs in most cases. Specifically, ViT not only achieves better performance on OOD data but also has smaller generalization gaps between IID and OOD datasets. **2)** As the model scale increases, ViTs strengthen these biases and thus gradually narrow the IID and OOD generalization gaps, especially in the case of corruption shifts and background shifts. In other words, larger ViTs are better at diminishing the effect of local changes. **3)** ViTs trained with larger patch size deal with texture shifts better, yet are inferior in other cases.

After validating the superiority of ViTs in dealing with OOD data, we focus on further improving their generalization capacity. Specifically, we design Generalization-Enhanced ViTs (GE-ViTs) from the perspectives of adversarial training [10], information theory [24] and self-supervised learning [34]. Equipped with GE-ViTs, we achieve significant performance boosts towards OOD data by 4% from vanilla ViTs. By performing an in-depth investigation on different models, we draw the following conclusions: **1)** For the enhanced transformer models, larger ViTs still benefit more for the OOD generalization. **2)** GE-ViTs are more sensitive to the hyper-parameters than their corresponding CNN models.

2. Related Work

Vision Transformers. Recently, Transformers have been applied to various vision tasks including image classification [7, 8, 27], object detection [4, 36], segmentation [31] and image processing [6]. Among them, the Vision Transformer (ViT) [8] is the first fully-transformer model applied for image classification and competitive with state-of-the-art CNNs. It heavily relies on large-scale datasets for model pre-training, requiring huge computation resources. Later, [27] propose the Data-efficient image Transformer (DeiT), which achieves competitive results against the state-of-the-art CNNs on ImageNet without external data by simply changing training strategies from ViT. Due to its efficiency, we use this family of models to investigate generalizations of Vision Transformers in this paper.

Table 1. **Illustration of our taxonomy of DS.** We build the taxonomy upon what kinds of semantic concepts are modified from the original image and divide the DS into four cases: background shifts, corruption shifts, texture shifts, and style shifts. ✓ denotes the unmodified vision cues under certain type of DS.

Shift Type	background	foreground			
		pixel	texture	shape	structure
Background Shift		✓	✓	✓	✓
Corruption Shift			✓	✓	✓
Texture Shift				✓	✓
Style Shift					✓

Out-of-distribution Generalization. Attracting much attention recently, various works have been proposed for OOD generalization under different settings. Most domain adaptation literatures aim at promoting the model's performance under distribution shift with access to the unlabeled target data [10,19,26]. Another setting for OOD generalization concentrates on learning representations without access to target data, commonly referred as domain generalization [9,17,30]. In addition, some recent works model OOD generalization on their newly-built benchmarks [11, 13, 14]. Though recent works [1, 5, 20, 21] have studied several properties of ViTs, the generalization of ViTs is still under explored.

3. Distribution Shifts and Evaluation Protocols

3.1. Taxonomy of Distribution Shifts

To make an extensive study on OOD generalization, we build the taxonomy of DS upon what kinds of semantic concepts are modified from the original image. Therefore, we divide the DS into four cases: background shifts, corruption shifts, texture shifts and style shifts, as shown in Tab. 1. The elaborately divided DS permit us to investigate model biases towards every visual cue respectively.

- **Background Shifts.** Image backgrounds are usually regarded as auxiliary cues in assigning images to corresponding labels in the image classification task. However, previous works have demonstrated that backgrounds may dominate in prediction [2, 23], which is undesirable to us. We focus on the model's invariance towards background change and thus define the background shifts. *ImageNet-9* [33] is adopted for background shifts.
- **Corruption Shifts.** The concept of corruption was proposed in [14], which stands for those naturally occurring vicinal impurities mixed in images. These corruptions either come from environmental influence during the shooting stage or from the image processing stage. We define these cases as corruption shifts, which only impact on object pixel-level elements while can still cause models obvious performance decrease. *ImageNet-C* [14] is used to examine generalization ability under corruption shifts.

- **Texture Shifts.** Generally, the texture gives us information about the spatial arrangement of the colors or intensities in an image, which is critical for the classifiers in obtaining a correct prediction. Thus, a replacement of object textures can influence model prediction. We define these variations as texture shifts. *Cue Conflict Stimuli* and *Stylized-ImageNet* [11] are used to investigate generalization under texture shifts.
- **Style Shifts.** Typically, style is a complicated concept determined by the characteristics that describe the artwork, such as the form, color, composition, etc. The variance of style often reflects in multiple concept levels, including texture, shape and object part, etc. *ImageNet-R* [13] and *DomainNet* [22] are used for the case of style shifts.

3.2. Model Zoo

- **Vision Transformer.** We follow the implementation in DeiT [27] and choose a range of models with different scales for experiments. The ViT architecture takes as input a grid of non-overlapping contiguous image patches of resolution $N \times N$. In this paper we typically use $N = 16$ ("/16") or $N = 32$ ("/32"). Besides the official DeiT models, we also utilize the data-efficient training scheme to train ViT-L/16 and ViT-B/32 and rename them DeiT-L/16 and DeiT-B/32.
- **Big Transfer.** Big Transfer models [15] are build on ResNet-V2 models. We select BiT-S-R50X1 based on a ResNet-50 backbone. Besides the official implementation, we also train a version using the identical data augmentation strategy from DeiTs for comparison. We respectively name them BiT and BiT$_{da}$.

3.3. Evaluation Protocols

In image classification tasks, a model generally consists of a feature encoder F and a classifier C. Suppose the model is trained on a training set $\mathscr{D}_{train} = \{(\mathrm{x}_i, y_i)\}_{i=1}^{N_{train}}$. We respectively introduce a set of independent identically distributed (IID) validation data $\mathscr{D}_{iid} = \{(\mathrm{x}_i, y_i)\}_{i=1}^{N_{iid}}$ and a set of out-of-distributed (OOD) data $\mathscr{D}_{ood} = \{(\mathrm{x}_i, y_i)\}_{i=1}^{N_{ood}}$ in the same semantic space. $N_{train}, N_{iid}, N_{ood}$ represent the number of data in $\mathscr{D}_{train}, \mathscr{D}_{iid}, \mathscr{D}_{ood}$ respectively. Then we use the following evaluations.

- **Accuracy on OOD Data.** A direct measurement is to calculate the accuracy on the OOD dataset:

$$Acc(F, C; \mathscr{D}_{ood}) = \frac{1}{|\mathscr{D}_{ood}|} \sum_{(\mathrm{x},y)\in\mathscr{D}_{ood}} \mathbb{1}(C(F(\mathrm{x})) = y), \tag{1}$$

  where $\mathbb{1}$ is the indicator function.
- **IID/OOD Generalization Gap.** In this paper, we also focus on how well a model could behave towards the OOD data compared with the IID data. Hence, we use the IID/OOD generalization gap to measure the performance difference caused by the distribution shift:

$$Gap(F, C; \mathscr{D}_{iid}, \mathscr{D}_{ood}) = Acc(F, C; \mathscr{D}_{iid}) - Acc(F, C; \mathscr{D}_{ood}). \tag{2}$$

4. Generalization-Enhanced ViTs

After investigating the OOD generalization properties of ViTs, it is natural to figure out strategies to further improve them. Thus we further design Generalization-Enhanced ViTs (GE-ViTs) from the perspectives of adversarial training [10], information theory [24] and self-supervised learning [34], named as T-ADV, T-MME, and T-SSL respectively. By making a full comparison of these three designs, we figure out the most suitable strategy for GE-ViTs.

4.1. Adversarial Learning

To learn domain-invariant representations, we introduce a domain discriminator [10] to promote the backbone to produce domain-confused features by adversarial training. Specifically, as shown in Fig 1 (a), the network consists of a shared feature encoder F, a label predictor C, and a domain classifier D. The feature encoder aims at minimizing the domain confusion loss $\mathcal{L}_{\mathrm{ADV}}$ for all samples and label prediction loss $\mathcal{L}_{\mathrm{CLS}}$ for labeled source samples while the domain classifier focus on maximizing the domain confusion loss $\mathcal{L}_{\mathrm{ADV}}$. The overall objectives are:

$$\mathcal{L}_{\mathrm{CLS}} = \sum_{(\mathrm{x},y)\in\mathscr{D}_s} \mathcal{H}(\sigma(C(F(\mathrm{x}))), y), \tag{3}$$

$$\mathcal{L}_{\mathrm{ADV}} = \sum_{(\mathrm{x},y_d)\in\mathscr{D}_s,\mathscr{D}_t} \mathcal{H}(\sigma(D(F(\mathrm{x}))), y_d), \tag{4}$$

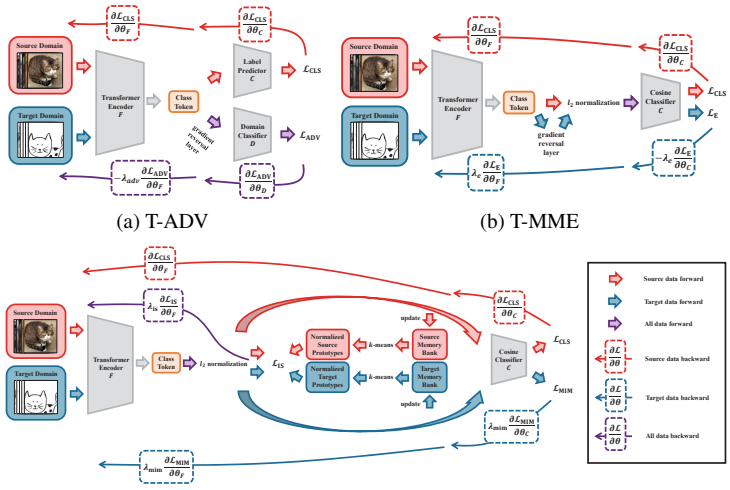
$$(\hat{\theta}_F, \hat{\theta}_C) = \arg\min_{\theta_F,\theta_C} \mathcal{L}_{\mathrm{CLS}} + \lambda_{adv}\mathcal{L}_{\mathrm{ADV}}, \tag{5}$$

$$\hat{\theta}_D = \arg\max_{\theta_D} \mathcal{L}_{\mathrm{ADV}}, \tag{6}$$

where y and y_d denote the class label and binary domain label respectively. $\sigma(\cdot)$ stands for the Softmax function and $\mathcal{H}(\cdot,\cdot)$ returns the cross-entropy of two input distributions. λ_{adv} is an adaptive coefficient that gradually changed from 0 to 1 by the schedule proposed in [10]. Furthermore, to facilitate training, a gradient reversal layer (GRL) is applied to implement the opposite objective of two parts.

4.2. Minimax Entropy

We leverage the minimax process on the conditional entropy of target data [24] to reduce the distribution gap while learning discriminative features for the task. As the pipeline is shown in Fig. 1 (b), a cosine similarity-based



(a) T-ADV (b) T-MME

(c) T-SSL

Figure 1. **A framework overview of the three designed generalization-enhanced ViTs.** All networks use a ViT F as feature encoder and a label prediction head C. Under this setting, the inputs to the models have labeled source examples and unlabeled target examples. **a) T-ADV** promotes the network to learn domain-invariant representations by introducing a domain classifier D for domain adversarial training. **b) T-MME** leverage the minimax process on the conditional entropy of target data to reduce the distribution gap while learning discriminative features for the task. The network uses a cosine similarity-based classifier architecture C to produce class prototypes. **c) T-SSL** is an end-to-end prototype-based self-supervised learning framework. The architecture uses two memory banks V^s and V^t to calculate cluster centroids. A cosine classifier C is used for classification in this framework.

classifier architecture C is exploited to produce class prototypes. The cosine classifier C consists of weight vectors $\mathrm{W} = [\mathrm{w}_1, ..., \mathrm{w}_{n_c}]$, where n_c denotes the total number of classes, and a temperature T. C takes ℓ_2 normalized $\frac{F(\mathrm{x})}{\|F(\mathrm{x})\|}$ as an input and output $\frac{1}{T}\frac{\mathrm{W}^\mathrm{T}F(\mathrm{x})}{\|F(\mathrm{x})\|}$. The key idea is to minimize the distance between the class prototypes and neighboring unlabeled target samples, thus extracting discriminative target features. To overcome the dominant impact of labeled source data on prototypes, prototypes are moved towards the target by maximizing the entropy $\mathcal{L}_\mathrm{E}$ of unlabeled target examples. Meanwhile, the feature extractor aims at minimizing the entropy of the unlabeled examples, to make them better clustered around the prototypes. Therefore, a minimax process is formulated between the weight vectors and the feature extractor. Additionally, the label prediction loss $\mathcal{L}_\mathrm{CLS}$ is also utilized on source samples. The overall objectives are:

$$\mathcal{L}_\mathrm{CLS} = \sum_{(\mathrm{x},y)\in\mathscr{D}_s} \mathcal{H}(\sigma(C(F(\mathrm{x}))), y), \tag{7}$$

$$\mathcal{L}_\mathrm{E} = \sum_{\mathrm{x}\in\mathscr{D}_t} \mathcal{H}(\sigma(C(F(\mathrm{x})))), \tag{8}$$

$$\hat{\theta}_F = \arg\min_{\theta_F} \mathcal{L}_\mathrm{CLS} + \lambda_e \mathcal{L}_\mathrm{E}, \tag{9}$$

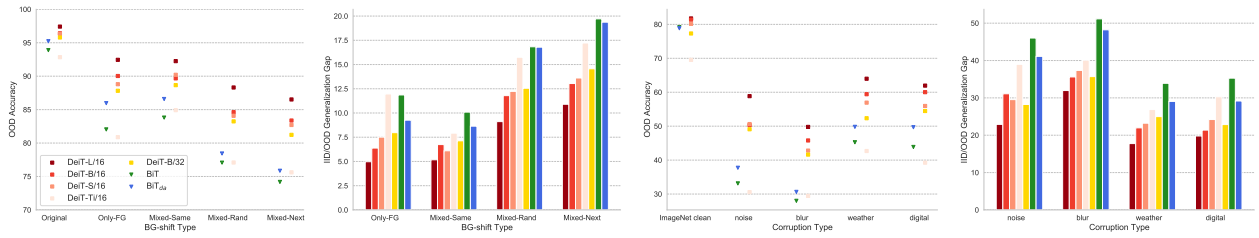
$$\hat{\theta}_C = \arg\min_{\theta_C} \mathcal{L}_\mathrm{CLS} - \lambda_e \mathcal{L}_\mathrm{E}, \tag{10}$$

where $\mathcal{H}(\cdot,\cdot)$ returns the cross-entropy of two input distributions and $\mathcal{H}(\cdot)$ returns the entropy. λ_e is a coefficient to balance two loss terms.

4.3. Self-Supervised Learning

We integrate an end-to-end prototypical self-supervised learning framework [34] into ViT. As shown in Fig. 1 (c), the framework also uses a cosine classifier C as introduced in Sec. 4.2. It first encodes semantic structure of data into the embedding space. ProtoNCE [18] is respectively applied in source and target domains. Specifically, two memory banks V^s and V^t are maintained to store feature vectors of every sample from source and target. These vectors are updated with momentum after each batch. k-means clustering is performed on memory banks to generate normalized prototypes $\{\mu_j^s\}_{j=1}^k$ and $\{\mu_j^t\}_{j=1}^k$. Then the similarity distribution vector between ℓ_2 normalized source feature vectors $f_i^s = \frac{F(\mathrm{x}_i^s)}{\|F(\mathrm{x}_i^s)\|}$ from current batch and normalized source prototypes $\{\mu_j^s\}_{j=1}^k$ as $P_i^s = [P_{i,1}^s, ..., P_{i,k}^s]$ with $P_{i,j}^s = \frac{\exp(\mu_j^s \cdot f_i^s/\phi)}{\sum_{r=1}^k \exp(\mu_r^s \cdot f_i^s/\phi)}$, where ϕ is a temperature value. Then the in-domain prototypical self-supervision loss is formed as: $\mathcal{L}_\mathrm{IS} = \sum_{i=1}^{|\mathscr{D}_s|} \mathcal{H}(P_i^s, c_s(i)) + \sum_{i=1}^{|\mathscr{D}_t|} \mathcal{H}(P_i^t, c_t(i))$, where $c_s(\cdot)$ and $c_t(\cdot)$ return the cluster index of the sample, and $|\cdot|$ returns the cardinal of the set. $\mathcal{H}(\cdot,\cdot)$ returns the cross-entropy of two input distributions.

In addition, since a network is desired to have high-confident and diversified predictions, an objective is set for maximizing the mutual information between the input image and the network prediction. This objective is split into two terms: entropy maximization of expected network prediction and entropy minimization on the network output. Therefore, the objective is formulated as: $\mathcal{L}_\mathrm{MIM} = \mathbb{E}_\mathrm{x}[\mathcal{H}(p(y|\mathrm{x};\theta)] -$



(a) (b) (c) (d)

Figure 2. **Results on ImageNet-9 and ImageNet-C.** (a)-(b) and (c)-(d) respectively illustrate the OOD Accuracy and IID/OOD Generalization Gap for different models on ImageNet-9 and ImageNet-C datasets. From (a) and (b), we conclude that **1)** ViTs perform with a weaker background-bias than CNNs, **2)** a larger ViT extracts a more background-irrelevant representation. From (c) and (d), we draw the conclusions that **1)** ViTs deal with corruption shifts better than CNNs and generalize better along with model size scaling up, **2)** ViTs do benefit from diverse augmentation in enhancing generalization towards vicinal impurities, but their architectural advantage cannot be overlooked as well, **3)** patch size for training has little influence on ViTs' generalization ability.

$\mathcal{H}(\mathbb{E}_{\mathrm{x}\in\mathscr{D}_s\cup\mathscr{D}_t}[p(y|\mathrm{x};\theta])$. The last term of training objective is the supervision loss on source domain measured by cross-entropy: $\mathcal{L}_{\mathrm{CLS}} = \sum_{(\mathrm{x},y)\in\mathscr{D}_s} \mathcal{H}(\sigma(C(F(\mathrm{x}))), y)$.

Finally, the overall learning objective is formulated as:

$$(\hat{\theta}_F, \hat{\theta}_C) = \arg\min_{\theta_F,\theta_C} \mathcal{L}_{\mathrm{CLS}} + \lambda_{\mathrm{is}}\mathcal{L}_{\mathrm{IS}} + \lambda_{\mathrm{mim}}\mathcal{L}_{\mathrm{MIM}}, \quad (11)$$

where λ_{is} and λ_{mim} denotes the coefficients of corresponding loss terms.

5. Systematic Study on ViTs Generalization

In-Distribution Generalization. We first examine the in-distribution generalization of different models on the ImageNet benchmark. As results are shown in Fig. 2 (c) column 1, we have the following observations. **1)** With the data-efficient training scheme, DeiT models tend to perform better as scales increase from *tiny* to *large*, but the gain of scale growth gradually dwindles. **2)** Having almost the same parameters and both trained without external data, DeiT-S/16 could beat BiT and BiT$_{da}$.

5.1. Background Shifts Generalization Analysis

We utilize ImageNet-9, a variety of foreground-background recombination plans, to investigate model bias towards background signal. These datasets empower us to investigate to what extent model decisions rely on the background signal. The OOD accuracy and IID/OOD gap results of four varieties of background shifts are illustrated in Fig. 2 (a) and (b) respectively.

- ViTs perform with a weaker background-bias than CNNs. By calculating accuracy gaps between *Mixed-Same* with class-relevant backgrounds and *Mixed-Rand* with neutral background signals, we can measure classifiers' reliance on the correct background. From Fig. 2 (a), the lower gaps achieved by ViTs indicate that ViTs depend less on corresponding background signals when the correct foreground is present. Likewise, it can be concluded that ViTs are less misled by the conflict background based on the accuracy gaps between *Mixed-Same* and *Mixed-Next*. In addition, comparing two BiT models, BiT$_{da}$ outperforms normal BiT in OOD accuracy and achieves lower IID/OOD Gaps, indicating that diverse augmentation during training exerts a salutary effect on model generalization on background-shifted data. Nonetheless, it is worth noticing that BiT$_{da}$ obtain a larger *Same-Rand* gap and *Same-Next* gap, which demonstrates that the augmentation training scheme cannot alleviate the model's dependence on correct background information. Therefore, ViTs perform with a weaker background bias than CNNs, and such property is brought by their architectures.

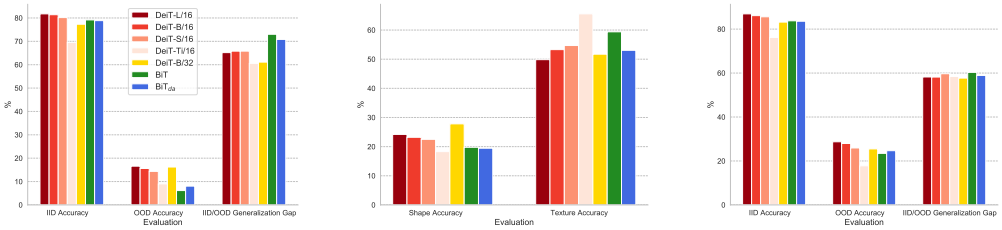
- A larger ViT extracts a more background-irrelevant representation. Via comparing ViTs of different sizes, we can observe that a larger ViT architecture contributes to a better OOD performance as well as a smaller IID/OOD gap. Even DeiT-L/16 could further narrow the gap by about 2% from DeiT-B/16, while they achieve almost the same in distribution accuracy results. Meanwhile, a larger ViT also achieves a lower *Same-Rand* gap and *Same-Next* gap, showing that there exists a positive correlation between the ViT scale and their ability to exclude distraction provided by irrelevant or conflict backgrounds. Hence, it is clear that a larger ViT tends to focus more attention on the foreground and learn a more background-irrelevant representation.

5.2. Corruption Shifts Generalization Analysis

The corruption results of 4 categories averaged over all subclasses and all severities, are shown in Fig. 2 (c) and (d).

- ViTs deal with corruption shifts better than CNNs and generalize better along with model size scaling up. There exist similar phenomena with the background shifts cases that most ViTs lead the BiT models to a large extent under both evaluations in all situations, and that a larger ViT architecture achieves a better OOD performance and narrows the IID/OOD generalization gap.

- ViTs benefit from diverse augmentation in enhancing generalization towards vicinal impurities, but their ar-



(a) Stylized-ImageNet (b) Cue Conflict Stimuli (c) ImageNet-R

Figure 3. **Results on Stylized-ImageNet, Cue Conflict Stimuli and ImageNet-R.** (a), (b) and (c) respectively illustrate the OOD Accuracy and IID/OOD Generalization Gap for different models on Stylized-ImageNet, Cue Conflict Stimuli, and ImageNet-R data sets. From (a) and (b) we could draw the following conclusions that **1)** ViTs' stronger bias towards shape enables them to generalize better under texture shifts and their shape biases have a positive correlation with their sizes, **3)** ViTs with larger patch size exhibit a stronger bias towards the shape. From (c) we observe that most ViTs beat BiTs in OOD accuracy while having little difference in the IID/OOD generalization gap.

(a) DeiT-B/16 (b) DeiT-S/16 (c) BiT (d) BiT$_{da}$

Figure 4. **Results on DomainNet.** From the results, we can conclude that **1)** DeiT-S/16 performs better on the small-scale datasets in IID conditions. Thus, the model easily outperforms BiTs in OOD accuracy, **2)** when inspecting the IID/OOD generalization gap, the results differ a lot. When models are trained on clipart and painting, there is no obvious difference of gap between DeiT-S/16 and BiTs.

chitectural advantage cannot be overlooked. Compared with BiT, BiT$_{da}$ constantly achieves about 4% better in OOD performances and IID/OOD gaps, emphasizing the contribution of diverse augmentation to model insensitivity towards pixel-level shifts. However, most ViT models are still ahead of BiT$_{da}$ under both evaluations, manifesting ViTs' performance can be partially attributed to the architecture design.

- Patch size for training has little influence on ViTs' generalization ability. Though DeiT-B/16 achieves higher OOD accuracy than DeiT-B/32, its counterpart trained with a larger patch size 32×32, there is little difference between their IID/OOD gaps. Therefore, patch size for training exert a peripheral effect on generalization ability from in-distribution data to out-of-distribution data, but only act on the model in-distribution generalization.

5.3. Texture Shifts Generalization Analysis

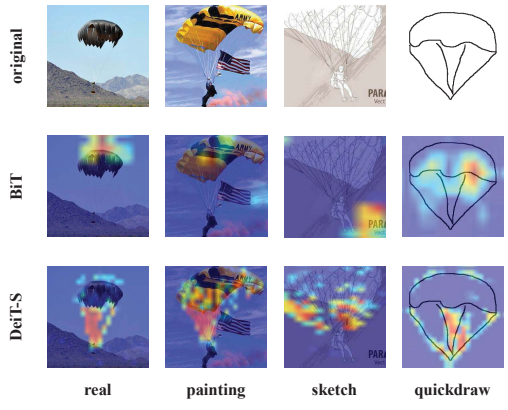
The results on Stylized-ImageNet and Cue Conflict Stimuli are shown in Fig. 3 (a) and (b).

- ViTs' stronger bias towards shape enables them to generalize better under texture shifts and their shape biases have a positive correlation with their sizes. It could be observed from results on Stylized-ImageNet that ViTs lead BiT models under both evaluations and a larger ViT architecture achieves a better OOD performance, which indicates that ViTs deal with the texture shifts better and a larger ViT contributes to better leveraging global semantic features (such as shape and object parts) and less affected by local changes. These phenomena reappear in results on Cue Conflict Stimuli that most ViTs achieve higher shape accuracy and lower texture accuracy than BiTs, which demonstrates that ViTs' insensitivities towards texture shifts are owed to their stronger bias on shape than CNNs. Meanwhile, there exists an uptrend of shape accuracy and a downtrend of texture accuracy as the ViT size increases. Hence, ViTs' shape biases have a positive correlation with their sizes.

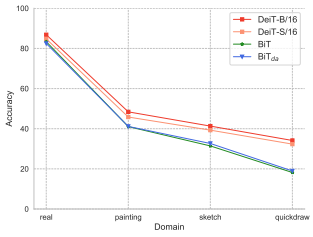
- ViTs with larger patch size exhibit a stronger bias towards the shape. On Stylized-ImageNet, DeiT-B/32 behaves better than DeiT-B/16 in OOD accuracy and IID/OOD generalization gap, which is opposite to their performances on ImageNet. Meanwhile, on Cue Conflict Stimuli, DeiT-B/32 is less affected by the misleading texture than DeiT-B/16, resulting in a higher shape accuracy. Therefore, ViTs with larger patch size rely less on local texture features and focus more on global high-level features, *i.e.* they show stronger bias towards shape lower bias towards texture.

5.4. Style Shifts Generalization Analysis

- ViTs have diverse performance on IID/OOD generalization gap under Style shifts. The results on ImageNet-R are shown in Fig. 3 (c). As ImageNet-R only contains 200 classes of ImageNet, we follow [13] to record accuracy on the ImageNet subset (ImageNet-200) and regard it as the IID result. When focusing on the accuracy on ImageNet-R, we observe most ViTs beat BiTs in OOD accuracy, while having similar performance in IID/OOD generalization gap. Accordingly, ViTs do not have competitive edges in generalizing from real images to art renditions. For Domain-

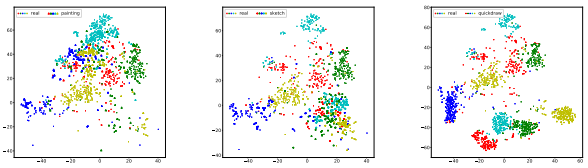


(a)

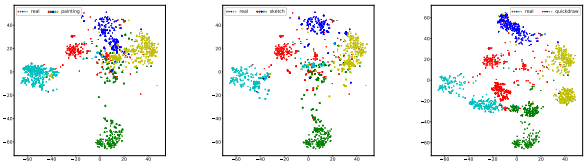


(b)

Figure 5. **Structure bias investigation.** (a) illustrates examples of class *parachute* of four domains and the Grad-CAM [25] attention maps of both BiT and DeiT-S. We shall observe that, as the color, texture, and shape cues become less and less informative from *real* to *quickdraw* and even there is only abstract structure preserved in *quickdraw*, DeiT-S constantly concentrates on the key structural information of parachutes while BiT fails to capture such essential feature. (b) shows the accuracies of models trained with real on different domains. From the results, we can see that the gap between ViTs and CNNs is getting larger when the tested domain contains fewer visual cues (*i.e.* from real to quickdraw). Therefore, we can conclude that ViTs are less affected by the shift of color, texture, and shape features, indicating that ViTs focus more on structures.



(a) rel vs. pnt, L8 (b) rel vs. skt, L8 (c) rel vs. qcd, L8



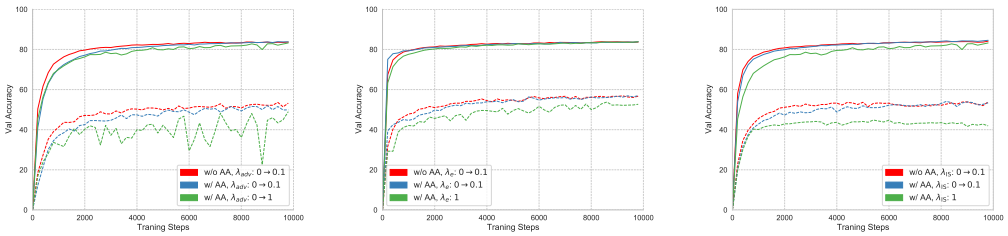
(d) rel vs. pnt, L12 (e) rel vs. skt, L12 (f) rel vs. qcd, L12

Figure 6. **T-SNE visualization results.** (a)-(c) and (d)-(e) repectively illustrates the comparison of visualizing Class Token data of *real vs. painting*, *real vs. sketch* and *real vs. quickdraw* of layer 8 and layer 12 from four domains. Please zoom for better view.

Net, we mainly compare the models with the same scale, i.e. DeiT-S/16 and BiTs, whose results are shown in Fig. 4. We observe DeiT-S/16 performs better on the small-scale datasets under IID and thus the model easily outperforms BiTs in OOD accuracy. When inspecting the IID/OOD generalization gap, the results differ a lot. When models are trained on *clipart* and *painting*, there is no obvious difference between DeiT-S/16 and BiTs. But for *real*, DeiT-S/16 leads BiTs over 4%, which can be explained as ViTs utilize the knowledge from pre-train data better if the pre-train data and downstream data are from similar distributions.

- ViTs shows stronger bias towards object structure. We further investigate how models shall behave as the other available visual cues come to degrade until there only remains structural information. We illustrate examples of class *parachute* of four domains and the Grad-CAM [25] attention maps of both BiT and DeiT-S in Fig. 5 (a). We shall observe that, as the color, texture, and shape cues become less and less informative from *real* to *quickdraw* and even there is only abstract structure preserved in *quickdraw*, DeiT-S constantly concentrates on the key structural information of parachutes while BiT fails to capture such essential feature. In addition, we test the accuracies of models trained with *real* on different domains. Since there are a considerable number of unrecognizable data in *quickdraw*, we exclude classes on which both ViTs and CNNs achieve accuracies less than 10%. We show the results in Fig. 5 (b), from which we can see that the gap between ViTs and CNNs are getting larger when the tested domain contain less visual cues (*i.e.* from *real* to *quickdraw*). Based on observations and analyses above, we can conclude that ViTs are less effected by the shift of color, texture, and shape features, indicating that ViTs focus more on structures.

- ViTs will eliminate different levels of DS in different layers. We select a set of classes of four domains in DomainNet shown in Fig. 5 (a) and the class lists are shown in the supplementary material. By extracting the intermediate Class Token and implementing dimensionality reduction via T-SNE technique [28], we generate the visualizations of Class Token data of layer 8 and layer 12 from four domains and respectively show the comparison of *real vs. painting*, *real vs. sketch* and *real vs. quickdraw*. As shown in Fig. 6, we can first observe from pictures in the first row that data from different domains are clustered together to a certain extent only in the *real vs. painting* condition at layer 8. As for *real vs. sketch*, the data become well clustered until at layer 12 (Fig. 6 (e)), whereas the *real vs. quickdraw* condition fails to mix up data from different domains together but there exists the decision boundary that can well divide data of different classes for both domains at layer 12 (Fig. 6 (f)). From the above analysis, we conclude that ViTs will eliminate different levels of DS in different layers.



(a) T-ADV (b) T-MME (c) T-SSL

Figure 7. **Investigation of Generalization-enhanced methods with different training strategies.** (a)-(c) show training curves on both source domain and target domain. From the results, we can conclude that classical training strategies (the green lines) on CNNs are not suitable for ViTs, which need smoother strategies (the red lines) to align features in both domains.

Table 2. **Results of Generalization-enhanced methods.** Specifically, we compare three types of GE-ViTs with their corresponding CNNs. From the results we could conclude that 1) equipped with GE-ViTs, we achieve significant performance boosts towards out-of-distribution data by 4% from vanilla ViTs. 2) three GE-ViTs have almost the same improvement from vanilla models on OOD accuracy. 3) for the enhanced transformer models, larger ViTs still benefit more for the out-of-distribution generalization.

Model	Method	R to C	R to P	P to C	C to S	S to P	R to S	P to R	Avg.
DeiT-B/16	-	54.64	48.40	40.37	45.69	36.75	41.31	55.33	46.07
	T-ADV	58.19	50.85	41.91	51.18	46.12	47.47	55.65	50.20
	T-MME	**60.59**	**51.98**	42.30	50.32	45.79	**47.92**	54.87	50.54
	T-SSL	56.80	49.06	**45.96**	**51.79**	**46.95**	45.95	**60.98**	**51.07**
DeiT-S/16	-	50.60	45.82	36.09	43.39	35.24	39.29	52.08	43.22
	T-ADV	53.60	47.84	37.99	47.10	41.61	41.94	52.82	46.13
	T-MME	**56.86**	**49.15**	38.97	46.48	42.95	**42.07**	52.49	47.00
	T-SSL	53.86	46.71	**42.79**	**47.25**	**43.01**	40.94	**57.07**	**47.37**
BiT	-	42.18	41.14	30.72	37.01	28.23	32.64	48.54	36.78
	DANN [10]	45.20	42.86	32.96	40.44	36.63	35.26	49.25	40.37
	MME [24]	50.21	**44.61**	34.75	40.27	38.41	37.83	47.58	41.95
	SSL [34]	**52.55**	42.80	**39.03**	**45.72**	**39.08**	**39.65**	**56.07**	**44.98**
VGG-16	-	39.39	37.32	26.36	32.96	25.55	27.79	45.70	33.58
	DANN [10]	43.26	40.09	28.68	36.22	31.63	**35.45**	44.73	37.15
	MME [24]	42.65	**42.46**	27.41	**36.93**	33.94	32.58	**45.87**	37.41
	SSL [34]	**43.79**	41.88	**32.19**	35.73	**36.99**	31.05	55.18	**39.54**

6. Studies on Generalization-Enhanced ViTs

Settings. We use DomainNet [22] for the following experiments. Following [24], we focus on the 7 scenarios listed in Tab. 2. To make a full comparison, we implement these enhancing techniques on two representative CNNs VGG-16 and BiT, and two ViTs including DeiT-S/16 and DeiT-B/16. We explore their performance on both the vanilla version and the generalization-enhanced version. Implementation details can be found in supplementary materials.

Performance Analysis. The results of three GE-ViTs comparing with CNNs are shown in Tab. 2. From the results we have the following observations: **1)** equipped with GE-ViTs, we achieve significant performance boosts towards out-of-distribution data by 4% from vanilla ViTs. **2)** Three GE-ViTs have almost the same improvement from vanilla models on OOD accuracy. In contrast, CNNs benefit more from the self-supervised learning method than the others. **3)** DeiT-B/16 has a larger gain on those enhancing methods than DeiT-S/16. Therefore, we conclude that **1)** ViTs and CNNs share many characteristics, and both can be beneficial from the generalization-enhancement methods. **2)** For the enhanced transformer models, larger ViTs still benefit more for the out-of-distribution generalization.

Smooth Feature Alignment. Fig. 7 shows the performance of GE-ViTs with different training strategies. The green line represents the same training strategies used in CNNs. The other two lines use smoother strategies. From the comparison of these strategies, we observe that **1)** the generally used automated augmentation schemes shall cause performance degradation on T-ADV while they have little influence on T-MME and T-SSL. **2)** smoother learning strategies are significant for ViT convergence, especially in the adversarial training mode. As for T-MME and T-SSL, smoothness of auxiliary losses also significantly improves the performance. Based on these observations, we conclude that GE-ViTs are more sensitive to the hyper-parameters than their corresponding CNN models.

7. Discussion and Conclusion

We provide a comprehensive study on the OOD generalization of ViTs, with the following contributions: **1)** We define a taxonomy on data distribution shifts according to the modified semantic concepts in images. **2)** We perform a comprehensive study on OOD generalization and inductive bias properties of ViTs under the five categorized distribution shifts. Several valuable observations are obtained. **3)** We further improve the OOD generalization of ViTs by designing GE-ViTs through adversarial learning, information theory, and self-supervised learning with smoother training strategies. Our work serves as an early attempt, thus there is plenty of room for developing more powerful GE-ViTs.

Broader Impacts. Some models utilized in this paper demand a large number of computing resources for training procedures. The consumption of electricity may exert an environmental impact.

Acknowledgements

This work is supported by NTU NAP, and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Yutong Bai, Jieru Mei, Alan Yuille, and Cihang Xie. Are transformers more robust than cnns? *arXiv preprint arXiv:2111.05464*, 2021. 1, 2

[2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32:9453–9463, 2019. 2

[3] Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000. 1

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020. 1, 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 1, 2

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. *arXiv preprint arXiv:2012.00364*, 2020. 1, 2

[7] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020. 1, 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[9] Qi Dou, Daniel C Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *arXiv preprint arXiv:1910.13580*, 2019. 2

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2, 3, 8

[11] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 1, 2, 3

[12] Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *arXiv preprint arXiv:2011.15091*, 2020. 1

[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *International Conference on Computer Vision*, pages 8320–8329, 2021. 1, 2, 3, 6

[14] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 1, 2

[15] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, pages 491–507. Springer, 2020. 3

[16] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995. 2

[17] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. 2

[18] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020. 4

[19] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 2

[20] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *arXiv preprint arXiv:2105.10497*, 2021. 1, 2

[21] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. *arXiv preprint arXiv:2105.07581*, 2021. 1, 2

[22] 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/CVF International Conference on Computer Vision*, pages 1406–1415, 2019. 3, 8

[23] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018. 2

[24] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019. 2, 3, 8

[25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[26] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019. 2

[27] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 1, 2, 3

[28] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 2

[30] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5339–5349, 2018. 2

[31] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. *arXiv preprint arXiv:2011.14503*, 2020. 2

[32] David H Wolpert, William G Macready, et al. No free lunch theorems for search. Technical report, Technical Report SFI-TR-95-02-010, Santa Fe Institute, 1995. 1

[33] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. 2

[34] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. *arXiv preprint arXiv:2103.16765*, 2021. 2, 3, 4, 8

[35] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1

[36] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 1, 2