

# Similarity of Neural Architectures using Adversarial Attack Transferability

Jaehui Hwang<sup>1,2,†</sup> Dongyoon Han<sup>3</sup> Byeongho Heo<sup>3</sup> Song Park<sup>3</sup>  
Sanghyuk Chun<sup>3,\*</sup> Jong-Seok Lee<sup>1,2,\*</sup>

<sup>1</sup> School of Integrated Technology, Yonsei University

<sup>2</sup>BK21 Graduate Program in Intelligent Semiconductor Technology, Yonsei University

<sup>3</sup> NAVER AI Lab

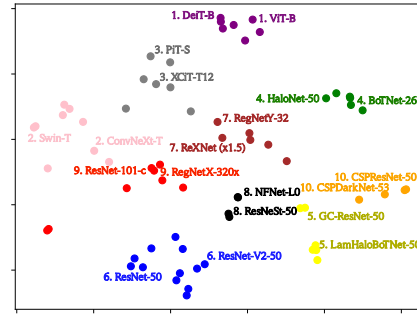
<sup>†</sup> Works done during an internship at NAVER AI Lab. \* Corresponding authors

**Abstract.** In recent years, many deep neural architectures have been developed for image classification. Whether they are similar or dissimilar and what factors contribute to their (dis)similarities remains curious. To address this question, we aim to design a quantitative and scalable similarity measure between neural architectures. We propose Similarity by Attack Transferability (SAT) from the observation that adversarial attack transferability contains information related to input gradients and decision boundaries widely used to understand model behaviors. We conduct a large-scale analysis on 69 state-of-the-art ImageNet classifiers using our SAT to answer the question. In addition, we provide interesting insights into ML applications using multiple models, such as model ensemble and knowledge distillation. Our results show that using diverse neural architectures with distinct components can benefit such scenarios.

**Keywords:** Architecture Similarity · Adversarial Attack Transferability

## 1 Introduction

The advances in deep neural networks (DNN) architecture design have taken a key role in their success by making the learning process easier (*e.g.*, normalization [3, 52, 116] or skip connection [42]), enforcing human inductive bias [60], or increasing model capability with the self-attention mechanism [106]. With different architectural components containing architectural design principles and elements, a number of different neural architectures have been proposed. They have different accuracies, but several researches have pointed out that their predictions are not significantly different [35, 71, 72].



**Fig. 1: t-SNE plot showing 10 clusters of 69 neural networks using our similarity function, SAT.**

By this, *can we say that recently developed DNN models with different architectural components are similar or the same?* The answer is *no*. It is because a model prediction is not the only characteristic to compare their similarities. Existing studies have found differences by focusing on different features, such as layer-by-layer network component [58, 82], a high-level understanding by visualization of loss surface [28], input gradient [89, 91], and decision boundary [90]. Researchers could understand the similarity between models through these trials; however, the similarity comparison methods from previous studies are insufficient for facilitating comprehensive studies because they do not satisfy two criteria that practical metrics should meet: (1) providing a quantitative similarity score and (2) being compatible with different base architectures (*e.g.*, CNN and Transformer). Recently, Tramèr et al. [103] and Somepalli et al. [90] suggested a quantitative similarity metric based on measuring differences in decision boundaries. However, these methods have limitations due to the non-tractable decision boundaries and limited computations as shown in Sec. 3.

We propose a quantitative similarity that is scalable and easily applicable to diverse architectures, named Similarity by Attack Transferability (SAT). We focus on adversarial attack transferability (AT), which indicates how generated adversarial perturbation is transferable between two different architectures. It is widely studied that the vulnerability of DNNs depends on their own architectural property or how models capture the features from inputs, such as the usage of self-attention [33], the stem layer [50], and the dependency on high or low-frequency components of input [4, 57]. Thus, if two different models are similar, the AT between the models is high because they share similar vulnerability [84]. Furthermore, AT can be a reliable approximation for comparing the input gradients [70], decision boundary [56], and loss landscape [26]. All of them are widely-used frameworks to understand model behavior and differences between models and used to measure the similarity of models in previous works [6, 18, 28, 64, 89, 90, 91, 94, 103, 113]; namely, SAT can capture various model properties.

We quantitatively measure pairwise SATs of 69 different ImageNet-trained neural architectures from [114]. We analyze what components among 13 architectural components (*e.g.*, normalization, activation, ...) that consist of neural architectures largely affect model diversity. Furthermore, we observe relationships between SAT and practical applications, such as ensemble and distillation.

## 2 Related Work

**Similarity between DNNs** has been actively explored recently. Several studies focused on comparing intermediate features to understand the behavior of DNNs. Raghu et al. [82] observed the difference between layers, training methods, and architectures (*e.g.*, CNN and ViT) based on **layer-by-layer comparison** [58]. Some studies have focused on **loss landscapes** by visualizing the loss of models on the parameter space [28, 64, 78]. Although these methods show a visual inspection, they cannot support quantitative measurements. On the other hand, our goal is to support a quantitative similarity by SAT.

Another line of research has been focused on **prediction-based statistics**, *e.g.*, comparing wrong and correct predictions [34, 35, 61, 86]. However, as recent complex DNNs are getting almost perfect, just focusing on prediction values can be misleading; Meding et al. [72] observed that recent DNNs show highly similar predictions. In this case, prediction-based methods will be no more informative. Meanwhile, our SAT can provide meaningful findings for 69 recent NNs.

**Input gradient** is another popular framework to understand model behavior by observing how a model will change predictions by local pixel changes [6, 88, 89, 91, 94]. If two models are similar, their input gradients will also be similar. These methods are computationally efficient, and no additional training is required; they can provide a visual understanding of the given input. However, input gradients are inherently noisy; thus, these methods will need additional pre-processing, such as smoothing, for a stable computation [18]. Also, these methods usually measure how the input gradient matches the actual foreground, *i.e.*, we need ground-truth foreground masks for measuring such scores. On the contrary, SAT needs no additional pre-processing and mask annotations.

**Comparing the decision boundaries** will provide a high-level understanding of how models behave differently for input changes and how models extract features from complicated data dimensions. Recent works [103, 113] suggested measuring similarity by comparing distances between predictions and decision boundaries. Meanwhile, Somepalli et al. [90] analyzed models by comparing their decision boundaries on the on-manifold plane constructed by three random images. However, these approaches suffer from inaccurate approximation, non-tractable decision boundaries, and finite pairs of inputs and predictions.

Finally, different behaviors of CNNs and Transformers have been studied in specific tasks, such as robustness [5, 74], layer-by-layer comparison [78, 82] or decision-making process [53]. Our work aims to quantify the similarity between general NNs, not only focusing on limited groups of architecture.

### 3 Similarity by Attack Transferability (SAT)

Here, we propose a quantitative similarity between two architectures using adversarial attack transferability, which indicates whether an adversarial sample from a model can fool another model. The concept of adversarial attack has effectively pointed out the vulnerabilities of DNNs by input gradient [36, 70, 95].

Interestingly, these vulnerabilities have been observed to be intricately linked to architectural properties. For example, Fu et al. [33] demonstrated the effect of the attention modules in architecture on attack success rate. Hwang et al. [50] analyzed that the stem layer structure causes models to have different adversarial vulnerable points in the input space, *e.g.*, video models periodically have vulnerable frames, such as every four frames. Namely, an adversarial sample to a model highly depends on the inherent architectural property of the model.

Another perspective emphasized the dissimilarities in dependencies on high-frequency and low-frequency components between CNN-based and transformer-based models, showing different vulnerabilities to different adversarial attacks

[4, 57]. Different architectural choices behave as different frequency filters (*e.g.*, the self-attention works as a low-pass filter, while the convolution works as a high-pass filter) [78]; thus, we can expect that the different architectural component choices will affect the model vulnerability, *e.g.*, vulnerability to high-frequency perturbations. If we can measure how the adversarial vulnerabilities of the models are different, we also can measure how the networks are dissimilar.

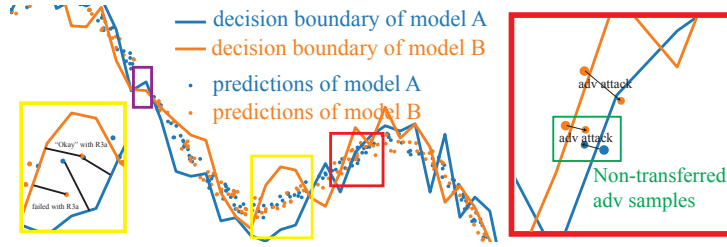
To measure how model vulnerabilities differ, we employ **adversarial attack transferability** (AT), where it indicates whether an adversarial sample from a model can fool another model. If two models are more similar, their AT gets higher [26, 65, 84]. On the other hand, because the adversarial attack targets vulnerable points varying by architectural components of DNNs [33, 49, 50, 57], if two different models are dissimilar, the AT between them gets lower. Furthermore, attack transferability can be a good approximation for measuring the differences in input gradients [70], decision boundaries [56], and loss landscape [26], where they are widely used techniques for understanding model behavior and similarity between models as discussed in the related work section. While previous approaches are limited to non-quantitative analysis, inherent noisy property, and computational costs, adversarial transferability can provide quantitative measures with low variances and low computational costs.

We propose a new similarity function that utilizes attack transferability, named **Similarity by Attack Transferability (SAT)**, providing a reliable, easy-to-conduct, and scalable method for measuring the similarity between neural architectures. Formally, we generate adversarial samples  $x_A$  and  $x_B$  of model  $A$  and  $B$  for the given input  $x$ . Then, we measure the accuracy of model  $A$  using the adversarial sample for model  $B$  (called  $\text{acc}_{B \rightarrow A}$ ). If  $A$  and  $B$  are the same, then  $\text{acc}_{B \rightarrow A}$  will be zero if the adversary can fool model  $B$  perfectly. On the other hand, if the input gradients of  $A$  and  $B$  differ significantly, then the performance drop will be neglectable because the adversarial sample is almost similar to the original image (*i.e.*,  $\|x - x_B\| \leq \varepsilon$ ). Let  $X_{AB}$  be the set of inputs where both  $A$  and  $B$  predict correctly,  $y$  be the ground truth label, and  $\mathbb{I}(\cdot)$  be the indicator function. We measure SAT between two different models by:

$$\text{SAT}(A, B) = \log \left[ \max \left\{ \varepsilon_s, 100 \times \frac{1}{2|X_{AB}|} \sum_{x \in X_{AB}} \{ \mathbb{I}(A(x_B) \neq y) + \mathbb{I}(B(x_A) \neq y) \} \right\} \right], \quad (1)$$

where  $\varepsilon_s$  is a small scalar value. If  $A = B$  and we have an oracle adversary, then  $\text{SAT}(A, A) = \log 100$ . In practice, a strong adversary (*e.g.*, PGD [70] or AutoAttack [23]) can easily achieve a nearly-zero accuracy if a model is not trained by an adversarial attack-aware strategy [22, 70]. Meanwhile, if the adversarial attacks on  $A$  are not transferable to  $B$  and vice versa, then  $\text{SAT}(A, B) = \log \varepsilon_s$ .

Ideally, we aim to define a similarity  $d$  between two models with the following properties: (1)  $n = \arg \min_m d(n, m)$ , (2)  $d(n, m) = d(m, n)$  and (3)  $d(n, m) > d(n, n)$  if  $n \neq m$ . If the adversary is perfect, then  $\text{acc}_{A \rightarrow A}$  will be zero, and it will be the minimum because accuracy is non-negative. “ $\text{acc}_{A \rightarrow B} + \text{acc}_{B \rightarrow A}$ ” is symmetric thereby SAT is symmetric. Finally, SAT satisfies  $d(n, m) \geq d(n, n)$  if  $n \neq m$  where it is a weaker condition than (3).



**Fig. 2: How SAT works?** Conceptual figure to understand SAT by the lens of the decision boundary. Each line denotes the decision boundary of a binary classification model, and each dot denotes individual prediction for given inputs.

*Comparison with other methods.* Here, we compare SAT with prediction-based measurements [34, 35, 61, 86] and similarity measurements by comparing decision boundaries (Tramèr et al. [103] and Somepalli et al. [90]). We first define two binary classifiers  $f$  and  $g$  and their predicted values  $f_p(x)$  and  $g_p(x)$  for input  $x$  (See Fig. 2).  $f$  classifies  $x$  as positive if  $f_p(x) > f_d(x)$  where  $f_d(x)$  is a decision boundary of  $f$ . We aim to measure the difference between decision boundaries, namely  $\int_x |f_d(x) - g_d(x)| dx$  to measure differences between models. However, DNNs have a non-tractable decision boundary function, thus,  $f_d$  and  $g_d$  are not tractable. Furthermore, the space of  $x$  is too large to compute explicitly. Instead, we may assume that we only have finite and sparingly sampled  $x$ .

In this scenario, we can choose three strategies. First, we can count the number of samples whose predicted labels are different for given  $x$ , which is *prediction-based measurements* or Somepalli et al. [90]. As we assumed sparsity of  $x$ , this approach cannot measure the area of uncovered  $x$  domain, hence, its approximation will be incorrect (purple box in Fig. 2) or needs too many perturbations to search uncovered  $x$ . In Appendix A.1, we empirically show that Somepalli et al. [90] suffers from the high variance even with a large number of samples while SAT shows a low variance with a small number of samples.

Second, we can measure the minimum distance between  $f_p(x)$  and  $f_d$  as Tramèr et al. [103]. This only measures the distance to its closest decision boundary without considering the other model. The yellow box of Fig. 2 shows if two predictions are similar at  $x$ , it would compute an approximation of  $|f_d(x) - g_d(x)|$  for  $x$ . However, if two predictions are different, it will compute a wrong approximation. Moreover, in practice, searching  $\epsilon$  is unstable and expensive.

Lastly, we can count the number of non-transferred adversarial samples (red box in Fig. 2), which is *our method, SAT*. If we have an oracle attack method that exactly moves the point right beyond the decision boundary, our SAT will measure the  $\ell_0$  approximation of  $\min(|f_d(x) - g_d(x)|, \epsilon)$  for given  $x$ . Namely, SAT can measure whether two decision boundaries are different by more than  $\epsilon$  for each  $x$ . If we assume that the difference between decision boundaries is not significantly large and  $\epsilon$  is properly chosen, SAT will compute an approximated decision boundary difference. We also compare SAT and other methods from the viewpoint of stability and practical usability in Sec. 5.1 and Appendix A.

*Discussions.* In practice, we do not have an oracle attack method. Instead, we employ the PGD attack [70] as the adversarial attack method. In Appendix B.1, we investigate the robustness of SAT to the choice of the attack methods. In summary, SAT measured by PGD shows a high correlation with SAT measured by various attacks, *e.g.*, AutoAttack [23], attacks designed for enhancing attack transferability, such as MIFGSM [29] and VMIFGSM [112], low-frequency targeted attacks, such as low-frequency PGD [38], method-specific attacks, such as PatchPool [33], or generative model-based attacks, such as BIA [129].

Also, SAT assumes an optimal attack with proper  $\epsilon$ . However, this assumption can be broken under the adversarial training setting when we use a practical attacker. Also, as shown by Tsipras et al. [105] and Ilyas et al. [51], adversarial training will lead to a different decision boundary from the original model. In Appendix B.2, we empirically investigate the effect of adversarial training to SAT. We observe that different adversarial training methods make as a difference as different training techniques, which we will discuss in Sec. 4.2.

*Analyzing 69 models.* Now, we analyze 69 recent ImageNet classifiers using SAT by focusing on two questions. (1) Which network component contributes to the diversity between models? (2) Why do we need to develop various neural architectures? The full list of the architectures can be found in Appendix C. We use the PGD attack [70] for the adversary. We set the iteration to 50, the learning rate to 0.1, and  $\epsilon$  to 8/255. As we discussed earlier, we show that SAT is robust to the choice of the adversarial attack method. We select 69 neural architectures trained on ImageNet [85] from the PyTorch Image Models library [114]. To reduce the unexpected effect of a significant accuracy gap, the chosen model candidates are limited to the models whose top-1 accuracy is between 79% and 83%. We also ignore the models with unusual training techniques, such as training on extra training datasets, using a small or large input resolution (*e.g.*, less than 200 or larger than 300), or knowledge distillation. When  $A$  and  $B$  take different input resolutions, then we resize the attacked image from the source network for the target network. We also sub-sample 10% ImageNet validation images (*i.e.*, 5,000 images) to measure the similarity. This strategy makes our similarity score more computationally efficient.

## 4 Model Analysis by Network Similarity

### 4.1 Which Architectural Component Causes the Difference?

*Settings.* We list 13 key architecture components: normalization (*e.g.*, BN [52] and LN [3]), activations (*e.g.*, ReLU [60] and GeLU [83]), the existence of depth-wise convolution, or stem layer (*e.g.*,  $7 \times 7$  conv,  $3 \times 3$  conv, or  $16 \times 16$  conv with stride 16 – a.k.a. “*patchify*” stem [69]). The list of the entire components is shown in the Appendix. We then convert each architecture as a feature vector based on the listed sub-modules. For example, we convert ResNet as  $f_{\text{ResNet}} = [\text{Base arch} = \text{CNN}, \text{Norm} = \text{BN}, \text{Activation} = \text{ReLU}, \dots]$ . The full list of components of 69 architectures can be found in Appendix C.

**Table 1: Clusters by SAT.** All the architectures here are denoted by the aliases defined in their respective papers. We show the top-5 keywords for each cluster based on TF-IDF. InRes, SA, and CWA denote input resolution, self-attention, and channel-wise attention, respectively. The customized model details are described in the footnote<sup>†</sup>.

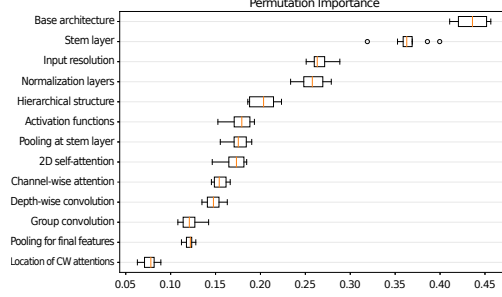
No. Top-5 Keywords	Architecture
1 Stem layer: 16×16 conv w/ s16, No Hierarchical, GeLU, LN, Final GAP	ConViT-B [31], CrossViT-B [13], DeiT-B [100], DeiT-S [100], ViT-S (patch size 16) [30], ResMLP-S24 [101], gMLP-S [67]
2 Stem layer: 4×4 conv w/ s4, LN, GeLU, Transformer, No pooling at stem	Twins-PCPVT-B [20], Twins-SVT-S [20], CoaT-Lite Small [25], NesT-T [131], Swin-T [68], S3 (Swin-T) [14], ConvNeXt-T [69], ResMLP-B24 [101]
3 Transformer, Final GAP, GeLU, Pooling at stem, InRes: 224	XCiT-M24 [1], XCiT-T12 [1], HaloRegNetZ-B** <sup>†</sup> , TNT-S [41], Visformer-S [17], PiT-S [46], PiT-B [46]
4 Stem layer: stack of 3×3 conv, 2D SA, InRes: 256, Pooling at stem, SiLU	HaloNet-50 [107], LambdaResNet-50 [7], BoTNet-26 [92], GC-ResNeXt-50 [12], ECAHaloNet-50** <sup>†</sup> , ECA-BoTNet-26** <sup>†</sup>
5 Stem layer: stack of 3×3 convs, InRes: 256, 2D SA, CWA: middle of blocks, CNN	LamHaloBoTNet-50** <sup>†</sup> , SE-BoTNet-33** <sup>†</sup> , SE-HaloNet-33** <sup>†</sup> , Halo2BoTNet-50** <sup>†</sup> , GC-ResNet-50 [12], ECA-Net-33 [111]
6 Stem layer: 7×7 conv w/ s2, ReLU, Pooling at stem, CNN, BN	ResNet-50 [42], ResNet-101 [42], ResNeXt-50 [117], Wide ResNet-50 [124], SE-ResNet-50 [48], SE-ResNeXt-50 [48], ResNet-V2-50 [43], ResNet-V2-101 [43], ResNet-50 (GN) [116], ResNet-50 (BlurPool) [130], DPN-107 [16], Xception-65 [19]
7 NAS, Stem layer: 3×3 conv w/ s2 CWA: middle of blocks, CWA, DW Conv	EfficientNet-B2 [97], FBNetV3-G [24], ReXNet (×1.5) [40], RegNetY-32 [81], MixNet-XL [98], NF-RegNet-B1 [10]
8 Input resolution: 224, Stem layer: stack of 3×3 convs, Group Conv, Final GAP, 2D SA	NFNet-L0** <sup>†</sup> , ECA-NFNet-L0** <sup>†</sup> , PoolFormer-M48 [121], ResNeSt-50 [126], ResNet-V2-50-D-EVOS** <sup>†</sup> , ConvMixer-1536/20 [104]
9 ReLU, Input resolution: 224, DW Conv, BN, 2D self-attention	ViT-B (patch size 32) [30], R26+ViT-S [93], DLA-X-102 [119], eSE-VoVNet-39 [63], ResNet-101-C [45], RegNetX-320 [81], HRNet-W32 [110]
10 ReLU + Leaky ReLU, InRes: 256, Stem layer: 7×7 conv, CNN, Pooling at stem	CSPResNet-50 [109], CSPResNeXt-50 [109], CSPDarkNet-53 [8], NF-ResNet-50 [10]

*Feature important analysis.* Now, we measure the feature importance by fitting a gradient boosting regressor [32] on the feature difference (*e.g.*,  $f_{\text{ResNet-50}} - f_{\text{DeiT-base}}$ ) measured by Hamming distance and the corresponding similarity. The details of the regressor are described in Appendix. We use the permutation importance [9] that indicates how the trained regression model changes the prediction according to randomly changing each feature. The feature importance of each architectural component is shown in Fig. 3. We first observe that the choice of base architecture (*e.g.*, CNN [60], Transformer [106], and MLP-Mixer [99]) contributes to the similarity most significantly. Fig. 3 also shows that the design choice of the input layer (*i.e.*, stem layer design choice or input resolution) affects the similarity as much as the choice of basic components such as normalization layers, activation functions, and the existence of attention layers. On the other hand, we observe that the modified efficiency-aware convolution operations, such as depth-wise convolution [19], are ineffective for diversity.

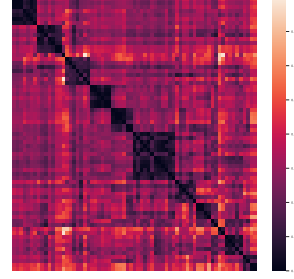
*Clustering analysis.* We additionally provide a clustering analysis based on the architectural similarities. We construct a pairwise similarity graph with adjacency matrix  $A$  between all 69 architectures where its vertex denotes an architecture, and its edge denotes the similarity between two networks. We perform

<sup>†</sup> Customized models by [114]: HaloRegNetZ = HaloNet + RegNetZ; ECA-BoTNet = ECA-Net + HaloNet + ResNeXt; ECA-BoTNet = ECA-Net + BoTNet + ResNeXt; LamHaloBoTNet = LambdaNet + HaloNet + BoTNet; SE-BoTNet = SENet + BoTNet; SE-HaloNet = SENet + HaloNet; Halo2BoTNet = HaloNet + BoTNet; NFNet-L0 = an efficient variant of NFNet-FO [11]; ECA-NFNet-L0 = ECA-Net + NFNet-L0; ResNet-V2-D-EVOS = ResNet-V2 + EvoNorms [66].





**Fig. 3: Importance of architectural components to network similarity.** 13 components are sorted by the contribution to the similarities. The larger feature importance means the component contributes more to the network similarity.



**Fig. 4: Pairwise distances of spectral features.** Rows and columns are sorted by the clustering index. More details are described in Appendix D.3.

the spectral clustering [75] on  $A$  where the number of clusters  $K$  is set to 10: We compute the Laplacian matrix of  $A$ ,  $L = D - A$  where  $D$  is the diagonal matrix and its  $i$ -th component is  $\sum_j A_{ij}$ . Then, we perform K-means clustering on the  $K$ -largest eigenvectors of  $L$ . The pairwise distances of spectral features (*i.e.*, 10-largest eigenvectors of  $L$ ) of 69 neural architectures are shown in Fig. 4. The rows and columns of Fig. 4 are sorted by the clustering index (Tab. 1). More details with model names are described in Appendix D.2. We can see the block-diagonal patterns, *i.e.*, in-clusters similarities are more significant than between-clusters similarities. More details are in Appendix D.3.

Tab. 1 shows the clustering results on 69 networks and the top-5 keywords for each cluster based on term frequency-inverse document frequency (TF-IDF) analysis. Specifically, we treat each model feature as a word and compute TF and IDF by treating each architecture as a document. Then we compute the average TF-IDF for each cluster and report top-5 keywords. Similar to Fig. 3, the base architecture (*e.g.*, CNN in Cluster 5, 6, 10 and Transformer in Cluster 2, 3) and the design choice for the stem layer (*e.g.*, Cluster 1, 2, 4, 5, 6, 7, 8, 10) repeatedly appear at the top keywords. Especially, we can observe that the differences in base architecture significantly cause the diversity in model similarities, *e.g.*, non-hierarchical Transformers (Cluster 1), hierarchical networks with the patchification stem (Cluster 2), hierarchical Transformers (Cluster 3), CNNs with 2D self-attention (Cluster 4, 5), ResNet-based architectures (Cluster 6), and NAS-based architectures (Cluster 7).

## 4.2 The Relationship between Training Strategy and SAT

The architectural difference is not the only cause of the model diversity. We compare the impact by different architecture choices (*e.g.*, ResNet and ViT) and by different training strategies while fixing the model architecture, as follows: **Dif-**



**Table 2: SAT within the same architecture.** We compare the average similarity within the same architecture but trained with different procedures, “All” denotes the average similarity of 69 architectures.

Architecture	ResNet-50	ViT-S
Init	4.23	4.21
Hparam	4.05	4.22
Tr. Reg.	3.27	3.44
All	2.73	

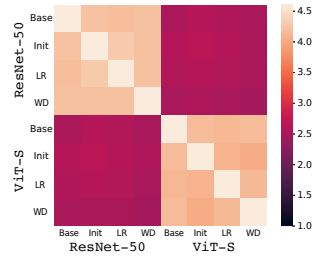
**Table 3: Ensemble performance with diverse architectures.** We report the error reduction rate by varying the number of ensemble models and the diversity of the ensemble models (related to Fig. 7a). “rand” indicates the random choice of models.

	less diverse ← # of clusters → more diverse					rand
	1	2	3	4	5	
# of models						
2	7.13	<b>7.84</b>				7.78
3	10.17	10.84	<b>11.20</b>			11.11
4	11.70	12.45	12.80	<b>13.00</b>		12.90
5	12.58	13.41	13.79	13.99	<b>14.11</b>	14.01

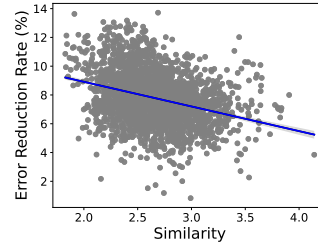
**ferent initializations** can affect the model training by the nature of the stochasticity of the training procedure. For example, Somepalli et al. [90] showed that the decision boundary of each architecture could vary by different initializations. We also consider **different optimization hyper-parameters** (*e.g.*, learning rate, weight decay). Finally, we study the effect of **different training regimes** (*e.g.*, augmentations, type of supervision). For example, the choice of data augmentation [122, 125] or label smoothing [96] can theoretically or empirically affect adversarial robustness [21, 77, 87, 127]. We also investigate the effect of supervision, such as self-supervision [15, 37, 44] or semi-weakly supervised learning [118]. Note that the training strategies inevitably contain the former ones. For example, when we train models with different training regimes, models have different initialization seeds and different optimization hyper-parameters. Comparing 69 different architectures also contains the effect of different initialization and optimization hyper-parameters and parts of different training regimes. This is necessary for achieving high classification performance.

Tab. 2 shows the comparison of similarity scores between the same architecture but different learning methods (a smaller similarity means more diversity). We report two architectures, ResNet-50 and ViT-S, and their training settings are in Appendix E. We also show the average SAT between all 69 architectures. In the table, we first observe that using different random initialization or different optimization hyper-parameters shows high correlations with each other (almost  $\geq 4.2$ ) while the average similarity score between various neural architectures is 2.73. In other words, the difference in initializations or optimization hyper-parameters does not significantly contribute to the model diversity.

Second, we observe that using different learning techniques remarkably affects SAT (3.27 for ResNet and 3.44 for ViT), but is not as significant as the architectural difference (2.73). Furthermore, the change of SAT caused by different initializations or hyper-parameters is less marked than the change caused by different architecture (Fig. 5). These observations provide two insights. First, the diversity resulting from various training strategies is not significant enough



**Fig. 5: Pairwise distance of spectral features by different optimizations.** Init, LR, and WD are randomly chosen from models trained with different settings of initialization, learning rate, and weight decay in Tab. 2.



**Fig. 6: Correlation between pairwise SAT and ensemble performance.** The trend line and its 90% confidence interval are shown.

compared to the diversity of architecture. Second, designing new architecture is more efficient in achieving diverse models rather than re-training the same one.

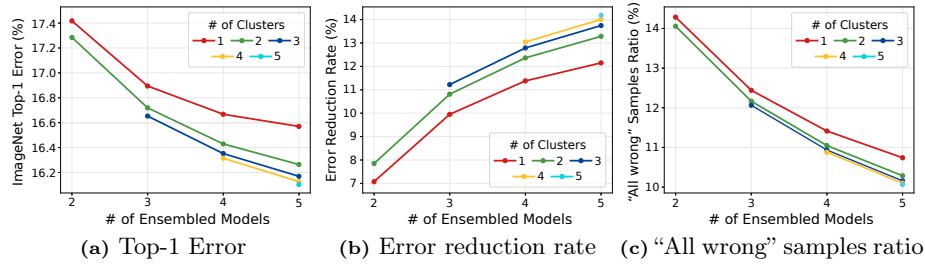
## 5 SAT Applications with Multiple Models

Here, we analyze how SAT is related to downstream tasks involving more than one model. First, we show that using more diverse models will lead to better ensemble performance. Second, we study the relationship between knowledge distillation and SAT. Furthermore, we can suggest a similarity-based guideline for choosing a teacher model when distilling to a specific architecture. Through these observations, we can provide insights into the necessity of diverse models.

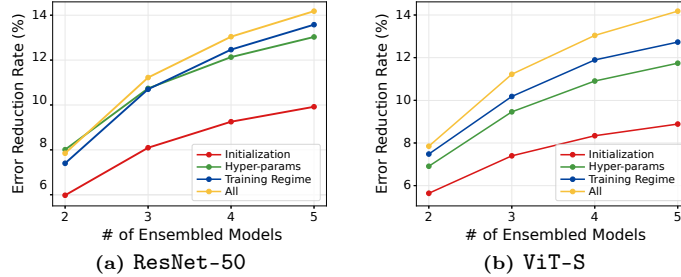
### 5.1 Model Diversity and Ensemble

*Settings.* The model ensemble is a practical technique for achieving high performance. However, only few works have studied the relationship between ensemble performance and model similarity, particularly for large-scale complex models. Previous studies are mainly conducted on tiny datasets and linear models [61]. We investigate the change of ensemble performance by the change of similarity based on the unweighted average method [55] (*i.e.*, averaging the logit values of the ensembled models). Because the ensemble performance is sensitive to the original model performances, we define Error Reduction Rate (ERR) as  $1 - \frac{\text{Err}_{\text{ens}}(M)}{\frac{1}{|M|} \sum_{m \in M} \text{Err}(m)}$ , where  $M$  is the set of the ensembled models,  $\text{Err}(m)$  denotes the top-1 ImageNet validation error of model  $m$ , and  $\text{Err}_{\text{ens}}(\cdot)$  denotes the top-1 error of the model ensemble results.

*Results.* We first measure the 2-ensemble performances among the 69 architectures (*i.e.*, the number of ensembles is  $\binom{69}{2} = 2346$ ). We plot the relationship between SAT and ERR in Fig. 6. We observe that there exists a strong negative



**Fig. 7: Model diversity and ensemble performance.** We report ensemble performances by varying the number of ensembled models ( $N$ ) and the diversity of the models. The diversity is controlled by choosing the models from  $k$  different clusters.



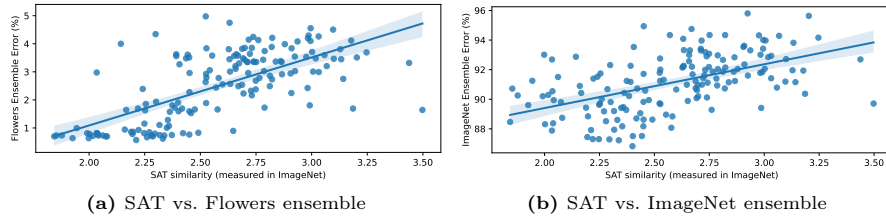
**Fig. 8: Diversity by training techniques and ensemble.** We report the the same metrics as Fig. 7 for various ResNet-50 and ViT-S models in Tab. 2.

correlation between the model similarity and the ensemble performance (Pearson correlation coefficient  $-0.32$  with p-value  $\approx 0$  and Spearman correlation  $-0.32$  with p-value  $\approx 0$ , *i.e.*, more diversity leads to better ensemble performance).

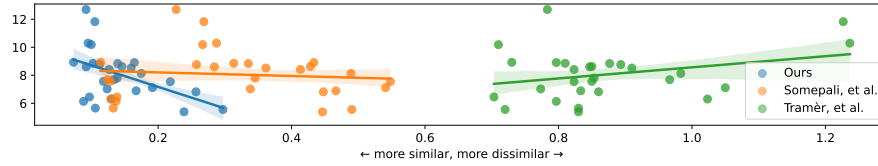
We also conduct  $N$ -ensemble experiments with  $N \geq 2$  based on our clustering results in Tab. 1. We evaluate the average ERR of the ensemble of models from  $k$  clusters, *i.e.*, if  $N = 5$  and  $k = 3$ , the ensembled models are only sampled from the selected 3 clusters while ignoring the other 7 clusters. We investigate the effect of model diversity and ensemble performance by examining  $k = 1 \dots N$  (*i.e.*, larger  $k$  denotes more diverse ensembled models). We report the result with ImageNet top-1 error and ERR in Fig. 7a and Fig. 7b.

In all metrics, we observe that the ensemble of more diverse models shows better performance. Interestingly, Fig. 7b shows that when the number of clusters for the model selection ( $k$ ) is decreased, the ensemble performance by the number of ensembled models ( $N$ ) quickly reaches saturation. Tab. 3 shows that the ensemble performances by choosing the most diverse models via SAT always outperform the random ensemble. Similarly, Fig. 7c shows that the number of wrong samples by all models is decreased by selecting more diverse models.

*Training Strategy vs. Architecture in the ensemble scenario?* Remark that Tab. 2 showed that the different training strategies are not as effective as different ar-



**Fig. 9: Cross-dataset SAT results.** SAT measured on ImageNet also has a positive correlation with ensemble performances on Flowers-102 [76].



**Fig. 10: Empirical comparison.** Relationship between the model similarity, including SAT, Somepalli et al. [90] and Tramèr et al. [103], and 2-ensemble performance.

chitectures for diversity. To examine this on the ensemble scenario, we report the ensemble results of different training strategies, *i.e.*, the same **ResNet-50** and **ViT-S** in Tab. 2. For comparison with different architectures, we also report the ensemble of different architectures where all ensembled models are from different clusters (*i.e.*,  $N=k$  in Fig. 7). Fig. 8 shows that although using diverse training regimes (blue lines) improves ensemble performance compared to other techniques (red and green lines), the improvements by using different architectures (yellow lines) are more significant than the improvements by using different training regimes (blue lines) with large gaps.

*Generalizability to other datasets.* We examine whether more diverse architectures in ImageNet SAT also lead to better ensemble performances on the other datasets. We fine-tuned all 69 architectures to the Flowers-102 dataset [76], and filter out low performing models ( $< 95\%$  top-1 accuracy). After the filtering, we have 16 fine-tuned models. Using the fine-tuned models, we plot the relationship between the Flowers-102 ensemble performance and SAT score measured in ImageNet. Fig. 9 shows that SAT also highly correlates with Flowers ensemble performances, despite that SAT is measured on ImageNet. This experimental result supports that SAT similarity can be applied in a cross-domain manner.

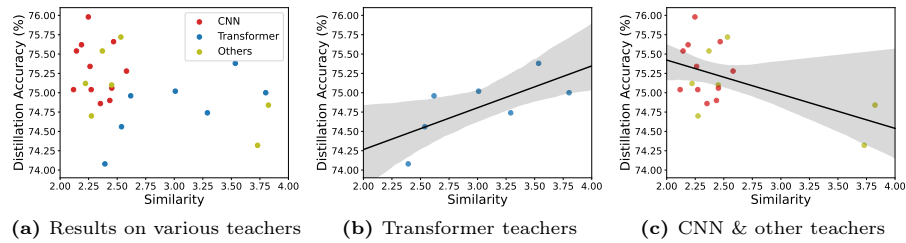
*Comparison of different similarity functions in the ensemble scenario.* Finally, we compare the impact of the choice of the similarity function and the ensemble performance when following our setting. We compare SAT with Somepalli et al. [90] and Tramèr et al. [103] on the 2-ensemble scenario with 8 out of 69 models due to the stability issue of Tramèr et al. [103]. Fig. 10 shows the relationship between various similarity functions and the 2-ensemble performance. We observe

that SAT only shows a strong positive correlation (blue line), while the others show an almost random or slightly negative correlation. Finally, in Appendix A.2, we compare SAT with Somepalli et al. [90] and a naive architecture-based clustering using our features under the same setting of Fig. 7 and 8. Similarly, SAT shows the best ensemble performance against the comparison methods.

## 5.2 Model Diversity and Knowledge Distillation

Knowledge distillation (KD) [47] is a training method for transferring rich knowledge of a well-trained teacher network. Intuitively, KD performance affects a lot by choice of the teacher network; however, the relationship between similarity and KD performance has not yet been explored enough, especially for ViT. This subsection investigates how the similarity between teacher and student networks contributes to the distillation performance. There are several studies showing two contradictory conclusions; Jin *et al.* [54] and Mirzadeh *et al.* [73] showed that a similar teacher leads to better KD performance; Touvron *et al.* [100] reports that distillation from a substantially different teacher is beneficial for ViT.

We train 25 ViT-Ti models with different teacher networks from 69 models that we used by the hard distillation strategy [47]. Experimental details are described in Appendix. Fig. 11a illustrates the relationship between the teacher-student similarity and the distillation performance. Fig. 11a tends to show a not significant negative correlation between teacher-student similarity and distillation performance ( $-0.32$  Pearson correlation coefficient with  $0.12$  p-value). However, if we only focus on when the teacher and student networks are based on the same architecture (*i.e.*, Transformer), we can observe a strong positive correlation (Fig. 11b) –  $0.70$  Pearson correlation coefficient with  $0.078$  p-value. In this case, our observation is aligned with [54, 73]: a teacher similar to the student improves distillation performance. However, when the teacher and student networks are based on different architectures (*e.g.*, CNN), then we can observe a stronger negative correlation (Fig. 11c) with  $-0.51$  Pearson correlation coefficient and  $0.030$  p-value. In this case, a more dissimilar teacher leads



**Fig. 11: Model diversity and distillation performance.** (a) We show the relationship between teacher-student similarity and distillation performance of 25 DeiT-S models distilled by various teacher networks. We show the relationship when the teacher and student networks are based on (b) Transformer and (c) otherwise.

to better distillation performance. We also test other factors that can affect distillation performance in Appendix; We observe that distillation performance is not correlated to teacher accuracy in our experiments.

Why do we observe contradictory results for Transformer teachers (Fig. 11b) and other teachers (Fig. 11c)? Here, we conjecture that when the teacher and student networks differ significantly, distillation works as a strong regularizer. In this case, using a more dissimilar teacher can be considered a stronger regularizer (Fig. 11c). On the other hand, we conjecture that if two networks are similar, then distillation works as easy-to-follow supervision for the student network. In this case, a more similar teacher will work better because a more similar teacher will provide more easy-to-follow supervision for the student network (Fig. 11b). Our experiments show that the regularization effect improves distillation performance better than easy-to-follow supervision (*i.e.*, the best-performing distillation result is by a CNN teacher). Therefore, in practice, we recommend using a significantly different teacher network for achieving better distillation performance (*e.g.*, using RegNet [81] teacher for ViT student as [100]).

## 6 Discussion

In Appendix G, we describe more discussions related to SAT. We first propose an efficient approximation of SAT when we have a new model; instead of generating adversarial samples from all models, only generating adversarial samples from the new model can an efficient approximation of SAT (Appendix G.1). We also show that SAT and the same misclassified samples have a positive correlation in Appendix G.2. Appendix G.3 demonstrates that we can estimate the similarity with a not fully trained model (*e.g.*, a model in an early stage). Finally, we describe more possible applications requiring diverse models (Appendix G.4).

## 7 Conclusion

We have explored similarities between image classification models to investigate what makes the model similar or diverse and whether developing and using diverse models is required. For quantitative and model-agnostic similarity assessment, we have suggested a new similarity function, named SAT, based on attack transferability demonstrating differences in input gradients and decision boundaries. Using SAT, we conduct a large-scale and extensive analysis using 69 state-of-the-art ImageNet models. We have shown that macroscopic architectural properties, such as base architecture and stem architecture, have a more significant impact on similarity than microscopic operations, such as types of convolution, with numerical analysis. Finally, we have provided insight into the ML applications using multiple models based on SAT, *e.g.*, model ensemble or knowledge distillation. Overall, we suggest using SAT to improve methods with multiple models in a practical scenario with a large-scale training dataset and a highly complex architecture.

## Acknowledgement

We thank Taekyung Kim and Namuk Park for comments on the self-supervised pre-training. This work was supported by an IITP grant funded by the Korean Government (MSIT) (RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University)) and by the Yonsei Signature Research Cluster Program of 2024 (2024-22-0161).

## Author Contributions

This work is done as an internship project by J Hwang under the supervision of S Chun. S Chun initialized the project idea: understanding how different architectures behave differently by using an adversarial attack. J Hwang, S Chun, and D Han jointly designed the analysis toolbox. J Hwang implemented the analysis toolbox and conducted the experiments with input from S Chun, D Han, and J Lee. J Hwang, D Han, B Heo, and S Chun contributed to interpreting and understanding various neural architectures under our toolbox. The initial version of “model card” (Tab. C.3 and C.4) was built by J Hwang, S Park, and verified by D Han and B Heo. B Heo contributed to interpreting distillation results. All ResNet and ViT models newly trained in this work were trained by S Park. J Lee supervised J Hwang and verified the main idea and experiments during the project. S Chun and J Hwang wrote the initial version of the manuscript. All authors contributed to the final manuscript.

## A Empirical Comparison of SAT and Other Methods

**Table A.1: Comparison of stability of measurements.** We compare the stability of method by [90] and SAT. Stability is indicated by the standard deviation (std). The numbers in  $(\cdot)$  mean the sampling ratio to all possible combinations to compute the exact value. “cost” denotes relative costs compared to the total forward costs for the 50K ImageNet validation set: Somepalli *et al.* needs  $\binom{50K}{3} = 2.1 \times 10^{13}$  and SAT needs 50K. Here, we assume that forward and backward computations cost the same.

Somepalli et al. [90]			SAT (ours)		
# triplets	std	cost	# images	std	cost
10 ( $4.8 \times 10^{-13}$ )	4.49	1.0	500 (0.01)	1.88	1.0
20 ( $9.6 \times 10^{-13}$ )	3.28	2.0	1000 (0.02)	1.05	2.0
50 ( $2.4 \times 10^{-12}$ )	1.63	5.0	2500 (0.05)	0.91	5.2
100 ( $4.8 \times 10^{-12}$ )	1.54	10.0	5000 (0.1)	0.77	10.2



### A.1 Comparison with Somepalli et al. in the Variance of Similarity

Somepalli et al. [90] proposed a sampling-based similarity score for comparing decision boundaries of models. SAT has two advantages over Somepalli *et al.*: computational efficiency and the reliability of the results. First, SAT involves sampling 5,000 images and using 50-step PGD; the computation cost is  $[5,000 \text{ (sampled images)} \times 50 \text{ (PGD steps)} + 5,000 \text{ (test to the other model)}] \times 2 \text{ (two models)}$ . Meanwhile, Somepalli et al. [90] sample 500 triplets and generate 2,500 points to construct decision boundaries. In this case, the total inference cost is  $[500 \text{ (sampled triplets)} \times 2,500 \text{ (grid points)}] \times 2 \text{ (two models)}$ , which is 4.9 times larger than SAT. Secondly, Somepalli *et al.* sampled three images of different classes. As the original paper used CIFAR-10 [59], 500 triplets can cover all possible combinations of three classes among the ten classes ( $\binom{10}{3} = 120 < 500$ ). However, it becomes computationally infeasible to represent all possible combinations of three classes among many classes (*e.g.*, ImageNet needs  $\binom{1000}{3} = 164,335,500$ , 130 times greater than its training images). Also, we find that the similarity of [90] is unreliable when the number of sample triplets is small. In Tab. A.1, we calculate the similarity scores between ConvNeXt-T [69] and Swin-T [68] from ten different sets with varying sample sizes. SAT exhibits significantly better stability (*i.e.*, low variances) than Somepalli *et al.* Note that we use our similarity measurement as the percentage degree without the logarithmic function and control the scale of samples to maintain similar computation complexity between SAT and the compared method for a fair comparison.

### A.2 Ensemble Performance Comparison of SAT and Other Methods

The purpose of our study is to provide a new lens for model similarity through adversarial attack transferability. Since we do not have the ground truth of the “similarity” between architectures, comparing different similarity functions is not really meaningful. Instead, we indirectly compare SAT, Somepalli et al. [90] and naive architecture feature-based clustering on the ensemble benchmark. More specifically, the naive architecture clustering is based on our architecture features proposed in Sec. 4.1; we apply the K-means clustering algorithm to get clusters. We note that the other comparison methods cannot be applied due to the expensive computations. The results are shown in Tab. A.2 and Tab. A.3.

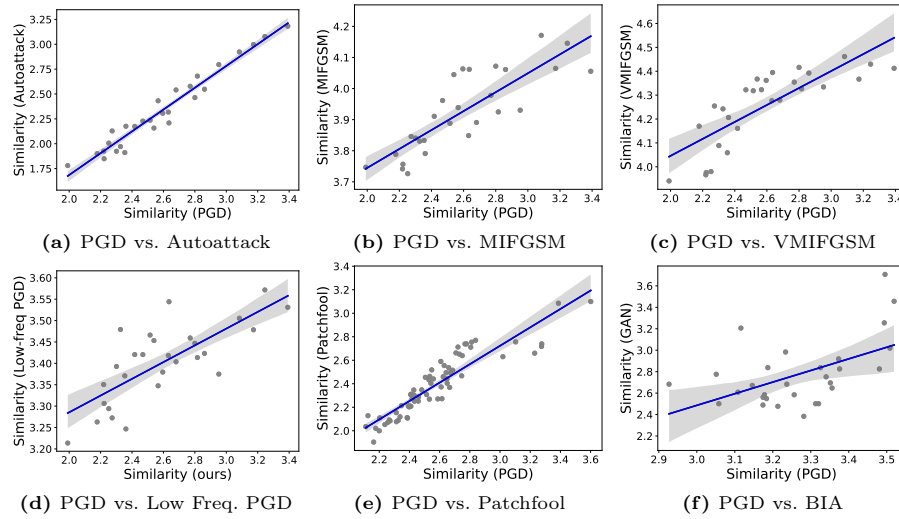
**Table A.2:** Somepalli *et al.*

	← # of clusters →				
	1	2	3	4	5
2	7.54	<b>7.79</b>			
3	10.85	11.03	<b>11.13</b>		
4	12.63	12.79	12.87	<b>12.93</b>	
5	13.74	13.89	13.95	14.01	<b>14.05</b>

**Table A.3:** Arc-based clustering

	← # of clusters →					rand	SAT
	1	2	3	4	5		
2	7.60	<b>7.80</b>				7.78	<b>7.84</b>
3	11.02	11.08	<b>11.12</b>			11.11	<b>11.20</b>
4	13.03	12.92	12.89	12.90		12.90	<b>13.00</b>
5	14.37	14.13	14.03	14.00	14.00	14.01	<b>14.11</b>

In the tables, SAT is the best-performing model similarity score on the ensemble task. We also tried to Tramèr et al. [103] in the same setting, but Tramèr



**Fig.B.1: Effect of different adversarial attack methods to SAT.** The trend line and its 90% confidence interval are shown. We show the relationship between our SAT using PGD [70] and SAT using other attacks, (a) Autoattack [23] (b) MIFGSM [29] (c) VMIFGSM [112] (d) low-frequency PGD [38] (e) Patchfool attack [33], and (f) BIA [129].

et al. [103] often failed to converge and show very small differences. Hence, we couldn’t use Tramèr et al. [103] for measuring all 69 arches used in the paper. Instead, as we reported in the main paper, we compare Tramèr et al. [103] and other model similarity variants on 8 architectures used in Fig. B.1. Among the candidate methods, we observe that SAT shows the strongest correlation with the ensemble performance using two models.

## B Discussions

### B.1 Robustness of SAT to the choice of the attack methods.

Our goal is not to design an attack-free method but to show the potential of using adversarial attack transferability (AAT) for measuring quantitative model similarity. However, we have checked that SAT scores are robust to the choice of the attack methods, even for stronger attacks, such as AutoAttack [23], attacks designed for enhancing AAT, such as MIFGSM [29] and VMIFGSM [112], low-frequency targeted attacks, such as low-frequency PGD [38], method-specific attacks, such as PatchPool [33], or generative model-based attacks, such as BIA [129]. We sample 8 representative models among 69 models for testing the effect of attacks on SAT; ViT-S, CoaT-Lite Small, ResNet-101, LamHaloBotNet50, ReXNet-150, NFNet-L0, Swin-T and Twins-ppvt. Fig. B.1a shows the high correlation between SAT scores using PGD and Autoattack; it shows a correlation

coefficient of 0.98 with a p-value of  $1.43 \times 10^{-18}$ . For testing the Patchfool attack, we only generate adversarial perturbations on ViT-S and get attack transferability to all other models (68 models) because it only targets Transformers. Fig. B.1b and B.1c show similar results: SAT shows consistent results even for the attacks designed for better AAT. The results show that SAT score is robust to the choice of attack methods if the attack is strong enough. Also, the low-frequency attack [38] shows a similar result, i.e., the frequency-targeted attack does not affect the similarity results. In Fig. B.1e, Patchfool also shows a high correlation compared to the PGD attack (correlation coefficient 0.91 with p-value  $3.62 \times 10^{-27}$ ). We additionally provide a result for generative model-based attack, BIA [129]. As BIA needs to train a new generative model for a different architecture, we only show the pre-trained models provided by the authors. We also get a similar result with previous results for BIA. Note that the compared attack methods are not model agnostic or computationally expensive than PGD, *e.g.*, PatchFool needs a heavy modification on the model code to extract attention layer outputs manually, and BIA needs to train a new generator for a new architecture. As SAT shows consistent rankings across the attack methods, we use PGD due to its simplicity.

## B.2 Impact of Adversarial Training to SAT

While our main analyses are based on ImageNet-trained models, in this subsection, we use CIFAR-10-trained models for two reasons. First, it is still challenging to achieve a high-performing adversarially trained model on the ImageNet scale. On the other hand, in the CIFAR-10 training setting, a number of adversarially-trained models are available and comparable. Second, adversarial training models show lower clean accuracy than normally trained models [105]. Adversarial robustness and accuracy are in a trade-off, and there is no ImageNet model with accuracy aligned with our target models yet.

We choose five adversarial training ResNet-18 from the AutoAttack repository [23] and measure SAT using the models. The average SAT between adversarial training models is **3.15**, slightly lower than the similarity score with different training strategies for ImageNet ResNet-50 (3.27 – See Tab. 2). In other words, we can confirm that different adversarial training methods make as a difference as different training techniques.

## C Details of Architectures Used in the Analyses

We use 69 models in our research to evaluate the similarity between models and to investigate the impact of model diversity. In the main paper, we mark the names of models based on their research paper and PyTorch Image Models library (`timm`; 0.6.7 version) [114]. Tab. C.1 shows the full list of the models based on their research paper and `timm` alias.

We show brief information of the architectural components in Tab. C.2. The full network specification is shown in Tab. C.3 and Tab. C.4. We follow the corresponding paper and `timm` library to list the model features.

**Table C.1: Lists of 69 models and their names based on their research paper and timm library.**

in timm	in paper	in timm	in paper	in timm	in paper
botnet26t_256	BoTNet-26	gluon_xception65	Xception-65	resnet50_gn	ResNet-50 (GN)
coat_lite_small	Coat-Lite Small	gmip_s16_224	gMLP-S	resnetblur50	ResNet-50 (BlurPool)
convit_base	ConvViT-B	halo2botnet50ts_256	Halo2BoTNet-50	resnetv2_101	ResNet-V2-101
convmixer_1536_20	ConvMixer-1536/20	halonet50ts	HaloNet-50	resnetv2_50	ResNet-V2-50
convnext_tiny	ConvNeXt-T	haloregnetz_b	HaloRegNetZ	resnetv2_50d_evos	ResNet-V2-50-EVOS
crossvit_base_240	CrossViT-B	hrnet_w64	HRNet-W32	resnext50_32x4d	ResNeXt-50
cspdarknet53	CSPDarkNet-53	jx_nest_tiny	NesT-T	rexnet_150	ReXNet ( $\times 1.5$ )
cspresnet50	CSPResNet-50	lambda_resnet50ts	LambdaResNet-50	sebotnet33ts_256	SEBotNet-33
cspresnext50	CSPResNeXt-50	lamhalobotnet50ts_256	LamHaloBoTNet-50	sehalonet33ts	SEHaloNet-33
deit_base_patch16_224	DeiT-B	mixnet_xl	MixNet-XL	seresnet50	SEResNet-50
deit_small_patch16_224	DeiT-S	nf_regnet_b1	NF-RegNet-B1	seresnext50_32x4d	SEResNeXt-50
dla102x2	DLA-X-102	nf_resnet50	NF-ResNet-50	swin_s3_tiny_224	S3 (Swin-T)
dpn107	DPN-107	nfnet_10	NFNet-L0	swin_tiny_patch4_window7_224	Swin-T
eca_botnext26ts_256	ECA-BoTNeXt-26	pit_b_224	PiT-B	tnt_s_patch16_224	TNT-S
eca_halonext26ts	ECA-HaloNeXt-26	pit_s_224	PiT-S	twins_pcpvt_base	Twins-PCPVT-B
eca_nfnet_l0	ECA-NFNet-L0	poolformer_m48	PoolFormer-M48	twins_svt_small	Twins-SVT-S
eca_resnet33ts	ECA-ResNet-33	regnetx_320	RegNetX-320	visformer_small	VisFormer-S
efficientnet_b2	EfficientNet-B2	regnety_032	RegNetY-32	vit_base_patch32_224	ViT-B
ese_vovnet39b	eSE-VoVNet-39	resmlp_24_224	ResMLP-S24	vit_small_patch16_224	ViT-S
fbnetv3_g	FBNetV3-G	resmlp_big_24_224	ResMLP-B24	vit_small_r26_s32_224	R26+ViT-S
gcrsnet50t	GCRsNet-50	resnet50d	ResNeSt-50	wide_resnet50_2	Wide ResNet-50
gcrsnext50ts	GCRsNeXt-50	resnet101	ResNet-101	xcit_medium_24_p16_224	XCiT-M24
gluon_resnet101_v1c	ResNet-101-C	resnet50	ResNet-50	xcit_tiny_12_p8_224	XCiT-T12

**Table C.2: Overview of model elements.** We categorize each architecture with 13 different architectural components. The full feature list is in the Appendix.

Components Elements
Base architecture CNN, Transformer, MLP-Mixer, Hybrid (CNN + Transformer), NAS-Net
Stem layer 7×7 conv with stride 2, 3×3 conv with stride 2, 16×16 conv with stride 16, ...
Input resolution 224×224, 256×256, 240×240, 299×299
Normalization layer BN, GN, LN, LN + GN, LN + BN, Normalization-free, ...
Using hierarchical structure Yes ( <i>e.g.</i> , CNNs, Swin [68]), No ( <i>e.g.</i> , ViT [30])
Activation functions ReLU, HardSwish, SiLU, GeLU, ReLU + GeLU, ReLU + SiLU or GeLU ...
Using pooling at stem Yes, No
Using 2D self-attention Yes ( <i>e.g.</i> , [7], [92], [107]), No
Using channel-wise (CW) attention Yes ( <i>e.g.</i> , [48], [12], [111]), No
Using depth-wise convolution Yes, No
Using group convolution Yes, No
Type of pooling for final feature Classification (CLS) token, Global Average Pool (GAP)
Location of CW attentions At the end of each block, in the middle of each block, ...

**Table C.3:** Description of features of 69 models. “s” in “Stem layer” indicates the stride of a layer in the stem, and the number before and after “s” are a kernel size and size of stride, respectively. For example, “3s2/3/3” means that the stem is composed of the first layer having  $3 \times 3$  kernel with stride 2, the second layer having  $3 \times 3$  kernel with stride 1, and the last layer having  $3 \times 3$  with stride 1.

Model name	Base architecture	Hierarchical structure	Stem layer	Input resolution	Normalization	Activation
botnet26t_256	CNN	Yes	3s2/3/3	256 × 256	BN	ReLU
convmixer_1536_20	CNN	Yes	7s7	224 × 224	BN	GeLU
convnext_tiny	CNN	Yes	4s4	224 × 224	LN	GeLU
csdparknet53	CNN	Yes	3s1	256 × 256	BN	Leaky ReLU
cspronet50	CNN	Yes	7s2	256 × 256	BN	Leaky ReLU
cspronet50	CNN	Yes	7s2	256 × 256	BN	Leaky ReLU
dla102x2	CNN	Yes	7s1	224 × 224	BN	ReLU
dpn107	CNN	Yes	7s2	224 × 224	BN	ReLU
eca_botnet26ts_256	CNN	Yes	3s2/3/3	256 × 256	BN	SiLU
eca_halonet26ts	CNN	Yes	3s2/3/3	256 × 256	BN	SiLU
eca_nfnet_l0	CNN	Yes	3s2/3/3/3s2	224 × 224	Norm-free	SiLU
eca_resnet33ts	CNN	Yes	3s2/3/3s2	256 × 256	BN	SiLU
ese_vovnet39b	CNN	Yes	3s2/3/3s2	224 × 224	BN	ReLU
gcrsnet50t	CNN	Yes	3s2/3/3s2	256 × 256	LN + BN	ReLU
gcrsnet50ts	CNN	Yes	3s2/3/3	256 × 256	LN + BN	ReLU + SiLU
gluon_resnet101_v1c	CNN	Yes	3s2/3/3	224 × 224	BN	ReLU
gluon_xception65	CNN	Yes	3s2/3	299 × 299	BN	ReLU
halo2botnet50ts_256	CNN	Yes	3s2/3/3s2	256 × 256	BN	SiLU
halonet50ts	CNN	Yes	3s2/3/3	256 × 256	BN	SiLU
hrnet_w64	CNN	Yes	3s2/3s2	224 × 224	BN	ReLU
lambda_resnet50ts	CNN	Yes	3s2/3/3	256 × 256	BN	SiLU
lamhalobotnet50ts_256	CNN	Yes	3s2/3/3s2	256 × 256	BN	SiLU
nf_resnet50	CNN	Yes	7s2	256 × 256	Norm-free	ReLU
nfnet_l0	CNN	Yes	3s2/3/3/3s2	224 × 224	Norm-free	ReLU + SiLU
poolformer_m48	CNN	Yes	7s4	224 × 224	GN	GeLU
resnet50d	CNN	Yes	3s2/3/3	224 × 224	BN	ReLU
resnet101	CNN	Yes	7s2	224 × 224	BN	ReLU
resnet50	CNN	Yes	7s2	224 × 224	BN	ReLU
resnet50_gn	CNN	Yes	7s2	224 × 224	GN	ReLU
resnetb1ur50	CNN	Yes	7s2	224 × 224	BN	ReLU
resnetv2_101	CNN	Yes	7s2	224 × 224	BN	ReLU
resnetv2_50	CNN	Yes	7s2	224 × 224	BN	ReLU
resnetv2_50d_evos	CNN	Yes	3s2/3/3	224 × 224	EvoNorm	-
resnext50_32x4d	CNN	Yes	7s2	224 × 224	BN	ReLU
sebotnet33ts_256	CNN	Yes	3s2/3/3s2	256 × 256	BN	ReLU + SiLU
sehalonet33ts	CNN	Yes	3s2/3/3s2	256 × 256	BN	ReLU + SiLU
seresnet50	CNN	Yes	7s2	224 × 224	BN	ReLU
seresnet50_32x4d	CNN	Yes	7s2	224 × 224	BN	ReLU
wide_resnet50_2	CNN	Yes	7s2	224 × 224	BN	ReLU
convit_base	Transformer	No	16s16	224 × 224	LN	GeLU
crossvit_base_240	Transformer	Yes	16s16	240 × 240	LN	GeLU
deit_base_patch16_224	Transformer	No	16s16	224 × 224	LN	GeLU
deit_small_patch16_224	Transformer	No	16s16	224 × 224	LN	GeLU
jx_nest_tiny	Transformer	Yes	4s4	224 × 224	LN	GeLU
pit_s_224	Transformer	Yes	16s8	224 × 224	LN	GeLU
swin_tiny_patch4_window7_224	Transformer	Yes	4s4	224 × 224	LN	GeLU
tnt_s_patch16_224	Transformer	Yes	7s4	224 × 224	LN	GeLU
vit_base_patch32_224	Transformer	No	32x32	224 × 224	LN	GeLU
vit_small_patch16_224	Transformer	No	16s16	224 × 224	LN	GeLU
gnmlp_s16_224	MLP-Mixer	Yes	16s16	224 × 224	LN	GeLU
resmlp_24_224	MLP-Mixer	No	16s16	224 × 224	Affine transform	GeLU
resmlp_big_24_224	MLP-Mixer	Yes	8s8	224 × 224	Affine transform	GeLU
swin_s3_tiny_224	NAS (TFM)	Yes	4s4	224 × 224	LN	GeLU
efficientnet_b2	NAS (CNN)	Yes	3s2	256 × 256	BN	SiLU
fnnetv3_g	NAS (CNN)	Yes	3s2	240 × 240	BN	HardSwish
haloregnetz_b	NAS (CNN)	Yes	3s2	224 × 224	BN	ReLU + SiLU
mixnet_xl	NAS (CNN)	Yes	3s2	224 × 224	BN	ReLU + SiLU
nf_regnet_b1	NAS (CNN)	Yes	3s2	256 × 256	Norm-free	ReLU + SiLU
regnetx_320	NAS (CNN)	Yes	3s2	224 × 224	BN	ReLU
regnety_032	NAS (CNN)	Yes	3s2	224 × 224	BN	ReLU
rexnet_150	NAS (CNN)	Yes	3s2	224 × 224	BN	ReLU + SiLU + ReLU6
coat_lite_small	Hybrid	Yes	4s4	224 × 224	LN	GeLU
pit_b_224	Hybrid	Yes	14s7	224 × 224	LN	GeLU
twins_pcpvt_base	Hybrid	Yes	4s4	224 × 224	LN	GeLU
twins_svt_small	Hybrid	Yes	4s4	224 × 224	LN	GeLU
visformer_small	Hybrid	Yes	7s2	224 × 224	BN	GeLU + ReLU
vit_small_r26_s32_224	Hybrid	No	7s2	224 × 224	LN + GN	GeLU + ReLU
xcit_medium_24_p16_224	Hybrid	No	3s2/3s2/3s2/3s2	224 × 224	LN + BN	GeLU
xcit_tiny_12_p8_224	Hybrid	No	3s2/3s2/3s2	224 × 224	LN + BN	GeLU

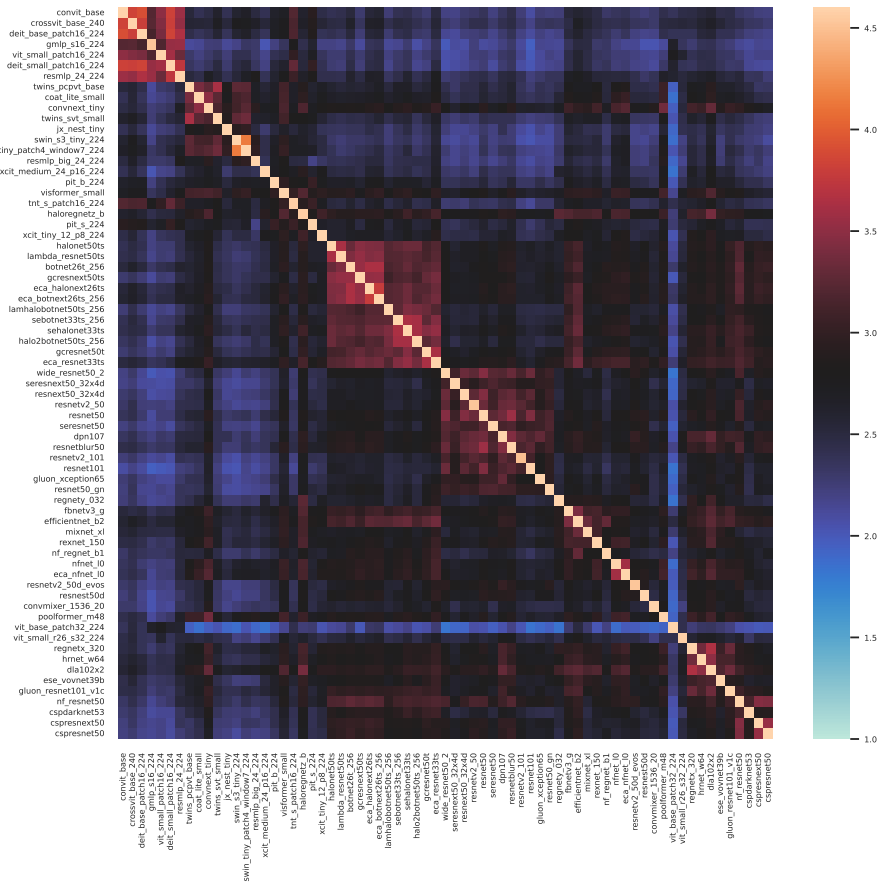
**Table C.4:** Description of features of 69 models. “Pooling (stem)” and “Pooling (final)” denote “Pooling at the stem” and “Pooling for final feature”, respectively. “SA”, “CW”, and “DW” means “Self-attention”, “Channel-wise”, and “Depth-wise”, respectively.

Model name	Pooling (stem)	Pooling (final)	2D SA	CW attention	Location of CW attention	DW conv	Group conv
botnet26t_256	Yes	GAP	Yes (BoT)	No		No	No
convmixer_1536_20	No	GAP	No	No		Yes	No
convnext_tiny	No	GAP	No	No		Yes	No
csdparknet53	No	GAP	No	No		No	No
csdpresnet50	Yes	GAP	No	No		No	No
csdpresnet50	Yes	GAP	No	No		No	Yes
dla102s2	No	GAP	No	No		No	Yes
dpa107	Yes	GAP	No	No		No	Yes
eca_botnet26ts_256	Yes	GAP	Yes (BoT)	Yes (ECA)	Middle	No	Yes
eca_halonext26ts	Yes	GAP	Yes (Halo)	Yes (ECA)	Middle	No	Yes
eca_nfnet_l0	No	GAP	No	Yes (ECA)	End	No	Yes
eca_resnet33ts	No	GAP	No	Yes (ECA)	Middle	No	No
ese_vovnet39b	No	GAP	No	Yes (ESE)	End	No	No
gcrsnet50t	No	GAP	No	Yes (GCA)	Middle	Yes	No
gcrsnext50ts	Yes	GAP	No	Yes (GCA)	Middle	Yes	Yes
gluon_resnet101_v1c	Yes	GAP	No	No		No	No
gluon_xception65	No	GAP	No	No		Yes	No
halo2botnet50ts_256	No	GAP	Yes (Halo, BoT)	No		No	No
halonet50ts	Yes	GAP	Yes (Halo)	No		No	No
hrnet_w64	No	GAP	No	No		No	No
lambda_resnet50ts	Yes	GAP	Yes (Lambda)	No		No	No
lamhalobotnet50ts_256	No	GAP	Yes (Lambda, Halo, BoT)	No		No	No
lrf_resnet50	Yes	GAP	No	No		No	No
nfnet_l0	No	GAP	No	Yes (SE)	End	No	Yes
poolformer_m48	No	GAP	No	No		No	No
resnest50d	Yes	GAP	No	Yes		No	Yes
resnet101	Yes	GAP	No	No		No	No
resnet50	Yes	GAP	No	No		No	No
resnet50_gn	Yes	GAP	No	No		No	No
resnetblur50	Yes	GAP	No	No		No	No
resnetv2_101	Yes	GAP	No	No		No	No
resnetv2_50	Yes	GAP	No	No		No	No
resnetv2_50d_evoe	Yes	GAP	No	No		No	No
resnext50_32x4d	Yes	GAP	No	No		No	Yes
sebotnet33ts	No	GAP	Yes (BoT)	Yes (SE)	Middle	No	No
sehalonet33ts	No	GAP	Yes (Halo)	Yes (SE)	Middle	No	No
seresnet50	Yes	GAP	No	Yes (SE)	End	No	No
seresnext50_32x4d	Yes	GAP	No	Yes (SE)	End	No	Yes
wide_resnet50_2	Yes	GAP	No	No		No	No
convit_base	No	CLS token	No	No		No	No
crossvit_base_240	No	CLS token	No	No		No	No
deit_base_patch16_224	No	CLS token	No	No		No	No
deit_small_patch16_224	No	CLS token	No	No		No	No
jx_nest_tiny	No	GAP	No	No		No	No
pit_s_224	No	CLS token	No	No		Yes	No
swin_tiny_patch4_window7_224	No	GAP	No	No		No	No
tnt_s_patch16_224	No	CLS token	No	No		No	No
vit_base_patch32_224	No	CLS token	No	No		No	No
vit_small_patch16_224	No	CLS token	No	No		No	No
gmip_s16_224	No	GAP	No	No		No	No
resmlp_24_224	No	GAP	No	No		No	No
resmlp_big_24_224	No	GAP	No	No		No	No
swin_s3_tiny_224	No	GAP	No	No		No	No
efficientnet_b2	No	GAP	No	Yes (SE)	Middle	Yes	No
fbnetv3_g	No	GAP	No	Yes (SE)	Middle	Yes	No
haloregnetz_b	No	GAP	Yes (Halo)	Yes (SE)	Middle	No	Yes
mixnet_s1	No	GAP	No	Yes (SE)	Middle	Yes	No
nf_regnet_b1	No	GAP	No	Yes (SE)	Middle	Yes	Yes
regnetx_320	No	GAP	No	No		No	Yes
regnety_032	No	GAP	No	Yes (SE)	Middle	No	Yes
rexnet_150	No	GAP	No	Yes (SE)	Middle	Yes	No
coat_lite_small	No	CLS token	No	No		Yes	No
pit_b_224	No	CLS token	No	No		Yes	No
twins_pcpvt_base	No	GAP	No	No		Yes	No
twins_svt_small	No	GAP	No	No		Yes	No
visformer_small	No	GAP	No	No		No	Yes
vit_small_r26_s32_224	Yes	CLS token	No	No		No	No
xcit_medium_24_p16_224	No	CLS token	No	No		Yes	No
xcit_tiny_12_p8_224	No	CLS token	No	No		Yes	No

## D Feature Importance and Clustering Details

### D.1 Feature Importance Analysis Details

We fit a gradient boosting regressor [32] based on `Scikit-learn` [80] and report the permutation importance of each architectural component in Fig. 3 of the main paper. The number of boosting stages, maximum depth, minimum number of samples, and learning rate are set to 500, 12, 4, and 0.02, respectively. Permutation importance is computed by permuting a feature 10 times.



**Fig.D.1: Pairwise similarity among 69 models.** Rows and columns are sorted by the clustering index in Tab. 2.  $(n, n)$  component of pairwise similarity is close to 4.6 (log 100) because the attack success rate is almost 100% when a model used for generating adversarial perturbation and attacked model are the same.

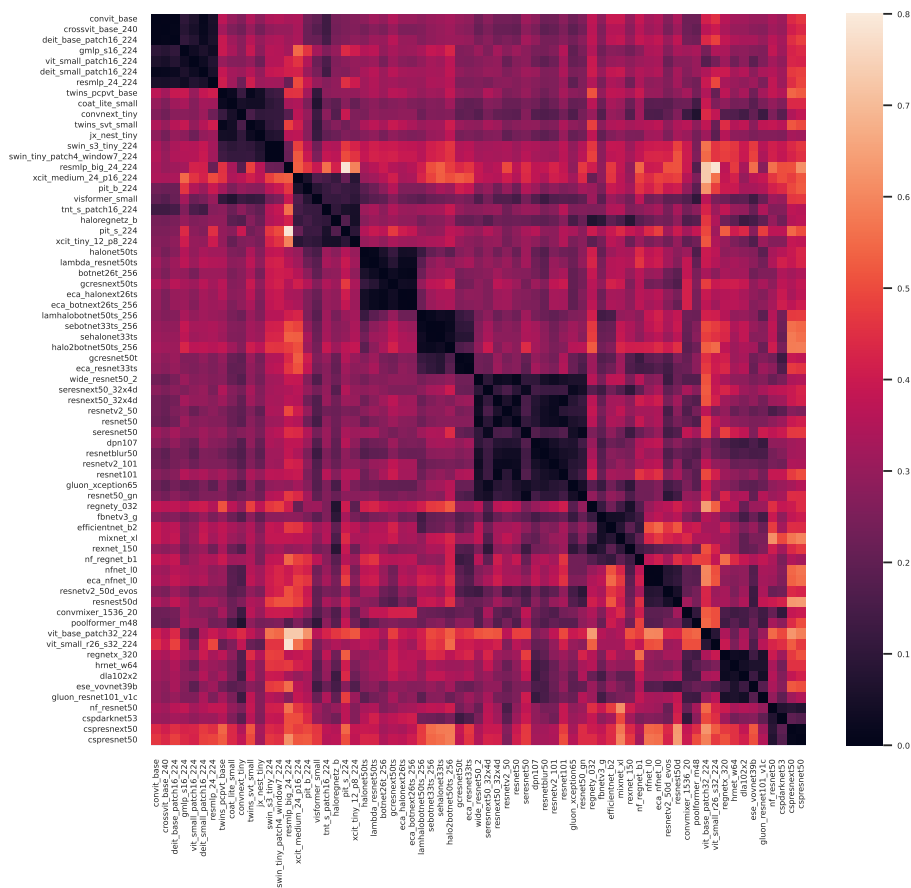


## D.2 Pairwise Similarities

Fig. D.1 indicates the pairwise similarity among 69 models. We can observe a weak block pattern around clusters, as also revealed in Fig. D.2.

### D.3 Spectral Clustering Details

We use the normalized Laplacian matrix to compute the Laplacian matrix. We also run K-means clustering 100 times and choose the clustering result with the best final objective function to reduce the randomness by the K-means clustering.



**Fig. D.2: Spectral features of 69 architectures.** The  $K$ -th largest eigenvectors of the Laplacian matrix of the pairwise similarity graph of 69 architectures are shown ( $K = 10$  in this figure). Rows and columns are sorted by the clustering index in Tab. 2. We denote the model name in `timm` for each row and column.

We visualize the pairwise distances of the spectral features (*i.e.*,  $K$ -largest eigenvectors of  $L$ ) of 69 architectures with their model names in Fig. D.2. This figure is an extension of Fig. 4, now including the model names. Note that rows and columns of Fig. D.2 are sorted by the clustering results. Fig. D.2 shows block diagonal patterns, *i.e.*, in-cluster similarities are large while between-cluster similarities are small.

## E Training Settings for Models Used in Analyses

*Models with various training methods for Sec. 4.1.* We train 21 ResNet-50 models and 16 ViT-S from scratch individually by initializing each network with different random seeds. We further train 28 ResNet-50 models by randomly choosing learning rate ( $\times 0.1$ ,  $\times 0.2$ ,  $\times 0.5$ ,  $\times 1$ ,  $\times 2$ , and  $\times 5$  where the base learning rate is 0.1), weight decay ( $\times 0.1$ ,  $\times 0.2$ ,  $\times 0.5$ ,  $\times 1$ ,  $\times 2$ , and  $\times 5$  where the base weight decay is  $1e-4$ ), and learning rate scheduler (step decay or cosine decay). Similarly, we train 9 ViT-S models by randomly choosing learning rate ( $\times 0.2$ ,  $\times 0.4$ , and  $\times 1$  where the base learning rate is  $5e-4$ ) and weight decay ( $\times 0.2$ ,  $\times 0.4$ , and  $\times 1$  where the base weight decay is 0.05). Note that the DeiT training is unstable when we use a larger learning rate or weight decay than the base values. Finally, we collect 22 ResNet-50 models with different training regimes: 1 model with standard training by PyTorch [79]; 4 models trained by GluonCV [39]<sup>1</sup>; a semi-supervised model and semi-weakly supervised model on billion-scale unlabeled images by Yalniz *et al.* [118]<sup>2</sup>; 5 models trained by different augmentation methods (Cutout [27], Mixup [125], manifold Mixup [108], CutMix [122], and feature CutMix<sup>3</sup>; 10 optimized ResNet models by [115]<sup>4</sup>. We also collect 7 ViT-S models with different training regimes, including the original ViT training setup [30]<sup>5</sup>, a stronger data augmentation setup in the Deit paper [100]-3<sup>5</sup>, the training setup with distillation [100]-3<sup>5</sup>, an improved DeiT training setup [102]-3<sup>5</sup>, and self-supervised training fashions by MoCo v3 [15]<sup>6</sup>, MAE [44]<sup>7</sup> and BYOL [37]<sup>8</sup>. We do not use adversarially-trained networks because the adversarial training usually drops the standard accuracy significantly [105].

*Distillation models for Sec. 5.* We train ViT-Ti student models with 25 different teacher models using hard distillation strategy. We follow the distillation training setting of DeiT official repo<sup>9</sup>, only changing the teacher model. Note that we

<sup>1</sup> gluon\_resnet50\_v1b, gluon\_resnet50\_v1c, gluon\_resnet50\_v1d, and gluon\_resnet50\_v1s from timm library.

<sup>2</sup> ssl\_resnet50 and swsl\_resnet50 from timm library.

<sup>3</sup> We use the official weights provided by <https://github.com/clovaai/CutMix-PyTorch>.

<sup>4</sup> We use the official weights provided by <https://github.com/rwightman/pytorch-image-models/releases/tag/v0.1-rsb-weights>

<sup>5</sup> deit\_small\_patch16\_224, vit\_small\_patch16\_224, deit\_small\_distilled\_patch16\_224, and deit3\_small\_patch16\_224 from timm library.

<sup>6</sup> We train the ViT-S model by following <https://github.com/facebookresearch/moco-v3>

<sup>7</sup> We train the ViT-S model by following <https://github.com/facebookresearch/mae>

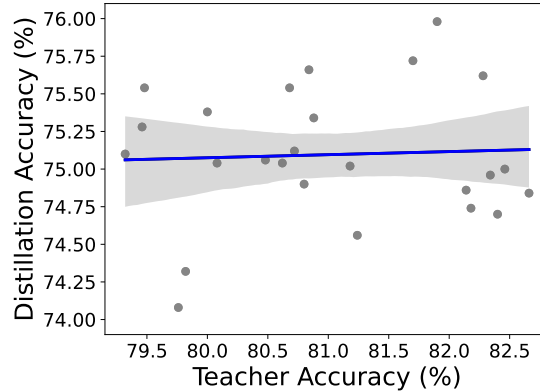
<sup>8</sup> We train the ViT-S model by following <https://github.com/lucidrains/byol-pytorch>

<sup>9</sup> <https://github.com/facebookresearch/deit>.

resize the input images to the input size the teacher model requires if the input sizes of student and teacher models differ. If a teacher model needs a different input resolution, such as  $240 \times 240$  and  $256 \times 256$ , we resize the input image for distilling it. Because `DeiT-Ti` has low classification accuracy compared to teacher models, the similarity score is calculated between `DeiT-S` and 25 models. The 25 teacher models are as follows: `BoTNet-26`, `CoaT-Lite_Small`, `ConViT-B`, `ConvNeXt-B`, `CrossViT-B`, `CSPDarkNet-53`, `CSPResNeXt-50`, `DLA-X-102`, `DPN-107`, `EfficientNet-B2`, `FBNetV3-G`, `GC-ResNet-50`, `gMLP-S`, `HaloRegNetZ`, `MixNet-XL`, `NFNet-L0`, `PiT-S`, `RegNetY-032`, `ResMLP-24`, `ResNet-50`, `SEHaloNet33`, `Swin-T`, `TNT-S`, `VisFormer-S`, and `XCiT-T12`.

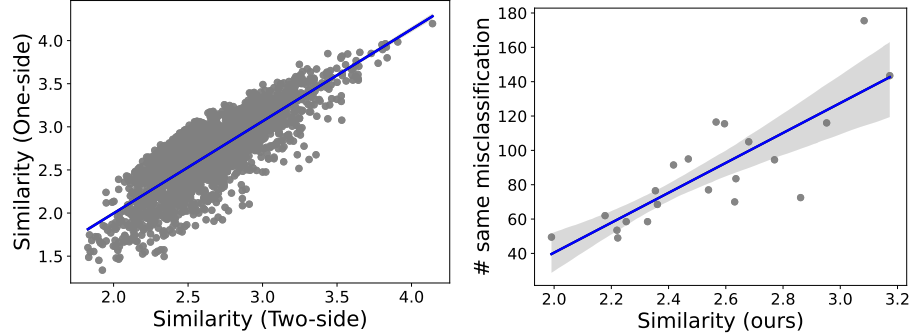
## F Knowledge Distillation

### F.1 Teacher Accuracy and Distillation Performance



**Fig. F.1: Teacher network accuracy and distillation performance.** There is no significant correlation between teacher accuracy and distillation performance.

The similarity between teacher and student networks may not be the only factor contributing to distillation. For example, a stronger teacher can lead to better distillation performance [123]. In Fig. F.1, we observe that if the range of the teacher accuracy is not significantly large enough (*e.g.*, between 79.5 and 82.5), then the correlation between teacher network accuracy and distillation performance is not significant; 0.049 Pearson correlation coefficient with 0.82 p-value. In this case, we can conclude that the teacher and student networks similarity contributes more to the distillation performance than the teacher performance.



(a) Approximated (one-side) SAT vs. Similarity (Two-side). (b) SAT vs. The number of same misclassification.

Fig. G.1: Additional Analysis for SAT.

## G Limitations and Discussions

### G.1 Efficient Approximation of SAT for a Novel Model

We can use our toolbox for designing a new model; we can measure SAT between a novel network and existing  $N$  architectures; a novel network can be assigned to clusters (Tab. 1) to understand how it works. However, it requires generating adversarial samples for all  $N + 1$  models (*e.g.*, 70 in our case), which is computationally inefficient. Instead, we propose the approximation of Eq. 1 by omitting to compute the accuracy of the novel network on the adversarial samples of the existing networks. It will break the symmetricity of SAT, but we found that the approximated score and the original score have high similarity – 0.82 Pearson coefficient with almost 0 p-value – as shown in Fig. G.1a.

As an example, we tested `Gluon-ResNeXt-50` [39] and the distilled version of `DeiT-S` [100]. As observed in Tab. 2 and Fig. 5, models with the same architecture have high similarity compared to models with different architectures; hence, we expect that `Gluon-ResNeXt-50` is assigned to the same cluster with `ResNeXt-50`, and distilled `DeiT-S` is assigned to the same cluster with `DeiT-S`. As we expected, each network is assigned to the desired cluster. Therefore, we suggest using our efficient approximation for analyzing a novel network with our analysis toolbox.

### G.2 Adversarial Attack Transferability and Direction of Misclassification

Waseda et al. [113] showed that adversarial attack transferability is highly related to the direction of the misclassification. We examine if SAT is related to the misclassification. Fig. G.1b shows the relationship between SAT and the number of the same misclassification by the attack. We observe that they are highly correlated, *i.e.*, we confirmed that SAT is also related to the misclassification.

### G.3 Change of SAT During Training

We check the adversarial attack transferability between the fully trained model and less trained models on CIFAR-10 with 180 training epochs. A model trained with only 20 epochs shows high similarity over different initializations (4.23 in Tab. 2 of the main paper). Note that SAT considers models having similar clean accuracy; namely, there is room to explore this further in future work.

**Table G.1:** Change of SAT between fully-trained model (epoch 180) and models on various epochs.

Epoch	0	20	40	60	80	100	120	140	160	180
SAT	2.84	4.46	4.56	4.58	4.59	4.59	4.60	4.60	4.60	4.60

### G.4 More Possible Applications Requiring Diverse Models

In the main paper, we introduce several applications with multiple models, such as model ensemble, knowledge distillation, and novel model development. As another example, we employ SAT-based diverse model selection for improving the dataset distillation (DD) task with random network selection [128]. DD task [62, 120, 128, 132, 133] aims to synthesize a small (usually less than 5 images per class) but informative dataset that prevents a significant drop from the original performance. Acc-DD [128] employs multiple random networks for DD, where each network is randomly selected during the training. In this study, we show that a more diverse network selection can help synthesize more informative and diverse condensed images. We replace the random selection of Acc-DD (**Rand**) with the selection by the probability proportional to (1) the similarity ( $P_{sim}$ ) or (2) the inverse of similarity ( $P_{sim^{-1}}$ ). More specifically, we first (a) select a network randomly and (b) select the next network by (1) or (2) with the current network. We repeat (b) similar to K-means++ [2]. We report the CIFAR-10 results by setting images per class as 1 using 50 CNNs. **Rand** shows 48.6 top-1 accuracy, while  $P_{sim}$  and  $P_{sim^{-1}}$  show 48.7 and **49.4**, respectively. Namely, a more diverse network selection ( $P_{sim^{-1}}$ ) helps Acc-DD.

## References

- [1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. In *Adv. Neural Inform. Process. Syst.*, 2021. 7
- [2] David Arthur, Sergei Vassilvitskii, et al. k-means++: The advantages of careful seeding. In *Soda*, pages 1027–1035, 2007. 27
- [3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1, 6
- [4] Jiawang Bai, Li Yuan, Shu-Tao Xia, Shuicheng Yan, Zhifeng Li, and Wei Liu. Improving vision transformers by revisiting high-frequency components. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 2, 4
- [5] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in neural information processing systems*, 34:26831–26843, 2021. 3
- [6] Naman Bansal, Chirag Agarwal, and Anh Nguyen. Sam: The sensitivity of attribution methods to hyperparameters. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 3
- [7] Irwan Bello. Lambdanetworks: Modeling long-range interactions without attention. In *Int. Conf. Learn. Represent.*, 2021. 7, 19
- [8] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 7
- [9] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 7
- [10] Andrew Brock, Soham De, and Samuel L Smith. Characterizing signal propagation to close the performance gap in unnormalized resnets. In *Int. Conf. Learn. Represent.*, 2021. 7
- [11] Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *Int. Conf. Mach. Learn.*, 2021. 7
- [12] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Int. Conf. Comput. Vis. Worksh.*, 2019. 7, 19
- [13] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Int. Conf. Comput. Vis.*, 2021. 7
- [14] Minghao Chen, Kan Wu, Bolin Ni, Houwen Peng, Bei Liu, Jianlong Fu, Hongyang Chao, and Haibin Ling. Searching the search space of vision transformer. In *Adv. Neural Inform. Process. Syst.*, 2021. 7
- [15] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *Int. Conf. Comput. Vis.*, 2021. 9, 24
- [16] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. In *Adv. Neural Inform. Process. Syst.*, 2017. 7
- [17] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Int. Conf. Comput. Vis.*, 2021. 7
- [18] Junsuk Choe, Seong Joon Oh, Sanghyuk Chun, and Hyunjung Akata, Zeynepand Shim. Evaluation for weakly supervised object localization: Protocol, metrics, and datasets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 2, 3

- [19] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 7
- [20] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *Adv. Neural Inform. Process. Syst.*, 2021. 7
- [21] Sanghyuk Chun, Seong Joon Oh, Sangdoo Yun, Dongyoon Han, Junsuk Choe, and Youngjoon Yoo. An empirical evaluation on robustness and uncertainty of regularization methods. In *Int. Conf. Mach. Learn. Worksh.*, 2019. 9
- [22] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Int. Conf. Mach. Learn.*, 2019. 4
- [23] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Int. Conf. Mach. Learn.*, 2020. 4, 6, 17, 18
- [24] Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Bichen Wu, Zijian He, Zhen Wei, Kan Chen, Yuandong Tian, Matthew Yu, Peter Vajda, et al. Fbnetv3: Joint architecture-recipe search using predictor pretraining. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 7
- [25] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. In *Adv. Neural Inform. Process. Syst.*, 2021. 7
- [26] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th USENIX security symposium (USENIX security 19)*, pages 321–338, 2019. 2, 4
- [27] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 24
- [28] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Int. Conf. Mach. Learn.*, 2017. 2
- [29] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 6, 17
- [30] 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 *Int. Conf. Learn. Represent.*, 2021. 7, 19, 24
- [31] Stéphane d’Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *Int. Conf. Mach. Learn.*, 2021. 7
- [32] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001. 7, 22
- [33] Yonggan F Fu, Shang Wu, Yingyan Lin, et al. Patch-fool: Are vision transformers always robust against adversarial perturbations? *Int. Conf. Learn. Represent.*, 2022. 2, 3, 4, 6, 17
- [34] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Adv. Neural Inform. Process. Syst.*, 2018. 3, 5



- [35] Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In *Adv. Neural Inform. Process. Syst.*, 2020. 1, 3, 5
- [36] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Int. Conf. Learn. Represent.*, 2015. 3
- [37] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Adv. Neural Inform. Process. Syst.*, 2020. 9, 24
- [38] Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. *UAI*, 2019. 6, 17, 18
- [39] Jian Guo, He He, Tong He, Leonard Lausen, Mu Li, Haibin Lin, Xingjian Shi, Chenguang Wang, Junyuan Xie, Sheng Zha, Aston Zhang, Hang Zhang, Zhi Zhang, Zhongyue Zhang, Shuai Zheng, and Yi Zhu. Gluoncv and gluonnlp: Deep learning in computer vision and natural language processing. *Journal of Machine Learning Research*, 21(23):1–7, 2020. 24, 26
- [40] Dongyoon Han, Sangdoo Yun, Byeongho Heo, and YoungJoon Yoo. Rethinking channel dimensions for efficient model design. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 7
- [41] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. In *Adv. Neural Inform. Process. Syst.*, 2021. 7
- [42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 1, 7
- [43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Eur. Conf. Comput. Vis.*, 2016. 7
- [44] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 9, 24
- [45] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 7
- [46] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Int. Conf. Comput. Vis.*, 2021. 7
- [47] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. In *Adv. Neural Inform. Process. Syst. Worksh.*, 2015. 13
- [48] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 7, 19
- [49] Jisung Hwang, Younghoon Kim, Sanghyuk Chun, Jaejun Yoo, Ji-Hoon Kim, and Dongyoon Han. Where to be adversarial perturbations added? investigating and manipulating pixel robustness using input gradients. *ICLR Workshop on Debugging Machine Learning Models*, 2019. 4
- [50] Jaehui Hwang, Jun-Hyuk Kim, Jun-Ho Choi, and Jong-Seok Lee. Just one moment: Structural vulnerability of deep action recognition against one frame attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7668–7676, 2021. 2, 3, 4
- [51] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Adv. Neural Inform. Process. Syst.*, 2019. 6

- [52] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Int. Conf. Mach. Learn.*, 2015. 1, 6
- [53] Mingqi Jiang, Saeed Khorram, and Li Fuxin. Comparing the decision-making mechanisms by transformers and cnns via explanation methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9546–9555, 2024. 3
- [54] Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. Knowledge distillation via route constrained optimization. In *Int. Conf. Comput. Vis.*, 2019. 13
- [55] Cheng Ju, Aurélien Bibaut, and Mark van der Laan. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, 45(15):2800–2818, 2018. 10
- [56] Hamid Karimi and Jiliang Tang. Decision boundary of deep neural networks: Challenges and opportunities. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, 2020. 2, 4
- [57] Gihyun Kim and Jong-Seok Lee. Analyzing adversarial robustness of vision transformers against spatial and spectral attacks. *arXiv preprint arXiv:2208.09602*, 2022. 2, 4
- [58] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Int. Conf. Mach. Learn.*, 2019. 2
- [59] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 16
- [60] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Adv. Neural Inform. Process. Syst.*, 2012. 1, 6, 7
- [61] Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003. 3, 5, 10
- [62] Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. Dataset condensation with contrastive signals. In *International Conference on Machine Learning (ICML)*, 2022. 27
- [63] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 7
- [64] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Adv. Neural Inform. Process. Syst.*, 2018. 2
- [65] Yuanchun Li, Ziqi Zhang, Bingyan Liu, Ziyue Yang, and Yunxin Liu. Modeldiff: Testing-based dnn similarity comparison for model reuse detection. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 139–151, 2021. 4
- [66] Hanxiao Liu, Andy Brock, Karen Simonyan, and Quoc Le. Evolving normalization-activation layers. In *Adv. Neural Inform. Process. Syst.*, 2020. 7
- [67] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to mlps. In *Adv. Neural Inform. Process. Syst.*, 2021. 7
- [68] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, 2021. 7, 16, 19

- [69] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 6, 7, 16
- [70] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Int. Conf. Learn. Represent.*, 2018. 2, 3, 4, 6, 17
- [71] Horia Mania, John Miller, Ludwig Schmidt, Moritz Hardt, and Benjamin Recht. Model similarity mitigates test set overuse. *Adv. Neural Inform. Process. Syst.*, 32, 2019. 1
- [72] Kristof Meding, Luca M Schulze Buschoff, Robert Geirhos, and Felix A Wichmann. Trivial or impossible—dichotomous data difficulty masks model differences (on imagenet and beyond). In *Int. Conf. Learn. Represent.*, 2022. 1, 3
- [73] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 13
- [74] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021. 3
- [75] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Adv. Neural Inform. Process. Syst.*, 2001. 8
- [76] M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 12
- [77] Chanwoo Park, Sangdoo Yun, and Sanghyuk Chun. A unified analysis of mixed sample data augmentation: A loss function perspective. In *Adv. Neural Inform. Process. Syst.*, 2022. 9
- [78] Namuk Park and Songkuk Kim. How do vision transformers work? In *Int. Conf. Learn. Represent.*, 2022. 2, 3, 4
- [79] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, 2019. 24
- [80] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 22
- [81] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollar. Designing network design spaces. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 7, 14
- [82] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021. 2, 3
- [83] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 6
- [84] Shahbaz Rezaei and Xin Liu. A target-agnostic attack on deep models: Exploiting security vulnerabilities of transfer learning. 2020. 2, 4

- [85] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 2015. 6
- [86] Luca Scimeca, Seong Joon Oh, Sanghyuk Chun, Michael Poli, and Sangdoo Yun. Which shortcut cues will dnns choose? a study from the parameter-space perspective. In *Int. Conf. Learn. Represent.*, 2022. 3, 5
- [87] Ali Shafahi, Amin Ghiasi, Furong Huang, and Tom Goldstein. Label smoothing and logit squeezing: a replacement for adversarial training? *arXiv preprint arXiv:1910.11585*, 2019. 9
- [88] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Int. Conf. Learn. Represent. Worksh.*, 2014. 3
- [89] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *Int. Conf. Mach. Learn. Worksh.*, 2017. 2, 3
- [90] Gowthami Somepalli, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richard Baraniuk, Micah Goldblum, and Tom Goldstein. Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 3, 5, 9, 12, 13, 15, 16
- [91] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *Int. Conf. Learn. Represent. Worksh.*, 2015. 2, 3
- [92] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 7, 19
- [93] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. In *Transactions on Machine Learning Research*, 2022. 7
- [94] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Int. Conf. Mach. Learn.*, 2017. 2, 3
- [95] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. 2014. 3
- [96] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 9
- [97] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Int. Conf. Mach. Learn.*, 2019. 7
- [98] Mingxing Tan and Quoc V Le. Mixconv: Mixed depthwise convolutional kernels. In *Brit. Mach. Vis. Conf.*, 2019. 7
- [99] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. In *Adv. Neural Inform. Process. Syst.*, 2021. 7
- [100] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Int. Conf. Mach. Learn.*, 2021. 7, 13, 14, 24, 26
- [101] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve,

- Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 7
- [102] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022. 24
- [103] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017. 2, 3, 5, 12, 16, 17
- [104] Asher Trockman and J. Zico Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022. 7
- [105] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *Int. Conf. Learn. Represent.*, 2019. 6, 18, 24
- [106] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, 2017. 1, 7
- [107] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 7, 19
- [108] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *Int. Conf. Mach. Learn.*, 2019. 24
- [109] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2020. 7
- [110] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3349–3364, 2020. 7
- [111] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 7, 19
- [112] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1924–1933, 2021. 6, 17
- [113] Futa Waseda, Sosuke Nishikawa, Trung-Nghia Le, Huy H Nguyen, and Isao Echizen. Closer look at the transferability of adversarial examples: How they fool different models differently. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 2, 3, 26
- [114] Ross Wightman. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, 2019. 2, 6, 7, 18
- [115] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. In *Adv. Neural Inform. Process. Syst. Worksh.*, 2021. 24
- [116] Yuxin Wu and Kaiming He. Group normalization. In *Eur. Conf. Comput. Vis.*, 2018. 1, 7
- [117] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 7

- [118] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 9, 24
- [119] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 7
- [120] Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *arXiv preprint arXiv:2301.07014*, 2023. 27
- [121] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 7
- [122] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Int. Conf. Comput. Vis.*, 2019. 9, 24
- [123] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 25
- [124] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Brit. Mach. Vis. Conf.*, 2016. 7
- [125] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Int. Conf. Learn. Represent.*, 2018. 9, 24
- [126] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2022. 7
- [127] Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. How does mixup help with robustness and generalization? In *Int. Conf. Learn. Represent.*, 2021. 9
- [128] Lei Zhang, Jie Zhang, Bowen Lei, Subhabrata Mukherjee, Xiang Pan, Bo Zhao, Caiwen Ding, Yao Li, and Dongkuan Xu. Accelerating dataset distillation via model augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11950–11959, 2023. 27
- [129] Qilong Zhang, Xiaodan Li, Yuefeng Chen, Jingkuan Song, Lianli Gao, Yuan He, and Hui Xue. Beyond imagenet attack: Towards crafting adversarial examples for black-box domains. *arXiv preprint arXiv:2201.11528*, 2022. 6, 17, 18
- [130] Richard Zhang. Making convolutional networks shift-invariant again. In *Int. Conf. Mach. Learn.*, 2019. 7
- [131] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, Sercan Ö Arik, and Tomas Pfister. Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 7
- [132] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, pages 12674–12685. PMLR, 2021. 27
- [133] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*, 2020. 27