The Principle of Diversity: Training Stronger Vision Transformers Calls for Reducing All Levels of Redundancy

Tianlong Chen[1], Zhenyu Zhang[1], Yu Cheng[2], Ahmed Awadallah[2], Zhangyang Wang[1]
[1]University of Texas at Austin, [2]Microsoft Research
{tianlong.chen, zhenyu.zhang, atlaswang}@utexas.edu, {yu.cheng, hassanam}@microsoft.com

Abstract

Vision transformers (ViTs) have gained increasing popularity as they are commonly believed to own higher modeling capacity and representation flexibility, than traditional convolutional networks. However, it is questionable whether such potential has been fully unleashed in practice, as the learned ViTs often suffer from over-smoothening, yielding likely redundant models. Recent works made preliminary attempts to identify and alleviate such redundancy, e.g., via regularizing embedding similarity or re-injecting convolution-like structures. However, a "head-to-toe assessment" regarding the extent of redundancy in ViTs, and how much we could gain by thoroughly mitigating such, has been absent for this field. This paper, for the first time, systematically studies the ubiquitous existence of redundancy at all three levels: patch embedding, attention map, and weight space. In view of them, we advocate a principle of diversity for training ViTs, by presenting corresponding regularizers that encourage the representation diversity and coverage at each of those levels, that enabling capturing more discriminative information. Extensive experiments on ImageNet with a number of ViT backbones validate the effectiveness of our proposals, largely eliminating the observed ViT redundancy and significantly boosting the model generalization. For example, our diversified DeiT obtains $0.70\% \sim 1.76\%$ *accuracy boosts on ImageNet with highly reduced similarity. Our codes are fully available in* https://github.com/VITA-Group/Diverse-ViT.

1. Introduction

Transformer [57], as the *de facto* neural architecture in natural language processing (NLP) [4, 19], recently revolutionizes modern computer vision applications such as image classification [21, 26, 53], object detection [5, 17, 74, 80], and image generation [10, 31, 46]. Rather than relying on convolution-like inductive bias, vision transformers [21] (ViTs) leverage the self-attention [57] to aggregate image patches across all spatial positions and model their global-range relationships, which are believed to improve model expressiveness and representation flexibility.

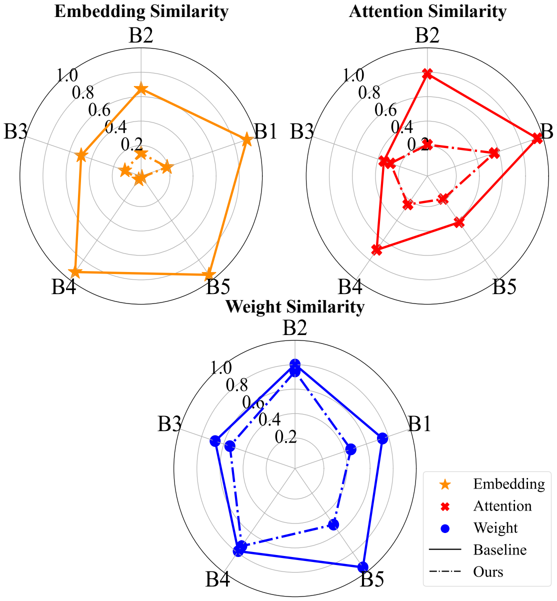

Figure 1. Relative similarity comparisons in embedding, attention, and weight spaces of DeiT-Small on ImageNet. The larger number indicates severer correlation/redundancy. B1∼B5 donate the blocks in the DeiT-Small model. *Cosine*, (normalized) *MSE*, 1 - (normalized) *reconstruction loss* are adopted to measure embedding, attention, and weight similarity. The former two are computed with $10,000$ images from the ImageNet training set without data augmentation, following the standard in [23].

Despite their promising potentials, the ViT training still suffers from considerable instability, especially when going deeper [23, 55]. One of the major reasons [23] is that the global information aggregation among all patches encourages their representations to become overly similar, causing substantially degraded discrimination ability. This phenomenon, known as over-smoothening, suggests a high degree of "redundancy" or ineffective usage of the ViT expressiveness and flexibility, and has been studied by a few prior arts [23,55,75,76]. Several initial attempts strive to fill the gap from different aspects. For example [23] proposes contrastive-based regularization to diversity patch embeddings, and [76] directly refines the self-attention maps via convolution-like aggregation to augment local patterns.

This paper aims to comprehensively study and mitigate the ViT redundancy issue. We first systematically demonstrate the ubiquitous existence of redundancy **at all three levels**: *patch embedding, attention map, and weight space*, for current state-of-the-art (SOTA) ViTs. That is even the case for those equipped with strong data augmentations (i.e., DeiT [54]) or sophisticated attention mechanisms (i.e., Swin [43]), e.g, as shown in Figure 1. In view of such collapse, we advocate a **principle of diversity** for training ViTs, by proposing corresponding regularizers that encourage the representation diversity and coverage at each of those levels, that unleashes the true discriminative power and representation flexibility of ViTs. We find each level's regularizers to provide generalization gains, and applying them altogether consistently yields superior performance. Our contributions lie in the following aspects:

- We provide the first comprehensive investigation of redundancy in ViTs by demonstrating its ubiquitous existence in all three levels of patch embeddings, attentions, and weights, across SOTA ViT models.
- For each of the three levels, we present diversity regularizers for training ViTs, which demonstrate complementary effects in eliminating redundancy, encouraging diversity, and enhancing generalization.
- We conduct extensive experiments with vanilla ViT, DeiT, and Swin transformer backbones on the ImageNet datasets, showing consistent and significant performance boost gains by addressing the tri-level redundancy issues with our proposed regularizers. Specifically, our proposals improve DeiT and Swin, by $0.70\% \sim 1.76\%$ and $0.15\% \sim 0.32\%$ accuracy.

2. Related Works

Vision transformer. Transformer [57] emerges from NLP applications with prevailing successes, motivating its adaptation to the computer vision scenarios [21]. ViTs encodes an image into a sequence of patches and feeds them to transformer encoders. Such self-attention based models get rid of common inductive biases, e.g., locality in convolutional neural networks (CNNs), and the global interactions among embedding grant ViTs a stronger learning capacity. ViT's empirical successes in various computer vision tasks include image classification [11, 21, 24], object detection [5, 17, 74, 80], segmentation [60, 63, 73], enhancement [9, 68], image generation [10, 31, 46], video and vision-language understanding [15, 36–38, 44, 51, 52, 71, 77, 78].

However, the global information aggregation of ViTs also results in over-smoothed and redundant representations [23, 76]. That makes the effective learning capacity of ViTs "collapsed", which prohibits ViTs from practically achieving higher capacity and sophisticated representations. Existing works have taken two angles: (i) (re-)injecting locality via convolution-like structures and fusing global and local contexts, for self-attention [2, 27, 32, 43, 55, 56, 64, 65, 67, 69, 76]; (ii) adopting patch-wise contrastive or mixing loss to boost diversity, for patch embeddings. Aiming to connect and expand those isolated efforts, our work is the first to target the full-scale redundancy in ViTs at embedding, attention, and weight levels. Note that our tri-level diversification framework is compatible with existing approaches by plugging them in the corresponding level. Detailed investigations are presented in Section 4.3.

Diversity regularization. Diversity constraints are designed to learn discriminative patterns for improved feature coverage and generalization [12, 23, 41]. Representative regularizers include cosine similarity-based [23], the margin or distance-based [13, 23, 33, 42, 49, 50], the hyperspherical uniformity-based [39–41], and the orthogonality-based [1,3,12,25,30,34,48,58,72]. Most of them are applid to CNNs, with [23] making the recent attempt in ViTs.

3. Methodology

3.1. Examining the Tri-Level Redundancy in ViTs

Preliminaries. Revisit that transformer architectures [21, 57] usually contain the multi-head self-attention modules (MHA) and feed-forward networks (FFN). In MHA, keys, queries, and values are linearly transformed for computing the attention heads, and then all the heads are aggregated by another linear transformation. FFNs are also built on two linear transformations with activations, as shown in Fig. 2.

Here, we use $\boldsymbol{W}^{\mathrm{MHA}}$ and $\boldsymbol{W}^{\mathrm{FFN}}$ to denote the weights in MHA and FFN modules, respectively. $\boldsymbol{A}$ represents the attention map (or the affinity matrix). It is calculated by $\boldsymbol{A} = \mathrm{softmax}(\alpha \boldsymbol{Q}\boldsymbol{K}^{\top})$, where $\boldsymbol{Q}$ is the query matrix, $\boldsymbol{K}$ is the key matrix, and α is a scale (typically $\frac{1}{\sqrt{d}}$ and d is the dimension of the keys and the queries). Let $\boldsymbol{e}^l = [\boldsymbol{e}^l_{\text{class}}, \boldsymbol{e}^l_1, \cdots, \boldsymbol{e}^l_n]$ be the feature embedding of layer l $(1 \leq l \leq \mathrm{L})$, where n is the total number of image patches. Without loss of generality, we take image recognition as an example. Then, the vision transformer is optimized by minimizing a classification loss $\mathcal{L}(\mathcal{C}(\boldsymbol{e}^{\mathrm{L}}_{\text{class}}), y)$, where $\mathcal{C}$ is the classification head and y is the label of input samples.

Redundancy in patch embeddings. We investigate the redundancy of the feature embedding by calculating the token-wise cosine similarity. It is depicted as follows:

$$\mathcal{R}^{s}_{\text{cosine}}(\boldsymbol{h}) := \frac{1}{n(n-1)} \sum_{i \neq j} \frac{|h_i^{\top} h_j|}{\|h_i\|_2 \|h_j\|_2}, \tag{1}$$

$$\mathcal{R}^{d}_{\text{cosine}}(\boldsymbol{h}^{l_1}, \boldsymbol{h}^{l_2}) := \frac{1}{n} \sum_{i} \frac{|h_i^{l_1^{\top}} h_i^{l_2}|}{\|h_i^{l_1}\|_2 \|h_i^{l_2}\|_2}, \tag{2}$$

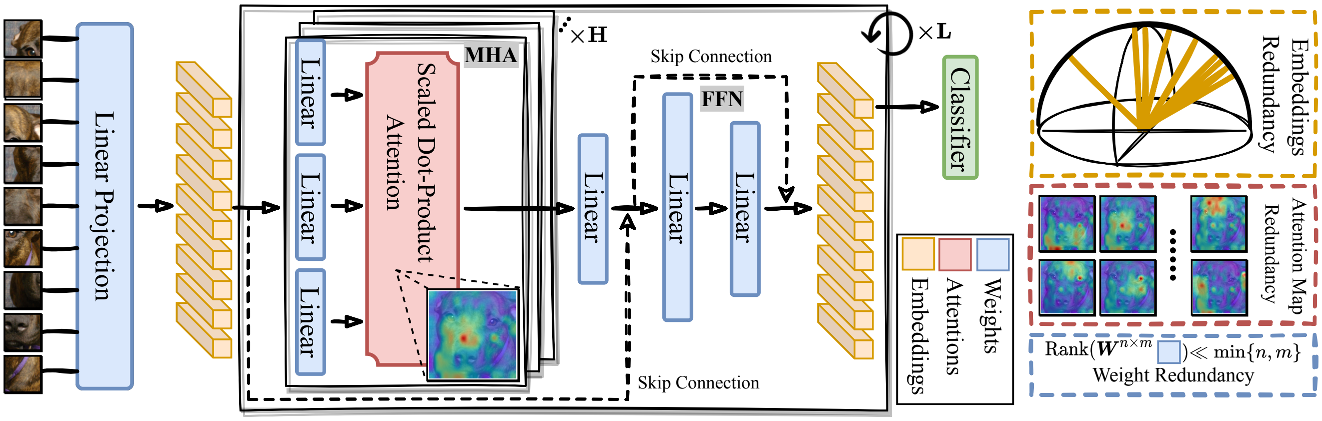

Figure 2. (*Left*) A overall pipeline of vision transformers [21,54]. Each image is divided into patches and transformed into embeddings via a linear projection layer. Then, embeddings are fed to the transformer encoder consists of MHA and FFN modules. Other operations like softmax and normalization are omitted here. (*Right*) An illustration of the redundancy of embedding, attention, and weight.

where $\boldsymbol{h}$ is the feature embedding $\boldsymbol{e} = [\boldsymbol{e}_{\text{class}}, \boldsymbol{e}_1, \cdots, \boldsymbol{e}_n]$ (superscript l is omitted for simplicity), and n is the total number of tokens.

Notably, $\mathcal{R}^s_{\text{cosine}}(\boldsymbol{h})$ and $\mathcal{R}^d_{\text{cosine}}(\boldsymbol{h}^{l_1}, \boldsymbol{h}^{l_2})$ denote the cosine similarity of the feature embedding within the same layer and across two different layers l_1, l_2, respectively. The larger cosine similarity suggests more redundancy. Intuitively, the within-layer redundancy hinders ViT from capturing different tokens' features; and the cross-layer redundancy hurts the learning capacity of ViTs since highly correlated representations actually collapse the effective depth of ViT to fewer or even single transformer layer.

Redundancy in attentions. We consider $\mathcal{R}_{\text{cosine}}(\boldsymbol{A})$ to measure the cosine similarity of attention maps within the same layer. Similarly, $\mathcal{R}_{\text{MSE}}(\boldsymbol{A}) := \frac{1}{n(n-1)}\sum_{i\neq j}\|A_i - A_j\|_2^2$ can also be used for the redundancy quantification. In contrast to these two metrics which show the similarity across attention heads, we further use the standard deviation statistics to indicate the element-wise variance within an attention head.

Redundancy in model weights. If the parameter space is highly redundant, then the weight matrix will fall approximately into a low-rank parameters subspace. Thus, we use the reconstruction error to depict the weight redundancy:

$$\mathcal{R}_{\text{PCA}}(\boldsymbol{W}) := \|\boldsymbol{W} - \tilde{\boldsymbol{W}}\|_2^2 \qquad (3)$$

where $\tilde{\boldsymbol{W}}$ is the reconstructed weight matrix by the principal component analysis (PCA) with the top-k principal components. Given a fixed reconstruction error, the larger k implies better diversity. In other words, given k, the larger reconstruction error means less weight redundancy. [12,41] also dissect the weight redundancy from the view of rank.

3.2. Eliminating the Tri-Level Redundancy in ViTs

To mitigate the observed redundancy, we introduce three groups of regularization to encourage the diversity of $i)$ learned feature embeddings; $ii)$ attention maps; $iii)$ model weights in the training of vision transformers.

Patch embedding diversity. To diversify patch feature embeddings, we use the cosine angle regularization $\mathcal{R}^s_{\text{cosine}}(\boldsymbol{e})$ and $\mathcal{R}^d_{\text{cosine}}(\boldsymbol{e}^{l_1}, \boldsymbol{e}^{l_2})$ to constrain within-layer and cross-layer embedding, respectively. Similar methods are leveraged to obtain diversified representations in vision [23], language [22], and graph [8] scenarios. Meanwhile, we adopt the contrastive regularization $\mathcal{R}^d_{\text{contrastive}}(\boldsymbol{e}^{l_1}, \boldsymbol{e}^{l_2})$ to boost cross-layer embedding diversity, which is presented as follows:

$$\mathcal{R}_{\text{contrastive}}(\boldsymbol{e}^{l_1}, \boldsymbol{e}^{l_2}) := -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp(e_i^{l_1\top} e_i^{l_2})}{\exp(e_i^{l_1\top} e_i^{l_2}) + \exp(e_i^{l_1\top}(\frac{1}{n-1}\sum_{j\neq i} e_j^{l_2}))}, \qquad (4)$$

where l_1 and l_2 are two different layer indexes. Note that the contrasitve regularizer is not applicable for the within-layer embedding diversification since the lack of positive pairs.

▷ *Rationale.* As pointed out by [23, 45], the cosine angle regularization can function like minimizing the upper bound of the largest eigenvalue of patch embedding $\boldsymbol{e}$, hence bringing improvements of expressiveness [23] and diversity to learned representations. For contrastive regularization, it pulls embeddings corresponding to the same patch together and simultaneously pushes apart embeddings belonging to different patches, reducing the feature correlation between different layers. As a result, it enables to learn separable patch embedding and maintain tolerance to semantically similar patches [59, 61], improving the representation qualities and the ViT performance.

Attention diversity. In the same way, the cosine regularization $\mathcal{R}^s_{\text{cosine}}(\boldsymbol{A})$ can be applied to remove the redundancy of attention, where $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_{\text{H}}]$ and H is the number of attention heads within one layer. Inspired by the orthogonality regularization's empirical effectiveness in vision [12, 35, 47] and language tasks [72], we investigate it under the context of ViTs. We adopt the canonical soft orthogonal regularization (SO) [3] as follows:

$$\mathcal{R}_{\text{SO}}(\boldsymbol{A}) := \|\boldsymbol{A}^\top \boldsymbol{A} - \boldsymbol{I}\|^2_{\text{F}}, \tag{5}$$

where $\|\cdot\|_{\text{F}}$ is the Frobenius norm and $\boldsymbol{I}$ is the identity matrix sharing the same size as $\boldsymbol{A}^\top \boldsymbol{A}$.

We also try an alternative Conditional number orthogonal regularization (CNO) [12] as follows:

$$\mathcal{R}_{\text{CNO}}(\boldsymbol{A}) = \|\lambda_1(\boldsymbol{A}^\top \boldsymbol{A}) - \lambda_2(\boldsymbol{A}^\top \boldsymbol{A})\|^2. \tag{6}$$

It enforces the orthogonality via directly regularizing the conditional number $\kappa = \frac{\lambda_1}{\lambda_2}$ to 1, where λ_1 and λ_2 are the largest and smallest eigenvalues of the target matrix $\boldsymbol{A}^\top \boldsymbol{A}$. To make it computationally more tractable and stable, we alternatively constrain the difference between λ_1 and λ_2.
▷ *Rationale.* These regularizations (i.e., SO and CNO) encourage diverse attention maps by constraining them to be orthogonal with each other, which actually upper-bounds the Lipschitz constant of learned function mappings [72], leading to robust and informative representations. As illustrated in [72], introducing an orthogonal diversity regularizer to the attention map also stabilizes the transformer training and boosts its generalization on NLP tasks.

Weight diversity. Similarly, the orthogonality regularization, e.g., $\mathcal{R}_{\text{CNO}}(\boldsymbol{W})$, can be easily plugged in and promote the diversity in ViT's weight space. Compared to orthogonality, hyperspherical uniformity is another more general diversity regularization demonstrated in [41]. Although it has been explored in CNNs, its study in ViTs has been absent so far. We study the minimum hyperspherical separation (MHS) regularizer, which maximizes the separation distance (or the smallest pairwise distance) as follows:

$$\max_{\{\hat{\boldsymbol{w}}_1,\cdots,\hat{\boldsymbol{w}}_m\}\in\mathbb{S}^{t-1}}\{\mathcal{R}_{\text{MHS}}(\hat{\boldsymbol{W}}) := \min_{i\neq j}\rho(\hat{\boldsymbol{w}}_i, \hat{\boldsymbol{w}}_j)\}, \tag{7}$$

where $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2 \cdots, \boldsymbol{w}_m]$, $\hat{\boldsymbol{w}}_i := \frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|}$ is the ith weight vector projected onto a unit hypersphere $\mathbb{S}^{t-1} := \{\hat{\boldsymbol{w}} \in \mathbb{R}^t | \|\hat{\boldsymbol{w}}\| = 1\}$, $\rho(\cdot,\cdot)$ is the geodesic distance on the unit hypersphere. As indicated in Equation 7, it is formulated as a max-min optimization and we solve it with alternative gradient ascent/descent.

Furthermore, we examine another maximum gram determinant (MGD) regularizer $\mathcal{R}_{\text{MGD}}(\hat{\boldsymbol{W}})$ as follows:

$$\max_{\{\hat{\boldsymbol{w}}_1,\cdots,\hat{\boldsymbol{w}}_m\}\in\mathbb{S}^{t-1}}\text{logdet}\big(\boldsymbol{G} := (\mathcal{K}(\hat{\boldsymbol{w}}_i, \hat{\boldsymbol{w}}_j))^m_{i,j=1}\big), \tag{8}$$

where $\det(\boldsymbol{G})$ is the determinant of the kernel gram matrix $\boldsymbol{G} \in \mathbb{R}^{m\times m}$ and $\mathcal{K}(\boldsymbol{u}, \boldsymbol{v}) := \exp(-\sum^t_{i=1}\epsilon^2(u_i - v_i)^2)$ denotes the kernel function with a scale $\epsilon > 0$. By maximizing the $\det(\boldsymbol{G})$ of weights $\hat{\boldsymbol{W}}$, MGD forces weight vectors to uniformly dispersed over the hypersphere.

▷ *Rationale.* As demonstrated in [39–41], the hyperspherical uniformity regularizations (i.e., MHS and MGD) characterizes the diversity of vectors on a unit hypersphere, which encodes a strong inductive bias with relational information. We believe it to benefit ViT training from two perspectives [41]: (i) eliminating weight redundancy and improving the representative capacity; (ii) learning better optimization and generalization by reducing the spurious local minima, evidenced in [39–41, 66].

4. Experiment

Implementation details. We conduct extensive experiments on the ImageNet-1k [18] dataset with ViT [21], DeiT [53] and Swin transformers [43]. All the hyperparameters of our introduced diversity regularizations are carefully tuned by a grid search, and the best configurations are provided in Section A2. Tesla V100-SXM2-32GB GPUs are used as our computing resources. Specifically, each experiment is ran with 8 V100s for $1 \sim 4$ days.

For vanilla ViT models, we consider two architectures with 12 layers, i.e., ViT-Small and ViT-Base, which contains 6 and 12 heads for the multi-head self-attention block in each layer, respectively. We train each model for 300 epochs with a batch size of 4096. An AdamW optimizer is adopted with 0.3 weight decay, and the learning rate starts from 4×10^{-3} with 4 epochs for warm-up and decays by a cosine annealing schedule. We maintain all training settings the same as the original ones in [21].

For DeiT architectures, we choose DeiT-Small, DeiT-Small24, and DeiT-Base. Specifically, both DeiT-Small and DeiT-Base contain 12 layers, while DeiT-Small24 has 24 layers. In each layer, DeiT-Small and DeiT-Small24 have 6 heads for the self-attention module, and DeiT-Base has 12 heads. Following [53], we train the models for 300 epochs with the batch size of 1024. We use an AdamW optimizer of 0.05 weight decay. The initial learning rate is 1×10^{-3} with 5 epochs of warm-up, which reduces by a cosine annealing schedule. More details about data augmentation and other training tricks can be found in [53],

For Swin transformer, we start from the official Swin-Small and Swin-Base pre-trained models and then fine-tune them for another 30 epochs, in which we use a constant 1×10^{-5} learning rate, 1×10^{-8} weight decay and a batch size of 1024 [23]. We keep other training details the same as [43] and compare the fine-tuning performance with/without diversity regularizations.

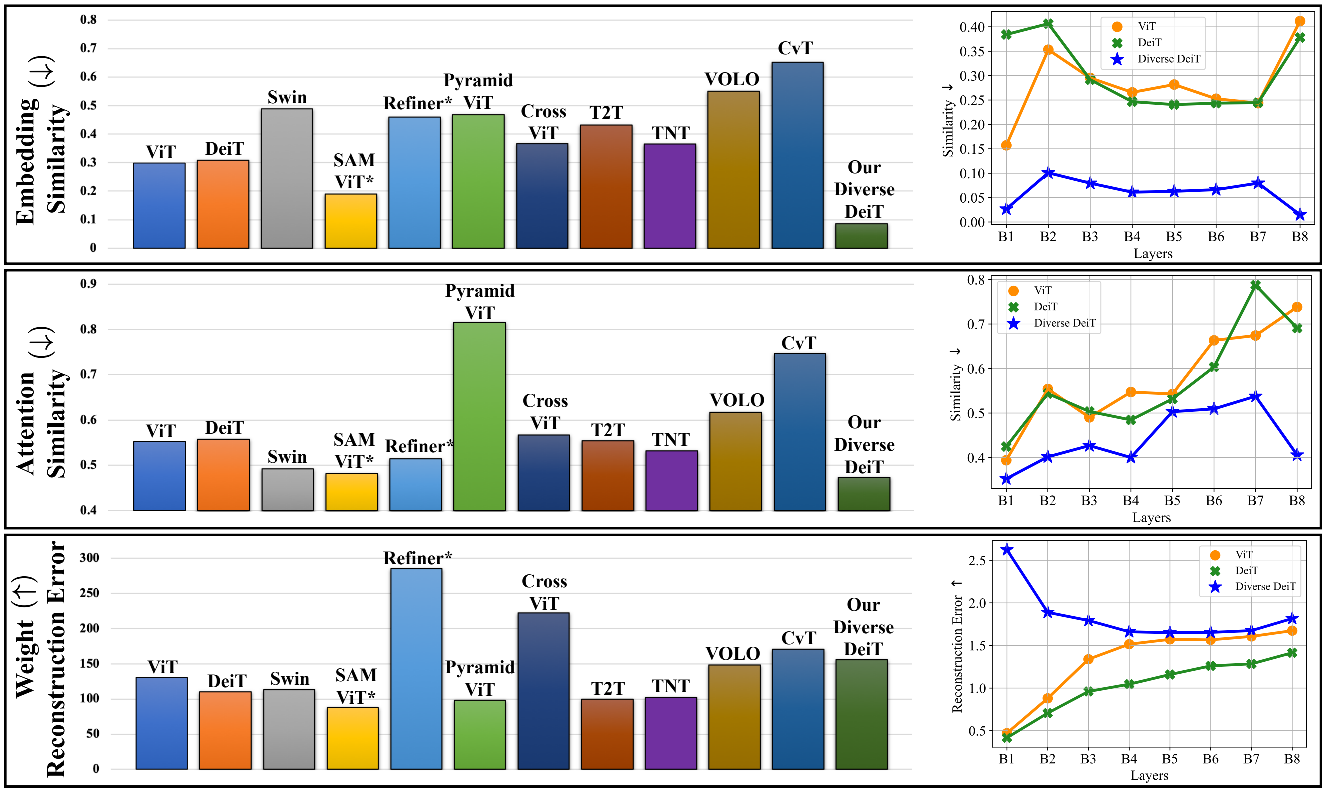

Figure 3. (*Left*) Redundancy comparisons in embedding, attention, and weight spaces of ViT [21], DeiT [53], Swin [21], SAM-ViT [14], Refiner [76], Pyramid-ViT [62], Cross-ViT [7], T2T [69], TNT [27], VOLO [70], CvT [64], and our diverse DeiT on ImageNet. We use publicly available pre-trained models for benchmarking their all levels of redundancy. For a fair comparison, most of selected pre-trained transformers share similar parameter counts, i.e., 19M $\sim$ 27M, while even the smallest released SAM-ViT and Refiner have 83M and 78M. ↑/↓ denote that the larger/smaller number indicates better diversity. *Cosine*, *cosine*, (normalized) *reconstruction error* are adopted to measure embedding, attention, and weight similarity. The former two are computed with $10,000$ sub-sampled images from the ImageNet training set without data augmentation, following the standard in [23]. (*Right*) The layer-wise similarity/reconstruction error of ViT, DeiT, and Swin, which B1$\sim$B8 is the corresponding transformer blocks (or layers).

4.1. Redundancy in Current ViTs

In this section, we conduct a thorough investigation to reveal the broadly existing redundancy in ViT's patch embedding, attention map and weight spaces. Specifically, 11 current SOTA ViTs and our diversified DeiT are examined on ImageNet, which can be grouped into four categories according to their proposed approaches on ViT [21]: (i) improving training techniques such as data augmentation in DeiT [53] and flatness-aware regularizer in SAM-ViT [14]; (ii) introducing convolution layers or crafting convolution-like operations like CvT [64] and Refiner [76] (termed as CNN + ViT); (iii) designing hierarchical structures to capture multi-scale information such as Swin [21], Pyramid-ViT [62], and Cross-ViT [7]; (iv) exploring finer-level features or encoding local contexts into featuring like T2T [69], TNT [27], and VOLO [70]. From Figure 3, we observe that:

❶ At the patch embedding level, most of current SOTA ViTs exclude SAM-ViT have amplified embedding redundancy compared to vanilla ViT. It seems to suggest that the flatness-aware regularizer can enable ViT to generate more diverse embeddings. Furthermore, our diverse DeiT presents a consistent and substantial reduction on both layer-wise and overall embedding similarity, outperforming all other ViT variants.

❷ At the attention level, our diverse DeiT again achieves superior diverse attention maps with the least redundancy. In addition, self-attention with shifted windows in Swin [21] and convolution-like aggregation in Refiner [76] effectively reduce the attention map correlation. Also, flatness-aware SAM-ViT brings more diversity across attention heads.

❸ At the weight level, with similar parameter counts, Cross-ViT [7] with a dual-branch structure for extracting multi-scale features has the largest weight reconstruction error, implying more weight diversity. Besides, our diverse DeiT still consistently outperforms its vanilla counterpart as shown in Figure 3 (*right*), which validates our proposal's effectiveness in eliminating the weight space redundancy.

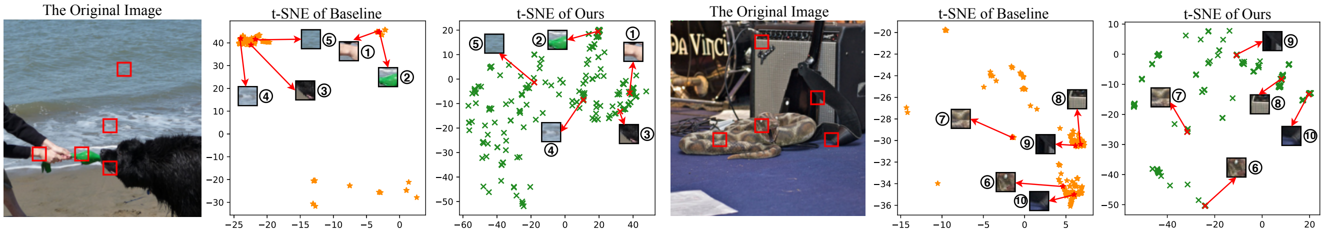

Figure 4. t-SNE visualization for patch embedding (i.e., 196 tokens) of images randomly sampled from ImageNet. The patch embedding in {2nd, 5th}/{3rd, 6th} columns are generated from the Deit baseline and our diversified variant, respectively. Patches ①∼⑩ are corresponding to the red box in the original images.

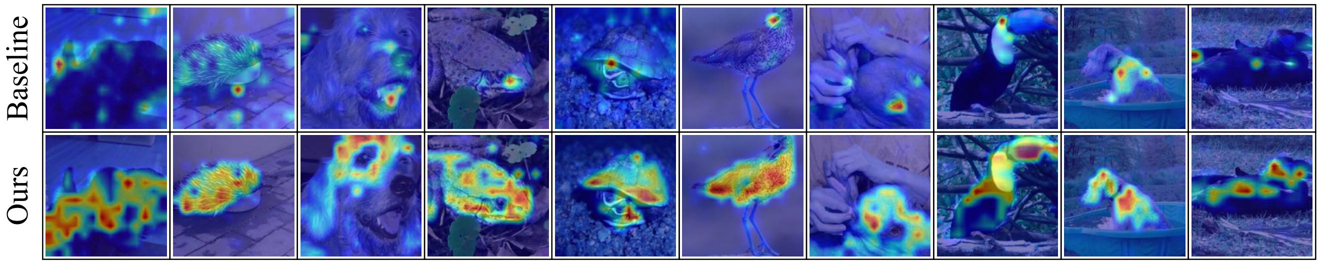

Figure 5. Attention visualizations of DeiT on ImageNet. Heatmaps in the 1st and 2nd rows are from baseline and our diversified variant. Tools of interpreting ViTs in [6] are adopted to produce visualization.

Table 1. Top-1 accuracies (%) of ViTs with or without all-level diversification on ImageNet. ↑ indicates the performance gains from our diversity regularizers compared with its vanilla version. All ViT and DeiT transformers are trained from scratch for 300 epochs with the default setup in [21] and [53]. Swin Transformers are fine-tuned from released checkpoints from [43] for 30 epochs.

Settings & Methods		Size	Accuracy
(CNN)	ResNet-152 [28]	230M	78.10
	DenseNet-201 [29]	77M	77.60
(CNN + ViT)	CVT-21 [64]	32M	82.50
(ViT + DiversePatch♣)	DeiT-Small [23]	22M	80.43
(ViT + DLA♠)	Refiner [76]	86M	81.20
(Vanilla ViT 12 layers)	ViT-Small [21]	22M	76.54
	ViT-Small + Ours	22M	78.60 (↑ 2.04)
	ViT-Base [21]	86M	77.90
	ViT-Base + Ours	86M	79.96 (↑ 2.06)
(DeiT 12 layers)	DeiT-Small [53]	22M	79.78
	DeiT-Small + Ours	22M	80.61 (↑ 0.83)
	DeiT-Base [53]	86M	80.98
	DeiT-Base + Ours	86M	81.68 (↑ 0.70)
(Swin 12 layers)	Swin-Small [43]	50M	83.18
	Swin-Small + Ours	50M	83.33 (↑ 0.15)
	Swin-Base [43]	88M	83.40
	Swin-Base + Ours	88M	83.72 (↑ 0.32)
(DeiT 24 layers)	DeiT-Small24 [53]	43M	80.03
	DeiT-Small24 + Ours	43M	81.79 (↑ 1.76)

♣ DiversePatch is the diversification on the token embedding level in [23].
♠ DLA is the diversification on the attention level in [76].

4.2. Enhanced ViTs by Introducing Diversity

Superior generalization of our proposal. In this section, we demonstrate that improving diversity in ViT training achieves superior generalization. In specific, our experiments consider six representative transformer backbones, i.e., vanilla ViT-Small/Base, DeiT-Small/Base and Swin-Small/Base. As shown in Table 1, several consistent observations can be drawn: ❶ Compared to ViT, DeiT[1] and Swin baselines, our diversified variants obtain $\sim 2\%$, $0.70\% \sim 1.76\%$, and $0.15\% \sim 0.32\%$ accuracy boosts respectively, which evidence the effectiveness of our tri-level diversity regularization. ❷ Vanilla ViTs tend to benefit more from our diversity-ware training, by $3 \sim 7$ times improvements on DeiT and Swin. A possible explanation is that the data augmentation in DeiT and self-attention with shifted windows in Swin have already injected a certain level of diversity, as suggested by the reduced attention and weight correlation in Figure 3. ❸ Compared with existing methods on the same ViT backbone, our diverse ViTs gain 0.18% and 0.48% accuracy improvements over DiversePatch [23] and DLA [76] respectively, which indicate the necessity of reducing all levels of redundancy. ❹ Deep ViTs receive more benefits from diversity regularized training. Particularly, DeiT-Small24 achieves a 1.76% accuracy increase, while its 12-layer variants have 0.83% accuracy gains. It's within expectation since deeper transformers usually suffer harsher over-smoothing and representation redundancy [20, 79], leaving more potential advancement for diversification. Similar conclusions can be observed in [23].

Effectively reduced redundancy. In order to further validate the superiority of our diverse ViTs, we provide extensive qualitative and quantitative visualizations for all levels of patch embedding, attention maps, and model weights.

[1] We disable the repeated augmentation in DeiT-Base's [54] training schemes due to the well-known loss NAN issue of the original implementation (https://github.com/facebookresearch/deit/issues/29), which will result in slight performance degradation.

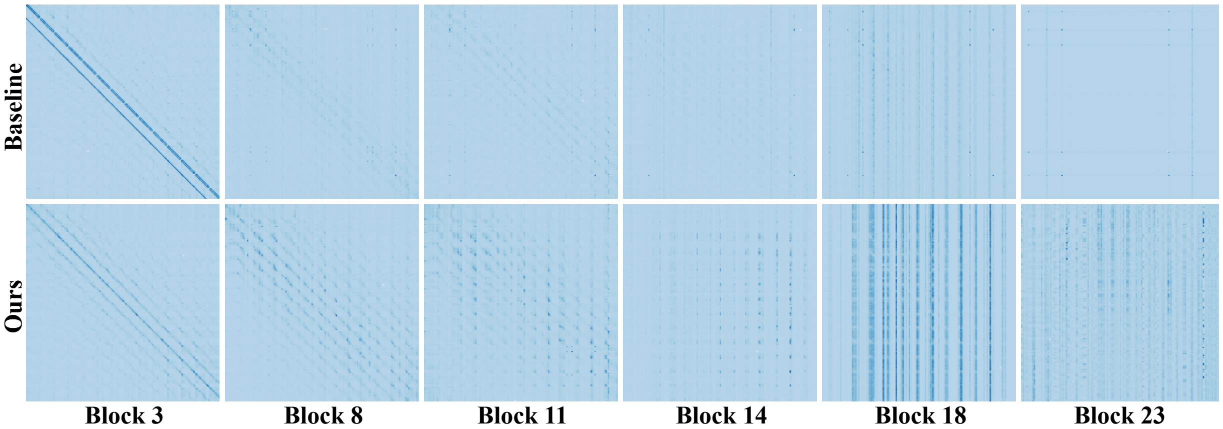

Figure 6. Comparison of the attention map (i.e., affinity matrix $\boldsymbol{A}$) between DeiT baseline (1st row) and our diversified variant (2nd row).

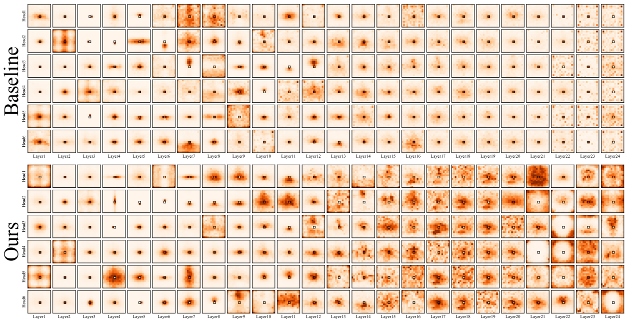

Figure 7. Attention probabilities for DeiT baseline (1st row) and our diversified variant (2nd row) with 24 layers (columns) and 6 heads (rows), visualized by [16]. Attention maps are averaged over 100 test images from ImageNet. The black square is the query pixel. Zoom-in for better visibility.

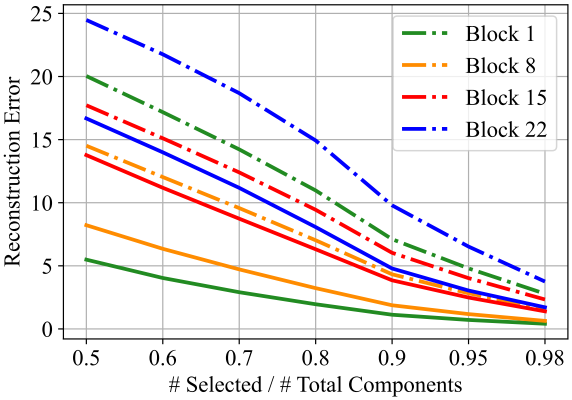

Figure 8. The weight reconstruction errors over the number of selected principle components. The dash and solid lines represent our diverse ViT and its vanilla version, respectively. The smaller error suggest a larger redundancy in the original weight space.

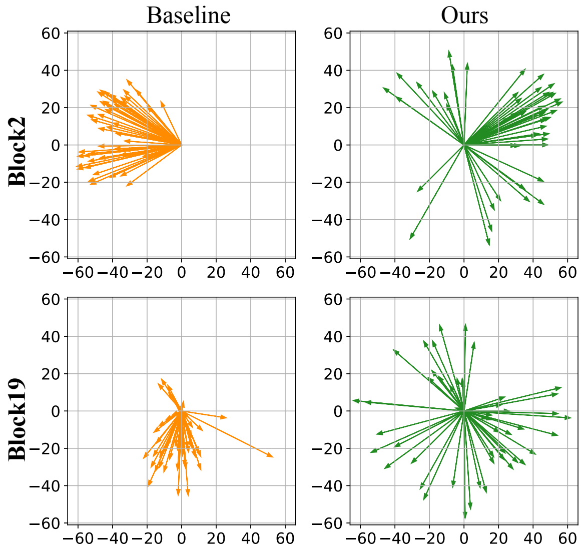

Figure 9. t-SNE of weight principle components (top-50) from Deit baseline (1st col.) and our diversified variant (2nd col.).

❶ *The patch embedding level.* As presented in Figure 3 (*right*) and Figure A10, our diverse DeiT obtains significantly reduced similarity for both within-layer and cross-layer embedding, compared to the baseline DeiT. Figure 4 visualizes the all patch embedding of 196 tokens for randomly selected images from ImageNet. Our methods show a much diverse embedding distribution and improved discriminative power. Specifically, the patch ① and ② from the left image are belong two distinct objects (i.e., human hands and dog toys) and also have totally different visual contents. However, the baseline ViT produces highly correlated patch embedding (almost fully overlapped points in Figure 4) and fails to distinguish them. Fortunately, these embedding redundancy and semantic ambiguity are well addressed by our diversity-aware ViT training. Similar observations can be drawn from patch pairs (③,④), (③,⑤), (⑥,⑩), and (⑧,⑨).

❷ *The attention map level.* From Figure 3 (right), Figure A11, and Figure A12, our proposals consistently reach an enhanced diversity in terms of within-layer/cross-layer attention heads and the standard deviation within attention maps. Furthermore, we offer illustrative visualizations to show the improved representation flexibility of ViTs. Figure 5 is generated by an advanced ViT interpretable approach [6]. The heatmaps from our methods display more diverse and accurate focuses. Specifically, ours concentrate attention on more focus points that spread across the whole

Table 2. Ablation on the different combination of diversity regularizers. Experiments are conducted on DeiT-Small with ImageNet, and top-1 testing accuracy (%) is reported.

Ablation on ↓	Other - Mixing Loss	Within-layer Embedding	Cross-layer Embedding	Attention Maps	Weight	Accuracy
none: DeiT-Small	✗	✗	✗	✗	✗	79.78
Diversity Regularization	✓	✗	✗	✗	✗	79.99
	✓	✓	✗	✗	✗	80.12
	✓	✗	✓	✗	✗	80.13
	✓	✓	✓	✗	✗	80.43
	✓	✓	✓	✓	✗	80.53
All levels diversity	✓	✓	✓	✓	✓	**80.61**

objects like heads, legs, and main bodies. Figure 6 visualizes the attention affinity matrices $\boldsymbol{A}$. We find that the attention maps from our diverse ViT training become less uniform compared to the baseline ViT, and show stronger local patterns, which provides another possible interpretation of our beneficial diversity regularization. In the end, Figure 7 dissects the fine-grained attention behaviors of certain query pixels, in which our approach shows more heterogeneous attention responses, especially in the ViT's later layers.

❸ *The weight level.* As shown in Fig. 3 (right) and Fig. 8, given a fixed number of selected principal components, the diversified ViTs have significantly larger reconstruction errors across almost all transformer layers, implying eliminated weight redundancy and improved representative power. Consistent observations are demonstrated in Fig. 9, where our diverse ViT can widely span its weight principle components on the 2-dim space.

4.3. Ablation Study

Multi-level v.s. single-level diversity. To verify the effects of diversity regularizers at different levels, we conduct an incremental evaluation on DeiT-Small with ImageNet. Achieved results are included in Table 2. We observe that: (i) Diversifying patch embedding brings the most accuracy benefits ($\sim 0.4\%$) and the other two levels contribute similarly ($\sim 0.1\%$). It is worthy of mentioning that if adopting within-layer or cross-layer embedding diversity regularization only gains $\sim 0.1\%$ accuracy while combining them leads to extra performance boosts. (ii) Reducing tri-level redundancy in patch embedding, attention maps, and weights establishes superior performance, which validates the effectiveness of multi-level diversity compared to single-level diversity. (iii) The previous useful data-level diversification, i.e., mixing loss from [23], can be easily plugged in our training framework and plays a complementary role in improving ViT's generalization ability. More details about the mixing loss are referred to Section A1.

How to choose different diversity regularizations. As discussed in Section 3, there are various regularizer options for different levels in our diverse ViTs. We implement comprehensive comparisons among these six kinds of regularizations. N.A. denotes that it is not applicable in certain levels because of unmatched design motivations. From Table 3, first, we find that constraining layerwise attention maps in a contrastive manner might be too aggressive to hurt the performance. While patch embedding is more amenable to diversification, all examined regularizers at the embedding level boost the ViT's performance. Secondly, compared to strict orthogonality regularization, the general hypersphere uniformity is more preferred by the weight level of ViTs. Lastly, at the attention level, SO outperforms all other choices. A possible reason is that orthogonal attention maps not only enjoy an unleashed representation flexibility but also stabilize and smoothen ViT training due to reduced Lipschitz constant, as suggested by [72].

Table 3. Ablation on different categories of diversity regularization. Experiments are conducted on DeiT-Small with ImageNet, and top-1 testing accuracy (%) is reported.

DeiT-Small (79.78)		Weight	Attention	Embedding
Similarity Regularization	Consine	N.A.	79.95	**80.20**
	Contrastive		69.98	80.11
Uniformity Regularization	MHS	**80.05**	N.A.	80.10
	MGD	79.96		79.92
Orthogonality Regularization	CondO	80.01	79.90	79.96
	SO	79.80	**80.03**	80.09

Training time analyses. Since the computation of our regularizers is quite cheap (e.g., 2 steps power iteration), the extra time overhead is moderate, as indicated below of per epoch training time on eight Quadro RTX 6000 GPUs.

Settings	Baseline	+ Embedding Reg.	+ Attention Reg.	+ Weight Reg.	+ All Reg.
Training Time (s)	582	590	595	677	695

5. Conclusion and Broader Impact

In this paper, we, for the first time, systematically reveal the broad existence of redundancy at all token embedding, attention map, and weight levels in vision transformers, which limits the ViT's expressiveness and flexibility. We address this issue following the principle of diversity during ViT training. Comprehensive experiments on ImageNet across diverse ViT backbones demonstrate that equipped diversity regularizers effectively eliminate the redundant representation and lead to superior generalization.

For the limitation of this work, we only focus on vision transformers rather than general transformers in natural language processing, which we leave for future works. Meanwhile, although our paper is scientific in nature, it might amplify the existing societal risk of applying ViTs since we have no control of anyone who can get access to our improved training algorithms. A potential solution is to issue licenses and limit the abuse.

References

[1] Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *International Conference on Machine Learning*, pages 1120–1128. PMLR, 2016. 2

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 2

[3] Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep cnns? *arXiv preprint arXiv:1810.09102*, 2018. 2, 4

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1, 2

[6] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. 6, 7

[7] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *arXiv preprint arXiv:2103.14899*, 2021. 5

[8] Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3438–3445, 2020. 3

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

[10] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 13–18 Jul 2020. 1, 2

[11] Tianlong Chen, Yu Cheng, Zhe Gan, Lu Yuan, Lei Zhang, and Zhangyang Wang. Chasing sparsity in vision transformers: An end-to-end exploration. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[12] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abdnet: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2, 3, 4

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[14] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pretraining or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021. 5

[15] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020. 2

[16] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations*, 2020. 7

[17] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. *arXiv preprint arXiv:2011.09094*, 2020. 1, 2

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[20] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. *arXiv preprint arXiv:2103.03404*, 2021. 6

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 3, 4, 5, 6

[22] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*, 2019. 3

[23] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu. Vision transformers with patch diversification, 2021. 1, 2, 3, 4, 5, 6, 8, A12

[24] Jianyuan Guo, Kai Han, Han Wu, Chang Xu, Yehui Tang, Chunjing Xu, and Yunhe Wang. Cmt: Convolutional neural networks meet vision transformers. *arXiv preprint arXiv:2107.06263*, 2021. 2

[25] Chuchu Han, Ruochen Zheng, Changxin Gao, and Nong Sang. Complementation-reinforced attention network for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3433–3445, 2020. 2

[26] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020. 1

[27] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021. 2, 5

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[29] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 6

[30] Lei Huang, Xianglong Liu, Bo Lang, Adams Wei Yu, Yongliang Wang, and Bo Li. Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[31] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong gan. *arXiv preprint arXiv:2102.07074*, 2021. 1, 2

[32] Zihang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. *arXiv preprint arXiv:2104.10858*, 2021. 2

[33] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 2

[34] Mingu Lee, Jinkyu Lee, Hye Jin Jang, Byeonggeun Kim, Wonil Chang, and Kyuwoong Hwang. Orthogonality constrained multi-head attention for keyword spotting. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 86–92. IEEE, 2019. 2

[35] José Lezama, Qiang Qiu, Pablo Musé, and Guillermo Sapiro. Ole: Orthogonal low-rank embedding-a plug and play geometric loss for deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8109–8118, 2018. 4

[36] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020. 2

[37] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[38] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2

[39] Rongmei Lin, Weiyang Liu, Zhen Liu, Chen Feng, Zhiding Yu, James M Rehg, Li Xiong, and Le Song. Regularizing neural networks via minimizing hyperspherical energy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6917–6927, 2020. 2, 4

[40] Weiyang Liu, Rongmei Lin, Zhen Liu, Lixin Liu, Zhiding Yu, Bo Dai, and Le Song. Learning towards minimum hyperspherical energy. *arXiv preprint arXiv:1805.09298*, 2018. 2, 4

[41] Weiyang Liu, Rongmei Lin, Zhen Liu, Li Xiong, Bernhard Schölkopf, and Adrian Weller. Learning with hyperspherical uniformity. In *International Conference On Artificial Intelligence and Statistics*, pages 1180–1188. PMLR, 2021. 2, 3, 4

[42] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7, 2016. 2

[43] 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. *arXiv preprint arXiv:2103.14030*, 2021. 2, 4, 6

[44] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 2

[45] Jorma Kaarlo Merikoski. On the trace and the sum of elements of a matrix. *Linear algebra and its applications*, 1984. 3

[46] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018. 1, 2

[47] Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, and Fahad Shahbaz Khan. Orthogonal projection loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12333–12343, 2021. 4

[48] Pau Rodríguez, Jordi Gonzalez, Guillem Cucurull, Josep M Gonfaus, and Xavier Roca. Regularizing cnns with locally constrained decorrelations. *arXiv preprint arXiv:1611.01967*, 2016. 2

[49] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[50] Kihyuk Sohn. Improved deep metric learning with multiclass n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 2

[51] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[52] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[53] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 1, 4, 5, 6

[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 2, 3, 6

[55] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. *arXiv preprint arXiv:2103.17239*, 2021. 1, 2

[56] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. *arXiv preprint arXiv:2102.10662*, 2021. 2

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1, 2

[58] Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, and Chris Pal. On orthogonality and learning recurrent networks with long term dependencies. In *International Conference on Machine Learning*, pages 3570–3578. PMLR, 2017. 2

[59] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2021. 3

[60] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. *arXiv preprint arXiv:2012.00759*, 2020. 2

[61] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 3

[62] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021. 5

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

[64] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021. 2, 5, 6

[65] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 2

[66] Bo Xie, Yingyu Liang, and Le Song. Diverse neural network learns true target functions. In *Artificial Intelligence and Statistics*, pages 1216–1224. PMLR, 2017. 4

[67] Jiangtao Xie, Ruiren Zeng, Qilong Wang, Ziqi Zhou, and Peihua Li. So-vit: Mind visual tokens for vision transformer. *arXiv preprint arXiv:2104.10935*, 2021. 2

[68] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020. 2

[69] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 2, 5

[70] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *arXiv preprint arXiv:2106.13112*, 2021. 5

[71] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *European Conference on Computer Vision*, pages 528–543. Springer, 2020. 2

[72] Aston Zhang, Alvin Chan, Yi Tay, Jie Fu, Shuohang Wang, Shuai Zhang, Huajie Shao, Shuochao Yao, and Roy Ka-Wei Lee. On orthogonality constraints for transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 375–382, Online, Aug. 2021. Association for Computational Linguistics. 2, 4, 8

[73] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. *arXiv preprint arXiv:2012.09164*, 2020. 2

[74] Minghang Zheng, Peng Gao, Xiaogang Wang, Hongsheng Li, and Hao Dong. End-to-end object detection with adaptive clustering transformer. *arXiv preprint arXiv:2011.09315*, 2020. 1, 2

[75] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021. 1

[76] Daquan Zhou, Yujun Shi, Bingyi Kang, Weihao Yu, Zihang Jiang, Yuan Li, Xiaojie Jin, Qibin Hou, and Jiashi Feng. Refiner: Refining self-attention for vision transformers. *arXiv preprint arXiv:2106.03714*, 2021. 1, 2, 5, 6

[77] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020. 2

[78] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. 2

[79] Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. Bert loses patience: Fast and robust inference with early exit. *arXiv preprint arXiv:2006.04152*, 2020. 6

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