A1. More Technical Details

Token-wise mixing loss. As an effective data-level diversification from [23], it mixes the input patches from two different images and leverages an additional shared classifier to output patch embedding for the classification of each patch. It can be described as follows:

$$\mathcal{R}_{\text{mixing}}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{\text{XE}}(g(\boldsymbol{e}_i^{\text{L}}), y_i), \tag{9}$$

where $\boldsymbol{e}_i^{\text{L}}$ is the patch representation in the last layer, g represents the additional shared linear classifier, y_i denotes the label of the corresponding patch, and $\mathcal{L}_{\text{XE}}$ stands for the cross entropy loss.

A2. More Implementation Details

Hyperparameters of our diversity regularizers. Table A4 summarizes our adopted hyperparameters during diversity-aware ViT training.

Table A4. Detailed hyperparamters of our diversity regularization.

Settings	Mixing loss	Weight	Attention	Embedding	
				Within-layer	Cross-layer
ViT-Small	1	5×10^{-4}	1×10^{-4}	0.5	0.5
ViT-Base	1	5×10^{-5}	1×10^{-5}	0.5	0.5
DeiT-Small	1	5×10^{-4}	1×10^{-4}	0.5	0.5
DeiT-Small24	1	5×10^{-4}	1×10^{-4}	0.5	0.5
DeiT-Base	1	1×10^{-6}	5×10^{-6}	0.5	0.5
Swin-Small	1×10^{-3}	1×10^{-6}	1×10^{-3}	0.9	-
Swin-Base	1	1×10^{-6}	1×10^{-3}	0.5	-

A3. More Experiment Results

Cross-layer diversity on patch embedding. Figure A10 shows that our methods substantially shrink the similarity of cross-layer patch embedding.

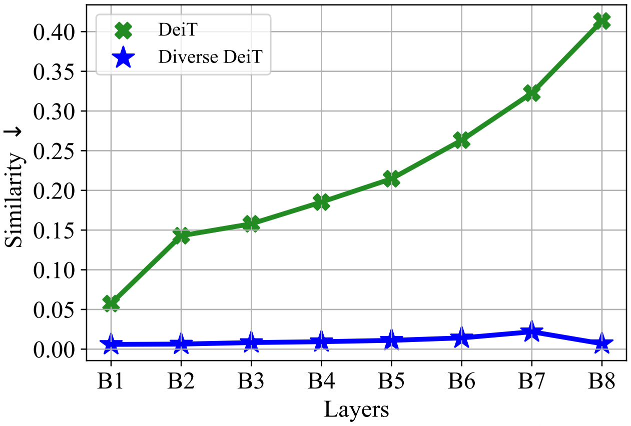

Figure A10. The cross-layer **patch embedding similarity** of DeiT-Small and its diversified version on ImageNet. We calculate the *cosine* similarity between embedding from each layer and the final layer. The smaller number indicates better diversity.

Cross-layer diversity on attention. As shown in Figure A11, our diverse ViT obtains a consistently lower correlation of cross-layer attention maps.

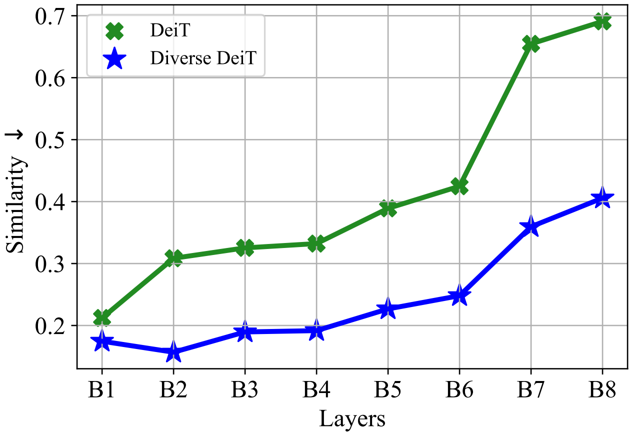

Figure A11. The cross-layer **attention maps similarity** of DeiT-Small and its diversified version on ImageNet. We calcluate the *cosine* similarity between embedding from each layer and the final layer. The smaller number indicates better diversity.

Standard deviations within attention maps. From Figure A12, we observe that the averaged standard deviations within attention maps are amplified by our approaches, suggesting an enhanced diversity. Note that we do not explicitly regularize the standard deviations of attention.

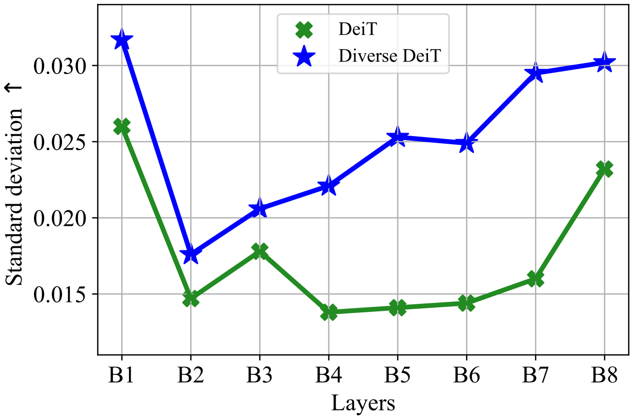

Figure A12. The averaged standard deviation within attention maps of DeiT-Small and its diversified version on ImageNet. The larger number indicates better diversity.