Supplementary Materials for "Robustifying Token Attention for Vision Transformers"

Yong Guo, David Stutz, Bernt Schiele
Max Planck Institute for Informatics, Saarland Informatics Campus
guoyongcs@gmail.com, {david.stutz,schiele}@mpi-inf.mpg.de

A. Overview and Outline

In our paper, we seek to address the token overfocusing issue of vision transformers and improve their overall robustness. To this end, we propose two general techniques, the *Token-aware Average Pooling (TAP)* module and the *Attention Diversification Loss (ADL)*. In this supplementary material, we conduct additional discussions on both techniques and provide complementary experiments. We organize the supplementary as follows:

- In Section B, we discuss the computational complexity of our Token-aware Average Pooling (TAP) module. Based on the considered baseline FAN-B-Hybrid, TAP only takes around 2% of the number of floating-point operations (FLOPs) in each self-attention layer, and less than 1% of the whole model.
- In Section C, besides corruption robustness, we demonstrate that the proposed methods also obtain promising improvement in terms of adversarial robustness. Then, we study the effect of the attention threshold τ in computing our Attention Diversification Loss (ADL). In addition, we provide more results for image classification and semantic segmentation.
- In Section D, we provide more visualization results to study the stability of attention maps in vision transformers. We demonstrate that the token overfocusing issue is particularly severe in relatively deep layers. We also highlight that this issue can be observed across diverse architectures (e.g., RVT [18] and FAN [31]) and tasks (including image classification and semantic segmentation). In addition, we provide more visual comparisons for the predicted segmentation masks.

B. Computational Complexity Analysis of TAP

As mentioned in the main paper [7], we introduce our TAP into each basic block to improve the robustness of the attention mechanism. We already demonstrated that our TAP only adds minimal computational overhead. Thus, in the following, we evaluate the computational complexity in terms of floating-point operations (FLOPs) to justify our argument. Given the input tokens $z \in \mathbb{R}^{H\times W\times C}$ with the spatial resolution of $H \times W$ and feature dimension of C, the complexity of a standard self-attention layer is[1]:

$$\mathrm{O}(\text{SelfAttention}) = 4HWC^2 + 2(HW)^2C \tag{i}$$

Based on the standard self-attention layer, we propose to introduce an additional TAP that exploits a dilation predictor (a two-layer convolutional module) to predict the weights for K branches and mix the features in a weighted sum manner. In this sense, the overall complexity consists of the complexities of the dilation predictor ($9HWCK + 9HWK^2$), K average pooling operations ($HWCK$), and the weighted sum operation ($HWCK$). Thus, the complexity introduced by TAP is:

$$\mathrm{O}(\text{TAP}) = 11HWCK + 9HWK^2 \tag{ii}$$

When combining TAP with the standard self-attention together, the overall complexity is:

$$\mathrm{O}(\text{TAP}-\text{SelfAttention}) = 11HWCK + 9HWK^2 + 4HWC^2 + 2(HW)^2C \tag{iii}$$

As for our best model built upon FAN-B-Hybrid [31], we have $H{=}W{=}14$, $C{=}448$, and $K{=}4$. By substituting them into the above equations, the cost introduced by our TAP only takes around 2% of the complexity of each self-attention block. When considering the whole model that consists of both convolutional modules/heads and self-attention blocks, the additional complexity is less than 1% in practice, showing that our TAP only introduces minimal computational overhead.

[1] A typical self-attention module consists of a fully-connected layer before and after the attention module, respectively. We compute the overall complexity of all the involved layers, which has been discussed and reported in [14].

C. More Discussions and Quantitative Results

In this paper, we mainly focus on improving the robustness against common corruptions. Besides this, we additionally investigate whether the improvement can also generalize to adversarial robustness. In this part, we report the robustness against adversarial attacks and demonstrate that both our TAP and ADL greatly improve adversarial robustness. Moreover, we provide additional comparison results on both image classification and semantic segmentation tasks.

Comparison of adversarial robustness. We also evaluate the robustness against adversarial attacks [17, 25]. We follow the settings of RVT [18] to construct the adversarial examples with the number of steps $t = 5$ and step size $\alpha = 0.5$, namely PGD-5. As shown in Table I, compared to the improvement against image corruptions, the proposed methods also obtain comparable improvement against adversarial attacks.

Method	ImageNet ↑	ImageNet-C (mCE) ↓	PGD-5 ↑
FAN-B-Hybrid [31]	83.9	46.1	30.5
+TAP	84.3	44.9 (-1.2)	31.4 (+0.9)
+ADL	84.0	44.4 (-1.7)	31.8 (+1.3)
+TAP & ADL	**84.3**	**43.7 (-2.4)**	**32.2 (+1.7)**

Table I. Comparisons of adversarial robustness against PGD attacks on ImageNet. We demonstrate that both our TAP and ADL also obtain promising improvement in terms of adversarial robustness.

More results for image classification. In this part, we provide more comparisons on diverse benchmarks. From Table II, besides RVT and FAN, we additionally compare more convolutional models/methods [8, 20, 10, 15] and a variety of transformer architectures [21, 3, 1, 14, 22, 12, 27]. We highlight that our models significantly outperform all the compared methods in terms of both clean accuracy and robustness. To be specific, based on FAN-B-Hybrid, our model outperforms a strong convolutional baseline ConvNeXt-B that contains more parameters by 0.5% in accuracy on ImageNet and 3.1% in mCE on ImageNet-C. This phenomenon can also be observed when compared with a popular transformer model DeiT-B. Furthermore, we also report the corruption error of each individual corruption type of ImageNet-C in Table III. Unsurprisingly, our model equipped with TAP and ADL yields the lowest (best) corruption error on most corruption types.

Method	#Params (M)	#FLOPs (G)	ImageNet ↑	ImageNet-C ↓	ImageNet-P ↓	ImageNet-A ↑	ImageNet-R ↑
ResNet50 [8]	25.6	4.1	76.1	76.7	58.0	0.0	36.1
Inception v3 [20]	27.2	5.7	77.4	80.6	61.3	10.0	38.9
EWS [5]	25.6	4.1	77.3	58.7	30.9	5.9	48.5
DeepAugment [10]	25.6	4.1	75.8	60.6	32.1	3.9	46.7
ConvNeXt-B [15]	88.6	15.4	83.8	46.8	-	36.7	51.3
DeiT-B [21]	86.6	17.6	82.0	48.5	32.1	27.4	44.9
ConViT-B [3]	86.5	17.7	82.4	46.9	32.2	29.0	48.4
XCiT-S24 [1]	47.7	9.1	82.6	49.4	-	27.8	45.5
Swin-B [14]	87.8	15.4	83.4	54.4	32.7	35.8	46.6
PVT-Large [22]	61.4	9.8	81.7	59.8	39.3	26.6	42.7
PiT-B [12]	73.8	12.5	82.4	48.2	-	33.9	43.7
T2T-ViT_t-24 [27]	64.1	15.0	82.6	48.0	31.8	28.9	47.9
RSPC (FAN-B-Hybrid) [6]	50.4	17.7	84.2	44.5	30.0	41.1	-
RVT-B [18]	91.8	17.7	82.6	46.8	31.9	28.5	48.7
+ TAP	92.1	17.9	83.0 (+0.4)	45.5 (-1.3)	30.6 (-1.3)	30.0 (+1.5)	49.4 (+0.7)
+ ADL	91.8	17.7	82.6 (+0.0)	45.2 (-1.6)	30.2 (-1.7)	30.8 (+2.3)	49.8 (+1.1)
+ TAP & ADL	92.1	17.9	**83.1 (+0.5)**	**44.7 (-2.1)**	**29.6 (-2.3)**	**32.7 (+4.2)**	**50.2 (+1.5)**
FAN-B-Hybrid [31]	50.4	11.7	83.9	46.1	31.3	39.6	52.7
+ TAP	50.7	11.8	84.3 (+0.4)	44.9 (-1.2)	30.3 (-1.0)	41.0 (+1.4)	53.9 (+1.2)
+ ADL	50.4	11.7	84.0 (+0.1)	44.4 (-1.7)	29.8 (-1.5)	41.4 (+1.8)	54.2 (+1.5)
+ TAP & ADL	50.7	11.8	**84.3 (+0.4)**	**43.7 (-2.4)**	**29.2 (-2.1)**	**42.3 (+2.7)**	**54.6 (+1.9)**

Table II. Comparisons on ImageNet and diverse robustness benchmarks. We report the mean corruption error (mCE) on ImageNet-C and mean flip rate (mFR) on ImageNet-P. For these metrics, lower is better. Moreover, we directly report the accuracy on ImageNet-A and ImageNet-R. Based on the considered two baselines, our models consistently improve the accuracy and robustness on diverse benchmarks.

Model	mCE ↓	Noise ↓			Blur ↓				Weather ↓				Digital ↓			
		Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brightness	Contrast	Elastic	Pixelate	JPEG
FAN-B-Hybrid [31]	46.1	40	39	37	52	64	48	55	40	44	37	37	34	62	53	52
+TAP	44.9	36	36	34	53	65	46	55	40	42	36	37	33	63	49	49
+ADL	44.1	36	36	34	**51**	64	45	54	38	40	38	37	33	**61**	48	**49**
+TAP & ADL	**43.7**	**34**	**34**	**32**	52	**62**	**45**	**54**	**38**	**40**	**35**	**36**	**33**	63	**47**	50

Table III. Corruption error on ImageNet-C. Following [11], we compute the corruption error for each corruption type by dividing by AlexNet's error. The mean corruption error (mCE) is a simple average over the corruption errors on all corruption types. Note that, for both corruption error and mCE, lower is better. Empirically, our TAP and ADL significantly reduce the corruption errors. More critically, combining TAP and ADL together obtains the lowest errors on most corruption types, yielding the best mCE in practice.

Model	Cityscapes	Average mIoU	Blur				Noise				Digital				Weather			
		on Cityscapes-C	Motion	Defoc	Glass	Gauss	Gauss	Impul	Shot	Speck	Bright	Contr	Satur	JPEG	Snow	Spatt	Fog	Frost
DeepLabv3+ (R50) [2]	76.6	36.8	58.5	56.6	47.2	57.7	6.5	7.2	10.0	31.1	58.2	54.7	41.3	27.4	12.0	42.0	55.9	22.8
DeepLabv3+ (R101) [2]	77.1	39.4	59.1	56.3	47.7	57.3	13.2	13.9	16.3	36.9	59.2	54.5	41.5	37.4	11.9	47.8	55.1	22.7
DeepLabv3+ (X65) [2]	78.4	42.7	63.9	59.1	52.8	59.2	15.0	10.6	19.8	42.4	65.9	59.1	46.1	31.4	19.3	50.7	63.6	23.8
DeepLabv3+ (X71) [2]	78.6	42.5	64.1	60.9	52.0	60.4	14.9	10.8	19.4	41.2	68.0	58.7	47.1	40.2	18.8	50.4	64.1	20.2
ICNet [28]	65.9	28.0	45.8	44.6	47.4	44.7	8.4	8.4	10.6	27.9	41.0	33.1	27.5	34.0	6.3	30.5	27.3	11.0
FCN8s [16]	66.7	27.4	42.7	31.1	37.0	34.1	6.7	5.7	7.8	24.9	53.3	39.0	36.0	21.2	11.3	31.6	37.6	19.7
DilatedNet [26]	68.6	30.3	44.4	36.3	32.5	38.4	15.6	14.0	18.4	32.7	52.7	32.6	38.1	29.1	12.5	32.3	34.7	19.2
PSPNet [29]	78.8	34.5	59.8	53.2	44.4	53.9	11.0	15.4	15.4	34.2	60.4	51.8	30.6	21.4	8.4	42.7	34.4	16.2
ConvNeXt-T [15]	79.0	54.4	64.1	61.4	49.1	62.1	34.9	31.8	38.8	56.7	76.7	68.1	76.0	51.1	25.0	58.7	74.2	35.1
SETR [30]	76.0	55.5	61.8	61.0	59.2	62.1	36.4	33.8	42.2	61.2	73.1	63.8	69.1	49.7	41.2	60.8	63.8	32.0
Swin-T [14]	78.1	47.5	62.1	61.0	48.7	62.2	22.1	24.8	25.1	42.2	75.8	62.1	75.7	33.7	19.9	56.9	72.1	30.0
Segformer-B0 [24]	76.2	48.9	59.3	58.9	51.0	59.1	25.1	26.6	30.4	50.7	73.3	66.3	71.9	31.2	22.1	52.9	65.3	31.2
Segformer-B1 [24]	78.4	52.6	63.8	63.5	52.0	29.8	23.3	35.4	56.2	76.3	70.8	74.7	36.1	56.2	28.3	60.5	70.5	36.3
Segformer-B2 [24]	81.0	55.8	68.1	67.6	58.8	68.1	23.8	23.1	27.2	47.0	79.9	76.2	78.7	46.2	34.9	64.8	76.0	42.1
Segformer-B5 [24]	82.4	65.8	69.1	68.6	64.1	69.8	57.8	63.4	52.3	72.8	81.0	77.7	80.1	58.8	40.7	68.4	78.5	49.9
FAN-B-Hybrid [31]	82.3	67.3	70.0	69.0	64.3	69.3	55.9	60.4	61.1	70.9	81.2	76.1	80.0	57.0	54.8	72.5	78.4	52.3
+TAP	82.7	69.2 (+1.9)	70.1	**69.2**	**66.6**	69.8	61.2	67.1	65.6	73.5	81.3	76.5	80.4	62.3	55.7	74.7	79.2	**54.9**
+ADL	82.4	69.4 (+2.1)	70.1	68.6	65.3	69.7	62.6	**68.5**	66.1	73.8	81.7	77.3	80.8	63.3	55.3	74.3	79.7	52.8
+TAP & ADL	**82.9**	**69.7 (+2.4)**	**70.4**	68.8	65.6	**69.8**	**63.0**	68.4	**67.1**	**74.1**	**81.8**	**77.4**	**80.9**	**63.5**	**56.9**	**74.9**	**80.0**	53.0

Table IV. Comparisons of mIoU on individual corruption type of Cityscapes-C based on FAN-B-Hybrid. We obtain the best results on most of the corruption types when combining our TAP and ADL together.

Model	Cityscapes	ACDC				
		Fog	Night	Rain	Snow	Average
RefineNet [13]	73.6	46.4	29.0	52.6	43.3	43.7
DeepLabv2 [23]	71.4	33.5	30.1	44.5	40.2	38.0
DeepLabv3+ (R101) [2]	77.1	45.7	25.0	50.0	42.0	41.6
DANet [4]	81.5	34.7	19.1	41.5	33.3	33.1
HRNetV2-W48 [19]	81.6	38.4	20.6	44.8	35.1	35.3
Segformer-B5 [24]	82.4	63.2	47.8	66.4	63.7	62.0
FAN-B-Hybrid [31]	82.2	64.0	45.9	67.8	64.5	60.6
FAN-B-Hybrid (TAP & ADL)	**82.9**	**64.5**	**51.0**	**70.5**	**68.2**	**63.6**

Table V. Comparisons of mIoU on individual adverse conditions of ACDC based on FAN-B-Hybrid. When equipped with TAP and ADL, we obtain the best results on all the conditions compared to the baseline FAN-B-Hybrid and existing approaches.

More results for semantic segmentation. For semantic segmentation, we provide more quantitative results and detailed comparisons, including the results on individual corruption types of Cityscapes-C in Table IV and the results on individual adverse conditions of ACDC in Table V. Unlike Table 4 in the main paper, we include more popular methods for comparisons on Cityscapes-C. We demonstrate that our TAP and ADL greatly improve the mIoU on Cityscapes-C by 1.9% and 2.1%, respectively, along with improved mIoU on the clean Cityscapes dataset. Moreover, our models outperform all the compared methods and achieve the best tradeoff between clean performance and robustness. In addition, we also show the detailed results on individual corruption types. As shown in Table IV, our best model yields the largest improvement mainly on Noise corruptions by $>3.2\%$ in terms of mIoU, while obtaining a relatively smaller improvement on blur corruptions. In addition, using TAP alone performs better on some corruption types, including defocus blur, glass blur, and frost. When combining TAP and ADL, we are able to obtain the best results on most of the corruption types. Similarly, we observe the same phenomenon on the adverse conditions of ACDC, as shown in Table V. For example, when using TAP and ADL, our model consistently obtains the best results on all the individual adverse conditions, yielding the best average results as well. These results indicate that both the proposed TAP and ADL are general techniques that are able to improve the robustness on diverse tasks and corruption types.

D. More Visualization Results

In this part, we provide additional visualization results of intermediate attention maps of vision transformers. We demonstrate that the token overfocusing issue can be observed across different layers in a model, different architectures, and the models on semantic segmentation tasks. Then, we show more visual comparisons of segmentation results.

Visualization of intermediate attention maps. In the main paper, we illustrate the overfocusing issue based on the attention maps of the last layer. Indeed, this issue can be observed across most of the layers. As shown in Figure I, for the baseline model, the overfocusing issue becomes more and more obvious from 7-th layer to the last layer. More critically, we highlight that all the deep layers focus on the same set of important tokens. When facing image corruptions, e.g., Gaussian noise, we observe a severe attention shift across all the intermediate layers, indicating that the standard self-attention is very fragile. In contrast, our model adopts a diagonal attention pattern in all the layers and exhibits significantly better stability against image corruptions. We hypothesize that the diagonal pattern plays an important role in stabilizing the attention since we inherently encourage the tokens to preserve most of their own information when aggregating information from other tokens. Although the information from other tokens is relatively weak in each layer, the model is able to gradually extract discriminative features in the end by stacking multiple self-attention layers. We also highlight that the diagonal attention pattern follows a similar fusion manner with the residual architecture [9], which preserves the original information using an identity mapping and extracts new features in the residual branch.

Alleviating token overfocusing issue on top of diverse architectures. We have shown the effectiveness of our methods in alleviating the token overfocusing issue based on FAN-B-Hybrid. Here, we additionally take another transformer RVT-B to verify the generalization ability of our methods. From Figure II, we obtain several important observations. First, the token overfocusing issue also exists in RVT-B and becomes much more serious than that in FAN-B-Hybrid (see the second column of Figure II). To be specific, the model often relies on less than 5 tokens to compute the self-attention. Second, our TAP and ADL exhibit consistent attention patterns between both RVT and FAN architectures. Clearly, TAP encourages more tokens to take part in the attention mechanism and ADL adopts a diagonal attention pattern in which the attention diversity among rows is high enough. When combing them together, the attention becomes much more stable against image corruptions, sharing a similar observation with Figure III. These results verify our argument that the proposed methods are general techniques that can be applied to diverse architectures.

Alleviating token overfocusing issue on semantic segmentation tasks. Besides image classification models, we additionally show the effectiveness of our methods in alleviating the overfocusing issue on semantic segmentation models. In Figure IV, we take FAN-B-Hybrid as the backbone to build a segmentation model and show the attention maps of the last layer. Note that the number of tokens in attention maps becomes much larger than that of image classification models due to the extremely large resolution of input images, e.g., often with 2048x1024 in Cityscapes. Because of the increased number of tokens, we observe a slightly different attention pattern in the baseline model such that more tokens are relied upon by the attention mechanism. Nevertheless, the overfocusing issue is still very obvious and the vulnerability of attention against common corruptions can also be observed. When applying our TAP and/or ADL to the segmentation model, we observe a similar attention pattern to that on image classification models shown in Figure III. These results indicate that our method can generalize well to semantic segmentation tasks.

More visual comparisons of semantic segmentation. In the main paper, we have shown some examples to demonstrate the superiority of our methods in improving the robustness of semantic segmentation models. Here, we additionally provide the visualization results of more examples. To be specific, we show more visualization results on ACDC and Cityscapes-C in Figure V and Figure VI, respectively. In Figure V, we study the robustness of segmentation models against all four adverse conditions in ACDC dataset, including night, fog, rain, and snow. For clarity, we use the red box to highlight the major differences between different segmentation masks. As for the night example (first row), the baseline model cannot detect the car on the left under the insufficient lighting condition, while our model is still able to detect the car. This phenomenon can be also observed in the fog weather in the second example. In the third and fourth examples, we show that the rain and snow conditions often cause misclassification of the road part. We highlight that road detection plays in important role in autonomous driving scenarios and the robustness against adverse weather conditions becomes critical. Moreover, we also compare the segmentation results against diverse common corruptions in Cityscapes-C. In Figure VI, we further investigate the robustness against common corrutpions. Besides Gaussian noise that we considered in previous experiments, we also show the visualization results on top of other corruptions, including spatter, pixelated, impulse noise, and defocus blur. As for the first two examples in Figure VI, the baseline model cannot accurately detect the bike regardless of whether there is a person on it. In the last two examples, when facing noise and blur, the baseline model cannot detect the whole body of the person, while our model accurately detects the person. Overall, these results demonstrate the effectiveness of the proposed methods in improving the robustness of semantic segmentation models.

Figure I. Attention maps of intermediate layers based on FAN-B-Hybrid. We demonstrate that the token overfocusing issue can be observed in most layers and becomes gradually more serious with the increase of depth. When facing common corruptions, e.g., Gaussian noise, the attention mechanism becomes extremely fragile by focusing on entirely different important tokens. By contrast, our model follows a similar attention pattern (diagonal) across layers and exhibits better stability against corruptions.

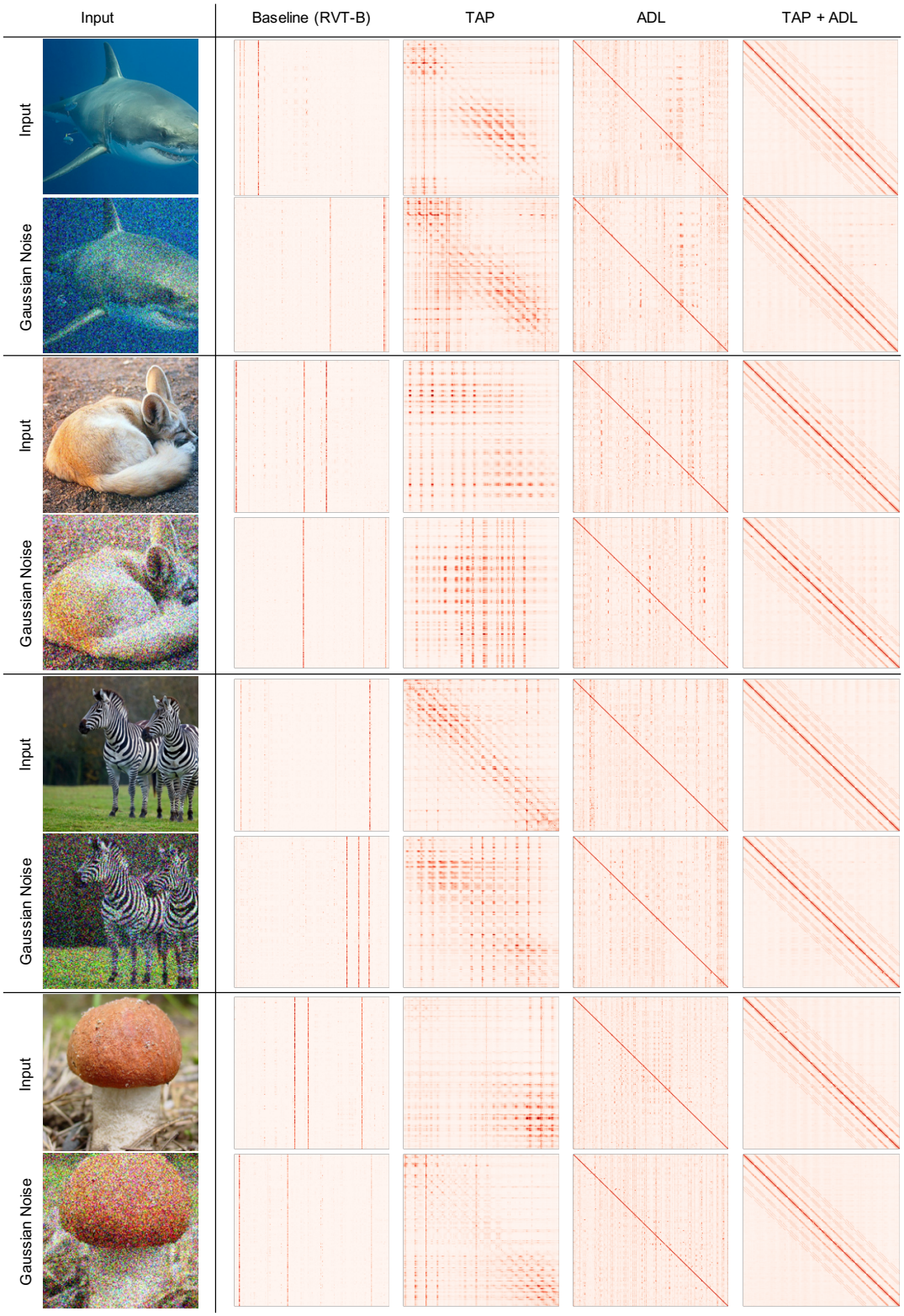

Figure II. Attention maps of the last layer based on RVT-B. Compared with FAN-B-Hybrid, the overfocusing issue becomes much more serious in RVT-B since the attention relies on fewer important tokens, e.g., often less than 5 tokens. Nevertheless, our TAP and ADL exhibit a similar attention pattern to that on FAN-B-Hybrid, indicating that the proposed methods can generalize well to diverse architectures.

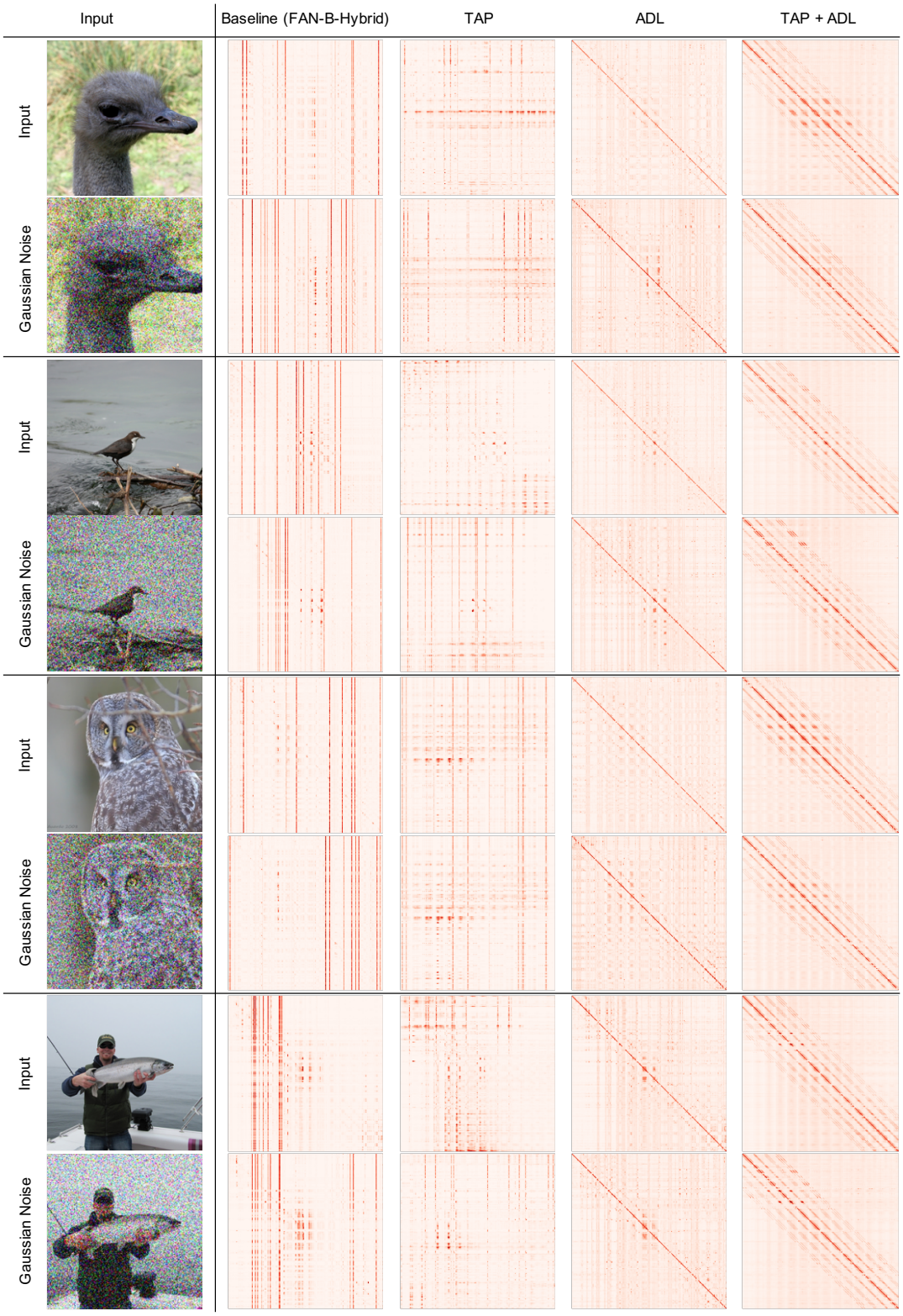

Figure III. Attention maps of the last layer based on FAN-B-Hybrid. We demonstrate that the baseline model tends to rely on very few tokens in the attention mechanism. By contrast, combining both our TAP and ADL obtains more balanced attention across tokens (columns) and diverse attention across rows. More importantly, the attention of our model is very stable against common corruptions.

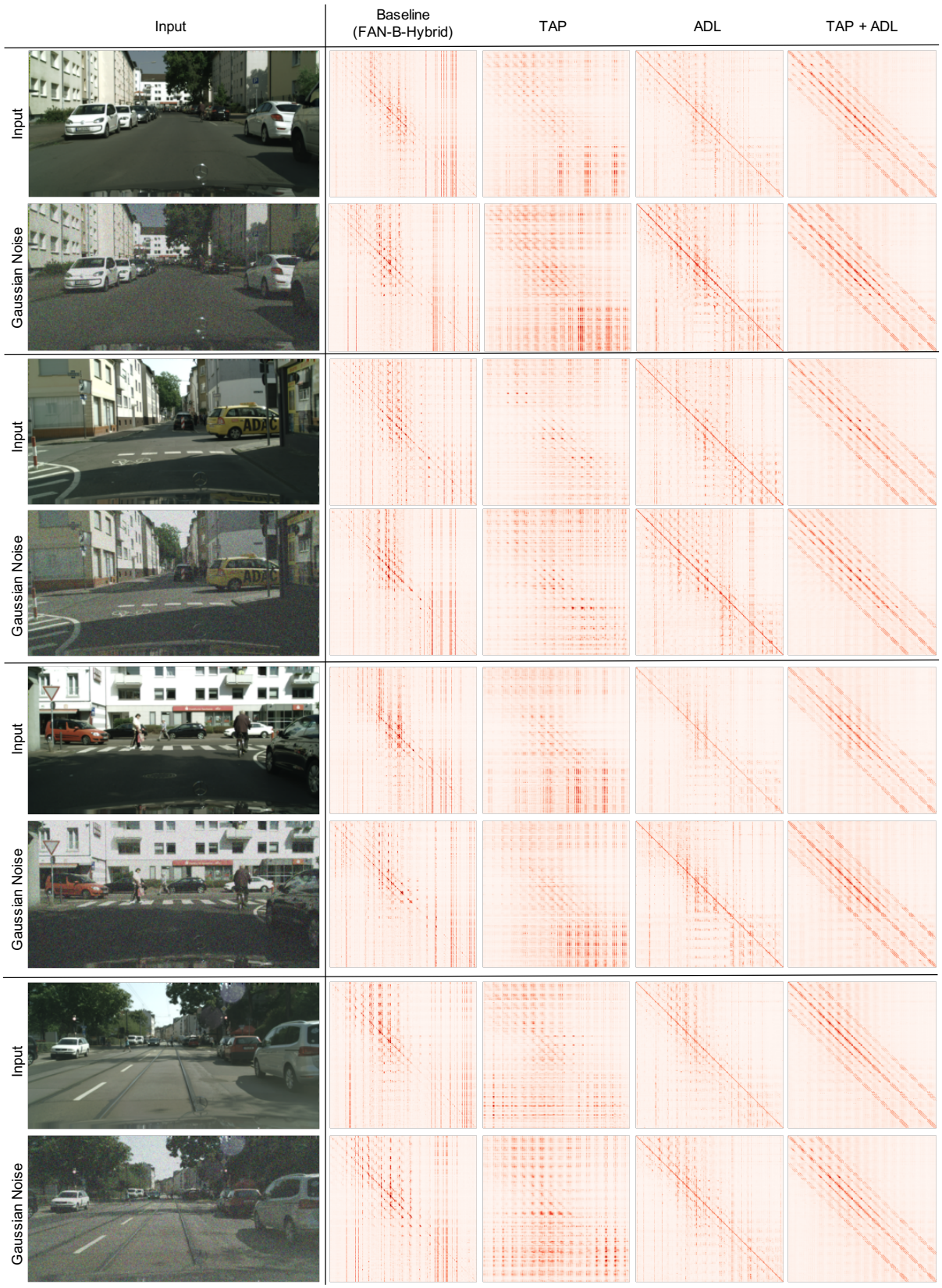

Figure IV. Attention maps of the last layer in the segmentation model based on FAN-B-Hybrid backbone. We show that the token overfocusing issue also exists in segmentation models and the attention mechanism is very fragile to image corruptions. By contrast, our TAP and ADL obtains consistent attention pattern for both image classification and semantic segmentation models. These results indicate the generalization ability of our approaches to the semantic segmentation tasks.

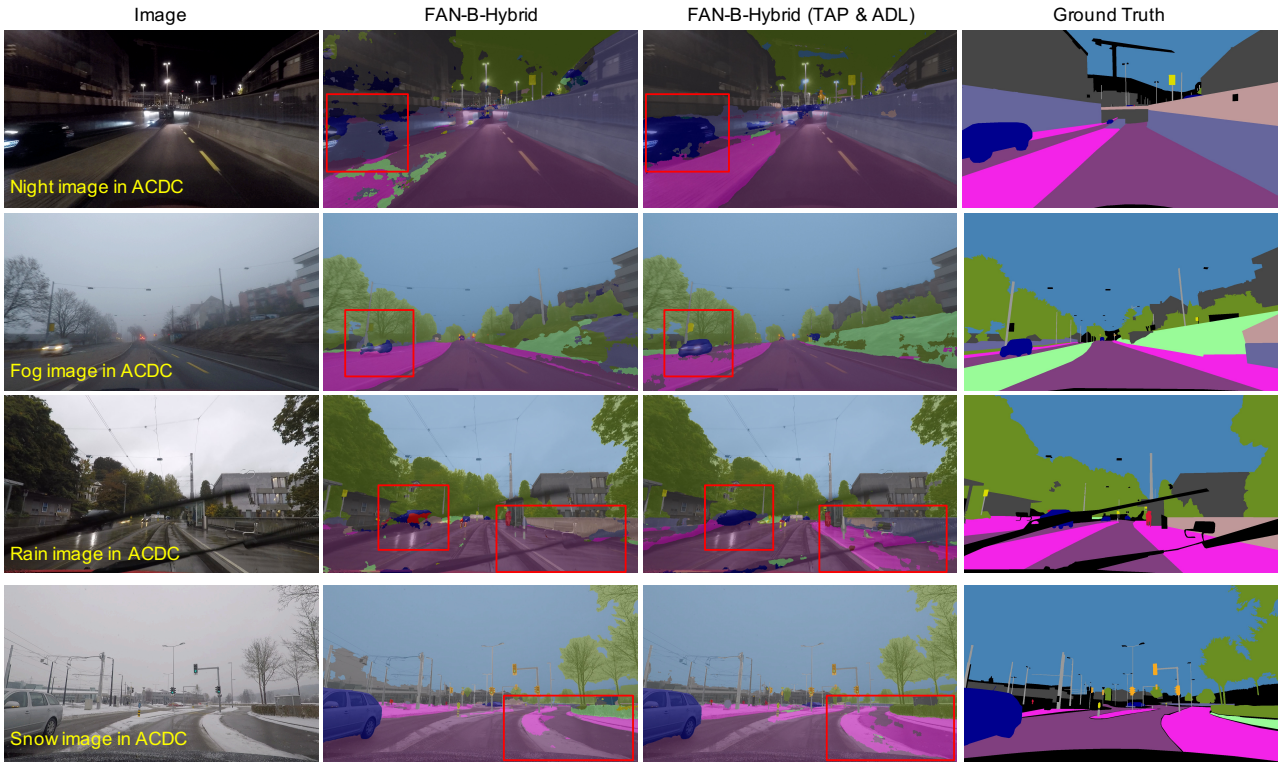

Figure V. Visual comparisons of segmentation results on ACDC. When facing adverse conditions, the baseline FAN-B-Hybrid model often fails to detect cars (in the first three examples) or roads (in the last two examples). By contrast, our model is much more robust against these adverse conditions than the baseline model.

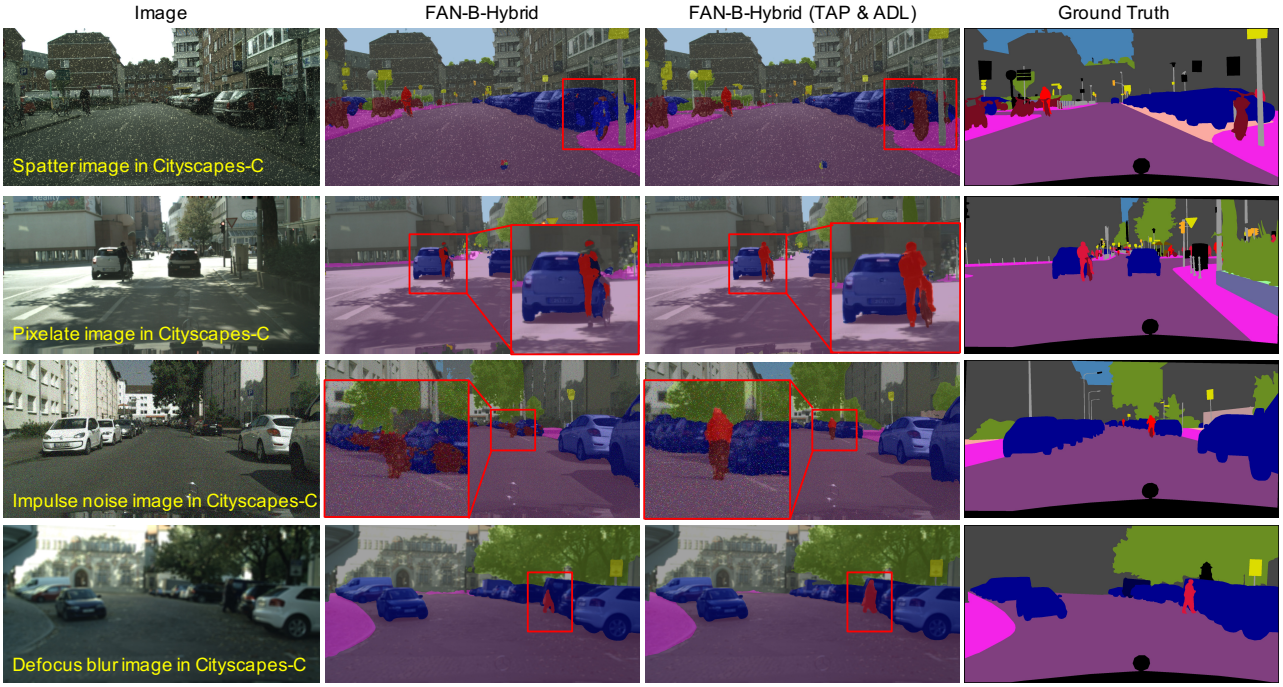

Figure VI. Visual comparisons of segmentation results on Cityscapes-C. When facing image corruptions, the baseline FAN-B-Hybrid model cannot detect the bike (in the first two exampls) and/or the whole body of a person (in the last two examples). By contrast, our model is much more robust against these corruptions.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. 2

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[3] 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 *Proc. of the International Conference on Machine Learning (ICML)*, pages 2286–2296. PMLR, 2021. 2

[4] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3146–3154, 2019. 3

[5] Yong Guo, David Stutz, and Bernt Schiele. Improving robustness by enhancing weak subnets. In *European Conference on Computer Vision*, pages 320–338. Springer, 2022. 2

[6] Yong Guo, David Stutz, and Bernt Schiele. Improving robustness of vision transformers by reducing sensitivity to patch corruptions. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4108–4118, 2023. 2

[7] Yong Guo, David Stutz, and Bernt Schiele. Robustifying token attention for vision transformers. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[10] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 2

[11] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2019. 3

[12] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 11936–11945, 2021. 2

[13] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1925–1934, 2017. 3

[14] 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 *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 1, 2, 3

[15] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. 2, 3

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 3

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2018. 2

[18] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[19] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 3

[20] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[21] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proc. of the International Conference on Machine Learning (ICML)*, 2021. 2

[22] 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. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 2

[23] Mark Weber, Huiyu Wang, Siyuan Qiao, Jun Xie, Maxwell D Collins, Yukun Zhu, Liangzhe Yuan, Dahun Kim, Qihang Yu, Daniel Cremers, et al. Deeplab2: A tensorflow library for deep labeling. *arXiv preprint arXiv:2106.09748*, 2021. 3

[24] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:12077–12090, 2021. 3

[25] Bingna Xu, Yong Guo, Luoqian Jiang, Mianjie Yu, and Jian Chen. Downscaled representation matters: Improving image rescaling with collaborative downscaled images. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 2

[26] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2016. 3

[27] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 558–567, 2021. 2

[28] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 405–420, 2018. 3

[29] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017. 3

[30] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6881–6890, 2021. 3

[31] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 27378–27394. PMLR, 2022. 1, 2, 3