Supplementary for "Improving Robustness of Vision Transformers by Reducing Sensitivity to Patch Corruptions"

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

A. Overview and Outline

In the main paper, we study the vulnerability of transformers by investigating the stability of self-attention layers against patch-based corruptions. Since the existing self-attention mechanism tends to be overly sensitive to patch-based corruptions, we propose to improve the stability by explicitly **reducing sensitivity to patch corruptions (RSPC)**. In the supplementary, we provide additional ablations and complementary experiments. We organize the supplementary as follows:

- In Section B, we conduct additional ablation studies on the effect of occluding patches with noise and the effect of our feature alignment loss in generating patch-based corruptions. We show that, compared with dropping patches, occluding patches with noise greatly benefits the feature alignment process. Moreover, generating the corrupted examples using our feature alignment loss encourages the model to obtain larger robustness improvement than the cross-entropy loss.
- In Section C, we investigate the impact of two hyperparameters used in our RSPC method, including the occlusion ratio ρ and the weight of alignment loss λ. Moreover, we further apply our method on top of diverse architectures. We observe that our RSPC yields consistent improvement based on diverse architectures.
- In Section D, we evaluate our RSPC models against patch corruptions/perturbations, adversarial attacks, and each individual corruption type of ImageNet-C. We show that RSPC outperforms the baseline model in terms of both corruption robustness and adversarial robustness. Moreover, our method yields better results on most corruption types, alongside the better overall robustness on the whole dataset.
- In Section E, we include more qualitative results regarding the stability of attention layers. Experiments show the superiority of our RSPC over the RVT baseline in improving the stability of intermediate attention maps.

B. More Discussions on Generating Patch-based Corruptions

In this section, we conduct further ablations of the proposed method based on RVT-B on ImageNet. Specifically, we first compare occluding patches with noise and simply dropping patches when generating the patch-based corruptions. Then, we also demonstrate the superiority of our feature alignment loss over the cross-entropy loss in identifying vulnerable patches.

Occluding patches with noise. As mentioned in the main paper, we construct the patch-based corruptions by occluding patches with noise. Actually, one can also directly drop these patches, namely PatchDrop. Here, we empirically compare these two methods based on RVT-B model on ImageNet(-C). As shown in Table I, performing feature alignment against both PatchDrop and the occlusion with noise consistently improves the clean accuracy. Nevertheless, occluding patches with noise encourages the model to obtain better robustness than PatchDrop on ImageNet-C. The main reason is that, directly dropping patches often imposes relatively weak impact on the model and thus comes with limited performance improvement. By contrast, occluding patches with noise provides more severe impact on the intermediate features, making the feature alignment more challenging and also more effective.

Effect of feature alignment loss in generating patch-based corruptions. We further empirically compare the proposed feature alignment loss $\mathcal{L}_{\text{align}}$ with the cross-entropy loss in generating the patch-based corruptions. As shown in Eqn. (**??**), we seek to maximize $\mathcal{L}_{\text{align}}$ to identify those vulnerable patches that would greatly distract the intermediate attention layers. Indeed, we can also directly maximize the cross-entropy loss $\mathcal{L}_{\text{ce}}$ on the final prediction to perform patch selection. We compare these two approaches and show the results in Table II. Clearly, maximizing $\mathcal{L}_{\text{align}}$ yields significantly better robustness than maximizing $\mathcal{L}_{\text{ce}}$. The main reason is that the cross-entropy loss only focuses on the final layer and may fail

Method	ImageNet ↑	ImageNet-C (mCE) ↓
Baseline (RVT-B)	82.6	46.8 (-0.0)
RSPC (PatchDrop)	**82.8**	46.3 (-0.5)
RSPC (Occlusion with Noise)	**82.8**	**45.7 (-1.1)**

Table I. Comparisons between occluding patches with noise and dropping patches on ImageNet and ImageNet-C. We show that stabilizing the models against the occluded patches with noise yields significantly better robustness than stabilizing against PatchDrop.

to mislead the intermediate attention layers. In contrast, maximizing $\mathcal{L}_{\text{align}}$ explicitly distracts all the attention layers in the model (also including the final layer) and would potentially encourage the whole model to be more robust when performing feature alignment against the generated patch corruptions.

Patch Selection Strategy	ImageNet ↑	ImageNet-C (mCE) ↓
Baseline (RVT-B)	82.6	46.8 (-0.0)
RSPC with Patch Selection ($\max \mathcal{L}_{\text{ce}}(\hat{x})$)	82.7	46.2 (-0.6)
RSPC with Patch Selection ($\max \mathcal{L}_{\text{align}}(x, \hat{x})$)	**82.8**	**45.7 (-1.1)**

Table II. Comparisons between cross-entropy loss and the feature alignment loss in generating patch-based corruptions based. We take RVT-B as the baseline in this experiment. Generating patch corruptions by maximizing the feature alignment loss explicitly distracts intermediate attention layers. Thus, stabilizing the model against these patch corruptions encourages the model to obtain significantly better robustness than the cross-entropy loss.

C. Impact of Hyperparameters and Effectiveness on Diverse Architectures

We conduct more ablations on two hyperparameters of our RSPC, including the occlusion ratio ρ and the weight of alignment loss λ. In addition, we further investigate the effect of RSPC on more architectures, such as DeiT [7] and Swin [4].

Occlusion ratio ρ and the weight of alignment loss λ. We train the RVT-S model using our RSPC on CIFAR-10 to investigate the impact of the occlusion ratio ρ and the weight of alignment loss λ. In Figure I (left), we change the value ρ within the range between 0% and 50%. In practice, we obtain the best robustness with the ratio of 10% and thus set ρ=10% in our experiments. In Figure I (right), we discuss the weight of our alignment loss λ. Given a set of candidate values as shown in this figure, we observe that λ=0.005 yields the best robustness along with competitive clean accuracy. We highlight that these hyper-parameters also generalize well on CIFAR-100 and ImageNet. We adopt these settings in all the experiments.

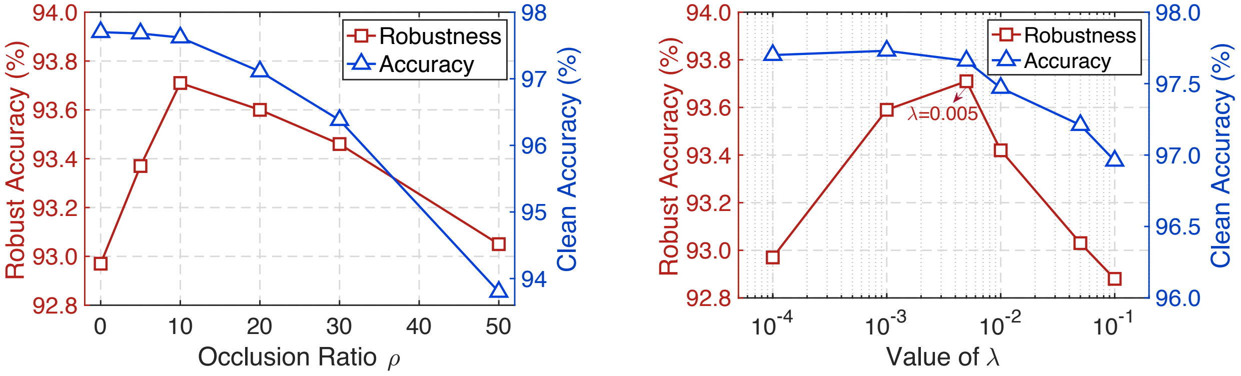

Figure I. Accuracy and corruption robustness of RSPC-RVT-S on CIFAR-10 plotted against the occlusion ratio ρ (left) and the weight of the alignment loss (right). *Left:* $\rho = 10\%$ obtains the best robustness along with comparable accuracy. *Right:* A too small or too large value of λ hampers the robustness of RSPC models. We found that $\lambda = 0.005$ performs best in most cases.

Effectiveness on top of diverse architectures. Recently, many efforts have been made to improve the robustness of deep networks [2,3,6,8]. Along this line of research, designing robust architectures becomes one of the most effective approaches, e.g., RVT [6] and FAN [8]. Nevertheless, we highlight that our approach can improve robustness not only on top of robust

Model	ImageNet	mCE on ImageNet-C ↓
DeiT-Ti [7]	72.2	71.1 (-0.0)
+RSPC (Ours)	**72.6**	**69.7 (-1.4)**
Swin-T [4]	81.2	62.0 (-0.0)
+RSPC (Ours)	**81.4**	**61.0 (-1.0)**
RVT-Ti [6]	79.2	57.0 (-0.0)
+RSPC (Ours)	**79.5**	**55.7 (-1.3)**
FAN-T-Hybrid [8]	80.1	58.3 (-0.0)
+RSPC (Ours)	**80.3**	**57.2 (-1.1)**

Table III. Effect of our RSPC based on various architectures. We report the accuracy and mean corruption error (mCE) on ImageNet and ImageNet-C, respectively. Our RSPC based models consistently improve the corruption robustness across different architectures while preserving comparable or higher clean accuracy.

architectures but also on general transformer architectures. To be specific, we additionally apply our RSPC on top of more transformer architectures, including DeiT [7] and Swin [4]. In this experiment, we report the accuracy on ImageNet and robustness in terms of mCE (the lower the better) on ImageNet-C. As shown in Table III, based on DeiT-Ti, we greatly improve the corruption robustness by reducing the mCE by 1.4% and yield a promising improvement of 0.4% on clean data. As for Swin-T, we obtain a similar observation that our RSPC is particularly effective in improving corruption robustness, reducing mCE from 62.0% to 61.0%. These results indicate that our RSPC can generalize well across diverse architectures.

D. More Comparisons of Model Robustness

Robustness against patch corruptions & perturbations. We also evaluate the robustness against patch-based corruptions (with the corruptions on both randomly selected patches and adversarial selected patches) and adversarial attacks. As for adversarial robustness, we report the accuracy on Patch-Fool [1] and PGD attack [5]. Here, we consider the vanilla Patch-Fool (single patch, i.e. 1P) without constraints and follow the hyper-parameters of the original paper. As for PGD attack, we follow the settings of RVT [6] to construct the adversarial examples with the number of steps $t = 5$ and step size $\alpha = 0.5$. From Table IV, our RSPC models consistently outperform the baseline models on both patch-based corruptions and adversarial attacks.

Method		RVT-B	FAN-B-Hybrid
Patch-based Corruption (Randomly Selected Patch)	Vanilla	71.4 (+0.0)	73.7 (+0.0)
	RSPC	**73.6 (+2.2)**	**75.2 (+1.5)**
Patch-based Corruption (Generated by $\mathcal{C}$)	Vanilla	61.4 (+0.0)	62.9 (+0.0)
	RSPC	**64.1 (+2.7)**	**65.2 (+2.3)**
Patch-Fool	Vanilla	69.3 (+0.0)	71.2 (+0.0)
	RSPC	**70.9 (+1.6)**	**72.5 (+1.3)**
PGD-5	Vanilla	29.9 (+0.0)	30.5 (+0.0)
	RSPC	**30.8 (+0.9)**	**31.7 (+1.2)**

Table IV. Comparisons of robustness against patch-based corruptions with the occlusion ratio $\rho = 10\%$, patch-based perturbations (e.g., Patch-Fool [1]), and PGD attack on ImageNet. Our RSPC models consistently outperform the baselines against them.

Robustness on individual corruption type. We compare the corruption error on each individual corruption type of ImageNet-C between RVT-Ti and our RSPC-RVT-Ti. In Figure II, our RSPC model yields lower corruption error than the baseline model in most of the corruption types. Although we introduce random noise when generating the patch-based corruptions, the major improvement of corruption robustness does not come from the noise related corruptions. Instead, the improved robustness can generalize well to other corruptions, e.g., yielding clearly lower error on the corruptions of snow, frost, and fog.

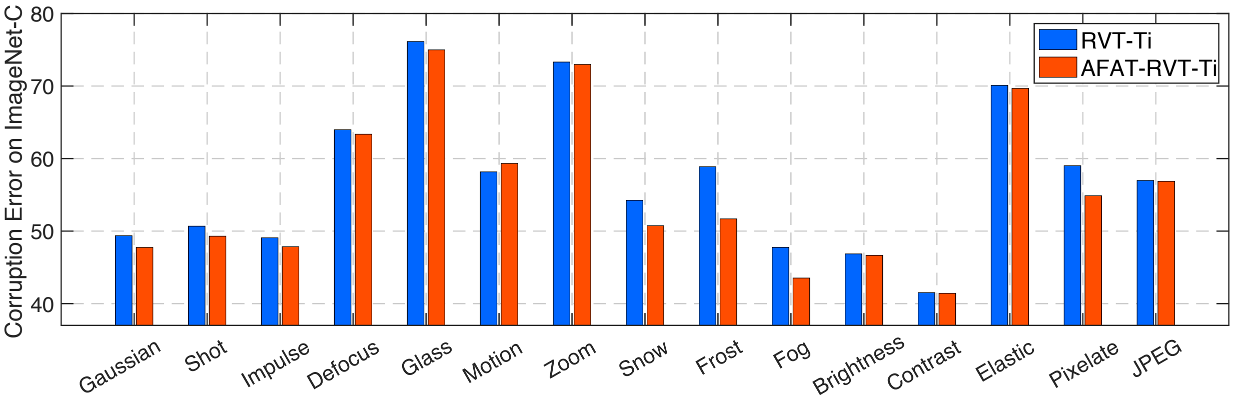

Figure II. Comparisons of corruption error (the lower, the better) on individual corruption type of ImageNet-C between RVT-Ti and RSPC-RVT-Ti. Our RSPC model outperforms the RVT baseline model on most of the corruption types.

E. More Visualization Results of Attention Stability

In the main paper, we have shown some examples to demonstrate the superiority of our RSPCs in improving the stability of attention maps. Here, we additionally provide the visualization results of more examples. As highlighted by Figure 5 of the paper, our RSPCs yield significantly more stable attention, in terms of cosine similarity (Cos-Sim), than the baseline RVTs across the whole ImageNet dataset. Interestingly, we observe that the visual stability of attention is highly correlated with Cos-Sim. To be specific, our RSPC often obtains very stable attention on the examples with a Cos-Sim larger than 0.8, as shown in Figure III. In addition, we also show some examples with the Cos-Sim lower than 0.8 in Figure IV.

As shown in Figure III, following [1], we average the attention maps across all the attention heads in each layer and visualize the attention map for a query token, e.g., the center token highlighted by the red box. Given the randomly occluded examples, RVT often incurs significant changes in the attention maps. By contrast, our RSPC effectively preserves most of the regions with relatively high attention scores across layers. In addition, as occluding different patches may have different impacts, we show the distribution of Cos-Sim over 1000 randomly sampled occlusion masks for each image in Figure III (last column). Across different examples, our RSPC yields significantly better attention stability than RVT both qualitatively and quantitatively. In Figure IV, we also show two examples with a Cos-Sim lower than 0.8. These occluded examples distract the attention of both RVT and our RSPC across layers. Nevertheless, our RSPC still yields better quantitative results than RVT. Overall, these results demonstrate the effectiveness of RSPC in improving the stability of self-attention.

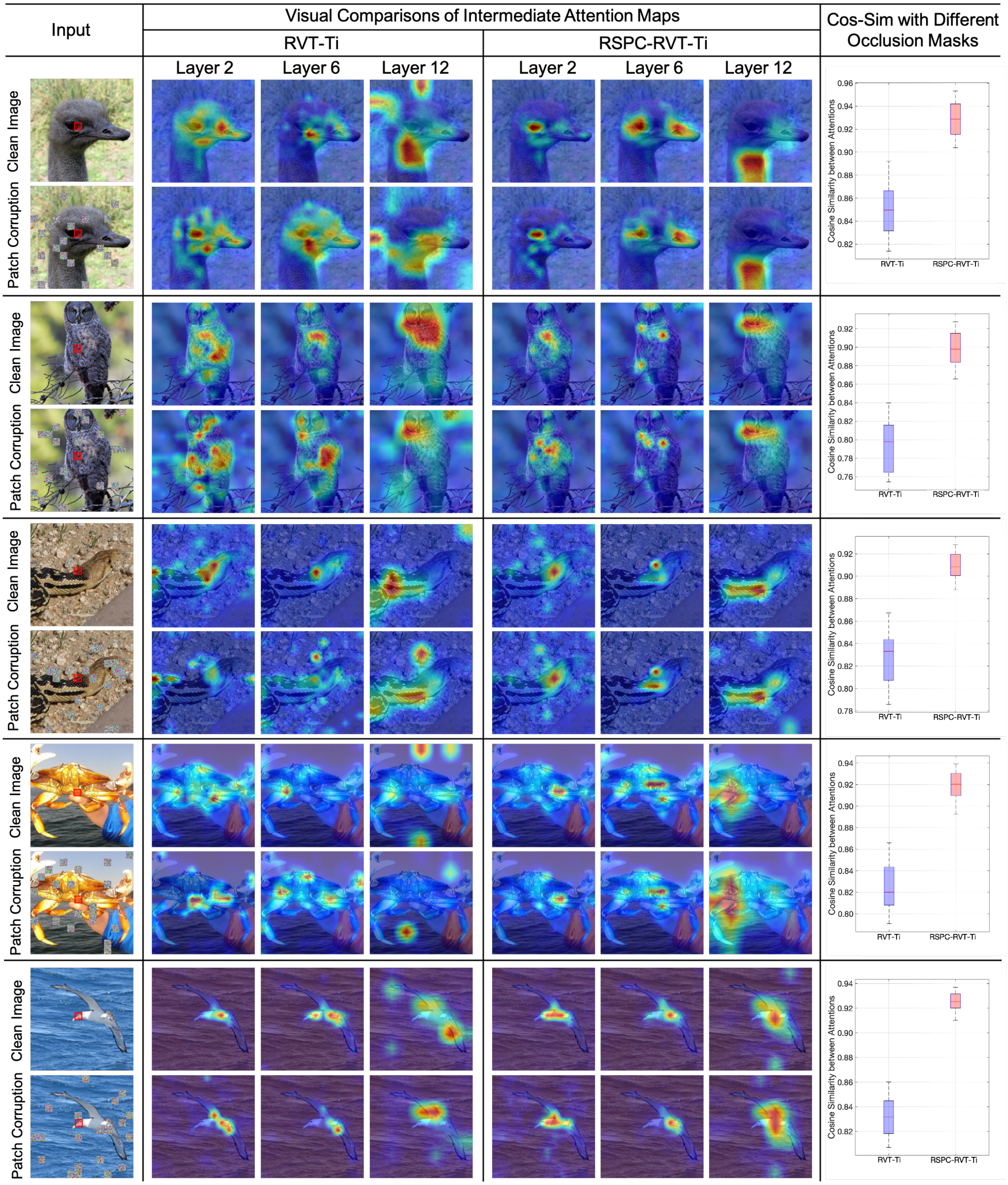

Figure III. Comparisons of attention stability between RVT-Ti and our RSPC-RVT-Ti on the examples with relatively high cosine similarity (Cos-Sim). In the last column, we also investigate the impact of different occlusion masks (1000 random masks) on each example in terms of Cos-Sim. Clearly, our RSPC yields much more stable attention maps both qualitatively and quantitatively.

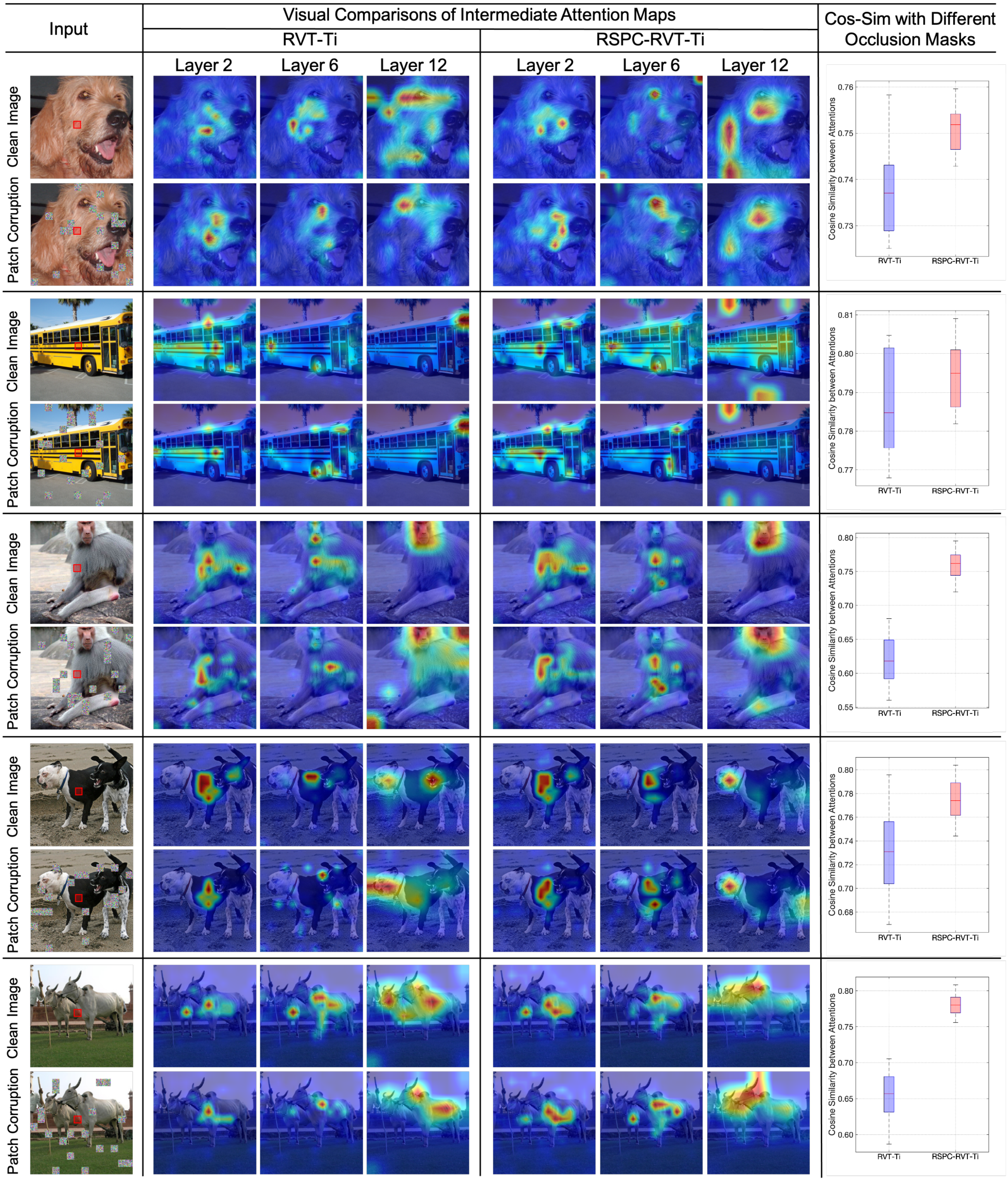

Figure IV. Comparisons of attention stability between RVT-Ti and our RSPC-RVT-Ti on the examples with relatively low cosine similarity (Cos-Sim). The occluded examples distract the attention of both methods but our RSPC still yields higher Cos-Sim than RVT.

References

[1] Yonggan Fu, Shunyao Zhang, Shang Wu, Cheng Wan, and Yingyan Lin. Patch-fool: Are vision transformers always robust against adversarial perturbations? In *Proc. of the International Conference on Learning Representations (ICLR)*, 2022. 3, 4

[2] Yong Guo, David Stutz, and Bernt Schiele. Improving robustness by enhancing weak subnets. In *Proc. of the European Conference on Computer Vision (ECCV)*. Springer, 2022. 2

[3] Yong Guo, David Stutz, and Bernt Schiele. Robustifying token attention for vision transformers. *arXiv preprint arXiv:2303.11126*, 2023. 2

[4] 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. 2, 3

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *Proc. of the International Conference on Learning Representations (ICLR)*, 2018. 3

[6] 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. 2, 3

[7] 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, 3

[8] 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. 2, 3