Scale-space Tokenization for Improving the Robustness of Vision Transformers

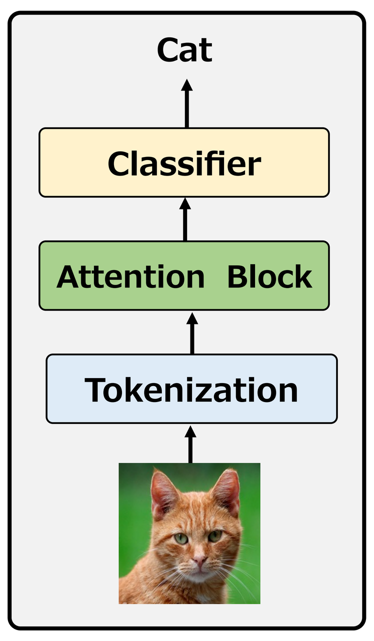
Lei Xu, Rei Kawakami, Nakamasa Inoue
Tokyo Institute of Technology

ACM MM 2023

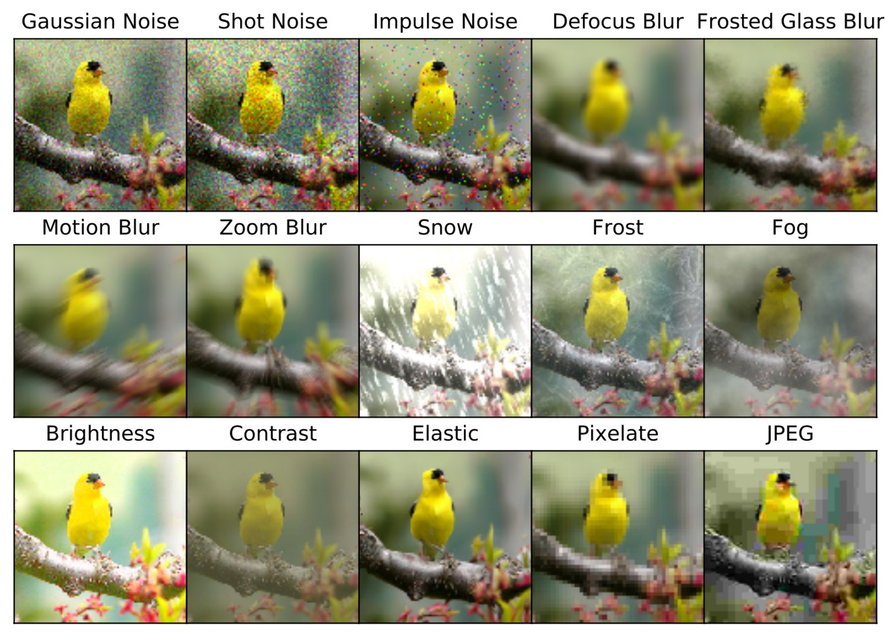
Introduction

The performance of the Vision Transformer (ViT) model and its variants in most vision tasks has surpassed traditional CNNs in terms of in-distribution accuracy. However, ViTs still have significant room for improvement in their robustness to input perturbations.

Robustness benchmark
Adversarial robustness: FGSM [Goodfellow+ 15], PGD [Madry+ 19]
Out-of-distribution robustness: ImageNet-C [Hendrycks+ 19]



ViT model

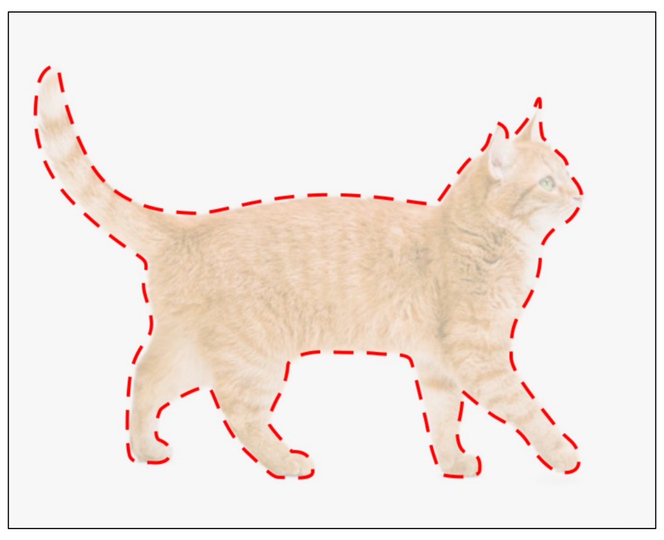


Robustness benchmark

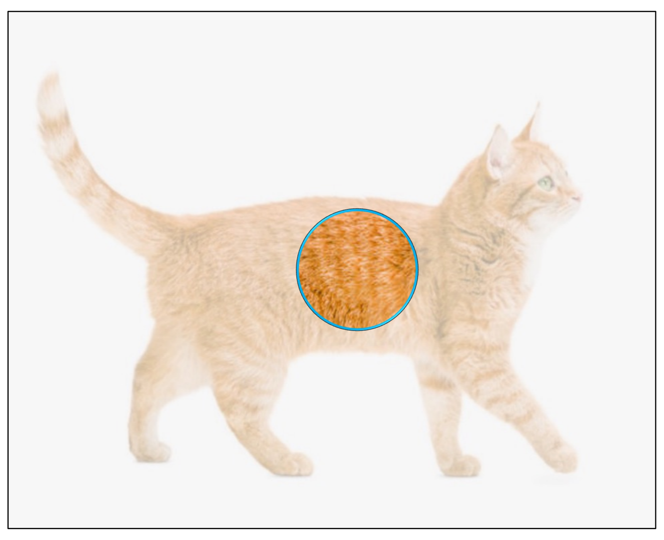
Proposed Method: Scale-space Tokenization for Vision Transformer Models

We propose scale-space tokenization for improving the robustness of vision transformers. Our key idea is to increase shape bias and predispose vision transformers to strike a certain degree of balance between the learning of texture-based and shape-based features by the fine-to-coarse image structures characterized by the scale space.

Motivation: Shape bias and texture bias

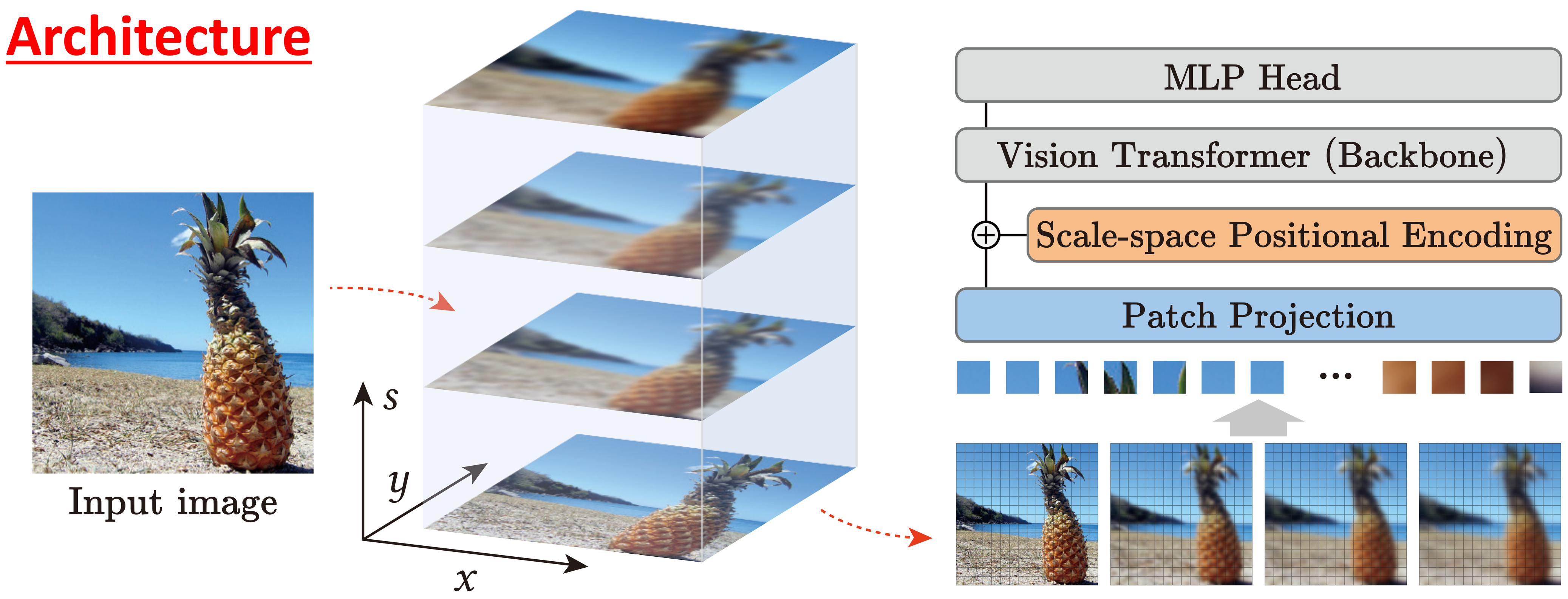


Shape bias

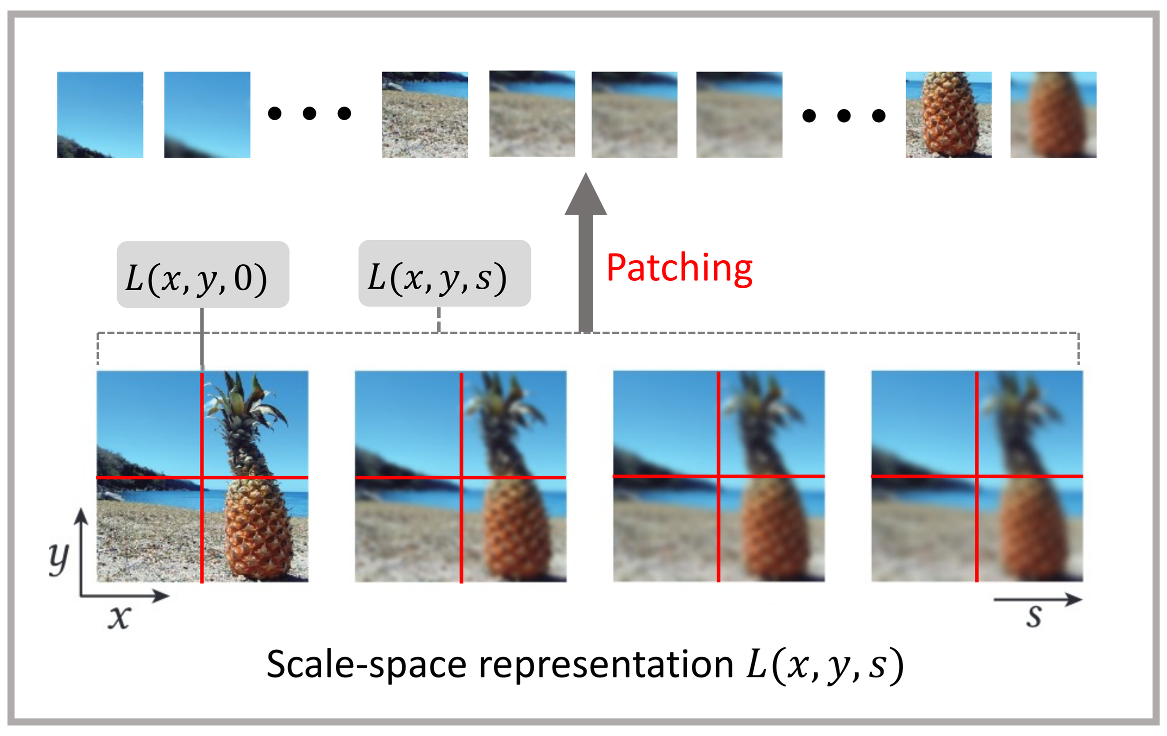


Texture bias

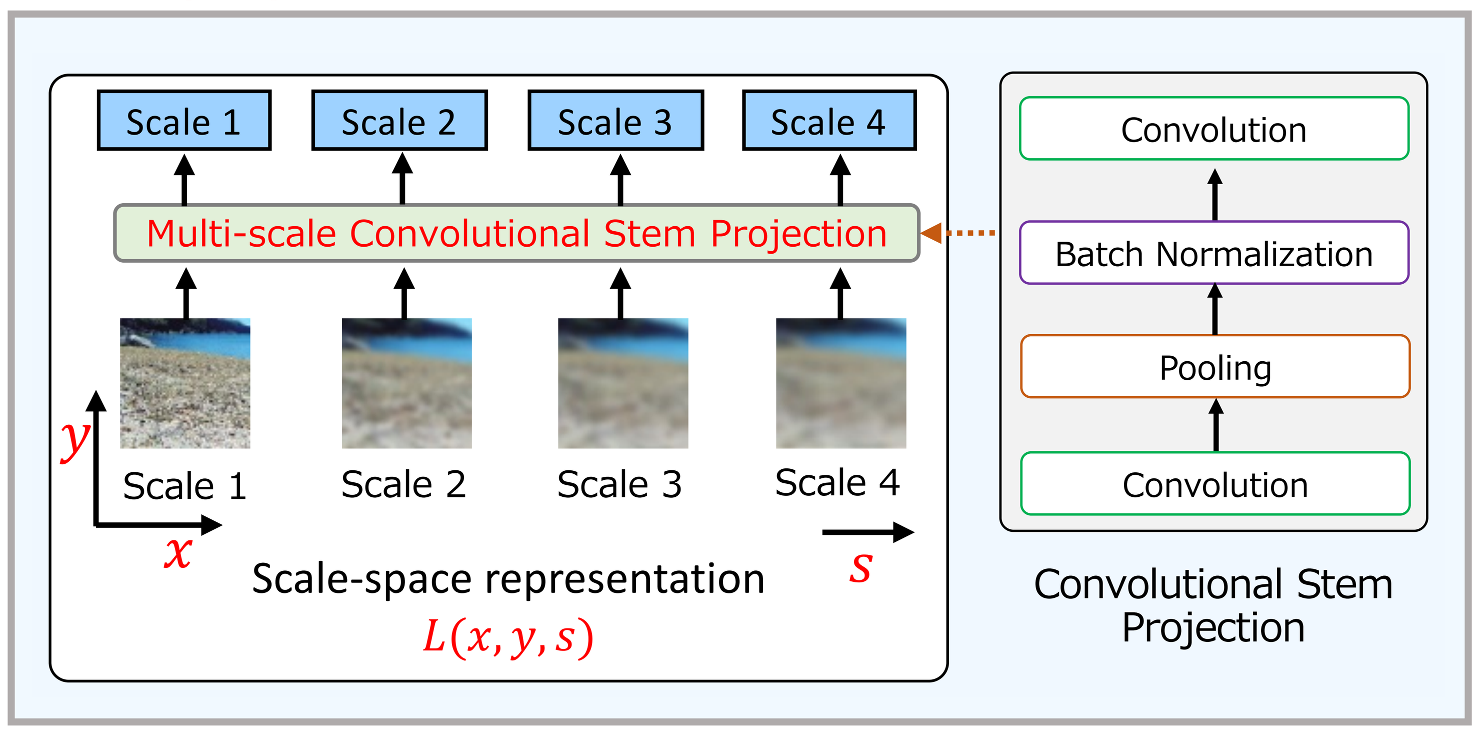
Architecture



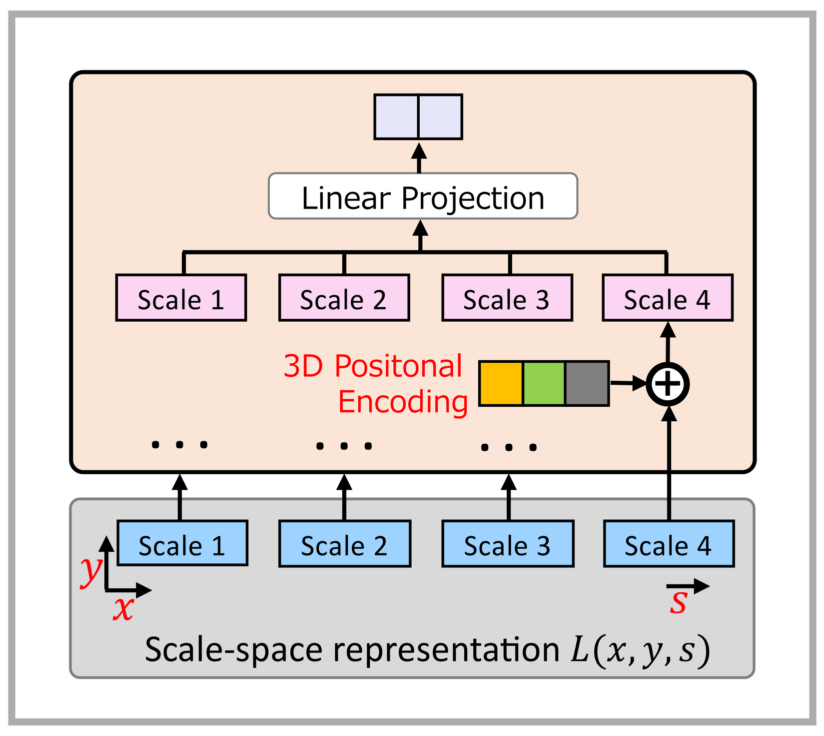
Patching



Patch projection



Scale-space positional Encoding



Experiments

We report the standard accuracy and robustness performance on ImageNet-1k.

Our model: Scale-space-based Robust Vision Transformer (SRVT)
Baseline model: Robust Vision Transformer (RVT) [Mao+ 22]
Evaluation metrics: Top-1 Accuracy (clean, FGSM and PGD), mCE on ImageNet-C
Model sizes: SRVT-Ti (8.6M), SRVT-S (22.1M), SRVT-M (49.1M)
Results: Our model architecture design significantly enhances the robustness against adversarial perturbations and common corruptions.

Model	mCE↓	Gauss.	Shot	Imp.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Bright	Cont.	Elas.	Pixel	JPEG
RVT-Ti	58.2	**48.4**	**50.0**	49.3	66.8	78.5	61.8	73.2	53.8	59.2	48.1	48.4	42.3	73.2	**58.5**	58.8
SRVT-Ti	**56.8**	48.5	50.4	**48.7**	**64.4**	**76.3**	**61.1**	**72.0**	**53.6**	**55.7**	**46.9**	**47.2**	**39.6**	**70.5**	59.2	**58.1**
RVT-S	50.1	41.5	43.4	**41.3**	58.4	71.4	**55.2**	67.0	47.2	50.0	38.8	40.8	35.2	64.6	**47.0**	50.2
SRVT-S	**49.3**	**41.3**	**43.2**	41.7	**57.1**	**68.7**	56.0	**66.7**	**45.8**	**45.7**	**37.4**	**39.8**	**33.7**	**63.9**	48.9	**49.9**
SRVT-M	48.4	40.8	42.1	38.9	56.3	67.6	50.7	49.4	44.6	48.0	38.7	40.2	34.2	62.6	47.9	49.7

Corruption error on ImageNet-C

Model	Params (M)	IN-1k Top-1	Robustness Benchmarks FGSM↑	PGD↑	IN-C↓
ResNet50 [19]	25.6	76.1	12.2	0.9	76.7
Inception-v3 [52]	27.2	77.4	22.5	3.1	80.6
RegNetY-4GF [48]	20.6	79.2	15.4	2.4	68.7
EfficientNet-B4 [53]	19.3	83.0	44.6	18.5	71.1
ResNeXt50 [64]	25.0	79.8	34.7	13.5	64.7
DeepAugment [21]	25.6	75.8	27.1	9.5	53.6
ANT [50]	25.6	76.1	17.8	3.1	63.0
AugMix [23]	25.6	77.5	20.2	3.8	65.3
AA CNN [72]	25.6	79.3	32.9	13.5	68.1
Debiased CNN [36]	25.6	76.9	20.4	5.5	67.5
DeiT-S [54]	22.1	79.9	40.7	16.7	54.6
ConViT-S [7]	27.8	81.5	41.0	17.2	49.8
Swin-T [39]	28.3	81.2	33.7	7.3	62.0
PVT-Small [58]	24.5	79.9	26.6	3.1	66.9
PiT-S [25]	23.5	80.9	41.0	16.5	52.5
TNT-S [18]	23.8	81.5	33.2	4.2	53.1
T2T-ViT_t-14 [68]	21.5	81.7	40.9	11.4	53.2
RVT-S [45]	22.1	81.7	51.3	26.2	50.1
SRVT-S (Ours)	22.1	**82.0**	**55.5**	**32.9**	**49.3**

Comparison with SOTA models

Conclusion and Future Work

We introduced a simple yet effective approach, scale-space tokenization, to improve adversarial and out-of-distribution robustness. Future work could explore other factors that enhance the robustness and incorporate them into the scale-space representation.