A APPENDIX

A.1 DETAILS OF THE EDGE DETECTOR

The edge details are extracted by the CNN-based edge detector using richer convolutional features (RCF) proposed in Liu et al. (2017), which produces edge maps as side outputs of different layers of the network. The level of details of the edge maps gradually become coarser from the early layers to the final layers of the network. We employ here a ResNet101-based RCF edge detector which was trained on the BSDS500 dataset Arbelaez et al. (2010) and use the pre-trained network to extract edge maps from images. The ResNet-101 architecture is composed of four stages, and we select the side output of the second stage to generate the edge maps, as we observed that these edge maps contain the essential shape details, as shown in Figure A1.

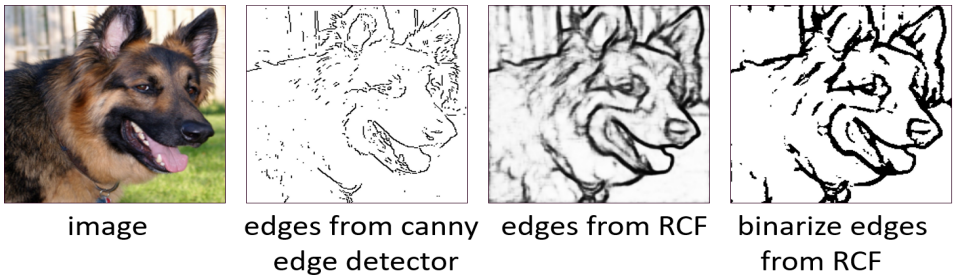


Figure A1: Illustration of a natural image, and edge maps extracted with Canny edge detector Canny (1986), with RCF Liu et al. (2017), and with binarized edges from RCF.

A.2 IMAGENET20 DATASET

We use a subset of 20 classes from ImageNet dataset to study the influence of the shape bias on robustness towards corruptions. It comprises a total of 25784 training and 1000 validation images. This subset consists of animal classes (`african elephant`, `german shepherd`, `tabby cat`, `arabian camel`, `tailed frog`, `scorpion`), birds (`king penguin`, `albatross`), insects (`fly`, `sulphur butterfly`), man made objects (`tea pot`, `stop watch`, `teddy bear`, `fur coat`), automobile (`sports car`, `trolley bus`, `life boat`) and edible items (`mushroom`, `bell pepper`, `pretzel`).

A.3 TECHNIQUES FOR THE EVALUATION OF SHAPE BIAS

Figure A2 depicts the two techniques that are used in our work to evaluate the shape bias. Figure A2a points to the **shuffled image patches** technique that perturb the shape details by preserving the local texture. On the other hand, Figure A2b show examples of **texture-shape cue conflict images** that test the network's bias towards shape or texture.

A.4 CONVOLUTIONAL FILTERS

We show in Table 1 and Table 2 that the network E demonstrates stronger shape bias than IN and SIN. In Figure A3, we visualize the filters of first convolutional layer of networks IN, SIN and E to understand the behavior of these networks. As seen in the figure, filters of E strongly resembles the Gabor filters for edge detection compared to other networks. These results suggest that E extracts features that corresponds to the shape information in the form of edges and effectively based its decision on shape details. The filters of E are non-colored because E is trained on edge maps whereas IN and SIN are trained on natural and stylized images respectively.

A.5 ADDITIONAL EXPERIMENTAL RESULTS

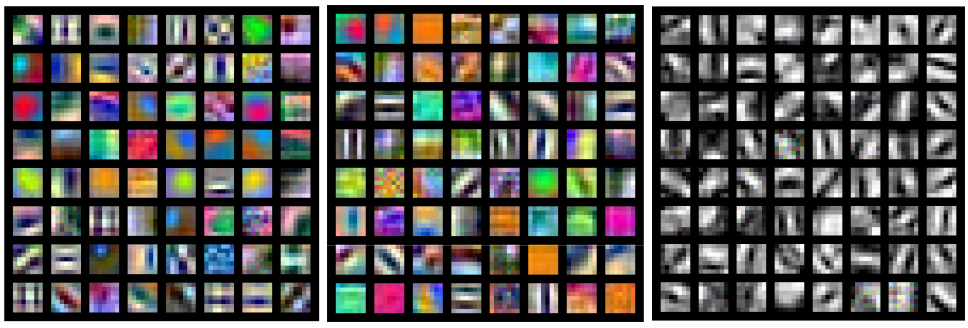
A.5.1 PATCH SHUFFLED RESULTS ON THE WHOLE VALIDATION DATA

We extend the results of *shuffled image patches* in Table 1 and 2 on the whole validation data and are shown in Table A1 and A2 respectively.

(a) The image is divided into $n \times n$ patches and the patches are randomly shuffled. The global object shape is increasingly perturbed with larger n.

(b) Images with conflicting shape and texture cues. The images are obtained by applying style transfer with a texture image as style source.

Figure A2: Techniques to evaluate the shape bias of networks



(a) Standard CNN (IN) (b) Stylized CNN (SIN) (c) Edge CNN (E)

Figure A3: Filters of first convolutional layer of Standard CNN (IN), Stylized CNN (SIN), Edge CNN (E) (ours). The filters of E strongly resembles Gabor filters for edge detection than other two networks.

A.5.2 RESNET18 ON IMAGENET200

We present an additional evaluation on ImageNet200 (200 TinyImageNet classes but all the images in full resolution from ImageNet) in Table A3: E-SIN has higher shape bias but demonstrates lower mean Corruption Accuracy(mCA) than SIN. This reinforces our core finding of shape bias not implying corruption robustness on a larger dataset. Here, all models have clean validation accuracy of about 70%.

A.5.3 RESNET18 WITH BATCHNORM ON IMAGENET20

We provide results on ResNet18 wih BatchNorm in Table A4 showing that it leads to the same findings as the ones reported in the paper for Group Normalization + Weight Standardization: Both networks E and E-SIN exhibit more shape bias but lower mean Corruption Accuracy (mCA) than SIN. Superposition (SE+IN) has comparable mean corruption accuracy with SIN despite showing lower shape bias on patch shuffled and cue conflict evaluation. Here, all models below exhibit validation accuracies of 89%-90%. Note that the architecture ResNet18 wih BatchNorm does not include Group Normalization and Weight Standardization layers.

Network	shuffled image patches 4×4 (%)		
	No styling	style blending	style randomization
IN	57.8	44.9	36
SIN	33.1	32.2	29.8
E	28.3	28.4	**23.43**

Table A1: Comparison of different feature space style augmentation methods on 4×4 shuffled image patches on whole validation data.

Network	shuffled image patches (%)		
	2×2	4×4	8×8
IN	66.7	36	26.7
SIN	63.5	29.8	17.6
E	61.8	23.4	9.8
SE	**57**	23.7	10
E-SIN	59	**20.1**	**9.1**

Table A2: Comparison of models trained on different datasets on shuffled image patches evaluated on whole validation data.

A.5.4 RESNET50 ON IMAGENET20

We present our analysis on deeper architecuture like ResNet50 with Group Normalization + Weight Standardization on ImageNet20 in Table A5: E and E-SIN with higher shape bias have lower mean corruption accuracy than SIN, whereas SE+IN with lower shape bias reach similar accuracy as SIN. These results show that our findings apply to deeper architectures. Here, all models have 86%-87% clean validation accuracy.

A.5.5 DENSENET121 AND MOBILENETV2 ON IMAGENET20

We extend our findings from ResNet to other architectures like DenseNet121 and MobileNetV2 on ImageNet20 with style randomization. Table A6 and Table A7 present the shape-based evaluation and mean corruption accuracy (mCA) of DenseNet121 and MobileNetV2 on different settings, respectively. Similar to the results on ResNet, we find that E-SIN exhibits a higher shape bias than other settings but still has a lower mCA than SIN. On the other hand, Superposition (SE+IN) shows a lower shape bias with similar mCA to SIN. These results show that our conclusion is also valid using different neural architectures.

Note that these results also include additional intermediate settings E-IN and E+IN. The explanation of these two settings is as follows:

- E-IN: This is a network training setting similar to E-SIN. In E-SIN, network is pretrained on Edge dataset (E) in the first stage of training for 75 epochs and later finetune on both the stylized images (SIN) and original images (IN) for another 75 epochs in the second stage (please refer Section **??** for training details). Similarly, E-IN also pretrains the network on Edge dataset (E) in the first stage but later finetune only on the original images (IN) in the second stage. This setting also shares similarity with network setting E, where network pretrains on edge maps during the first stage and later finetune on both edge maps and original images in the second stage of training.
- E+IN: Similar to Superposition (SE+IN), this dataset variant E+IN interpolates edge maps I_{E} from Edge dataset (E) with images I_{IN} from ImageNet20 (IN): $I_{\text{E+IN}} := (1 - \alpha) \cdot I_{\text{E}} + \alpha \cdot I_{\text{IN}}$. We set $\alpha = 0.5$. Similar to SE+IN, this setting also pretrains the network on E+IN images in the

Network	shuffled image patches(%)			Cue conflict shape #880	Mean corruption accuracy (%)
	2×2	4×4	8×8		
IN	78.6	45.7	15.6	90	35.8
SIN	61.7	17.7	3.7	273	**52.4**
E-SIN	**54.3**	**10.6**	**1.8**	**337**	47.7

Table A3: Shape based evaluation & corruption accuracies with ResNet18 on 200 classes of ImageNet. Patch shuffled evaluation is conducted on 5474 correctly classified validation images by all the networks.

Network	shuffled image patches(%)			Cue conflict shape #400	Mean corruption accuracy (%)
	2×2	4×4	8×8		
IN	89.9	72.9	41.9	58	57.8
SIN	**81.3**	47.7	19.6	151	**72.9**
E	**81.3**	**39**	**9.5**	128	50.8
E-SIN	84.3	41	11.2	**169**	67.5
SE+IN	86.1	60.8	33.5	91	**72**

Table A4: Shape based evaluation & corruption accuracies of ResNet18 networks with BatchNorm. Patch shuffled evaluation is conducted on 775 correctly classified validation images by all the networks.

first stage of training and later finetunes on both E+IN and also on the original images (IN) in the second stage.

A.5.6 Evaluation on edge maps of validation set

We evaluate the performance of models trained under different training settings on edge map based validation set of ImageNet20 and the validation accuracy results are presented in Table A8. Note that among different settings in our experiments, edge maps of training data are directly used in E throughout the training, also used for pretraining the network in E-SIN, and stylized edge maps of training data are used during training of Superposition (SE+IN) (please refer Section A.3 for training details). On the other hand, edge maps are not used in any way for training in IN and SIN. From the results of Table A8, we can observe that the setting E has the highest edge map-based validation accuracy among all others as the edge maps have been used in the entire training process.

A.6 Significance of interpolation parameter in superposition

In Section 4 under *Stylization variants*, we discussed about studying the **role of natural image statistics** by interpolating images I_{SE} from SE with images I_{IN} from IN: $I_{\text{SE+IN}} := (1-\alpha) \cdot I_{\text{SE}} + \alpha \cdot I_{\text{IN}}$. We show in Table 3 that such setup (SE+IN) with $\alpha = 0.5$ outperforms SIN despite having lower shape bias than all the other networks. In Figure A4, we show the performance of SE+IN at different values of α on the mean corruption accuracy on all 15 ImageNet-C corruptions at different severity levels. Here, $\alpha = 0$ corresponds to the network trained only on SE whereas $\alpha = 1$ on IN. As shown in the figure, the two extreme set of values of α i.e., α being very smaller or very larger result drop in performance on corruptions. This implies that the images with balanced details of natural image statistics and style variations is essential for improved performance on corruptions.

Network	shuffled image patches(%)			Cue conflict shape #400	Mean corruption accuracy (%)
	2×2	4×4	8×8		
IN	88.8	74.1	50.7	65	58
SIN	83.2	46	20.5	141	**78.2**
E	**72.5**	34.5	13.7	164	52.2
E-SIN	73.8	**25.7**	**8.4**	**209**	72.3
SE+IN	85.5	66.4	36.5	102	**78.1**

Table A5: Shape based evaluation & corruption accuracies with ResNet50 architecture.

Network	shuffled image patches(%)			Cue conflict shape #400	Mean corruption accuracy (%)
	2×2	4×4	8×8		
IN	**62.9**	37.9	23.5	55	62.7
SIN	68.4	42.5	21.7	154	**80.3**
E	65.2	35.8	14.0	167	60.2
E-SIN	66.0	35.4	**12.0**	**219**	77.7
SE+IN	68.3	51.7	32.0	118	79.0
E-IN	64.5	**33.7**	12.7	164	62.4
E+IN	69.0	49.4	31.1	69	65.0

Table A6: Shape based evaluation & corruption accuracies with DenseNet121 architecture. All models exhibit clean validation accuracies of 88%-91%. Patch shuffled evaluation is conducted on 761 correctly classified validation images by all the networks. It can be observed from these results that the SE+IN have similar mCA to SIN despite having lower shape bias than both SIN and E-SIN.

A.7 Performance of different networks on common corruptions

We show in Section 6 that networks E, SE, E-SIN performs poorly on corruptions despite having stronger shape bias than SIN. We also show that superposition of SE with natural images IN (SE+IN) slightly outperforms SIN even having lower shape bias respectively. In Figure A5, we show that these results are consistent across all 15 ImageNet-C distortions at different severity levels.

A.8 Finetuning affine parameters on different corruptions

As mentioned in Section 7, affine parameters of normalization layers in pre-trained IN are fine-tuned on corruptions from ImageNet-C separately. Here, fine-tuning only the affine parameters of IN on a respective corruption greatly improves the mean corruption accuracy on the same corruption across different severity levels. The affine parameters are fine-tuned on `speckle noise`, `gaussian blur`, `snow`, `frost`, `fog`, `brightness`, `contrast`, `elastic transform`, `pixelate`, `jpeg compression` separately and resulting 10 different IN networks. Each of these fine-tuned networks are evaluated on the same corruption or similar category of corruptions. For e.g, a network fine-tuned on `frost` is evaluated only on `frost` and network fine-tuned on `speckle noise` is also evaluated on the other set of noises like `gaussian noise`, `shot noise`, `impulse noise`. Note that training data of ImageNet20 is augmented with respective corruptions to fine-tune

Network	shuffled image patches(%)			Cue conflict shape #400	Mean corruption accuracy (%)
	2×2	4×4	8×8		
IN	53.2	27.5	21.8	90	56.5
SIN	50.6	20.4	9.0	189	**72.2**
E	52.6	18.7	11.4	160	56.0
E-SIN	**49.1**	**16.1**	**6.0**	**212**	69.0
SE+IN	57.4	31.6	20.0	138	70.0
E-IN	52.4	19.0	9.0	148	56.6
E+IN	56.4	33.0	20.0	94	60.3

Table A7: Shape based evaluation & corruption accuracies with MobileNetV2 architecture. All models exhibit clean validation accuracies of 86%-88.5%. Patch shuffled evaluation is conducted on 700 correctly classified validation images by all the networks. Here, SE+IN shows high corruption robustness close to the one of SIN despite having a lower shape bias than SIN and E-SIN. The reason for the gap between SIN and SE+IN corruption accuracy can be explained as follows: stylized dataset SIN is seen as strongly augmented dataset than the superposition SE+IN (notice perceptual differences in Figure 2) and unlike larger architectures like ResNet18/DenseNet121, the compact nature of the MobileNetV2 architecture is shown to benefit from such stronger data augmentation than the superposition.

Network	Architectures		
	ResNet18 (%)	DenseNet121 (%)	MobileNetV2 (%)
IN	18.5	22.1	31.8
SIN	48.6	51.2	57.8
E	**77.9**	**80.1**	**74.5**
E-SIN	72.4	71.9	67.2
SE+IN	47.8	51.6	54.8

Table A8: Evaluation of different networks on edge map based validation dataset. It can observed that the network E has the highest edge map-based validation accuracy as the edge maps have been used in the entire training process. Here, SE+IN has higher validation accuracy than IN, comparable to SIN. The reason SE+IN has higher validation accuracy is that the stylized edges are used during training of the SE+IN setting.

the affine parameters. Each training sample in a mini-batch is augmented with the corresponding corruption at a randomly chosen severity level. The severity parameters that are already pre-defined for every severity level in ImageNet-C are used. A total of 50 epochs are used for fine-tuning the affine parameters on a corruption, starting with learning rate 0.01 and reduce it to 0.001 after 45 epochs. Performance of the networks that are fine-tuned and evaluated on the same or similar category of corruptions at different severity levels are presented in Figure A6. As shown in the figure, adapting just the distribution on the learned representations from the standard ImageNet20 is sufficient to achieve high performance on respective distribution of corruptions.

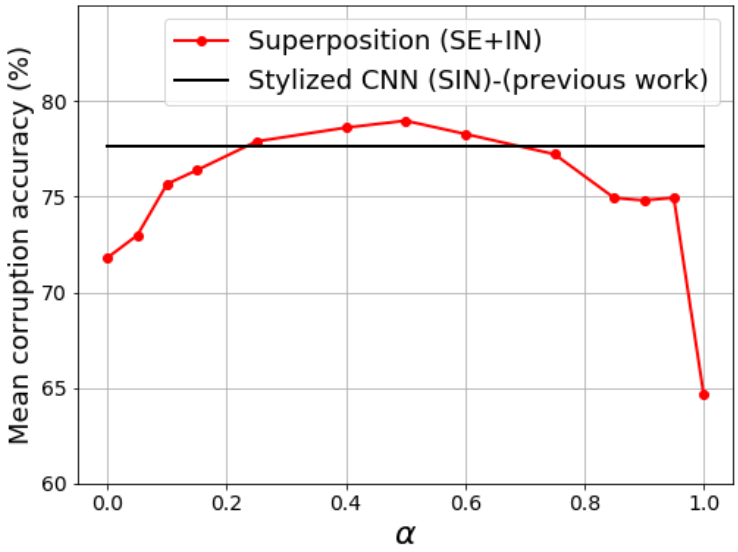


Figure A4: Mean corruption accuracy on ImageNet-C corruptions at different values of α in a stylized dataset SE+IN. Here $\alpha \in \{0, 0.05, 0.1, 0.15, 0.25, 0.4, 0.5, 0.6, 0.75, 0.85, 0.9, 0.95, 1\}$. The solid black line represents the performance of the baseline SIN.

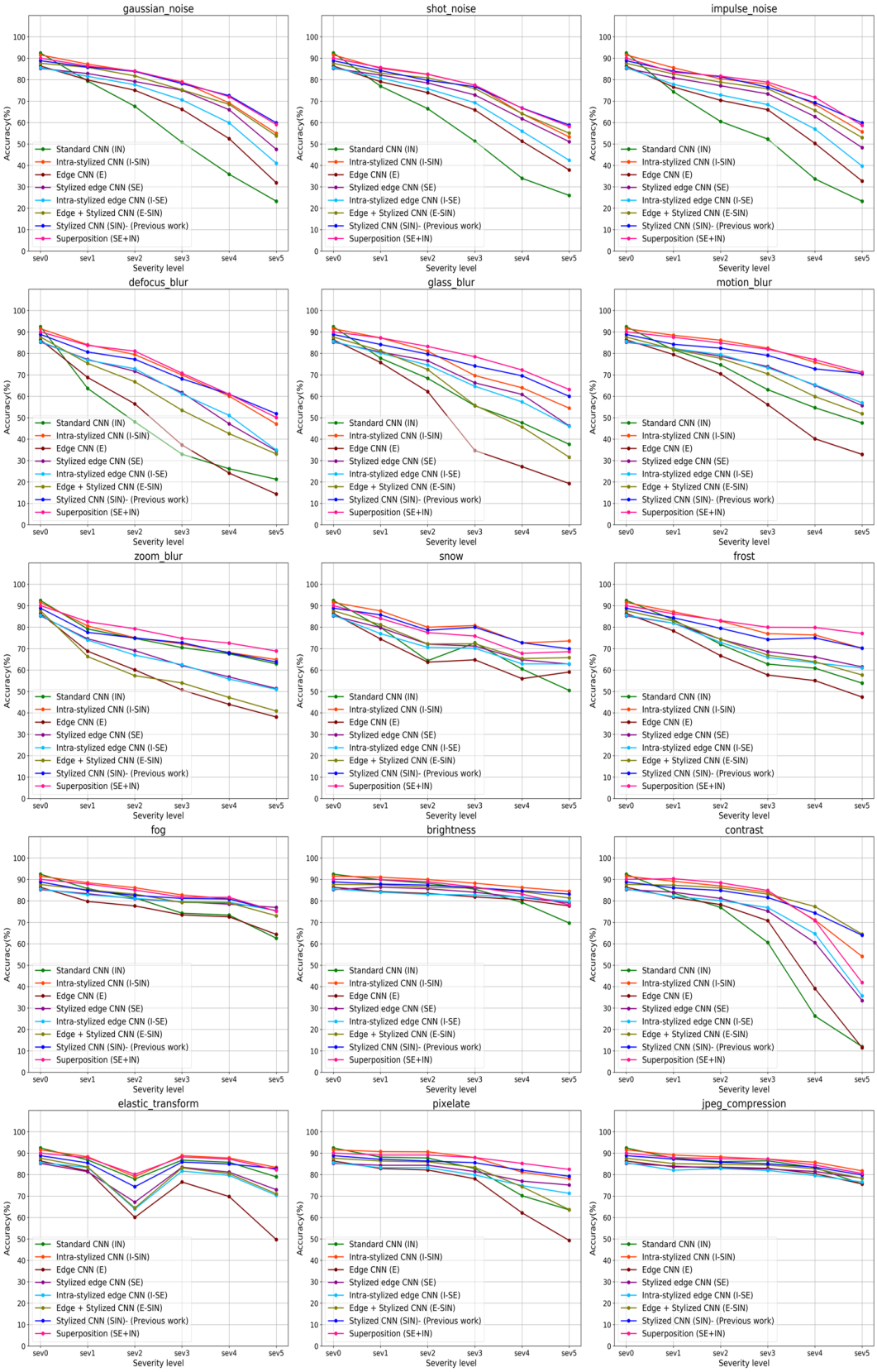


Figure A5: Performance of different networks on ImageNet-C corruptions at different severity levels

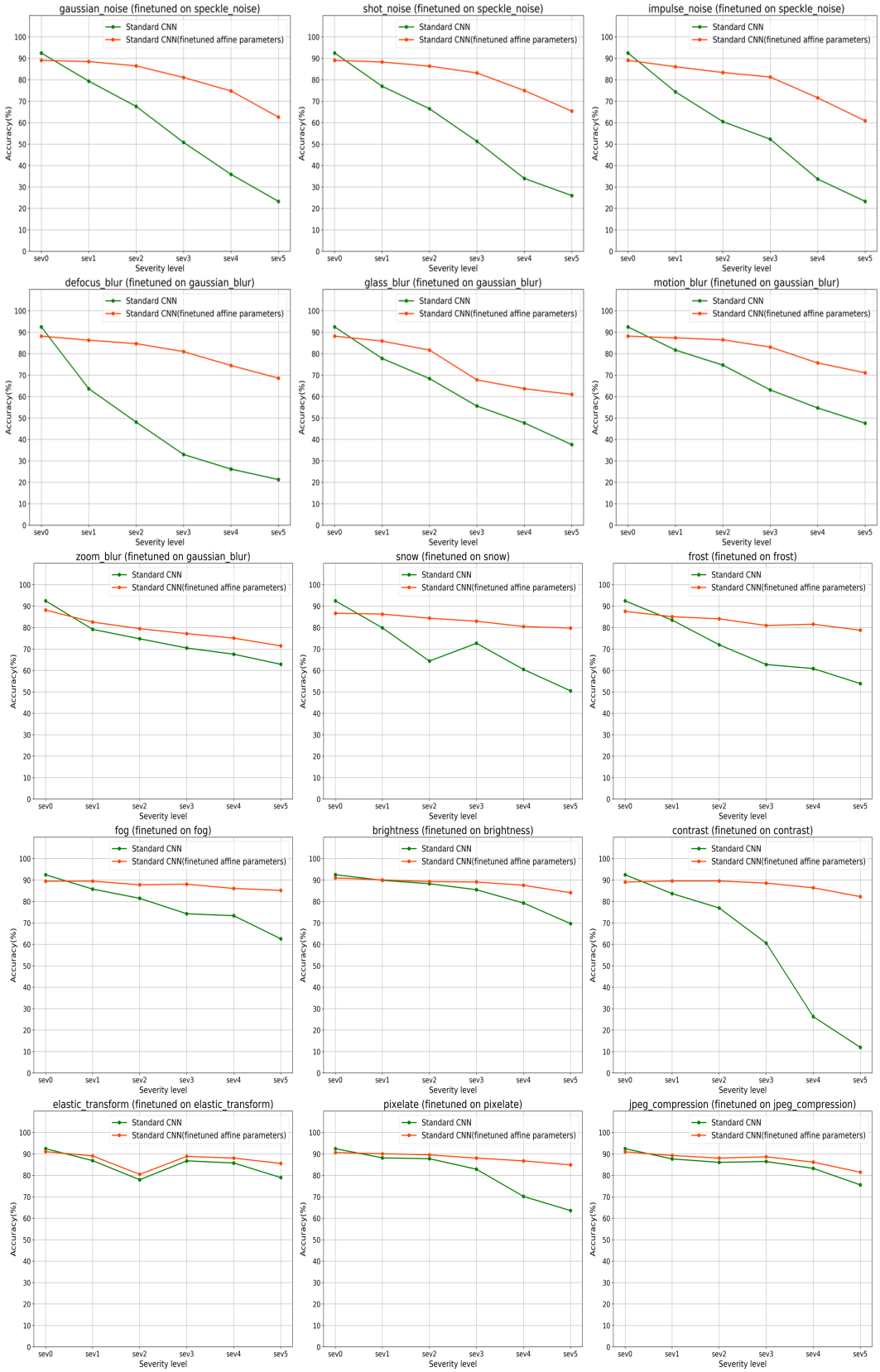


Figure A6: Performance of standard network IN on ImageNet-C corruptions when finetuned the affine parameters on the same corruptions