Robustness Analysis on Foundational Segmentation Models

Supplementary Material

The supplementary will provide more details about the benchmark datasets, models, and additional results. In Section 1, we provide more details on how our real-world perturbations were generated for the MS COCO-P and ADE20K-P dataset. In Section 2 we provide additional results from our benchmark.

1. Distribution Shift Perturbations

We used 17 types of algorithmically generated corruptions to generate a perturbed dataset. These corruption are from different categories like noise, blur, environment, digital and camera. We have given a overview plot for all perturbation in Figure 10,11,12.

Noise. We have Gaussian, shot, impulse, and speckle noise in the noise category. Gaussian noise is modeled by adding random values sampled from a Gaussian distribution to the pixel intensities of a clean image, the standard deviation of the Gaussian noise determines the severity. Shot noise is modeled by applying Poisson distribution to the pixel values of the clean image. Impulse noise is modeled by adding salt and pepper noise to the clean image, and the density of the noise determines the severity. Speckle noise is generated by adding a normal noise distribution whose intensities are proportional to the clean image pixel intensities, and the standard deviation of the noise determines the severity.

Blur. We have defocus, motion, and zoom blur in the Blur category. Defocus blur is modeled by convolving the clean image with a blur kernel, here the blur kernel is a circular Gaussian blur kernel, and the blur radius determines the severity of noise. Zoom blur is modeled by averaging multiple zoomed images generated by scaling up the image and cropping out of the boundary region to maintain the original shape. Here, the list of scaling factors used determines the severity.

Compression. In the digital category, we have jpeg and pixelate corruption. Jpeg corruption is generated by saving the image in jpeg format by reducing the quality, and the quality determines the severity. Pixelate corruption is modeled by upsampling a low-resolution image, and the severity is controlled by how much it was downsampled before upsampling.

Digital Contrast corruption is generated by blending a clean image with another image in which all pixel values are set to the mean value of the clean image. Here the blending factor determines the severity. Shear corruption is generated with the help of imgaug [25].

Camera In the geometric category, we have translate and rotate. Both translate and rotate are implemented with the help of imgaug [25] library to generate corrupted images and their corresponding annotations.

Environment Darkness corruption is modeled by blending a black image with a clean image with a blending factor determined by severity. We additionally have snow and fog corruption which are algorithmically generated images that try to mimic real-life fog and snow.

Fog, Snow, motion blur, brightness, and shear perturbations are all implemented using imgaug [25] library.

2. Additional Results

Here we provide additional results and more details on the robustness scores and performance of the selected models. Absolute Robustness (γ^a) scores are additionally included here and are the absolute drop in performance while Relative Robustness (γ^r) is the relative drop based on the original model score.

2.1. Instance Segmentation

Table 1 and Table 2 respectively shows results for absolute robustness scores γ^a and relative robustness scores γ^r for the selected models. γ^a measures the absolute drop in performance as compared to γ^r which measures relative drop to original performance of a given model. These results are averaged across severity for each corruption type. One observation is that when comparing results for ADE20K-P where ODISE and SAM are evaluated zero-shot, absolute robustness is much higher than relative. This indicates that while models may start with lower performance overall, they show more consistent results across perturbations. More details on model behavior across severity for instance segmentation are shown in Figure 1 on MS COCO-P and Figure 2 for ADE20K-P where multimodal models are evaluated on zero-shot. On MS COCO-P, we see very similiar trends across all corruptions except for compression-based. For both JPEG and Pixelate, we see a some different trends for ODISE showing a sudden drop at severity 3. For ADE20K, where multimodal are evaluated zero-shot, we see more consistent results across severity and more declines from the Mask2Former model. This supports the conclusion that of the selected multimodal models, while their zero-shot performance is low, their absolute robustness across severity is good and performance consistent. Table 3 presents the object super-category wise robustness scores for both γ^a and γ^r. We observe that multimodal models are noticeabley more relatively robust in certain object categories.

Table 1. **Absolute Robustness scores (γ^a) for instance segmentation** on models on the MS COCO-P and ADE20K-P dataset. Models with the least relative drop in performance are in bold, and models that are second least are underlined.

COCO (γ^a)	**Environment**				**Digital**		**Compression**		**Pixel Noise**				**Camera**		**Blur**		
	dark	bright	snow	fog	shear	contrast	jpeg	pixel.	speckle	gauss.	shot	impulse	rotate	translate	motion	defocus	zoom
Mask2Former+R50	0.99	0.96	0.78	0.93	0.94	0.96	0.81	0.79	0.89	0.79	0.91	0.74	0.85	0.96	0.78	0.78	0.77
MaskDINO+R50	0.99	0.96	0.77	0.93	0.94	0.96	0.81	0.78	0.88	0.78	0.91	0.76	0.85	0.96	0.78	0.79	0.76
Mask2Former+swinL	0.99	0.97	0.89	0.97	0.94	0.98	0.89	0.88	0.94	0.87	0.95	0.88	0.89	0.95	0.81	0.81	0.77
MaskDINO+swinL	0.99	0.97	0.90	0.97	0.94	0.97	0.90	0.89	0.94	0.87	0.95	0.88	0.89	0.95	0.82	0.81	0.77
VitL$_a dapter$	1.00	0.98	0.88	0.96	0.94	0.98	0.92	0.90	0.94	0.85	0.95	0.86	0.89	0.95	0.83	0.82	0.79
ODISE+Caption	**1.00**	0.98	0.89	0.97	0.96	0.98	0.86	0.87	0.95	0.88	0.96	0.87	0.92	0.98	0.86	0.85	0.83
ODISE+Label	1.00	0.98	0.88	0.97	0.95	0.97	0.83	0.85	0.94	0.87	0.95	0.86	0.90	0.97	0.83	0.82	0.79
Prompt+SAM	1.00	0.98	0.92	0.98	0.95	0.98	0.86	0.89	0.94	0.87	0.95	0.89	0.90	0.96	0.84	0.84	0.82
InternImage-XL	0.99	0.98	0.89	0.98	0.94	0.98	0.91	0.88	0.94	0.88	0.95	0.89	0.88	0.95	0.83	0.82	0.79
PAINTER	1.00	0.98	0.90	0.97	**0.98**	**0.99**	**0.94**	**0.95**	**0.96**	**0.91**	**0.97**	**0.91**	**0.94**	**0.99**	**0.92**	**0.90**	**0.89**
GroundedSam+swinB	1.00	**0.98**	**0.92**	**0.98**	0.96	0.98	0.89	0.90	0.94	0.88	0.96	0.89	0.91	0.97	0.86	0.86	0.84
ADE20K (γ^a)	**Environment**				**Digital**		**Compression**		**Pixel Noise**				**Camera**		**Blur**		
	dark	bright	snow	fog	shear	contrast	jpeg	pixel.	speckle	gauss.	shot	impulse	rotate	translate	motion	defocus	zoom
Mask2Former+swinL	0.97	0.98	1.00	0.85	0.97	0.87	0.88	0.93	0.86	0.93	0.86	0.97	0.94	0.86	0.92	0.99	0.84
Mask2Former+R50	0.96	0.98	1.00	0.85	0.93	0.81	0.79	0.92	0.84	0.90	0.80	0.93	0.88	0.79	0.86	0.98	0.85
ODISE+Caption	**0.99**	**0.99**	**1.00**	**0.95**	**0.99**	**0.96**	**0.95**	**0.95**	**0.95**	**0.96**	**0.96**	**0.99**	**0.99**	0.95	0.98	**1.00**	**0.95**
ODISE+Label	0.99	0.99	1.00	0.95	0.99	0.95	0.95	0.95	0.95	0.96	0.95	0.99	0.98	**0.95**	**0.98**	1.00	0.94
GroundedSam+swinB	0.98	0.99	1.00	0.92	0.98	0.93	0.93	0.94	0.92	0.95	0.93	0.99	0.97	0.93	0.96	0.99	0.91

Table 2. **Relative Robustness scores (γ^r) for instance segmentation** on models on the MS COCO-P and ADE20K-P. Models with the least relative drop in performance are in bold, and models that are second least are underlined.

COCO (γ^r)	**Environment**				**Digital**		**Compression**		**Pixel Noise**				**Camera**		**Blur**		
	dark	bright	snow	fog	shear	contrast	jpeg	pixel.	speckle	gauss.	shot	impulse	rotate	translate	motion	defocus	zoom
Mask2Former+R50	0.98	0.90	0.50	0.83	0.86	0.91	0.57	0.53	0.74	0.52	0.80	0.41	0.66	0.91	0.49	0.49	0.47
MaskDINO+R50	0.98	0.90	0.49	0.84	0.86	0.92	0.56	0.51	0.73	0.52	0.79	0.45	0.65	0.90	0.50	0.52	0.47
Mask2Former+swinL	0.99	0.95	0.79	0.94	0.89	0.95	0.78	0.76	0.88	0.73	0.90	0.76	0.78	0.91	0.63	0.62	0.54
MaskDINO+swinL	0.99	0.95	0.81	0.94	0.88	0.95	0.79	0.78	0.87	0.73	0.90	0.76	0.78	0.90	0.64	0.63	0.55
VitL$_a dapter$	0.99	0.95	0.73	0.92	0.87	0.96	0.82	0.78	0.87	0.68	0.89	0.70	0.76	0.90	0.64	0.62	0.55
ODISE+Caption	**1.00**	0.95	0.72	0.93	0.91	0.94	0.64	0.67	0.87	0.69	0.89	0.67	**0.79**	0.95	0.63	0.62	0.55
ODISE+Label	0.99	0.95	0.75	0.94	0.89	0.95	0.63	0.67	0.88	0.71	0.90	0.69	0.78	0.92	0.62	0.62	0.55
Prompt+SAM	1.00	0.95	**0.81**	**0.96**	0.89	0.96	0.69	0.76	0.87	0.71	0.90	0.74	0.78	0.91	0.65	**0.65**	0.59
InternImage-XL	0.99	0.95	0.77	0.95	0.88	**0.96**	**0.82**	0.76	**0.88**	**0.75**	**0.91**	**0.78**	0.76	0.90	0.64	0.64	0.57
PAINTER	0.98	0.94	0.65	0.89	**0.93**	0.96	0.78	**0.82**	0.86	0.69	0.89	0.69	0.78	**0.97**	**0.72**	0.65	**0.60**
GroundedSam+swinB	0.99	**0.96**	0.78	0.95	0.90	0.96	0.71	0.74	0.86	0.69	0.88	0.71	0.76	0.92	0.63	0.63	0.58
ADE20K (γ^r)	**Environment**				**Digital**		**Compression**		**Pixel Noise**				**Camera**		**Blur**		
	dark	bright	snow	fog	shear	contrast	jpeg	pixel.	speckle	gauss.	shot	impulse	rotate	translate	motion	defocus	zoom
Mask2Former+swinL	0.92	0.95	1.00	0.56	0.91	0.61	0.64	**0.80**	0.59	**0.79**	0.58	0.93	0.82	0.60	0.77	0.98	0.54
Mask2Former+R50	0.84	0.91	0.99	0.44	0.74	0.27	0.20	0.68	0.41	0.63	0.24	0.75	0.56	0.22	0.47	0.93	0.42
ODISE+Caption	**0.92**	**0.95**	**1.01**	**0.65**	**0.94**	0.68	0.67	0.67	**0.68**	0.74	**0.68**	**0.95**	**0.90**	0.64	0.85	**0.99**	**0.61**
ODISE+Label	0.91	0.93	0.99	0.63	0.92	**0.69**	**0.68**	0.67	0.64	0.74	0.67	0.92	0.88	**0.67**	**0.86**	0.99	0.59
GroundedSam+swinB	0.87	0.93	0.98	0.55	0.91	0.57	0.60	0.67	0.57	0.70	0.59	0.93	0.80	0.60	0.75	0.95	0.50

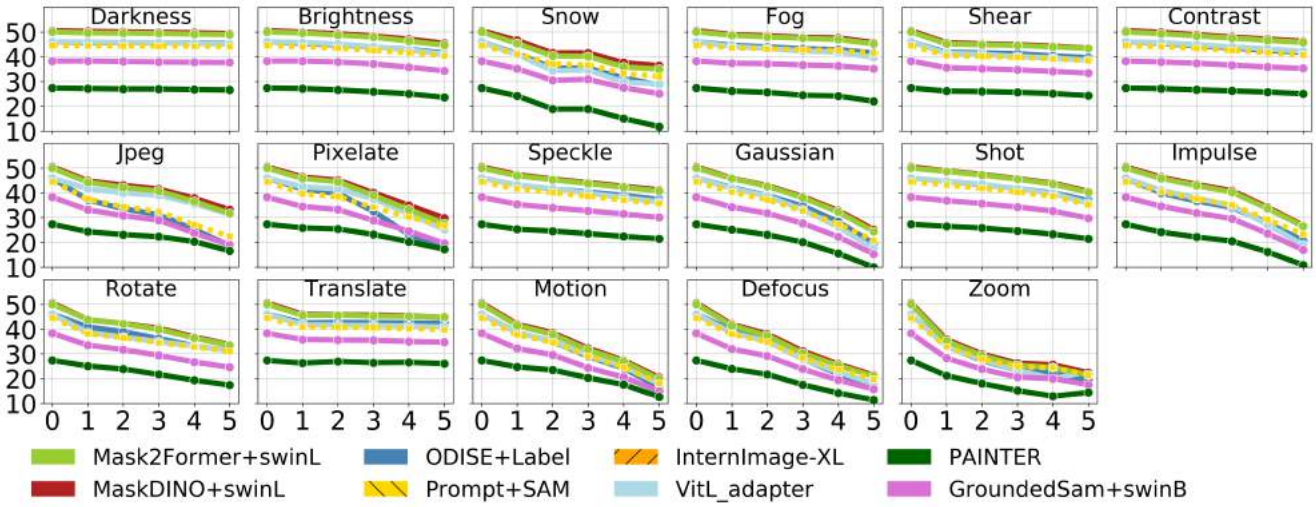

Figure 1. **Results for each corruption and each severity for instance segmentation measured by average precision (AP) on the MS COCO-P dataset**. x-axis: Severity ranges from 0 (no corruption) to 5 (most corruption). y-axis: AP results for instance segmentation.

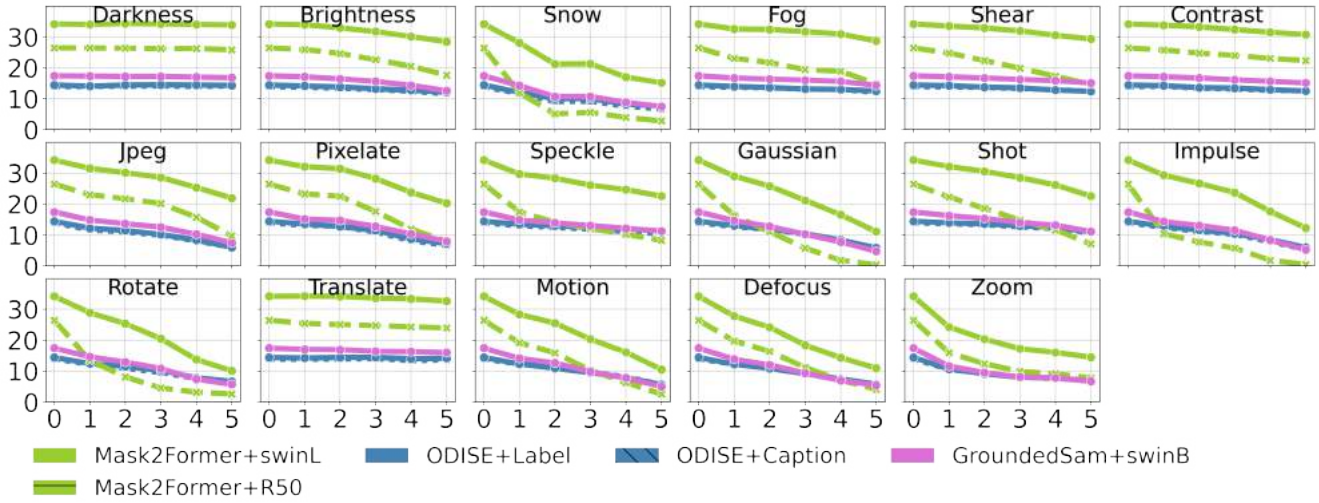

Figure 2. **Results for each corruption and each severity for instance segmentation measured by average precision (AP) on the ADE20K-P dataset.** x-axis: Severity ranges from 0 (no corruption) to 5 (most corruption). y-axis: AP results for instance segmentation.

Table 3. **Relative robustness scores (γ^r) and Absolute robustness scores (γ^a) for each object super-category for instance segmentation on MS COCO-P.** Here the scores are averaged across all corruptions, severity for each model.

γ^r	accessory	animal	appliance	electronic	food	furniture	indoor	kitchen	outdoor	person	sports	vehicle
Mask2Former+R50	0.62	0.71	0.65	0.68	0.65	0.68	0.61	0.58	0.71	0.73	0.62	0.68
MaskDINO+R50	0.63	0.71	0.65	0.68	0.64	0.68	0.61	0.58	0.73	0.74	0.64	0.68
Mask2Former+swinL	0.74	0.86	0.80	0.80	0.82	0.84	0.76	0.72	0.82	0.82	0.74	0.81
MaskDINO+swinL	0.75	**0.86**	0.83	**0.81**	0.83	0.85	0.75	0.71	0.82	0.82	0.74	0.82
VitL$_a$*dapter*	0.73	0.85	0.81	0.78	0.81	0.81	0.73	0.71	0.82	0.81	0.74	0.80
ODISE+Caption	0.73	0.81	**1.02**	0.75	0.81	0.81	0.75	0.70	0.80	0.82	0.71	0.80
ODISE+Label	0.75	0.82	0.78	0.77	0.78	0.82	0.74	0.70	0.80	0.83	0.71	0.81
Prompt+SAM	0.76	0.83	0.88	0.81	**0.87**	**0.92**	0.74	0.71	0.82	0.77	0.74	0.77
InternImage-XL	0.75	0.85	0.86	0.80	0.82	0.84	0.75	**0.72**	0.83	0.82	0.75	**0.82**
PAINTER	**0.77**	0.83	0.76	0.76	0.78	0.78	**0.80**	0.70	0.84	**0.87**	0.76	0.81
GroundedSam+swinB	0.71	0.85	0.83	0.76	0.83	0.89	0.76	0.70	**1.10**	0.82	**0.93**	0.78
γ^a	accessory	animal	appliance	electronic	food	furniture	indoor	kitchen	outdoor	person	sports	vehicle
Mask2Former+R50	0.87	0.83	0.83	0.83	0.87	0.89	0.88	0.87	0.87	0.87	0.86	0.86
MaskDINO+R50	0.87	0.83	0.83	0.83	0.86	0.89	0.88	0.86	0.89	0.87	0.86	0.86
Mask2Former+swinL	0.89	0.92	0.88	0.89	0.92	0.94	0.91	0.89	0.92	0.90	0.89	0.91
MaskDINO+swinL	0.89	0.92	0.90	0.89	0.92	0.94	0.90	0.88	0.91	0.90	0.89	0.91
VitL$_a$*dapter*	0.90	0.91	0.90	0.88	0.93	0.93	0.90	0.89	0.92	0.90	0.90	0.91
ODISE+Caption	0.92	0.89	0.93	0.89	0.94	0.95	0.93	0.93	0.93	0.91	0.90	0.92
ODISE+Label	0.91	0.89	0.89	0.88	0.92	0.93	0.91	0.90	0.91	0.91	0.89	0.91
Prompt+SAM	0.90	0.90	**0.94**	0.89	0.95	**0.98**	0.91	0.90	0.92	0.89	0.90	0.90
InternImage-XL	0.90	0.91	0.92	0.89	0.92	0.94	0.91	0.89	0.92	0.90	0.90	0.92
PAINTER	**0.96**	**0.93**	0.88	**0.92**	**0.96**	0.95	**0.97**	**0.96**	**0.96**	**0.96**	**0.95**	**0.94**
GroundedSam+swinB	0.91	0.92	0.91	0.88	0.94	0.97	0.92	0.91	0.95	0.92	0.79	0.91

2.2. Semantic Segmentation

Table 4 and 5 show relative robustness (γ^r) and absolute robustness (γ^a) scores for the selected models on semantic segmentation. While comparing, we do find that all selected models are typically more robust to semantic segmentation as opposed to instance segmentation, but still multimodal models perform poorly for compression, snow and certain noises in comparison to transformer based unimodal in MS COCO-P dataset. We observe that CNN models on the ADE20K-P dataset are even less robust compared to MS COCO-P. Additionally, the ODISE model is more relatively robust on ADE20K-P, where it is evaluated zero-shot, as compared to MS COCO-P. To better observe performance across varying severity for each corruption, we visualize results for COCO-P in 3 and ADE20K-P in 4. For both datasets, but especially ADE20K-P, we see CNN-based backbones for unimodal models have a much steeper decline as corruption severity increases, most noticeably for noise-based corruptions. This may indicate CNN-based ar-

chitectures are more sensitive to noise-based corruptions.

2.3. Fine-tuning on Corrupted Dataset

The fine-tuning dataset comprises a subset of the ADE20K training dataset, consisting of 8000 images, which is consistent for all fine-tuning. The first 2000 are clean, while the remaining 6000 are randomly augmented using perturbations from the specific category we are targeting. Figure respectively shows the performance of Mask2Former and ViT-Adapter.

2.4. Qualitative Examples

We show examples of model predictions in Figures 7, 9 and 8. Figure 7 shows an image from the COCO-P dataset under *JPEG* compression with severity 1, 3, and 5. As severity increases, mask quality and the number of objects decreases. This is more noticeable with ODISE where it additionally classifies objects. Figure 8 shows the same but under the *snow* corruption. Models are typically more robust to *snow* as compared to *JPEG*, but show some decrease in performance as severity increases as shown in Figure 1. Here we see mask quality persist but the number of smaller objects classified and masked decrease. Figure 9 shows the same but for *zoom blur*, a corruption all models are low in robustness to. Again we see as severity increases, ODISE misclassifies some objects. However, even with the low robustness to blur, we see the mask quality is still visually higher when compared to *JPEG*.

Table 4. **Relative Robustness scores (γ^r) for models on the MS COCO-P and ADE20K-P dataset for semantic segmentation**. Models with the least relative drop in performance are in bold, and models that are second least are underlined.

COCO (γ^r)	**Environment**				**Digital**		**Compression**		**Pixel Noise**				**Camera**		**Blur**		
	dark	bright	snow	fog	shear	contrast	jpeg	pixel.	speckle	gauss.	shot	impulse	rotate	translate	motion	defocus	zoom
Mask2Former+R50	0.98	0.92	0.46	0.82	0.97	0.95	0.64	0.60	0.75	0.56	0.81	0.43	0.78	0.98	0.70	0.72	0.69
MaskDINO+R50	0.98	0.91	0.45	0.81	0.96	0.95	0.61	0.61	0.73	0.54	0.79	0.44	0.77	0.98	0.70	0.73	0.68
Mask2Former+swinL	1.00	**0.97**	0.83	**0.96**	0.99	0.97	0.91	0.91	0.93	**0.84**	**0.95**	0.86	**0.93**	0.99	0.89	**0.86**	0.83
MaskDINO+swinL	0.99	0.97	**0.86**	0.96	**0.99**	0.97	**0.91**	**0.92**	**0.93**	0.84	0.94	**0.86**	0.93	0.98	**0.89**	0.86	**0.83**
ODISE+Caption	0.99	0.96	0.76	0.93	0.98	0.97	0.77	0.82	0.90	0.76	0.92	0.73	0.89	0.99	0.83	0.80	0.79
ODISE+Label	**1.00**	0.97	0.78	0.95	0.99	**0.98**	0.77	0.82	0.91	0.78	0.93	0.75	0.91	**0.99**	0.84	0.82	0.80
PAINTER	0.99	0.95	0.65	0.91	0.99	0.97	0.85	0.90	0.89	0.76	0.92	0.76	0.88	0.99	0.87	0.80	0.78
ADE20K (γ^r)	**Environment**				**Digital**		**Compression**		**Pixel Noise**				**Camera**		**Blur**		
	dark	bright	snow	fog	shear	contrast	jpeg	pixel.	speckle	gauss.	shot	impulse	rotate	translate	motion	defocus	zoom
Mask2Former+R50	**1.00**	0.92	0.29	0.78	0.88	0.97	0.80	0.71	0.53	0.32	0.62	0.26	0.36	0.98	0.54	0.62	0.55
MaskDINO+R50	0.99	0.90	0.30	0.78	0.84	0.96	0.76	0.69	0.48	0.29	0.56	0.24	0.33	0.95	0.54	0.63	0.55
Mask2Former+swinL	1.00	**0.97**	0.71	**0.95**	0.98	**0.99**	0.92	0.91	0.87	0.77	0.91	0.79	0.73	1.00	0.82	0.79	0.74
VitL-adapter	1.00	0.97	0.78	0.95	**0.98**	0.98	**0.92**	**0.93**	0.92	0.78	0.94	0.79	0.79	0.99	0.84	0.81	0.76
ODISE+Label	0.99	0.97	**0.78**	0.94	0.97	0.98	0.84	0.88	**0.95**	**0.83**	**0.95**	0.80	**0.84**	**1.01**	**0.85**	**0.82**	**0.79**
InternImage-H	1.00	0.96	0.73	0.94	0.97	0.98	0.90	0.91	0.89	0.79	0.92	**0.81**	0.80	0.98	0.82	0.76	0.78
PAINTER	0.98	0.91	0.51	0.86	0.96	0.96	0.89	0.91	0.87	0.76	0.90	0.77	0.65	0.97	0.82	0.74	0.72

Table 5. **Absolute Robustness scores (γ^a) for models on the MS COCO-P and ADE20k-P dataset for semantic segmentation**. Models with the least relative drop in performance are in bold, and models that are second least are underlined.

COCO (γ^a)	**Environment**				**Digital**		**Compression**		**Pixel Noise**				**Camera**		**Blur**		
	dark	bright	snow	fog	shear	contrast	jpeg	pixel.	speckle	gauss.	shot	impulse	rotate	translate	motion	defocus	zoom
Mask2Former+R50	0.99	0.95	0.66	0.89	0.98	0.97	0.78	0.76	0.85	0.73	0.88	0.65	0.86	0.99	0.81	0.83	0.81
MaskDINO+R50	0.99	0.95	0.67	0.89	0.98	0.97	0.76	0.76	0.84	0.72	0.88	0.66	0.86	0.99	0.82	0.83	0.81
Mask2Former+swinL	1.00	**0.98**	0.89	**0.98**	0.99	0.98	0.94	0.94	0.95	**0.89**	**0.96**	0.90	**0.96**	0.99	0.92	**0.91**	0.88
MaskDINO+swinL	1.00	0.98	**0.90**	0.97	**0.99**	0.98	**0.94**	**0.95**	**0.95**	0.89	0.96	**0.91**	0.95	0.99	**0.92**	0.91	0.88
ODISE+Caption	1.00	0.98	0.87	0.96	0.99	0.98	0.88	0.90	0.95	0.88	0.96	0.86	0.94	0.99	0.91	0.90	**0.89**
ODISE+Label	**1.00**	0.98	0.86	0.97	0.99	0.98	0.85	0.88	0.94	0.86	0.96	0.84	0.94	**1.00**	0.90	0.88	0.87
PAINTER	1.00	0.97	0.80	0.95	0.99	**0.99**	0.91	0.94	0.94	0.86	0.96	0.86	0.93	0.99	0.92	0.88	0.87
ADE20K (γ^a)	**Environment**				**Digital**		**Compression**		**Pixel Noise**				**Camera**		**Blur**		
	dark	bright	snow	fog	shear	contrast	jpeg	pixel.	speckle	gauss.	shot	impulse	rotate	translate	motion	defocus	zoom
Mask2Former+R50	**1.00**	0.96	0.67	0.90	0.94	0.99	0.91	0.87	0.79	0.69	0.83	0.66	0.71	0.99	0.79	0.83	0.79
MaskDINO+R50	1.00	0.95	0.66	0.89	0.92	0.98	0.89	0.85	0.75	0.65	0.79	0.63	0.67	0.98	0.77	0.82	0.78
Mask2Former+swinL	1.00	0.98	0.84	0.97	0.99	0.99	**0.96**	0.95	0.93	0.87	0.95	0.89	0.85	1.00	0.90	0.88	0.86
VitL-adapter	1.00	0.98	0.87	0.97	0.99	0.99	0.96	0.96	0.96	0.87	0.96	0.88	0.88	0.99	0.91	0.89	0.86
ODISE+Caption	1.00	0.99	0.91	0.98	0.99	**1.00**	0.95	0.96	0.97	0.94	0.98	0.94	0.94	**1.00**	**0.96**	0.94	**0.94**
ODISE+Label	1.00	**0.99**	**0.93**	**0.98**	**0.99**	0.99	0.95	**0.97**	**0.98**	**0.95**	**0.98**	**0.94**	**0.95**	1.00	0.96	**0.95**	0.94
InternImage-H	1.00	0.98	0.84	0.96	0.98	0.99	0.94	0.95	0.94	0.87	0.95	0.89	0.88	0.99	0.89	0.86	0.87
PAINTER	0.99	0.95	0.76	0.93	0.98	0.98	0.95	0.95	0.94	0.88	0.95	0.88	0.82	0.99	0.91	0.87	0.86

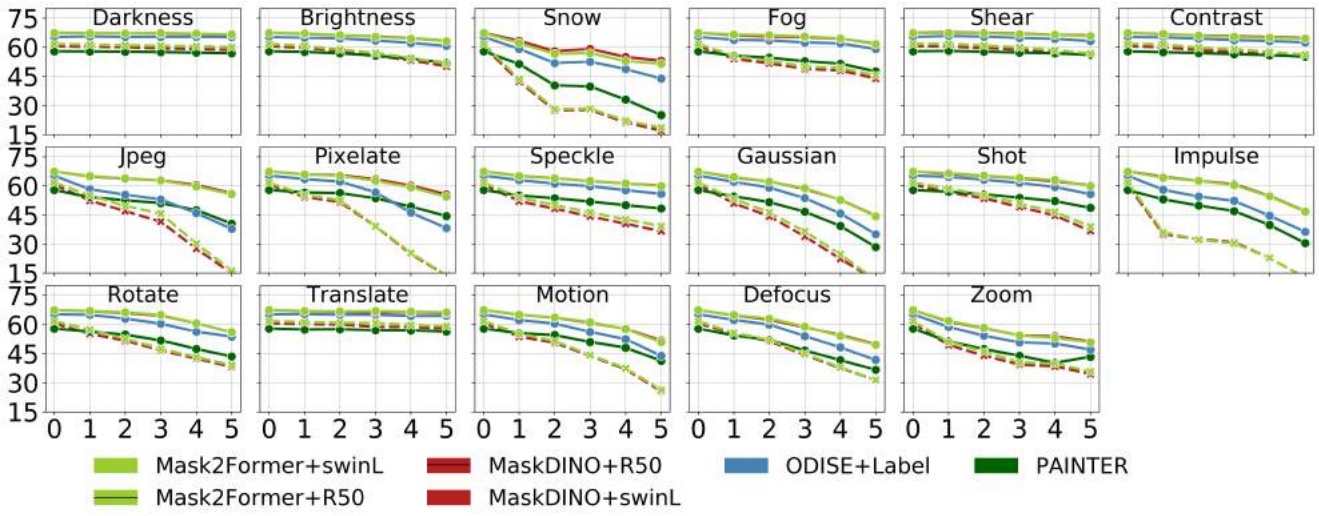

Figure 3. **Results for each corruption and each severity for semantic segmentation measured on the MS COCO-P dataset**. x-axis: Severity ranges from 0 (no corruption) to 5 (most corruption). y-axis: model performance measured by mean intersection-over-union (mIoU).

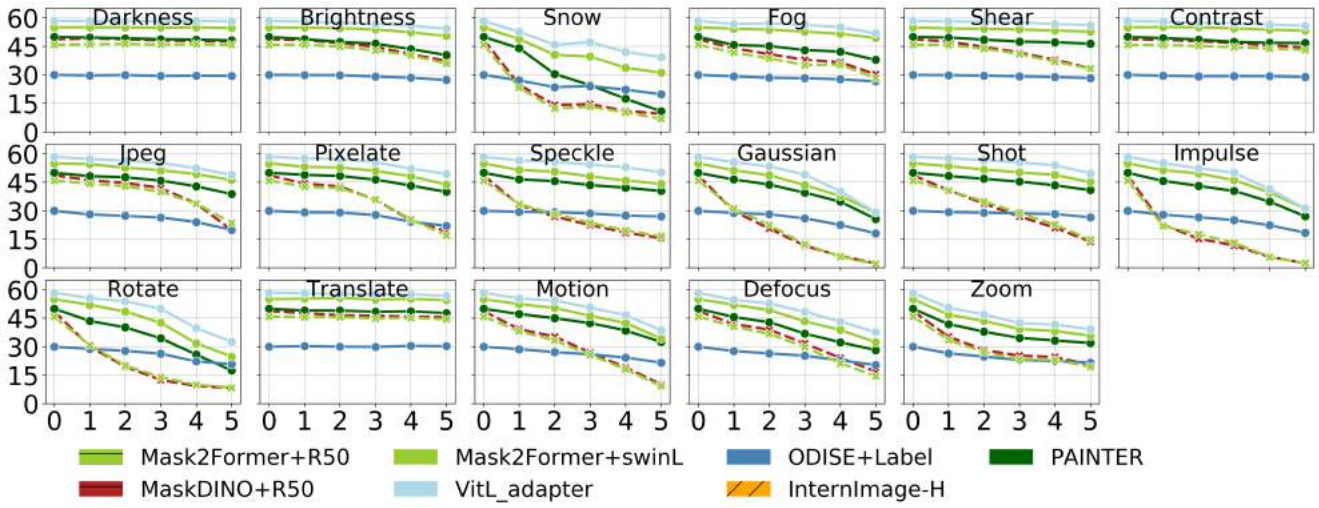

Figure 4. **Results for each corruption and each severity for semantic segmentation measured on the ADE20K-P dataset**. x-axis: Severity ranges from 0 (no corruption) to 5 (most corruption). y-axis: model performance measured by mean intersection-over-union (mIoU).

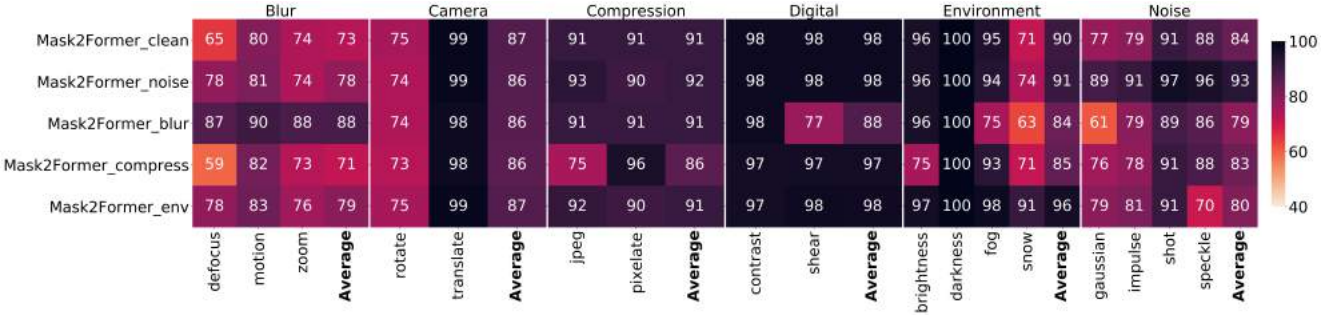

Figure 5. **Fine-tuned performance of Mask2Former on semantic segmentation for the augmented ADE20K-P dataset**. Here y-axis denotes augmentation used for fine-tuning (expect first row) and x-axis denotes models' relative robustness γ^r for each corruption averaged over severity.

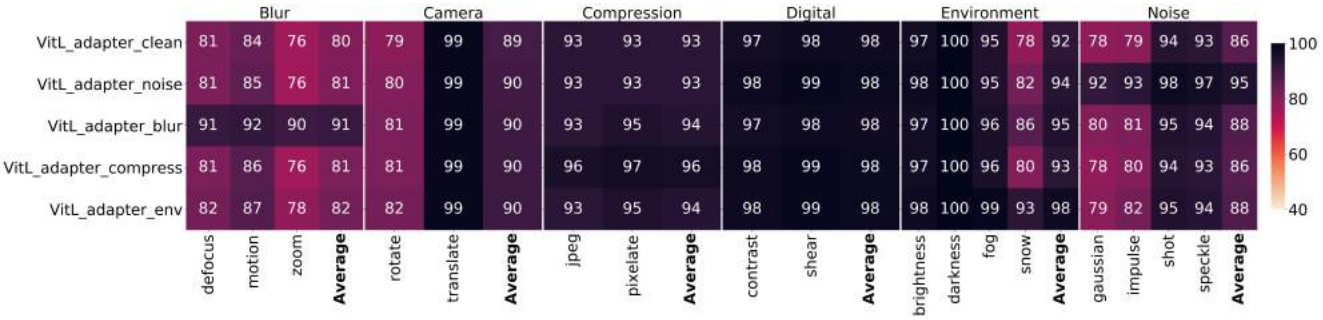

Figure 6. **Fine-tuned performance of ViT-Adapter on semantic segmentation for the ADE20K-P dataset on**. Y-axis: Augmentation used for fine-tuning (expect first row). X-axis: model Relative Robustness score for each corruption averaged over severity.

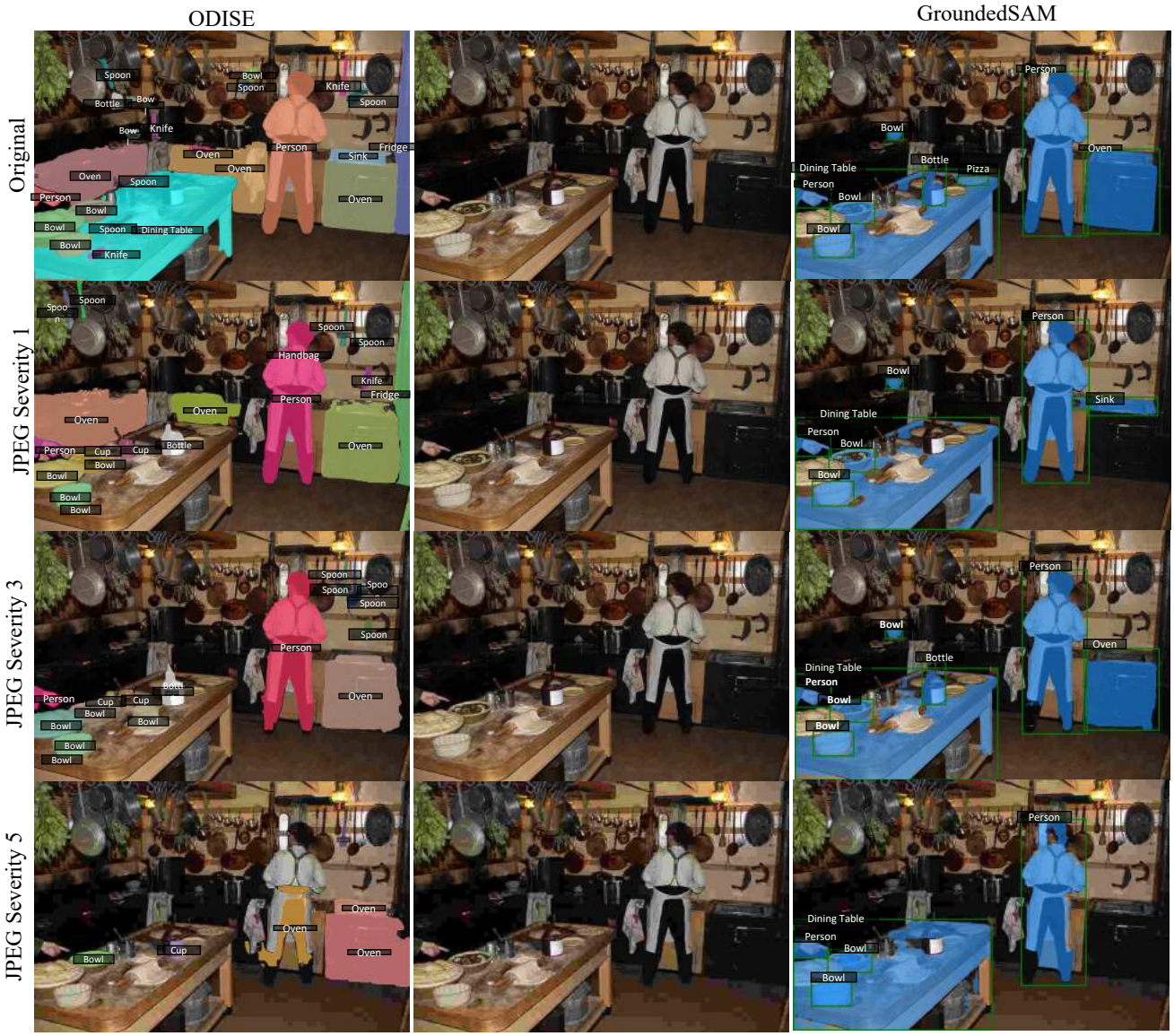

Figure 7. **Visual example from the COCO-P dataset for JPEG compression under varying levels of severity.** The left shows results for ODISE, middle shows the original images, and the right shows GroundedSAM. We again see as severity increases, both models mask quality decreases but ODISE additionally misclassifies objects, such as “person” to “oven”.

Figure 8. **Visual example from the MS COCO-P dataset for Snow corruption under varying levels of severity**. The left shows results for ODISE, middle shows the original images, and the right shows GroundedSAM. We again see as severity increases, both models mask quality decreases but ODISE additionally incorrectly classifies objects, such as "person" to "oven".

Figure 9. **Visual example from the COCO-P dataset for Zoom Blur corruption under varying levels of severity.** The left shows results for ODISE, middle shows the original images, and the right shows GroundedSAM. We again see as severity increases, both models mask quality decreases but ODISE additionally incorrectly classifies objects, such as "person" to "oven".

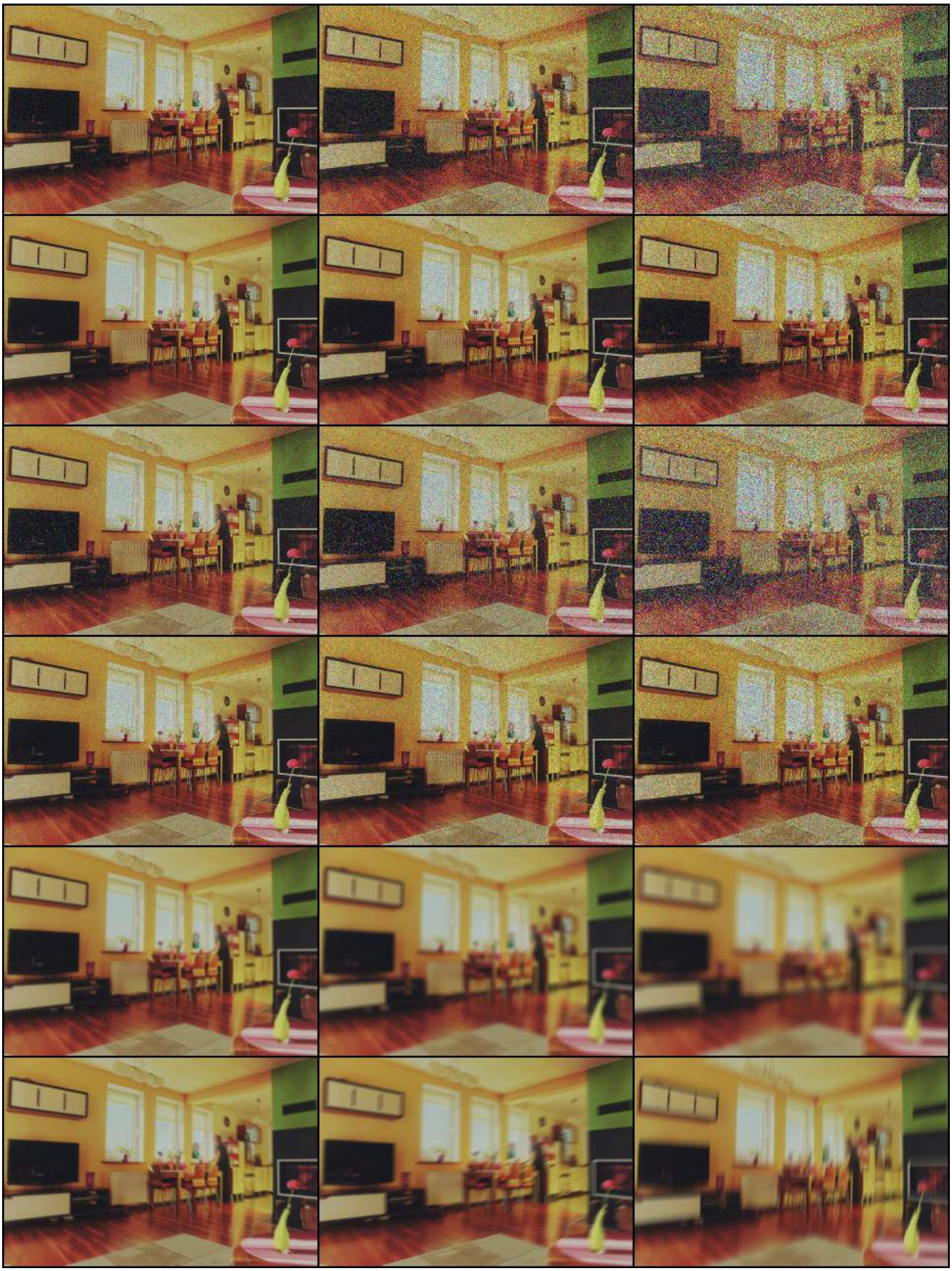

Figure 10. **Visual example from the MS COCO-P dataset for perturbations *gaussian, shot, impulse, speckle, defocus, motion*** across 1, 3, 5 severity.

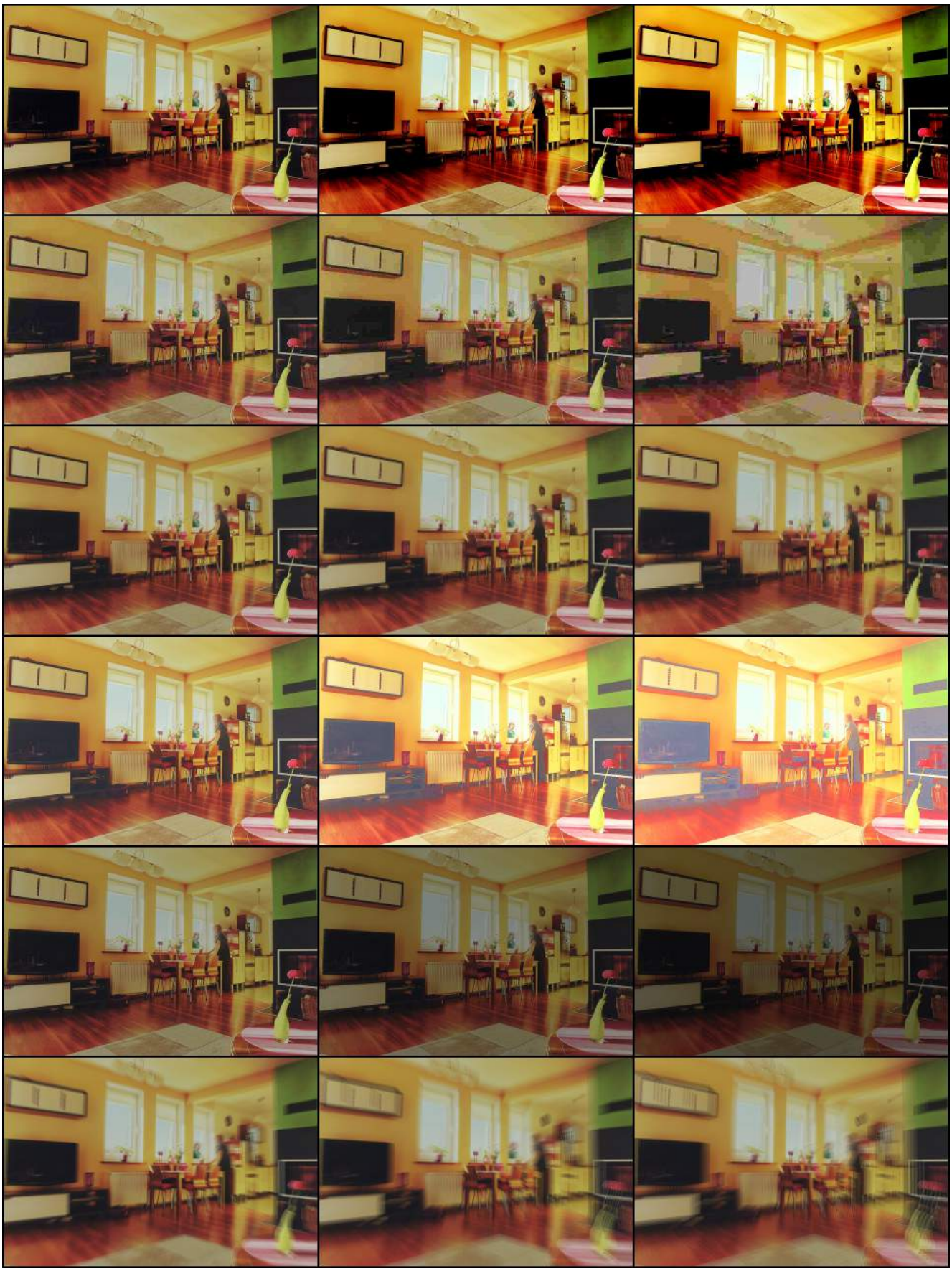

Figure 11. **Visual example from the MS COCO-P dataset for perturbations *contrast, jpeg, pixelate, brightness, darkness, zoom*** across 1, 3, 5 severity.

Figure 12. **Visual example from the MS COCO-P dataset for perturbations *fog, snow, rotate, translate, shear*** across 1, 3, 5 severity.