Supplementary Material
Benchmarking Segmentation Models with Mask-Preserved Attribute Editing

Zijin Yin[1] Kongming Liang[1*] Bing Li[2] Zhanyu Ma[1] Jun Guo [1]
[1] Beijing University of Posts and Telecommunications
[2] King Abdullah University of Science and Technology
[1]{yinzijin2017@, liangkongming@, mazhanyu@, guojun@}bupt.edu.cn [2]bing.li@kaust.edu.sa

1. More details on Mask-Preserved Attribute Editing Pipeline

1.1. Text Manipulation details

In Table 1, we introduce exhaustive prompts used to instruct GPT-3.5 Turbo [1], and edited sentence examples with different attribute variations. We find that adding pre-defined roles (*the professional linguistic assistant*) and examples in text prompts can drastically improve performances.

1.2. Mask-Guided Diffusion details

We leverage the state-of-the-art text-to-image algorithm Latent Diffusion Model [16], a.k.a Stable Diffusion (SD), in which the diffusion process performs in low-dimensional latent space where semantic information can better transfer. It consists of a variational autoencoder network to encode and decode between latent space and pixel space, and denoising network U-Net [17] architecture conditioned on the guiding text prompt to achieve diffusion process. And we integrate our Mask-Guided Attention and ControlNet [25] to the text-guided Image-to-Image Translation approach PnP [19]. In all our results, the Mask-Guided Attention and ControlNet block [25] is integrated into all decoder layers of Stable Diffusion. For integration duration in the denoising process, we utilize two thresholds (i) $\tau_m \in [0, 1]$ is the sample step until Mask-Guided Attention is integrated, (ii) $\tau_c \in [0, 1]$ is the sample step until which ControlNet block [25] is integrated. We set $\tau_m = 0$ since we need to ensure the irrelevant region is not affected in every step of the denoising process. We set $\tau_c = 0.5$ since a large value will diminish the spatial constraint effects of semantic layout labels, and a small value will reduce the reality of edited images. More discussion on τ_c is presented in Sec 5.2. The detailed parameter values in our pipeline are presented in Table 2.

It is noteworthy that the Mask-Guided Attention serves in local attribute editing and does not serve in global attribute editing. This is because in local editing we need to identify the area of edition by the object mask, while in global editing the whole image needs to be changed.

* Corresponding author.

2. More Evaluation Results

2.1. Per-Class analysis

Figure 2 shows the per-class mIoU drop of two segmentation models OCRNet [23] and SEEM [27]. It can be observed that creature classes like dog, cat, and horse are more easily disturbed than inanimate object classes such as bus and train.

2.2. Robustness comparison

To exclude the impact of the performance improvement in original data on our benchmark, we evaluate model robustness by calculating the segmentation accuracy decline in Section 4. In this part, we additionally plot the mIoU accuracy on original Pascal VOC [8] *vs.* our Pascal-EA benchmark. As shown in Figure 1, the CATSeg [4] exhibits the greatest robustness than others.

2.3. More qualitative results

Figures 9, 10 and 11 illustrate qualitative results of segmentation methods under different attribute variations.

3. Analysis of Pascal-EA

In contrast to previous benchmarks like COCO-O [15] and ImageNet-C [9] providing samples from out-of-distribution domains, our Pascal-EA consists of image variations in editable attributes within the in-distribution domain. We measure the extent to which the distribution of our Pascal-EA shifts from that of the original Pascal VOC dataset [8] by the widely-used out-of-distribution detection approach GradNorm [12]. Experimental results in twelve different attribute variations are shown in Figure 3, the x-axis is gradient norms scores and the y-axis is the density of each

Table 1. Illustrations of text prompts used to instruct GPT-3.5 Turbo [1] and edited sentence examples. Edited parts are colored red.

Variation type	Text Prompt	Examples
Local color	*You are a professional linguistic assistant. I will provide you with a sentence. You should first identify the subject of the sentence, and then generate a variation by altering or adding the color attributes of the subject. You should only return the sentence variation. For example, change "a photo of an airplane on the ground" to "a photo of a blue airplane on the ground"*	Input: *a photo of a white and red train.* Output: *a photo of a blue and yellow train.*
Local material	*You are a professional linguistic assistant. I will provide you with a sentence. You should first identify the subject of the sentence, and then generate a variation by changing the material attribute of the subject. The material should be selected from "wooden", "paper", "metallic" and "paper". You should only return the sentence variation. For example, change "a photo of an airplane on the ground" to "a photo of a wooden airplane on the ground"*	Input: *a photo of a white and red train.* Output: *a photo of a wooden white and red train.*
Local pattern	*You are a professional linguistic assistant. I will provide you with a sentence. You should first identify the subject of the sentence, and then generate a variation by changing the material attribute of the subject. The type of patterns should be selected from "dotted", "striped" and "lettered". You should only return the sentence variation. For example, change "a photo of an airplane on the ground" to "a photo of an airplane with stripes on the surface on the ground".*	Input: *a photo of a white and red train.* Output: *a photo of a striped white and red train.*
Global domain:	*You are a professional linguistic assistant. You should generate one possible edition by changing the provided sentence's data domain without changing the content. The data domain should be selected from "oil pastel", "painting", "and sketch" For example, change "a photo of an airplane on the ground" to "a painting of an airplane on the ground"*	Input: *a photo of a white and red train.* Output: *a sketch of a white and red train.*
Global weather:	*You are a professional linguistic assistant. You should generate one possible edition by only adding a weather description to the provided sentence without changing the content. The weather should be selected from "snow", "rain", and "fog". For example: change "a photo of an airplane on the ground" to "a photo of an airplane on the ground on a snowy day".*	Input: *a photo of a white and red train.* Output: *a photo of a red and white train against a backdrop of falling snow*

Table 2. Parameter settings of Mask-Guided Diffusion.

Parameters	Values
Image resolution	512×512
SD version	1.5
Seed	1
Guidance scale	7.5
Inversion timesteps	1000
Diffusion timesteps	50
τ_f [19]	0.8
τ_A [19]	0.5
τ_c (ours)	0.5
τ_m (ours)	0.0

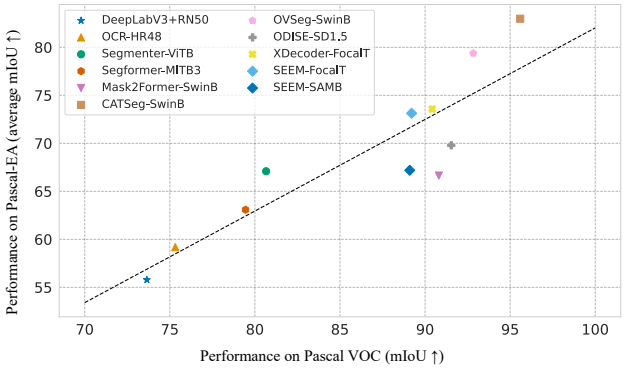

Figure 1. The average performances on our Pascal-EA *vs.* Performances on original Pascal VOC [8]. The black dashed line indicates the linear fit of all segmentation methods.

score, and we calculate the overlap area of two distributions in which heavier overlap indicates closer to original distribution. We observe that our Pascal-EA can proximity the original distribution meanwhile having object attribute variations, which implies can provide reliable evaluations.

To explore whether the diffusion process will degrade the reliability of our evaluations for segmentation, we create a reference set in which we reconstruct the original validation set using our pipeline with non-edited texts. This operation first adds noises to the original images and then denoise them with non-edited texts. We compare model performances on the original validation set and our reference set in Table 5. It is obvious that the diffusion process induces subtle disturbing on segmentation performances which is negligible compared to attribute variation themselves. Such results also serve as evidence that our pipeline's robustness to potential errors in generated caption. In our all experiments in Sec 4, we replace results in the validation set with those in our reconstructed set to remove the effects of perturbations.

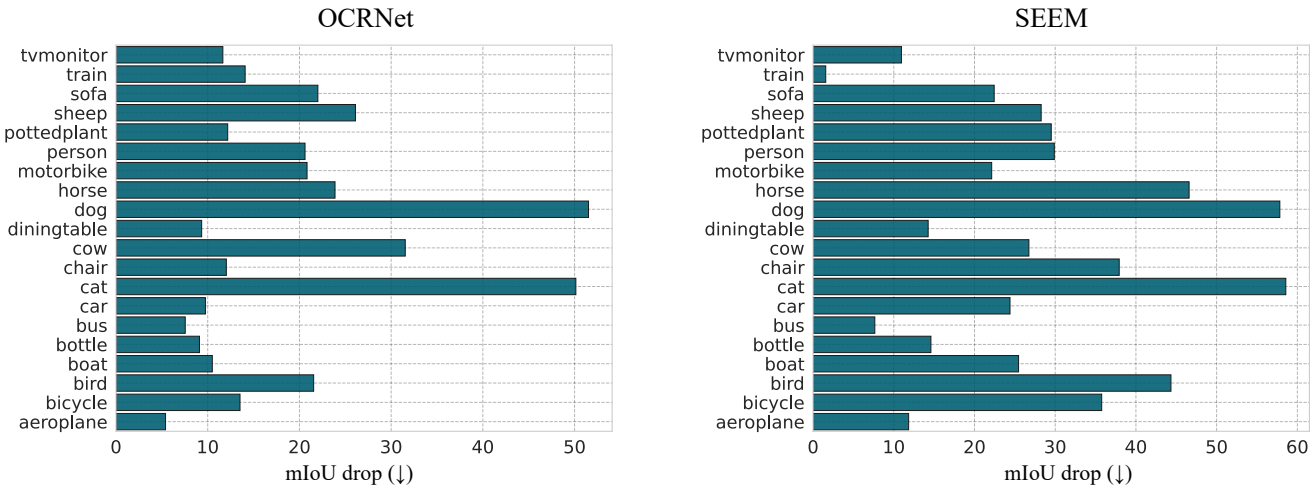

Figure 2. Average mIoU drop (↑) in each class of OCRNet [23] and SEEM [27] under our Pascal-EA.

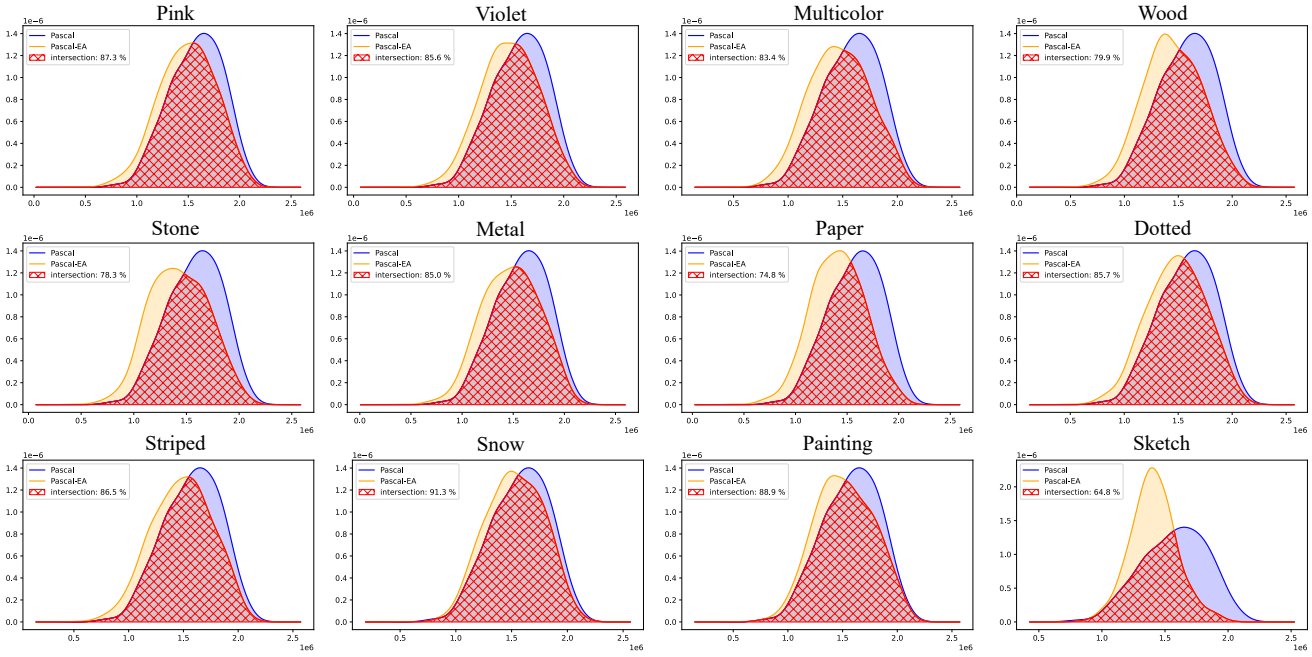

Figure 3. Different distribution of our edited images and original Pascal VOC [8] in terms of the quantities in GradNorm.

Additional qualitative results of the Pascal-EA benchmark are illustrated in Figure 7 and Figure 8. The red lines delineate the boundaries of objects in ground truths, our mask-preserved pipeline can ensure the correctness of original labels in attribute-edited images.

4. Additional applications of our pipeline

Improving the generalization ability of segmentation methods. We show that our mask-preserved attribute editing pipeline can be exploited to generate training images for improving the robustness ability of segmentation models. We construct edited training sets with four adverse conditions (fog, snow, rain, and night) using the Cityscapes dataset [5], and use them to train models. Following the previous setting of [13], we report performances from Cityscapes to ACDC domain generalization, the results of comparative approaches are directly from [13]. The quantitative results in Table 3 and 4 exhibit that model training with our data has competitive performances, and consistently gains improvement across all datasets and scenarios.

Method	HRNet [20] CS	Rain	Fog	Snow	Night	Avg.	Segformer [21] CS	Rain	Fog	Snow	Night	Avg.
Baseline	70.47	44.15	58.68	44.20	18.90	41.48	67.90	50.22	60.52	48.86	28.56	47.04
CutOut [7]	71.39	40.29	57.70	43.98	16.55	39.63	68.93	47.68	60.34	46.98	26.49	45.37
CutMix [24]	72.68	42.48	58.63	44.50	17.07	40.67	69.23	49.53	61.58	47.42	27.77	46.57
Weather [9]	69.25	**50.78**	60.82	38.34	22.82	43.19	67.41	**54.02**	64.74	49.57	28.50	49.21
StyleMix [11]	57.40	40.59	49.11	39.14	19.34	37.04	65.30	53.54	63.86	49.98	28.93	49.08
Ours	65.77	46.40	**61.61**	**49.78**	**28.49**	**46.57**	63.48	52.25	**69.54**	**56.20**	**30.12**	**52.03**
Oracle	-	65.67	75.22	72.34	50.39	65.90	-	63.67	74.10	67.97	48.79	63.56

Table 3. Comparison of different methods from Cityscapes (source) to ACDC (target) using the mIoU (↑) metric. The results are reported on the Cityscapes (CS) validation set, four individual scenarios of ACDC, and the average (Avg). The best performances are bold. Oracle indicates the supervised training on ACDC, serving as an upper bound for the other methods.

Vanilla	CutOut [7]	CutMix [24]	Digital Corruption [9]	AugMix [10]	**Our**
90.81	91.44	92.01	91.19	91.88	**92.57**

Table 4. The results of different data augmentation techniques using Mask2Former [3] on Pascal VOC dataset [8].

Table 5. The mIoU (↑) of different methods on Pascal VOC [8] and our Pascal-EA. The performance of all methods drops on our Pascal-EA.

Method	Pascal VOC	Pascal-EA
DeepLabV3+ [2]	75.33	73.65 (-1.68)
OCRNet [23]	76.92	75.32 (-1.60)
Segmenter [18]	82.25	80.66 (-1.59)
Segformer [21]	81.01	79.45 (-1.56)
Mask2Former [3]	91.98	90.81 (-1.17)
CATSeg [4]	96.60	95.59 (-1.1)
OVSeg [14]	94.49	92.83 (-1.66)
ODISE [22]	93.22	91.54 (-1.68)
X-Decoder [26]	91.77	90.42 (-1.35)
SEEM [27]	91.06	89.21 (-1.85)

In Table 3, since CutOut [7] and CutMix [24] just combine local visual contents, it exhibits improvement in in-distribution performance while deterioration on global style shifts. On account of unreal images, Hendrycks-Weather [9] degrade performance in snow, and StyleMix [11] has declined in all scenarios.

5. More discussion

5.1. Text quality evaluation

As generating variations of text descriptions of images by LLM [1], our framework inevitably imports text perturbations. To extensively evaluate the quality of target texts in comparison to the source, we adopt several metrics: (1) Perplexity to measure sentence quality, (2) CIDEr and SPICE to measure the fidelity and semantic meanings respectively, (3) BERT score [6] which calculate cosine similarity between texts and category labels of images to measure the consistency with class ground truth. The results are shown in Table.6. The results indicate a reduction in both the quality and semantic consistency of generated texts, but the decrement is in a reasonable range. Thus, we infer that while the LLM introduces additional noise to texts, it is still adequate as a text editor in our framework.

Table 6. The quantitative assessments of text variations to original ones in Pascal VOC dataset [8].

	Perplexity (↓)	CIDEr (↑)	SPICE (↑)	BERT-Score (↑)
Source	103.71	10.00	1.00	0.71
Text variations				
Color	107.70	7.34	0.85	0.71
Material	118.14	3.58	0.56	0.65
Pattern	124.26	6.10	0.71	0.62
Style	129.54	7.35	0.67	0.69

5.2. Impact of threshold τ_c

The threshold $\tau_c \in [0, 1]$ defines the duration of ControlNet [25] injection during the denoising process, the smaller value indicates a longer adoption duration. We perform additional experiments to explore its effects on edited image quality. We translate images with complex city scenes in the Cityscapes dataset [5] to that on adverse weather (snow, rain, fog, night). Figure 4 illustrates several qualitative results. We observe that, as ControlNet [25] injection duration decreases, the structural consistency decreases while the realism of edited images increases. To make a trade-off between realism and structural consistency, we further compute their FID score as shown in Figure 5. We can notice that as $\tau_c = 0.5$ we achieve the best performances on all weather editing scenarios. Therefore, we adopt $\tau_c = 0.5$ in our pipeline.

5.3. Failure cases of edited images

Inheriting the innate limitations of diffusion models, our pipeline has unpleasant performances in several scenarios.

Figure 4. Resulting images edited by our method using different values of τ_c

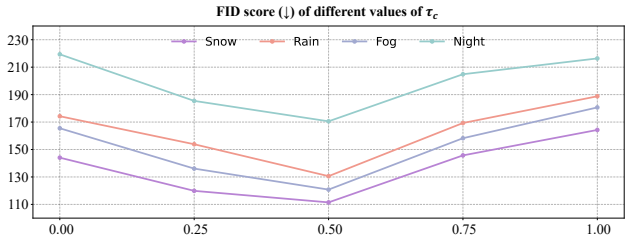

Figure 5. FID score (↓) of our generation results using different values of τ_c. When $\tau_c = 0.5$, our method achieves the best performances in all weather conditions.

The qualitative failure cases of edited images are shown in Figure 6. As multiple objects overlap, our edited images violate the original structures of inside objects. Moreover, since the inherent failure modes in diffusion model [16], our edited images could deviate from the original face of creatures.

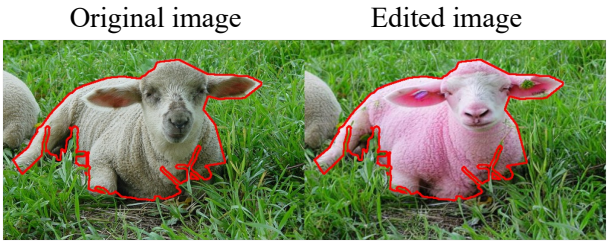

Figure 6. Illustration of our failure cases. Our method does not faithfully preserve the appearance of the face due to the limitations of the diffusion model; however, our method effectively preserves the global structure, which ensures the ground-truth mask of the edited image is consistent with that of the original one.

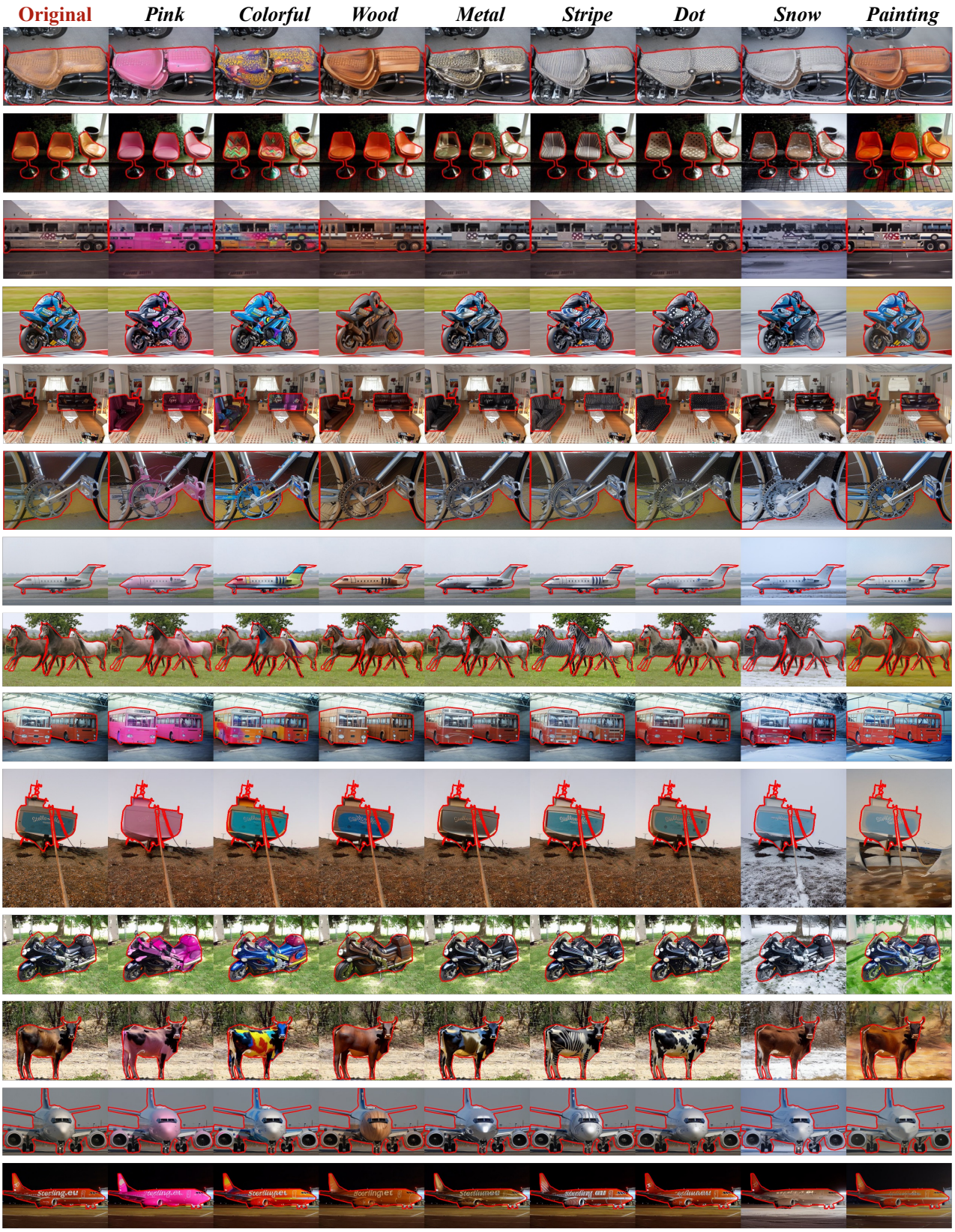

Figure 7. Resulting images edited by our method. Our method effectively converts various objects into versions with different attributes, while ensuring their segmentation mask is consistent with the original ones.

Figure 8. Resulting images edited by our method. Our method effectively converts various objects into versions with different attributes, while ensuring their segmentation mask is consistent with the original ones.

Figure 9. Qualitative segmentation results of OCRNet [23] under object color attribute variations.

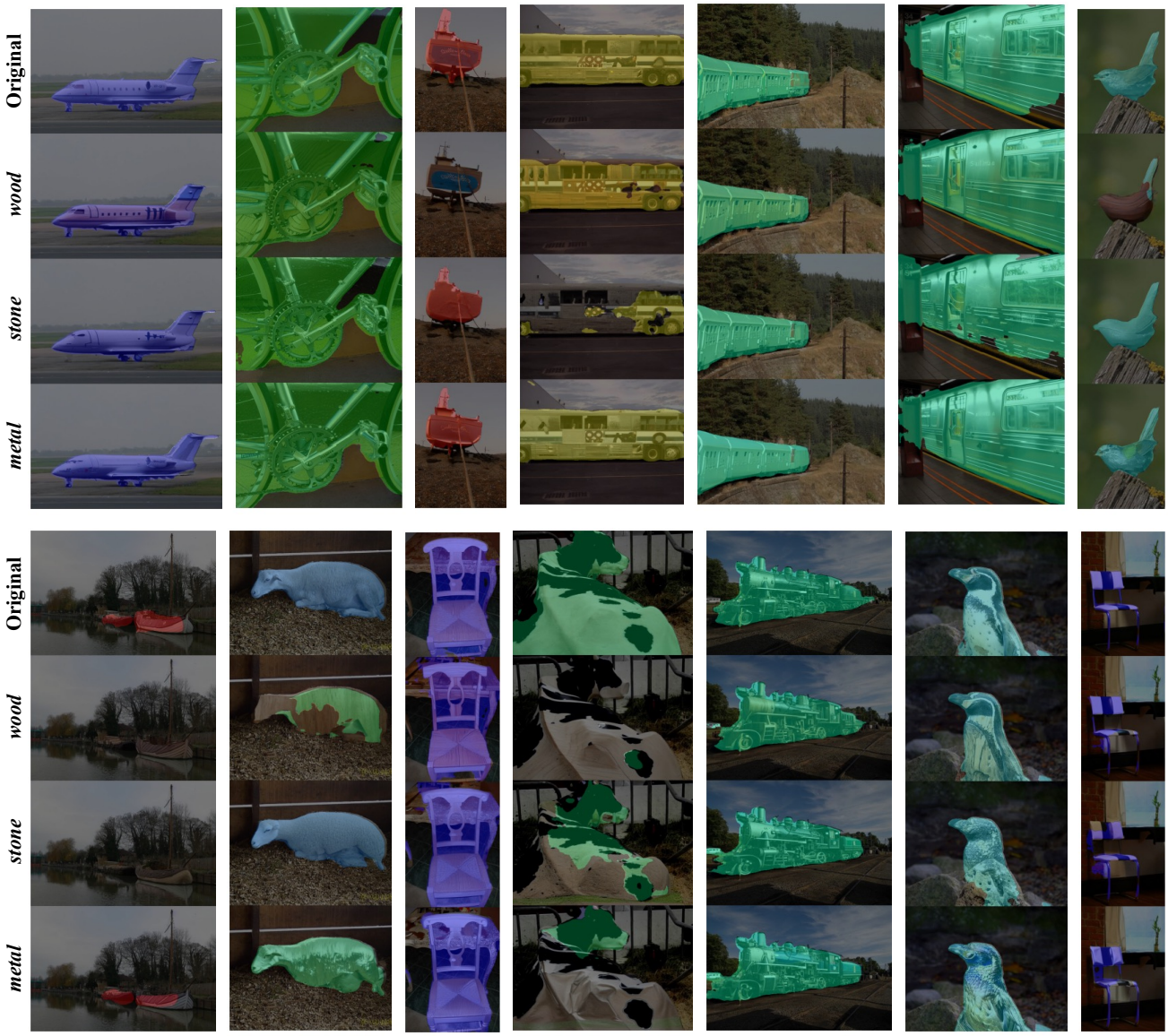

Figure 10. Qualitative results of OCRNet [23] under object material attribute variations.

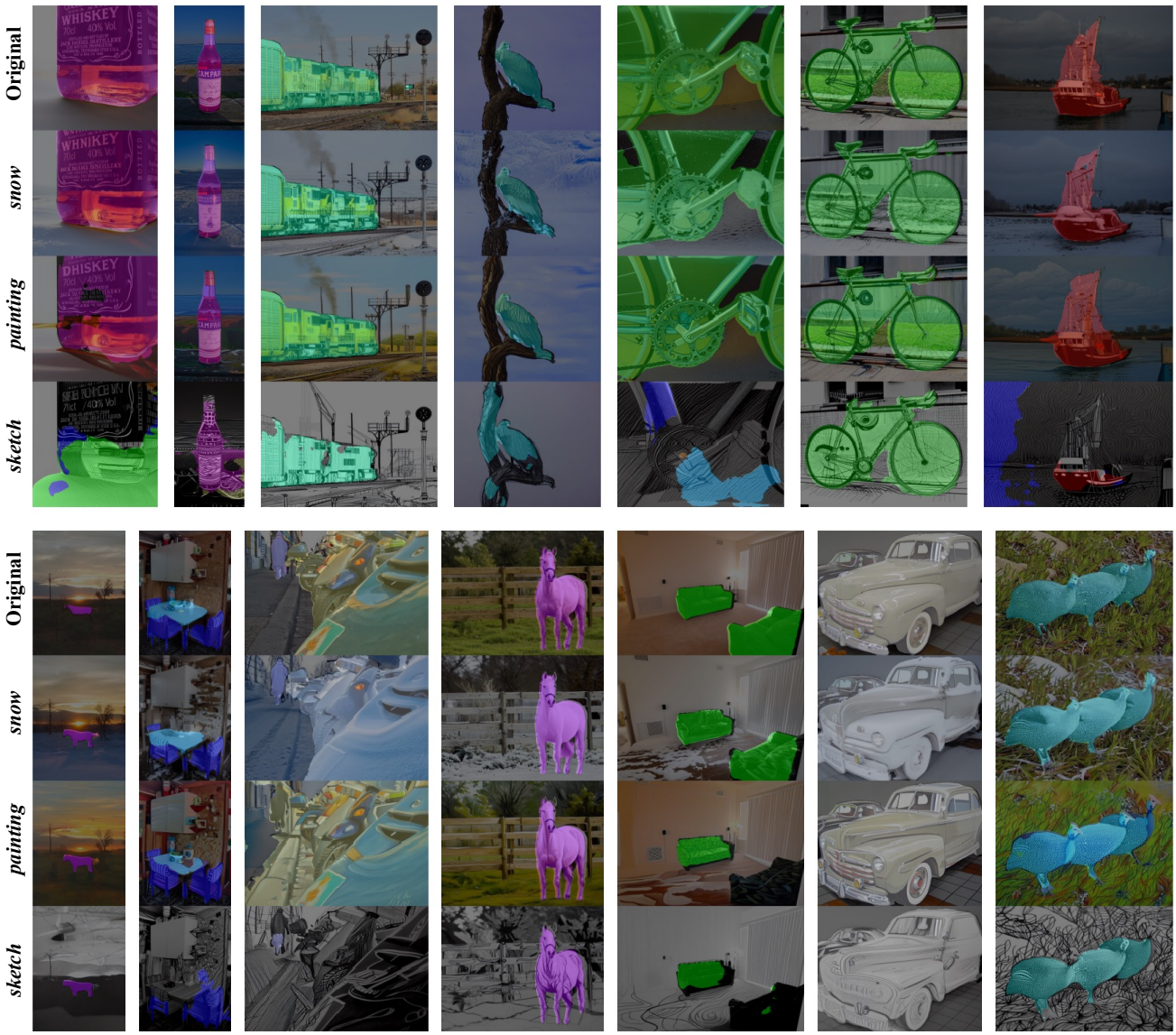

Figure 11. Qualitative segmentation results of OCRNet [23] under image style attribute variations.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2, 4

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 4

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 4

[4] Seokju Cho, Heeseong Shin, Sunghwan Hong, Seungjun An, Seungjun Lee, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2303.11797*, 2023. 1, 4

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3, 4

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[7] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 4

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1, 2, 3, 4

[9] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1, 4

[10] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 4

[11] Minui Hong, Jinwoo Choi, and Gunhee Kim. Stylemix: Separating content and style for enhanced data augmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14862–14870, 2021. 4

[12] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, 2021. 1

[13] Yumeng Li, Dan Zhang, Margret Keuper, and Anna Khoreva. Intra-& extra-source exemplar-based style synthesis for improved domain generalization. *arXiv preprint arXiv:2307.00648*, 2023. 3

[14] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 4

[15] Xiaofeng Mao, Yuefeng Chen, Yao Zhu, Da Chen, Hang Su, Rong Zhang, and Hui Xue. Coco-o: A benchmark for object detectors under natural distribution shifts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6339–6350, 2023. 1

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 5

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1

[18] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 4

[19] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 1, 2

[20] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 4

[21] 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*, 34:12077–12090, 2021. 4

[22] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 4

[23] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020. 1, 3, 4, 8, 9, 10

[24] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 4

[25] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 4

[26] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 4

[27] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 1, 3, 4