Reducing Noise Pixels and Metric Bias in Semantic Inpainting on Segmentation Map-Appendix

Jianfeng He[†], Bei Xiao[‡], Xuchao Zhang[†], Shuo Lei[†], Shuhui Wang[+*], Chang-Tien Lu[†]
[†]Sanghani Center for Artificial Intelligence and Data Analytics, Virginia Tech, Falls Church, VA, USA
[‡]Department of Computer Science, American University, Washington, DC, USA
[+] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
[†]{jianfenghe, xuczhang, slei, ctlu}@vt.edu, [‡]bxiao@american.edu, [+]wangshuhui@ict.ac.cn

1. Organization of Appendix

In this appendix, we show the supplementary material of the paper "Reducing Noise Pixels and Metric Bias in Semantic Inpainting on Segmentation Map". Firstly, we present more implementation details. Then, the statistics for Fig. 6 in main scrip is listed. Finally, additional qualitative experiment results are illustrated.

1.1. Additional Implementation Details

Besides the experiment setup introduced in Sec. 4.1 of main scripts, we introduce more implementation details as below.

Noises Number. Since current metrics on SISM do not consider noise pixels in the testing process, we design a simple metric Noises Number (NN) to count average number of noise pixels. Specifically, for each testing sample, we extract a set Q^c of pixel values of the inpainted area from the respective ground truth. Similarly, we have a set $\hat{Q}$ of pixel values of the inpainted area from the respective SISM result. Then, we have a difference set $\hat{Q} - Q^c$, which contains values of unique pixels (noises) in $\hat{Q}$. The Noises Number is an average of total number of pixels with values in the difference set for each testing sample.

Inpainting categories. Based on the rank of object numbers, we choose 8 movable objects categories for SISM on Cityscape, 'person', 'rider', 'car', 'truck', 'bus', 'train', 'motorcycle', and 'bicycle'. For ADE20K, we choose 7 categories ("bed(165)", "table(2684)", "lamp(1395)", "picture(1735)", "window(3055)", "pillow(1869)", and "curtain(687)") for SISM. The numbers in the brackets are the class IDs in ADE20K, we provide them for better reproducing our results.

Image size. The training and generated image resolutions are 256×128 and 256×256 respectively for Cityscapes and ADE20K.

Selection of model weights. For SISM, we train each dataset for 200 epochs and select the model weights from the last epoch as our final model for the testing process. This is the same as [2, 4, 1].

Downstream Model. The downstream model for Cityscapes is trained for 300 epochs, based on the code provided by [2]. And the downstream model for ADE20K is pretrained by [2], which is applied in our experiment directly.

Random seed. For all testing results, we fix the random seed as "679" to get the quantitative and qualitative results of SISM and testing results about Sem. For the training, we do not set a fixed random seed.

Data split. For the data split, our training data and testing data of the Cityscapes are same to [4]. And the data split of ADE20K is same to [2]. For PS-COCO, we apply [3] provided training set as training data and provided validation set as testing set.

Data preprocess. For SISM experiments, we exclude the object bounding boxes that are too small to carry significant scene content. For the two datasets, we set the size threshold as 0.02, which filters bounding boxes that are smaller than 2% of the size of inputted images. For the images that include no object bounding box satisfying the size threshold, we skip the images in both the training and testing processes.

For Sem related experiments, an example of extraction, described in Sec. 4.1, is shown in Fig. 1. The extracted binary instance segmentations are then resized to 224×224 as our binary target object map $\mathbf{O^c}$, which is the first-channel of input to our shape classifier, described in Sec. 3.4 of main script. We set 224×224, as it is the image sizes applied in EfficientNet. We extract their respective mask areas by calculating their bounding boxes, which is the second-channel of input to our shape classifier.

Link to downloadable version of the datasets. The Cityscapes is from https://www.cityscapes-

* Corresponding author.

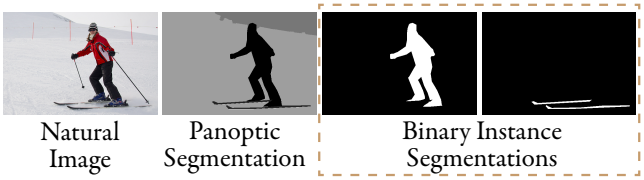

Figure 1. Diagram of extraction for binary instance segmentations, which are further resized into 224×224 as the first-channel of the input to our shape classifier.

dataset.com/. And the ADE20K can be downloaded from https://groups.csail.mit.edu/vision/datasets/ADE20K/. The PS-COCO can be found from https://cocodataset.org/download.

Data Samples and Source Code. We will provide the data samples and source code in the Github soon.

Computing infrastructure. We do SISM experiments and Sem testing on two GPUs, which both are GTX 1080Ti. The RAM in our machine is 64 GB. The Sem training is finished on 4 V100 for around 65 hours.

1.2. Statistics Of Fig. 6 In Main Script

The statistics of Fig.6 in main script are listed in Tab. 1. The Change of $\{\mathrm{A}\}$ in the last three columns are the results by respective statistics of $\{\mathrm{A}\}$ of the previous one level subtracting those of current level. From the table, we can see that the change of 1-Sem is around 50% less compared with that of hamm, when we flip the target objects. Since flip keeps the original semantics, it verifies that Sem can better quantify semantic divergence between the generated and ground-truth target objects.

1.3. Additional Qualitative Results

Besides the qualitative results provided in Sec. 4.3 in the main scripts. We also provide more qualitative results in the Fig. 2, Fig. 3, Fig. 4 and Fig. 5. From them, the conclusion similar to the main scripts are concluded as below,

(1) The TwoSM+DA(Train) can effectively remove the noise pixels without disturbing the semantics. Nearly all TwoSM+DA(Train) results, in Fig. 2, Fig. 3, Fig. 4 and Fig. 5, achieve denoise without destroying original semantics generated by TwoSM. This shows the stability of our DA.

(2) Implementing DA in the training process performs better than that in the testing process. The results in Fig. 2 and Fig. 3 are all examples to verify better performance by TwoSM+DA(Train). Specifically, the first row of Fig. 2 shows a case, where the TwoSM+DA(Both) denoises better compared with that of TwoSM+DA(Train). This shows the effect of applying DA in the training and testing processes.

(3) Some natural images from the downstream task cannot reflect the improvement of SISM results. From the natural images in Fig. 4, we cannot clearly see the effects of noise pixels, though the noise is removed by our DA. This further indicates that the current downstream models do not perform well on the image inpainting, even when the noise pixels are removed. Thus, using downstream results to evaluate the SISM results brings new bias by the limitation of model performance.

(4) Some noise pixels lead to obvious disturbances to the natural images from downstream models. All results of the Fig. 5 show prominent disturbances in the natural images from the noise of SISM results. Specifically, the last row of Fig. 5 shows a case where TwoSM+DA(Train) repairs the shape of table, even though there is no noise in TwoSM.

References

[1] Jianfeng He, Xuchao Zhang, Shuo Lei, Shuhui Wang, Qingming Huang, Chang-Tien Lu, and Bei Xiao. Semantic editing on segmentation map via multi-expansion loss. *arXiv preprint arXiv:2010.08128*, 2020.

[2] Seunghoon Hong, Xinchen Yan, Thomas S Huang, and Honglak Lee. Learning hierarchical semantic image manipulation through structured representations. In *Advances in Neural Information Processing Systems*, pages 2708–2718, 2018.

[3] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.

[4] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

	Level	Sem	1-Sem	hamm	tIOU	C-1-Sem	C-hamm	C-tIOU
Gaussian noisewith various frequencies	0	0	1	1	1	-	-	-
	1	0.7418	0.2582	0.8751	0.8751	0.7418	0.1249	0.1249
	2	0.7298	0.2702	0.7501	0.7501	-0.012	0.125	0.125
	3	0.7302	0.2698	0.5	0.5	0.0004	0.2501	0.2501
Gaussian noisewith various means	0	0	1	1	1	-	-	-
	1	0.8034	0.1966	0.8751	0.8751	0.8034	0.1249	0.1249
	2	0.7636	0.2364	0.8751	0.8751	-0.0398	0	0
	3	0.7394	0.2606	0.8751	0.8751	-0.0242	0	0
Erosionwith various kernel sizes	0	0	1	1	1	-	-	-
	1	0.1615	0.8385	0.9358	0.8782	0.1615	0.0642	0.1218
	2	0.2732	0.7268	0.8778	0.7687	0.1117	0.058	0.1095
	3	0.3672	0.6328	0.8266	0.6724	0.0939	0.0512	0.0963
Dilationwith various kernel sizes	0	0	1	1	1	-	-	-
	1	0.1852	0.8148	0.9427	0.903	0.1852	0.0573	0.097
	2	0.3395	0.6605	0.8956	0.8374	0.1543	0.0471	0.0656
	3	0.4697	0.5303	0.8557	0.7892	0.1302	0.0399	0.0482
Flipwith various frequencies	0	0	1	1	1	-	-	-
	1	0.0417	0.9584	0.9151	0.8602	0.0416	0.0849	0.1398
	2	0.0831	0.9169	0.8311	0.7348	0.0414	0.084	0.1255
	3	0.164	1	0.6703	0.5302	0.0809	0.1608	0.2045
Rotationwith various degrees	0	0	1	1	1	-	-	-
	1	0.5885	0.4115	0.5912	0.3228	0.5885	0.4088	0.6772
	2	0.6092	0.3908	0.6224	0.4775	0.0207	-0.0312	-0.1547
	3	0.6765	0.3236	0.5565	0.2849	0.0673	0.0659	0.1925

Table 1. The statistics of Fig. 6 in main script. The the C-{A} in the last three columns are the abbreviations of the changes of {A}, which are the results by respective statistics of {A} of the previous one level subtracting those of current level.

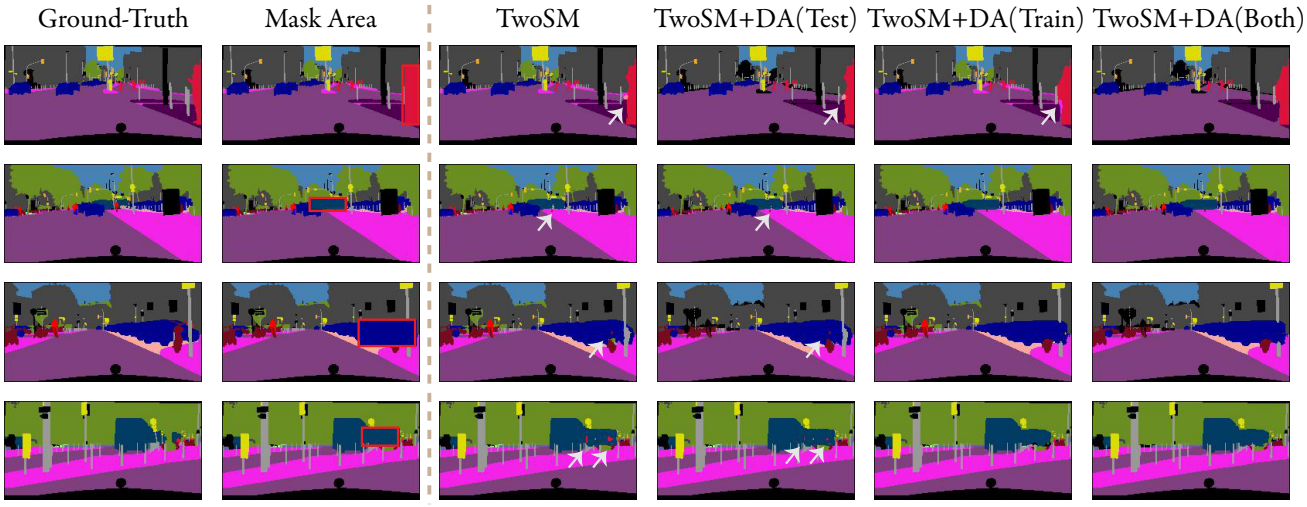

Figure 2. Examples of SISM results of TwoSM, TwoSM+DA(Test), TwoSM+DA(Train), and TwoSM+DA(Both) on the Cityscapes. The ground-truths segmentation maps, incomplete segmentation maps are shown on the left, where the red rectangles are mask areas. The right four columns show the inpainted segmentation maps and inpainted natural images of the four methods respectively. The white arrows point to the visible noise pixels. Please zoom in for better vision.

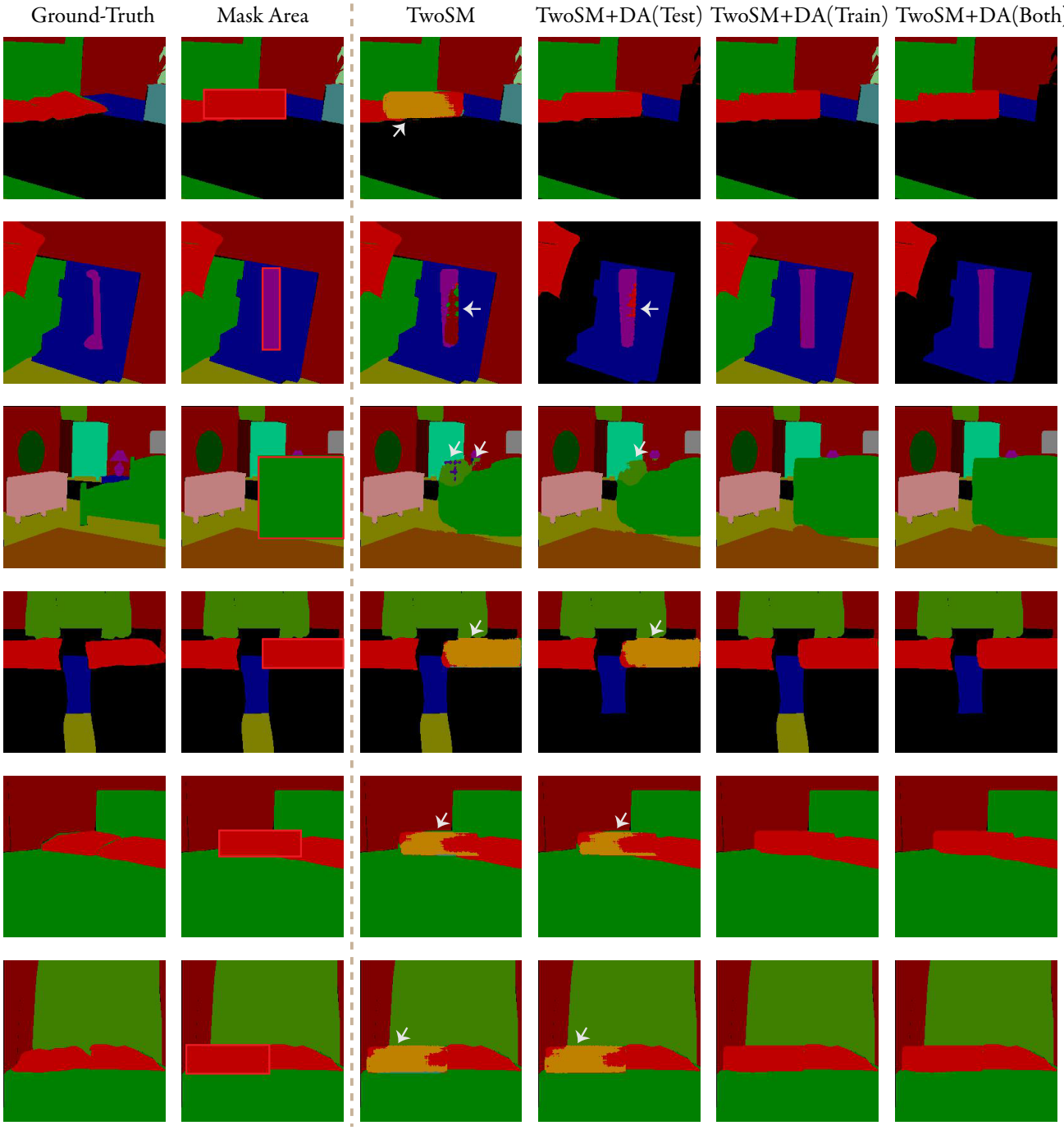

Figure 3. Examples of SISM results of TwoSM, TwoSM+DA(Test), TwoSM+DA(Train), and TwoSM+DA(Both) on the Cityscapes. The images are organized in a similar way as Fig. 2. Please zoom in for better vision.

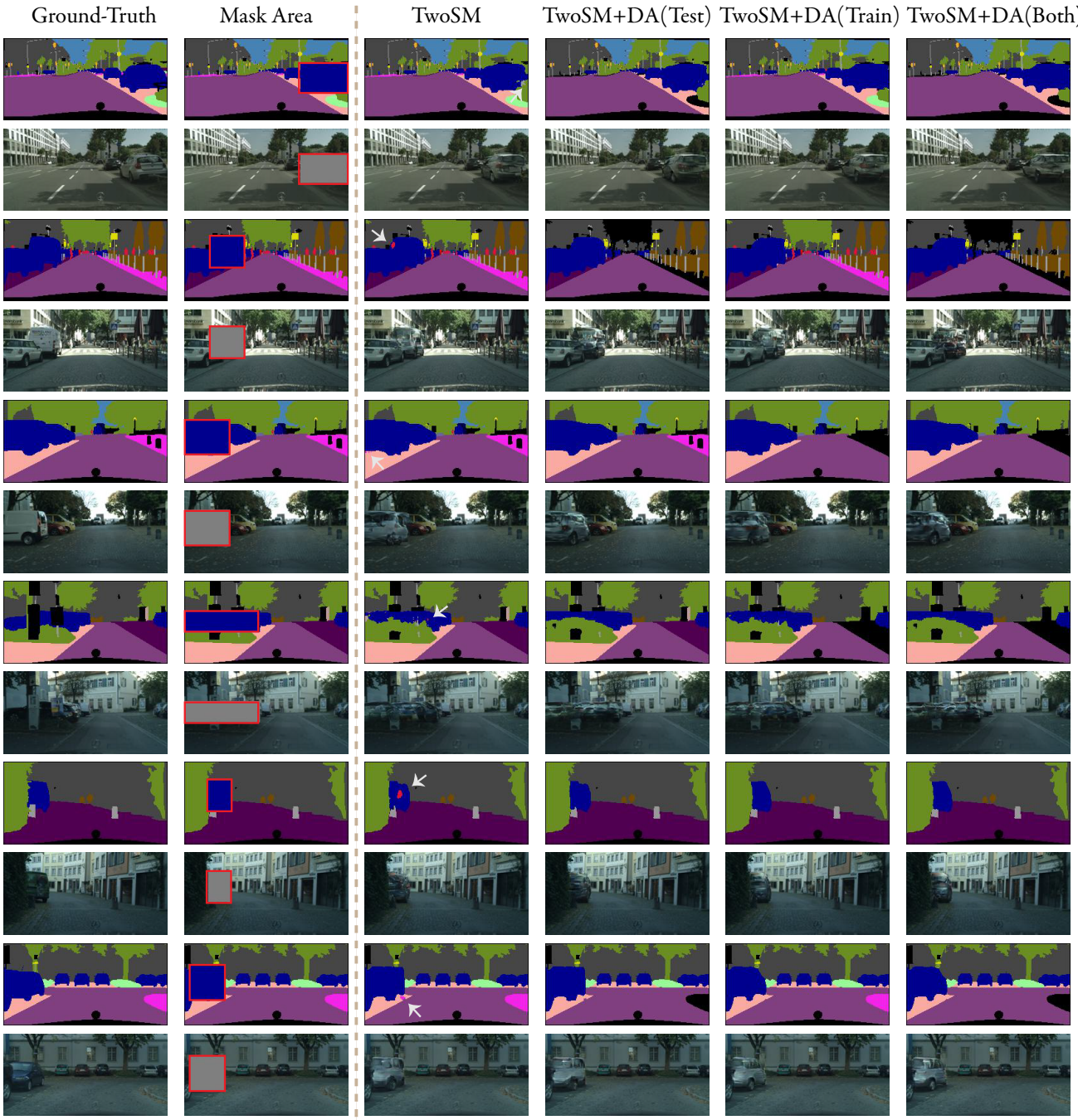

Figure 4. Examples of SISM and downstream model results of TwoSM, TwoSM+DA(Test), TwoSM+DA(Train), and TwoSM+DA(Both) on the Cityscapes. The ground-truths segmentation maps, incomplete segmentation maps and ones for natural images are shown on the left, where the red rectangles are mask areas. The right four columns show the inpainted segmentation maps and inpainted natural images of the four methods respectively. The white arrows point to the visible noise pixels. From the figure, we can see that the results of downstream model cannot fairly evaluate the SISM results, because the current downstream model is not well performed to figure out the difference of SISM results. Please zoom in for better vision.

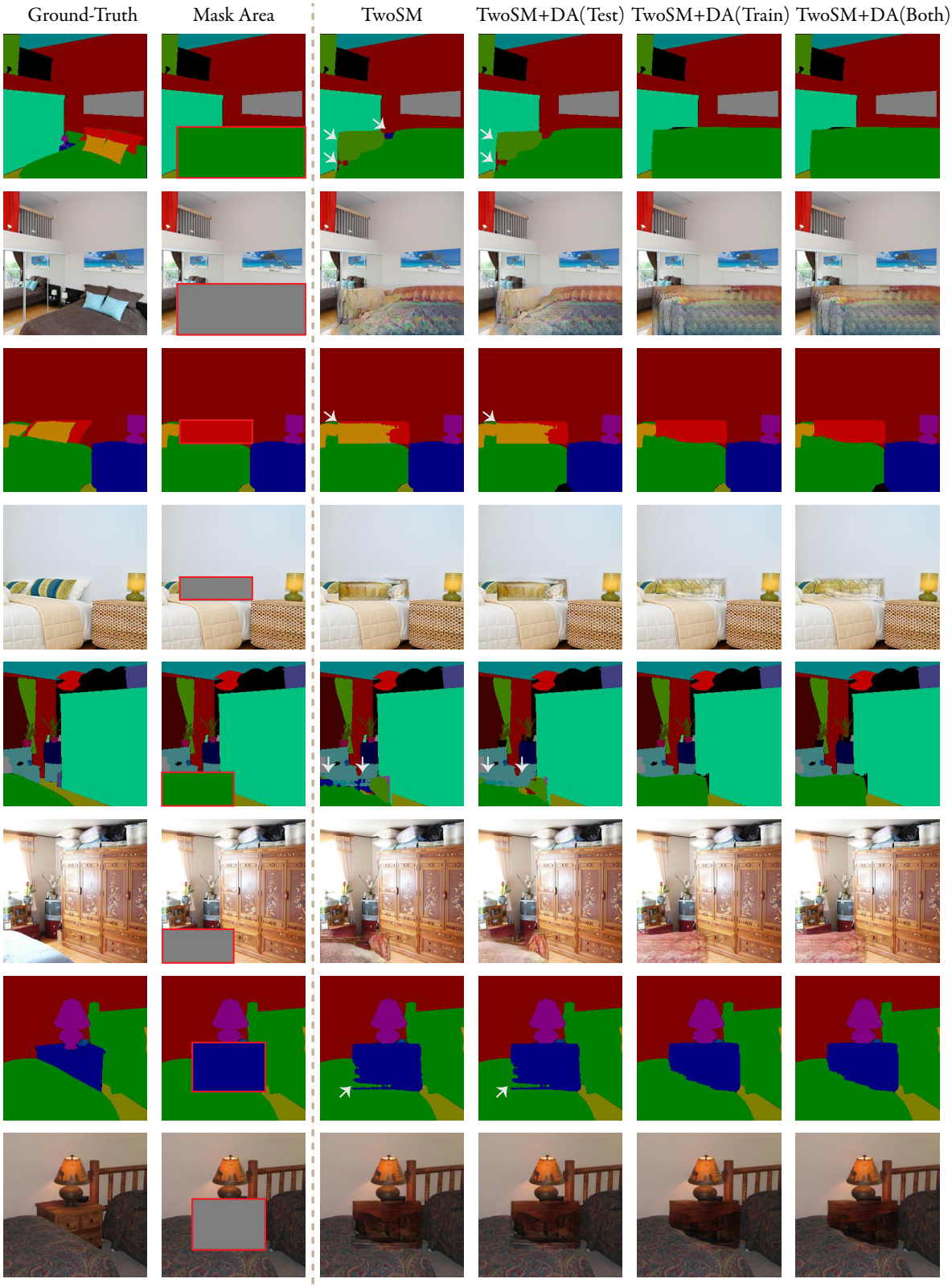

Figure 5. Examples of SISM and downstream model results of TwoSM, TwoSM+DA(Test), TwoSM+DA(Train), and TwoSM+DA(Both) on the ADE20K. The images are organized in a similar way as Fig. 4. Please zoom in for better vision.