Supplementary Material for "Guidance and Evaluation: Semantic-Aware Image Inpainting for Mixed Scenes"

Liang Liao[1,2][0000−0002−2238−2420], Jing Xiao[1,2][0000−0002−0833−5679], Zheng Wang[2][0000−0003−3846−9157], Chia-Wen Lin[3][0000−0002−9097−2318], and Shin'ichi Satoh[2]

[1] National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
[2] National Institute of Informatics
[3] Department of Electrical Engineering, National Tsing Hua University
{liang, wangz, satoh}@nii.ac.jp; jing@whu.edu.cn; cwlin@ee.nthu.edu.tw

1 Implementation Details

In addition to Section 3, we introduce the detailed architecture of the proposed SGE-Net. The main network of SGE-Net is composed of an encoder, a decoder, and a Context Inference Module (CIM) between them. The encoder takes 3-channels image and 1-channel mask as input, and gradually down-samples the contextual feature. We build the encoder based on ResNet-50 with five blocks (Conv1, Conv2_x, Conv3_x, Conv4_x, and Conv5_x), which is pre-trained on the ImageNet classification task. The CIM is used to initially infer the contextual feature extracted by the encoder to the feature of a complete image. To better infer and update the contextual feature, we adopt dilated convolution layers to expand the receptive field. The decoder gradually updates and refines the inferred contextual feature from low-resolution to high-resolution. At every resolution scale, an inpainting branch and a segmentation branch are used to generate the inpainted image and the segmentation map from the contextual feature. To progressively update the contextual feature from a corrupted image to a complete image using semantic information, Segmentation Confidence Evaluation Module (SCEM) and Semantic-Guided Inference Module+ (SGIM+) are developed between the contextual features of different scales in the decoder.

We implement this network using the Pytorch toolbox and optimize the SGE-Net and the discriminator using the Adam algorithm with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a learning rate of 0.0001. In all experiments, we use a batch size of 4 and set the training iterations to 1,000,000. The loss weight λ_p, λ_a and λ_s are set to 1, 0.01, and 5 respectively. The thresholds of proportion to decide the τ^l are set to 0.25, 0.5, 0.75 for scale 2 to scale 4. Taking the 0.25 in scale 2, for example, it means the thresholds τ^2 equals to the value, which is greater than 25% of the values in the max-possibility map.

For data augmentation, random mirroring and random crop are applied during training. All the results in the paper and supplementary are output directly from the trained models without any post-processing.

2 Additional Results

2.1 Comparison with State-of-the-art

In this section, we show more inpainting results by the proposed SGE-Net are also shown in Fig. 1.

2.2 Performance Comparison of Scene Complexity

To validate the performance improvement of SGE-Net on the scenes involving multiple semantic components over the baselines with structural information, we present visual comparisons (see Fig. 2 and Fig. 3) with the baseline methods on the mixed scenes with various numbers of semantics. The test images and the division method are the same as that in user study (Please refer it in Section 4.2).

The visual results shown in the figures represent that with the increment of the semantics in the scene, the performance of the SGE-Net can be better than the baselines to a higher extent, The visual results also agree with the conclusion of the following user study on different levels of scene complexity.

2.3 Automatic Segmentation vs. Human-labeled Semantics

In Section 4.3, we study the impact on the inpainting performance of SGE-Net by replacing the human-labeled segmentation maps for training SGE-Net with the maps generated by CNN-based state-of-the-art segmentation models (i.e., DPN [2] for **Outdoor Scenes** and Deeplab v3+ [1] for **CityScape**). The objective quality comparison is shown in Table 4 in section 4.3.

Fig. 4 and Fig. 5 show the visual quality comparisons of the inpainting results by SGE-Net trained, respectively, on the human-labeled segmentation maps and on the segmentation model-generated maps. As can be observed, SGE-Net trained on imperfect semantic annotation achieves comparable inpainting performance with SGE-Net trained on human-labeled semantics.

2.4 More Results on Places2 Dataset

More comparisons with those baselines on **Places2** dataset are conducted. We use our model trained on **Outdoor Scenes** to complete the images with similar scenes in **Places2**. The GC and EC models are the released models pretrained on **Places2**. Please refer to Fig. 7 for more visual results.

2.5 More Results on Paris StreetView Dataset

The **Paris StreetView** dataset is also a commonly evaluated one for image inpainting methods. Moreover, the categories of the dataset is similar with that of **Outdoor Scene** dataset. Therefore, we apply our model trained on **Outdoor Scene** to complete the **Paris StreetView**. It can also achieve reasonable results. Please refer to Fig. 6 for more visual results.

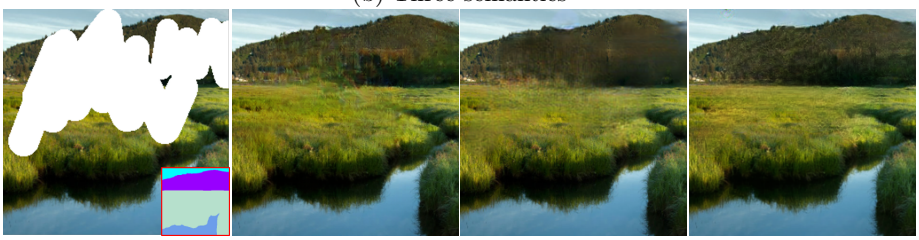
References

1. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
2. Liu, Z., Li, X., Luo, P., Loy, C.C., Tang, X.: Deep learning markov random field for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(8), 1814–1828 (2017)
3. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212 (2019)
4. Song, Y., Yang, C., Shen, Y., Wang, P., Huang, Q., Kuo, C.C.J.: Spg-net: Segmentation prediction and guidance network for image inpainting. arXiv preprint arXiv:1805.03356 (2018)

Fig. 1: More inpainting results from **Outdoor Scenes** (upper two rows) and **Cityscapes** (lower two rows).

(a) Two semantics

(b) Three semantics



(c) Four semantics

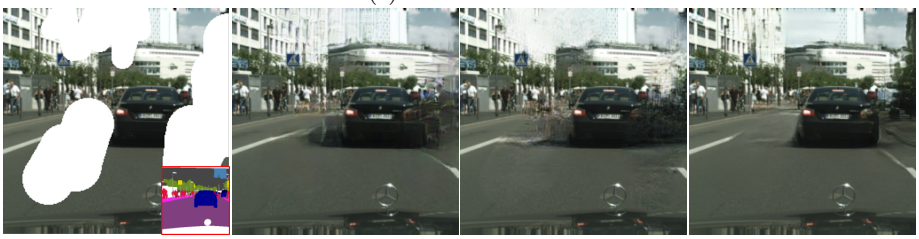
(d) Five semantics

Fig. 2: Subjective quality comparison of test results on image samples of 2 to 5 semantics from **Outdoor Scenes**. From left to right: Corrupted image, EC [3], SPG [4] and SGE-Net (ours). We only count the number of dominant semantics in a image.

(a) Four semantics

(b) Five semantics

(c) Six semantics



(d) Seven semantics

Fig. 3: Subjective quality comparison of test results on image samples of various number of semantics from **Cityscapes**. From left to right: Corrupted image, EC [3], SPG [4] and SGE-Net (ours). We only count the number of dominant semantics in a image.

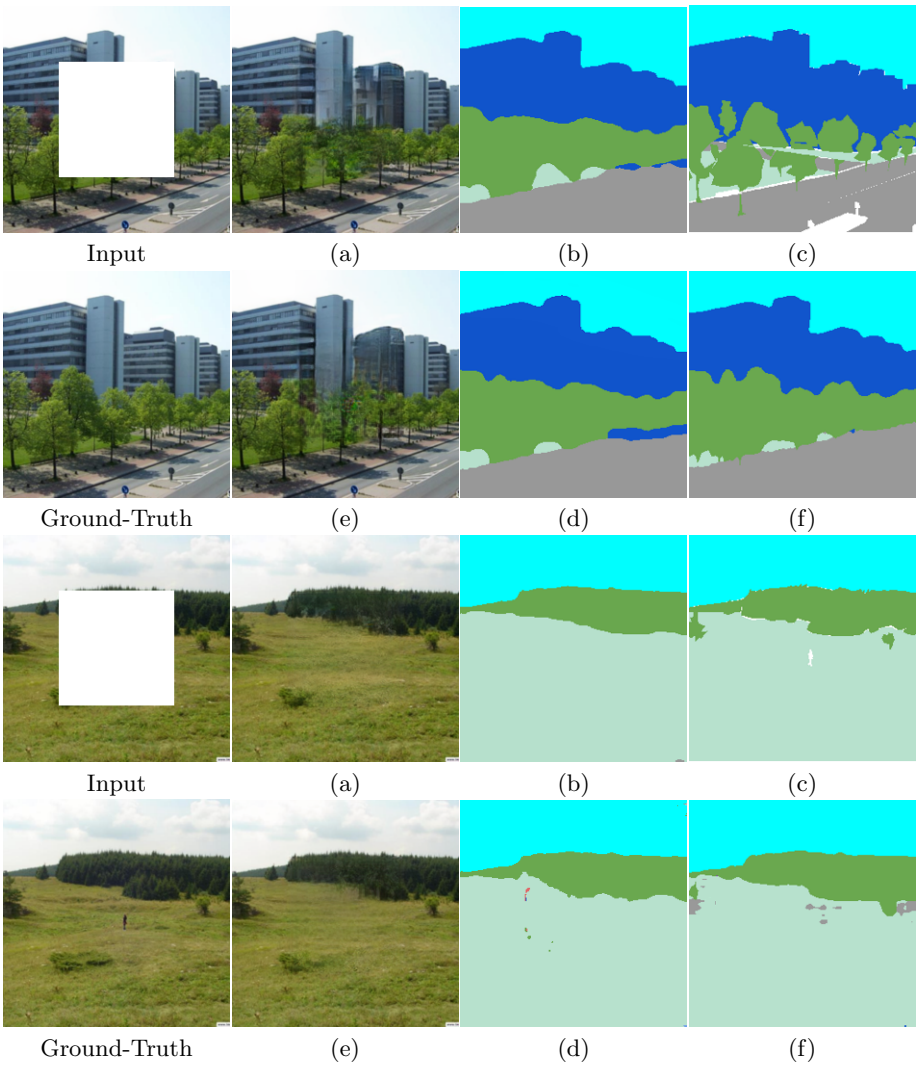


Fig. 4: Impact of segmentation accuracy on **Outdoor Scene**. (a) the inpainting results by the model trained on human-labeled segmentation maps; (b) the predicted segmentation maps by the model trained on human-labeled segmentation maps; (c) the human-labeled segmentation maps; (d) the inpainting results by the model trained on the DPN-extracted segmentation maps; (e) the predicted segmentation maps by the model trained on the DPN-extracted segmentation maps; (f) the segmentation maps extracted by DPN model.

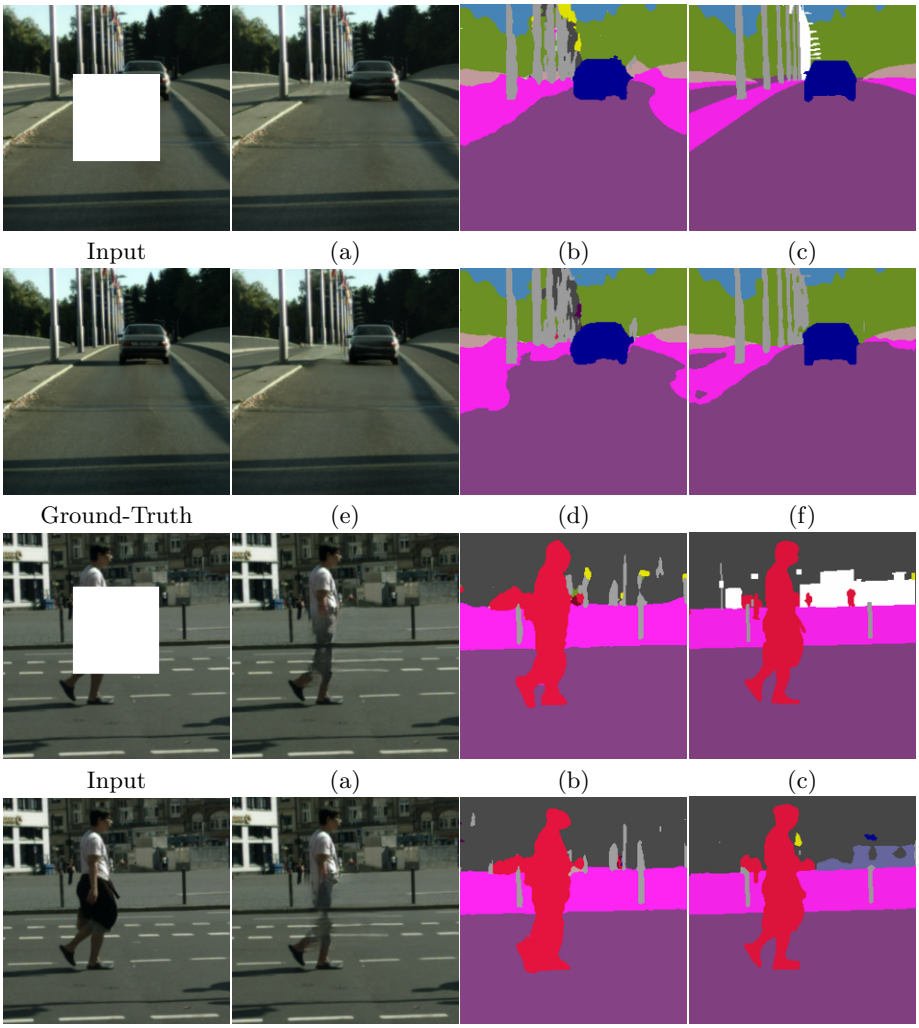


Fig. 5: Impact of segmentation accuracy on **Cityscapes**. (a) the inpainting results by the model trained on human-labeled segmentation maps; (b) the predicted segmentation maps by the model trained on human-labeled segmentation maps; (c) the human-labeled segmentation maps; (d) the inpainting results by the model trained on the Deeplab-extracted segmentation maps; (e) the predicted segmentation maps by the model trained on the Deeplab-extracted segmentation maps; (f) the segmentation maps extracted by Deeplab model.

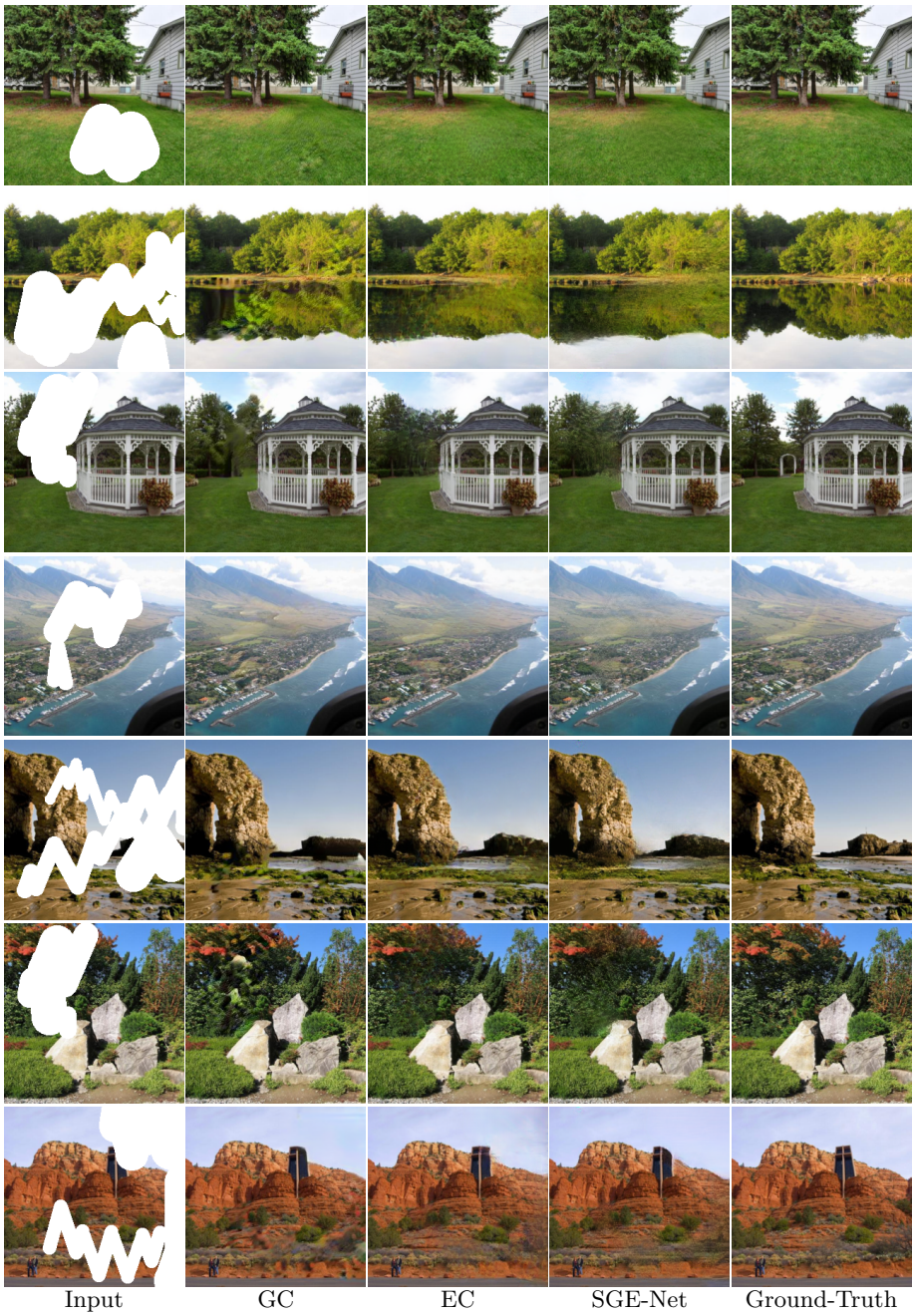


Fig. 6: Subjective quality comparison on image samples from **Places2**. SGE-Net is trained on **Outdoor Scenes** dataset

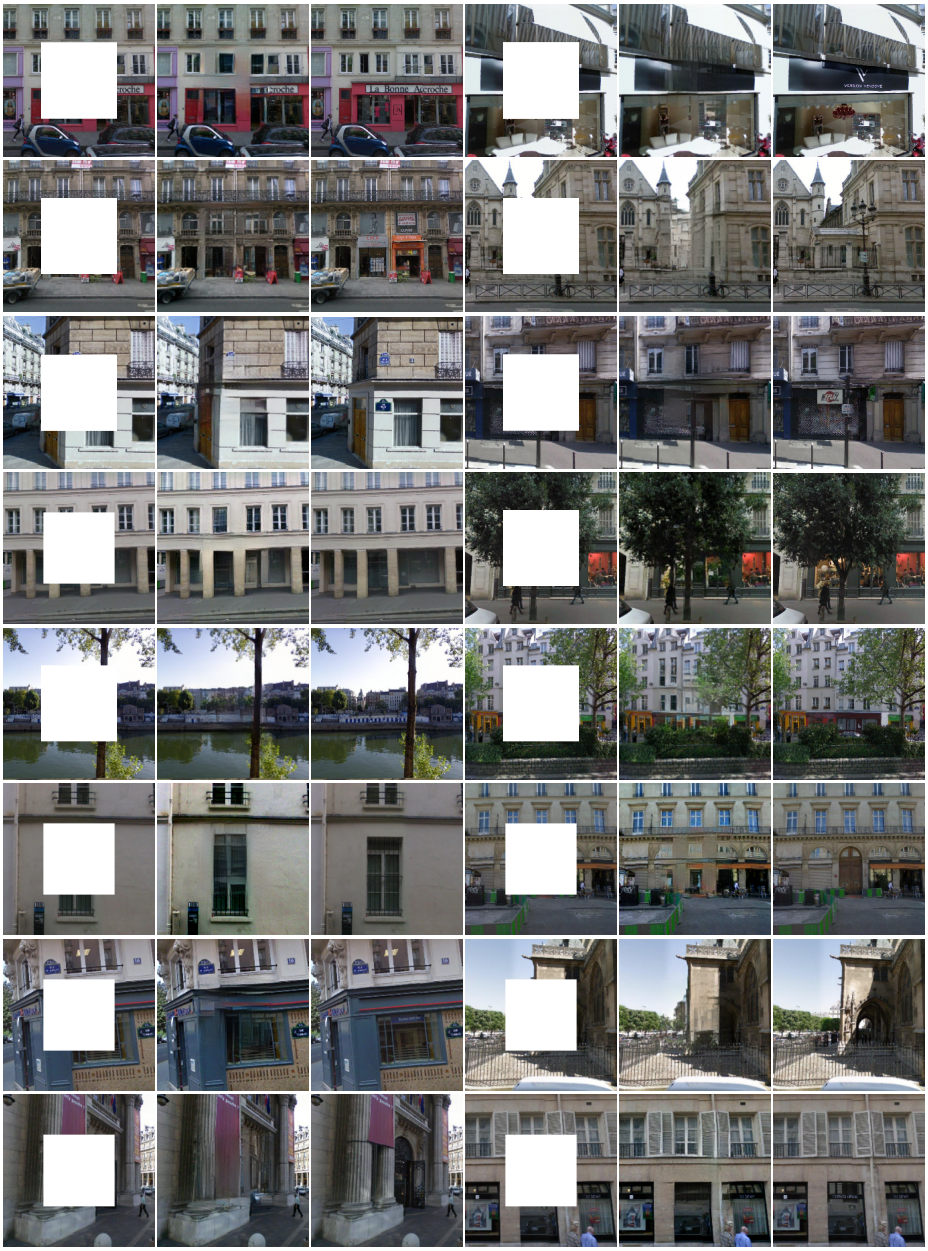


Fig. 7: Subjective quality comparison on image samples from **Paris StreetView**. SGE-Net is trained on **Outdoor Scenes** dataset. From left to right: Input, completed results of SGE-Net, Ground-Truth