Hallucinating Visual Instances in Total Absentia – *Supplementary Material* –

Jiayan Qiu[1], Yiding Yang[2], Xinchao Wang[2], and Dacheng Tao[1]

[1] UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia
{jqiu3225@uni.sydney.edu.au,dacheng.tao@sydney.edu.au}
[2] Stevens Institute of Technology, Hoboken, NJ 07030, USA
{yyang99,xinchao.wang}@stevens.edu

In this document, we provide the details and results that we can not fit into the main manuscript due to the page limit. Specifically, we show the effect of the mask size and that of the Autoencoder on the GCN module, comparison with SOTA iimage inpainting method, more visual results on COCO, NYU v2, Visual Genome and multi-instance hallucination, *background* hallucination, as well as results of a generalization test.

Please note that, the aim of our work is, as discussed in the main manuscript, to show the feasibility of hallucinating absent objects from a masked image, rather than beating the state-of-the-art Graph Convolutional Neural Networks, Generative Adversarial Networks, and image inpainting methods. We have thus endeavored to develop lightweight networks for our framework. More sophisticated networks, as long as end-to-end trainable, can be readily applied to substitute the corresponding modules in the framework.

As state-of-the-art GAN models are not able to well account for semantic constraints, like the orientation of a moving car and the legitimate pose of a person, into the image generation process, we test our model that also relies on GAN, on 20 out of 81 classes in COCO for all the datasets we use. Hallucinating visual instances with explicit hard-constraint reasoning is left for our future work.

The figures in this document follow the same layout as done for the main manuscript: the first column corresponds to the masked image, the second shows our restored image, and the third depicts the ground truth.

1 Effect of Mask Size

In this experiment, we test the effect of the mask size on the GCN module. For the experiments in main paper, the mask area is determined by their ground-truth bounding box. Here, we increase the mask size gradually and show its effect on the GCN module.

As can be seen from Tab. 1, the GCN module preserves a good accuracy ($> 80\%$) when the enlarging factor is smaller than 1.5. When the factor is set larger than 1.5, however, the accuracy decreases visibly, indicating that the size of the hallucinated object plays an important role on the prediction results.

Term	1.0X	1.1X	1.2X	1.3X	1.4X	1.5X	1.6X	1.7X	1.8X	2.0X
Accuracy(%)	87.93	87.34	86.25	84.54	82.02	79.33	76.60	73.80	70.98	66.01
Balanced Acc(%)	85.62	85.64	83.59	81.57	79.20	76.33	73.56	70.72	67.97	63.06

Table 1: Effect of the mask size on GCN module. **1.nX (or 2.0X)** denotes the enlarging factor of the mask with respect to the ground-truth size.

2 Effect of Autoencoder

In our approach, we train the GCN module with the global feature from the Autoencoder. Here, we first test the effect of Autoencoder's reconstruction quality on our GCN module.

Term	20.0dB	20.5dB	21dB	21.5dB	22.0dB	22.5dB	23.0dB	23.5dB	24.0dB	25dB
Accuracy(%)	76.33	77.90	79.62	81.62	84.95	86.41	87.93	87.93	87.95	87.96
Balanced Acc(%)	73.12	73.92	76.45	78.94	81.80	83.72	85.62	85.63	85.64	85.64

Table 2: Effect of the Autoencoder's reconstruction quality on the GCN module. **xx.x dB** denotes the PSNR measurement on the reconstruction result.

As can be seen from Tab. 2, the accuracy of the GCN module increases as the quality of reconstruction does. It is worth noting that when the reconstruction quality is larger than 23.0dB, the accuracy of the GCN module no longer increases. This shows that when the reconstruction quality reaches a certain level, the information from the global feature is sufficient for the hallucinating process. For the experiments in main paper, we adopt an Autoencoder with 23.0 dB.

In the final end-to-end training stage, we only involve the GCN and the GAN module. Although Autoencoder can also be involved, we find this involvement provides no improvement on GCN's prediction accuracy and GAN's visual reality. As can be seen from Tab. 3, when involving Autoencoder into the end-to-end training process, the prediction accuracy of GCN module remains almost the same. Therefore, we do not include Autoencoder in the our end-to-end training, in which way we can increase the batch size for better visual authenticity.

Term	End-to-end trained with Autoencoder	Trained with fixed Autoencoder
Accuracy(%)	87.94	87.93
Balanced Acc(%)	85.62	85.62

Table 3: Comparing the prediction accuracy of the GCN module when trained with and without Autoencoder end-to-end.

3 Comparison

In this section, we show comparison between images restored by our framework and those by the state-of-the-art image inpainting method [1]. The comparison shows in Fig. 1.

Fig. 1: Comparison on visual results. For each group of images, the first one is the masked image, the second is the inpainted image by [1], the third one is our restored image and the fourth one is the original image.

4 More Visual Examples on COCO, Visual Genome, and NYU V2

For the experiments in the main paper, we test the proposed framework on 20 classes, including apple, backpack, bottle, bowl, cellphone, clock, cup, donuts, frisbee, handbag, keyboard, kite, microwave, mouse, orange, remote, sports ball, stop sign, suitcase, and television.

We provide in this document visual examples from each class of the COCO dataset in Figs. 2 and 3. More visual results from Visual Genome are provided in Fig. 4, and those from NYU V2 are shown in Fig. 5

5 Multi-instance hallucination.

The proposed approach can be readily applied to images with multiple holes, where we hallucinate the holes one by one. We show qualitative and quantitative results in Fig. 6, Fig. 7 and Tab. 4 respectively. As the number of holes in an image increases, the accuracies drop gradually, as expected, since the inference tasks on the graph becomes more challenging with fewer scene objects available.

Missing Nodes	**1**	**2**	**3**	**4**	**5**
Accuracy	87.93	80.12	65.19	38.32	7.71
Balanced Accuracy	85.62	79.93	60.01	35.11	7.10

Table 4: GNN classification accuracies with multiple instances absent.

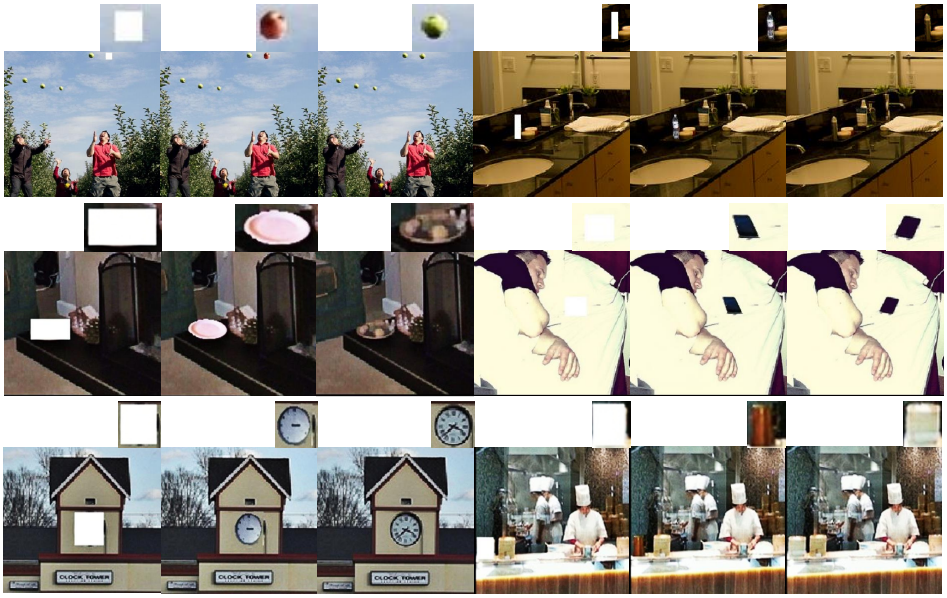

Fig. 2: Visual examples for apple, bottle, bowl, cellphone, clock, and cup hallucination on the COCO dataset.

6 Effect of L_{valid} in RN.

L_{valid} ensures the hole background to be consistent with its surroundings. Fig. 7 shows the results withouts L_{valid}, where discontinuities are observed.

7 Effect of spatial information.

We show in Fig. 8 examples where spatial information is removed in the training and testing of our GAN module. The model produces a semantically reasonable but structurally incorrect prediction.

8 *Background* Hallucination

The proposed method is able to judge whether a missing hole belongs to background, since the *background* class is taken into consideration during the training process. We show some visual results in Fig. 9.

9 Generalization Test

Here, we conduct an interesting experiment to show the generalization capability of our proposed approach, in which, the *background* class is "turn off" during

the training and testing process. Specifically, given an image, we manually mask a region, which *do not* hold an absent object, and then apply our approach for hallucination. As now our approach is trained without the background class, it is enforced to hallucinate a novel object in the designated region.

We show some visual results in Fig. 10, where we see our hallucination results are indeed reasonable. For example, in the right column of the middle row, we are given an image with a mask on the wall. Our network, in this scenario, hallucinates a clock on the wall, which truly makes much sense despite the ground truth shows only the background in the masked region.

References

1. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5505–5514 (2018)

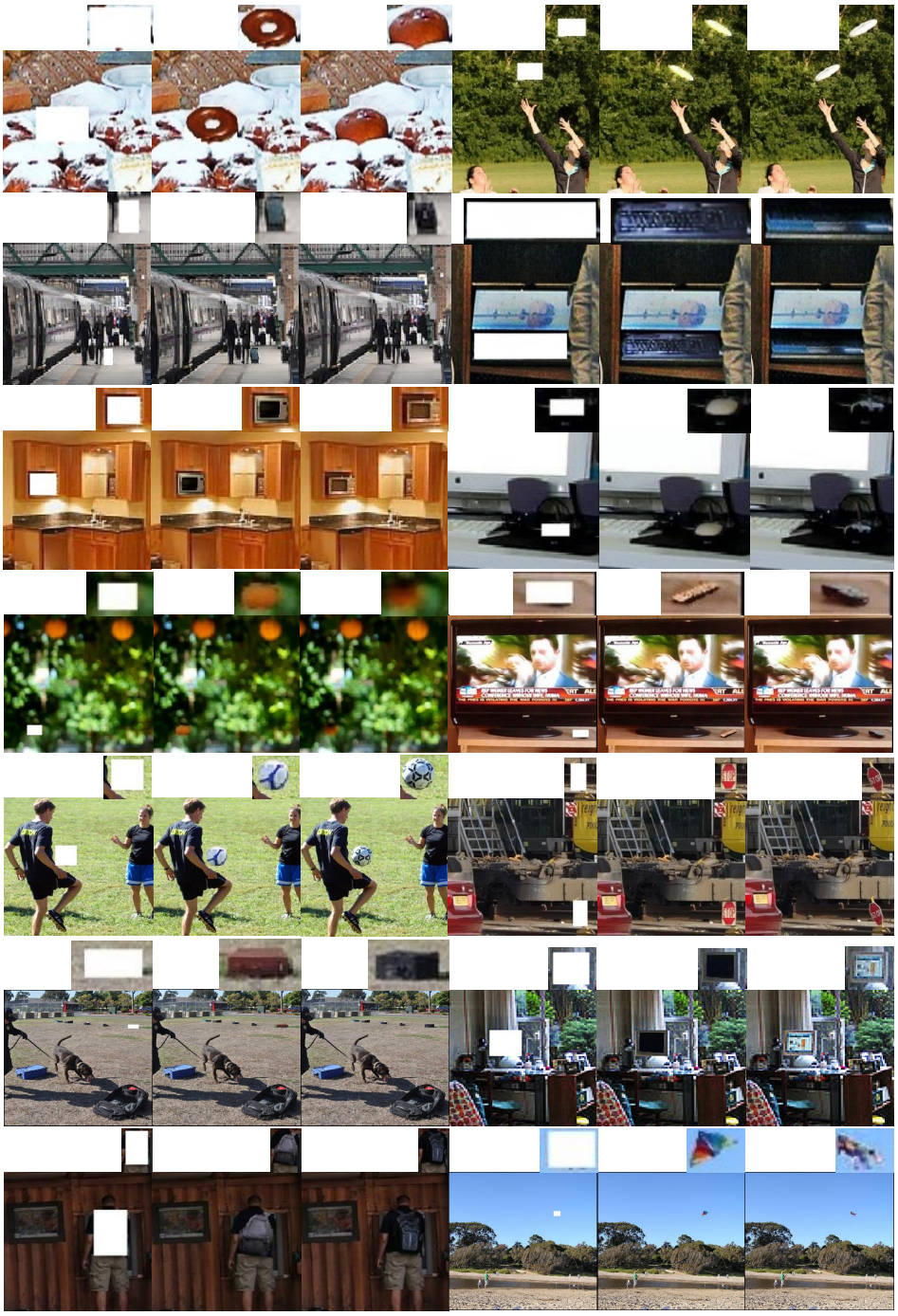

Fig. 3: Visual examples for donuts, frisbee, handbag, keyboard, microwave, mouse, orange, remote, sports ball, stop sign, suitcase, television, backpack, and kite hallucination on the COCO dataset.

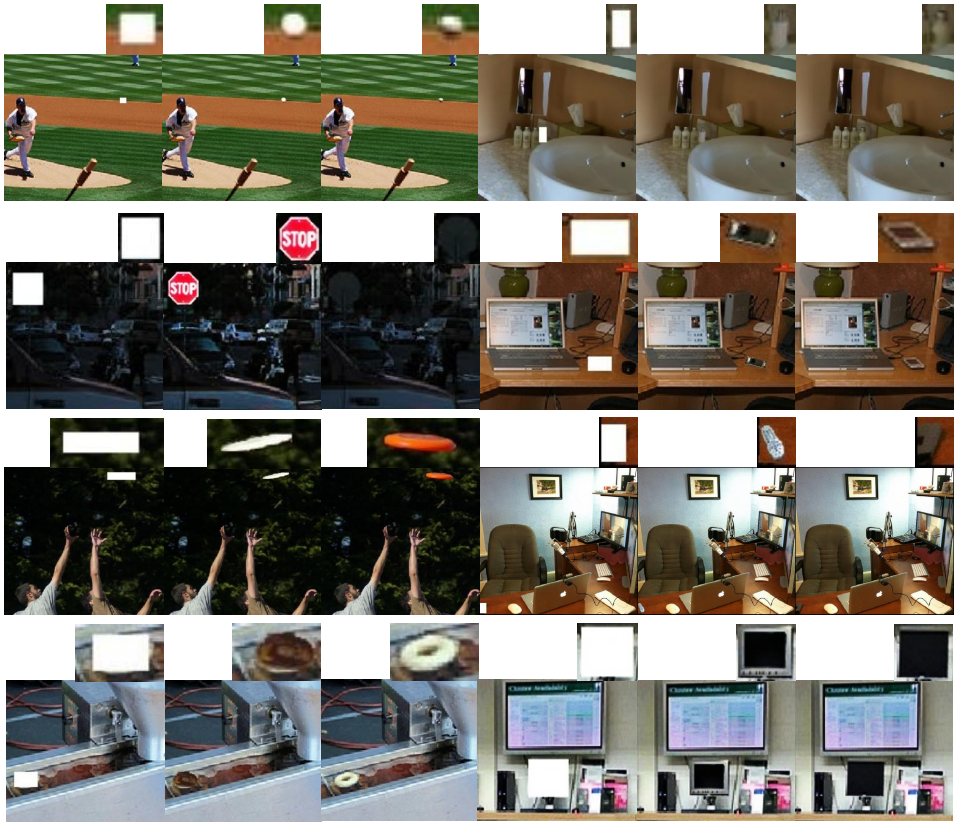

Fig. 4: Visual examples on the Visual Genome dataset.

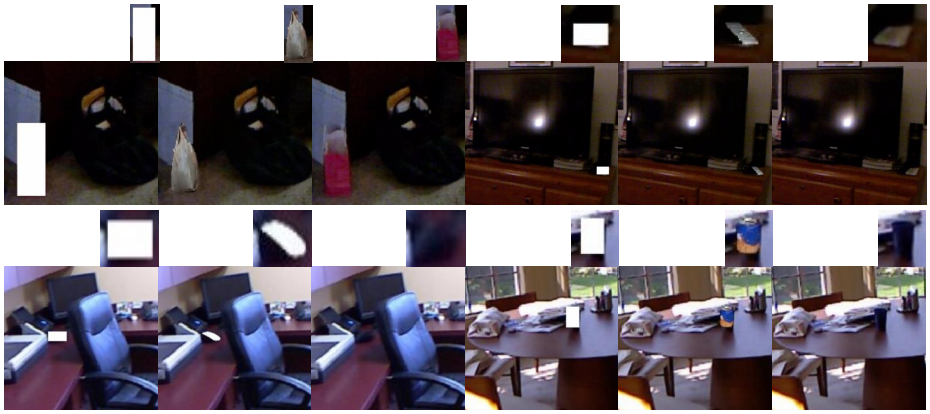

Fig. 5: Visual examples on the NYU V2 dataset.

Fig. 6: Visual examples for multi-instance hallucination.

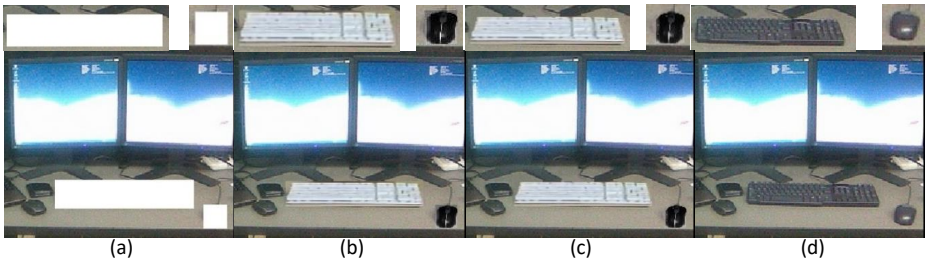

Fig. 7: An example of hallucinating multiple holes. (a), (b), (c) and (d) denote the masked image, the hallucinated image using RN trained without L_{valid}, the hallucinated image using RN trained with L_{valid}, and the original image, respectively.

Fig. 8: An example without spatial information. (a), (b), (c) and (d) denote the masked image, the hallucinated image without spatial information, the hallucinated image with spatial information, and the original image, respectively.

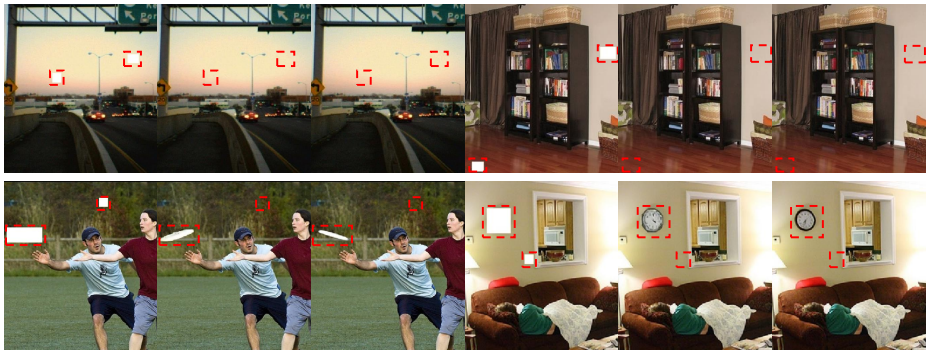

Fig. 9: Visual results of background judgement.

Fig. 10: Testing the generalization capability of the proposed approach. The masked region is manually chosen.