1. RELATED WORK

1.1. Image Inpainting

Conventional image inpainting methods *et al*. [1, 2, 3] fill the holes by borrowing existing content from the known region. These methods cannot generate entirely new content that does not exist in the input image. In recent years, driven by the success of deep generative models, extensive research efforts have been put into data-driven deep learning based approaches [4, 5, 6, 7]. This branch of work usually formulates image completion as an image generation problem conditioned on the existing pixels in known regions. They can generate plausible new content and have shown significant improvements in filling holes in complex images.

Some works attempt to allow users to provide more guidance to reduce the ambiguity of image inpainting and improve the results. Different types of guidance have been explored, such as examplar images, sketches, label maps, and text. Yu *et al*. [8] propose DeepFillV2 for sketch-guided image inpainting. Park [9] explore face inpainting with sketch and color as guidance. Zhang *et al*. [10] propose to inpaint the missing part of an image according to text guidance provided by users. Ardino *et al*. [11] propose to use label maps as guidance for image inpainting. Although the guided inpainting methods [10] and [11] might be able to generate an entire object if the text or label map about the object is given as guidance, they require the users to provide the external guidance explicitly. In comparison, our method only takes the incomplete image and hole mask as input. Compared with text/class conditional inpainting methods, which focus on controlling semantic attributes of the objects, our method provides a more flexible way and allows users to control both the shape and category of the object to inpaint.

1.2. Semantic Image Synthesis

Semantic image synthesis is a subclass of conditional image generation aimed at generating photorealistic images from user-specified semantic layouts. It was first introduced by Isola *et al*. [12], who proposed an image-to-image translation framework, called Pix2Pix, to generate images from label maps or edge maps. Zhu *et al*. [13] propose CycleGAN where an image translation model can be trained on unpaired data with a cycle consistency constraint. Park *et al*. [14] propose spatially-adaptive normalization for semantic image synthesis, which modulates the activations using semantic layouts to propagate semantic information throughout the network. Chen *et al*. [15] propose cascaded refinement networks and use perceptual losses for semantic image synthesis. Wang *et al*. [16] propose Pix2PixHD which improves the quality of synthesized images using feature matching losses, multiscale discriminators, and an improved generator. Our method takes inspiration from semantic image synthesis methods to design the top-down stream of the contextual object generator. Unlike semantic image synthesis, where the semantic layouts or label maps are known, our semantic object maps are derived by combining the predicted class and the hole mask.

1.3. Background-based Object Recognition

Object recognition is the task of categorizing images according to visual content. In recent years, the availability of large-scale datasets and powerful computers made it possible to train deep CNNs, which achieved breakthrough success for object recognition [17]. Normally, an object recognition model categorizes an object primarily by recognizing visual patterns in the foreground region. However, recent research has shown that a deep network can produce reasonable object results with only background available. Zhu *et al*. [18] find that the AlexNet model [17] trained on a pure background without objects achieves a highly reasonable recognition performance that beats human recognition in the same situations. Xiao *et al*. [19] analyze the performance of state-of-the-art architectures on object recognition with foreground removed in different ways. It is reported that the models can achieve over 70% test accuracy in a no-foreground setting where the foreground objects are masked. These works aim to predict only the class of an object from background. In this paper, we show that the entire object can be generated based on the background.

2. ABLATION STUDY

Object prior. Unlike previous work on image inpainting which generates the training data using random masks, we construct the specialized training data for object inpainting to incorporate object prior. Without this prior, the trained inpainting model usually has the bias towards background generation and will not generate objects when filling a missing region, as shown in Fig. 1 (b).

The predictive class embedding (PCE). PCE extracts information related to the class from the context. Without this module, the model trained on object data might be able to produce object-like content. However, it is challenging to generate a semantically reasonable object without knowing the object's class. As shown in Fig. 1 (c), usually the appearance of the generated objects are simply taken from the nearby regions. For instance, in the second row, the model without PCE generates an object of zebra shape but with the texture of a nearby giraffe. By default, the dimension of the class embedding is set to 1024 for all experiments. We find that the performance is not very sensitive to the dimension of the class embedding. The second column of Table 3 reports the results with the 512-dimensional class embedding, which is close to the results with the 1024-dimensional embedding.

Top-down stream. The top-down stream takes the semantic object mask as input, which provides a stronger spatial semantic guidance for object generation. Without this infor-

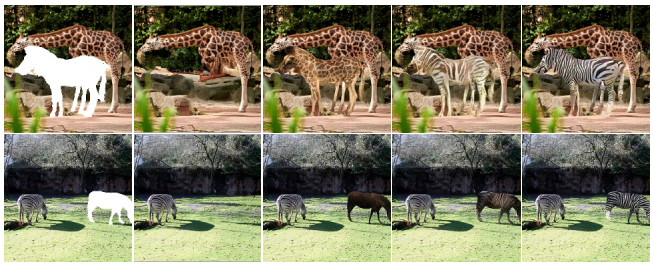


Fig. 1: From left to right are: (a) input, (b) without object training data, (c) without predictive class embedding, (d) without top-down stream, (e) full model.

Table 1: Effect of each component in terms of FID and LPIPS.

Object Data	PCE	Top-down	FID	LPIPS
√			6.144	0.1066
√	√		5.434	0.1081
√	√	√	4.700	0.1049

mation, the model can only access class-related information from PCE, which is insufficient for hallucinating object appearance. Hence the model will still rely on the appearance of the surrounding area. As shown in Fig. 1 (d), although the model without the top-down stream can produce some zebra strikes, the color of the zebra seems to be from the surrounding background area. Table 1 reports FID scores with and without each component. We can see that the predictive class embedding and the incorporation of the top-down stream can significantly reduce the FID by providing class-related information. To further demonstrate the effect of the top-down stream, we present the results with altered class labels in Fig. 3 (c). Specifically, in this experiment, when constructing the semantic object maps, we manually assign a class label rather than using the predicted class. For instance, for the example in the left of the first row, the predicted class is giraffe as shown in (b). To obtain the result in (c), we manually assign the zebra class to the semantic object map. From the results in Fig. 3 we can see that an object generated with altered class has lighting and style similar to the original result but has the class-related feature of the assigned class.

Shape Guidance. In shape-guided object inpainting, the guidance is given implicitly by the shape of a missing region. To explore the effect of shape guidance, we train the models using object masks of different precision: precise masks, coarse masks obtained by dilating the original masks with 21 pixels, and square masks. The FID, LIPIS as well as the recognition accuracy of the missing objects are shown in Table 2. We can see that dilated masks lead to a performance drop within a reasonable range, and square masks lead to a large drop in both image quality and recognition accuracy.

SC AdaIN. We use positional normalization following [20]

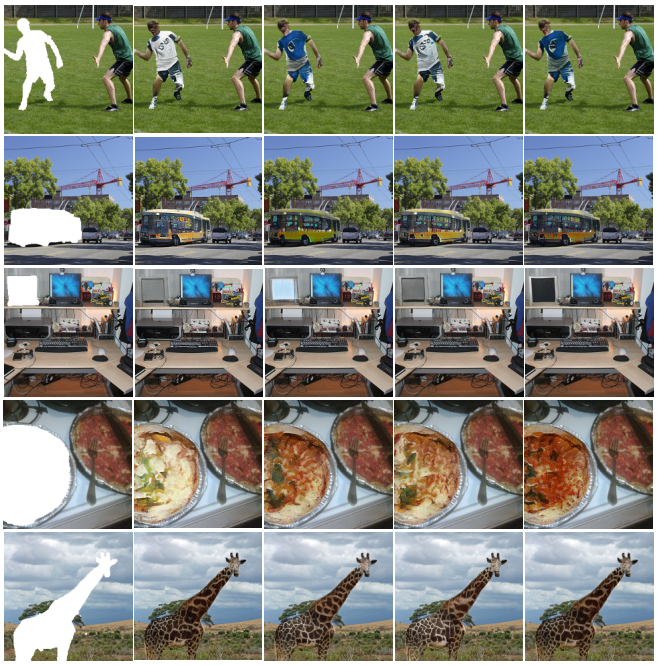


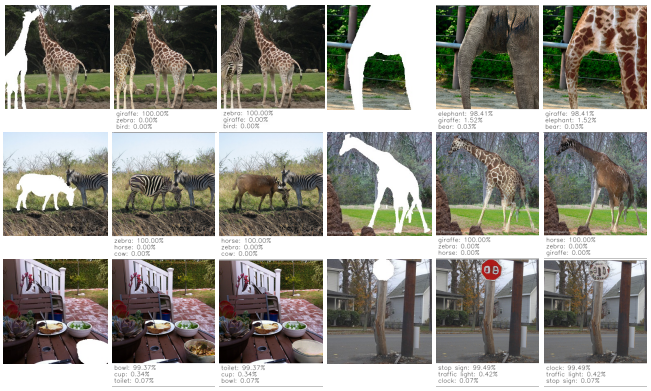
Fig. 2: Our method can produce multiple diverse object inpainting results for the same input image by using different random latent code z.

Table 2: Effect of imprecise masks. Acc1: top 1 accuracy; Acc3: top 3 accuracy.

Masks	FID	LPIPS	Acc1	Acc3
Precise Masks	7.693	0.1122	0.5100	0.7180
Dilated Masks	10.25	0.1371	0.4104	0.6760
Square Masks	15.70	0.4192	0.1740	0.4760

as it computes the statistics at each spatial position and can better preserve the structure information. The fourth and fifth columns of Table 3 report the results obtained by aggregating encoder feature maps with instance normalization and concatenation. Using positional normalization (second column) yields better results than the other choices.

User-drawn masks. Our method is robust to imperfect masks as long as the shapes in masks are recognizable. As indicated in Table 2, using square hole masks leads to an obvious performance drop, while the performance with dilated masks is comparable to accurate masks. Therefore, for object insertion, imperfect user-drawn masks are also acceptable. Fig. 4 shows example results with user-drawn masks as input. As the proposed method predicts classes based on both shape and context information, it can still generate reasonable objects when one of them is ambiguous, *i.e.* when the object masks are imprecise or with mismatched masks and context, *e.g.* a giraffe mask on a beach.



(a) (b) (c) (a) (b) (c)

Fig. 3: Results obtained by altering the class of the semantic object map. The texts under the images indicate the predicted class or the manually assigned class. (a) input; (b) results with the predicted classes; (c) results with the assigned classes.

Table 3: Quantitative results on COCO validation set. Ours: ours default setting; Ours-512: ours with 512 dimensional class embedding; Ours-IN: results with instance normalization; Ours-Concat: results with concatenation.

Settings	Ours	Ours-512	Ours-IN	Ours-Concat
FID↓	4.700	4.742	5.492	5.226
LPIPS↓	0.1049	0.1052	0.1075	0.1077

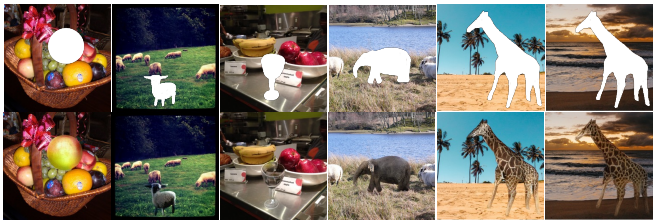


Fig. 4: Object inpainting results with user-drawn masks.

3. REFERENCES

[1] Alexei A Efros and Thomas K Leung, "Texture synthesis by non-parametric sampling," in *International Conference on Computer Vision*. IEEE, 1999, vol. 2, pp. 1033–1038. 1

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 24, 2009. 1

[3] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE Transaction on Image Process.*, vol. 10, no. 8, pp. 1200–1211, 2001. 1

[4] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–14, 2017. 1

[5] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang, "Generative image inpainting with contextual attention," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[6] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *European Conference on Computer Vision*, 2018. 1

[7] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu, "High-resolution image inpainting with iterative confidence feedback and guided upsampling," in *European Conference on Computer Vision*. Springer, 2020. 1

[8] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang, "Free-form image inpainting with gated convolution," in *International Conference on Computer Vision*, 2019. 1

[9] Youngjoo Jo and Jongyoul Park, "Sc-fegan: Face editing generative adversarial network with user's sketch and color," in *International Conference on Computer Vision*, 2019, pp. 1745–1753. 1

[10] Lisai Zhang, Qingcai Chen, Baotian Hu, and Shuoran Jiang, "Text-guided neural image inpainting," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1302–1310. 1

[11] Pierfrancesco Ardino, Yahui Liu, Elisa Ricci, Bruno Lepri, and Marco De Nadai, "Semantic-guided inpainting network for complex urban scenes manipulation," IEEE, 2021. 1

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134. 1

[13] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *International Conference on Computer Vision*, 2017, pp. 2223–2232. 1

[14] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346. 1

[15] Qifeng Chen and Vladlen Koltun, "Photographic image synthesis with cascaded refinement networks," 2017, pp. 1511–1520. 1

[16] 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," 2018, pp. 8798–8807. 1

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012. 1

[18] Zhuotun Zhu, Lingxi Xie, and Alan L Yuille, "Object recognition with and without objects," in *International Joint Conference on Artificial Intelligence*, 2017. 1

[19] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry, "Noise or signal: The role of image backgrounds in object recognition," in *International Conference on Learning Representations*, 2020. 1

[20] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *CVPR*. 2