PANOPTICALLY GUIDED IMAGE INPAINTING WITH IMAGE-LEVEL AND OBJECT-LEVEL SEMANTIC DISCRIMINATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent image inpainting methods have made great progress. However, the existing approaches often struggle to hallucinate realistic object instances in natural scenes. Such a limitation is partially due to the lack of semantic-level constraints inside the hole as well as the lack of a mechanism to enforce the realism of local objects. To tackle the challenging object inpainting task, we propose a new panoptically guided image inpainting task that leverages a panoptic segmentation map to guide the completion of object instances. To enforce the realism of the generated objects, we propose a semantic discriminator that leverages pretrained visual features to improve the generated semantics. Furthermore, we propose object-level discriminators that take aligned instances as input to enforce the realism of individual objects. Experiments on the large-scale Places2 dataset demonstrate the significant improvement by our method on object completion, verified in both quantitative and qualitative evaluation. Furthermore, our framework is flexible and can be generalized to other inpainting tasks including segmentation-guided inpainting, edge-guided inpainting, as well as standard image inpainting without guidance. Consequently, our approach achieves new state-of-the-art performance on the various inpainting tasks and impressive results on object completion.

1 INTRODUCTION

Image inpainting refers to the task of completing missing regions of an image. As a fundamental research problem, it has many practical applications such as background or object completion, image re-targeting, compositing, and editing. Given the development of generative adversarial networks (GANs) and the impressive power of GANs on hallucinating local textures and simple semantic structures, recent works on image inpainting (Yu et al., 2018; Zeng et al., 2020; Suvorov et al., 2021; Zhao et al., 2021; Zheng et al., 2022) have shown impressive results on removing distracting objects while completing the missing region with background pixels. However, limited by the capacity of current generative models in synthesizing complex natural scenes with randomly placed objects (Sauer et al., 2022), inpainting a large missing region to produce a reasonable semantic layout and realistic object instances remains a huge challenge. As Fig. 1 depicts, the existing inpainting approaches often lead to obvious structural artifacts such as distorted objects and degenerated semantic layout, which significantly impact the inpainting quality. As such, how to inpaint large missing regions while maintaining a reasonable semantic layout and realistic object instances remains an open and essential problem for image completion.

One way to enhance large hole completion is to provide a guidance map as an additional input so that the image completion process follows the provided structural hint or guidance. In the literature, such a scheme known as guided inpainting (Nazeri et al., 2019; Zeng et al., 2021) often leverage guidance maps such as an edge map (Nazeri et al., 2019; Zeng et al., 2021; Yu et al., 2019), semantic map (Ntavelis et al., 2020; Song et al., 2018) or color map (Portenier et al., 2018; Jo & Park, 2019) to provide structural clues for better completion. Despite the promising results, the existing methods still suffer from generating unnatural objects or semantic layout due to the limited capacity on generating semantically coherent structure and realistic objects. Furthermore, the existing semantic label map or edge map guidance do not offer the fine-grained and instance-level semantics information that is critical for completing complex scenes, e.g., a group of interacting people.

Figure 1: We propose a panoptically guided inpainting task that leverages fine-grained and instance-level panoptic segmentation to tackle the challenging use case of object instance inpainting. To enable photo-realistic inpainting, we propose a novel semantic discriminator design and object-level discriminators. Compared to CM-GAN (Zheng et al., 2022), the recent state of the art inpainting models and the re-trained CM-GAN$*$ for the panoptic-guided task, our panoptically guided approach generates much higher quality results on objects.

In this work, we tackle a challenging large-hole guided image completion task where the goal is to complete whole or a large part of objects that are arbitrarily located in natural scene. Different from the well-established guided inpainting methods that leverage semantic (Ntavelis et al., 2020; Song et al., 2018) or edge map guidance (Nazeri et al., 2019; Zeng et al., 2021; Yu et al., 2019), we first propose a new *panoptically guided inpainting* task that leverages a panoptic segmentation map (Kirillov et al., 2019) to provide fine-grain and instance-level semantic clue inside the hole, avoiding the confusion caused by overlapped instances with the same semantic class. However, naively applying the state-of-the-art inpainting models (Zhao et al., 2021; Suvorov et al., 2021; Zheng et al., 2022) while treating the panoptic guidance as condition often leads to poor results, i.e. distorted object instances and degenerated semantic layout as shown in Fig. 4. In this work, we found that semantic and object-level modeling are crucial to the guided-inpainting quality. Consequently, we propose a novel learning scheme that leverages *a semantic discriminator* and *an object-level discriminator* to enforce both semantic-level realism and quality of generated objects. In particular, our semantic discriminators leverage the semantic understanding capacity of pretrained visual models (Radford et al., 2021) to enhance the generation of semantic layout. Meanwhile, the object-level discriminators take the aligned and cropped object as input to determine the quality of fine-grained objects instances at a local scale. As shown in Fig. 1, our panoptically guided inpainting system with the semantic and object-level discriminators significantly boost the realism of the completed objects and leads to significant gain over the current architectures on the panoptically guided task. Moreover, our proposed framework is versatile and can be applied to other guided-inpainting tasks, including semantics-guided image inpainting (Ntavelis et al., 2020) and edge-guided image inpainting (Zeng et al., 2021). Furthermore, with slight modification, our trained model can be applied to the standard image inpainting task while showing significant improvement over the recent methods (Suvorov et al., 2021; Zhao et al., 2021; Zheng et al., 2022). With the newly introduced components, our methods significantly boost the generation quality of all four inpainting tasks and produce very promising results for large-hole image completion and object completion.

Our contributions are three-fold:

- A novel panoptically guided inpainting task to facilitate the completion of object instances for image inpainting.
- A new semantic discriminator design that leverages the pretrained visual features to encourage the semantic consistency of the generated contents and a novel object-level discriminator framework that enforces the realism of the generated local objects.
- State-of-the-art results on the Places2 dataset for various tasks including panoptically guided inpainting, semantic-guided inpainting, edge-guided inpainting, as well as standard image inpainting.

2 METHOD

As depicted in Fig. 2, our approach for the panoptically guided image inpainting task is based on conditional Generative Adversarial Networks (Mirza & Osindero, 2014) to complete the missing region of an image $\boldsymbol{X}$ annotated by a binary mask $\boldsymbol{M}$ according to a guidance panoptic segmentation map $\boldsymbol{P}$. Unlike color (Portenier et al., 2018), edge map (Yu et al., 2019; Nazeri et al., 2019; Xiong et al., 2019) or semantic map (Song et al., 2018; Ntavelis et al., 2020) guided inpainting tasks,

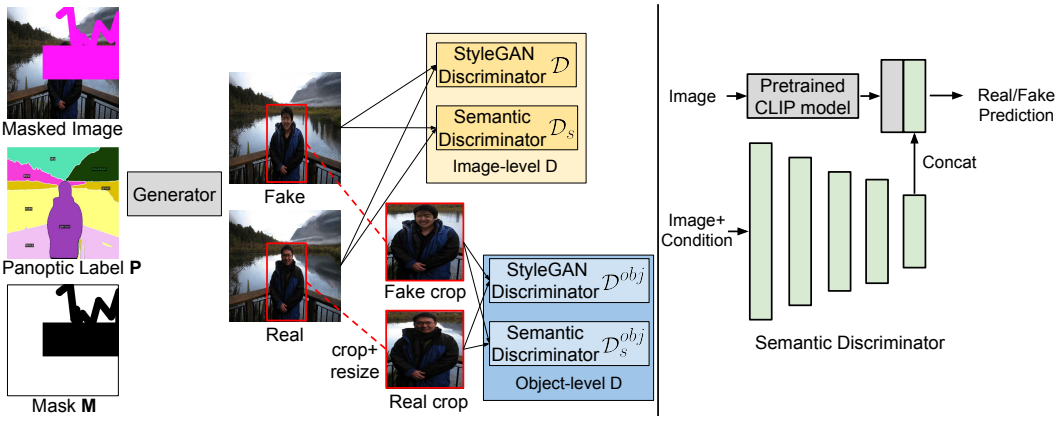

Figure 2: Left: Our guided generative inpainting model leverage a combination of vanilla StyleGAN discriminator (Karras et al., 2020b) and the proposed semantic discriminators at both image level and object level to enforce semantic and object coherency. The object-level discriminators take the resized object crop as inputs to enforce realism of object instances. Right: the semantic discriminators leverage the semantic knowledge of the pretrained CLIP (Radford et al., 2021) model to enforce the realism of generated semantic.

panoptic segmentation (Kirillov et al., 2019) provides fine-grained semantic annotation and instance-level contour to facilitate the completion of individual objects inside the hole.

Specifically, the panoptic guidance label $\boldsymbol{P}_i$ at each pixel i is denoted as a tuple (l_i, z_i), where $l_i \in \{0, \cdots, L-1\}$ represents the semantic class of pixel i and $z_i \in \mathbb{N}$ represents its instance id. To better adopt the panoptic annotation as conditions inputs to our model, we convert the panoptic annotation $\boldsymbol{P}$ to a semantic label map $\boldsymbol{L}$ that indicates the semantic layout and a binary edge map $\boldsymbol{E}$ that represents the boundary of the panoptic segmentation as shown in Fig. 2. Following such a formulation, our generator $\mathcal{G}$ takes an incomplete image $\boldsymbol{X} \odot (1-\boldsymbol{M})$, a mask $\boldsymbol{M}$, the semantic label map $\boldsymbol{L}$ and the edges map $\boldsymbol{E}$ to predict the completed image $\hat{\boldsymbol{X}} = \mathcal{G}(\boldsymbol{X} \odot (1-\boldsymbol{M}), \boldsymbol{M}, \boldsymbol{L}, \boldsymbol{E})$; Furthermore, a discriminator $\mathcal{D}$ predicts a score $\hat{y} = \mathcal{D}(\hat{\boldsymbol{X}}, \boldsymbol{M}, \boldsymbol{L}, \boldsymbol{E})$ that indicates how likely $\hat{\boldsymbol{X}}$ is the ground-truth.

2.1 Network Architecture

2.1.1 The Generator

Recently, Cascaded-Modulation GAN (CM-GAN) (Zheng et al., 2022) has shown significant improvement in standard image inpainting tasks thanks to the architecture design that cascades modulation blocks for better global context modeling. Therefore, we adopt the CM-GAN generator to our guided inpainting task to leverage the strong inpainting capacity of the CM-GAN generator. However, we pass the additional panoptic guidance to the generator to leverage the panoptic guidance. Specifically, we decompose the panoptic map $\boldsymbol{P}$ into a semantic label map $\boldsymbol{L}$ and an edges map $\boldsymbol{E}$. Then, we pass the semantic label map $\boldsymbol{L}$ to an embedding layer following ℓ_2 normalization to produce a normalized semantic embedding $\boldsymbol{S}$. Finally, the concatenation of the incomplete image $\boldsymbol{X} \odot (1-\boldsymbol{M})$, mask $\boldsymbol{M}$, semantic embedding $\boldsymbol{S}$, and the edges map $\boldsymbol{E}$ are passed to the generator to predict the completed image.

2.1.2 The Semantic and Object Discriminators

Following recent inpainting works (Zheng et al., 2022; Zhao et al., 2021) that leverage StyleGAN discriminator (Karras et al., 2020b) for adversarial learning, we adopt a panoptically conditioned discriminator $\mathcal{D}$ that takes the concatenation of generated image $\hat{\boldsymbol{X}}$ and the condition $\boldsymbol{M}$, $\boldsymbol{S}$, $\boldsymbol{E}$ as inputs to output the discriminator score $\hat{y}$:

$$\hat{y} = \mathcal{D}(\hat{\boldsymbol{X}}, \boldsymbol{M}, \boldsymbol{S}, \boldsymbol{E}). \tag{1}$$

We found that such a adversarial learning scheme indeed achieves leading results comparing to other baseline models. However, due to the lack of semantic-level supervision and constraint on objects,

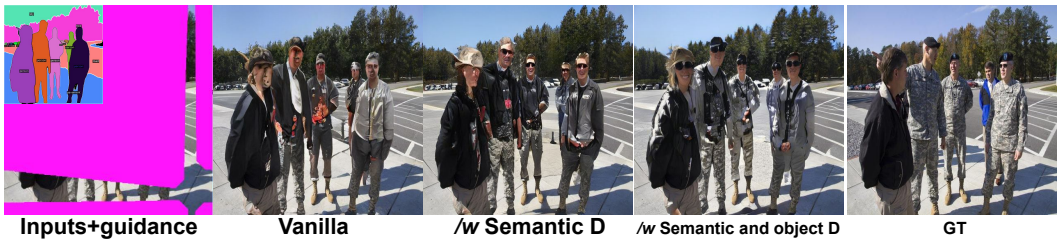

Figure 3: The image-level semantic discriminator and the object-level discriminators progressively improve the photo realism of the generated image (e.g., face and body) in comparison to the baseline trained with only the StyleGAN discriminator (Karras et al., 2020b).

the generator trained with the conditioned StyleGAN discriminator tends to hallucinate distorted objects or degenerated semantic layout as depicted in Fig. 3, which seriously impact the inpainting quality. Hence, we propose a novel *semantic discriminator* for improving semantic coherency of completion and *object-level discriminators* for enhancing the photo realism of the individually generated objects.

Semantic Discriminator. To generate realistic object instances and a complex semantic layout, a discriminator should distinguish whether the generated contents $\hat{\boldsymbol{X}}$ is realistic and conformed to the given semantic layout. However, Kumari et al. (2021); Wang et al. (2020) show that discriminator may potentially focus on artifacts that are imperceptible to humans but obvious to a classifier and that the learned visual feature may cover only parts of the visual concept (Sauer et al., 2021a) while ignoring other parts. Therefore, with the regular adversarial learning between $\mathcal{G}$ and $\mathcal{D}$, it is challenging for the generator to discover complex semantic concepts or hallucinate realistic objects.

To tackle this issue, we propose a semantic discriminator $\mathcal{D}_s$ that leverages the visual representation extracted by a pretrained vision model (Radford et al., 2021) to discriminate the semantic-level realism. Benefiting from the comprehensive semantic concepts captured by pretrained vision models (Bau et al., 2020), our semantic discriminator better captures high-level visual concepts, and in turn improves the realism of the generated semantic layout, c.f. Fig. 3. Specifically, our semantic discriminator $\mathcal{D}_s$ takes the generated image and the panoptic condition as inputs:

$$\hat{y}_s = \mathcal{D}_s(\hat{\boldsymbol{X}},\, \boldsymbol{M},\, \boldsymbol{S},\, \boldsymbol{E}), \tag{2}$$

and output the semantic-level realism prediction. As shown in Fig. 2 (right), the semantic discriminator is based on the two branches of the encoder to extract complementary features: a pretrained ViT model branch (Radford et al., 2021) produces visual feature of the completed image, and a trainable encoder based on strided convolution to extract condition feature from the concatenation of the condition $\hat{\boldsymbol{X}}, \boldsymbol{M}, \boldsymbol{S}, \boldsymbol{E}$. Finally, the pretrained feature and the encoder feature at the final scale are concatenated to produce the final discriminator prediction. As the semantic discriminator are designed to classify the high-level structure at the semantic level, we found that combining StyleGAN discriminator (Kumari et al., 2021) improves the generated local textures.

Object-level Discriminators. Recent progress on image generation (Karras et al., 2019; 2020b) demonstrates impressive results on generating objects such as face, car, animal (Karras et al., 2017) or body (Ma et al., 2017) in an aligned setting where objects are carefully placed or registered in the center of the image. However, generating unaligned objects in complex natural scene is known challenging (Sauer et al., 2022) for various tasks including inpainting (Park et al., 2019; Zhao et al., 2021) and semantic image generation (Park et al., 2019). Although semantic discriminator can improve quality of the generated objects, generating photo-realistic instances is still challenging. To improve the realism of completed objects, we found that the object-level alignment mechanism for discriminator has a profound impact on improving inpainting quality. Consequently, we propose novel object-level discriminators that are dedicated to model the hierarchical composition of aligned objects for predicting the object-level realism. In particular, as shown in Fig. 2, given an object instance and its bounding box $\boldsymbol{b} = (x_0, y_0, x_1, y_1)$, an object-level discriminator takes the crop-and-resized image $\hat{\boldsymbol{X}}_a$ and the corresponding crop-and-resized condition maps $\boldsymbol{M}_a, \boldsymbol{L}_a, \boldsymbol{E}_a$ as inputs to predict the realism of the object. In addition, an object-level discriminator also takes a binary map $\boldsymbol{I}_{crop}$ as input to indicate the shape of the instance. To enhance the capacity of discriminator for

Figure 4: Qualitative comparisons on the panoptically guided inpainting task on Places2-person. We compare our model against LaMa* (Suvorov et al., 2021), CoModGAN* (Zhao et al., 2021), CM-GAN* (Zheng et al., 2022) whereas $*$ denotes models re-trained with the additional panoptic segmentation condition for panoptically guided inpainting. Best viewed by zoom-in on screen.

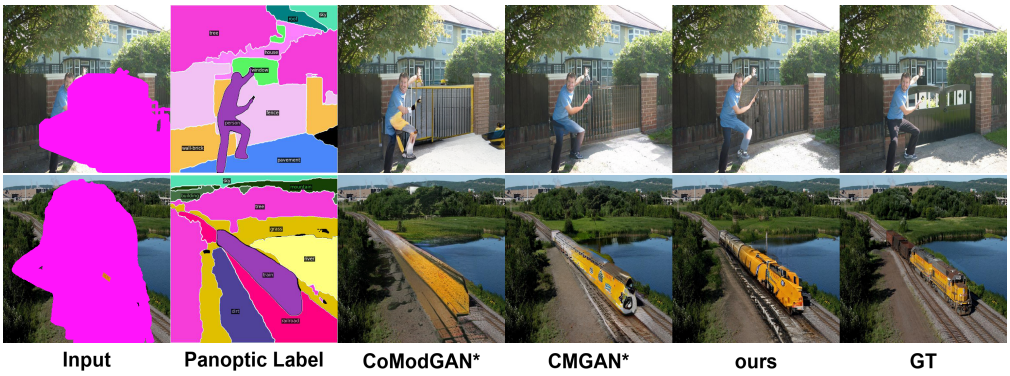

Figure 5: Qualitative comparisons on the panoptically guided inpainting task on Places2-object. We compare our model against retrained CoModGAN$*$ (Zhao et al., 2021), CM-GAN$*$ (Zheng et al., 2022) whereas the $*$ symbol denotes models re-trained with the additional panoptic segmentation condition for panoptically guided inpainting. Best viewed by zoom-in on screen.

object modeling, following the image-level discriminators, our object-level discriminators depicted in Fig. 2 are implemented as a combination of an object-level discriminator $\hat{y}^{obj}$

$$\hat{y}^{obj} = \mathcal{D}^{obj}(\hat{\boldsymbol{X}}_a, \boldsymbol{M}_a, \boldsymbol{L}_a, \boldsymbol{E}_a, \boldsymbol{I}_a), \tag{3}$$

and an object-level semantic discriminator $\mathcal{D}_s^{obj}$

$$\hat{y}_s^{obj} = \mathcal{D}_s^{obj}(\hat{\boldsymbol{X}}_a, \boldsymbol{M}_a, \boldsymbol{L}_a, \boldsymbol{E}_a, \boldsymbol{I}_a), \tag{4}$$

where $\hat{y}^{obj}$ and $\hat{y}_s^{obj}$ represent how likely the object instance is the ground-truth object patch. The object-level discriminator $\mathcal{D}^{obj}$ follows the implementation of image-level StyleGAN discriminator $\mathcal{D}$ while the object-level semantic discriminator $\mathcal{D}_s^{obj}$ follows the implementation of image-level semantic discriminator $\mathcal{D}_s$.

2.1.3 Training Objective

Our training objective is a summation of non-saturating adversarial loss (Goodfellow et al., 2014) for the generator $\mathcal{G}$ and a set of StyleGAN and semantic discriminators at both image level and

Table 1: Evaluation of panoptically guided inpainting, semantic-guided inpainting and edge-guided inpainting on Places2-person. Methods with $*$ are re-trained on the guided-inpainting tasks.

	CoModGAN masks			Object masks		
Methods	FID↓	U-IDS (%)↑	P-IDS (%)↑	FID↓	U-IDS (%)↑	P-IDS (%)↑
inpainting w/ panoptic segm.						
SESAME$*$ (Ntavelis et al., 2020)	12.0061	7.41	0.24	8.3656	9.91	0.41
LaMa$*$ (Suvorov et al., 2021)	7.0563	21.74	4.83	4.8156	26.28	7.39
CoModGAN$*$ (Zhao et al., 2021)	5.0168	29.58	14.94	4.4232	31.59	16.75
CM-GAN$*$ (Zheng et al., 2022)	2.8470	33.20	18.50	2.7246	34.42	19.66
ours	**2.0720**	**36.96**	**25.90**	**1.8682**	**37.90**	**26.30**
inpainting w/ semantic segm.						
SESAME (Ntavelis et al., 2020)	12.2308	7.09	0.22	8.3940	9.75	0.40
ours	**2.3860**	**33.11**	**19.25**	**2.1565**	**34.85**	**21.12**
inpainting w/ edge						
EdgeConnect (Nazeri et al., 2019)	41.7631	3.18	0.04	22.9517	3.98	0.06
SketchEdit (Zeng et al., 2021)	16.1271	13.02	1.58	8.7878	19.77	3.21
ours	**2.6909**	**33.43**	**20.45**	**2.1873**	**36.19**	**23.46**

Table 2: Quantitative evaluation of panoptically guided inpainting on Places2-object.

Methods	FID↓	U-IDS (%)↑	P-IDS (%)↑	Methods	FID↓	U-IDS (%)↑	P-IDS (%)↑
CoModGAN$*$	5.9140	31.39	15.44	SESAME$*$	7.6420	11.92	0.64
LaMa$*$	4.1189	31.05	11.35	CM-GAN$*$	3.3929	36.02	20.92
ours	**3.2126**	**37.58**	**25.80**				

object level $\mathbf{D} = \{\mathcal{D}, \mathcal{D}_s, \mathcal{D}^{obj}, \mathcal{D}_s^{obj}\}$:

$$\mathcal{L}_{adv} = \sum_{\mathcal{D}\in\mathbf{D}} \log \mathcal{D}(x) + \log(-\mathcal{D}(\hat{\boldsymbol{X}})), \tag{5}$$

where $\hat{\boldsymbol{X}}$ is the generated image. To improve the generated the textures while stabilizing the training, we incorporate perceptual loss (Johnson et al., 2016) as the additional reconstruction loss $\mathcal{L}_{rec} = \sum_{l=1}^{L} \|\Phi^{(l)}(\hat{\boldsymbol{X}}) - \Phi^{(l)}(\boldsymbol{X}))\|_1$, where $\Phi^{(l)}$ is the feature representation of a pretrained network at scale $l \in \{1, \cdots, L\}$ whereas $L = 4$. We use a pretrained segmentation model with high receptive field to improve large-mask inpainting (Suvorov et al., 2021).

3 Experiments

3.1 Implementation Details

Datasets and evluation. We collect two large-scale object-centric datasets named Places2-person and Places2-object from the Places2 dataset (Zhou et al., 2017) for evaluating various object inpainting task in various settings. Specifically, Places2-person and Places2-object are subsets of Places2 dataset that contain at least one person or general object instances, respectively. We leverage pretrained PanopticFCN model (Li et al., 2021) to generate panoptic segmentation annotations for both datasets and apply the random stroke mask (Zhao et al., 2021) and object-shaped masks (Zeng et al., 2020) for model evaluation. We report the numerical metrics on test sets using the mask scheme of CoModGAN (Zhao et al., 2021) and the object masks of (Zeng et al., 2020) and report *Frchet Inception Distance* (FID) (Heusel et al., 2017) and the *Paired/Unpaired Inception Discriminative Score* (P-IDS/U-IDS) (Zhao et al., 2021) for evaluation.

Inpainting Tasks and compared methods. We evaluate our model on the *panoptically guided*, *semantic guided* and *edge guided* inpainting tasks. Furthermore, Sec. 3.4 proposes a variant of our model on the *standard inpainting* task. For the panoptically guided task, we compare our method with the recent inpainting and guided-inpainting methods including SESAME$*$ (Ntavelis et al., 2020), LaMa$*$ (Suvorov et al., 2021), CoModGAN$*$ (Zhao et al., 2021) and CM-GAN$*$ (Zheng et al., 2022), where $*$ symbol denotes models retrained for the panoptically guided task. All the retrained models are trained on 8 A100 GPUs for at least three days and until convergence to ensure fair comparisons. For semantic-guided inpainting, we compare our method with SESAME (Ntavelis et al., 2020) and for the edge-guided inpainting, we compare our method with Edge-connect (Nazeri et al., 2019) and SketchEdit (Zeng et al., 2021).

Figure 6: Qualitative comparisons on the standard inpainting task (Sec. 3.4). Compared to the existing methods, our method can generate high-quality and photo-realist object instances.

Table 3: Ablation studies of our model. *Adv.*, *perc.*, *sem. D*, *obj. D* are abbreviations of adversarial loss, perceptual loss, semantic discriminator and object-level discriminator, respectively.

	CoModGAN masks			Object masks		
Methods	FID↓	U-IDS (%)↑	P-IDS (%)↑	FID↓	U-IDS (%)↑	P-IDS (%)↑
ours *w/* adv.	10.5587	20.60	5.56	9.3800	22.70	6.38
ours *w/* adv. + perc.	2.8470	33.20	18.50	2.7246	34.42	19.66
ours *w/* adv. + perc. + sem. D	2.2705	35.41	22.30	2.1636	36.12	22.98
ours *w/* adv. + perc. + sem. D + obj. D (full)	**2.0720**	**36.96**	**25.90**	**1.8682**	**37.90**	**26.30**
ours full *w/* semantic segm.	2.3860	33.11	19.25	2.1565	34.85	21.12

Network Details. We leverage the pretrained CLIP model (Radford et al., 2021) as backbone to extract feature for the semantic discriminators. Please refer to the appendix for more details about the network structure and training configuration.

3.2 Quantitative and Qualitative Evaluation

Tab. 1 presents the evaluation on the panoptic, semantic and edge-guided inpainting tasks on Places2-person. For all the tasks, our method achieves significantly gain compared to the existing methods. In addition, we observe that our panoptically guided model achieves better FID scores compared the semantic-guided or edge-guided counterparts thanks to the instance-level semantic information provided by the panoptic guidance. However, our semantic-guided and edge-guided model still achieves impressive FID scores compared to the existing methods, showing the flexible and robustness of our approach. Furthermore, Tab. 2 presents the evaluation of the panoptically guided task on Places2-object where our model improves the the existing methods and show generalization capacity to the general object classes.

To understand the visual effect of our approach, we present visual comparisons of our method with the state-of-the-art methods on the guided tasks. Specifically, Figs. 4 and 5 present the qualitative comparison of our method on the panoptically guided inpainting task on Places2-person and Places-object against the retrained SESAME (Ntavelis et al., 2020), LaMa (Suvorov et al., 2021), CoModGAN (Zhao et al., 2021) and CM-GAN (Zheng et al., 2022). Moreover, visual comparison on semantic-guided and the edge-guided task (Fig. 8) demonstrate the clear advantage of our method on generating realistic object instances in comparison to the most recent works including Ntavelis et al. (2020); Zeng et al. (2021).

3.3 Ablation Study

We perform a set of ablation experiments to show the importance of each component of our model. Quantitative results are shown in Tab. 3 and the visual comparisons are shown in Fig. 7. Below we describe the ablation experiment in the following aspects:

Perceptual Loss We start with the conditional CM-GAN as the baseline for the panoptically guided inpainting task. We find that the model trained with only the StyleGAN discriminator loss (abbreviated as adv.) suffers from slow convergence and sometimes produces color blobs, while the perceptual loss model (perc.) improves the performance and reduce the FID to 2.8470 and 2.2746, respectively, on the two masks. This finding is consistent with the observations of CM-GAN (Zheng et al., 2022) and LaMa (Suvorov et al., 2021).

Figure 7: The visual effect of various ablation models. Our full panoptically guided model achieves the best visual results with realistic semantic and generated objects. Best viewed by zoom-in on screen.

Figure 8: Qualitative comparisons on semantic-guided inpainting (left) edge-guided inpainting (right).

Semantic Discriminator Based on the perceptual loss model, we insert the semantic discriminator (sem. D) at the image-level only for model training. As shown in Tab. 3, the semantic discriminator improves the FID, which is coherent to the improvement on object generation such as face. However, the semantic discriminator model still suffers from object distortion.

Object-level Discriminators We further add the object-level StyleGAN discriminator and semantic discriminator on top of the image-level discriminators. The visual results show that object-level discriminators significantly improve the quality of the generated object and the FID scores.

Semantic Label Map Guidance We compare the panoptically guided task with the semantics-guided task using our trained models. The panoptically guided task achieves better FID scores and generates better object boundaries when instances overlap (e.g. the two overlapping persons in the first row of Fig. 7). Our semantic-guided model produces high-quality results for disjoint instances.

3.4 Panoptically Guided Model Improves Image Inpainting

The existing image inpainting methods (Zhao et al., 2021; Zheng et al., 2022) often struggle to generate realistic objects inside the hole. In Tab. 4, we show that our trained panoptically guided model can significantly improve the traditional image inpainting that requires no guidance. Specifically, we use an off-the-shelf inpainting method CM-GAN (Zheng et al., 2022) to generate the initial completed image and apply PanopticFCN (Li et al., 2021) to generate a panoptic layout of the output. Finally, the panoptic layout prediction and the masked image are passed to our model for panoptically guided inpainting. As visualized in Fig. 6, such a pipeline significantly improves the object generation capacity of the existing approaches with decreased FID.

Table 4: Quantitative evaluation on the standard inpainting task that does not require guidance input. Results are evaluated on Places2-person.

	CoModGAN masks			Object masks		
Methods	FID↓	U-IDS (%)↑	P-IDS (%)↑	FID↓	U-IDS (%)↑	P-IDS (%)↑
LaMa (Suvorov et al., 2021)	32.9607	8.08	0.66	13.5481	15.91	2.46
CoModGAN (Zhao et al., 2021)	12.0215	19.02	5.46	10.4286	20.59	5.38
CM-GAN (Zheng et al., 2022)	11.6727	19.56	5.53	9.0216	23.05	15.18
ours	**4.5402**	**29.65**	**22.42**	**3.1960**	**33.98**	**28.24**

4 RELATED WORK

4.1 IMAGE INPAINTING AND GUIDED IMAGE INPAINTING

Early image inpainting methods leverage patch-based copy-pasting (Efros & Freeman, 2001; Kwatra et al., 2005; Barnes et al., 2009; Cho et al., 2008; Darabi et al., 2012) or color propagation (Ballester et al., 2001; Chan & Shen, 2001; Shen & Chan, 2002; Criminisi et al., 2004) to fill in the target hole. Those methods can produce high-quality textures while completing simple shapes but cannot hallucinate new semantic structures. Recently, deep generative models have shown promising results on image inpainting. Inspired by Pathak et al. (2016) that trains an encoder-decoder network to completes the missing region of an image, numerous approaches have been proposed to improve the learning-based hole filling. The proposed mechanisms including multi-stage networks (Yu et al., 2018; Liu et al., 2020; Yi et al., 2020; Wan et al., 2021; Zeng et al., 2020; Nazeri et al., 2019; Xiong et al., 2019; Yang et al., 2020; Song et al., 2018), attention mechanism (Yu et al., 2018), dilated convolution (Iizuka et al., 2017; Yu et al., 2018), Fourier convolution (Chi et al., 2020; Suvorov et al., 2021) expands the receptive field of the generative models, allowing better contextual feature propagating while enabling better inpainting quality. Recently, modulation-based methods (Zhao et al., 2021; Zheng et al., 2022) further improve the global context modeling for image inpainting. In addition, probabilistic diffusion models (Ho et al., 2020; Saharia et al., 2021; Lugmayr et al., 2022) and vision transformers (Wan et al., 2021; Zheng et al., 2021a;b) have also shown promising results on image inpainting. Guided inpainting leverages additional structural guidance to improve image completion of complex semantic structures while providing tools for users to manipulate the inpainting outcome, In the literature, edge (Yu et al., 2019; Nazeri et al., 2019; Xiong et al., 2019), color (Portenier et al., 2018; Jo & Park, 2019) or gradient maps (Yang et al., 2020) are used to guide the image completion process. To provide semantic-level control for inpainting, semantic segmentation (Song et al., 2018; Hong et al., 2018; Ntavelis et al., 2020) is also used a guidance map for more controlled inpainting.

4.2 DISCRIMINATORS FOR GANS

The initial Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) leverages a multi-layer perceptron (MLP) to predict the realism of the generated images. Since then, there have been rapid advance to achieve photo-realistic image synthesis. Specifically, patch discriminator (Isola et al., 2017; Li & Wand, 2016; Yu et al., 2019) is proposed to predict the realism of the generated patches in local regions. Later, the StyleGAN-based discriminators (Karras et al., 2019; 2020b) leverages strided convolution combined with a fully connected layer to generate realistic images. Motivated by contrasitive learning (Chen et al., 2020), several works (Kang & Park, 2020; Zhang et al., 2021; Jeong & Shin, 2021; Yu et al., 2021) design discriminators to predict the pairwise relations between different modalities or the real and fake samples. Likewise, the patch co-occurrence discriminator (Park et al., 2020) predicts the similarity between the output patches and the reference style image. To enhance the recognition capacity of a discriminator, recent works (Sauer et al., 2021b; Kumari et al., 2021) leverages pretrained visual features to enhance the semantic understanding of a discriminator for unconditional generation.

5 CONCLUSION

Aiming at inpainting realistic objects, we investigate a panoptically guided image inpainting task that leverages panoptic segmentation to assist image inpainting. Our approach is based on a new semantic discriminator design that leverages pretrained visual features to improve the semantic consistency of the generated contents. We further propose object-level discriminators to enhance the realism of the generated content. Our approach shows significant improvements on the generated object and leads to new state-of-the-art performance on various tasks, including panoptically guided inpainting, semantic-guided inpainting, edge-guided inpainting, and standard inpainting without guidance.

REFERENCES

Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8):1200–1211, 2001.

Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020.

Tony F Chan and Jianhong Shen. Nontexture inpainting by curvature-driven diffusions. *Journal of visual communication and image representation*, 12(4):436–449, 2001.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33, 2020.

Taeg Sang Cho, Moshe Butman, Shai Avidan, and William T Freeman. The patch transform and its applications to image editing. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, 2008.

Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212, 2004.

Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. Image melding: Combining inconsistent images using patch-based synthesis. *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012.

Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 341–346. ACM, 2001.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Seunghoon Hong, Xinchen Yan, Thomas S Huang, and Honglak Lee. Learning hierarchical semantic image manipulation through structured representations. *Advances in Neural Information Processing Systems*, 31, 2018.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Jongheon Jeong and Jinwoo Shin. Training gans with stronger augmentations via contrastive discriminator. *arXiv preprint arXiv:2103.09742*, 2021.

Youngjoo Jo and Jongyoul Park. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1745–1753, 2019.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Minguk Kang and Jaesik Park. Contragan: Contrastive learning for conditional image generation. *Advances in Neural Information Processing Systems*, 33:21357–21369, 2020.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020b.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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*, pp. 9404–9413, 2019.

Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. *arXiv preprint arXiv:2112.09130*, 2021.

Vivek Kwatra, Irfan Essa, Aaron Bobick, and Nipun Kwatra. Texture optimization for example-based synthesis. In *ACM SIGGRAPH 2005 Papers*, pp. 795–802. 2005.

Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European conference on computer vision*, pp. 702–716. Springer, 2016.

Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 214–223, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Hongyu Liu, Jiang Bin, Yibing Song, Huang Wei, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Proceedings of the European Conference on Computer Vision*, 2020.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. *arXiv preprint arXiv:2201.09865*, 2022.

Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.

Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. Sesame: Semantic editing of scenes by adding, manipulating or erasing objects. In *European Conference on Computer Vision*, pp. 394–411. Springer, 2020.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *arXiv preprint arXiv:2007.00653*, 2020.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Tiziano Portenier, Qiyang Hu, Attila Szabo, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. Faceshop: Deep sketch-based face image editing. *arXiv preprint arXiv:1804.08972*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A Lee, Jonathan Ho, Tim Salimans, David J Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *arXiv preprint arXiv:2111.05826*, 2021.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34, 2021a.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34, 2021b.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. *arXiv preprint arXiv:2202.00273*, 2022.

Jianhong Shen and Tony F Chan. Mathematical models for local nontexture inpaintings. *SIAM Journal on Applied Mathematics*, 62(3):1019–1043, 2002.

Yuhang Song, Chao Yang, Yeji Shen, Peng Wang, Qin Huang, and C-C Jay Kuo. Spg-net: Segmentation prediction and guidance network for image inpainting. *arXiv preprint arXiv:1805.03356*, 2018.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021.

Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. *arXiv preprint arXiv:2103.14031*, 2021.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695–8704, 2020.

Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5840–5848, 2019.

Jie Yang, Zhiquan Qi, and Yong Shi. Learning to incorporate structure knowledge for image inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12605–12612, 2020.

Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7508–7517, 2020.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5505–5514, 2018.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4471–4480, 2019.

Ning Yu, Guilin Liu, Aysegul Dundar, Andrew Tao, Bryan Catanzaro, Larry S Davis, and Mario Fritz. Dual contrastive loss and attention for gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6731–6742, 2021.

Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. *arXiv preprint arXiv:2005.11742*, 2020.

Yu Zeng, Zhe Lin, and Vishal M Patel. Sketchedit: Mask-free local image manipulation with partial sketches. *arXiv preprint arXiv:2111.15078*, 2021.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 833–842, 2021.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021.

Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic free-from image completion. *International Journal of Computer Vision*, pp. 1–20, 2021a.

Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai, and Dinh Phung. Bridging global context interactions for high-fidelity image completion, 2021b.

Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connelly Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsi, and Jiebo Luo. Cm-gan: Image inpainting with cascaded modulation gan and object-aware training. In *Proceedings of the European conference on computer vision (ECCV)*, 2022.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

A Dataset Details

We construct two large-scale datasets named Places2-person and Places2-object from the Places2 dataset (Zhou et al., 2017) for evaluating the object inpainting task in various settings. Specifically, Places2-person is the subsets of Places2 dataset that contains at least one person instances and it includes 1.28M images for training and 62748 images for testing. The Places2-object is the Places2 subset that contains at least one object instances and it includes 2.75M images for training and 127567 images for testing. We leverage the pretrained PanopticFCN model (Li et al., 2021) trained on COCO-Stuff (Lin et al., 2014) to generate the panoptic segmentation annotation on both datasets. We use the inpainting mask of CoModGAN (Zhao et al., 2021) to generate the mask for training and the mask of CoModGAN and the object mask of (Zeng et al., 2020) for evaluation. During training, we augment input images by random cropping and random horizontal flipping.

B ARCHITECTURE AND TRAINING DETAILS

We leverage the pretrained CLIP model (Radford et al., 2021) as the backbone to extract feature for the semantic discriminators. During training, we randomly sample one object instance that overlaps the mask, then crop and resize object instances to resolution 224×224 for training the object-level StyleGAN and semantic discriminators. We use the panoptic segmentation annotation to generate the bounding box of objects. Specifically, for each instances, we take the minimal bounding box corresponding to the instance as the bounding box for cropping. To reduce aliasing and ringing artifacts of generated objects, we following the data augmentation practice of StyleGAN-ada (Karras et al., 2020a) and upscale the global image by a factor of 2 and then apply a band-limited low-pass filter before the cropping and resizing operations. To generate the cropped patch of discrete condition such as semantic map and edge map, we follow the same upscale-cropping-resizing routine that is used for generating object patches. However, we apply nearest sampling operation instead of filtering to produce discrete label map and edge map. Our codebase is implemented based on the Pytorch implementation of StyleGAN2-ada (Karras et al., 2020a). Our model is trained with the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001 and a batch size of 32. The model training takes 3.5 days to converge on a server with 8 A100 GPUs.

C MORE VISUAL COMPARISONS

Fig. 9 presents the visual comparison between the standard inpainting results generated by the state of the art inpainting models, i.e., CoModGAN and CMGAN against the results generated by our panoptically guided inpainting model. Fig. 10 and Fig. 11 present the visual comparisons of various panoptically guided inpainting models. Fig. 12, Fig. 13 and Fig. 14 present the additional visual results of our approach on semantic-guided, edge-guided and standard inpainting tasks, respectively.

D ANALYSIS OF THE LPIPS SCORES (ZHANG ET AL., 2018)

We provide analysis on the LPIPS scores (Zhang et al., 2018) on the images respectively generated by LaMa* (Suvorov et al., 2021), CoModGAN* (Zhao et al., 2021), CM-GAN* (Zheng et al., 2022), and our approach, in Fig. 15. We found that LPIPS is not a good metric for indicating the object-level realism as LPIPS tends to prefer faded out structures and give higher distance prediction to image completion results with better object-level realism such as face and body. Therefore, we do not evaluate the LPIPS metric in our further experiments.

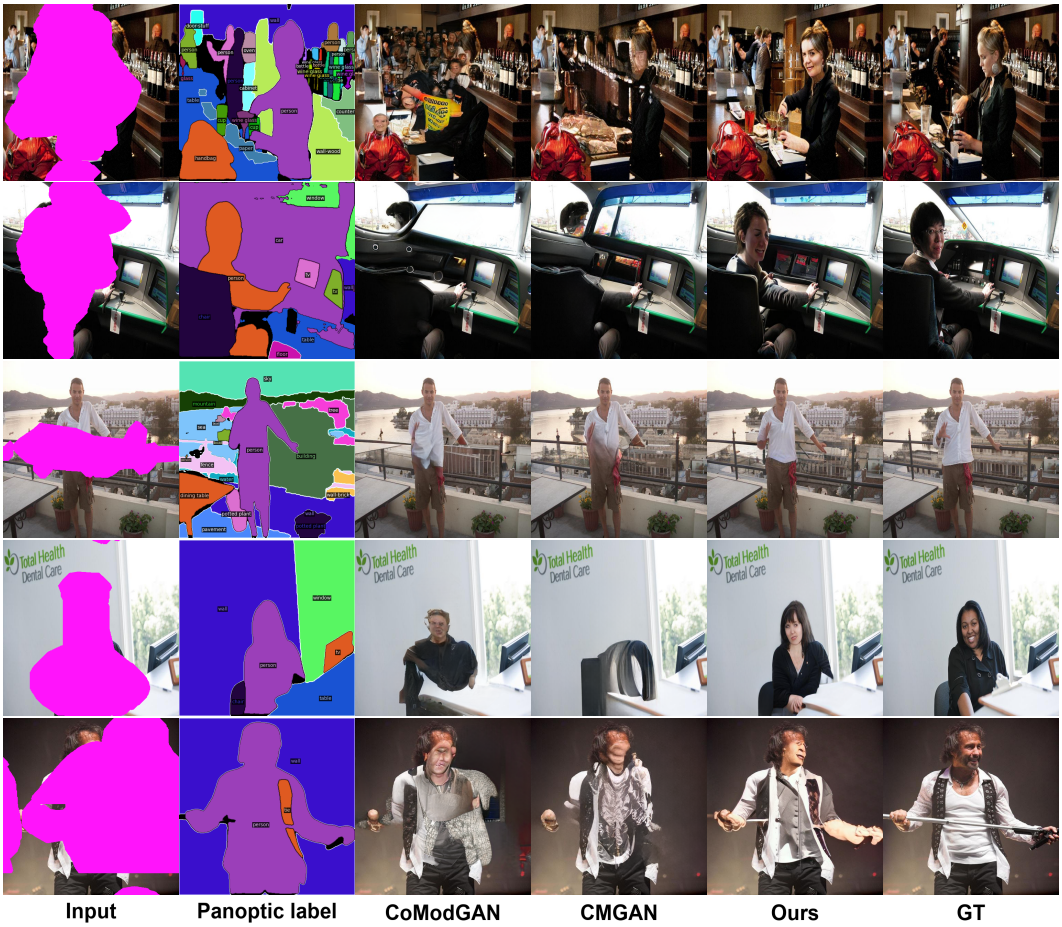

Figure 9: Visual comparisons on between the standard inpainting task, i.e. *CoModGAN, CMGAN* against result generated by our proposed panoptically guided task. Compared to the standard inpainting method, panoptically guided inpainting provides more control over the generated contents and our approach can generate high-quality and photo-realist completion results. Best viewed by zoom-in on screen.

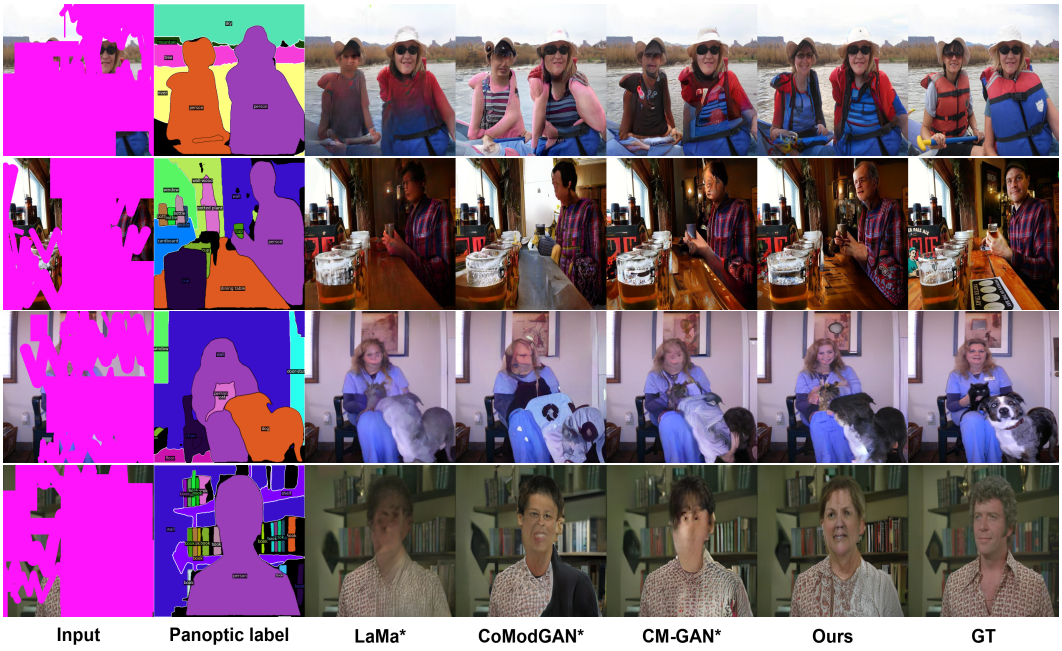

Figure 10: Visual comparisons on the panoptically guided inpainting task on Places2-person. We compare our model against LaMa* (Suvorov et al., 2021), CoModGAN* (Zhao et al., 2021), CM-GAN* (Zheng et al., 2022) whereas $*$ denotes models re-trained with the additional panoptic segmentation condition for panoptically guided inpainting. Best viewed by zoom-in on screen.

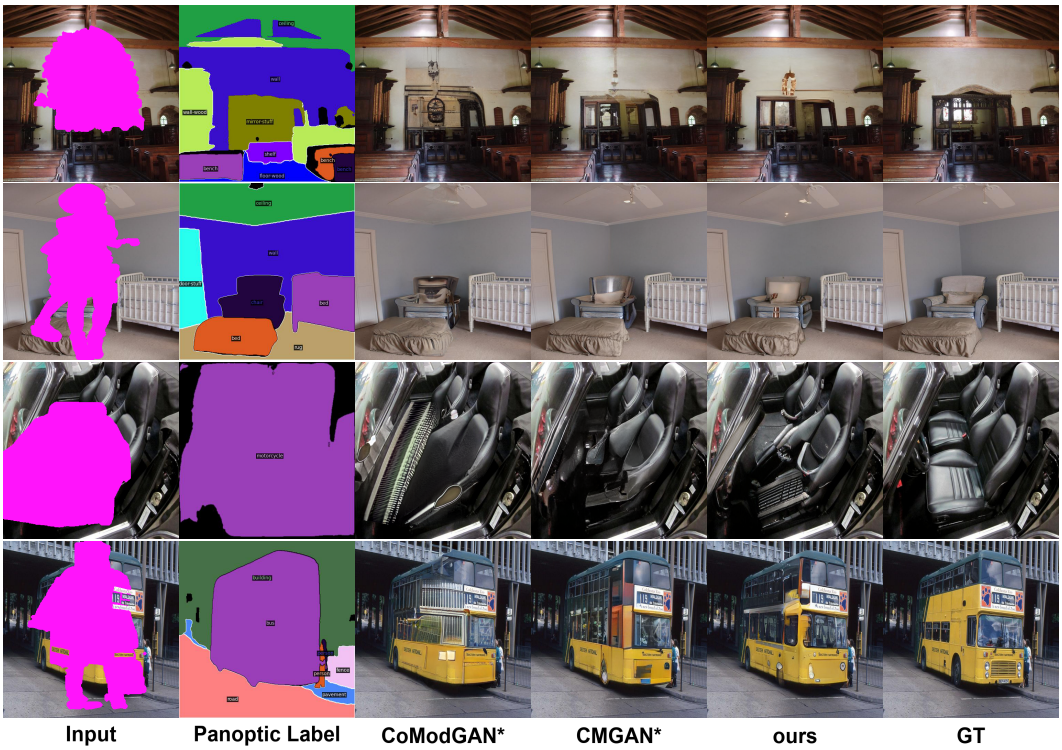

Figure 11: Qualitative comparisons on the panoptically guided inpainting task on Places2-object. We compare our model against retrained CoModGAN$*$ (Zhao et al., 2021), CM-GAN$*$ (Zheng et al., 2022) whereas the $*$ symbol denotes models re-trained with the additional panoptic segmentation condition for panoptically guided inpainting. Best viewed by zoom-in on screen.

Figure 12: Qualitative comparisons on semantic-guided inpainting. Best viewed by zoom-in on screen.

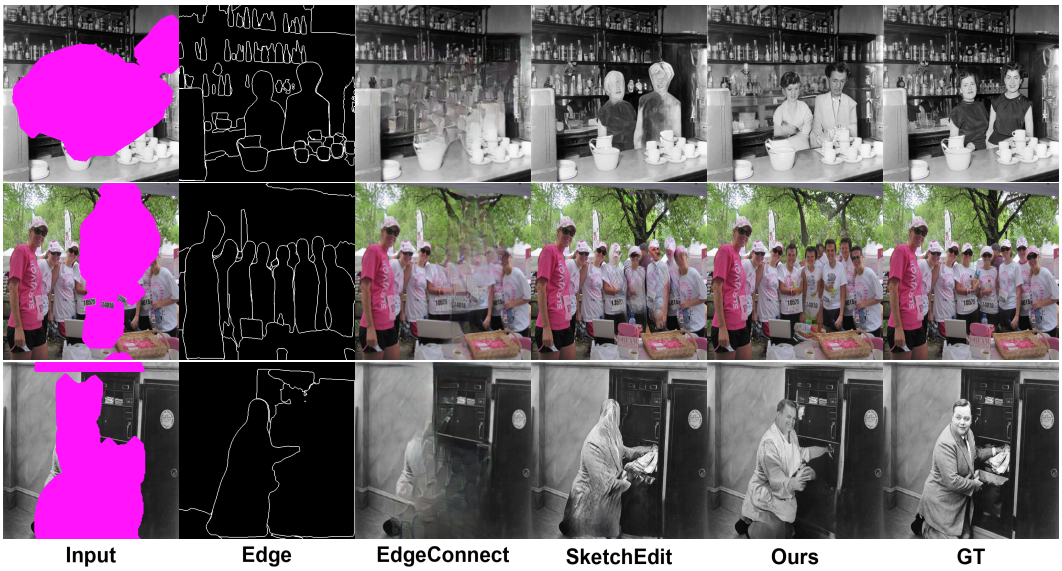

Figure 13: Qualitative comparisons on edge-guided inpainting. Best viewed by zoom-in on screen.

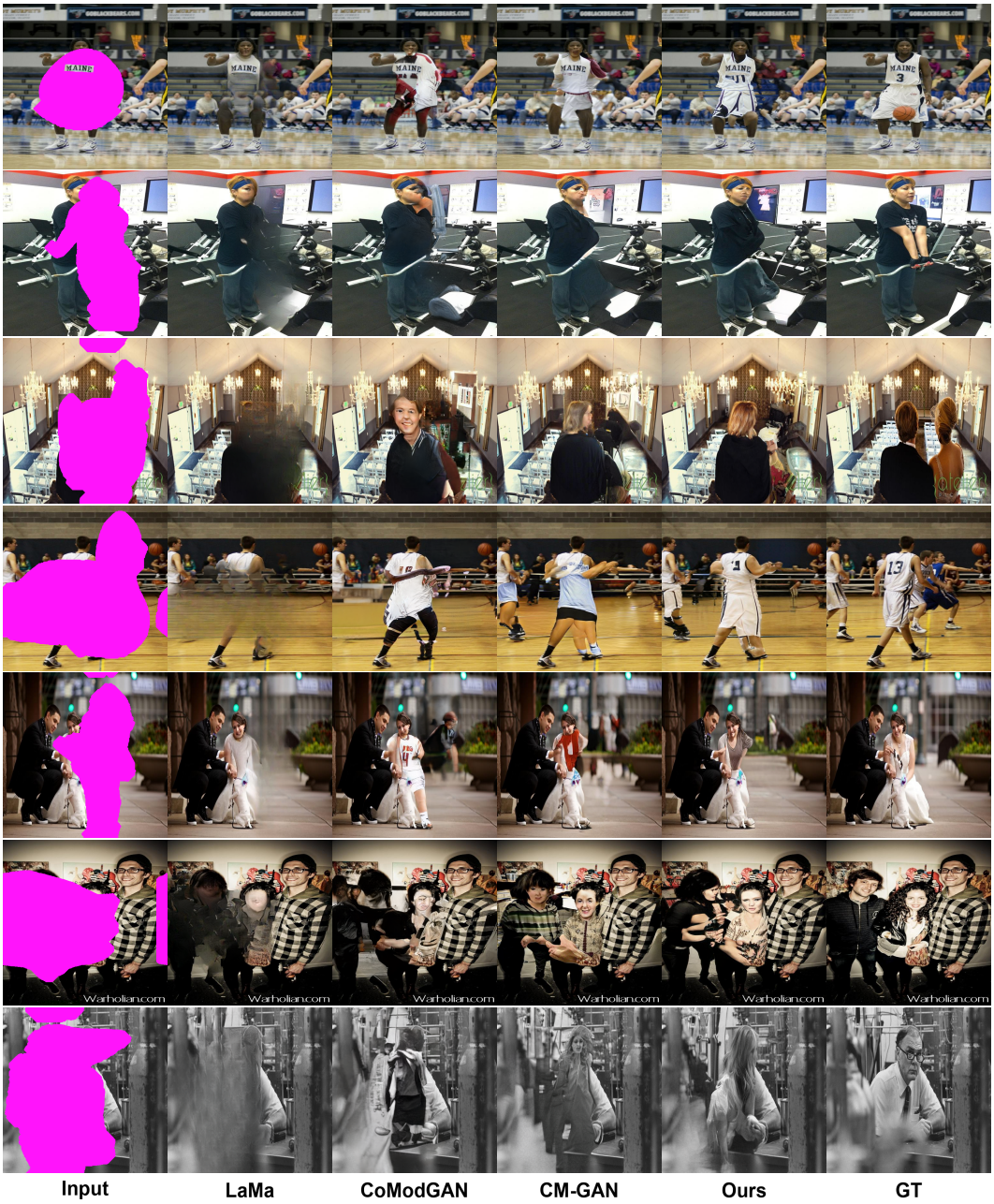

Figure 14: Qualitative comparisons on the standard inpainting task. Compared to the existing methods, our method can generate high-quality and photo-realist object instances.

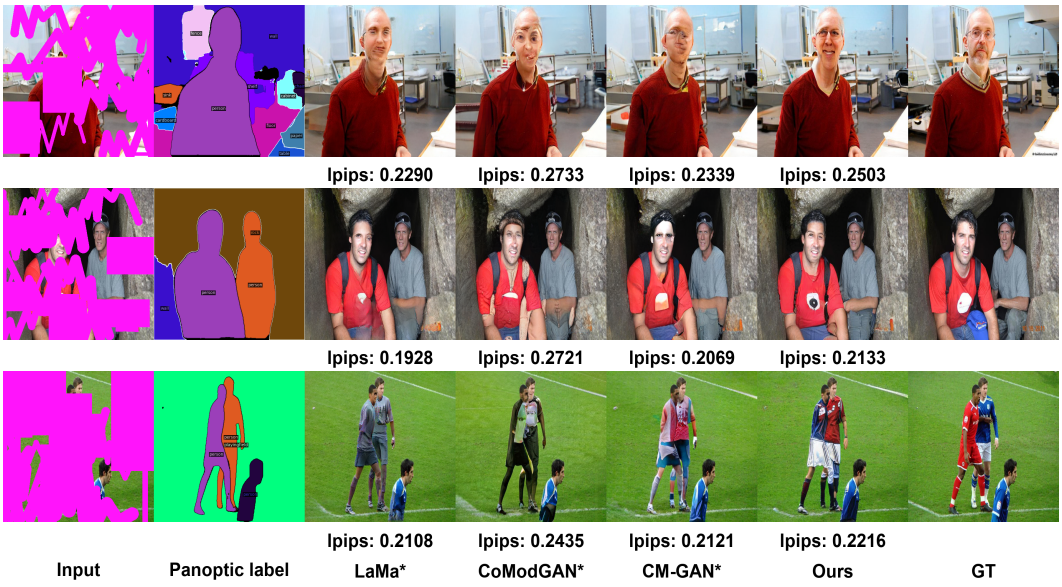

Figure 15: The Learned Perceptual Image Patch Similarity (LPIPS) metric (Zhang et al., 2018) on individual images generated by LaMa* (Suvorov et al., 2021), CoModGAN* (Zhao et al., 2021), CM-GAN* (Zheng et al., 2022), and our approach. We found LPIPS favors averaged and overly-smooth outputs with faded structure details rather than results with realistic instances. Best viewed by zoom-in on screen.