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

Abstract. Completing a corrupted image with correct structures and reasonable textures for a mixed scene remains an elusive challenge. Since the missing hole in a mixed scene of a corrupted image often contains various semantic information, conventional two-stage approaches utilizing structural information often lead to the problem of unreliable structural prediction and ambiguous image texture generation. In this paper, we propose a Semantic Guidance and Evaluation Network (SGE-Net) to iteratively update the structural priors and the inpainted image in an interplay framework of semantics extraction and image inpainting. It utilizes semantic segmentation map as guidance in each scale of inpainting, under which location-dependent inferences are re-evaluated, and, accordingly, poorly-inferred regions are refined in subsequent scales. Extensive experiments on real-world images of mixed scenes demonstrated the superiority of our proposed method over state-of-the-art approaches, in terms of clear boundaries and photo-realistic textures.

Keywords: Image Inpainting, Semantic Guidance, Segmentation Confidence Evaluation, Mixed Scene.

1 Introduction

Image inpainting refers to the task of filling the missing area in a scene with synthesized content. Due to its wide applications in photo editing, de-caption, damaged image repairing, error concealment in data transmission, etc., it has drawn great attention in the field of computer vision and graphics [1, 2, 5, 10, 28]. Recent learning-based methods have achieved great success in filling large missing regions with plausible contents of various simple scenes [23, 31, 34, 41, 42, 44, 45, 47]. However, these existing methods still encounter difficulties while completing images of a mixed scene, that composes of multiple objects with different semantics.

⋆ Corresponding author: Jing Xiao

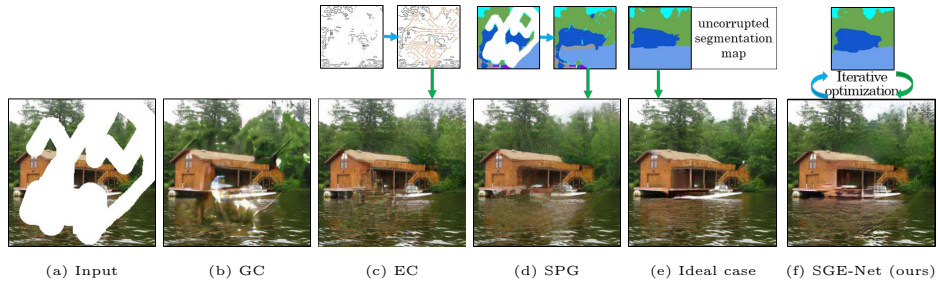| (a) Input | (b) GC | (c) EC | (d) SPG | (e) Ideal case | (f) SGE-Net (ours) |

Fig. 1: Comparison of the inpainting results for a mixed scene: (b) GC [42] without structural information; (c) EC [21] with predicted edges; (d) SPG [27] with less reliable predicted semantic segmentation; (e) semantic-guided inpainting with an uncorrupted segmentation map; and (f) the proposed SGE-Net with iteratively optimized semantic segmentation. [Best viewed in color]

Existing learning-based image inpainting methods typically fill missing regions by inferring the context of corrupted images [12, 23, 35, 41, 42]. However, in a mixed scene, the prior distributions of various semantics are different and various semantic regions also contribute differently to pixels in the missing regions, thus uniformly mapping different semantics onto a single manifold in the context-based methods often leads to unrealistic semantic content as illustrated in Fig. 1(b).

To address this issue, low to mid-level structural information [16, 27, 36, 46] was introduced to assist image inpainting. These methods extract and reconstruct the edges or contours in the first stage and complete an image with the predicted structural information in the second stage. The spatial separation by the structures helps to alleviate the blurry boundary problem. These methods, however, ignore the modeling of semantic content, which may result in ambiguous textures at the semantic boundaries. Moreover, the performance of the two-stage inpainting process highly relies on the reconstructed structures from the first stage, but the unreliability of the edge or contour connections largely increases in a mixed scene (Fig. 1(c)). As revealed in [34] that human beings perceive and reconstruct the structures under the semantic understanding of a corrupted image, it is natural to involve semantic information in the process of image inpainting.

In this paper, we show how semantic segmentation can effectively assist image inpainting of a mixed scene based on two main discoveries: **semantic guidance and segmentation confidence evaluation**. Specifically, a semantic segmentation map carries pixel-wise semantic information, providing the layout of a scene as well as the category, location and shape of each object. It can assist the learning of different texture distributions of various semantic regions. Moreover, the intermediate confidence score derived from the segmentation process can offer a self-evaluation for an inpainted region, under the assumption that ambiguous semantic contents usually cannot lead to solid semantic segmentation results.

To the best of our knowledge, a similar work making use of semantic segmentation information for image inpainting is SPG proposed in [27], which is also

a two-stage process. It extracts and reconstructs a segmentation map, and then utilizes the map to guide image inpainting. Thanks to the helpful semantic information carried in the segmentation map, SPG can effectively improve inpainting performance compared to those methods without a semantic segmentation map. Nevertheless, it is hard to predict reliable semantics about a region when its context information is largely missing, especially in mixed scene. As a result, its performance can be significantly degraded by such unreliable semantic region boundaries and labels predicted by the semantic segmentation. Such performance degradation is evidenced in Fig. 1(d), from which we can observe blurry and incorrect inpainted textures generated by SPG. By contrast, segmentation-guided inpainting can achieve high-quality image completion provided that a reliable segmentation map (i.e., the segmentation map of uncorrupted image) is given as illustrated in Fig. 1(e). Therefore, to make the best use of semantic information carried in the segmentation map for image inpainting, how to predict a reliable semantic segmentation map, even if part of an image is corrupted, is the key.

To address the above problems, we advocate that the interplay between the two tasks, semantic segmentation and image inpainting, can effectively improve the reliability of the semantic segmentation map from a corrupted image, which will in turn improve the performance of inpainting as illustrated in Fig. 1(f). To this end, we propose a novel **S**emantic **G**uidance and **E**valuation **Net**work (**SGE-Net**) that makes use of the interplay between semantic segmentation and image inpainting in a coarse-to-fine manner. Experiments conducted on the datasets containing mixtures of multiple semantic regions demonstrated the effectiveness of our method in completing a corrupted mixed scene with significantly improved semantic contents.

Our contributions are summarized as follows:

1) We show that the interplay between semantic segmentation and image inpainting in a coarse-to-fine manner can effectively improve the performance of image inpainting by simultaneously generating an accurate semantic guidance from merely an input corrupted image.

2) We are the first to propose a self-evaluation mechanism for image inpainting through segmentation confidence scoring to effectively localize the predicted pixels with ambiguous semantic meanings, which enables the inpainting process to update both contexts and textures progressively.

3) Our model outperforms the state-of-the-art methods, especially on mixed scenes with multiple semantics, in the sense of generating semantically realistic contexts and visually pleasing textures.

2   Related Work

2.1   Deep Learning-Based Inpainting

Deep learning-based image inpainting approaches [15, 23, 39] are generally based on generative adversarial networks (GANs) [9, 24, 33] to generate the pixels of a missing region. For instance, Pathak *et al.* introduced Context Encoders [23],

which was among the first approaches in this kind. The model was trained to predict the context of a missing region but usually generates blurry results. Based on the Context Encoders model, several methods were proposed to better recover texture details through the use of well-designed loss functions [7, 12, 15], neural patch synthesis [38], residual learning [6, 40], feature patch matching [26, 37, 41, 43], content and style disentanglement [8, 32, 34], and others [20, 29, 31]. Semantic attention was further proposed to refine the textures in [18]. However, most of the above methods were designed for dealing with rectangular holes, but cannot effectively handle large irregular holes. To fill irregular holes, Liu *et al.* [17] proposed a partial convolutional layer, which calculates a new feature map and updates the mask at each layer. Later, Yu *et al.* [42] proposed a gated convolutional layer based on the models in [41] for irregular image inpainting. While these methods work reasonably well for one category of objects or background, they can easily fail if the missing region contains multiple categories of scenes.

2.2   Structural Information-Guided Inpainting

Recently, structural information was introduced in learning-based framework to assist the image inpainting process. These methods are mostly based on two-stage networks, where missing structures are reconstructed in the first stage and then used to guide the texture generation in the second stage. Edge maps were first introduced by Liao *et al.* [16] as a structural guide to the inpainting network. This idea is further improved by Nazeri *et al.* [21] and Li *et al.* [14] in terms of better edge prediction. Similar to edge information, object contours were used by Xiong *et al.* [36] to separately reconstruct the foreground and background areas. Ren *et al.* [25] proposed using smoothed images to carry additional image information other than edges as prior information. Considering semantic information for the modeling of texture distributions, SPG proposed in [27] predicts the semantic segmentation map of a missing region as a structural guide. The above-mentioned methods show that the structure priors effectively help improve the quality of the final completed image. However, how to reconstruct correct structures remains challenging, especially when the missing region becomes complex.

3   Approach

As illustrated in Figs. 1(d)-(f), the success of semantic segmentation-guided inpainting depends on a reliable segmentation map, which is hard to obtain from an image with a corrupted mixed scene. To address this issue, we propose a novel method to progressively predict a reliable segmentation map from a corrupted image through the interplay between semantic segmentation and image inpainting in a coarse-to-fine manner. To verify how semantic information boosts image inpainting, two networks are proposed. As a baseline, the first one uses only semantic guidance on image inpainting. Moreover, the semantic evaluation is added in the second network as an advanced strategy.
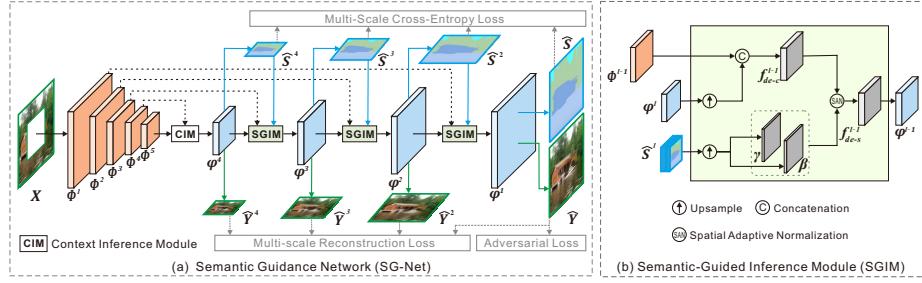
Fig. 2: Proposed Baseline: Semantic Guidance Network (SG-Net). It iteratively updates the contextual features in a coarse-to-fine manner. SGIM updates the predicted context features based on the segmentation map at the next scale.

We first introduce some notations used throughout this paper. Given a corrupted image X with a binary mask M (0 for holes), and the corresponding ground-truth image Y, the inpainting task is to generate an inpainted image $\hat{Y}$ from X and M. Given a basic encoder-decoder architecture of L layers, we denote the feature maps from deep to shallow in the encoder as ϕ^L, ϕ^{L-1}, ..., ϕ^l, ..., ϕ^1, and in the decoder as φ^L, φ^{L-1}..., φ^l, ..., φ^1.

3.1 Semantic Guidance Network (SG-Net)

The SG-Net architecture is shown in Fig. 2(a). The encoder is used to extract the contextual features of a corrupted image. The decoder then updates the contextual features to predict the semantic segmentation maps and inpainted images simultaneously in a multi-scale manner. Based on this structure, semantic guidance takes effect in two aspects. First, the semantic supervisions are added to guide the learning of contextual features at different scales of the decoder. Second, the predicted segmentation maps are involved in the inference modules to guide the update of the contextual features at the next scale. Being different from the two-stage process[16, 21, 27], the supervision of semantic segmentation on the contextual features enables them to carry the semantic information, that helps the decoder learn better texture models for different semantics.

The corrupted image is initially completed in the feature level through a Context Inference Module (CIM). After that, the image inpainting and semantic segmentation interplay with each other and are progressively updated across scales. Two branches are extended from the contextual features at each scale of the decoder to generate multi-scale completed images $\hat{Y}^{L-1}$, ..., $\hat{Y}^l$, ..., $\hat{Y}^1$ and their semantic segmentation maps $\hat{S}^{L-1}$, ..., $\hat{S}^l$, ..., $\hat{S}^1$.

$$\hat{Y}^l = h(\varphi^l), \quad \hat{S}^l = g(\varphi^l), \tag{1}$$

where $h(\cdot)$ and $g(\cdot)$ denote the inpainting branch and segmentation branch, respectively.
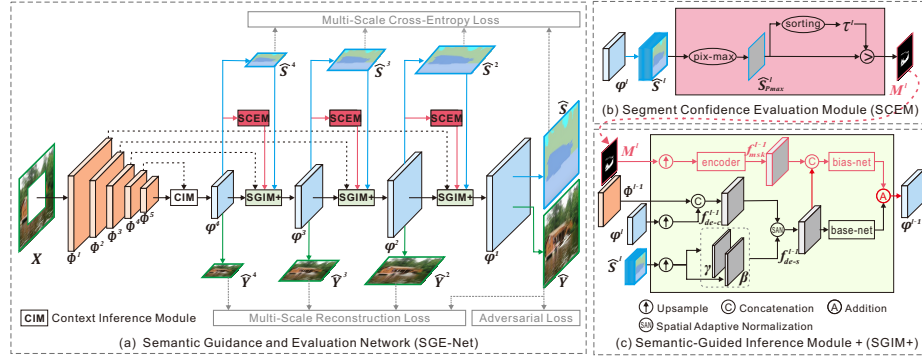
Fig. 3: Proposed Semantic Guidance and Evaluation Network (SGE-Net). It iteratively evaluates and updates the contextual features through the SCEM and SGIM+ modules in a coarse-to-fine manner, where SCEM identifies the pixels where the context needs to be corrected, while SGIM+ updates the predicted context features representing the incorrect pixels located by SCEM.

Semantic-Guided Inference Module (SGIM) SGIM is designed to make an inference and update the contextual features at the next scale φ^{l-1}. As shown in Fig. 2(b), SGIM takes three types of inputs: two of them are the current contextual features φ^l and the skip features of the next scale ϕ^{l-1} from the encoder. The third input is the segmentation map $\hat{S}^l$, which is used to formalize the textures under the assumption that those regions of the same semantic class should have similar textures. The inference process can be formulated as follows:

$$\varphi^{l-1} = infer(\varphi^l, \phi^{l-1}, \hat{S}^l), \tag{2}$$

where $infer(\cdot)$ is the process of updating the contextual features in SGIM.

To update the contextual features based on segmentation map $\hat{S}^l$, we follow the image generation approach in [22], which adopts spatial adaptive normalization to propagate semantic information into the predicted images for achieving effective semantic guidance. The contextual features f_{de-s}^{l-1} are updated as follows:

$$f_{de-s}^{l-1} = \gamma \odot \frac{f_{de-c}^{l-1} - \mu}{\sigma} + \beta, \tag{3}$$

where (γ, β) is a pair of affine transformation parameters modeled from segmentation map $\hat{S}^l$, μ and σ are the mean and standard deviation of each channel in the concatenated feature vector f_{de-c}^{l-1} generated from ϕ^l and φ^{l-1}. $\odot$ denotes element-wise multiplication.

3.2   Semantic Guidance and Evaluation Network (SGE-Net)

To deeply exploit how segmentation confidence evaluation can help correct the wrongly predicted pixels, we add the Segmentation Confidence Evaluation Module (SCEM) on each decoder layer of SG-Net. The evaluation is performed under

the assumption that predicted ambiguous semantic content would result in low confidence scores during the semantic segmentation process. Therefore, we introduce the **segmentation confidence scoring** after each decoding layer to self-evaluate the predicted region. The reliability mask is then feed to the next scale, which can be used to identify those to-be-updated contextual features that contribute to the unreliable area. This module enables the proposed method to correct the mistakes in those regions completed at the previous coarser scale. Fig. 3(a) illustrates the detailed architecture of SGE-Net.

Segmentation Confidence Evaluation Module (SCEM) The output of the semantic segmentation branch is a k-channel probability map. The confidence score at every channel of a pixel in the map signifies how the pixel looks like a specific class. Based on the scores, we assume that an inpainted pixel is unreliable if it has low scores for all semantic classes.

The framework of SCEM is depicted in Fig. 3(b). Taking the segmentation probability map at a certain scale $\hat{S}^l$, we generate a reliability mask M^l to locate those pixels which might have unreal semantic meaning. We first generate a max-possibility map $\hat{S}^l_{P_{max}}$ by assigning each pixel with the highest confidence score of k channels in $\hat{S}^l$. Then, the mask value of pixel (x, y) in the reliability mask is decided by judging whether the max-confidential score at each pixel location exceeds a threshold τ^l.

$$M^l(x,y) = \begin{cases} 1, & \hat{S}^l_{P_{max}} > \tau^l \\ 0, & \text{otherwise} \end{cases}, \tag{4}$$

where τ^l is decided by the percentile of the sorted confidence value.

Enhanced SGIM (SGIM+) In order to correct the pixels marked as unreliable from the SCEM, SGIM+ takes the reliability mask M^l as the fourth input to update the current context features (as shown in Fig. 3(c)). The formulation of the inference process can be updated as follows:

$$\varphi^{l-1} = infer(\varphi^l, \phi^{l-1}, \hat{S}^l, M^l). \tag{5}$$

To enable the dynamic corrections of semantics, we introduced a bias-net F^l_{bi} in correspondence to the original network branch between the feature f^{l-1}_{de-s} and φ^{l-1} in the previous version of SGIM, which we call base-net F^l_{ba}. The new reliability mask M^l is fed into the bias-net to learn residuals to rectify the basic contextual features from the base-net. The new contextual features at the next scale can be formulated as

$$\varphi^{l-1} = F^l_{ba}(f^{l-1}_{de-s}) + F^l_{bi}(f^{l-1}_{de-s} \oplus F(M^l)), \tag{6}$$

where $\oplus$ denotes the concatenation operation and F represents the convolutions to translate the reliability mask into feature maps.

3.3   Training Loss Function

The loss functions comprise loss terms for both image inpainting and semantic segmentation. For image inpainting, we adopt the multi-scale reconstruction loss

to refine a completed image and the adversarial loss to generate visually realistic textures. For semantic segmentation, we adopt the multi-scale cross-entropy loss to restrain the distance between the predicted and target class distributions of pixels at all scales.

Multi-scale Reconstruction Loss We use the $\mathcal{L}_1$ loss to encourage per-pixel reconstruction accuracy, and the perceptual loss [13] to encourage higher-level feature similarity.

$$\mathcal{L}_{re}^l(X, \hat{Y}^l) = \parallel X - up(\hat{Y}^l) \parallel_1 + \lambda_p \sum_{n=1}^{N} \parallel \Psi_n(X) - \Psi_n(up(\hat{Y}^l)) \parallel_1, \qquad (7)$$

where Ψ_n is the activation map of the n-th layer, $up(\cdot)$ is the operation to up-sample $\hat{Y}^l$ to the same size as X, λ_p is a trade-off coefficient. We use layered features $relu2\_2$, $relu3\_3$, and $relu4\_3$ in VGG-16 pre-trained on ImageNet to calculate those loss functions.

Adversarial Loss We use a multi-scale PatchGAN [34] to classify the global and local patches of an image at multiple scales. The multi-scale patch adversarial loss is defined as:

$$\mathcal{L}_{ad}(X, \hat{Y}) = \sum_{k=1,2,3} (E_{p_X^k \sim X^k}[\log D(p_X^k)] + E_{p_{\hat{Y}}^k \sim \hat{Y}^k}[1 - \log D(p_{\hat{Y}}^k)]), \qquad (8)$$

where $D(\cdot)$ is the discriminator, $p_{\hat{Y}}^k$ and $p_{\hat{Y}}^k$ are the patches in the k-th scaled versions of X and $\hat{Y}$.

Multi-scale Cross-Entropy Loss This loss is used to penalize the deviation of $\hat{S}^l$ at each position at every scale.

$$\mathcal{L}_{se}^l(S, \hat{S}^l) = -\sum_{i \in S} S_i \log(up(\hat{S}^l)). \qquad (9)$$

where i indicates each pixel in segmentation map S.

Final Training Loss The overall training loss of our network is defined as the weighted sum of the multi-scale reconstruction loss, adversarial loss, and multi-scale cross-entropy loss.

$$\mathcal{L}_{Final} = \sum_{l=0}^{4} \mathcal{L}_{re}^l(X, \hat{Y}^l) + \lambda_\alpha \mathcal{L}_{ad}(X, \hat{Y}) + \sum_{l=0}^{4} \lambda_s \mathcal{L}_{se}^l(S, \hat{S}^l), \qquad (10)$$

where λ_α and λ_s are the weights for the adversarial loss and the multi-scale cross-entropy loss, respectively.

4 Experiments

4.1 Setting

We evaluate our method on **Outdoor Scenes** [30] and **Cityscapes** [4] both with segmentation annotations. **Outdoor Scenes** contains 9,900 training images and

300 test images belonging to 8 categories. **Cityscapes** contains 5,000 street view images belonging to 20 categories. In order to enlarge the number of training images of this dataset, we use 2,975 images from the training set and 1,525 images from the test set for training, and test on the 500 images from the validation set. We resize each training image to ensure its minimal height/width to be 256 for **Outdoor Scenes** and 512 for **Cityscapes**, and then randomly crop sub-images of size 256×256 as inputs to our model.

We compare our method with the following three representative baselines:

• GC [42]: gated convolution for free-form image inpainting, without any auxiliary structural information.

• EC [21]: two-stage inpainting framework with edges as low-level structural information.

• SPG [27]: two-stage inpainting framework with a semantic segmentation map as high-level structural information.

In our experiments, we fine-tune the GC and EC models, pre-trained on Places2, on our datasets. We also re-implement and train the model of SPG by ourselves since there is no released code or model. We conduct experiments on both settings of centering and irregular holes. The centering holes are (128×128 for **Outdoor Scenes** and 96×96 for **Cityscapes**), and the irregular masks are obtained from [17].

4.2   Image Inpainting Results

In this section, we present the results of our model trained on human-annotated segmentation labels. We also verify our model trained on the segmentation labels predicted by a state-of-the-art segmentation model. The results can be found in Section 4.3.

Qualitative Comparisons The subjective visual comparisons of the proposed SG-Net and SGE-Net with the three baselines (GC, EC, SPG) on **Outdoor Scenes** and **Cityscapes** are presented in Fig. 4. The corrupted area is simulated by sampling a central hole (128×128 for **Outdoor Scenes** and 96×96 for **Cityscapes**), or placing masks with random shapes. As shown in the figure, the baselines usually generate unrealistic shape and textures. The proposed SG-Net generates more realistic textures than the baselines, but still has some flaws at the boundaries since its final result highly depends on the initial inpainting result. The proposed SGE-Net generates better boundaries between semantic regions and more consistent textures than SG-Net and all the baselines, thanks to its evaluation mechanism that can correct the wrongly predicted labels.

Quantitative Comparisons Table 1 shows the numerical results based on three quality metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) and Fréchet Inception Distance (FID) [11]. In general, the proposed SGE-Net achieves significantly better objective scores than the baselines, especially in PSNR and SSIM .

User Study We conduct the user study on 80 images randomly selected from both datasets. In total, 24 subjects with some background of image processing are involved to rank the subjective visual qualities of images completed by four

| Input | GC | EC | SPG | SG-Net | SGE-Net | GT |

Fig. 4: Subjective quality comparison of inpainting results on image samples from **Outdoor Scenes** and **Cityscapes**. **GT stands for Ground-Truth**.

Table 1: Objective quality comparison of five methods in terms of PSNR, SSIM, and FID on **Outdoor Scenes** and **Cityscapes** (↑: Higher is better; ↓: Lower is better). The two best scores are colored in red and blue, respectively.

| | Outdoor Scenes | | | | | | Cityscapes | | | | | |
| | centering holes | | | irregular holes | | | centering holes | | | irregular holes | | |
	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑	FID↓
GC [42]	19.06	0.73	42.34	19.27	0.81	40.31	21.13	0.74	20.03	17.42	0.72	40.57
EC [21]	19.32	0.76	41.25	19.63	0.83	44.31	21.71	0.76	19.87	17.83	0.73	38.07
SPG [27]	18.04	0.70	45.31	17.85	0.74	50.03	20.14	0.71	23.21	16.01	0.64	44.13
SG-Net (ours)	19.58	0.77	41.49	19.87	0.81	41.74	23.04	0.83	18.98	17.94	0.64	41.24
SGE-Net (ours)	20.53	0.81	40.67	20.02	0.83	42.47	23.41	0.85	18.67	18.03	0.75	39.93

inpainting methods (GC, EC, SPG, and our SGE-Net). As shown in Table 2, the study shows that 67.4% of subjects (1295 out of 1920 comparisons), 70.7% and 73.2% preferred our results over GC, EC, and SPG, respectively. Hence, our method outperforms the other methods.

Since our method mainly focuses on completing mixed scenes with multiple semantics, we also verify its performance on images with different scene complexities. We conduct this analysis by dividing all 80 images into three levels of semantic complexities: 1) low-complexity scenes containing 27 images with 1–2 semantics; 2) moderate-complexity scenes containing 32 images with 3–4 semantics; 3) high-complexity scenes containing 21 images with more than 4 semantics. As shown in Table 2, compared to the baselines, while our method achieves generally better results than the baselines for the simple- to moderate-complexity scenes (about from 60% to 70%), the preference rate increases significantly for

Table 2: Preference percentage matrix (%) of different scene complexities on **Outdoor Scenes** and **Cityscapes** datasets. Overall, low complexity, moderate complexity, and high complexity are colored in black, green, blue and red, respectively.

	GC [42]	EC [21]	SPG [27]	SGE-Net (ours)
GC [42]	–	(46.7)/41.5/47.7/52.0	(58.1)/57.3/59.8/56.7	(32.6)/37.8/34.8/22.4
EC [21]	(53.3)/58.5/52.3/48.0	–	(70.1)/68.7/69.3/73.0	(29.3)/35.0/31.8/18.1
SPG [27]	(41.9)/42.7/40.2/43.3	(29.9)/31.3/30.7/27.0	–	(26.8)/32.7/28.8/16.3
SGE-Net (ours)	(67.4)/62.2/65.2/77.6	(70.7)/65.0/68.2/81.9	(73.2)/67.3/71.2/83.7	–

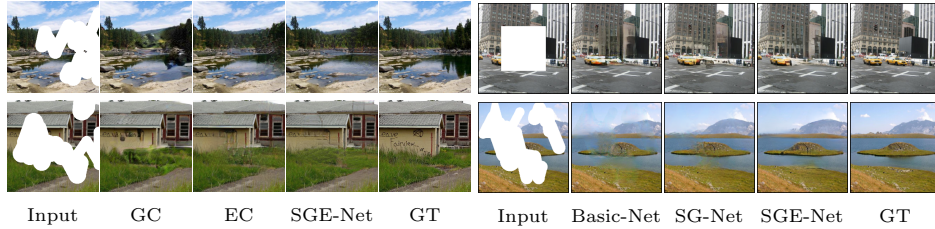| Input | GC | EC | SGE-Net | GT | | Input | Basic-Net | SG-Net | SGE-Net | GT |

Fig. 5: Subjective quality comparison on image samples from **Places2**.

Fig. 6: Subjective visual quality comparisons on the effects of SGIM and SCEM.

the complex scenes (from 77.6% to 83.7%). This verifies that our method is particularly powerful for completing mixed-scene images with multiple semantics, thanks to its mechanism for understanding and updating the semantics.

Additional Results on Places2 For fair comparison, we also test our method on **Places2** [48] to verify that SGE-Net can be applied to images without segmentation annotations. **Places2** was used for evaluation by both GC and EC. It contains images with similar semantic scenes to **Outdoor Scenes**. Therefore, we use our model trained on **Outdoor Scenes** to complete the images with similar scenes in **Places2**. The subjective results in Fig. 5 show that SGE-Net is still able to generate proper semantic structures, owing to the introduction of the semantic segmentation, which provides the prior knowledge about the scenes.

4.3   Ablation Study

Effectiveness of SGIM and SCEM In the proposed networks, the two core components of our method, semantic-guided inference and segmentation confidence evaluation, are implemented by SGIM and SCEM, respectively. In order to investigate their effectiveness, we conduct an ablation study on three variants: a) Basic-Net (without SGIM and SCEM); b) SG-Net (with SGIM but without SCEM); and c) SGE-Net (with both SGIM and SCEM).

The visual and numeric comparisons on **Outdoor Scenes** are shown in Fig. 6 and Table 3. In general, the inpainting performance increases with the added modules. Specifically, the multi-scale semantic-guided interplay framework does a good job for generating detailed contents, and the semantic segmentation map helps learn a more accurate layout of a scene. With SGIM, the spatial adaptive normalization helps generate more realistic textures based on the semantic

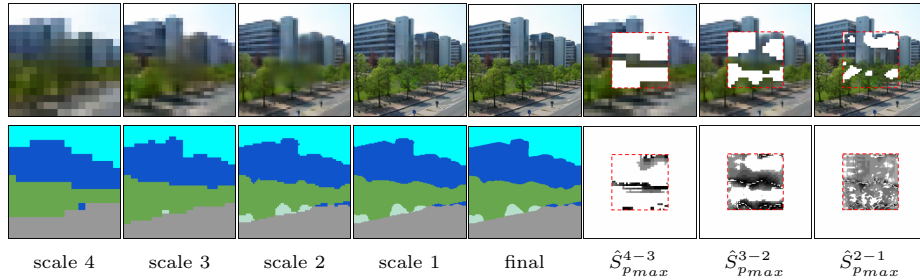| scale 4 | scale 3 | scale 2 | scale 1 | final | $\hat{S}^{4-3}_{pmax}$ | $\hat{S}^{3-2}_{pmax}$ | $\hat{S}^{2-1}_{pmax}$ |

Fig. 7: Illustration of multi-scale progressive refinement with SGE-Net. From left to right of the first 5 columns: the inpainted images (top row) and the segmentation maps (bottom row) from scale 4 to scale 1 and the final result. The last 3 columns show the the reliability maps (top row) and the confidence score maps (bottom row) of the inpainted area across scales (e.g., $\hat{S}^{4-3}_{Pmax}$ shows the confidence score increases from scale 4 to 3).

priors. Moreover, SCEM makes further improvements on completing structures and textures (fourth column in Fig. 6) by coarse-to-fine optimizing the semantic contents across scales.

To further verify the effectiveness of SCEM, we visualize a corrupted image and its segmentation maps derived from all decoding scales. As shown in the first five columns of Fig. 7, the multi-scale progressive-updating mechanism gradually refines the detailed textures as illustrated in the images and the segmentation maps at different scales. The last three columns of the top row show that the region of the unreliable mask gradually shrinks

Table 3: Objective quality comparison on the performances of SGIM and SCEM in terms of three metrics.

	SGIM	SCEM	PSNR↑	SSIM↑	FID↓
Basic-Net	✗	✗	19.14	0.71	43.43
SG-Net	✓	✗	19.58	0.77	41.49
SGE-Net	✓	✓	**20.53**	**0.81**	**40.67**

as well. Correspondingly, the bottom row shows the increase of the confidence scores of segmentation maps from left to right (e.g., $\hat{S}^{4-3}_{Pmax}$ showing the increase of the confidence score from scale 4 to scale 3). The proportion of the white region, which roughly indicates unreliable labels, also decreases significantly from left to right. The result evidently demonstrates the benefits of SCEM in strengthening the semantic correctness of contextual features.

Justification of Segmentation Confidence Scoring During the progressive refinement of image inpainting and semantic segmentation, the semantic evaluation mechanism of SCEM is based on the assumption that the pixel-wise confidence scores from the segmentation possibility map can well reflect the correctness of the inpainted pixel values. Here we attempt to justify this assumption. Some examples from both datasets are shown in Fig. 8. It can be seen that (except for the confidence scores at the region boundaries): a) The low confidence scores (the white area in row 3) usually appear in the mask area, indicating that the scores reasonably well reflects the reliability of inpainted image content; b) the confidence score becomes higher when the scale goes finer, and correspondingly the area of unreliable pixels reduces, meaning that our method can progressively refine the context feature towards correct inpainting.

| scale 4 | scale 3 | scale 2 | scale 1 | final | | scale 4 | scale 3 | scale 2 | scale 1 | final |

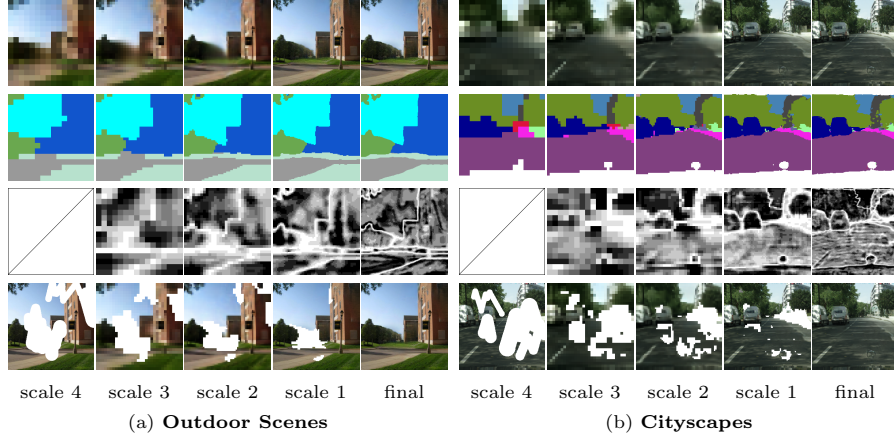(a) **Outdoor Scenes**   (b) **Cityscapes**

Fig. 8: Correspondence between the confidence score value and the reliability of inpainted image content. Row 1: Inpainted image. Row 2: Predicted segmetation map. Row 3: the confidence score map (darker color means higher confidence score, and vice versa). Row 4: unreliable pixel map (white pixels indicate unreliable pixels). Since the map at scale 4 is the same as the input mask, we put the input image for better comparison.
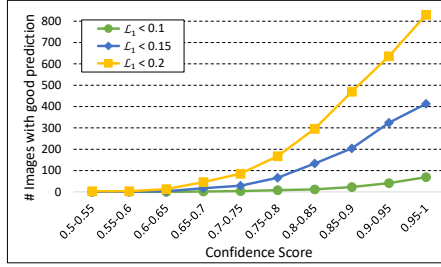

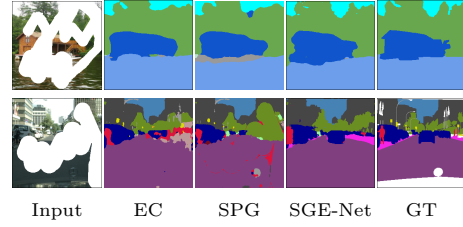Fig. 9: Correlation between the inpainting quality and confidence score.

Input   EC   SPG   SGE-Net   GT

Fig. 10: Visual comparison on semantic segmentation between SGE-Net and the segmentation-after-inpainting solutions. 'EC' and 'SPG' stand for EC+DPN/Deeplab and SPG+DPN/Deeplab, respectively.

We then further verify the effectiveness of the pixel-wise confidence scores by validating the correlation between the confidence scores and the $\mathcal{L}_1$ loss of the completed images with respect to the ground-truth that can be used to measure of the fidelity of inpainted pixels. We randomly select 9,000 images out of all the training and testing images from the two datasets with centering and irregular-hole settings, and calculate the average $\mathcal{L}_1$ loss and the confidence scores of all pixels in the missing region. As demonstrated in Fig. 9, the number of good-fidelity images with $\mathcal{L}_1$ loss increases with the segmentation confidence score (lower $\mathcal{L}_1$ means higher quality of the predicted image), implying the segmentation confidence score well serves the purpose of a metric of evaluating the accuracy of inpainted image content.

Impact of Semantic Segmentation The success of semantics-guided inpainting largely relies on the quality of semantic segmentation map. Here we in-

vestigate the impact of segmentation accuracy on image inpainting. We conduct comparison between SGE-Net with segmentation maps generated by state-of-the-art segmentation tools and SGE-Net with human-labeled maps. We utilize the DPN model [19] pre-trained on [30] as the segmentation tool for **Outdoor Scenes** as it is the only released model on the dataset. We select the Deeplab v3+ model [3] for **Cityscapes** due to its superior performance on that dataset.

As shown in Table 4, the performance degradation of our SGE-Net trained on imperfect semantic annotations is not significant, meaning that our model can still do a reasonably good job even trained on model-generated semantic annotation. More subjective quality comparisons are provided in supplementary material. Note that the segmentation maps, either human-annotated or model-generated, are only used in the training stage of our model. While completing an image, SGE-Net itself can automatically generate the inpainted image and segmentation map simultaneously, without the need of the semantic annotations.

We also conduct experiments to validate whether the iterative interplay between inpainting and semantic segmentation outperforms the traditional non-iterative segmentation-after-inpainting strategy in semantic segmentation. We compare the segmentation maps generated by SGE-Net itself with initial segmentation maps extracted from images completed by the baselines. As compared in Fig. 10, the results show that SGE-Net evidently beats the segmentation-after-inpainting methods since SGE-Net leads to more accurate semantic assignments and object boundaries, thanks to its joint-optimization of semantics and image contents.

Table 4: Objective quality comparison on model trained by automatic segmentation (Auto-segs) and Human-labeled semantics (Label-segs).

Outdoor Scenes		Cityscapes	
Methods	PSNR	Methods	PSNR
Auto-segs	20.19	Auto-segs	22.94
Label-segs	**20.53**	Label-segs	**23.41**

5   Conclusion

In this paper, a novel SGE-Net with semantic segmentation guided scheme was proposed to complete corrupted images of mixed semantic regions. To address the problem of unreliable semantic segmentation due to missing regions, we proposed a progressive multi-scale refinement mechanism to conduct interplay between semantic segmentation and image inpainting. Experimental results demonstrate that the mechanism can effectively refines poorly-inferred regions through segmentation confidence evaluation to generate promising semantic structures and texture details in a coarse-to-fine manner.

References

1. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graphics **28**(3), 24 (2009)
2. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proc. Conf. Comput. Graphics Interact. Tech. pp. 417–424 (2000)
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 801–818 (2018)
4. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016)
5. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. **13**(9), 1200–1212 (2004)
6. Demir, U., Unal, G.: Deep stacked networks with residual polishing for image inpainting. arXiv preprint arXiv:1801.00289 (2017)
7. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: NeurIPS. pp. 658–666 (2016)
8. Gilbert, A., Collomosse, J., Jin, H., Price, B.: Disentangling structure and aesthetics for style-aware image completion. In: CVPR. pp. 1848–1856 (2018)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS. pp. 2672–2680 (2014)
10. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Trans. Graphics **26**(3), 4 (2007)
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS. pp. 6626–6637 (2017)
12. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Trans. Graphics **36**(4), 107 (2017)
13. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV. pp. 694–711 (2016)
14. Li, J., He, F., Zhang, L., Du, B., Tao, D.: Progressive reconstruction of visual structure for image inpainting. In: ICCV. pp. 5962–5971 (2019)
15. Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. In: CVPR. pp. 3911–3919 (2017)
16. Liao, L., Hu, R., Xiao, J., Wang, Z.: Edge-aware context encoder for image inpainting. In: ICASSP. pp. 3156–3160 (2018)
17. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: ECCV. pp. 89–105 (2018)
18. Liu, H., Jiang, B., Xiao, Y., Yang, C.: Coherent semantic attention for image inpainting. In: ICCV. pp. 4169–4178 (2019)
19. Liu, Z., Li, X., Luo, P., Loy, C.C., Tang, X.: Deep learning markov random field for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **40**(8), 1814–1828 (2017)
20. Ma, Y., Liu, X., Bai, S., Wang, L., He, D., Liu, A.: Coarse-to-fine image inpainting via region-wise convolutions and non-local correlation. In: IJCAI. pp. 3123–3129 (2019)

21. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: Structure guided image inpainting using edge prediction. In: ICCVW. pp. 3265–3274 (2019)
22. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR. pp. 2337–2346 (2019)
23. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR. pp. 2536–2544 (2016)
24. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR (2016)
25. Ren, Y., Yu, X., Zhang, R., Li, T.H., Liu, S., Li, G.: Structureflow: Image inpainting via structure-aware appearance flow. In: ICCV. pp. 181–190 (2019)
26. Song, Y., Yang, C., Lin, Z., Liu, X., Huang, Q., Li, H., Jay Kuo, C.C.: Contextual-based image inpainting: Infer, match, and translate. In: ECCV. pp. 3–18 (2018)
27. Song, Y., Yang, C., Shen, Y., Wang, P., Huang, Q., Kuo, C.C.J.: Spg-net: Segmentation prediction and guidance network for image inpainting. In: BMVC. p. 97 (2018)
28. Sun, J., Yuan, L., Jia, J., Shum, H.Y.: Image completion with structure propagation. ACM Trans. Graphics **24**(3), 861–868 (2005)
29. Wang, N., Li, J., Zhang, L., Du, B.: Musical: multi-scale image contextual attention learning for inpainting. In: IJCAI. pp. 3748–3754 (2019)
30. Wang, X., Yu, K., Dong, C., Change Loy, C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: CVPR. pp. 606–615 (2018)
31. Wang, Y., Tao, X., Qi, X., Shen, X., Jia, J.: Image inpainting via generative multi-column convolutional neural networks. In: NeurIPS. pp. 331–340 (2018)
32. Wang, Z., Jiang, J., Wu, Y., Ye, M., Bai, X., Satoh, S.: Learning sparse and identity-preserved hidden attributes for person re-identification. IEEE Trans. Image Process. **29**(1), 2013–2025 (2019)
33. Wang, Z., Ye, M., Yang, F., Bai, X., Satoh, S.: Cascaded sr-gan for scale-adaptive low resolution person re-identification. In: IJCAI. pp. 3891–3897 (2018)
34. Xiao, J., Liao, L., Liu, Q., Hu, R.: Cisi-net: Explicit latent content inference and imitated style rendering for image inpainting. In: AAAI. pp. 354–362 (2019)
35. Xie, C., Liu, S., Li, C., Cheng, M.M., Zuo, W., Liu, X., Wen, S., Ding, E.: Image inpainting with learnable bidirectional attention maps. In: ICCV. pp. 8858–8867 (2019)
36. Xiong, W., Yu, J., Lin, Z., Yang, J., Lu, X., Barnes, C., Luo, J.: Foreground-aware image inpainting. In: CVPR. pp. 5840–5848 (2019)
37. Yan, Z., Li, X., Li, M., Zuo, W., Shan, S.: Shift-net: Image inpainting via deep feature rearrangement. In: ECCV. pp. 3–19 (2018)
38. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: CVPR. pp. 4076–4084 (2017)
39. Yeh, R.A., Chen, C., Yian Lim, T., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: CVPR. pp. 6882–6890 (2017)
40. Yi, P., Wang, Z., Jiang, K., Shao, Z., Ma, J.: Multi-temporal ultra dense memory network for video super-resolution. IEEE Transactions on Circuits and Systems for Video Technology (2019)
41. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: CVPR. pp. 5505–5514 (2018)
42. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: ICCV. pp. 4470–4479 (2019)

43. Zeng, Y., Fu, J., Chao, H., Guo, B.: Learning pyramid-context encoder network for high-quality image inpainting. In: CVPR. pp. 5961–5970 (2019)
44. Zhang, R., Ren, Y., Qiu, J., Li, G.: Base-detail image inpainting. In: BMVC. p. 195 (2019)
45. Zhang, S., He, R., Sun, Z., Tan, T.: Demeshnet: Blind face inpainting for deep meshface verification. IEEE Trans. Inf. Forensics Security **13**(3), 637–647 (2018)
46. Zhao, D., Guo, B., Yan, Y.: Parallel image completion with edge and color map. Applied Sciences **9**(18), 3856 (2019)
47. Zheng, C., Cham, T.J., Cai, J.: Pluralistic image completion. In: CVPR. pp. 1438–1447 (2019)
48. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1452–1464 (2017)