Boosted GAN with Semantically Interpretable Information for Image Inpainting

1st Ang Li
The University of Melbourne
angl4@student.unimelb.edu.au

2nd Jianzhong Qi
The University of Melbourne
jianzhong.qi@unimelb.edu.au

3rd Rui Zhang
The University of Melbourne
rui.zhang@unimelb.edu.au

4th Ramamohanarao Kotagiri
*The University of Melbourne*
kotagiri@unimelb.edu.au

Abstract—Image inpainting aims at restoring missing regions of corrupted images, which has many applications such as image restoration and object removal. However, current GAN-based inpainting models fail to explicitly consider the semantic consistency between restored images and original images. For example, given a male image with image region of one eye missing, current models may restore it with a female eye. This is due to the ambiguity of GAN-based inpainting models: these models can generate many possible restorations given a missing region. To address this limitation, our key insight is that semantically interpretable information (such as attribute and segmentation information) of input images (with missing regions) can provide essential guidance for the inpainting process. Based on this insight, we propose a boosted GAN with semantically interpretable information for image inpainting that consists of an inpainting network and a discriminative network. The inpainting network utilizes two auxiliary pretrained networks to discover the attribute and segmentation information of input images and incorporates them into the inpainting process to provide explicit semantic-level guidance. The discriminative network adopts a multi-level design that can enforce regularizations not only on overall realness but also on attribute and segmentation consistency with the original images. Experimental results show that our proposed model can preserve consistency on both attribute and segmentation level, and significantly outperforms the state-of-the-art models.

Index Terms—image inpainting, GAN, semantic information, image attribute, image segmentation

I. INTRODUCTION

Given an image where part of the image is missing, image inpainting aims to synthesize plausible contents that are coherent with non-missing regions. Figure 1 illustrates the problem, where Figure 1b shows two input images where a region is missing from each original image. The aim is to fill the missing regions, such that the filled regions contain contents that make the whole image look natural and undamaged (semantically consistent with the original images in Figure 1a). With the help of image inpainting, applications such as restoring damaged images or removing blocking contents from images can be realized.

Traditional inpainting methods [2] [3] [4] mostly rely on a strong assumption that patches (or textures) similar to those in the missing regions should appear in the non-missing regions. They fill the missing regions with background patches obtained from the non-missing regions. These methods only focus on low-level features and thus fail to deal with situations where the missing regions contain non-repetitive contents.
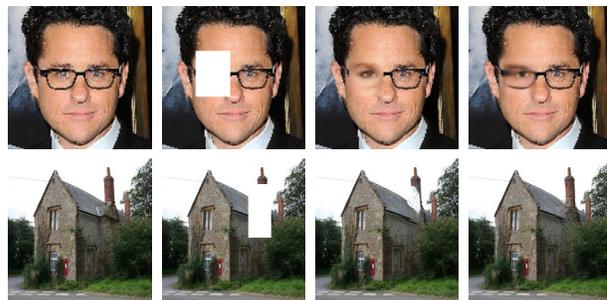
(a) Original   (b) Input   (c) GntIpt   (d) Ours

Fig. 1: Inpainting comparison with the state-of-the-art method GntIpt [1].

Recently, *Generative Adversarial Networks* (GAN) have shown a strong capability in image generation [5] [6] [7]. Inspired by the GAN model, image inpainting can be seen as a conditional image generation task given the context of input images [1] [8] [9] [10] [11]. A representative work following this model is *Context Encoder* [9], where a convolutional encoder-decoder network is trained with a combination of reconstruction loss and adversarial loss [5] to predict the missing contents. The reconstruction loss guides to recover the overall coarse structure of the missing region, while the adversarial loss guides to choose a specific distribution from the results and promotes natural-looking patterns.

In spite of the encouraging results, the restored images by current GAN-based inpainting methods may be perceptually contradictory with their corresponding ground truth images. For example in Figure 1c, the inpainting results generated by the latest state-of-the-art inpainting method *GntIpt* [1] encounter attribute mismatch (inconsistent on the attribute "wearing eyeglasses" in the first image) and segmentation structure misalignment (part of the building missing in the second image) with the non-missing contexts. Such limitations come from the ill-posed GAN-based formulation of the problem: the one-to-many mapping relationships between a missing region and its possible generated restorations. This can be ascribed to the limitation of commonly used combination of loss functions: a reconstruction loss for pixel-wise identity and an adversarial loss for determining overall image realness. The reconstruction loss mainly focuses on minimizing pixel-wise difference (very low-level), while the adversarial loss determines overall similarity with real images (very high-level). Intuitively, mid-level (between the pixel and

overall realness levels) regularizations which can provide more explicit guidances on semantics are needed. This limitation can lead to sub-optimal inpainting results as shown in Figure 1 due to the lack of specific regularizations on semantic consistency in inpainting model training.

To address this limitation, our key insight is that semantically interpretable information of input images (with missing regions) can provide essential guidance for the inpainting process. Here, *semantically interpretable information* refers to human-defined label information that is understandable and has explicit meanings. Arguably, the two most representative types of semantically interpretable information of images are image attribute information and image segmentation information: (1) Attribute information is usually represented as a vector representing multiple descriptive text labels related to an image, e.g., gender, hair color, and face shape of a facial image. It can provide direct and specific hint to generate image contents with the specified attributes [12] [13]. For example, for the first image in Figure 1b, a pretrained classification model can predict the image attribute vector to tell attributes such as "wearing eyeglasses", "male", "not smiling", etc. (2) Segmentation information is usually represented as a one-dimensional map that associates each pixel of an image with a region label, e.g., regions of eyes, mouth, and nose from a facial image. It can tell the spatial relationship and boundaries among different segmentation regions [14] [15], and therefore enforce the inpainting process to focus more on predicting and restoring object structures related to different segmentation regions. Note that a wide range of large-scale image datasets labeled with attribute or segmentation information [16] [17] [18] [19] are easily accessible nowadays. Therefore, we aim to utilize and incorporate these two essential information in our inpainting process. In order to effectively determine these two types of information for input images, we can pretrain state-of-the-art multi-label image classification models [20] [21] [22] [23] [24] and image semantic segmentation models [25] [26] [27] [28] [29] on auxiliary labeled datasets and incorporate them in image inpainting process. The auxiliary labeled datasets do *not* need to contain images that are exactly the same as those to be inpainted, as long as they are in similar categories. For example, both images in our inpainting training dataset and its corresponding auxiliary labeled datasets are facial images, but they do not have to be images of the same person.

Based on this insight, we propose a novel *boosted GAN with semantically interpretable information for image inpainting*. Our proposed model is based on a GAN structure that consists of an inpainting network and a discriminative network. The inpainting network employs two pretrained components: an attribute embedding network to predict attribute vectors of input images (images with missing regions) and a segmentation embedding network to predict segmentation maps of input images. Then, we combine input images with their attribute vectors and segmentation maps to provide guidance for the generator to restore missing regions of the input images on both attribute and segmentation levels. The generator follows an encoder-decoder structure with dilated convolutional layers used in the mid-layers to enlarge receptive fields [8]. To examine whether a restored image is similar to its ground truth image not only on overall realness but also on attribute and segmentation levels, we propose a multi-level discriminative network. It consists of three discriminators: a global discriminator that enforces overall realness, an attribute discriminator that regularizes attribute-level errors, and a segmentation discriminator that regularizes segmentation structure consistency with ground truth images. As shown in Figure 1d, our model produces restored images with much better semantic consistency than those of the state-of-the-art model GntIpt [1]. Our experimental results confirm that attribute regularization is essential for generating contents with correct attributes and sharp details, while segmentation regularization benefits alignment and consistency in segmentation structure. We also conduct ablation study to analyze the effects of attribute regularization and segmentation regularization when used independently or combined with different trade-off parameters. We find that the two regularizations have positive affects on each other and can achieve better results when incorporated together. Besides, we introduced a novel semantic-level quantitative evaluation for image inpainting based on image retrieval.

In summary, the contributions of our paper are:

- We propose a boosted GAN with semantically interpretable information for image inpainting that consists of an inpainting network and a discriminative network.
- The inpainting network utilizes an attribute embedding network and a segmentation network to discover the attribute and segmentation information of corrupted (input) images. These two types of information are incorporated into the inpainting process to provide explicit semantic guidance. We also study the impact of inpainting quality contributed by the attribute information and segmentation information independently.
- The discriminative network adopts a multi-level design with three different discriminators. The global discriminator enforces overall realness with ground truth images, while the attribute discriminator regularizes attribute-level errors and the segmentation discriminator regularizes segmentation structure consistency.
- Experiments show that our model can effectively reduce the semantic inconsistency and generate visually plausible results compared with state-of-the art models.

II. RELATED WORK

A. Generative Adversarial Networks

The Generative Adversarial Network (GAN) [5] [30] [6] model is built on a game scenario consisting of two competitively learning networks: a generator and a discriminator. The generator learns to generate samples that share similar distribution with the training data. In contrast, the discriminator examines samples to identify whether they are from the generator's distribution or the training data distribution. Both networks are trained alternatively and the competition drives

them to improve until the generated samples are indistinguishable from the genuine samples.

B. Image Inpainting

Traditional image inpainting methods are mostly based on patch matching [3] [31] or texture synthesis [4] [32]. They suffer in the quality of the generated images when dealing with large arbitrary missing regions. Recently, deep learning and GAN-based approaches have been employed to produce more promising inpainting results. Phatak *et al.* [9] first propose the *Context Encoder* (CE) model that has an encoder-decoder CNN structure and train CE with the combination of reconstruction loss and adversarial loss [5]. The CE model can generate better images comparing with those generated by traditional methods. However, the images generated by CE still tend to be blurry with evident artifacts and lack fine-grained details due to the limitation of the basic encoder-decoder generator structures. Later methods use post-processing on top of images inpainted with encoder-decoder models to improve the quality of generate images. For example, Yang *et al.* [10] take CE as the initial stage and refine inpainting results by propagating surrounding texture information. Iizuka *et al.* [8] make further improvement using global and local discriminators, followed by Poisson blending as post processing. Yu *et al.* [1] propose a refinement network based on contextual attention. These methods mainly focus on enhancing the resolution of inpainted images, while ignoring the semantic consistency between the inpainted contents and the existing image context. Their models start with a classical encoder-decoder structure, which can easily suffer from the ill-posed one-to-many ambiguity when generating initial results. Then the post-processing stage fails to take effect if given semantically incorrect intermediate results. Liu *et al.* [33] propose partial convolution in order to utilize only valid pixels in convolution, which also fails to address the semantic inconsistency problem.

We refer to two previous studies [11] [34] that contain the phrase "semantic image inpainting" in their titles. Their notion of "semantic" is different from ours. Specifically, they use "semantic" to refer to the process of restoring the missing regions based on the context of input images (this is what GAN-based inpainting process does). In contrast, we explore explicit human-understandable information (image attribute and segmentation information) from input images to boost and regularize the GAN-based inpainting process.

C. Semantic Regularization for Deep Encoder-Decoders

Efforts have been made to improve autoencoder-based image synthesis models using semantic regularization. Yan *et al.* [35] propose a conditional variational autoencoder with attribute-induced semantic regularization. Reed *et al.* [12] develop a generative adversarial model to generate images using text descriptions. Zhang *et al.* [13] propose a two-stage process to further improve the synthesis results. Our study differs from these studies in the following aspects: (1) We

focuses on image inpainting, while these studies aim at text-to-image image synthesis. (2) We introduce a unified semantic regularization strategy that can explore and integrate both attribute and segmentation information, while text-to-image synthesis tasks only use pre-labeled text description. (3) Our inpainting training dataset does not need to be labeled and only images are provided as input during testing, while datasets for text-to-image image synthesis tasks are supposed to be labeled with corresponding text description.

III. METHODOLOGY

Given an input image x with missing regions (filled with zeros), image inpainting aims to restore the missing regions so that the output image z can be consistent with the ground truth image y. In this work, we utilize semantically interpretable information (attribute and segmentation information) of the input images (with missing regions) with the help of auxiliary labeled datasets. Specifically, we set a *prerequisite* for our model: auxiliary dataset d_1 labeled with attribute information and auxiliary dataset d_2 labeled with segmentation information should be available. Images in these two datasets are supposed to have similar contents to those of the images in our inpainting training dataset d_0 (e.g., the images all contain human faces), but they do not have to be the same as those in d_0 (e.g., the images may be from different people).

A. Model Structure

We propose a *boosted GAN with semantically interpretable information for image inpainting*. Our model is based on a GAN structure that consists of an inpainting network and a discriminative network. The inpainting network (Section 3.1) utilizes two auxiliary embedding networks (pretrained on relevant auxiliary labeled datasets) to explore two types of semantically interpretable information (attribute information and segmentation information) of input images in order to provide explicit semantic-level guidance for the generator network. The discriminative network (Section 3.2) focuses on constraining not only the output images to be similar to ground truth ones, but also the attribute and segmentation of output images to be consistent with those of the input images. The architecture of our model is illustrated in Figure 2.

1) Inpainting Network: The inpainting network consists of three sub-networks: an *encoder-decoder generator network* G, an *attribute embedding network* W_a, and a *segmentation embedding network* W_s. Our generator network is built on the architecture proposed by Iizuka *et al.* [8], which follows a general encoder-decoder structure, with dilated convolutional layers used in the mid-layers to enlarge receptive fields. Since GAN-based image inpainting is inherently a one-to-many mapping problem, an autoencoder network alone may generate diverse contents that are in conflict with surrounding contexts. Unlike existing image inpainting methods that only take input images themselves as inputs, we propose to extract semantically interpretable information from these input images first and treat it as auxiliary information to boost inpainting
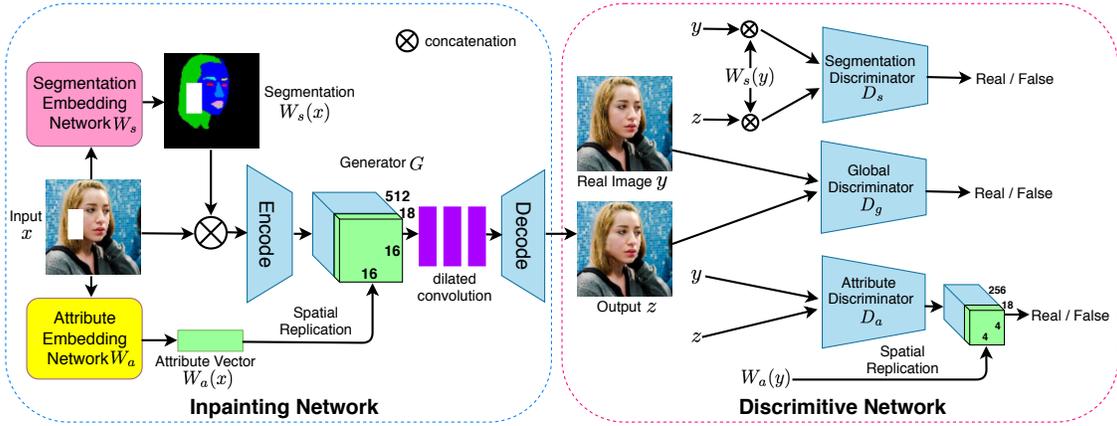
Fig. 2: Overview of our proposed model (best viewed in color).

process. This strategy can provide explicit guidance on attribute level and segmentation level to increase semantic consistency between restored contents and surrounding contexts. The following two networks are utilized to retrieve and embed the attribute and segmentation information with input images.

Attribute Embedding Network Attribute information provides detailed text guidance for image inpainting. The attribute embedding network is a multi-label classification network. It aims to predict an attribute vector for each input (corrupted) image. For example, in Figure 2, the attribute vector of the input image predicted by the attribute embedding network can tell attributes such as 'female', 'young', 'oval face', *etc*. We use the state-of-the-art multi-label classification model [36] as our attribute embedding network. We pretrain this network on an external multi-label image dataset in which the images are similar to the ones in our inpainting training dataset.

Segmentation Embedding Network Segmentation information tells the spatial relationship and boundaries among different segmentation regions. The segmentation embedding network is a semantic segmentation network. It aims to predict the segmentation maps for the input corrupted images. We use the state-of-the-art semantic segmentation model [27] as our attribute embedding network. It is pretrained on an external image segmentation dataset in which the images are also similar to the ones in our inpainting training dataset.

To extract attribute and segmentation information, we first feed an input image x into the attribute embedding network W_a and the segmentation embedding network W_s as shown in Figure 2. We obtain the attribute vector $W_a(x)$ and the segmentation map $W_s(x)$ of x respectively. The input image x is concatenated with its segmentation map $W_s(x)$, and fed into the generator network G which then becomes an intermediate feature map with a spatial size of $M_1 \times M_1$. Then the N_1 dimensional attribute vector $W_a(x)$ is spatially replicated to a $M_1 \times M_1 \times N_1$ tensor, and concatenated with the intermediate image feature map along the channel dimension as illustrated in the green and blue vectors of Figure 2. Finally, dilated convolutional layers and a series of upsampling layers (*i.e.*, decoder) are used to generate the restored image z, which can be represented as follows:

$$z = G(x, W_s(x), W_a(x)) \tag{1}$$

2) Discriminative Network: The problem of semantic inconsistency poses a new challenge to the GAN discriminator: it now not only needs to measure the general similarity between restored images and real ones but also to provide explicit feedback for the generator to reach better consistency on both attribute and segmentation levels for the input and generated (restored) images. To address this challenge, we propose a multi-level discriminative network. As illustrated in Figure 2, this network consists of three discriminators focusing on regularizations of different levels: a *global discriminator* D_g, an *attribute discriminator* D_a, and a *segmentation discriminator* D_s.

The global discriminator examines the overall coherence between restored images and real images. It is a binary classifier constructed by convolutional layers and fully connected layers to test whether its input image is fake (restored) or real.

The attribute discriminator first predicts the attribute vector $W_a(y)$ of the real image y and then spatially replicates it to a $M_2 \times M_2 \times N_1$ tensor. At the same time, the image (the generator output image z or the real image y) is fed into the attribute discriminator to form a feature map with a spatial dimension of $M_2 \times M_2$. Then this feature map is concatenated with the attribute vector over the channel dimension and further fed to the remaining part of the attribute discriminator to obtain the decision score.

For the segmentation discriminator, both the restored image and the real image are concatenated with the real image's segmentation map $W_s(y)$ before being fed into the segmentation discriminator. The structure of the segmentation discriminator is the same as that of the general discriminator.

We utilize the matching-based design [12] for both attribute discriminator and segmentation discriminator. During training, the attribute (or segmentation) discriminator treats real images and their predicted attribute vectors (or segmentation maps) as positive sample pairs, while negative pairs have two groups: (1) real images and mismatched attribute vectors (or segmentation maps); (2) restored images and the predicted attribute vectors (or segmentation maps) of real images.

B. Model Training

Our model is trained in an end-to-end procedure. In a single training epoch, the parameters of the discriminative

network are first updated and then followed by those of the inpainting network, since the inpainting network depends on the back propagation of loss from the discriminative network. The objective function of the discriminative network $\mathcal{L}_D$ is:

$$\mathcal{L}_D = \mathcal{L}_{D_g} + \lambda_a \mathcal{L}_{D_a} + \lambda_s \mathcal{L}_{D_s} \tag{2}$$

where $\mathcal{L}_{D_g}$, $\mathcal{L}_{D_a}$, $\mathcal{L}_{D_s}$ represent the losses of the global discriminator, the attribute discriminator ,and the segmentation discriminator, respectively. Tradeoff parameters λ_a and λ_s control the importance of the three terms. The objective function of the global discriminator follows the original GAN loss function:

$$\mathcal{L}_{D_g} = -E_{y_i \sim P_y}[\log D_g(y_i)] - E_{z_i \sim P_z}[\log(1 - D_g(z_i))] \tag{3}$$

where z_i and y_i represent a generated image and its corresponding real image; P_y and P_z are the real and generated data distributions, respectively.

For the attribute discriminator, we take the real image y_i and its corresponding predicted attribute vector $W_a(y_i)$ as a positive sample pair $\{y_i, W_a(y_i)\}$. Negative sample pairs consist of two situations: $\{z_i, W_a(y_i)\}$ and $\{y_i, \overline{W}_a(y_i)\}$, where $\overline{W}_a(y_i)$ represents the mismatched attribute vector of y_i. The objective function for the attribute discriminator is:

$$\begin{aligned} \mathcal{L}_{D_a} = -\, & E_{y_i \sim P_y}[\log D_a(y_i, W_a(y_i))] \\ & - E_{z_i \sim P_z}[\log(1 - D_a(z_i, W_a(y_i)))] \\ & - E_{y_i \sim P_y}[\log(1 - D_a(y_i, \overline{W}_a(y_i)))] \end{aligned} \tag{4}$$

Similarly for the segmentation discriminator, its objective function is formulated as follows:

$$\begin{aligned} \mathcal{L}_{D_s} = -\, & E_{y_i \sim P_y}[\log D_s(y_i, W_s(y_i))] \\ & - E_{z_i \sim P_z}[\log(1 - D_s(z_i, W_s(y_i)))] \\ & - E_{y_i \sim P_y}[\log(1 - D_s(y_i, \overline{W}_s(y_i)))] \end{aligned} \tag{5}$$

where $\overline{W}_s(y_i)$ represents mismatched segmentation maps of real images y_i.

For our inpainting network, the adversarial loss contains three terms that provide guidance on overall similarity, attribute level, and segmentation level ,respectively. Furthermore, we utilize l_2 distance as the reconstruction loss to minimize pixel-wise difference. Hence, the object function $\mathcal{L}_I$ of the inpainting network is:

$$\begin{aligned} \mathcal{L}_I =\, & E[\|z - y\|_2 - \beta(\log D_g(z) + \lambda_a \log D_a(z, W_a(y)) \\ & + \lambda_s \log D_s(z, W_s(y)))] \\ =\, & E_{(y_i, x_i) \sim P_{(y,x)}}[\|G(x_i, W_s(x_i), W_a(x_i)) - y_i\|_2 \\ & - \beta \log D_g(G(x_i, W_s(x_i), W_a(x_i))) \\ & - \beta \lambda_a \log D_a(G(x_i, W_s(x_i), W_a(x_i)), W_a(y_i)) \\ & - \beta \lambda_s \log D_s(G(x_i, W_s(x_i), W_a(x_i)), W_s(y_i))] \end{aligned} \tag{6}$$

where $P_{(y,x)}$ indicates the joint distribution of the ground truth images and the corresponding input corrupted images. β controls the balance between the reconstruction loss and the adversarial loss.

IV. Experiments

To validate our method, both qualitative and quantitative evaluations are performed in this section. We compare with four state-of-the-art inpainting methods: **CE** [9], **GL** [8], **PConv** [33], and **GntIpt** [1]. Following these studies [8] [1], we also use CelebA-HQ [37] and Places2 [17] (we select 20 scene categories) to train our model. The images are resized and cropped to 256×256.

Before training our model on the inpainting dataset, we pretrain our attribute embedding network and segmentation embedding network on two auxiliary labeled datasets, respectively. Both auxiliary datasets are supposed to have similar (not necessarily the same) contents with the inpainting training dataset. *Note that the inpainting dataset for training our full model does not need to be labeled.* For CelebA-HQ, we use CelebA [16] to pretrain the attribute embedding network (we select 18 attributes such as "age", "gender" out of a total of 40 attributes), and Helen Face dataset [19] [38] [19] to pretrain the segmentation embedding network. For Places2, we use SUN [39] to pretrain the attribute embedding network (we select 60 attributes: 20 scene categories that are the same as those of Places2 and 40 object categories), and ADE20k [18] to pretrain the segmentation embedding network.

Implementation Details By default, $M_1 = 16$ and $M_2 = 4$, while N_1 equals to the number of attributes we used to pretrain the attribute embedding network (18 for CelebA-HQ and 60 for Places2). Following existing work [1] [8], we set β as 0.01. Based on the analysis in Section IV-C, we empirically set $\lambda_a = 0.1$ and $\lambda_s = 0.1$. We mask an image with a rectangular region that has a random location and a random size (ranging from 80×80 to 160×160). The training takes one day on both CelebA [16] and Places2 [17] using an Nvidia GTX 1080Ti GPU.

A. Qualitative Comparisons

We show sample images in this section to highlight the performance difference between our methods and the four state-of-the-art methods on CelebA-HQ and Places2. As shown in Figure 3, CE [9] can generate overall reasonable contents but the contents may be coarse and semantically inconsistent with surrounding contexts. For example, the first row shows an image of a person that does not have heavy makeup. However, the generated content by CE is inconsistent on this attribute. CE also encounters segmentation misalignment for the images in the second (misplaced mouth part) and third (blurry facial edge) rows. This is due to the limitation of the original GAN model that only provides overall high-level guidance (whether the inpaining output is generally similar to the ground truth image) and lacks specific regularization in mid-level semantics such as attribute and segmentation. GL [8] and PConv [33] can generate better details than CE but still suffers from semantic inconsistency. GntIpt [1] encounters attribute mismatch in the first (heavy make-up) and second (not smiling) rows and segmentation misalignment in the third (distorted facial edge) row. Since the first stage of GntIpt resembles the overall structure of GL, if the coarse outputs
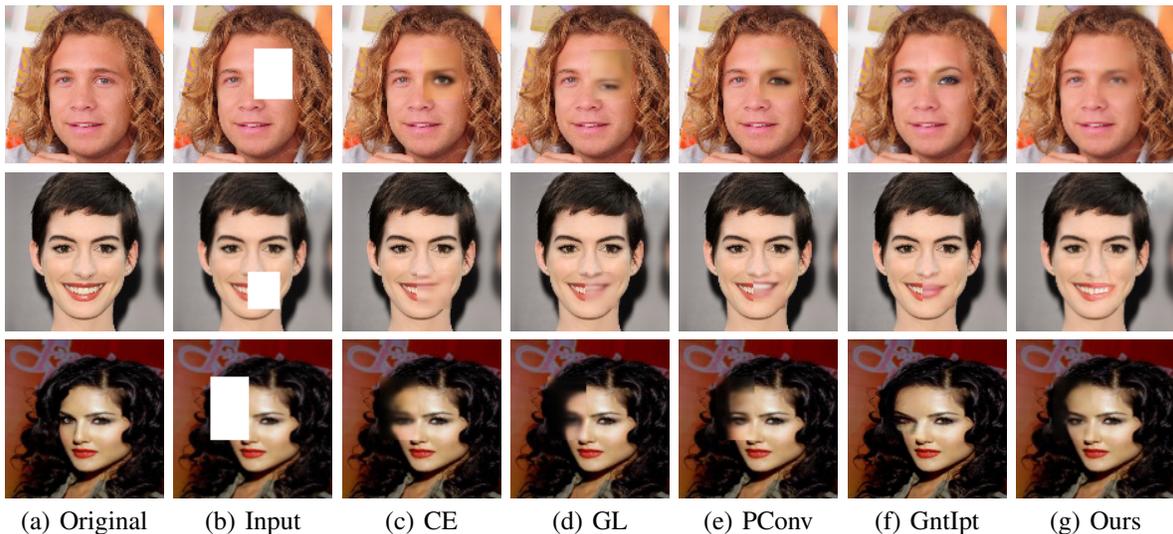
(a) Original   (b) Input   (c) CE   (d) GL   (e) PConv   (f) GntIpt   (g) Ours

Fig. 3: Qualitative comparisons on CelebA-HQ.

(a) Original   (b) Input   (c) CE   (d) GL   (e) PConv   (f) GntIpt   (g) Ours
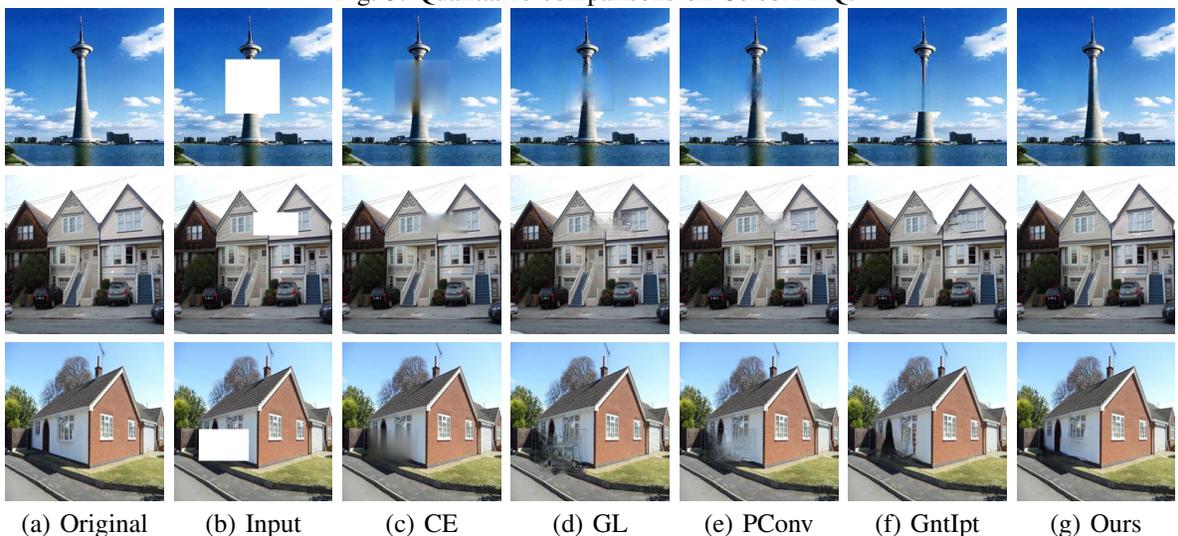
Fig. 4: Qualitative comparisons on Places2.

from the first stage of GntIpt contain semantically inconsistent contents with the ground truth images, the second stage of GntIpt (*i.e.*, refinement) does no help rescue the results and may even cause extra artifacts. In contrast, our model can take advantage of the attribute and segmentation information that already exists in the input images, and can give explicit guidance on both attribute and segmentation level during inpainting process. Therefore, our method can achieve better consistency over different attributes and better region boundary alignment compared with the four state-of-the-art methods.

We also evaluate the methods on Places2 as shown in Figure 4. Again our method performs favorably in generating semantically plausible and photo-realistic images.

B. Quantitative Comparisons

Following previous studies [1] [9] [10] [11], we compute the mean l_1 error, mean l_2 error, *Peak Signal-to-Noise Ratio* (PSNR) and *Structural Similarity* (SSIM) over the restored image and the corresponding testing image for quantitative comparison. As shown in Table I, our proposed method

outperforms the four methods among all the four metrics, and the results are consistent on both datasets.

TABLE I: Quantitative comparisons on CelebA-HQ and Places2.

Model	mean l_1 loss (smaller is better)	mean l_2 loss (smaller is better)	PSNR (larger is better)	SSIM (larger is better)
CelebA-HQ				
CE	10.51%	2.92%	17.78	0.906
GL	9.60%	2.57%	18.09	0.923
PConv	9.33%	2.41%	18.61	0.938
GntIpt	9.18%	2.28%	18.80	0.940
Ours	**8.92%**	**2.19%**	**19.11**	**0.943**
Places2				
CE	9.77%	2.69%	19.31	0.798
GL	9.38%	2.21%	19.56	0.812
PConv	8.36%	2.04%	21.79	0.839
GntIpt	8.61%	2.10%	21.08	0.836
Ours	**8.26%**	**1.97%**	**22.32**	**0.845**

C. Ablation Study/Parameter Analysis

We further conduct ablation experiments for our model on Places2. Specifically, we investigate different combinations of attribute regularization and segmentation regularization in the
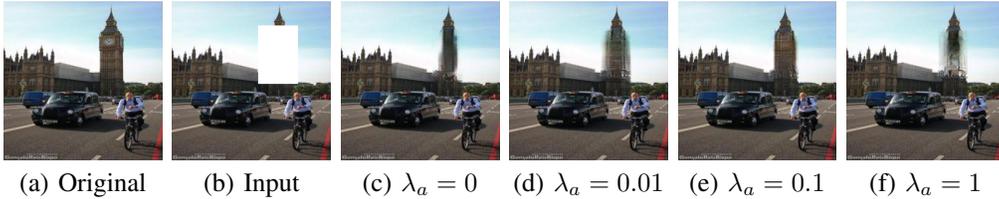
(a) Original (b) Input (c) $\lambda_a = 0$ (d) $\lambda_a = 0.01$ (e) $\lambda_a = 0.1$ (f) $\lambda_a = 1$

Fig. 5: Qualitative comparisons of different λ_a for attribute regularization. Here $\lambda_s = 0.1$.

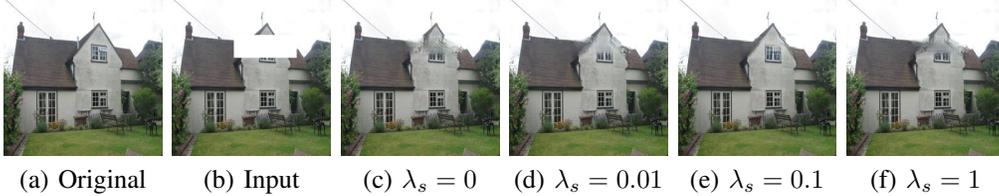(a) Original (b) Input (c) $\lambda_s = 0$ (d) $\lambda_s = 0.01$ (e) $\lambda_s = 0.1$ (f) $\lambda_s = 1$

Fig. 6: Qualitative comparisons of different λ_s for segmentation regularization. Here $\lambda_a = 0.1$.

loss function by (1) setting λ_a as 0 and varying $\lambda_s \in [0.01, 1]$, (2) setting λ_s as 0 and varying $\lambda_a \in [0.01, 1]$, (3) setting λ_a as 0.1 and varying $\lambda_s \in [0.01, 1]$, (4) setting λ_s as 0.1 and varying $\lambda_a \in [0.01, 1]$. Qualitative comparisons are shown in Figure 5 and 6. Based on the visual results, we see that attribute regularization is able to capture attribute coherency and help generate sharp details, while segmentation regularization helps preserve region boundaries and structures. Quantitative comparisons with 4 groups of parameter setting are given in Table II. We see that both attribute regularization and segmentation regularization contributes to the model performance when using them alone. Moreover, the collaboration of these two different semantic regularizations in our model can achieve complementary effects to improve inpainting quality. We empirically set $\lambda_a = 0.1$ and $\lambda_s = 0.1$ based on the ablation experiments.

TABLE II: Quantitative comparisons with different combinations of λ_a and λ_s on Places2.

$\lambda_a = 0$			
λ_s	0.01	0.1	1
PSNR	20.72	**21.21**	20.38
SSIM	0.831	**0.838**	0.832
$\lambda_s = 0$			
λ_a	0.01	0.1	1
PSNR	21.06	**21.80**	21.54
SSIM	0.816	**0.828**	0.820
$\lambda_a = 0.1$			
λ_s	0.01	0.1	1
PSNR	21.46	**22.32**	21.84
SSIM	0.835	**0.846**	0.840
$\lambda_s = 0.1$			
λ_a	0.01	0.1	1
PSNR	21.62	**22.32**	22.15
SSIM	0.835	**0.846**	0.838

D. Evaluation from Semantic Perspective

Recent studies [1] [10] find that inpainting results with high PSNR may turn out to be overly smooth and semantically unsatisfactory. Thus, instead of relying only on low-level metrics such as PSNR, we propose a semantic-level metric based on image retrieval. Specifically, given a specific image retrieval method m and a image retrieval benchmark dataset, *we first replace the ground truth images of each query image with the returned results of this query image by m*. Then we put same masks on the original query images, and a certain inpainting method i is implemented to recover these masked query images. Finally we use the restored images as query images and conduct image retrieval on the benchmark dataset with the same image retrieval method m to compute the mean average precision mAP_i, with respect to the retrieved results of original query images as our ground truth. Therefore, mAP_i can evaluate to what extent this inpainting method i can narrow down the gap between the original query images and the restored ones, which directly measures how successfully the inpainting method i restores the images from semantic level. This can actually be done on any dataset since we do not need manual-labeled ground truth (we use the retrieved results of original query images as our ground truth).

TABLE III: Comparison of the proposed mAP on benchmark datasets.

Method	Paris	Oxford	Caltech256
Masked	76.23%	71.82%	84.28%
CE	83.82%	80.88%	89.73%
GL	88.79%	83.42%	92.06%
PConv	93.21%	87.27%	95.14%
GntIpt	91.65%	87.36%	93.32%
Ours	**96.53%**	**93.68%**	**97.40%**

To compare this proposed mAP of the four methods and our method, we choose three image retrieval benchmark datasets (Paris [40], Oxford [41] and Caltech256 [42]). We resize the query images to 256×256 and employ a 128×128 missing region at the center of the images. We use a state-of-the-art image retrieval method DF.FC2 in [43], which uses the second fully connected layer of a pre-trained VGG-16 network [44] to obtain the image feature representation and Euclidean distance to measure the similarity between query images and ground truth images. Results of the mAP values of the different models are shown in Table III. Here, larger mAP values are preferred, which means that using recovered images recovered from our model for image retrieval produces the most similar query results as those produced by using the original query images for image retrieval. Our method outperforms the four state-of-the-art inpainting methods over all the three benchmark datasets. Here, "Masked" means the mAP value of directly using the masked query images as input for image retrieval.

V. CONCLUSION

We studied the semantic consistency problem in image inpainting and proposed a unified boosted GAN with semantically interpretable information for image inpainting that can generate contents consistent with the surrounding contexts both on attribute level and segmentation level. The proposed inpainting network utilizes two auxiliary components: an attribute embedding network and a segmentation embedding network to discover the attribute and segmentation information of corrupted (input) images and incorporates them into the inpainting process. The proposed multi-level discriminative network enforces regularizations not only on overall realness but also on attribute and segmentation consistency with the original images. We evaluated the proposed method from both pixel level and semantic level. The experimental results confirm that our proposed method outperforms the state-of-the-art methods consistently on real datasets. Our model can be improved further, for example, since our attribute and segmentation embedding networks are pretrained on auxiliary datasets, attribute vectors and segmentation maps might not be accurately predicted for input images, and this might result in an inaccurate supervision for inpainting. Our future work will explore domain-transfer-based strategies to improve the performance of our proposed method.

REFERENCES

[1] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," *CVPR 2018*.

[2] M. V. Afonso, J. M. Bioucas-Dias, and M. A. Figueiredo, "An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 681–695, 2011.

[3] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 24–1, 2009.

[4] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *ICCV 1999*.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS 2014*.

[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *ICLR 2016*.

[7] X. Wang, R. Zhang, Y. Sun, and J. Qi, "Kdgan: Knowledge distillation with generative adversarial networks," in *NIPS 2018*.

[8] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 107, 2017.

[9] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR 2016*.

[10] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," in *CVPR 2017*.

[11] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *CVPR 2017*.

[12] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *ICML 2016*.

[13] H. Zhang, T. Xu, and H. Li, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *ICCV 2017*.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR 2017*.

[15] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *CVPR 2018*.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV 2015*.

[17] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[18] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *CVPR 2017*.

[19] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *ECCV 2012*.

[20] J. Jin and H. Nakayama, "Annotation order matters: Recurrent image annotator for arbitrary length image tagging," in *ICPR 2016*.

[21] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun, "Semantic regularisation for recurrent image annotation," *CVPR 2017*.

[22] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *CVPR 2016*.

[23] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "Hcp: A flexible cnn framework for multi-label image classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1901–1907, 2016.

[24] H. Yang, J. Tianyi Zhou, Y. Zhang, B.-B. Gao, J. Wu, and J. Cai, "Exploit bounding box annotations for multi-label object recognition," in *CVPR 2016*.

[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[26] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[27] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *arXiv preprint arXiv:1802.02611*.

[28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR 2015*.

[29] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR 2017*.

[30] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[31] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," *IEEE Transactions on Image Processing*, vol. 12, no. 8, pp. 882–889, 2003.

[32] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *SIGGRAPH 2001*.

[33] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," *ECCV 2018*.

[34] H. Zhang, Z. Hu, C. Luo, W. Zuo, and M. Wang, "Semantic image inpainting with progressive generative networks," in *ACMMM 2018*.

[35] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2image: Conditional image generation from visual attributes," in *ECCV 2016*.

[36] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *ICCV 2017*.

[37] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *ICLR 2018*.

[38] B. M. Smith, L. Zhang, J. Brandt, Z. Lin, and J. Yang, "Exemplar-based face parsing," in *CVPR 2013*.

[39] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR 2010*.

[40] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR 2007*.

[41] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *CVPR 2008*.

[42] G. Griffin, A. Holub, and P. Perona, "Caltech256 image dataset," 2006.

[43] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *ACMMM*, 2014.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.