UCTGAN: Diverse Image Inpainting based on Unsupervised Cross-Space Translation

Lei Zhao*, Qihang Mo, Sihuan Lin, Zhizhong Wang,
Zhiwen Zuo, Haibo Chen, Wei Xing, Dongming Lu

College of Computer Science and Technology, Zhejiang University
{cszhl, moqihang, linsh, endywon, zzwcs, feng123, wxing, ldm}@zju.edu.cn

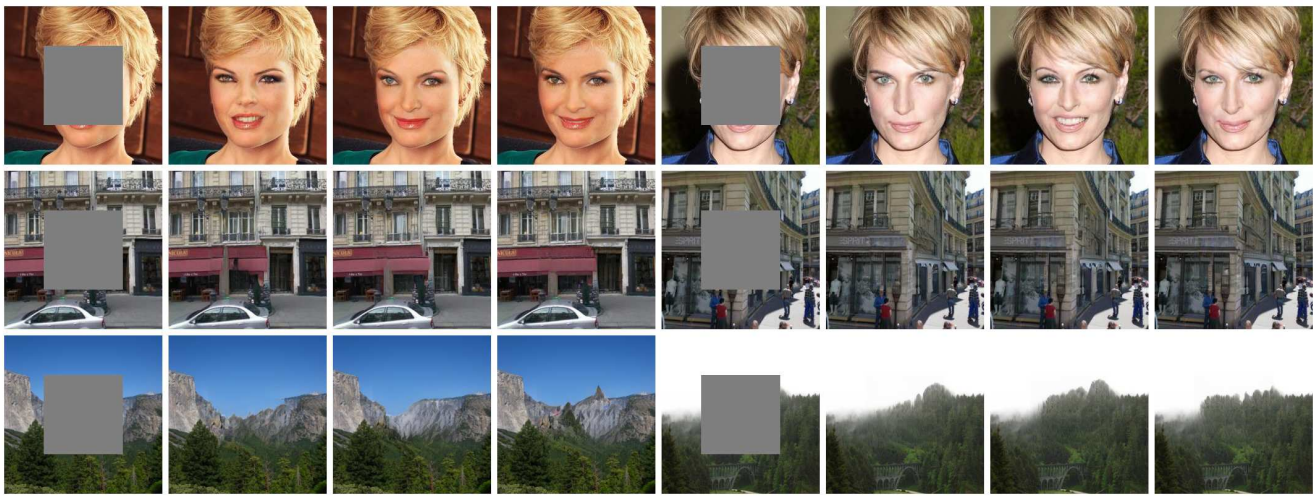

Figure 1: Exemplar inpainting results of our method on images of face (from CelebA-HQ [9]), building (from Paris [4]) and natural scene (from Places2 [30]). Missing regions are shown in gray. From left to right, we show the masked input image, the diverse and reasonable outputs of our model without any post-processing (zoom in to see the details).

Abstract

Although existing image inpainting approaches have been able to produce visually realistic and semantically correct results, they produce only one result for each masked input. In order to produce multiple and diverse reasonable solutions, we present Unsupervised Cross-space Translation Generative Adversarial Network (called UCTGAN) which mainly consists of three network modules: conditional encoder module, manifold projection module and generation module. The manifold projection module and the generation module are combined to learn one-to-one image mapping between two spaces in an unsupervised way by projecting instance image space and conditional completion image space into common low-dimensional manifold space, which can greatly improve the diversity of the repaired samples. For understanding of global information, we also introduce a new cross semantic attention layer that exploits the long-range dependencies between the known parts and the completed parts, which can improve realism and appearance consistency of repaired samples. Extensive experiments on various datasets such as CelebA-HQ, Places2, Paris Street View and ImageNet clearly demonstrate that our method not only generates diverse inpainting solutions from the same image to be repaired, but also has high image quality.

1. Introduction

Image inpainting (also known as image completion or image hole-filling) refers to using the known information of the images and a specific method to repair a partially damaged or missing image in an undetectable way. It fills the

* Corresponding author

missing part of an image to make it look natural (*i.e.*, visually realistic and semantically correct) according to some rules [1, 21, 2, 15, 20, 25, 13]. This task is a basic problem in the field of image processing and has drawn great attention for decades, because image inpainting can be used in various applications such as object removal, image editing and old photo restoration, to name a few. Its key problem is how to generate missing contents to maintain the integrity and consistency of the restored images, and avoid incomplete fusion between the filled contents and the known visible contents.

Image inpainting is an underdetermined inverse problem, where a large number of plausible solutions can satisfy the constraints of image restoration. In this paper, our main object is to produce multiple and diverse reasonable results when given a masked image, so we refer to this task as diverse image inpainting.

Early image inpainting approaches progressively fill in missing regions by searching and pasting the most similar image patches in the background regions under the hypothesis that the contents to be filled in come from the background areas [1, 21, 2]. This hypothesis is not always conformed with the real cases. Though these approaches work well for some cases, they can not generate semantically meaningful contents. More recently, some deep-learning based image inpainting approaches have been proposed to learn the essential distribution of training data, which is used to repair the masked images. However, these approaches can only generate one optimal result, and do not have the capacity to produce various semantically meaningful solutions.

Nowadays, typical GAN-based image generation approaches, such as [6, 19, 9, 3], have been able to generate novel and diverse image samples by mapping the noise of normal distribution to the image. However, they can not be directly applied to diverse image inpainting for the following reasons: 1) In diverse image inpainting scenario, the condition label is the masked image itself, and there is only one instance (*i.e.*, the ground truth image corresponding to the masked image) in the training set for each condition label. That is to say, there are no conditional training datasets which explicitly express conditional distribution. 2) The diverse image inpainting scenario has strong constraints (the repaired images should keep integrity and consistency in color and texture with the masked image), so it is more vulnerable to suffer from mode collapse than typical image generation.

As we know, the set of all possible repaired results for a given masked image expresses conditional probability distribution, the set of masked images expresses marginal probability distribution and the training dataset expresses joint probability distribution. So, the diverse image restoration can be regarded as the problem of finding conditional probability with known marginal probability and joint probability, which means that we can borrow some information from training data when traversing the conditional completion image space. Inspired by the above analyses, we present a conditional image-to-image translation network for instance-guided diverse image inpainting, conditioned on the masked image.

The main contributions of our work are:

- An instance-guided conditional image-to-image translation framework for diverse image inpainting that is able to learn conditional completion distribution when given a masked image.
- A new network structure with two branches, which learns one-to-one image mapping between instance image space and conditional completion image space in an unsupervised way. Our method has much higher sampling diversity as compared to existing methods.
- A novel cross semantic attention layer that exploits long-range global information to ensure appearance and structure consistency in the image domain.
- We demonstrate that our approach is able to generate multiple reasonable solutions that have significant diversity for a masked image input, such as those shown in Fig. 1.

2. Related Work

Non-deep-learning based inpainting

Non-deep-learning based inpainting approaches mainly utilize non-learning prior knowledge (*i.e.*, hand-crafted features), such as statistics of patch offsets and low rank to recover the image. Among them, patch-based methods [5] and diffusion-based methods [11] are the most typical. Patch-based methods were first introduced for texture synthesis [5]. They were then applied in image inpainting to fill missing parts at pixel level [21]. They usually search and borrow similar patches from image datasets or undamaged image background to generate missing parts based on distance metrics between patches [14]. Non-deep-learning methods for image inpainting are able to generate sharp results similar to context. However, it is difficult to produce semantically plausible solutions by patch-based approaches, due to the lack of high level semantic understanding of images.

Deep-learning based inpainting

Deep-learning based inpainting methods often use deep neural networks and GANs to adversarially generate pixels of missing parts [6, 15, 20, 25, 13]. The existing deep learning based inpainting methods are mainly divided into two categories: *single-solution inpainting methods* and *multiple-solution inpainting methods*.

Single-solution inpainting methods produce only one result for each masked input, although there may be many reasonable possibilities. These approaches, such as [28, 8, 15, 20, 24], often generate distorted structures and blurry textures inconsistent with the visible regions. In order to overcome these problems, researchers have done a lot of work, such as [27, 23, 26, 22, 13, 16, 12, 25].

Multiple-solution inpainting methods can produce multiple plausible results for each masked input. Zheng *et al*. [29] proposed a probabilistically principled framework with two parallel paths, which utilized prior-conditional lower bound coupling to generate multiple diverse results with reasonable content for a single masked input. Our method is similar to that of [29] in goal, both are to generate multiple diverse and reasonable results for a masked image input, but our approach uses a different route to improve the diversity and realism of the restored image.

Diverse image generation

Image generation methods produce novel diverse samples according to high-dimension data distributions learned from the image dataset. Currently, the most typical approaches are Variational Autoencoders (VAE) [19] and Generative Adversarial Networks (GAN) [6]. Cross domain image translation can also generate diverse images, such as BicycleGAN (BG) [31], MUNIT [7], DR [10], etc. BicycleGAN (BG) [31] explicitly encourages the connection between output and the latent code to be invertible, which helps prevent a many-to-one mapping from the latent code to the output during training, and produces more diverse results. MUNIT [7] and DR [10] use the content (or style) of one image as a guide, and combine with the style (or content) of another image to achieve diverse image-to-image translation. Inspired by them, we also adopt instance images of training dataset as a guide to perform diverse image inpainting. However, our approach is fundamentally different from MUNIT [7] and DR [10]. Our method does not decouple images into content code and style code. The disentangled representations of content and style are the bases of diverse image-to-image translation performed by MUNIT [7]. The cross adversarial training of two different domains is necessary in order to decouple the content and style of the image in MUNIT [7]. However, in the scenario of diverse image inpainting, the images to be repaired, instance images used as guides and corresponding completion images all belong to the same domain, so MUNIT [7] can not be used for diverse image inpainting since it can not realize the disentangled learning of content and style in a single domain.

3. Our Approach

Suppose we have an image from a training dataset, originally I_g, but degraded by a mask M to become I_m (the masked image) comprising the observed/visible pixels. Our goal is to produce multiple and diverse semantically reasonable and visually realistic completion images I_c for a masked image I_m. The set of all these completion images I_c is called conditional completion image space S_{cc} of a given masked image I_m. The instance image I_i for guidance comes from the training dataset, and the set of all instance images I_i is called instance image space S_i. The network model is prone to suffer from mode collapse in the diverse image inpainting scenario, which results in poor diversity of completion images. In order to improve the variance of the repaired images, our network learns a one-to-one mapping between instance image space S_i and conditional completion image space S_{cc} by an unsupervised way (unsupervised cross-space translation), which is implemented by projecting instance image space S_i and conditional completion image space S_{cc} into common low-dimensional manifold space S_m. The deep neural network of a specific structure is designed to learn a mapping MAP: $S_i \rightarrow S_{cc}$ such that $E_1(I_i) = E_1(I_c)$, where $E_1(\cdot)$ is a multivariate function which projects I_i or I_c into low-dimension manifold space, and $I_c = U(I_i, I_m)$, $U(\cdot)$ is the function expressed by our UCTGAN networks.

3.1. Probabilistic Analysis

Our network framework will maximize the conditional log-likelihood of the training instances, which involves variational lower bound:

$$\begin{aligned}\mathbf{log}p(I_c|I_m) \geq & -KL(f_\varphi(Z_c|I_i, I_m)\|f_\psi(Z_c|I_m)) \\ & +\mathbb{E}_{Z_c\sim f_\varphi(Z_c|I_i,I_m)}[\mathbf{log}g_\theta(I_c|Z_c, I_m)]\end{aligned} \tag{1}$$

where I_i, I_c and I_m are the instance image, the repaired image and the masked image, respectively. Z_c is the latent vector of I_i in space S_m. f_φ, f_ψ and g_θ are the posterior sampling function, the conditional prior and the likelihood, with φ, ψ and θ being the deep network parameters of their corresponding functions. The conditional prior is set as $f_\psi(Z_c|I_m) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The first term mainly projects instance image I_i into low-dimensional manifold vector Z_c which is shared by the completion image corresponding to the instance image.

3.2. Network Structure

Our network is trained in an end-to-end fashion, which consists of two branches, shown in Fig. 2, which mainly consists of three network modules: manifold projection module E_1, conditional encoder module E_2 and generation module G. The primary branch consists of a manifold projection module E_1 and a generation module G, which is responsible for learning one-to-one image mapping between two spaces in an unsupervised way by projecting instance image space S_i and conditional completion image space S_{cc} into one common latent manifold space S_m. The second

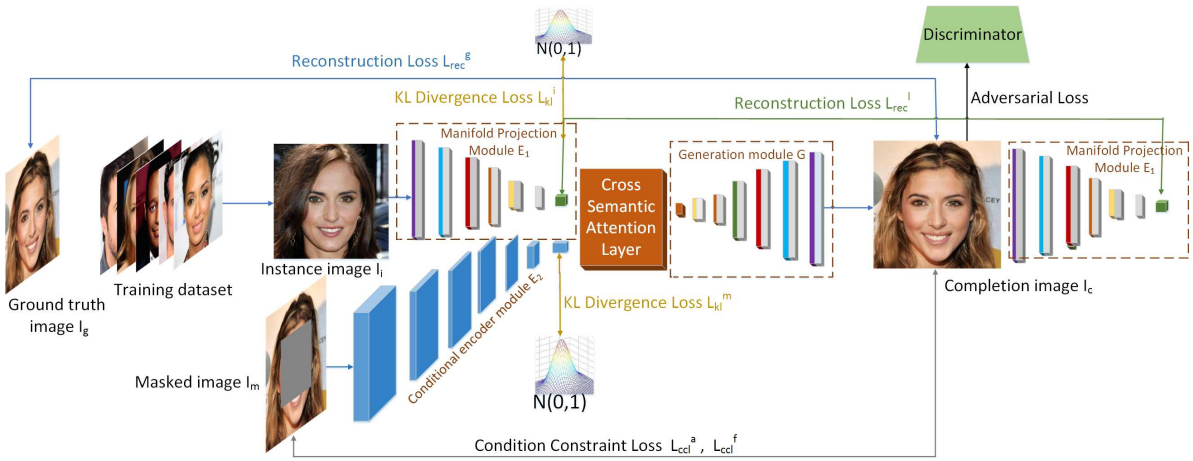

Figure 2: Overview of our architecture with two branches. The primary branch consists of a manifold projection module and a generation module, which is responsible for mapping the instance image space to the conditional completion image space. The secondary branch consists of a conditional encoder module, which acts as the conditional label.

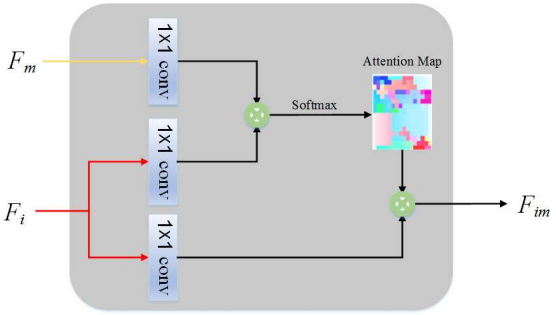

Figure 3: Our cross semantic attention layer. The attention map is computed on masked image features and instance image features on bottleneck layer .

branch consists of a conditional encoder module E_2, which acts as conditional constraint similar to the conditional label. For a masked image I_m, there is only one original image I_g that can be used as training data to maximize the likelihood in equation (1). That is to say, the mapping between instance images and completion images can only be obtained in an unsupervised way, which often results in mode collapse. In order to associate two spaces (instance image space and conditional completion image space) by one-to-one mapping, instance images I_i and their corresponding mapped restored images I_c should have the same representation in the low-dimension manifold space S_m.

3.3. Training Loss

Condition Constraint Loss. The multiple and diverse results generated by our network need to be consistent with the masked image, which requires that the appearance and perceptual features of the repaired images be as same as possible to those of the corresponding masked images in the known pixel regions. We define conditional constraint loss in terms of appearance and perceptual features.

$$\begin{aligned}\mathcal{L}_{ccl} &= \mathcal{L}_{ccl}^{a} + \mathcal{L}_{ccl}^{f} \\ &= \mathbb{E}_{I_i \sim p_{data}} \|(M \odot U(I_i, I_m)) - I_m\|_1 \\ &+ \mathbb{E}_{I_i \sim p_{data}} \|\varphi(M \odot U(I_i, I_m)) - \varphi(I_m)\|_1\end{aligned} \tag{2}$$

where M is the mask, $U(\cdot)$ is the function expressed by our network, p_{data} is the distribution of training dataset, φ is the pretrained feature extractor such as VGG16, $\mathcal{L}_{ccl}^{a}$ and $\mathcal{L}_{ccl}^{f}$ are appearance constraint loss and perceptual constraint loss, respectively.

KL Divergence Loss. The KL divergence loss $\mathcal{L}_{KL}$ is defined as:

$$\begin{aligned}\mathcal{L}_{KL} =& \mathcal{L}_{KL}^{i} + \mathcal{L}_{KL}^{m} \\ =& KL(E_1(Z_c|I_i)\|\mathcal{N}(\mathbf{0}, \mathbf{I})) \\ &+ KL(E_2(Z_m|I_m)\|\mathcal{N}(\mathbf{0}, \mathbf{I}))\end{aligned} \tag{3}$$

where $\mathcal{L}_{KL}^{i}$ and $\mathcal{L}_{KL}^{m}$ are responsible for projecting instance images and masked images into multivariate normal distribution space, E_1 and E_2 are functions represented by manifold projection module and conditional encoder module, respectively. Z_c and Z_m are the latent vector of I_i and I_m in multivariate normal distribution space, respectively.

Reconstruction Loss. Our network translates instance images into completion images in an unsupervised way. However, the instance image is different from the corresponding completion image in pixel level. It is desired that the instance image is the same as the corresponding completion image in low-dimensional manifold space. So the low-dimensional manifold loss is defined as

$$\mathcal{L}_{rec}^{l} = \mathbb{E}_{I_i \sim P_{data}} \|E_1(I_i) - E_1(G(E_1(I_i), E_2(I_m)))\|_1 \tag{4}$$

where I_m is the masked image, I_i is the instance image randomly sampled from training dataset, P_{data} is the dis-

Figure 4: Comparisons on Paris [4] test set for center region completion. For each group, from left to right the images are masked image, CE [15], CA [26], CSA [13], SF [16], SN [23], PN [29] in blue box and our method in red box, respectively.

tribution of training dataset, E_1 is the manifold projection module, E_2 is the conditional encoder module, G is the generation module.

For each masked image I_m, there is only one ground truth image I_g corresponding to it. When its corresponding ground truth image I_g is used as the guided instance image, the output of the generation module is I_g. Therefore, an identical reconstruction constraint is needed, which is defined as follows:

$$\mathcal{L}_{rec}^{g} = \|I_g - G(E_1(I_g), E_2(I_m))\|_1 \tag{5}$$

where I_m is the masked image, I_g is the ground truth image of I_m, E_1 is the manifold projection module, E_2 is the conditional encoder module, and G is the generation module.

The total reconstruction loss $\mathcal{L}_{rec} = \mathcal{L}_{rec}^{l} + \mathcal{L}_{rec}^{g}$.

Adversarial Loss. Our adversarial loss is defined as

$$\begin{aligned}\mathcal{L}_{adv} = &\min_{U}\max_{D}(\mathbb{E}_{I_i\sim p_{data}}\log D(I_i) \\ &+ \mathbb{E}_{I_i\sim p_{data}}\log(1 - D(U(I_i, I_m))))\end{aligned} \tag{6}$$

where p_{data} is the distribution of training dataset, D is the discriminator, and $U(\cdot)$ is our network (UCTGAN).

Full Objective. The total loss function $\mathcal{L}_{total}$ of our network (UCTGAN) consists of four groups of component losses:

$$\begin{aligned}\mathcal{L}_{total} =&\lambda_{rec}(\mathcal{L}_{rec}^{g} + \mathcal{L}_{rec}^{l}) + \lambda_{ccl}(\mathcal{L}_{ccl}^{a} + \mathcal{L}_{ccl}^{f}) \\ &+ \lambda_{adv}\mathcal{L}_{adv} + \lambda_{KL}(\mathcal{L}_{KL}^{i} + \mathcal{L}_{KL}^{m})\end{aligned} \tag{7}$$

where the $\mathcal{L}_{KL}$ group measures the matching degree of two probability distributions in terms of KL divergences, the condition constraint losses $\mathcal{L}_{ccl}^{a}$ and $\mathcal{L}_{ccl}^{f}$ encourage consistency and integrity between completion contents and known contents, reconstruction losses $\mathcal{L}_{rec}^{g}$ and $\mathcal{L}_{rec}^{l}$ encourage one-to-one mapping between the instance image and the repaired image and avoid falling into mode collapse, and adversarial loss $\mathcal{L}_{adv}$ makes repaired images fit in with the distribution of training dataset. The hyper-parameters λ_{rec}, λ_{ccl}, λ_{adv} and λ_{KL} control the relative importance of each group of component loss.

3.4. Cross Semantic Attention

Our proposed cross semantic attention module is shown in Fig. 3. It is added after the max pooling layer of bottleneck layer. Feature maps F_m of the masked image I_m and

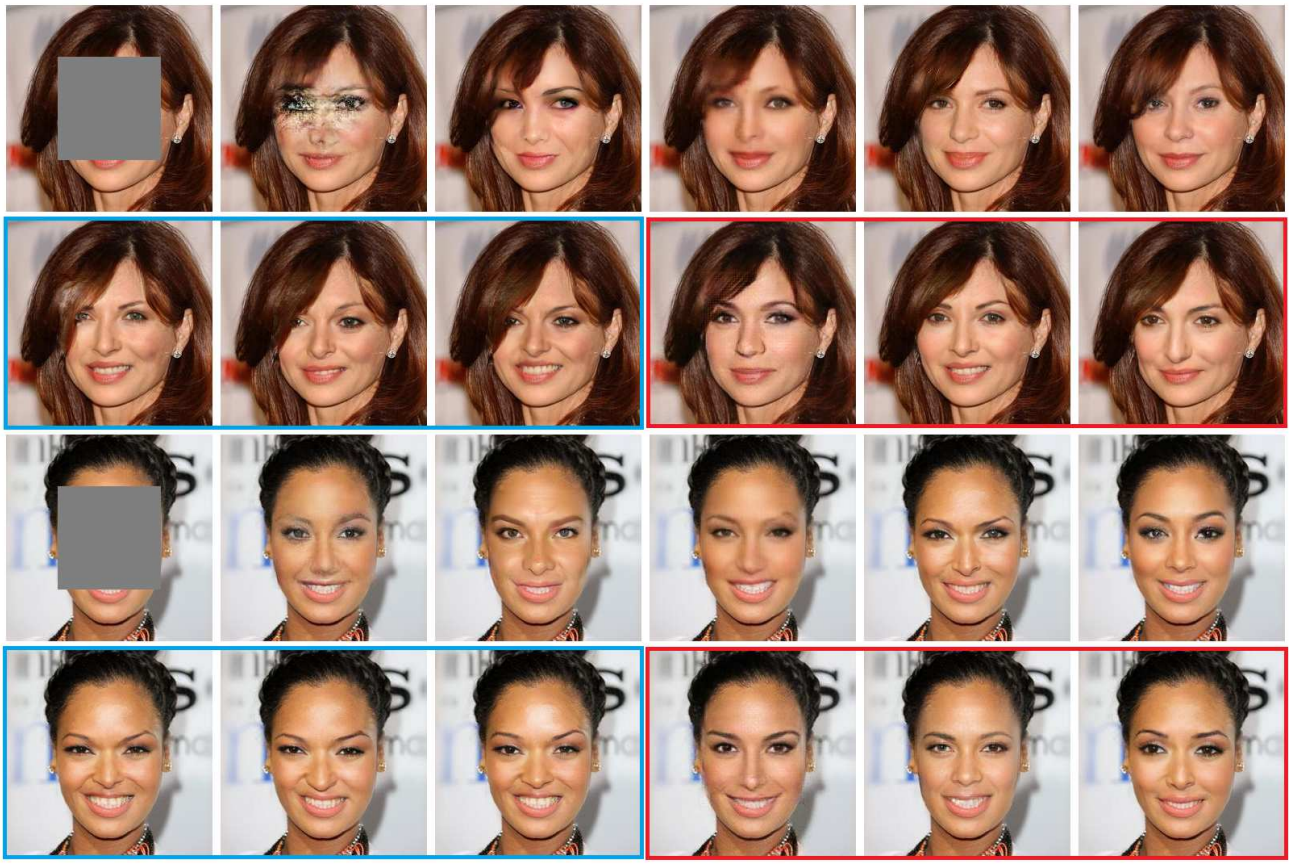

Figure 5: Comparisons on CelebA-HQ [9] test set for center region completion. For each group, from left to right the images are masked image, CE [15], CA [26], CSA [13], SF [16], SN [23], PN [29] in blue box and our method in red box, respectively.

Figure 6: Comparisons on Places2 [30] test set. For each group, from left to right the images are masked image, PN [29] in blue box and our method in red box, respectively.

F_i of the instance image I_i are transformed by 1x1 convolution filter into two feature spaces to evaluate the cross semantic attention F^n_{im} between $\overline{F_m}$ and $\overline{F_i}$.

$$F^n_{im} = \frac{1}{M(F)} \sum_{\forall j} \exp((\overline{F^n_m})^T (\overline{F^j_i}) d(F^j_i)) \qquad (8)$$

where $\overline{F_m} = W_f F_m$, $\overline{F_i} = W_k(F_i)$, $d(F_i) = W_d(F_i)$. The equation is normalized by a factor $M(F) = \sum_{\forall j} \exp((\overline{F^n_m})^T (\overline{F^j_i}))$. Here j is the index that enumerates all possible positions and n is the output position index, W_f, W_k, W_d are the learned weight matrices. Then the output

Algorithm 1 Training procedure of our framework

while G, E_1, E_2 have not converged **do**
  Sample batch images x from training data
  Sample instance images y for x from training data
  Replace the first 3 images of y with the ground truth image I_g of x ▷ batchsize is 8
  Generate random masks M for x
  Construct inputs $x \leftarrow x \odot M$
  Generate outputs $\overline{x} \leftarrow G(E_1(y), E_2(x))$
  Compute all the losses
  Update G, E_1, E_2 with $\mathcal{L}^g_{rec}$, $\mathcal{L}^m_{KL}$, $\mathcal{L}^i_{KL}$, $\mathcal{L}_{adv}$
  Update G with $\mathcal{L}^l_{rec}$, $\mathcal{L}^a_{ccl}$, $\mathcal{L}^f_{ccl}$ keeping E_1,E_2 fixed
  Update D with $\mathcal{L}_{adv}$
end while

F_i^O is:

$$F_i^O = \Gamma_d F_{im} + F_i \tag{9}$$

where Γ_d is a scale parameter for balancing the weights between F_{im} and F_i.

4. Experimental Results

We now prove the advantages of the proposed method by showing the results of diverse image inpainting on four datasets including Paris [4], CelebA-HQ [9], Places2 [30], and ImageNet [17].

Baselines. We compare with the following baselines: context encoders (CE) [15], contextual attention (CA) [26], coheret semantic attention (CSA) [13], structureflow (SF) [16], shift-net (SN) [23], CVAE [19], BicycleGAN (BG) [31], and PICNet (PN) [29].

Implementation details. Our model is learned using the training set and tested on the test set, following the experimental settings used by baselines for fair comparisons. We use images of resolution 256 × 256 with regular holes or irregular holes in random positions. We train our networks using Adam optimizer with β_1 = 0.5 and β_2 = 0.99, and all networks are initialized with Orthogonal Initialization. The learning rate is initialized at 0.0001 and it multiplies by 0.97 per 1K iterations. We train the network for 500K iterations. The batch size is 8. We choose low-dimensional manifold vector $|Z|$ = 512 across all the datasets. An overview of the training procedure can be seen in Algorithm 1.

4.1. Qualitative Comparison

We compare our method with existing methods on Paris [4], CelebA-HQ [9], Places2 [30], and ImageNet [17], respectively. As shown in Fig. 4, 5 and 6, our model produces various plausible results by sampling from the latent space of instance data. Our model can also be trained for arbitrary region completion.

4.2. Quantitative comparisons

We quantitatively compare our model with existing single-solution inpainting methods and multiple-solution inpainting methods, respectively.

Comparisons with Single-solution Inpainting Methods. Given a masked image input, our model can generate multiple and diverse reasonable solutions, while the single-solution inpainting methods can only generate one result. For fair comparison, we select top 5 samples (ranked by the discriminator) generated by our model to evaluate the average metric value.

In order to better measure the quality of the restored image, we introduce a Modified Inception Score (MIS), which is modified on the basis of Inception Score (IS) [18]. As we known from [18], IS is defined as

$$IS = \exp(H(p(y)) - \mathbb{E}_x H(p(y|x))) \tag{10}$$

where $H(\cdot)$ is an entropy function, $p(y)$ denotes the marginal probability function of image category distribution, $p(y|x)$ denotes the probability function of category distribution of the given image x. $H(p(y))$ is used to measure the diversity of generated image categories. However, in the scenario of image inpainting, there is only one kind of image. In addition, $p(y)$ often needs a lot of images to make sense. So we remove the item $H(p(y))$. The MIS is defined as

$$MIS = \exp(\mathbb{E}_{x\sim p_g} \sum_i (p(y_i|x) \log p(y_i|x))) \tag{11}$$

where p_g is the model distribution of image x. y is the label predicted by pre-trained classification models. The larger the value of MIS is, the better the image quality is. The maximum value of MIS is 1. Compared with IS [18], MIS is more suitable for evaluating image quality in the scenario of image inpainting. The comparison is conducted on CelebA-HQ 1000 test images, with quantitative measures of mean l_1 loss, peak signal-to-noise ratio (PSNR), structural similarity (SSIM), IS and MIS as shown in the Table 1. We used a 128 × 128 mask in the center.

Comparisons with Multiple-solution Inpainting Methods. We evaluate diversity scores using the LPIPS metric reported in [31]. The average score is calculated between 5K pairs generated from a sampling of 1K center-masked images. I_{out} and $I_{out(m)}$ are the full output and mask-region output, respectively. Our method obtains relatively higher diversity scores than other existing methods as shown in Table 2.

User Study. To better evaluate and compare with other methods, we randomly select 600 images from the CelebA-HQ [9] test set and randomly distribute these images to 20

Table 1: Quantitative comparison with the state-of-the-art approaches on CelebA-HQ dataset. Our model was trained on regular holes. † Lower is better. ⊎ Higher is better.

Method	PSNR⊎	SSIM⊎	IS⊎	MIS⊎	l_1 loss (%) †
SF [16]	25.9794	0.8835	2.8850	0.0156	1.69
CA [26]	24.2377	0.8671	2.8674	0.0151	2.35
CE [15]	26.1634	0.8910	2.8851	0.0149	25.20
CSA [13]	26.1920	**0.9021**	2.7997	0.0163	1.68
SN [23]	26.0732	0.8671	2.9981	0.0170	1.81
PN [29]	24.4229	0.8692	3.0097	0.0170	2.17
UCTGAN with noise	25.9700	0.8752	2.9012	0.0174	1.61
UCTGAN without attention	26.0223	0.8732	3.0011	0.0174	1.65
UCTGAN with attention	**26.3833**	0.8862	**3.0127**	**0.0178**	**1.51**

Table 2: Quantitative comparison of diversity with the state-of-the-art methods.

Method	LPIPS(I_{out})	LPIPS($I_{out(m)}$)
CVAE [19]	0.004	0.014
BG [31]	0.027	0.060
PN [29]	0.029	0.088
UCTGAN without $\mathcal{L}_{rec}^{l}$	0.017	0.032
UCTGAN with noise	0.029	0.062
UCTGAN	**0.030**	**0.092**

users. Each user is given 30 images with holes together with the inpainting results of PICNet (PN) [29] and ours. Each of them is asked to rank the results in non-increasing order (meaning they can say two results have similar quality). The statistics show that our model is ranked better most of time (71.15%) over PICNet (PN) [29].

4.3. Ablation Study

With and without cross semantic attention module. We train a complete UCTGAN on the CelebA-HQ dataset with cross semantic attention layer (called UCTGAN with attention) and one model that does not involve cross semantic attention layer (called UCTGAN without attention). Table 1 lists the evaluation results. From the results in Table 1, we can see that the cross semantic attention layer (UCTGAN with attention) improves image quality in several metrics such as MIS, IS and PSNR.

With and without guided instance. In order to test the effect of the manifold projection module, we replace the output of manifold projection module with the noise sampled from standard normal distribution. We train this model (called UCTGAN with noise) on the CelebA-HQ dataset. The evaluation results are shown in Table 1 and 2, we can see that the instance guided method (UCTGAN with attention) improves image quality and diversity.

With and without low dimension loss $\mathcal{L}_{rec}^{l}$. The low dimension loss $\mathcal{L}_{rec}^{l}$ is used to ensure that the instance image and the corresponding repaired image are projected onto the same low-dimensional manifold, which realizes one-to-one mapping between instance image space and conditional completion image space. In order to measure the effect of loss $\mathcal{L}_{rec}^{l}$ on the diversity of the generated repaired images, we train the model without $\mathcal{L}_{rec}^{l}$ (UCTGAN without $\mathcal{L}_{rec}^{l}$) on the CelebA-HQ dataset. The evaluation results are shown in Table 2, and we can see that the low dimension loss $\mathcal{L}_{rec}^{l}$ greatly improves image diversity.

5. Conclusion

In this paper, we propose a conditional image-to-image translation network (UCTGAN) to generate multiple and diverse semantically reasonable and visually realistic results for image inpainting. Our method learns the conditional distribution by unsupervised cross-space translation. Specifically, the proposed network realizes one-to-one mapping between instance image space and conditional completion image space, which can significantly reduce the possibility of mode collapse and improve the diversity of restored images. We also introduce a new cross semantic attention layer that exploits the long-range dependencies between the known parts and the completed parts, which improves realism and appearance consistency. As for future work, we plan to extend our method to other tasks, such as diverse intra-domain image generation based on instance images, diverse image super-resolution and diverse text-to-image generation.

Acknowledgments. This work was supported in part by the Zhejiang science and technology program (No:2019C03137), Zhejiang fund project (No:LGF18F020006, LY19F020049), and the key scientific research base for digital conservation of cave temples in Zhejiang university, state administration for cultural heritage of china.

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics (ToG)*, volume 28, page 24. ACM, 2009.

[2] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, and Stanley Osher. Simultaneous structure and texture image inpainting. *IEEE transactions on image processing*, 12(8):882–889, 2003.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[4] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. What makes paris look like paris? *Acm Transactions on Graphics*, 31(4):1–9, 2012.

[5] A Efros and W Freeman. Image quilting for texture synthesis. In *Proceedings of SIGGRAPH 2001*, volume 341, page 346, 2001.

[6] 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*, pages 2672–2680, 2014.

[7] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[8] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):107, 2017.

[9] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *IEEE Conference on Learning Representations*, 2018.

[10] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2018.

[11] Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics. In *null*, page 305. IEEE, 2003.

[12] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[13] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. *International Conference on Computer Vision*, 2019.

[14] David G Lowe et al. Object recognition from local scale-invariant features. In *iccv*, volume 99, pages 1150–1157, 1999.

[15] 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*, pages 2536–2544, 2016.

[16] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 181–190, 2019.

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[18] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[19] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015.

[20] 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. In *The British Machine Vision Conference*, 2018.

[21] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34, 2004.

[22] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 331–340, 2018.

[23] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–17, 2018.

[24] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5485–5493, 2017.

[25] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018.

[26] 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*, pages 5505–5514, 2018.

[27] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1486–1494, 2019.

[28] Haoran Zhang, Zhenzhen Hu, Changzhi Luo, Wangmeng Zuo, and Meng Wang. Semantic image inpainting with progressive generative networks. In *2018 ACM Multimedia*

Conference on Multimedia Conference, pages 1939–1947. ACM, 2018.

[29] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019.

[30] 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.

[31] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017.