Pluralistic Image Completion

Chuanxia Zheng Tat-Jen Cham Jianfei Cai
School of Computer Science and Engineering
Nanyang Technological University, Singapore
{chuanxia001,astjcham,asjfcai}@ntu.edu.sg

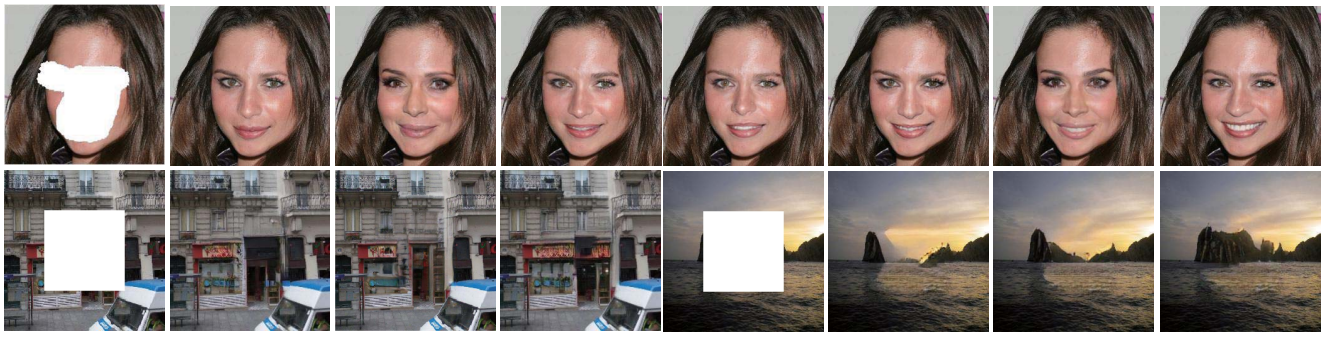


Figure 1. Example completion results of our method on images of a face, a building, and natural scenery with various masks (missing regions shown in white). For each group, the masked input image is shown left, followed by sampled results from our model without any post-processing. The results are diverse and plausible. (Zoom in to see the details.)

Abstract

Most image completion methods produce only one result for each masked input, although there may be many reasonable possibilities. In this paper, we present an approach for ***pluralistic image completion*** *– the task of generating multiple and diverse plausible solutions for image completion. A major challenge faced by learning-based approaches is that usually only one ground truth training instance per label. As such, sampling from conditional VAEs still leads to minimal diversity. To overcome this, we propose a novel and probabilistically principled framework with two parallel paths. One is a reconstructive path that utilizes the only one given ground truth to get prior distribution of missing parts and rebuild the original image from this distribution. The other is a generative path for which the conditional prior is coupled to the distribution obtained in the reconstructive path. Both are supported by GANs. We also introduce a new short+long term attention layer that exploits distant relations among decoder and encoder features, improving appearance consistency. When tested on datasets with buildings (Paris), faces (CelebA-HQ), and natural images (ImageNet), our method not only generated higher-quality completion results, but also with multiple and diverse plausible outputs.*

1. Introduction

Image completion is a highly subjective process. Supposing you were shown the various images with missing regions in fig. 1, what would you *imagine* to be occupying these holes? Bertalmio *et al*. [4] related how expert conservators would inpaint damaged art by: 1) imagining the semantic content to be filled based on the overall scene; 2) ensuring structural continuity between the masked and unmasked regions; and 3) filling in visually realistic content for missing regions. Nonetheless, each expert will independently end up creating substantially different details, even if they may universally agree on high-level semantics, such as general placement of eyes on a damaged portrait.

Based on this observation, our main goal is thus to generate *multiple* and *diverse* plausible results when presented with a masked image — in this paper we refer to this task as **pluralistic image completion** (depicted in fig. 1). This is as opposed to approaches that attempt to generate only a single "guess" for missing parts.

Early image completion works [4, 7, 5, 8, 3, 13] focus only on steps 2 and 3 above, by assuming that gaps should be filled with similar content to that of the background. Although these approaches produced high-quality texture-consistent images, they cannot capture global semantics and hallucinate new content for large holes. More recently, some learning-based image completion methods [29, 14, 39, 40, 42, 24, 38] were proposed that infer seman-

tic content (as in step 1). These works treated completion as a conditional generation problem, where the input-to-output mapping is one-to-many. However, these prior works are limited to generate only one "optimal" result, and do not have the capacity to generate a variety of semantically meaningful results.

To obtain a diverse set of results, some methods utilize conditional variational auto-encoders (CVAE) [34, 37, 2, 10], a conditional extension of VAE [19], which explicitly code a distribution that can be sampled. However, specifically for an image completion scenario, the standard single-path formulation usually leads to grossly underestimating variances. This is because when *the condition label is itself a partial image*, the number of instances in the training data that match each label is *typically only one*. Hence the estimated conditional distributions tend to have very limited variation since they were trained to reconstruct the single ground truth. This is further elaborated on in section 3.1.

An important insight we will use is that *partial images*, as a superset of full images, may also be considered as generated from *a latent space with smooth prior distributions*. This provides a mechanism for alleviating the problem of having scarce samples per conditional partial image. To do so, we introduce a new image completion network with two parallel but linked training pipelines. The first pipeline is a VAE-based reconstructive path that not only utilizes the full instance ground truth (*i.e*. both the visible partial image, as well as its complement — the hidden partial image), but also imposes smooth priors for the latent space of complement regions. The second pipeline is a generative path that predicts the latent prior distribution for the missing regions conditioned on the visible pixels, from which can be sampled to generate diverse results. The training process for the latter path does *not* attempt to steer the output towards reconstructing the instance-specific hidden pixels at all, instead allowing the reasonableness of results be driven by an auxiliary discriminator network [11]. This leads to substantially great variability in content generation. We also introduce an enhanced short+long term attention layer that significantly increases the quality of our results.

We compared our method with existing state-of-the-art approaches on multiple datasets. Not only can higher-quality completion results be generated using our approach, it also presents multiple diverse solutions.

The main contributions of this work are:

1. A probabilistically principled framework for image completion that is able to maintain much higher sample diversity as compared to existing methods;

2. A new network structure with two parallel training paths, which trades off between reconstructing the original training data (with loss of diversity) and maintaining the variance of the conditional distribution;

3. A novel self-attention layer that exploits short+long term context information to ensure appearance consistency in the image domain, in a manner superior to purely using GANs; and

4. We demonstrate that our method is able to complete the same mask with multiple plausible results that have substantial diversity, such as those shown in figure 1.

2. Related Work

Existing work on image completion either uses information from within the input image [4, 5, 3], or information from a large image dataset [12, 29, 42]. Most approaches will generate only one result per masked image.

Intra-Image Completion Traditional intra-image completion, such as diffusion-based methods [4, 1, 22] and patch-based methods [5, 7, 8, 3], assume image holes share similar content to visible regions; thus they would directly match, copy and realign the background patches to complete the holes. These methods perform well for background completion, *e.g*. for object removal, but cannot hallucinate unique content not present in the input images.

Inter-Image Completion To generate semantically new content, inter-image completion borrows information from a large dataset. Hays and Efros [12] presented an image completion method using millions of images, in which the image most similar to the masked input is retrieved, and corresponding regions are transferred. However, this requires a high contextual match, which is not always available. Recently, learning-based approaches were proposed. Initial works [20, 30] focused on small and thin holes. Context encoders (CE) [29] handled 64×64-sized holes using GANs [11]. This was followed by several CNN-based methods, which included combining global and local discriminators as adversarial loss [14], identifying closest features in the latent space of masked images [40], utilizing semantic labels to guide the completion network [36], introducing additional face parsing loss for face completion [23], and designing particular convolutions to address irregular holes [24, 41]. A common drawback of these methods is that they often create distorted structures and blurry textures inconsistent with the visible regions, especially for large holes.

Combined Intra- and Inter-Image Completion To overcome the above problems, Yang *et al*. [39] proposed multi-scale neural patch synthesis, which generates high-frequency details by copying patches from mid-layer features. However, this optimization is computational costly. More recently, several works [42, 38, 35] exploited spatial attention [16, 46] to get high-frequency details. Yu *et al*. [42] presented a contextual attention layer to copy similar features from visible regions to the holes. Yan *et al*. [38] and Song *et al*. [35] proposed PatchMatch-like ideas on feature domain. However, these methods identify similar fea-

tures by comparing features of holes and features of visible regions, which is somewhat contradictory as feature transfer is unnecessary when two features are very similar, but when needed the features are too different to be matched easily. Furthermore, distant information is not used for new content that differs from visible regions. Our model will solve this problem by extending self-attention [43] to harness abundant context.

Image Generation Image generation has progressed significantly using methods such as VAE [19] and GANs [11]. These have been applied to conditional image generation tasks, such as image translation [15], synthetic to realistic [44], future prediction [27], and 3D models [28]. Perhaps most relevant are conditional VAEs (CVAE) [34, 37] and CVAE-GAN [2], but these were not specially targeted for image completion. CVAE-based methods are most useful when the conditional labels are few and discrete, and there are sufficient training instances per label. Some recent work utilizing these in image translation can produce diverse output [47, 21], but in such situations the condition-to-sample mappings are more local (*e.g.* pixel-to-pixel), and only change the visual appearance. This is untrue for image completion, where the conditional label is itself the masked image, with only one training instance of the original holes. In [6], different outputs were obtained for face completion by specifying facial attributes (*e.g.* smile), but this method is very domain specific, requiring targeted attributes.

3. Approach

Suppose we have an image, originally $\mathbf{I}_g$, but degraded by a number of missing pixels to become $\mathbf{I}_m$ (the *masked partial image*) comprising the observed / visible pixels. We also define $\mathbf{I}_c$ as its *complement partial image* comprising the ground truth hidden pixels. Classical image completion methods attempt to reconstruct the ground truth unmasked image $\mathbf{I}_g$ in a deterministic fashion from $\mathbf{I}_m$ (see fig. 2 "Deterministic"). This results in only a single solution. In contrast, our goal is to *sample* from $p(\mathbf{I}_c|\mathbf{I}_m)$.

3.1. Probabilistic Framework

In order to have a distribution to sample from, a current approach is to employ the CVAE [34] which estimates a parametric distribution over a latent space, from which sampling is possible (see fig. 2 "CVAE"). This involves a variational lower bound of the conditional log-likelihood of observing the training instances:

$$\begin{aligned}\log p(\mathbf{I}_c|\mathbf{I}_m) \geq &- \mathrm{KL}(q_\psi(\mathbf{z}_c|\mathbf{I}_c,\mathbf{I}_m)||p_\phi(\mathbf{z}_c|\mathbf{I}_m)) \\ &+ \mathbb{E}_{q_\psi(\mathbf{z}_c|\mathbf{I}_c,\mathbf{I}_m)}[\log p_\theta(\mathbf{I}_c|\mathbf{z}_c,\mathbf{I}_m)] \quad (1)\end{aligned}$$

where $\mathbf{z}_c$ is the latent vector, $q_\psi(\cdot|\cdot)$ the posterior importance sampling function, $p_\phi(\cdot|\cdot)$ the conditional prior, $p_\theta(\cdot|\cdot)$ the likelihood, with ψ, ϕ and θ being the deep network parameters of their corresponding functions. This lower bound is maximized w.r.t. all parameters.

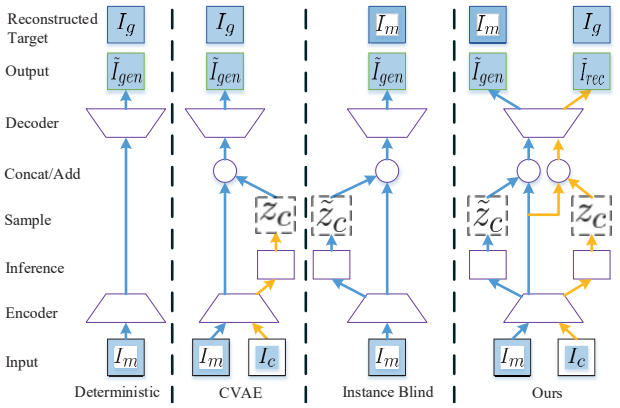


Figure 2. Completion strategies given masked input. (Deterministic) structure directly predicts the ground truth instance. (CVAE) adds in random sampling to diversify the output. (Instance Blind) only matches the visible parts, but training is unstable. (Ours) uses a generative path during testing, but is guided by a parallel reconstructive path during training. Yellow path is used for training.

For our purposes, the chief difficulty of using CVAE [34] directly is that the high DoF networks of $q_\psi(\cdot|\cdot)$ and $p_\phi(\cdot|\cdot)$ are not easily separable in (1) with the KL distance easily driven towards zero, and is approximately equivalent to maximizing $\mathbb{E}_{p_\phi(\mathbf{z}_c|\mathbf{I}_m)}[\log p_\theta(\mathbf{I}_c|\mathbf{z}_c,\mathbf{I}_m)]$ (the "GSNN" variant in [34]). This consequently learns a delta-like prior of $p_\phi(\mathbf{z}_c|\mathbf{I}_m) \rightarrow \delta(\mathbf{z}_c - \mathbf{z}_c^*)$, where $\mathbf{z}_c^*$ is the maximum latent likelihood point of $p_\theta(\mathbf{I}_c|\cdot,\mathbf{I}_m)$. While this low variance prior may be useful in estimating a single solution, sampling from it will lead to *negligible diversity* in image completion results (as seen in fig. 9). When the CVAE variant of [37], which has a fixed latent prior, is used instead, the network learns to ignore the latent sampling and directly estimates $\mathbf{I}_c$ from $\mathbf{I}_m$, also resulting in a single solution. This is due to the image completion scenario when there is only one training instance per condition label, which is a partial image $\mathbf{I}_m$. Details are in the supplemental section B.1.

A possible way to diversify the output is to simply not incentivize the output to reconstruct the instance-specific $\mathbf{I}_g$ during training, only needing it to fit in with the training set distribution as deemed by an learned adversarial discriminator (see fig. 2 "Instance Blind"). However, this approach is unstable, especially for large and complex scenes [35].

Latent Priors of Holes In our approach, we require that missing partial images, as a superset of full images, *to also arise from a latent space distribution*, with a smooth prior of $p(\mathbf{z}_c)$. The variational lower bound is:

$$\begin{aligned}\log p(\mathbf{I}_c) \geq &- \mathrm{KL}(q_\psi(\mathbf{z}_c|\mathbf{I}_c)||p(\mathbf{z}_c)) \\ &+ \mathbb{E}_{q_\psi(\mathbf{z}_c|\mathbf{I}_c)}[\log p_\theta(\mathbf{I}_c|\mathbf{z}_c)] \quad (2)\end{aligned}$$

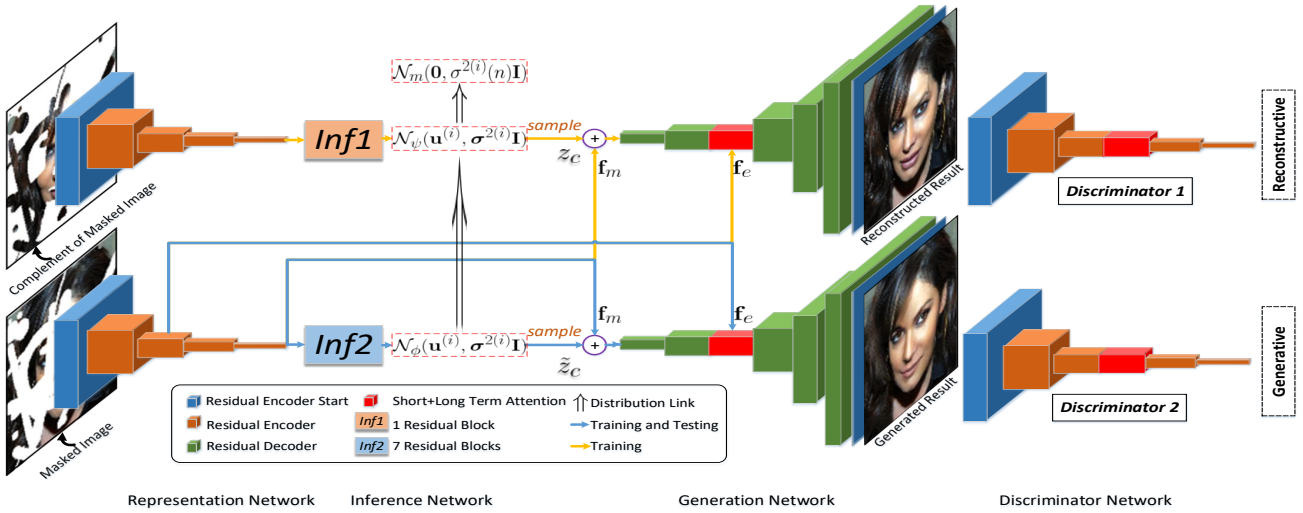


Figure 3. Overview of our architecture with two parallel pipelines. The **reconstructive** pipeline (yellow line) combines information from $\mathbf{I}_m$ and $\mathbf{I}_c$, which is used only for training. The **generative** pipeline (blue line) infers the conditional distribution of hidden regions, that can be sampled during testing. Both representation and generation networks share identical weights.

where in [19] the prior is set as $p(\mathbf{z}_c) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. However, we can be more discerning when it comes to partial images since they have different numbers of pixels. *A missing partial image $\mathbf{z}_c$ with more pixels (larger holes) should have greater latent prior variance than a missing partial image $\mathbf{z}_c$ with fewer pixels (smaller holes).* Hence we generalize the prior $p(\mathbf{z}_c) = \mathcal{N}_m(\mathbf{0}, \sigma^2(n)\mathbf{I})$ to adapt to the number of pixels n.

Prior-Conditional Coupling Next, we combine the latent priors into the conditional lower bound of (1). This can be done by assuming $\mathbf{z}_c$ is much more closely related to $\mathbf{I}_c$ than to $\mathbf{I}_m$, so $q_\psi(\mathbf{z}_c|\mathbf{I}_c, \mathbf{I}_m) \approx q_\psi(\mathbf{z}_c|\mathbf{I}_c)$. Updating (1):

$$\log p(\mathbf{I}_c|\mathbf{I}_m) \geq -\mathrm{KL}(q_\psi(\mathbf{z}_c|\mathbf{I}_c)||p_\phi(\mathbf{z}_c|\mathbf{I}_m)) + \mathbb{E}_{q_\psi(\mathbf{z}_c|\mathbf{I}_c)}[\log p_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)] \quad (3)$$

However, unlike in (1), notice that $q_\psi(\mathbf{z}_c|\mathbf{I}_c)$ *is no longer freely learned during training, but is tied to its presence in (2).* Intuitively, the learning of $q_\psi(\mathbf{z}_c|\mathbf{I}_c)$ is regularized by the prior $p(\mathbf{z}_c)$ in (2), while the learning of the conditional prior $p_\phi(\mathbf{z}_c|\mathbf{I}_m)$ is in turn regularized by $q_\psi(\mathbf{z}_c|\mathbf{I}_c)$ in (3).

Reconstruction vs Creative Generation One issue with (3) is that the sampling is taken from $q_\psi(\mathbf{z}_c|\mathbf{I}_c)$ during training, but is not available during testing, whereupon sampling must come from $p_\phi(\mathbf{z}_c|\mathbf{I}_m)$ which may not be adequately learned for this role. In order to mitigate this problem, we modify (3) to have a blend of formulations *with and without importance sampling*. So, with simplified notation:

$$\log p(\mathbf{I}_c|\mathbf{I}_m) \geq \lambda \left\{\mathbb{E}_{q_\psi}[\log p^r_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)] - \mathrm{KL}(q_\psi||p_\phi)\right\} + (1-\lambda)\,\mathbb{E}_{p_\phi}[\log p^g_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)] \quad (4)$$

where $0 \leq \lambda \leq 1$ is implicitly set by training loss coefficients in section 3.3. When sampling from the importance function $q_\psi(\cdot|\mathbf{I}_c)$, the full training instance is available and we formulate the likelihood $p^r_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)$ to be focused on *reconstructing* $\mathbf{I}_c$. Conversely, when sampling from the learned conditional prior $p_\phi(\cdot|\mathbf{I}_m)$ which does not contain $\mathbf{I}_c$, we facilitate *creative generation* by having the likelihood model $p^g_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) \cong \ell^g_\theta(\mathbf{z}_c, \mathbf{I}_m)$ be *independent of the original instance* of $\mathbf{I}_c$. Instead it only encourages generated samples to fit in with the overall training distribution.

Our overall training objective may then be expressed as jointly maximizing the lower bounds in (2) and (4), with the likelihood in (2) unified to that in (4) as $p_\theta(\mathbf{I}_c|\mathbf{z}_c) \cong p^r_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)$. See the supplemental section B.2.

3.2. Dual Pipeline Network Structure

This formulation is implemented as our dual pipeline framework, shown in fig. 3. It consists of two paths: the upper reconstructive path uses information from the whole image, *i.e.* $\mathbf{I}_g = \{\mathbf{I}_c, \mathbf{I}_m\}$, while the lower generative path only uses information from visible regions $\mathbf{I}_m$. Both representation and generation networks share identical weights. Specifically:

- For the upper reconstructive path, the complement partial image $\mathbf{I}_c$ is used to infer the importance function $q_\psi(\cdot|\mathbf{I}_c) = \mathcal{N}_\psi(\cdot)$ during training. The sampled latent vector $\mathbf{z}_c$ thus contains information of the missing regions, while the conditional feature $\mathbf{f}_m$ encodes the information of the visible regions. Since there is sufficient information, the loss function in this path is geared towards reconstructing the original image $\mathbf{I}_g$.
- For the lower generative path, which is also the test

path, the latent distribution of the holes $\mathbf{I}_c$ is inferred based only on the visible $\mathbf{I}_m$. This would be significantly less accurate than the inference in the upper path. Thus the reconstruction loss is only targeted at the visible regions $\mathbf{I}_m$ (via $\mathbf{f}_m$).

- In addition, we also utilize adversarial learning networks on both paths, which ideally ensure that the full synthesized data fit in with the training set distribution, and empirically leads to higher quality images.

3.3. Training Loss

Various terms in (2) and (4) may be more conventionally expressed as loss functions. Jointly maximizing the lower bounds is then minimizing a total loss $\mathcal{L}$, which consists of three groups of component losses:

$$\mathcal{L} = \alpha_{\text{KL}}(\mathcal{L}^r_{\text{KL}} + \mathcal{L}^g_{\text{KL}}) + \alpha_{\text{app}}(\mathcal{L}^r_{\text{app}} + \mathcal{L}^g_{\text{app}}) + \alpha_{\text{ad}}(\mathcal{L}^r_{\text{ad}} + \mathcal{L}^g_{\text{ad}}) \tag{5}$$

where the $\mathcal{L}_{\text{KL}}$ group regularizes consistency between pairs of distributions in terms of KL divergences, the $\mathcal{L}_{\text{app}}$ group encourages appearance matching fidelity, and while the $\mathcal{L}_{\text{ad}}$ group forces sampled images to fit in with the training set distribution. Each of the groups has a separate term for the reconstructive and generative paths.

Distributive Regularization The typical interpretation of the KL divergence term in a VAE is that it regularizes the learned importance sampling function $q_\psi(\cdot|\mathbf{I}_c)$ to a fixed latent prior $p(\mathbf{z}_c)$. Defining as Gaussians, we get:

$$\mathcal{L}^{r,(i)}_{\text{KL}} = -\text{KL}(q_\psi(\mathbf{z}|I^{(i)}_c)||\mathcal{N}_m(\mathbf{0}, \sigma^2(n)\mathbf{I})) \tag{6}$$

For the generative path, the appropriate interpretation is *reversed*: the learned conditional prior $p_\phi(\cdot|\mathbf{I}_m)$, also a Gaussian, is regularized to $q_\psi(\cdot|\mathbf{I}_c)$.

$$\mathcal{L}^{g,(i)}_{\text{KL}} = -\text{KL}(q_\psi(\mathbf{z}|I^{(i)}_c))||p_\phi(\mathbf{z}|I^{(i)}_m))) \tag{7}$$

Note that the conditional prior only uses $\mathbf{I}_m$, while the importance function has access to the hidden $\mathbf{I}_c$.

Appearance Matching Loss The likelihood term $p^r_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)$ may be interpreted as probabilistically encouraging appearance matching to the hidden $\mathbf{I}_c$. However, our framework also auto-encodes the visible $\mathbf{I}_m$ deterministically, and the loss function needs to cater for this reconstruction. As such, the per-instance loss here is:

$$\mathcal{L}^{r,(i)}_{\text{app}} = ||I^{(i)}_{\text{rec}} - I^{(i)}_g||_1 \tag{8}$$

where $I^{(i)}_{\text{rec}}$=$G(z_c, f_m)$ and $I^{(i)}_g$ are the reconstructed and original full images respectively. In contrast, for the generative path we ignore instance-specific appearance matching for $\mathbf{I}_c$, and only focus on reconstructing $\mathbf{I}_m$ (via $\mathbf{f}_m$):

$$\mathcal{L}^{g,(i)}_{\text{app}} = ||M * (I^{(i)}_{\text{gen}} - I^{(i)}_g)||_1 \tag{9}$$

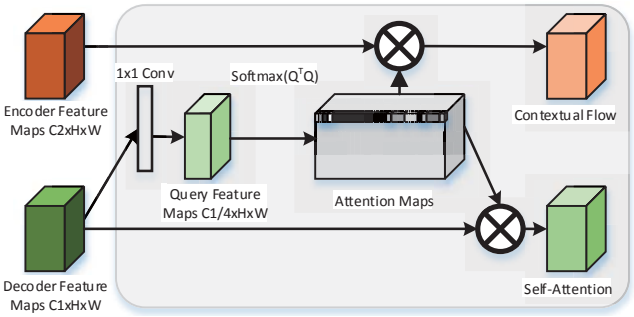


Figure 4. Our short+long term attention layer. The attention map is directly computed on the decoder features. After obtaining the self-attention scores, we use these to compute self-attention on decoder features, as well as contextual flow on encoder features.

where $I^{(i)}_{\text{gen}}$=$G(\tilde{z}_c, f_m)$ is the generated image from the $\tilde{z}_c$ sample, and M is the binary mask selecting visible pixels.

Adversarial Loss The formulation of $p^r_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)$ and the instance-blind $p^g_\theta(\mathbf{I}_c|\tilde{\mathbf{z}}_c, \mathbf{I}_m)$ also incorporates the use of adversarially learned discriminators D_1 and D_2 to judge whether the generated images fit into the training set distribution. Inspired by [2], we use a mean feature match loss in the reconstructive path for the generator,

$$\mathcal{L}^{r,(i)}_{\text{ad}} = ||f_{D_1}(I^{(i)}_{\text{rec}}) - f_{D_1}(I^{(i)}_g)||_2 \tag{10}$$

where $f_{D_1}(\cdot)$ is the feature output of the final layer of D_1. This encourages the original and reconstructed features in the discriminator to be close together. Conversely, the adversarial loss in the generative path for the generator is:

$$\mathcal{L}^{g,(i)}_{\text{ad}} = [D_2(I^{(i)}_{\text{gen}}) - 1]^2 \tag{11}$$

This is based on the generator loss in LSGAN [26], which performs better than the original GAN loss [11] in our scenario. The discriminator loss for both D_1 and D_2 is also based on LSGAN.

3.4. Short+Long Term Attention

Extending beyond the Self-Attention GAN [43], we propose not only to use the self-attention map within a decoder layer to harness *distant spatial context*, but also to further capture *feature-feature context* between encoder and decoder layers. Our *key novel insight* is: doing so would allow the network a choice of attending to the finer-grained features in the encoder or the more semantically generative features in the decoder, depending on circumstances.

Our proposed structure is shown in fig. 4. We first calculate the self-attention map from the features $\mathbf{f}_d$ of a decoder middle layer, using the attention score of:

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N}\exp(s_{ij})}, \text{where } s_{ij} = Q(f_{di})^T Q(f_{dj}), \tag{12}$$

N is the number of pixels, $Q(\mathbf{f}_d)$=$\mathbf{W}_q\mathbf{f}_d$, and $\mathbf{W}_q$ is a 1x1 convolution filter. This leads to the short-term intra-layer attention feature (**self-attention** in fig. 4) and the output $\mathbf{y}_d$:

$$c_{dj} = \sum_{i=1}^{N} \beta_{j,i} f_{di} \ , \qquad \mathbf{y}_d = \gamma_d \mathbf{c}_d + \mathbf{f}_d \tag{13}$$

where, following [43], we use a scale parameter γ_d to balance the weights between $\mathbf{c}_d$ and $\mathbf{f}_d$. The initial value of γ_d is set to zero. In addition, for attending to features $\mathbf{f}_e$ from an encoder layer, we have a long-term inter-layer attention feature (**contextual flow** in fig. 4) and the output $\mathbf{y}_e$:

$$c_{ej} = \sum_{i=1}^{N} \beta_{j,i} f_{ei} \ , \qquad \mathbf{y}_e = \gamma_e (1 - \mathbf{M}) \mathbf{c}_e + \mathbf{M}\mathbf{f}_e \tag{14}$$

As before, a scale parameter γ_e is used to combine the encoder feature $\mathbf{f}_e$ and the attention feature $\mathbf{c}_e$. However, unlike the decoder feature $\mathbf{f}_d$ which has information for generating a full image, the encoder feature $\mathbf{f}_e$ only represents visible parts $\mathbf{I}_m$. Hence, a binary mask $\mathbf{M}$ (holes=0) is used. Finally, both the short and long term attention features are aggregated and fed into further decoder layers.

4. Experimental Results

We evaluated our proposed model on four datasets including Paris [9], CelebA-HQ [25, 17], Places2 [45], and ImageNet [31] using the original training and test splits for those datasets. Since our model can generate multiple outputs, we sampled 50 images for each masked image, and chose the top 10 results based on the discriminator scores. We trained our models for both regular and irregular holes. For brevity, we refer to our method as **PICNet**. We provide PyTorch implementations and interactive demo.

4.1. Implementation Details

Our generator and discriminator networks are inspired by SA-GAN [43], but with several important modifications, including the short+long term attention layer. Furthermore, inspired by the growing-GAN [17], multi-scale output is applied to make the training faster.

The image completion network, implemented in Pytorch v0.4.0, contains 6M trainable parameters. During optimization, the weights of different losses are set to $\alpha_{\text{KL}} = \alpha_{\text{rec}}$=20, α_{ad}=1. We used Orthogonal Initialization [33] and the Adam solver [18]. All networks were trained from scratch, with a fixed learning rate of λ=10^{-4}. Details are in the supplemental section D.

4.2. Comparison with Existing Work

Quantitative Comparisons Quantitative evaluation is hard for the pluralistic image completion task, as our goal is to get diverse but reasonable solutions for one masked image. The original image is only one solution of many, and comparisons should not be made based on just this image.

However, just for the sake of obtaining quantitative measures, we will assume that one of our top 10 samples (ranked by the discriminator) will be close to the original ground truth, and select the single sample with the best balance of quantitative measures for comparison. The comparison is conducted on ImageNet $20,000$ test images, with quantitative measures of mean ℓ_1 loss, peak signal-to-noise ration (PSNR), total variation (TV), and Inception Score (IS) [32]. We used a 128×128 mask in the center.

Method	ℓ_1 loss	PSNR	TV loss	IS
GL [14]	15.32	19.36	13.97	24.31
CA [42]	13.57	19.22	19.55	**28.80**
PICNet-regular	**12.91**	**20.10**	**12.18**	24.90

Table 1. Quantitative comparison with state-of-the-art. For center masks, our model was trained on regular holes.

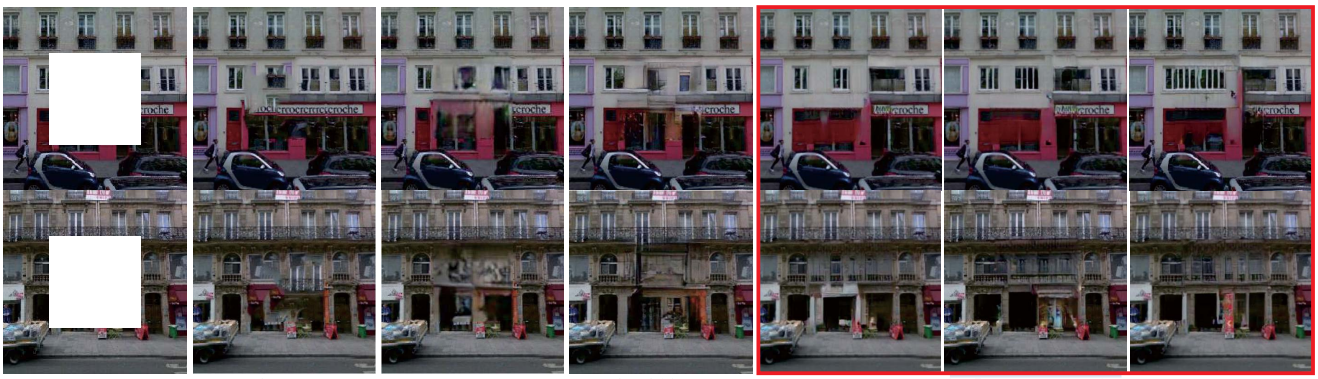
Qualitative Comparisons First, we show the results in fig. 5 on the Paris dataset [9]. For fair comparison among learning-based methods, we only compared with those trained on this dataset. PatchMatch [3] worked by copying similar patches from visible regions and obtained good results on this dataset with repetitive structures. Context Encoder (CE) [29] generated reasonable structures with blurry textures. Shift-Net [38] made improvements by feature copying. Compared to these, our model not only generated more natural images, but also with multiple solutions, *e.g.* different numbers of windows and varying door sizes.

Next, we evaluated our methods on CelebA-HQ face dataset, with fig. 6 showing examples with large regular holes to highlight the diversity of our output. Context Attention (CA) [42] generated reasonable completion for many cases, but for each masked input they were only able to generate a single result; furthermore, on some occasions, the single solution may be poor. Our model produced various plausible results by sampling from the latent space conditional prior.

Finally, we report the performance on the more challenging ImageNet dataset by comparing to the previous PatchMatch [3], CE [29], GL [14] and CA [42]. Different from the CE and GL models that were trained on the 100k subset of training images of ImageNet, our model is directly trained on original ImageNet training dataset with all images resized to 256×256. Visual results on a variety of objects from the validation set are shown in fig. 7. Our model was able to infer the content quite effectively.

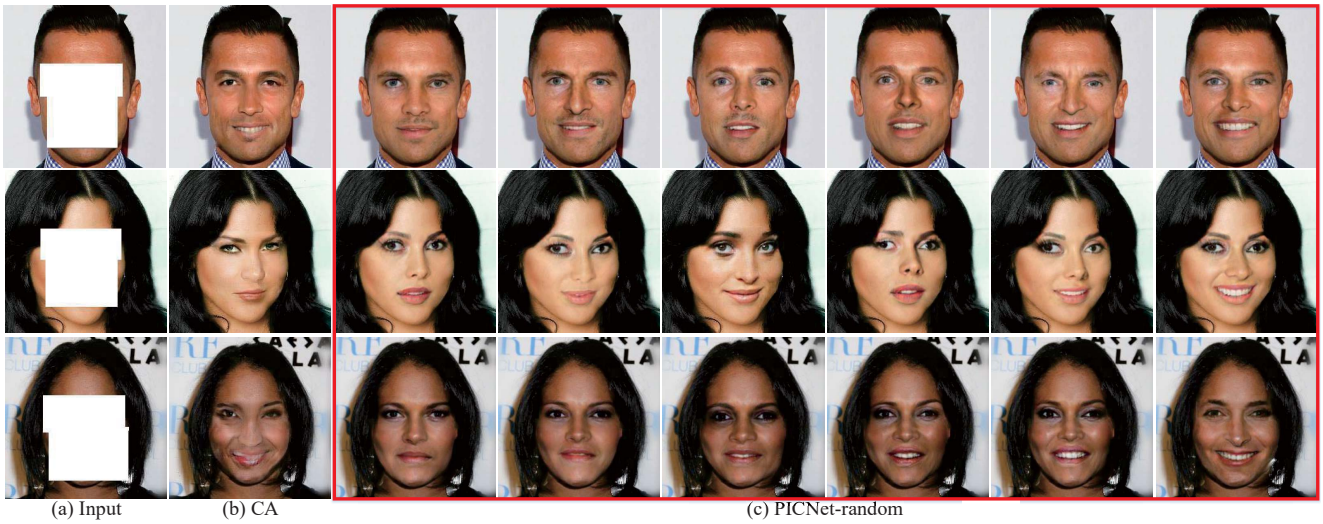
4.3. Ablation Study

Our PICNet vs CVAE vs "Instance Blind" vs BicycleGAN We investigated the influence of using our two-path



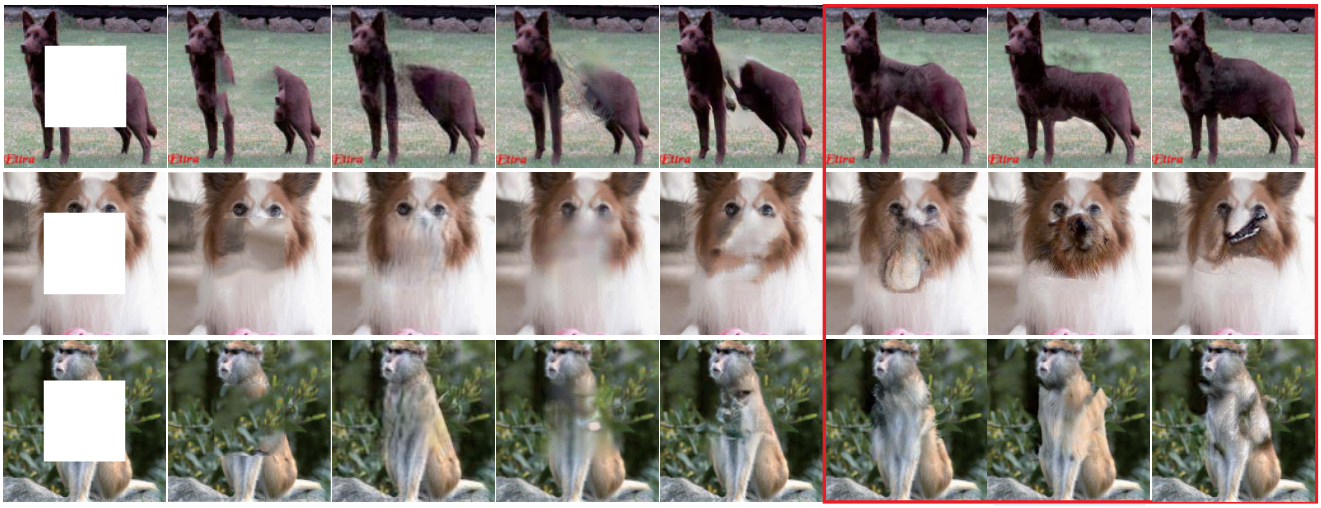
(a) Input (b) PM (c) CE (d) Shift-Net (e) PICNet-regular

Figure 5. Comparison of our model with PatchMatch(PM) [3], Context Encoder(CE) [29] and Shift-Net [38] on images taken from the Paris [9] test set for center region completion. Best viewed by zooming in.



(a) Input (b) CA (c) PICNet-random

Figure 6. Comparison of our model with Contextual Attention(CA) [42] on CelebA-HQ. Best viewed by zooming in.



(a) Input (b) PM (c) CE (d) GL (e) CA (f) PICNet-regular

Figure 7. Qualitative results and comparisons with the PM, CE, Global and Local(GL) [14] and CA on the ImageNet validation set.

Figure 8. Comparison of our Pluralistic model with BicycleGAN.

Figure 9. Comparison of training with different strategies: ours (top), CVAE (middle), instance-blind (bottom).

training structure in comparison to other variants such as the CVAE [34] and "instance blind" structures in fig. 2. We trained the three models using common parameters. As shown in fig. 9, for the CVAE, even after sampling from the latent prior distribution, the outputs were almost identical, as the conditional prior learned is narrowly centered at the maximum latent likelihood solution. As for "instance blind", if reconstruction loss was used only on visible pixels, the training may become unstable. If we used reconstruction loss on the full generated image, there is also little variation as the framework has likely learned to ignore the sampling and predicted a deterministic outcome purely from $\mathbf{I}_m$.

We also trained and tested BicycleGAN [47] for center masks. As is obvious in fig. 8, BicycleGAN is not directly suitable, leading to poor results or minimal variation.

Diversity Measure We computed diversity scores using the LPIPS metric reported in [46]. The average score is calculated between 50K pairs generated from a sampling of 1K center-masked images. $\mathbf{I}_{out}$ and $\mathbf{I}_{out(m)}$ are the full output and mask-region output, respectively. While [46] obtained relatively higher diversity scores (still lower than ours), most of their generated images look unnatural (fig. 8).

Short+Long Term Attention vs Contextual Attention We visualized our attention maps as in [43]. To compare to the contextual attention (CA) layer [42], we retrained CA on the Paris dataset via the authors' code, and used their publicly released face model. The CA attention maps are presented in their color-directional format. As shown in fig. 10, our short+long term attention layer borrowed features from different positions with varying attention weights, rather than directly copying similar features from just one visible position. For the building scene, CA's results were of similar high quality to ours, due to the repeated structures present. However for a face with a large mask, CA was unable to borrow features for the hidden content (*e.g.* mouth, eyes) from visible regions, with poor output. Our attention map is able to utilize both decoder features (which do not have masked parts) and encoder features as appropriate.

	Diversity (LPIPS)	
Method	$\mathbf{I}_{out}$	$\mathbf{I}_{out(m)}$
CVAE	0.004	0.014
Instance Blind	0.015	0.049
BicycleGAN [46]	0.027	0.060
PICNet-Pluralistic	**0.029**	**0.088**

Table 2. Quantitative comparisons of diversity.

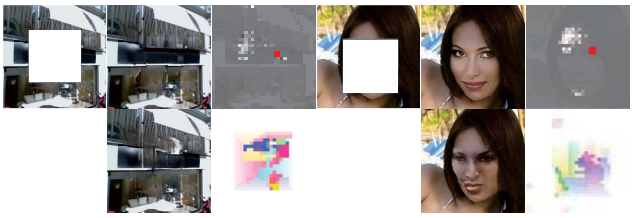


Figure 10. Visualization of attention map using different attention modules: ours (top), contextual attention (bottom). We highlight the most-attended regions for the query position (red point).

5. Conclusion

We proposed a novel dual pipeline training architecture for pluralistic image completion. Unlike existing methods, our framework can generate multiple diverse solutions with plausible content for a single masked input. The experimental results demonstrate this prior-conditional lower bound coupling is significant for conditional image generation. We also introduced an enhanced short+long term attention layer which improves realism. Experiments on a variety of datasets showed that our multiple solutions were diverse and of high-quality, especially for large holes.

Acknowledgements This research is supported by the BeingTogether Centre, a collaboration between Nanyang Technological University (NTU) Singapore and University of North Carolina (UNC) at Chapel Hill. The BeingTogether Centre is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centres in Singapore Funding Initiative. This research was also conducted in collaboration with Singapore Telecommunications Limited and partially supported by the Singapore Government through the Industry Alignment Fund - Industry Collaboration Projects Grant.

References

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

[2] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: Fine-grained image generation through asymmetric training. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2764–2773. IEEE, 2017.

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

[4] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424. ACM Press/Addison-Wesley Publishing Co., 2000.

[5] 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.

[6] Zeyuan Chen, Shaoliang Nie, Tianfu Wu, and Christopher G Healey. High resolution face completion with multiple controllable attributes via fully end-to-end progressive generative adversarial networks. *arXiv preprint arXiv:1801.07632*, 2018.

[7] Antonio Criminisi, Patrick Perez, and Kentaro Toyama. Object removal by exemplar-based inpainting. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE, 2003.

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

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

[10] S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

[11] 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.

[12] James Hays and Alexei A Efros. Scene completion using millions of photographs. In *ACM Transactions on Graphics (TOG)*, volume 26, page 4. ACM, 2007.

[13] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Image completion using planar structure guidance. *ACM Transactions on graphics (TOG)*, 33(4):129, 2014.

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

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017.

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

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

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

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[20] Rolf Köhler, Christian Schuler, Bernhard Schölkopf, and Stefan Harmeling. Mask-specific inpainting with deep neural networks. In *German Conference on Pattern Recognition*, pages 523–534. Springer, 2014.

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

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

[23] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5892–5900. IEEE, 2017.

[24] 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)*, September 2018.

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[26] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2813–2821. IEEE, 2017.

[27] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[28] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image

generation network for novel 3d view synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 702–711. IEEE, 2017.

[29] 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.

[30] Jimmy SJ Ren, Li Xu, Qiong Yan, and Wenxiu Sun. Shepard convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 901–909, 2015.

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

[32] 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.

[33] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[34] 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.

[35] Yuhang Song, Chao Yang, Zhe Lin, Xiaofeng Liu, Qin Huang, Hao Li, and CC Jay. Contextual-based image inpainting: Infer, match, and translate. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

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

[37] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision (ECCV)*, 2016.

[38] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[39] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017.

[40] 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 *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 6882–6890. IEEE, 2017.

[41] 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.

[42] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. *arXiv preprint arXiv:1801.07892*, 2018.

[43] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[44] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[45] 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, 2018.

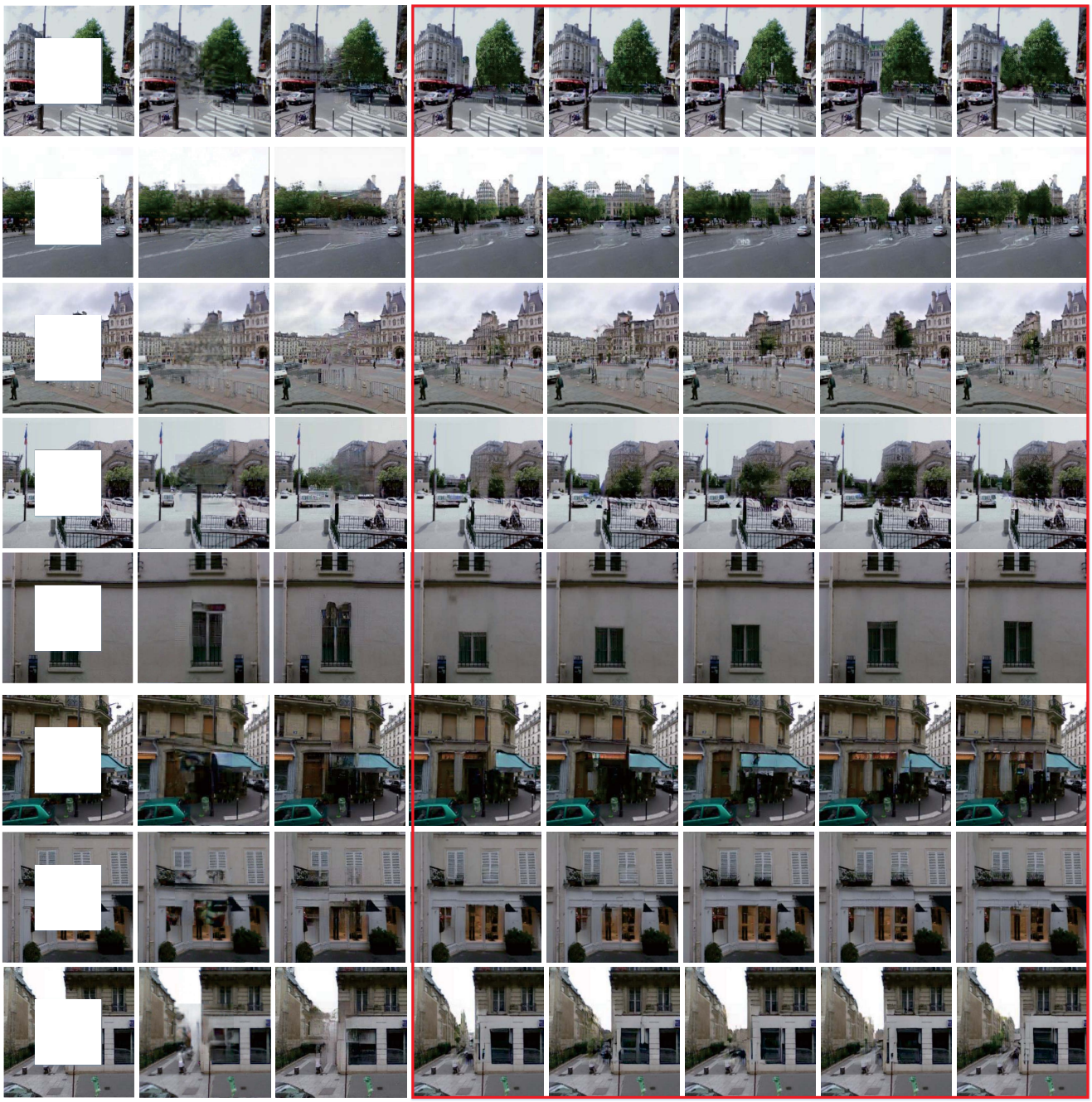
[46] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European conference on computer vision*, pages 286–301. Springer, 2016.

[47] 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.

A. Additional Examples

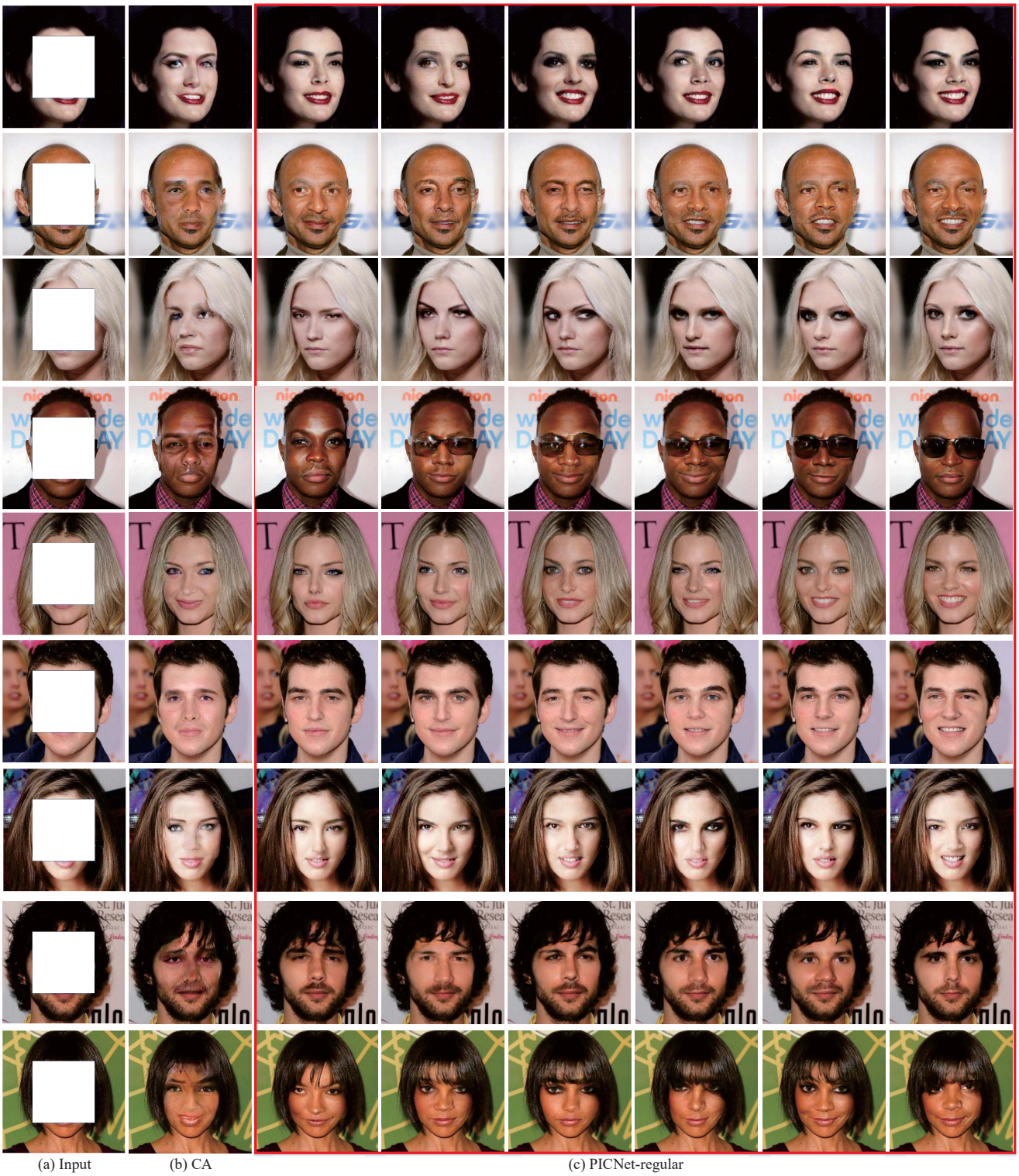
We first show our results on center hole completion, in relation to those from other methods trained on corresponding datasets. As for random irregular and regular holes, we simply present our results so that readers may appreciate the multiple diverse results we can get with differently sized and shaped holes. Finally, we show the interesting application on face editing.

A.1. Comparison with Existing Work on Center Hole Completion



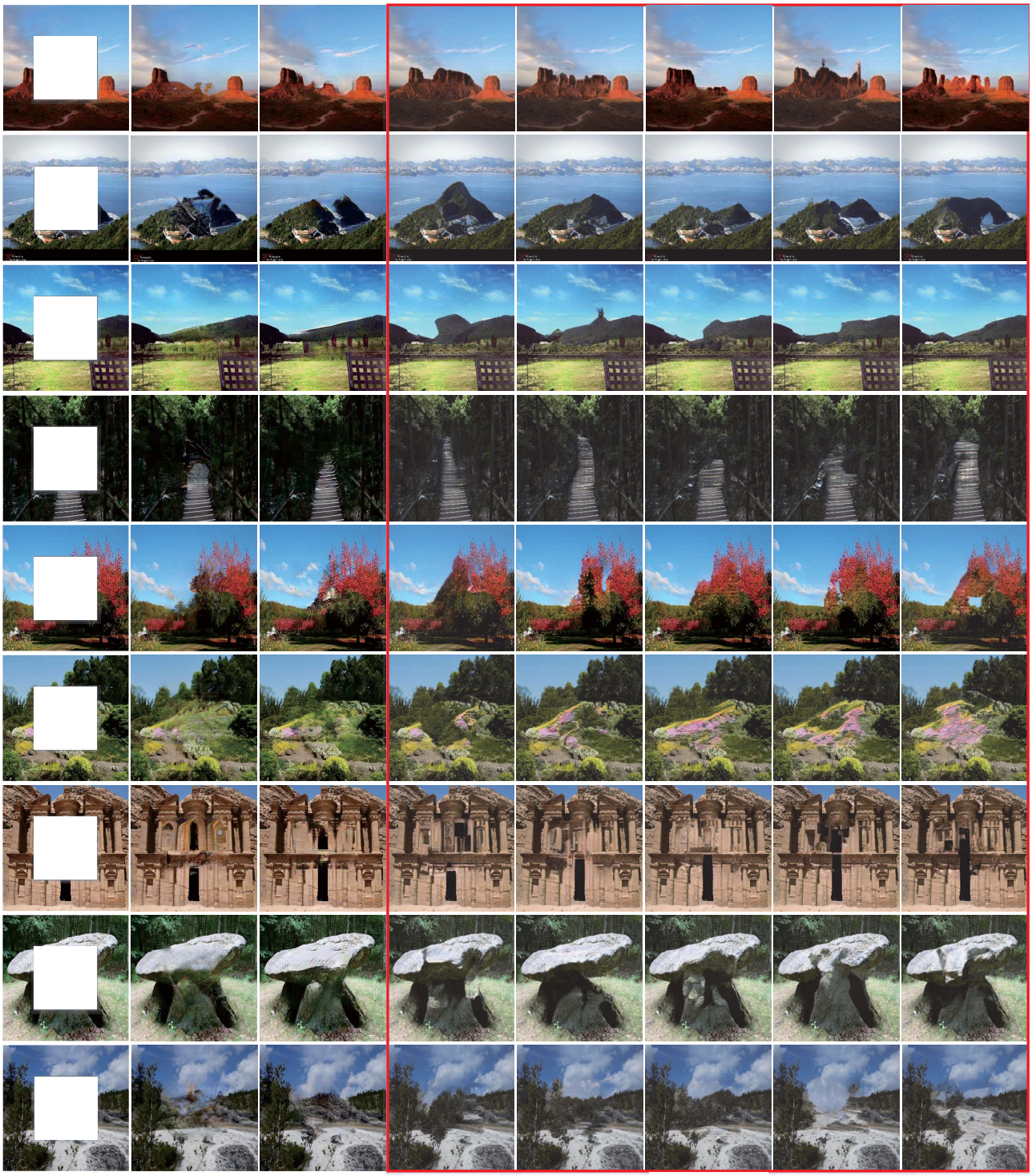
(a) Input (b) CE (c) Shift-Net (d) PICNet-regular

Figure A.1. Additional results on the Paris variation set for center hole completion. This variation dataset contains 100 images, for which we obtained generally more realistic results than the existing methods of CE and Shift-Net. Furthermore, our multiple results had a diverse range of sizes, shapes, colors and textures. Best viewed by zooming in.



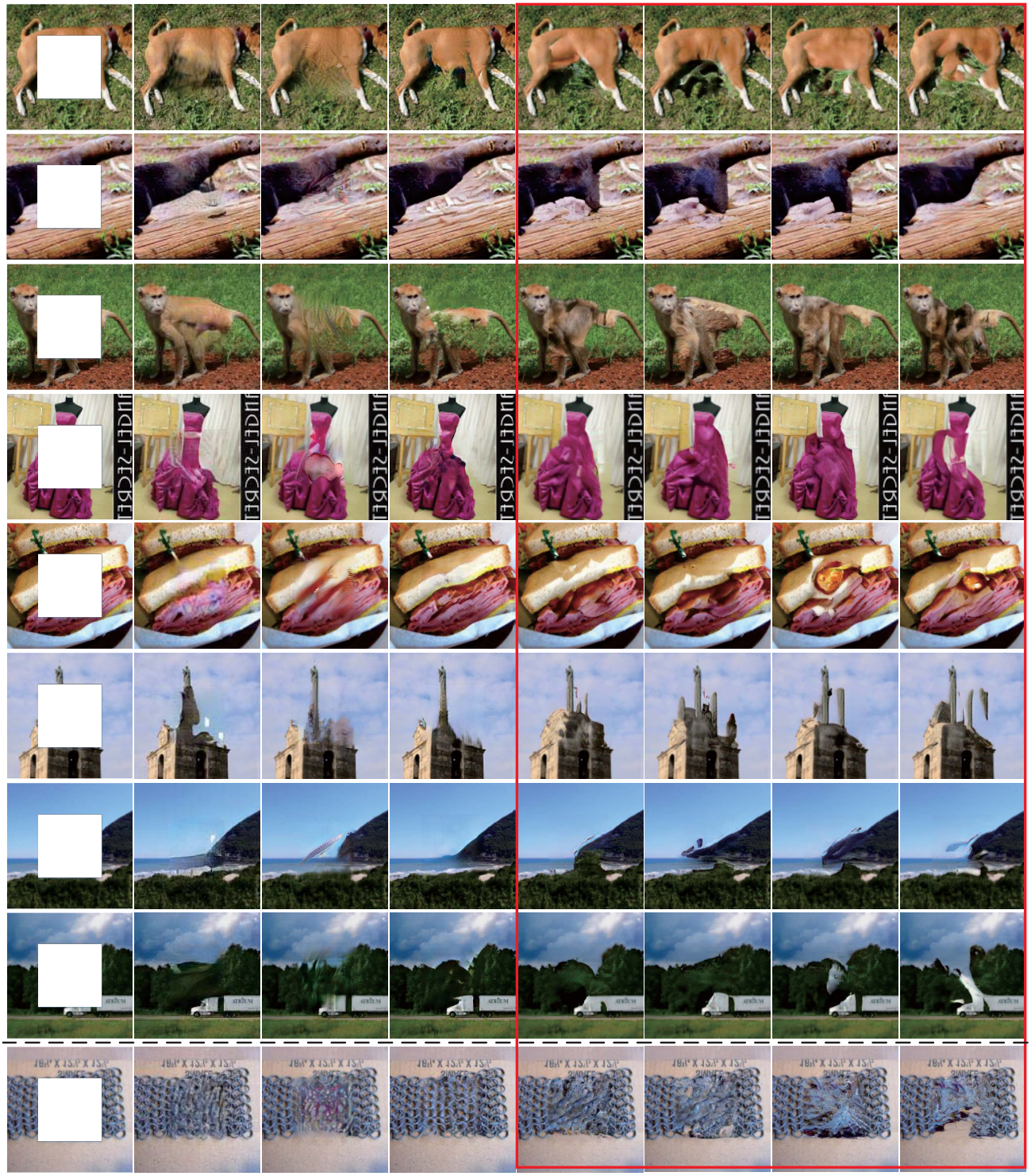
(a) Input (b) CA (c) PICNet-regular

Figure A.2. Additional results on the CelebA-HQ test set for center hole completion. Examples have different genders, skin tones, views and partial visible expressions. Since the occluded content in the large center holes was not repeated in visible regions, CA was unable to create results that were as visually realistic as ours. Moreover, our multiple outputs have different shapes, sizes and colors for eyes, noses and mouths. The details can be viewed by zooming in. Note that, no any other attribute labels (*e.g.* smile) were applied in our approach.



(a) Input (b) GL (c) CA (d) PICNet-regular

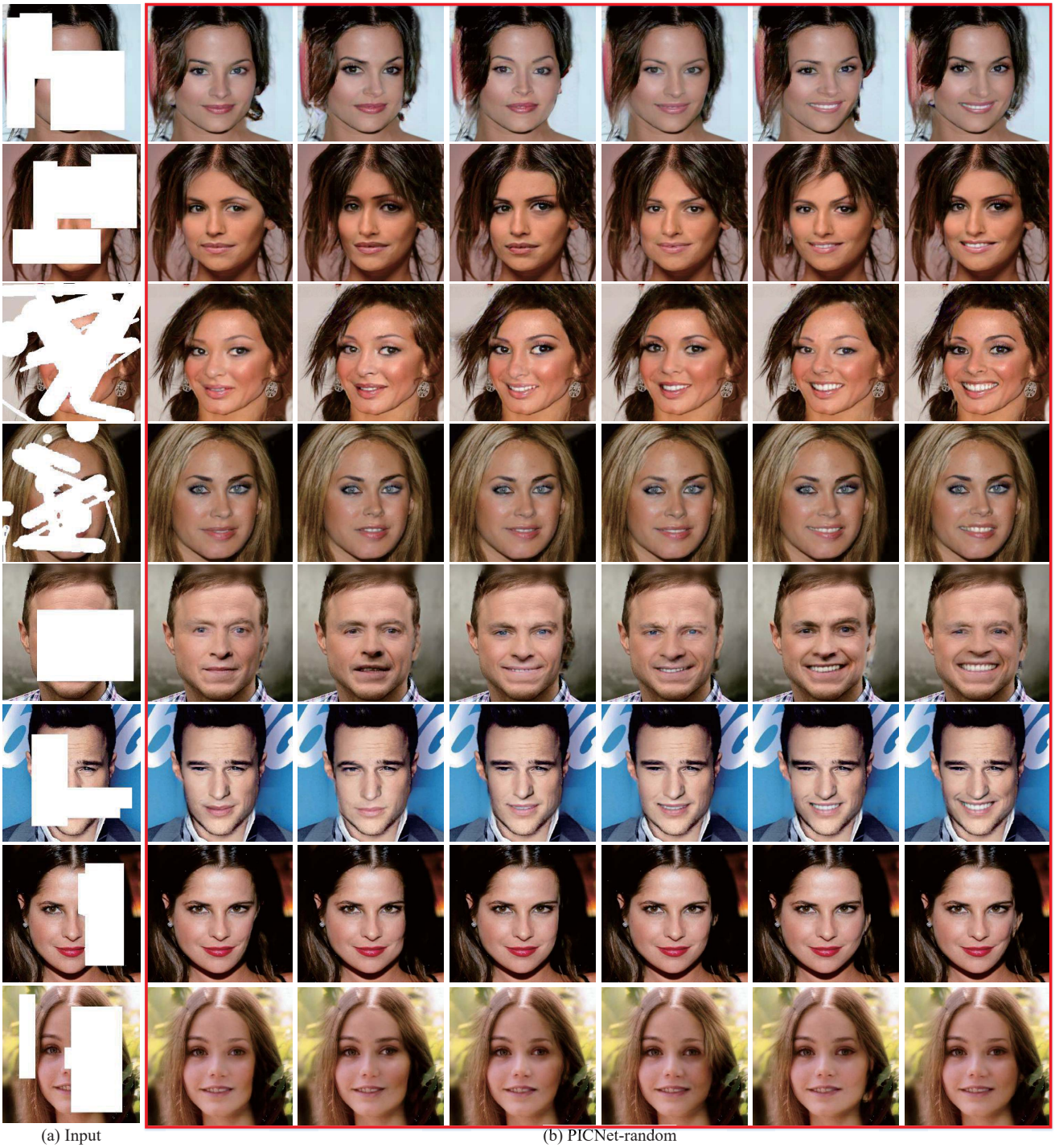
Figure A.3. Additional results on the Places2 variation set for center hole completion. Compared with existing state-of-the-art methods, our model not only generated completion results of comparable quality, but also provided multiple plausible results, with different shapes, colors, textures and content. The shape variations in generating the various prominent hills are obvious. Some changes were at finer scale, *e.g*. color changes of the flowers and texture changes in the boulder are better viewed by zooming in.



(a) Input (b) CE (c) GL (d) CA (e) PICNet-regular

Figure A.4. Additional results for center holes on the ImageNet variation set used in Context Encoder (CE). For our results, four completed images were selected and included failure cases in the last column. The first four rows show examples in which our model generated more visually realistic results than other methods. The next four rows show examples in which the methods all performed with similar realism, while the final row shows an example in which the Context Attention (CA) had the most realistic result.

A.2. Additional Results on Random and Irregular Hole Completion



(a) Input (b) PICNet-random

Figure A.5. Additional results on the CelebA-HQ test set for random and irregular hole completion. One interesting observation is that natural facial symmetry exerts a strong constraint on the completion results. In the examples where both eyes and/or mouth are masked out, the completion results exhibit substantial variation for those facial features when sampled. However, when only one eye is masked out or half of the mouth is visible (last three rows), the completion results for the other eye or the other half of mouth have little variation when sampled. Even when part of an eye is visible (fourth row), it exerts a strong constraint on the variation.

A.3. Additional Results on Free-Form Mask Using Our Interactive Demo

Figure A.6. Face image editing results from our online interactive demo. The white mask regions will be normalized to gray mask as the input. It shows that our PICNet can be used to object removal and face editing.

A.4. Video for Additional Results

Besides this document, we also included two video clips of additional results as part of the supplemental material. The first video, shows free-from mask results on various datasets. The second video consists of four parts to show multiple examples of center hole completion, random hole completion, comparison results with different training strategies and face editing of my self-portraits.

B. Mathematical Derivation and Analysis

B.1. Difficulties with Using the Classical CVAE for Image Completion

Here we elaborate on the difficulties encountered when using the classical CVAE formulation for pluralistic image completion, expanding on the shorter description in section 3.1.

B.1.1 Background: Derivation of the Conditional Variational Auto-Encoder (CVAE)

The broad CVAE framework of Sohn *et al*. [34] is a straightforward conditioning of the classical VAE. Using the notation in our main paper, a latent variable $\mathbf{z}_c$ is assumed to stochastically generate the hidden partial image $\mathbf{I}_c$. When conditioned on the visible partial image $\mathbf{I}_m$, we get the conditional probability:

$$p(\mathbf{I}_c|\mathbf{I}_m) = \int p_\phi(\mathbf{z}_c|\mathbf{I}_m) p_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) d\mathbf{z}_c \tag{B.1}$$

The variance of the Monte Carlo estimate can be reduced by importance sampling to get

$$\begin{aligned} p(\mathbf{I}_c|\mathbf{I}_m) &= \int q_\psi(\mathbf{z}_c|\mathbf{I}_c, \mathbf{I}_m) \frac{p_\phi(\mathbf{z}_c|\mathbf{I}_m)}{q_\psi(\mathbf{z}_c|\mathbf{I}_c, \mathbf{I}_m)} p_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) d\mathbf{z}_c \\ &= \mathbb{E}_{\mathbf{z}_c \sim q_\psi(\mathbf{z}_c|\mathbf{I}_c, \mathbf{I}_m)} \left[\frac{p_\phi(\mathbf{z}_c|\mathbf{I}_m)}{q_\psi(\mathbf{z}_c|\mathbf{I}_c, \mathbf{I}_m)} p_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) \right] \end{aligned} \tag{B.2}$$

Taking logs and apply Jensen's inequality leads to

$$\begin{aligned} \log p(\mathbf{I}_c|\mathbf{I}_m) &\geq \mathbb{E}_{\mathbf{z}_c \sim q_\psi(\mathbf{z}_c|\mathbf{I}_c, \mathbf{I}_m)} \left[\log p_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) - \log \frac{q_\psi(\mathbf{z}_c|\mathbf{I}_c, \mathbf{I}_m)}{p_\phi(\mathbf{z}_c|\mathbf{I}_m)} \right] \\ \mathcal{V} &= \mathbb{E}_{\mathbf{z}_c \sim q_\psi(\mathbf{z}_c|\mathbf{I}_c, \mathbf{I}_m)} \left[\log p_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) \right] - \mathrm{KL}\left(q_\psi(\mathbf{z}_c|\mathbf{I}_c, \mathbf{I}_m) || p_\phi(\mathbf{z}_c|\mathbf{I}_m) \right) \end{aligned} \tag{B.3}$$

The variational lower bound $\mathcal{V}$ totaled over all training data is jointly maximized w.r.t. the network parameters θ, ϕ and ψ in attempting to maximize the total log likelihood of the observed training instances.

B.1.2 Single Instance Per Conditioning Label

As is typically the case for image completion, there is only one training instance of $\mathbf{I}_c$ for each unique $\mathbf{I}_m$. This means that for the function $q_\psi(\mathbf{z}_c|\mathbf{I}_c, \mathbf{I}_m)$, $\mathbf{I}_c$ can simply be learnt into the network as a hardcoded dependency of the input $\mathbf{I}_m$, so $q_\psi(\mathbf{z}_c|\mathbf{I}_c, \mathbf{I}_m) \cong \hat{q}_\psi(\mathbf{z}_c|\mathbf{I}_m)$. Assuming that the network for $p_\phi(\mathbf{z}_c|\mathbf{I}_m)$ has similar or higher modeling power and there are no other explicit constraints imposed on it, then in training $p_\phi(\mathbf{z}_c|\mathbf{I}_m) \to \hat{q}_\psi(\mathbf{z}_c|\mathbf{I}_m)$, and the KL divergence in (B.3) goes to zero.

In this situation of zero KL divergence, we can rewrite the variational lower bound and replace $\hat{q}_\psi(\mathbf{z}_c|\mathbf{I}_m)$ with $p_\phi(\mathbf{z}_c|\mathbf{I}_m)$ without loss of generality, as

$$\mathcal{V} = \mathbb{E}_{\mathbf{z}_c \sim p_\phi(\mathbf{z}_c|\mathbf{I}_m)} \left[\log p_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) \right] \tag{B.4}$$

B.1.3 Unconstrained Learning of the Conditional Prior

We can analyze how$\mathcal{V}$ can be maximized, by using Jensen's inequality again (reversing earlier use)

$$\begin{aligned} \mathcal{V} &\leq \log \mathbb{E}_{\mathbf{z}_c \sim p_\phi(\mathbf{z}_c|\mathbf{I}_m)} \left[p_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) \right] \\ &= \log \int p_\phi(\mathbf{z}_c|\mathbf{I}_m) p_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) d\mathbf{z}_c \end{aligned} \tag{B.5}$$

By further applying Hölder's inequality (*i.e.* $\|fg\|_1 \leq \|f\|_p \|g\|_q$ for $1/p + 1/q = 1$), we get

$$\begin{aligned} \mathcal{V} &\leq \log \left[\left| \int |p_\phi(\mathbf{z}_c|\mathbf{I}_m)| \, d\mathbf{z}_c \right| \left| \int |p_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)|^\infty \, d\mathbf{z}_c \right|^{\frac{1}{\infty}} \right] \quad \text{(by setting } p = 1, q = \infty) \\ &= \log \left[1 \cdot \max_{\mathbf{z}_c} p_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) \right] = \max_{\mathbf{z}_c} \log p_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) \end{aligned} \tag{B.6}$$

Assuming that there is a unique global maximum for $\log p_\phi(\mathbf{z}_c|\mathbf{I}_m)$, the bound achieves equality when the conditional prior becomes a Dirac delta function centered at the maximum latent likelihood point

$$p_\phi(\mathbf{z}_c|\mathbf{I}_m) \rightarrow \delta(\mathbf{z}_c - \mathbf{z}_c^*) \quad \text{where } \mathbf{z}_c^* = \arg\max_{\mathbf{z}_c} \log p_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) \tag{B.7}$$

Intuitively, subject to the vagaries of stochastic gradient descent, the network for $p_\phi(\mathbf{z}_c|\mathbf{I}_m)$ without further constraints will learn a narrow delta-like function that sifts out maximum latent likelihood value of $\log p_\theta(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)$.

As mentioned in section 3.1, although this narrow conditional prior may be helpful in estimating a single solution for $\mathbf{I}_c$ given $\mathbf{I}_m$ during testing during testing, this is poor for sampling a diversity of solutions. In our framework, the (unconditional) latent priors are imposed for the partial images themselves, which prevent this delta function degeneracy.

B.1.4 CVAE with Fixed Prior

An alternative CVAE variant [37] assumes that conditional prior is independent of the $\mathbf{I}_m$ and fixed, so $p(\mathbf{z}_c|\mathbf{I}_m) \cong p(\mathbf{z}_c)$, where $p(\mathbf{z}_c)$ is a fixed distribution (*e.g.* standard normal). This means

$$p(\mathbf{I}_c|\mathbf{I}_m) = \int p(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) p(\mathbf{z}_c) d\mathbf{z}_c \tag{B.8}$$

Now we can consider the case for a fixed $\mathbf{I}_m = \mathbf{I}_m^*$, and rewrite (B.8) as

$$p_{\mathbf{I}_m^*}(\mathbf{I}_c) = \int p_{\mathbf{I}_m^*}(\mathbf{I}_c|\mathbf{z}_c) p(\mathbf{z}_c) d\mathbf{z}_c \tag{B.9}$$

Doing so makes it obvious we can then derive the standard (unconditional) VAE formulation from here. Thus an appropriate interpretation of this CVAE variant is that it uses $\mathbf{I}_m$ as a "switch" parameeter to choose between different VAE models that are trained for the specific conditions.

Once again, this is fine if there are multiple training instances per conditional label. However, in the image completion problem, there is only one $\mathbf{I}_c$ per unique $\mathbf{I}_m$, so the condition-specific VAE model will simply ignore the sampling "noise" and learn to predict the single instance of $\mathbf{I}_c$ from $\mathbf{I}_m$ directly, *i.e.* $p(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) \approx p(\mathbf{I}_c|\mathbf{I}_m)$, which incidentally achieves equality for the variational lower bound. This results in negligible variation of output despite now sampling from $p(\mathbf{z}_c) = \mathcal{N}(0, 1)$.

Our framework resolves this in part by defining all (unconditional) partial images of $\mathbf{I}_c$ as sharing a common latent space with adaptive priors, with the likelihood parameters learned as an unconditional VAE, and further coupling on the conditional portion (*i.e.* the generative path) to get a more distinct but regularized estimate for $p(\mathbf{z}_c|\mathbf{I}_m)$.

B.2. Joint Maximization of Unconditional and Conditional Variational Lower Bounds

The overall training loss function (5) used in our framework has a direct link to jointly maximizing the unconditional and unconditional variational lower bounds, respectively expressed by (2) and (4). Using simplified notation, we rewrite these bounds respectively as:

$$\begin{aligned} \mathcal{B}_1 &= \mathbb{E}_{q_\psi} \log p_\theta^r - \mathrm{KL}(q_\psi||p_{z_c}) \\ \mathcal{B}_2 &= \lambda \left(\mathbb{E}_{q_\psi} \log p_\theta^r - \mathrm{KL}(q_\psi||p_{z_c}) \right) + \mathbb{E}_{p_\phi} \log p_\theta^g \end{aligned} \tag{B.10}$$

To clarify, $\mathcal{B}_1$ is the lower bound related to the unconditional log likelihood of observing $\mathbf{I}_c$, while $\mathcal{B}_2$ relates to the log likelihood of observing $\mathbf{I}_c$ conditioned on $\mathbf{I}_m$. The expression of $\mathcal{B}_2$ reflects a blend of conditional likelihood formulations with and without the use of importance sampling, which are matched to different likelihood models, as explained in section 3.1. Note that the $(1 - \lambda)$ coefficient from (4) is left out here for simplicity, but there is no loss of generality since we can ignore a constant factor of the true lower bound if we are simply maximizing it.

We can then define a combined objective function as our maximization goal

$$\begin{aligned} \mathcal{B} &= \beta \mathcal{B}_1 + \mathcal{B}_2 \\ &= (\beta + \lambda) \mathbb{E}_{q_\psi} \log p_\theta^r + \mathbb{E}_{p_\phi} \log p_\theta^g - [\beta \mathrm{KL}(q_\psi||p_{z_c}) + \lambda \mathrm{KL}(q_\psi||p_\phi)] \end{aligned} \tag{B.11}$$

with $\beta \geq 0$.

To understand the relation between $\mathcal{B}$ in (B.11) and $\mathcal{L}$ in (5), we consider the equivalence of:

$$-\mathcal{B} \cong \mathcal{L} = \alpha_{\mathrm{KL}}(\mathcal{L}^r_{\mathrm{KL}} + \mathcal{L}^g_{\mathrm{KL}}) + \alpha_{\mathrm{app}}(\mathcal{L}^r_{\mathrm{app}} + \mathcal{L}^g_{\mathrm{app}}) + \alpha_{\mathrm{ad}}(\mathcal{L}^r_{\mathrm{ad}} + \mathcal{L}^g_{\mathrm{ad}}) \tag{B.12}$$

Comparing terms

$$\mathcal{L}^r_{\mathrm{KL}} \cong \mathrm{KL}(q_\psi || p_{z_c}), \quad \mathcal{L}^g_{\mathrm{KL}} \cong \mathrm{KL}(q_\psi || p_\phi) \quad \Rightarrow \beta = \lambda = \alpha_{\mathrm{KL}} \tag{B.13}$$

For the reconstructive path that involves sampling from the (posterior) importance function $q_\psi(\mathbf{z}_c|\mathbf{I}_c)$ of (3), we can substitute $(\beta + \lambda) = 2\alpha_{\mathrm{KL}}$ and get the reconstructive log likelihood formulation as

$$-\mathbb{E}_{q_\psi} \log p^r_\theta \cong \frac{\alpha_{\mathrm{app}}}{2\alpha_{\mathrm{KL}}}\mathcal{L}^r_{\mathrm{app}} + \frac{\alpha_{\mathrm{ad}}}{2\alpha_{\mathrm{KL}}}\mathcal{L}^r_{\mathrm{ad}} \tag{B.14}$$

Here, $\mathbf{I}_c$ is available, with $\mathcal{L}^r_{\mathrm{app}}$ reconstructing both $\mathbf{I}_c$ and $\mathbf{I}_m$ as in (8), while $\mathcal{L}^r_{\mathrm{ad}}$ involves GAN-based pairwise feature matching (10).

For the generative path that involves sampling from the conditional prior $p_\phi(\mathbf{z}_c|\mathbf{I}_m)$, we have the generative log likelihood formulation as

$$-\mathbb{E}_{p_\phi} \log p^g_\theta \cong \alpha_{\mathrm{app}}\mathcal{L}^g_{\mathrm{app}} + \alpha_{\mathrm{ad}}\mathcal{L}^g_{\mathrm{ad}} \tag{B.15}$$

As explained in sections 3.1 and 3.2, the generative path does not have direct access to $\mathbf{I}_c$, and this is reflected in the likelihood p^g_θ in which the instances of $\mathbf{I}_c$ are ignored. Thus $\mathcal{L}^g_{\mathrm{app}}$ is only for reconstructing $\mathbf{I}_m$ in a deterministic auto-encoder fashion as per (9), while $\mathcal{L}^g_{\mathrm{ad}}$ in (11) only tries to enforce that the generated distribution be consistent with the training set distribution (hence without per-instance knowledge), as implemented in the form of a GAN.

C. Architectural Details

Our **pluralistic image completion** network (**PICNet**) architecture is inspired by SA-GAN [43] and BigGAN, but features several important modifications that enable us to train for this image-conditional generation task. We first replace the batch normalization with instance normalization in the generation network (**ResBlock up** in Fig. C.7), and remove the batch normalization in our other networks, (*i.e.* the representation, inference and discriminator networks comprising **ResBlock start** and **ResBlock** in Fig. C.7), because different holes will affect the means and variances in each batch. **ResBlock down** is similar to **ResBlock**, in which we add the average pooling layer after Conv3 $\times$ 3 and Conv1 $\times$ 1.

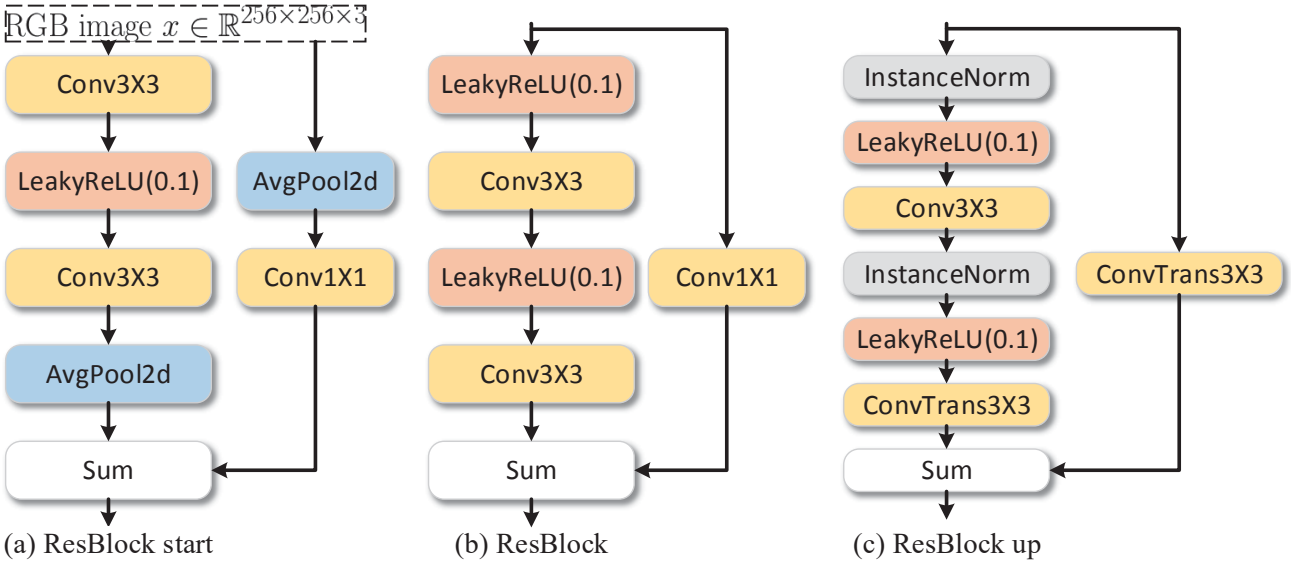
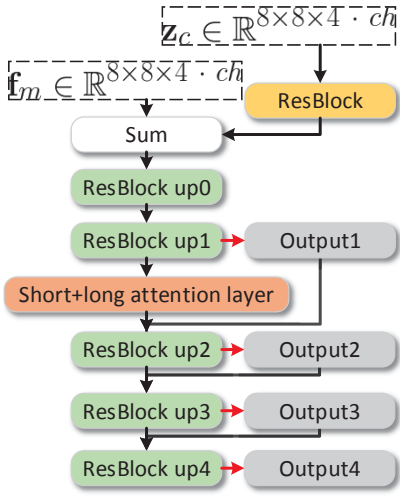


Figure C.7. Illustration of the Residual Block used in our model. (a) The starter Residual Block for the encoder (representation) and discriminator networks. (b) A Residual Block in the encoder (representation), inference and discriminator networks. (c) A Residual Block in the decoder (generator) network.



(a) Decoder (Generator)

RGB image $x \in \mathbb{R}^{256\times256\times3}$
ResBlock start $128 \times 128 \times 1 \cdot ch$
ResBlock down $64 \times 64 \times 2 \cdot ch$
ResBlock down $32 \times 32 \times 4 \cdot ch$
ResBlock down $16 \times 16 \times 4 \cdot ch$
ResBlock down $8 \times 8 \times 4 \cdot ch$

(b) Encoder (Representation)

RGB image $x \in \mathbb{R}^{256\times256\times3}$
ResBlock start $128 \times 128 \times 1 \cdot ch$
ResBlock down $64 \times 64 \times 2 \cdot ch$
ResBlock down $32 \times 32 \times 4 \cdot ch$
Self-Attention Layer $32 \times 32 \times 4 \cdot ch$
ResBlock down $16 \times 16 \times 4 \cdot ch$
ResBlock down $8 \times 8 \times 4 \cdot ch$
ResBlock $7 \times 7 \times 4 \cdot ch$
LeakyReLU(0.1), Conv, $6 \times 6 \times 1$

(c) Discriminator

Table C.1. Architectures for our framework, where ch represents the base channel width. For the output layer, we use the LeakyReLU(0.1), Conv3 $\times$ 3 and Tanh at all scales.

The **Infer1** network only consists of one Residual Block, for self-inferring the latent distribution of the ground truth $\mathbf{I}_c$ (treated as known in the reconstructive path), while the **Infer2** network consists of seven Residual Blocks, which are applied to predict the latent distribution of $\mathbf{I}_c$ (treated as unknown in the generative path) based on the visible pixels $\mathbf{I}_m$.

D. Experimental Details

Our network is implemented in Pytorch v0.4.0, and employs the architectures of Appendix C. To reduce memory cost, we restrained the feature channel width to $4 \cdot ch$ and selected $ch = 32$. We experimented with different channels with largest being $16 \cdot ch = 1024$, but found that the improvement was not obvious. In addition, we applied the self-attention layer of the discriminator and the short+long term attention layer of the generator on a 32×32 feature size. Spectral Normalization is used in all networks. All networks are initialized with Orthogonal Initialization and trained from scratch with a fixed learning rate of $\lambda = 10^{-4}$. We used the Adam optimizer with $\beta_1 = 0$ and $\beta_2 = 0.999$.

The final weights we used were $\alpha_{\text{KL}} = \alpha_{\text{app}}$=20, α_{ad}=1. The KL loss and appearance matching loss weights come from the variational *lower bound*. Since the appearance matching loss is used in four output scales, the final weight for the KL loss is $\alpha_{\text{KL}} = \alpha_{\text{KL}} \times N_{\text{scale}}$, where N_{scale} is the number of output scales. We also tried different values of α_{KL} and α_{app}, and found that the bigger the KL loss weight, the greater the diversity of the generated $\mathbf{I}_c^{'}$, but it was also harder to retain the appearance consistency of the generated $\mathbf{I}_c^{'}$ to the visible region $\mathbf{I}_m$. The values of α_{app} and α_{ad} were obtained from α-GAN. We experimented with the number of D steps per G step (varying it from 1 to 5), and found that one D step per G step gave the best results. When α_{app} is smaller than 1, we can use two or four D steps per G step, but the full generated $\mathbf{I}_g^{'}$ does not reconstruct the original conditional visible regions $\mathbf{I}_m$ well. When α_{app} is larger than 100, we needed two or four G steps per D step, if not the discriminator loss will become zero and the generated $\mathbf{I}_c^{'}$ will be blurry.

We trained each model on a single GPU, with a batch size of 20 on a GTX 1080TI (11GB) and 32 on a NVIDIA V100 (16GB). Training models for centered holes of Paris and CelebA-HQ takes roughly 3 days, while for ImageNet and Places2 it takes roughly 2 weeks. On the other hand, training models for random irregular and un-centered holes takes about twice the time compared to models for centered holes. Moreover, since the prior distribution of random holes $p(\mathbf{z}) = \mathcal{N}_m(\mathbf{0}, \sigma^2(n)\mathbf{I})$ is changed with the number of pixels in each hole n, the training loss may sometimes change abruptly due to the KL loss component.