International Journal of Computer Vision (2021) 129:2786–2805
https://doi.org/10.1007/s11263-021-01502-7

Pluralistic Free-Form Image Completion

Chuanxia Zheng[1] · Tat-Jen Cham[1] · Jianfei Cai[2]

Received: 21 November 2020 / Accepted: 7 July 2021 / Published online: 30 July 2021

Abstract

Image completion involves filling plausible contents to missing regions in images. Current image completion methods produce only one result for a given masked image, 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 free-form image completion. A major challenge faced by learning-based approaches is that usually only one ground truth training instance per label for this multi-output problem. 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 ground truth to get prior distribution of missing patches 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 adversarial learning. We then introduce a new short+long term patch attention layer that exploits distant relations among decoder and encoder features, to improve appearance consistency between the original visible and the generated new regions. Experiments show that our method not only yields better results in various datasets than existing state-of-the-art methods, but also provides multiple and diverse outputs.

Keywords Image completion · Multi-modal generative models · Image generation · Conditional variational auto-encoders

1 Introduction

Image completion involves the issues of filling alternative contents for the missing parts in images, which can be used for restoring the damaged painting, removing unwanted objects, and generating new contents for incomplete scenes. Many approaches have been proposed for this non-trivial task, including diffusion-based methods (Bertalmio et al. 2000; Ballester et al. 2001; Levin et al. 2003; Bertalmio et al. 2003), patch-based methods (Criminisi et al. 2003, 2004; Jia and Tang 2004; Barnes et al. 2009) and learning-based methods (Pathak et al. 2016; Iizuka et al. 2017; Yu et al. 2018; Liu et al. 2018; Nazeri et al. 2019; Yi et al. 2020). While these approaches rapidly improve the completion results, they produce only one "optimal" result for a given masked image and do *not* have the capacity to generate a variety of semantically meaningful results. It remains a challenging problem to provide *multiple* and *diverse* plausible results for this highly subjective process problem.

Supposing you were shown the images with various missing regions in Fig. 1, what would you *imagine* to be occupying these holes? Bertalmio et al. (2000) related how expert conservators would restore 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*, such as various shapes and colors of eyes, even if they may universally agree on high-level semantics, such as general placement of eyes and mouth on a damaged portrait.

Based on this observation, our main goal in this research is thus to generate *multiple* and *diverse* plausible results when presented with a masked image. We refer to this task as *plu-*

Communicated by Jian Sun.

✉ Chuanxia Zheng
chuanxia001@e.ntu.edu.sg

Tat-Jen Cham
astjcham@ntu.edu.sg

Jianfei Cai
Jianfei.Cai@monash.edu

[1] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

[2] Department of Data Science and AI, Monash University, Clayton, VIC, Australia

ralistic image completion (depicted in Fig. 1). This is as opposed to existing works that attempt to generate only a single "guess" for this ill-posed problem.

To obtain a diverse set of results for a given input, some methods utilize conditional variational auto-encoders (CVAE) (Sohn et al. 2015; Walker et al. 2016; Bao et al. 2017; Eslami et al. 2018), a conditional extension of variational auto-encoders (VAE) (Kingma and Welling 2013), 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 masked image*, the number of ground truth instances in the training data that match the label is *typically only one—the original complement of the masked image*. Hence the estimated original conditional distributions tend to have very limited variation since they were trained to reconstruct the single original image.

An important insight we will use is that *partial images (patches)*, as a superset of full images, may also be considered as generated from *a latent space with smooth prior distributions* (Shaham et al. 2019). This provides a mechanism for alleviating the problem of having scarce samples per conditional masked image. To do so, we introduce a Pluralistic Image Completion Network, called *PICNet*, 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, but also imposes smooth priors for the latent space of missing partial image. The second pipeline is a generative path that learns to predict the latent prior distribution for the missing regions only based 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 results at all, instead allowing the reasonableness of results being driven by an auxiliary discriminator network (Goodfellow et al. 2014). This leads to substantially great variability in generation.

To further utilize the information from the visible partial images as much as possible (Barnes et al. 2009; Yu et al. 2018), we also introduce an enhanced *short+long term patch attention* layer, a generic attention mechanism that allows information flowing from visible regions to missing holes. This scheme converges quickly and significantly increases the quality of our completed results.

We comprehensively evaluate and compare our approach with existing state-of-the-art methods on a large variety of scenes (Sect. 4.2), where various masks, including regular and free-form irregular masks, are used to erode the images. We additionally present many interesting applications of our model on free-form image editing (Sect. 4.3), e.g. object removal, face editing, and scene content-aware-move. The extensive experimental results demonstrate that our proposed PICNet not only generates higher-quality completion results, but also produces multiple diverse solutions for this subjective processing task.

In summary, in this paper we present:

1. A probabilistically principled framework for free-form image completion that is able to maintain much higher sample diversity as compared to existing methods;
2. a PICNet with two parallel training paths, which trades off between reconstructing the original training data and maintaining the variance of the conditional distribution;
3. a novel short+long term patch attention layer that exploits context information to ensure appearance consistency in the image domain, in a manner superior to purely using GANs;
4. we demonstrate that our method is able to complete the free-form mask with multiple plausible results that have substantial diversity.

A preliminary version of this manuscript was published in CVPR'19 (Zheng et al. 2019). In this journal extension, we improved the proposed image completion method, conducted a thorough analysis of each component, and presented many more extensive experiments. In particular, we restricted the distribution estimation in a separate training phase, and further extended the *Short + Long Term Attention* to patch level. Through these modifications, we successfully extended the probabilistically principled framework to *free-form* image completion, solving for arbitrary input masks. We also provided a thorough ablation study to analyze each proposed component. Moreover, we additionally evaluated our approach on various free-form masks and conducted many more experiments, with quantitative comparisons using two learning-based feature-level metrics and two types of user studies, and present qualitative results on handing high-resolution images and various image editing applications. Many recent works (Deng and Wang 2020; Zhao et al. 2020; Peng et al. 2021) have also begun to consistently use our method as a state-of-the-art benchmark for pluralistic image completion, and our framework has also been extended by other groups to different tasks (Hara and Harada 2020). Our code and interactive demo are also publicly available.[1]

The rest of the paper is structured as follows: We discuss the related work in Sect. 2. Next, we describe the proposed probabilistically principled framework in Sect. 3.1, and the improved attention module in Sect. 3.2 together with thorough analysis for each proposed component in the corresponding sections. We then describe and discuss the experiments in Sect. 4, and conclude in Sect. 5.

[1] Code: https://github.com/lyndonzheng/Pluralistic-Inpainting Demo: http://www.chuanxiaz.com/project/pluralistic.

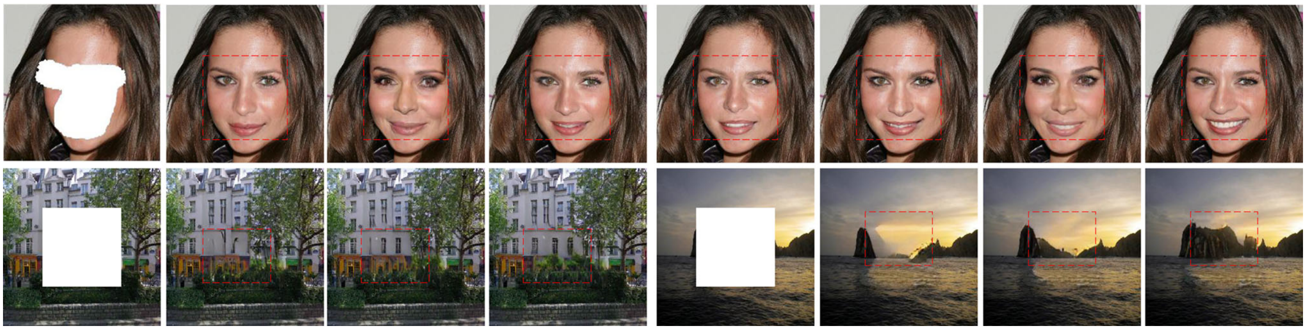


Fig. 1 Example completion results of our method on images of a face, a building, and natural scenery with various masks (masks shown in white only for visual purpose). For each group, the masked input image is shown left, followed by sampled results from our model without any post-processing. The results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

2 Related Work

Existing work on image completion either uses information from within the image (Bertalmio et al. 2000, 2003), or information from a large image dataset (Hays and Efros 2007; Pathak et al. 2016). Most approaches generate only one result per masked image, which is precisely the downside we want to address in this paper.

Intra-Image Completion Traditional intra-image completion works [also known as "inpainting" (Bertalmio et al. 2000)] mainly propagate, copy and realign the background regions to missing regions, focusing only on the steps 2 and 3 above, by assuming that the holes should be filled with similar appearance to that of the visible regions. One category of intra-image completion methods are diffusion-based image synthesis (Bertalmio et al. 2000; Ballester et al. 2001; Levin et al. 2003; Bertalmio et al. 2003). These methods fill the surrounded backgrounds to the missing regions by propagating the local colors. They only work well on the small and narrow holes. Another category of intra-image completion methods are patch-based approaches (Criminisi et al. 2003, 2004; Jia and Tang 2004; Barnes et al. 2009). They fill the holes by copying information from similar visible regions, which produce high-quality texture-consistent result. However, these intra-image methods cannot capture global semantics to hallucinate new content for large holes (as in step 1), which is significant for real image completion.

Inter-Image Completion To hallucinate semantically new content, inter-image completion borrows information from a large dataset. Hays and Efros (2007) first present an image completion method using millions of images. Recently, learning-based approaches are proposed. Initial works (Köhler et al. 2014; Ren et al. 2015) focus on small and thin holes. Then, Pathak et al. (2016) proposed the Context Encoders (CE) to handle 64 × 64-sized holes. Iizuka et al. (2017) built upon (Pathak et al. 2016) by combining global and local discriminators (GL) as adversarial loss. Wang et al. (2018) designed a Multi-column CNNs and a cosine similarity based loss for high quality image in painting. More recent, Liu et al. (2018) introduced "partial convolution" for free-form irregular mask image completion.

Some work has also explored additional information for semantically image completion. Yeh et al. (2017), the "closest" features in the latent space for the masked image are searched to generate an image. Li et al. (2017) introduced additional face parsing loss to ensure the semantic consistency of completed images. Song et al. (2018b) proposed SPG-Net that simultaneously does semantic map and RGB appearance completion. Moreover, sketches and color are used in the latest Faceshape (Portenier et al. 2018), DeepFillv2 (Yu et al. 2019), EdgeConnect (Nazeri et al. 2019) and SC-FEGAN (Jo and Park 2019). A common drawback of these methods is that they utilize the visible information only through local convolutional operations, which creates distorted structures and blurry textures inconsistent with the visible regions, especially for large holes.

Combined Intra- and Inter-Image Completion To mitigate the blurry problems, Yang et al. (2017) proposed multi-scale neural patch synthesis, which generates high-frequency details by copying patches from mid-layer features. More recently, several works (Yu et al. 2018; Yan et al. 2018; Song et al. 2018a; Yi et al. 2020) exploit spatial attention (Jaderberg et al. 2015; Zhou et al. 2016) to get high-frequency details. Yu et al. (2018) proposed a contextual attention layer to produce high-frequency details by copying similar features from visible regions to missing regions. Yan et al. (2018) and Song et al. (2018a) proposed PatchMatch-like ideas on feature domain. Yi et al. (2020) proposed contextual residual aggregation for very high resolution (8K) image inpainting. However, these methods identify similar features by comparing features of holes and 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

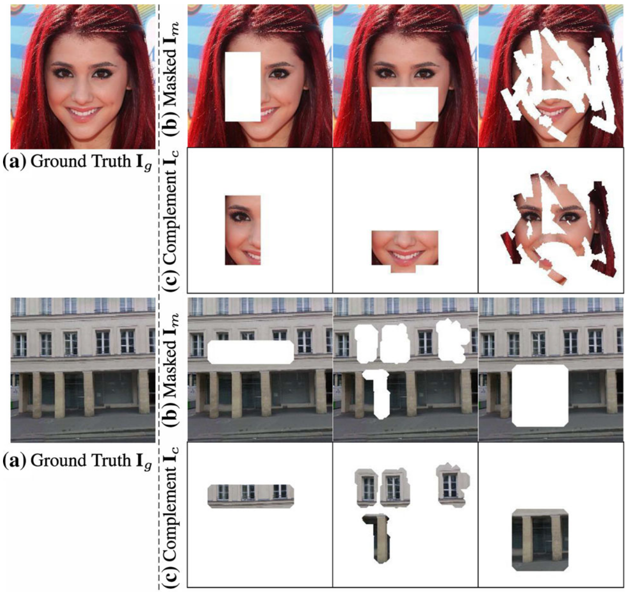

(a) Ground Truth $\mathbf{I}_g$
(b) Masked $\mathbf{I}_m$
(c) Complement $\mathbf{I}_c$
(a) Ground Truth $\mathbf{I}_g$
(b) Masked $\mathbf{I}_m$
(c) Complement $\mathbf{I}_c$

Fig. 2 Examples of different degraded images. **a** Ground truth image $\mathbf{I}_g$. **b** Masked image $\mathbf{I}_m$. **c** The corresponding complement image $\mathbf{I}_c$ to each top masked image $\mathbf{I}_g$. It is often not reasonable to strongly enforce the completed masked regions to be identical to the ground truth, especially in cases when large variations in the completed content can still be perfectly consistent to the visible regions, e.g. when the entire mouth or both eyes are masked

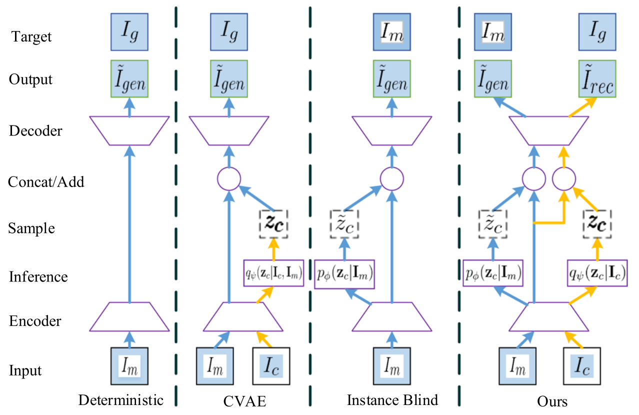


Fig. 3 Completion strategies given masked image. (Deterministic) structure directly predicts the ground truth instance. [CVAE (Walker et al. 2016)] adds in random sampling to diversify the output, but is still trained on the single ground truth. (Instance blind) only matches the masked instance, but training is unstable. (Ours) uses a generative path during testing, but is guided by a parallel reconstructive path during training. Note that, yellow path is only used for training (Color figure online)

for new content that differs from visible regions. Our model solves it by extending self-attention to harness abundant context.

Image Generation Image generation has progressed significantly using methods such as VAE (Kingma and Welling 2013) and GANs (Goodfellow et al. 2014). These have been applied to conditional image generation tasks, such as image translation (Isola et al. 2017; Zhu et al. 2017a), synthetic to realistic (Shrivastava et al. 2017; Zheng et al. 2018), future prediction (Mathieu et al. 2015), and 3D models (Park et al. 2017). Perhaps most relevant in spirit to us are conditional VAEs (CVAE) (Sohn et al. 2015; Walker et al. 2016) and CVAE-GAN (Bao et al. 2017), but these are 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 (Zhu et al. 2017b; Lee et al. 2018), but in such situations the condition-to-sample mappings are more local (e.g. pixel-to-pixel), and only change the visual appearance without generating new content. This is untrue for image completion, where the conditional label is the masked image itself, with only one training instance of the original holes. Chen et al. (2018), different outputs were obtained for face completion by specifying facial attributes (e.g. smile), but this method is very domain specific, requiring targeted attributes. In contrast, our proposed probabilistically principled framework produces multiple and diverse plausible in various datasets, which does not need any label information for training.

3 Approaches

Suppose we have an image, originally ground truth $\mathbf{I}_g$ (Fig. 2a), but degraded by a number of missing pixels to become $\mathbf{I}_m$ (Fig. 2b), *masked partial image* comprising the visible pixels. We also define $\mathbf{I}_c$ (Fig. 2c) as its *complement partial image* comprising the missing pixels.

Prior image completion methods (Yu et al. 2018; Pathak et al. 2016; Iizuka et al. 2017; Nazeri et al. 2019) attempt to reconstruct the original unmasked image $\mathbf{I}_g$ in a deterministic fashion from $\mathbf{I}_m$ (see Fig. 3 "Deterministic"). However, this rigid approach has several limitations. First, while it is fine to rebuild the original image $\mathbf{I}_g$ when visible regions tightly constrain the completed content, e.g. when only the left half of a face is masked in Fig. 2, it is unnecessarily limiting when visible regions allow a much greater range of perceptually consistent completion, e.g. with many different mouth expressions or building door appearances equally acceptable in Fig. 2. Second, deterministic methods can only generate a single solution and are not able to recover a richer distribution of reasonable possibilities. Instead, our goal is to *sample* from $p(\mathbf{I}_c|\mathbf{I}_m)$ and we reconstruct the original image only when the corresponding complement partial images $\mathbf{I}_c$ are provided during the training.

3.1 Pluralistic Image Completion Network

3.1.1 Probabilistic Framework

In order to have a distribution to sample from, an approach is to employ the CVAE (Sohn et al. 2015) which estimates a parametric distribution over a latent space, from which sampling is possible. This involves a variational lower bound of the conditional log-likelihood:

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

where $\mathbf{z}_c$ is the latent vector of missing patches, $q_\psi(\mathbf{z}_c|\mathbf{I}_c,\mathbf{I}_m)$ is the recognition network, $p_\phi(\mathbf{z}_c|\mathbf{I}_m)$ is the conditional prior, and $p_\theta(\mathbf{I}_c|\mathbf{z}_c,\mathbf{I}_m)$ is 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.

For our purposes, the chief difficulty of using CVAE (Sohn et al. 2015) directly is that the high DoF of recognition network $q_\psi(\mathbf{z}_c|\mathbf{I}_c,\mathbf{I}_m)$ and conditional prior network $p_\phi(\mathbf{z}_c|\mathbf{I}_m)$ are not easily separable in (1). Besides, since the conditional prior network $p_\phi(\mathbf{z}_c|\mathbf{I}_m)$ is sufficiently unconstrained in (1), it will lean a narrow delta-like prior distribution 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 $p_\theta(\mathbf{I}_c|\mathbf{z}_c,\mathbf{I}_m)$. In this way, the variance σ^2 of the learned latent distribution is easily driven towards zero. Then it 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 (Sohn et al. 2015), in which they directly set the recognition network the same as the prior network, i.e., $q_\psi(\mathbf{z}_c|\mathbf{I}_c,\mathbf{I}_m) = p_\phi(\mathbf{z}_c|\mathbf{I}_m)$. 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. When the CVAE variant of Walker et al. (2016), which assumes conditional prior $p_\phi(\mathbf{z}_c|\mathbf{I}_m) = \mathcal{N}(\mathbf{0},\mathbf{I})$, is used instead, the network learns to ignore the latent sampling and directly estimates $\mathbf{I}_c$ from $\mathbf{I}_m$ for a fixed ground truth, also resulting in similar solutions. A possible way to diversify the output is simply to 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 a learned adversarial discriminator (see Fig. 3 "Instance Blind"). However, this approach is unstable, especially for large and complex scenes (Song et al. 2018a). A detail analysis is presented in Sect. 3.1.3.

Latent Priors of Holes In our approach, we require that missing partial images (patches), 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:

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

where in Kingma and Welling (2013) 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. In particular, *a complement image* $\mathbf{I}_c$ *with more pixels (large holes for the masked image* $\mathbf{I}_m$*, as shown in the last column in* Fig. 2) *should have greater prior variance than a complement image* $\mathbf{I}_c$ *with fewer pixels (small holes)* and in fact a masked partial image $\mathbf{I}_m$ with no pixels missing should be completely deterministic! 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 missing pixels n, where $\sigma^2(n) = \frac{n}{H\times W} \in (0,1]$.

Prior-Conditional Coupling Next, we combine the latent priors into the conditional lower bound of (1). Since $\mathbf{z}_c$ represents the distributions of target missing partial image $\mathbf{I}_c$, $\mathbf{z}_c$ can be naturally inferred using the target missing image $\mathbf{I}_c$, that $q_\psi(\mathbf{z}_c|\mathbf{I}_c,\mathbf{I}_m) \approx q_\psi(\mathbf{z}_c|\mathbf{I}_c)$ when $\mathbf{I}_c$ is available in the training. 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, yet 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*.

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 function $q_\psi(\mathbf{z}_c|\mathbf{I}_c,\mathbf{I}_m)$, $\mathbf{I}_c$ can be learned 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 $p_\phi(\mathbf{z}_c|\mathbf{I}_m) \rightarrow \hat{q}_\psi(\mathbf{z}_c|\mathbf{I}_m)$, and the KL divergence in (1) goes to zero. Then we get the following function:

$$\log p(\mathbf{I}_c|\mathbf{I}_m) \geq \mathbb{E}_{p_\phi(\mathbf{z}_c|\mathbf{I}_m)}[\log p_\theta(\mathbf{I}_c|\mathbf{z}_c,\mathbf{I}_m)] \quad (4)$$

the "GSNN" version in Sohn et al. (2015). However, unlike Sohn et al. (2015), the variance σ^2 of the learned distribution $p_\phi(\mathbf{z}_c|\mathbf{I}_m)$ in our method will not be zero as mentioned

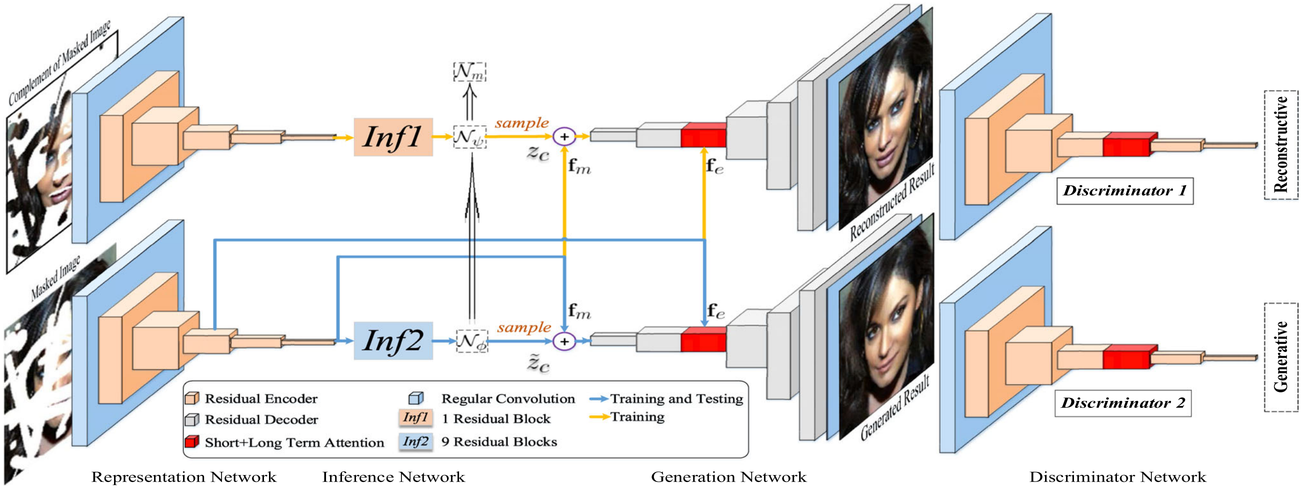


Fig. 4 Overview of our architecture with two parallel pipelines. The top *reconstructive* pipeline (yellow line) combines information from $\mathbf{I}_m$ and $\mathbf{I}_c$, which is used only for training. The bottom *generative* pipeline (blue line) infers the conditional distribution of hidden regions, that can be sampled during testing. The two representation networks and generation networks in top and bottom share identical weights (Color figure online)

above. This $\mathbf{z}_c$ for missing regions is sampling from the visible regions $\mathbf{I}_m$, we call this *without importance sampling*, contrary to the *importance sampling* $q_\psi(\mathbf{z}_c|\mathbf{I}_c)$. Finally, we combine (3) and (4) to obtain the reconstruction and creative generation function:

$$\log p(\mathbf{I}_c|\mathbf{I}_m) \geq \lambda \left\{\mathbb{E}_{q_\psi}[\log p_\theta^r(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)] - \mathrm{KL}(q_\psi||p_\phi)\right\} \\ + (1-\lambda)\,\mathbb{E}_{p_\phi}[\log p_\theta^g(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)] \qquad (5)$$

where $\lambda \in [0, 1]$ is implicitly set by training loss coefficients in Sect. 3.1.2. When sampling from the importance function $q_\psi(\mathbf{z}_c|\mathbf{I}_c)$, the missing instance information is available and we formulate the likelihood $p_\theta^r(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)$ to be focused on *reconstructing* $\mathbf{I}_c$. Conversely, when sampling from the learned distribution $p_\phi(\mathbf{z}_c|\mathbf{I}_m)$ which does not contain $\mathbf{I}_c$, we will facilitate *creative generation* by having the likelihood model $p_\theta^g(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m) \cong \ell_\theta^g(\mathbf{z}_c, \mathbf{I}_m)$ be *independent of the original instance* of $\mathbf{I}_c$. Instead it only *encourages generated samples to fit in with the overall training distribution.*

Joint Unconditional and Conditional Variational Lower Bounds Our overall training objective may then be expressed as jointly maximizing the lower bounds in (2) and (5). This can be done by unifying the likelihood in (2) to that in (5) as $p_\theta(\mathbf{I}_c|\mathbf{z}_c) \cong p_\theta^r(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)$, in which the $\mathbf{z}_c$ is sampling from the *important sampling* $q_\psi(\mathbf{z}_c|\mathbf{I}_c)$ that can be used for rebuild the original missing regions $\mathbf{I}_c$. We can then define a combine function as our maximization goal:

$$\begin{aligned}\mathcal{B} &= \beta\,\mathcal{B}_1 + \mathcal{B}_2 \\ &= -\left[\beta \mathrm{KL}(q_\psi||p_{z_c}) + \lambda \mathrm{KL}(q_\psi||p_\phi)\right] \\ &\quad + (\beta+\lambda)\mathbb{E}_{q_\psi}\log p_\theta^r + (1-\lambda)\mathbb{E}_{p_\phi}\log p_\theta^g \end{aligned} \qquad (6)$$

where $\mathcal{B}_1$ is the lower bound related to the unconditional log likelihood of missing partial image $\mathbf{I}_c$, and $\mathcal{B}_2$ relates to the log likelihood of missing regions $\mathbf{I}_c$ conditioned on $\mathbf{I}_m$. Note that this function holds a key different with hybrid objective function in Sohn et al. (2015) that *the conditional prior network* $p_\phi(\mathbf{z}_c|\mathbf{I}_m)$ *and the recognition network* $q_\psi(\mathbf{z}_c|\mathbf{I}_c)$ *are no longer freely learned, but are constrained by a mask related prior* $p(\mathbf{z}_c) = \mathcal{N}_m(\mathbf{0}, \sigma^2(n)\mathbf{I})$. Furthermore, our *without importance sampling*, also the testing sampling, does not learn to predict a fixed instance during the training, which encourages larger diversity.

3.1.2 Network Structure and Training Loss

The formula in (6) is implemented as our dual pipeline, illustrated in Fig. 4. This consists of representation, inference, generation, and auxiliary discriminator networks in two paths. The upper pipeline is the reconstruction path used in training that corresponds to the lower bound $\mathcal{B}_1$, in which $\mathbf{z}_c$ contains information of missing image $\mathbf{I}_c$. Hence when combined with the conditional feature $\mathbf{f}_m$, we can easily train this path to rebuild the original image $\mathbf{I}_g$. In contrast, the lower path, used in both training and testing, is responsible for the lower bound $\mathcal{B}_2$, where the missing information is inferred only from masked image $\mathbf{I}_m$, resulting in a less restrictive prediction.

We transfer the lower bound terms in (6) as the corresponding loss function. During training, jointly maximizing the lower bounds is then minimizing a total loss $\mathcal{L}$, which consists of three groups of component losses:

$$\mathcal{L} = \alpha_{\mathrm{KL}}(\mathcal{L}_{\mathrm{KL}}^r + \mathcal{L}_{\mathrm{KL}}^g) + \alpha_{\mathrm{app}}(\mathcal{L}_{\mathrm{app}}^r + \mathcal{L}_{\mathrm{app}}^g) \\ + \alpha_{\mathrm{ad}}(\mathcal{L}_{\mathrm{ad}}^r + \mathcal{L}_{\mathrm{ad}}^g) \qquad (7)$$

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 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, respectively.

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

$$\mathcal{L}_{\text{KL}}^{r,(i)} = \text{KL}(q_\psi(\mathbf{z}|I_c^{(i)})||\mathcal{N}_m(\mathbf{0}, \sigma^{2,(i)}(n)\mathbf{I})) \tag{8}$$

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

$$\mathcal{L}_{\text{KL}}^{g,(i)} = \text{KL}(q_\psi(\mathbf{z}|I_c^{(i)}))||p_\phi(\mathbf{z}|I_m^{(i)}))) \tag{9}$$

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

Appearance Matching Loss The likelihood term $p_\theta^r(\mathbf{I}_c|\mathbf{z}_c)$ is interpreted as probabilistically encouraging appearance matching to the missing regions $\mathbf{I}_c$. However, our framework also auto-encodes the masked image $\mathbf{I}_m$ (via $\mathbf{f}_m$) deterministically, and the loss function needs to cater for this reconstruction. As such, the per-instance loss here is:

$$\mathcal{L}_{\text{app}}^{r,(i)} = ||I_{\text{rec}}^{(i)} - I_g^{(i)}||_1 \tag{10}$$

where $I_{\text{rec}}^{(i)} = G(z_c, f_m)$ and $I_g^{(i)}$ are the reconstructed and original full images respectively. The purpose of this loss is to bias the representation towards the actual visible information. In contrast, for the generative path the latent distribution $\mathcal{N}_\phi$ of the missing regions $\mathbf{I}_c$ is inferred based only on the visible $\mathbf{I}_m$. This would be significantly less accurate than the inference in the upper path. Thus, we ignore instance-specific appearance matching for $\mathbf{I}_c$, and only focus on reconstructing $\mathbf{I}_m$:

$$\mathcal{L}_{\text{app}}^{g,(i)} = ||M * (I_{\text{gen}}^{(i)} - I_g^{(i)})||_1 \tag{11}$$

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

Adversarial Loss. The formulation of $p_\theta^r(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)$ and the instance-blind $p_\theta^g(\mathbf{I}_c|\mathbf{z}_c, \mathbf{I}_m)$ 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 (Bao et al. 2017), we use a mean feature match loss in the reconstructive path for the generator,

$$\mathcal{L}_{\text{ad}}^{r,(i)} = ||f_{D_1}(I_{\text{rec}}^{(i)}) - f_{D_1}(I_g^{(i)})||_2 \tag{12}$$

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:

$$\mathcal{L}_{\text{ad}}^{g,(i)} = [D_2(I_{\text{gen}}^{(i)}) - 1]^2 \tag{13}$$

This is based on the generator loss in LSGAN (Mao et al. 2017), which performs better than the original GAN loss (Goodfellow et al. 2014) in our scenario. The discriminator loss for both D_1 and D_2 is also based on LSGAN.

3.1.3 Analysis

Effect of Network Structure We first investigated the influence of using our two-path training structure in comparison to other variants such as the CVAE of (Walker et al. 2016) and the "Instance Blind" structures in Fig. 3. We also trained the state-of-the-art multi-model BicycleGAN (Zhu et al. 2017b) on Celeba-HQ dataset (Liu et al. 2015; Karras et al. 2017) by setting $\mathbf{A} = \mathbf{I}_m$, $\mathbf{B} = \mathbf{I}_c$ with center mask.

We first computed diversity score using the Learned Perceptual Image Patch Similarity (LPIPS) metric reported in Zhu et al. (2017b). LPIPS metric (Zhang et al. 2018b) calculates the average distance of samples in a deep feature domain. For each random pairs, a pre-trained deep network is used to extract the features of images. Then, the distance of two vectors is calculated using ℓ_1 distance. The larger distance indicates the results are much more diverse, as the generated pairs far from each other. For each method, we sampled 50 K pairs of randomly generated images from 1K center masked images. $\mathbf{I}_{out}$ and $\mathbf{I}_{out(m)}$ are the full output and the masked-regions' output, respectively. Furthermore, we used the popular Fréchet Inception Distance (FID) (Heusel et al. 2017) to assess the visual quality of completed images by comparing the distance between distributions of completed and real images in a deep feature domain. As for the traditional pixel-level and patch-level image quality metrics, including the mean ℓ_1 loss, structural similarity (SSIM), and peak signal-to-noise ratio (PSNR), we select the closest generated image to the ground truth image for calculation, as these metrics are based on one-to-one pairing.

Table 1 shows diversity and image quality analysis for different network structures. We note that our method not only improved the image quality significantly (relative 18% improvement for FID), but also generated multiple and diverse completion results. Here, BicycleGAN obtained relatively higher diversity scores than our baseline framework by using cycle loss instead of reconstruction loss. However, the completed images are of low quality (as shown in Fig. 5), which suggests that despite increased diversity, its network structure is not directly suitable for image completion.

Table 1 Quantitative comparisons of different network structures on CelebA-HQ testing set (Liu et al. 2015; Karras et al. 2017) with center masks

	Diversity (LPIPS)		Image quality ($\mathbf{I}_{out}$)			
	$\mathbf{I}_{out}$ ↑	$\mathbf{I}_{out(m)}$ ↑	ℓ_1 loss ↓	SSIM ↑	PSNR ↑	FID ↓
CA (Yu et al. 2018)	–	–	0.031	0.820	23.57	9.53
EC (Nazeri et al. 2019)	–	–	0.030	0.819	23.47	8.01
MEDFE (Liu et al. 2020)	–	–	0.028	0.830	24.38	7.85
CVAE (Sohn et al. 2015)	0.004	0.014	0.023	0.847	24.02	9.96
Instance Blind	0.015	0.049	0.025	0.852	23.77	9.48
BicycleGAN (Zhu et al. 2017b)	0.020	0.060	0.026	0.845	23.71	11.56
PICNet	**0.024**	**0.071**	**0.021**	**0.867**	**24.69**	**6.43**

The best results are highlighted in bold

↓ lower is better, ↑ higher is better. $\mathbf{I}_{out}$ is the completed output image and $\mathbf{I}_{out(m)} = (1 - M) \times \mathbf{I}_{out}$ is extracted for the missing regions. We report the traditional pixel-level and patch-level image quality metrics, including ℓ_1 loss, SSIM and PSNR. We also report the latest learning-based feature-level metrics, i.e. LPIPS and FID scores

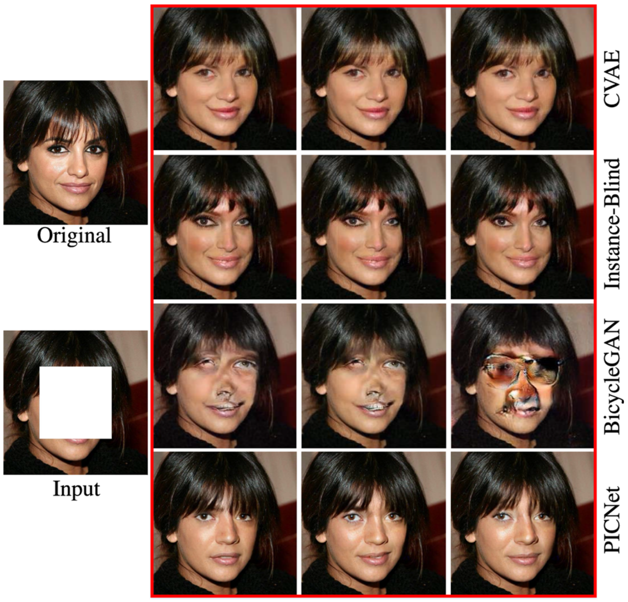


Fig. 5 Qualitative comparison results of different training strategies. *First column:* original and masked image. *Others:* the completed results of different methods. Our method provides diverse results, i.e. different hair styles and mouth expressions, with realistic appearance

Table 2 Quantitative comparisons of different variances σ^2 on ImageNet (Russakovsky et al. 2015) with free-form masks provided in Liu et al. (2018)

	Dynamic $\sigma^2(n)$		Fixed $\sigma^2 = 1$	
	LPIPS ↑	FID ↓	LPIPS ↑	FID ↓
[0.01, 0.1]	0.001	9.33	0.001	11.12
(0.1, 0.2]	0.004	15.93	0.005	17.75
(0.2, 0.3]	0.008	22.74	0.012	27.24
(0.3, 0.4]	0.015	36.23	0.021	40.98
(0.4, 0.5]	0.024	53.14	0.033	58.57
(0.5, 0.6]	0.045	78.53	0.059	86.91

The first column denotes the masked ratios. Our dynamic $\sigma^2(n)$ is adapted to the number of missing pixels n

Figure 5 shows some sampled examples of each structure. We observe that CVAE (Walker et al. 2016) obtains reasonable results, yet with little variation. The framework has likely learned to ignore the sampling and predicted a deterministic outcome as it always tries to rebuild the original ground truth during the training no matter what masks are used to degrade the input. As for "Instance Blind", If we enforced the generated image back to the original "ground truth" I_g, the experience will be similar to the CVAE (Walker et al. 2016). The visual results of BicycleGAN are much worse than other methods. In their model, the latent code $\mathbf{z}$ to the encoder is replicated from $1 \times 1 \times Z$ to $H \times W \times Z$, where the different spatial position holds the same random value that does not represent any semantic meaning. On the contrary, our latent code $\mathbf{z}$ is inferred from the visible pixels during the testing, which includes the predicted semantic information from the visible pixels.

Effect of Dynamic $\sigma^2(n)$ We compare the proposed dynamic variance $\sigma^2(n)$ to the previous fixed variance $\sigma^2 = 1$ used in VAE (Kingma and Welling 2013) and CVAE (Sohn et al. 2015). As shown in Table 2, the diversity is naturally related to the masked ratio in the first column, where larger masked regions resulted in higher diversity. We noted that the fixed version achieved higher diversity because those sampling vectors came from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ with a large range of variation, but it had much lower FID scores (on average 4.45 lower than ours). In contrast, our dynamic $\sigma^2(n)$ restrains the range of sampling vectors, which generates higher quality results with some diversity. The recent work (Peng et al. 2021) further aims to simultaneously improve the diversity and image quality.

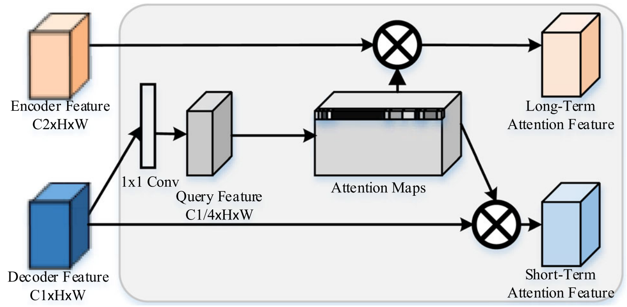


Fig. 6 Our short + long term patch attention layer. The attention map is directly computed on the decoder features to estimate the content similarity in the same domain. After obtaining the self-attention scores, we use these to compute self-attention on decoder features, as well as contextual flow on encoder features

3.2 Short + Long Term Patch Attention

A weakness of purely convolutional operations is that they have limited spatial ranges, and cannot efficiently exploit distant correlation. Extending beyond the Self-Attention in SAGAN (Zhang et al. 2018a), we propose a novel short + long term patch attention layer that not only to use the self-attention 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 visible features in the encoder or the more semantically generative features in the decoder, depending on circumstances. Our proposed structure is shown in Fig. 6.

3.2.1 Self-Patch-Attention Map

Feature attention has been widely used in image completion task (Yu et al. 2018; Yan et al. 2018; Song et al. 2018a; Yi et al. 2020). They calculate the attention map by comparing low-frequency decoder features of holes and high-frequency encoder feature of visible regions. Then, the high-frequency feature are copied from visible regions to the missing holes based on the similarity score. However, this is a little contradictory as *feature transfer is unnecessary when two features are very similar, but when needed the features are too difficult to be matched easily*.

To address this, we calculate the content similarity in itself feature domain, the decoder feature. Our attention map cal-

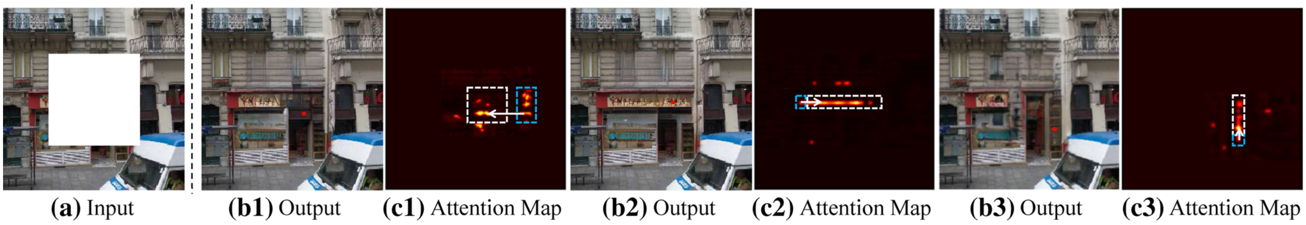
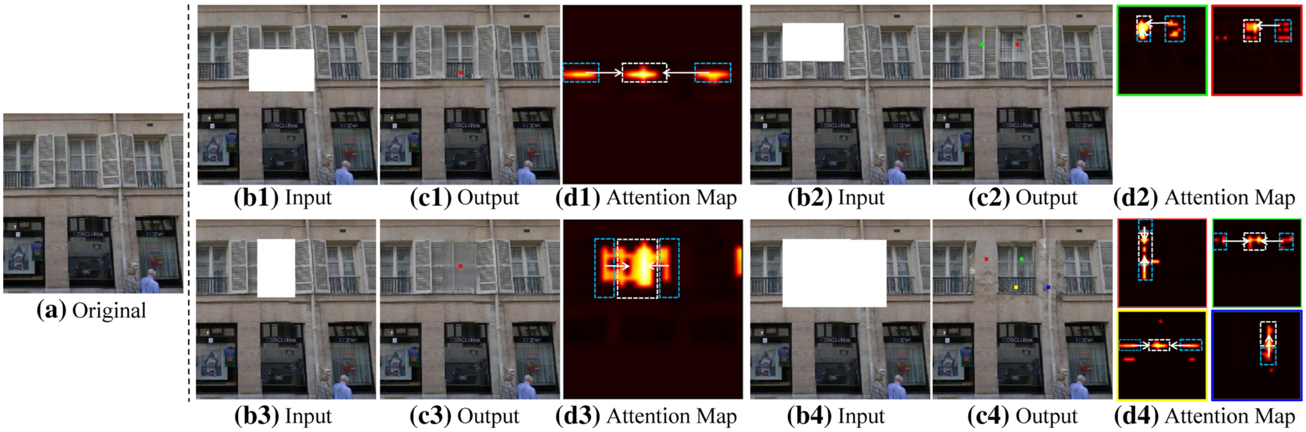

(a) Input **(b1)** Output **(c1)** Attention Map **(b2)** Output **(c2)** Attention Map **(b3)** Output **(c3)** Attention Map

Fig. 7 Texture flow (white arrow) for diversely generated contents with the same mask. **a** Masked input image. **b*** Multiple and diverse results as well as one query point (red dot). **c*** The corresponding attention maps (upsampled to original image size for visualization) for the query points in the output. The high-quality textures are copied from different visible regions (blue rectangles) to the generated regions (white rectangles), depending on what content has been generated. Here, we highlight some points with the highest attention scores (Color figure online)


(b1) Input **(c1)** Output **(d1)** Attention Map **(b2)** Input **(c2)** Output **(d2)** Attention Map

(a) Original

(b3) Input **(c3)** Output **(d3)** Attention Map **(b4)** Input **(c4)** Output **(d4)** Attention Map

Fig. 8 Texture flow (white arrow) for different masked regions. **a** Original image. **b*** Masked input images with different degraded regions. **c*** The completed results as well as query points (denoted by color dots). **d*** The corresponding attention maps for the query points in the output. The results attend to different visible regions (blue rectangles) based on the different visible content (Color figure online)

culates the response at a position in a sequence by paying attention to other position in the *same sequence*. Given the features $\mathbf{f}_d$ from the previous decoder layer, we first calculates the point attention score of:

$$\mathbf{A}_{j,i} = \frac{\exp(s_{i,j})}{\sum_{i=1}^{N}\exp(s_{i,j})}, \quad \text{where } s_{i,j} = \theta(f_{di})^{\top}\theta(f_{dj}), \tag{14}$$

where $\mathbf{A}_{j,i}$ represents the similarity of ith location to the jth location. $N = H \times W$ is the number of pixels, while θ is a 1x1 convolution filter for refining the feature.

Inspired by PatchMatch (Barnes et al. 2009), we further ensure the consistency of attention maps by fusing the similarity score in a square patch:

$$\hat{\mathbf{A}}_{j,i} = \sum_{j' \in U_j, i' \in U_i} \mathbf{A}_{j',i'} \tag{15}$$

where U_j and U_i are the neighborhood patch sets at jth and ith locations separately. We fixed the square size as 3×3 throughout this paper.

3.2.2 Short-Term Attention from Decoder Full Regions

After we obtain the attention map, the non-local information is fused in the decoder features. This leads to the short-term intra-layer attention feature (*Short-Term Attention* in Fig. 6) and the output $\mathbf{y}_d$:

$$c_{dj} = \sum_{i=1}^{N} \hat{\mathbf{A}}_{j,i} f_{di}, \quad \mathbf{y}_d = \gamma_d \mathbf{c}_d + \mathbf{f}_d \tag{16}$$

where, we use a scale parameter γ_d to balance the weights between attention feature $\mathbf{c}_d$ and decoder feature $\mathbf{f}_d$. The initial value of γ_d is set to zero.

3.2.3 Long-Term Attention from Encoder Visible Regions

In addition, specifically for image completion task, we not only need the high quality results for missing holes, but also need to ensure the appearance consistency of the generated patches of missing parts and the original patches of visible parts. Then, we introduce a long-term inter-layer attention feature (*Long-Term Attention* in Fig. 6), in which the response attends to visible encoded features $\mathbf{f}_e$. Therefore, the output $\mathbf{y}_e$ is given by:

$$c_{ej} = \sum_{i=1}^{N} \hat{\mathbf{A}}_{j,i} f_{ei}, \quad \mathbf{y}_e = \gamma_e(1 - M)\mathbf{c}_e + M\mathbf{f}_e \tag{17}$$

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 M (1 denotes visible regions, and 0 represents the holes) is used. In this way, the high-quality visible features are flowed to the holes based on the content similarity. Finally, both the short- and long-term attention features are aggregated and fed into further decoder layers.

3.2.4 Analysis

Readers may wonder why the proposed short-long term attention layer would achieve better performance than existing contextual attention layers (Yu et al. 2018; Yi et al. 2020). Here, we show that the proposed module is able to exploit non-local information from *both* visible and generated regions for the holes, instead of purely copying high-frequency information from visible regions.

In Figs. 7 and 8, completed results, along with corresponding attention maps for query points, are presented. Here, only points with the highest attention scores are highlighted. We use white arrows to explicitly show the texture flow, or how the attention layer copies information from high-quality visible features (blue rectangles) to the originally masked regions (white rectangles). In Fig. 7, we find that the proposed attention layer attends to different visible regions for differently generated content, as sampled from our model. In this way, the model ensures appearance consistency between the diversely generated appearance and the visible pixels. Figure 8 shows other examples of texture flow from visible regions to masked regions. When we mask different regions of the window, the proposed attention layer learns to copy high-quality pixels from corresponding visible regions (blue rectangles) to the missing holes (white rectangles).

We also compare the proposed attention layer to previous methods, including contextual attention (*CA*) (Yu et al. 2018) and self-attention (*SA*) (Zhang et al. 2018a) for image completion. As shown in Fig. 9, our proposed attention layer borrows features from different positions, rather than directly copying similar features from one visible position like CA. In the building scene, CA's result is of similar high quality to our method, due to the presence of repeated structures. However, in the case of faces, if the mask regions are large, both CA and SA are unable to generate high-quality results. It is worth mentioning that CA can copy high-quality pixels for skin (purple rectangle) from the visible skin, yet obtaining unrealistic eyes (blue rectangle). This is because when two eyes are masked, they cannot copy non-local similar patches from other visible parts. Conversely, SA only copies features in the decoder network, ignoring high-quality visible features. While it generates plausible appearances for skin and eyes, the generated skin is inconsistent to the visi-

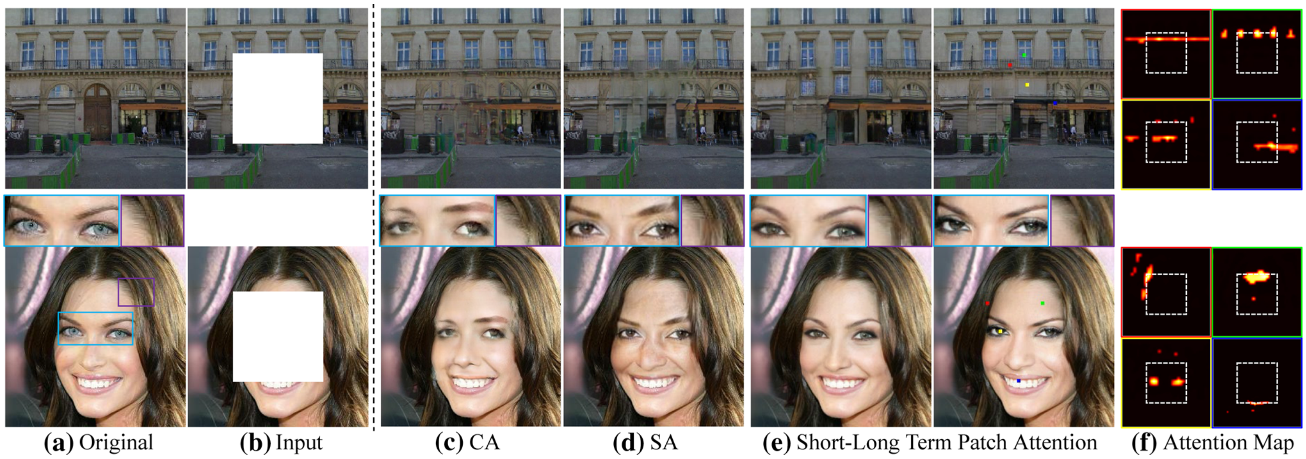


Fig. 9 Comparison of various attention modules. **a** Original image. **b** Masked input image. **c** Results of contextual attention (Yu et al. 2018). **d** Results of self-attention (Zhang et al. 2018a). **e** Multiple results of our method with short-long term patch attention. **f** The corresponding attention maps for the query points, e.g. hair (red), skin (green), eye (yellow) and teeth (blue) on the face. As can be seen, the hair point focuses more on the original visible regions (top-left attention map), while the left eye attends to the generated right eye (bottom-left attention map); the skin and teeth copy information from both visible and generated regions (right attention maps) (Color figure online)

ble skin. Our attention module is able to utilize both decoder features (which do not have masked parts) and encoder features appropriately. In completing the left eye, information is distantly shared from the decoded right eye. When it comes to completing a point in a masked hair region, it will focus on encoded features from visible hairs.

4 Experimental Results

4.1 Experimental Details

Datasets We evaluated the proposed *PICNet* with arbitrary mask types on various datasets, including Paris (Doersch et al. 2012), CelebA-HQ (Liu et al. 2015; Karras et al. 2017), ImageNet (Russakovsky et al. 2015) and Places2 (Zhou et al. 2018). Here, we only train one model to evaluate both the general free-form irregular masks and the center regular mask.

Metrics Quantitative evaluation is tricky for the pluralistic image completion task, as our goal is to get diverse but reasonable solutions for a given masked image. The original image is only one solution of many, and comparisons should not be made only based on this image. Therefore, we first used the Fréchet Inception Distance (FID) (Heusel et al. 2017) and Inception Score (IS) (Salimans et al. 2016) to assess the quality of completed image, as they are measured on learned features over the whole test set. Following Liu et al. (2018) and Nazeri et al. (2019), we then reported the traditional pixel- and patch-level image quality metrics, including ℓ_1 loss, structure similarity index (SSIM) and peak signal-to-noise ratio (PSNR). We additionally compared the visual realism of all results using human judgment, as previously proposed (Zhang et al. 2016) and widely adopted for image generation (Isola et al. 2017; Zhu et al. 2017a, b; Park et al. 2019; Nazeri et al. 2019).

Training PICNet is implemented in Pytorch v1.4. The missing regions take value 0 in the input. We highlight the missing regions as white in the figures only for visual purposes. Each mini-batch has 16 images per NVIDIA V100 GPU and each input has 1 reconstructive and 1 generative output. For the binary masks, we used randomly regular and irregular holes. However, allowing unrestricted mask sizes is more difficult than keeping to center masks as in our prior work (Zheng et al. 2019). In order to train the networks to convergence, two training steps were used: first, the completion network was trained using only the losses for the top reconstructive path, which has full information from both visible and missing regions. To do this, we estimated the missing regions' distributions that relate to different mask sizes. After we obtained the distribution of missing regions through the reconstructive path, the bottom generative path was trained to infer the distribution of missing holes based on the visible parts, from which we can generate multiple results. During optimization, the weights of different losses were set to α_{KL}=10, α_{rec}=10, α_{ad}=1.

Inference At test time, only the bottom generation path will be applied to generate *multiple* and *diverse* results based on the visible information. We sampled 50 images for each masked input image $\mathbf{I}_{\mathrm{m}}$. Note that the distribution we sampled from is also learned from the visible regions, rather than a fixed distribution used in previous works (Sohn et al. 2015; Walker

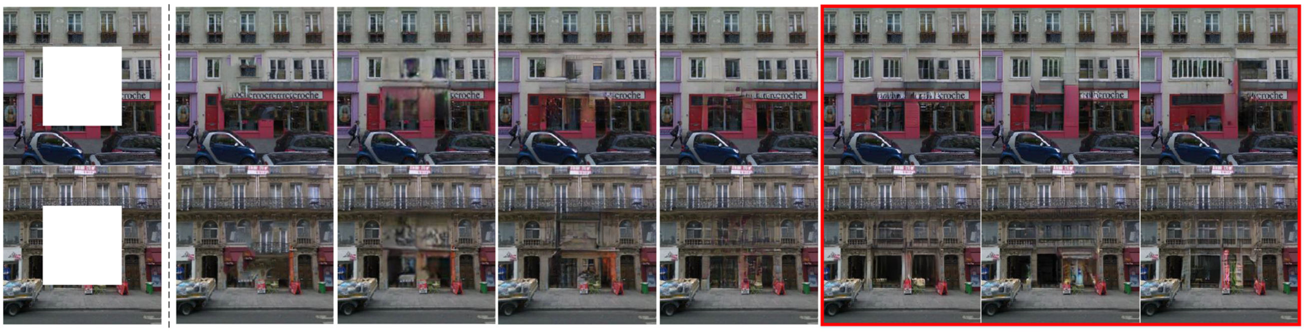


Fig. 10 Qualitative results on Paris validation set (Doersch et al. 2012) for center region completion. Here, we compare with *PM* (Barnes et al. 2009), *CE* (Pathak et al. 2016), *Shift-Net* (Yan et al. 2018) and *EC* (Nazeri et al. 2019). Note that, our *PICNet* generates different numbers of windows and varying door size with realistic appearance

et al. 2016). The visual results were automatically selected based on the higher discriminator scores.

4.2 Comparison with Existing Work

We mainly compare our method with 6 methods:

- *PM:* PatchMatch (Barnes et al. 2009), the state-of-the-art non-learning based approach.
- *CE*[2]: context Encoder (Pathak et al. 2016), the first learning-based method for large holes.
- *GL*[3]: globally and Locally (Iizuka et al. 2017), the first learning-based method for arbitrary regions.
- *CA*[4]: contextual attention (Yu et al. 2018), the first method combining learning- and patch-based methods.
- *PConv*[5]: partial convolution (Liu et al. 2018), the first learning-based method for free-form irregular holes.
- *EC*[6] *and GC*[7]: EdgeConnect (Nazeri et al. 2019) and Gated Convolution (Yu et al. 2019), the latest completion networks that use auxiliary edge information.

Compared to these approaches, our *PICNet* is the first work considering multiple solutions on various datasets for this ill-posed problem. For fair comparison among learning-based methods, we mainly reported the results with *each model trained on the corresponding dataset*. We consider the released models on the respective authors' websites to be their best performing models.

4.2.1 Center Region Completion

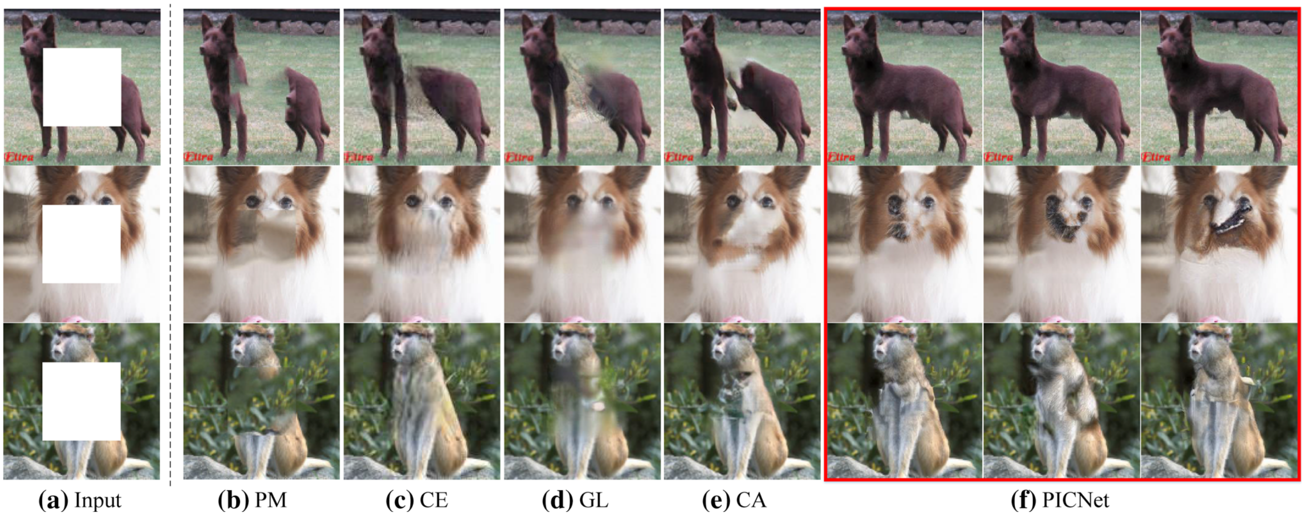
Qualitative Results In Fig. 10, we first show the visual results on the Paris dataset (Doersch et al. 2012). *PM* works by coping similar patches from visible regions and obtains good results on this dataset with repetitive structures. *CE* generates reasonable structures with blurry textures. *Shift-Net* produces better results by copying feature from visible regions to holes, which is similar to *CA* (*CA* did not release model for Paris). *EC* provides single reasonable solution. Compared to these, our *PICNet* model not only generates more natural images with high-quality, but also provides multiple results, e.g. different numbers of windows and varying door sizes.

Next, we report the performance on the more challenging ImageNet dataset (Russakovsky et al. 2015). For a fair comparison, we also used a subset of 100K training images of ImageNet to train our model as previous works (Iizuka et al. 2017). Visual results on a variety of objects from the validation set are shown in Fig. 11. These visual test images are those chosen in Iizuka et al. (2017). We note that, while learning-based methods *CE*, *GL* and *CA* provide correctly semantic results, our model is able to infer the content quite effectively. We observe that our model tries to generate full body for the first dog, and the mouth for the second dog. Meanwhile, our *PICNet* provides *multiple* and *diverse* results, from which we can choose different realistic results.

4.2.2 Free-form Region Completion

We further evaluate our model on various datasets with irregular holes proposed by Liu et al. (2018). In this testing dataset, they generate 6 categories of free-form masks with different hole-to-image area ratios: [0.01, 0.1], (0.1, 0.2], (0.2, 0.3], (0.3, 0.4], (0.4, 0.5], (0.5, 0.6]. Each has 2,000 irregular masks. Results are compared against the current state-of-the-art approaches both qualitatively and quantitatively. Results

[2] https://github.com/pathak22/context-encoder.
[3] https://github.com/satoshiiizuka/siggraph2017_inpainting.
[4] https://github.com/JiahuiYu/generative_inpainting.
[5] https://github.com/NVIDIA/partialconv.
[6] https://github.com/knazeri/edge-connect.
[7] https://github.com/JiahuiYu/generative_inpainting.



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

Fig. 11 Qualitative results and comparisons with the *PM* (Barnes et al. 2009), *CE* (Pathak et al. 2016), *GL* (Iizuka et al. 2017) and *CA* (Yu et al. 2018) on the ImageNet validation set (Russakovsky et al. 2015) . Our *PICNet* tries to generate some semantic result for the animals, even when the significant semantic information is missing

Table 3 Quantitative comparisons on ImageNet (Russakovsky et al. 2015) with free-form masks provided in Liu et al. (2018)

	Size	GL	CA	PConv	EC	PICNet
FID†	[0.01, 0.1]	10.40	12.63	11.59	**8.78**	9.33
	(0.1, 0.2]	26.42	24.63	26.46	16.75	**15.93**
	(0.2, 0.3]	50.37	39.87	47.32	28.37	**22.74**
	(0.3, 0.4]	79.01	57.44	77.16	43.74	**36.23**
	(0.4, 0.5]	108.37	76.10	91.29	63.15	**53.14**
	(0.5, 0.6]	125.41	93.55	113.62	93.43	**78.53**
IScore⋆	[0.01, 0.1]	34.66	37.33	**38.62**	38.57	38.18
	(0.1, 0.2]	31.94	34.95	31.97	**35.59**	35.36
	(0.2, 0.3]	24.26	28.79	25.53	31.06	**32.95**
	(0.3, 0.4]	17.00	22.52	18.43	26.27	**28.73**
	(0.4, 0.5]	12.13	18.35	12.43	18.94	**21.20**
	(0.5, 0.6]	8.12	13.37	10.2	12.84	**16.99**

The best results are highlighted in bold

†Lower is better

⋆Higher is better. Here, we used the top 10 samples (ranked by the discriminator score) in our models for the latest learning-based feature-level image quality evaluation

of *GL* and *CA* are obtained from their released models, which were trained only on regular random masks. Results of *EC* are also generated from their released model, which was trained on the same images and masks as ours. As *PConv* only provides the partial convolutional operation, we reproduced the model with the same masks.

Quantitative Results In Table 3, we first report the FID and IS results on the ImageNet test set (Russakovsky et al. 2015). In this setting, we used our top 10 samples of the 50 generated images for the evaluation (automatically voted using the discriminator score). As can be seen, while our multiple results are slight worse than *EC* on small mask sizes, we improve FID and IS significantly on large mask ratios, e.g. "78.53" *vs* "93.43" (*16% relative improvement*) FID for mask ratio (0.5, 0.6]. This suggests that when the mask ratios are small, it is sufficient to predict a *single* best result based on the neighboring visible pixels, yet it is not reasonable when the mask ratios are large. The latter requires our approach of generating multiple and diverse results that match the testing set distribution.

Traditional pixel- and patch-level comparison results are reported on the Places2 test set (Zhou et al. 2018) in Table 4. As these metrics require one-to-one matched images for the evaluation, we selected one sample from our multiple results, with the best balance of quantitative measures for comparison. Without bells and whistles, all instantiations of our model outperform the existing state-of-the-art models, indicating that our random samples include the close example to the original image. While the prior works (Iizuka et al. 2017; Yu et al. 2018; Liu et al. 2018; Nazeri et al. 2019; Yu et al. 2019) strongly enforce the generated images to be the same as the original images via a reconstruction loss, the testing images are not in the training set.

Qualitative Results Qualitative comparison results are visualized in Figs. 12, 13 and 14. Our PICNet is able to achieve good results for multiple solutions even under challenging conditions.

In Fig. 12, we show some results on Paris dataset. We can see that *PM* and *PConv* fail to synthesize semantic structure for large holes. The *EC* works well on the obvious structure

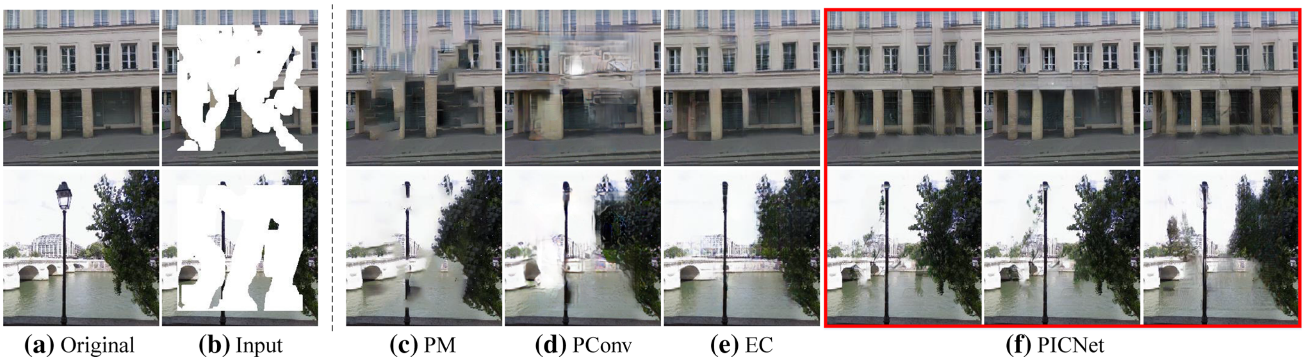


Fig. 12 Comparison of qualitative results on Paris validation set (Doersch et al. 2012) with free-form masks from PConv (Liu et al. 2018). (a) Original image. (b) Masked input. (c) Results of *PM* (Barnes et al. 2009). (d) Results of *PConv* (Liu et al. 2018). (e) Results of *EC* (Nazeri et al. 2019). (f) Our multiple and diverse results

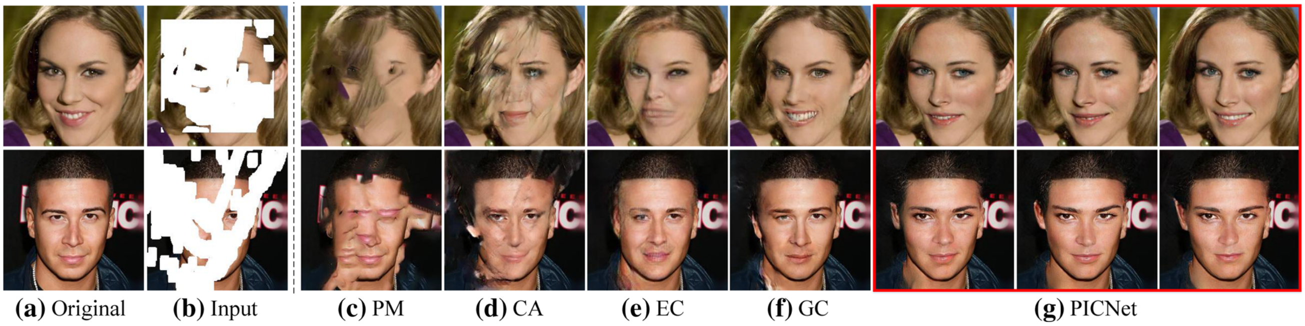


Fig. 13 Qualitative results on CelebA-HQ testing set (Liu et al. 2015; Karras et al. 2017) with free-form masks from PConv (Liu et al. 2018). **a** Original image. **b** Masked input. **c** Results of *PM* (Barnes et al. 2009). **d** Results of *CA* (Yu et al. 2018). **e** Results of *EC* (Nazeri et al. 2019). **f** Results of *GC* (Yu et al. 2019). **g** Our multiple and diverse results

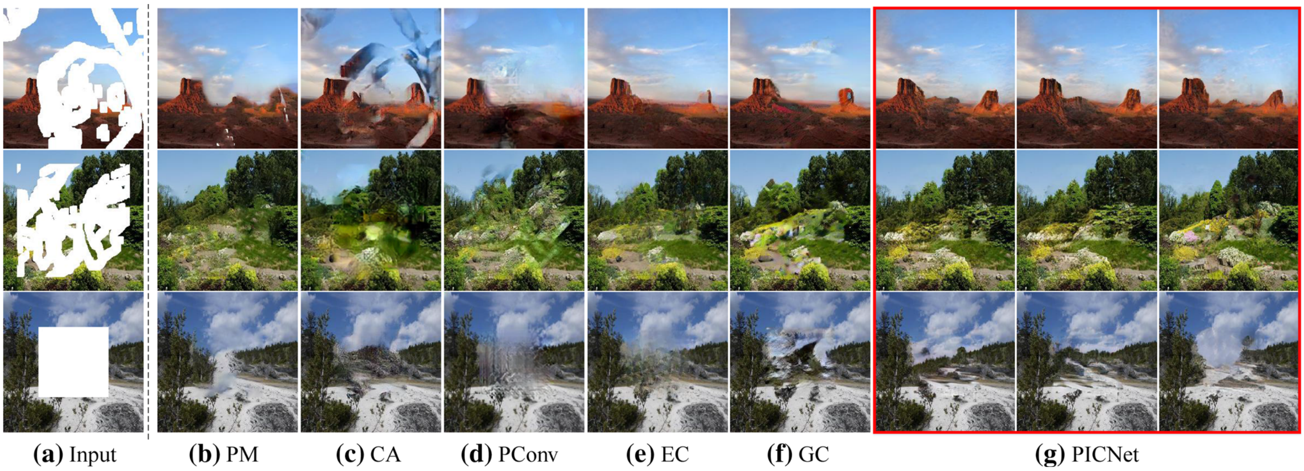


Fig. 14 Qualitative results on Place2 testing set (Zhou et al. 2018) with various masks. **a** Masked input. **b** Results of *PM* (Barnes et al. 2009). **c** Results of *CA* (Yu et al. 2018). **d** Results of *PConv* (Liu et al. 2018). **e** Results of *EC* (Nazeri et al. 2019). **f** Results of *GC* (Yu et al. 2019). **g** Our multiple and diverse results

Table 4 Quantitative comparisons over Places2 (Zhou et al. 2018) on free-form masks provided in Liu et al. (2018)

	Size	GL	CA	PConv	EC	GC	PICNet
$\ell_1 (\%)^\dagger$	[0.01, 0.1]	0.023	0.024	0.021	0.020	0.021	**0.010**
	(0.1, 0.2]	0.035	0.034	0.030	0.025	0.027	**0.016**
	(0.2, 0.3]	0.050	0.047	0.042	0.033	0.034	**0.025**
	(0.3, 0.4]	0.066	0.061	0.057	0.042	0.043	**0.035**
	(0.4, 0.5]	0.081	0.075	0.073	0.051	0.053	**0.046**
	(0.5, 0.6]	0.095	0.093	0.099	0.068	0.073	**0.064**
SSIM*	[0.01, 0.1]	0.915	0.908	0.917	0.923	0.926	**0.963**
	(0.1, 0.2]	0.853	0.845	0.859	0.878	0.886	**0.914**
	(0.2, 0.3]	0.767	0.765	0.782	0.820	0.832	**0.852**
	(0.3, 0.4]	0.682	0.691	0.704	0.760	0.771	**0.785**
	(0.4, 0.5]	0.600	0.613	0.622	0.693	0.707	**0.712**
	(0.5, 0.6]	0.529	0.532	0.513	0.599	0.603	**0.618**
PSNR*	[0.01, 0.1]	28.42	26.85	28.79	29.47	28.81	**32.26**
	(0.1, 0.2]	24.41	23.18	24.67	26.25	25.98	**27.33**
	(0.2, 0.3]	21.33	20.44	21.63	23.82	23.58	**24.44**
	(0.3, 0.4]	19.11	18.63	19.39	21.95	21.50	**22.32**
	(0.4, 0.5]	17.56	17.30	17.75	20.44	19.94	**20.71**
	(0.5, 0.6]	16.48	16.08	15.68	18.53	17.64	**18.72**

The best results are highlighted in bold

†Lower is better

*Higher is better. Here, the closest to the original ground truth samples in our method are selected for the traditional pixel- and patch-level image quality evaluation

by utilizing auxiliary edge. Our method was explicitly trained to copy information from visible parts, leading to better visual results on repetitive structures, e.g. the window in first row. Furthermore, our model provides multiple and diverse results for one given masked image.

Figure 13 shows some results on the Celeba-HQ dataset. We can see that the non-learning-based method *PM* is unable to generate reasonable semantic content in the images. While the *CA* is able to generate novel content on the face, it is not as suitable for large holes. GC further improves the results by using the learned gated convolution. *EC* results in reasonable semantic structure but blurry and inconsistent images. Our approach was explicitly trained for variable results, rather than strongly enforcing the completed image to be close to the original. Hence, our *PICNet* can provide multiple plausible results with different expressions. The online demo is also provided on our project page.[8]

In Fig. 14, we further show results on the more challenging Places2 dataset. The non-learning-based *PM* fills in reasonable pixels for natural scenes by copying similar patches from visible parts to missing holes. The *CA* only works well on regular masks as their released model was only trained on random regular masks. *EC* and *GC* generate content that is semantically reasonable but not realistic due to missing details. Instead, we can select plausible images from PICNet's multiple sampled results. Furthermore, it is hard to identify the filled-in areas in our completed images, as our short-long term patch attention copies non-local information from visible regions based on correctly predicted content.

4.2.3 Visual Turing Tests

We additionally compared the perceived visual fidelity of our model against existing approaches using human perceptual metrics, as proposed in Zhang et al. (2016). We conducted two types of user surveys: *2 alternative forced choice* (2AFCs) and *visual fidelity and perceived quality* (VFPQ). In particular, for 2AFCs, we randomly presented a generated image from an undisclosed method to the participants, and asked them to decide whether the presented image was real or fake. For quality control, we also inserted a number of real images to avoid negative testing. For VFPQ, we gave the participants a masked input and the corresponding results from all methods (blinded), and asked the participants to choose the image that was the most visually realistic. The participants were allowed to vote for multiple images simultaneously, if they felt the images were equally realistic. For each participant, we randomly presented 100 questions, consisting of 60 2AFCs examples and 40 VFPQ questions. We collected 47 valid surveys with 4700 answers.

We first show the 2FACs evaluation results in Table 5. Most participants correctly identified the real image dur-

[8] http://www.chuanxiaz.com/project/pluralistic/.

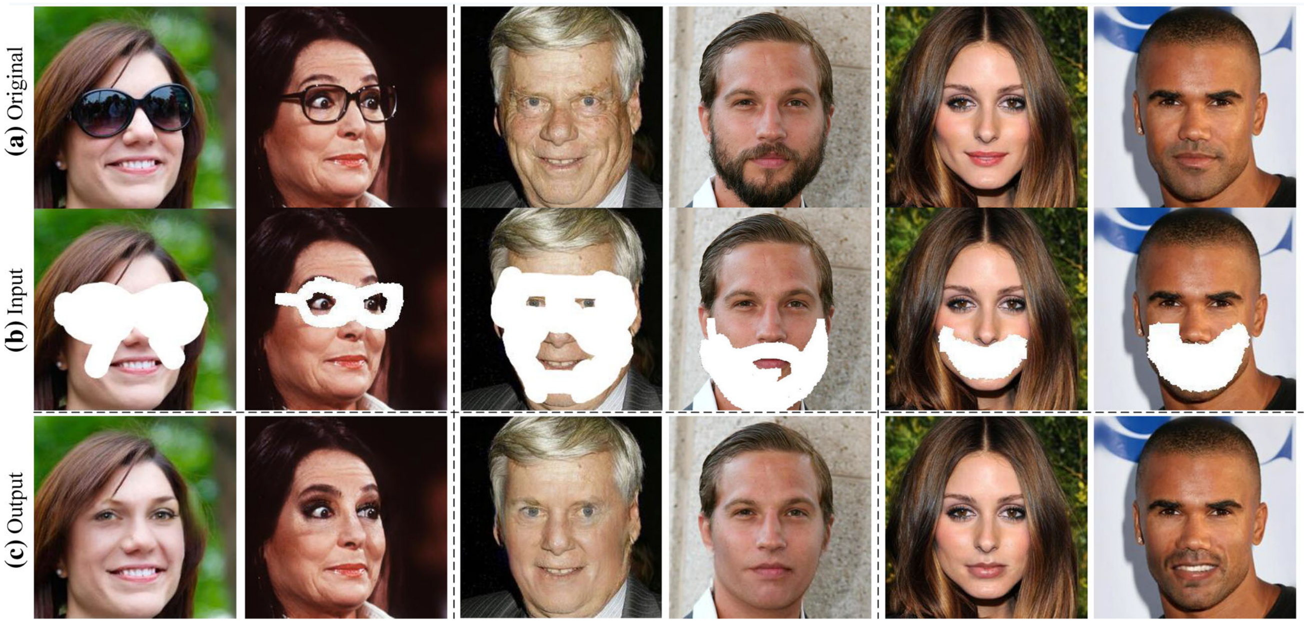


Fig. 15 Additional results of our *PICNet* on the CelebA-HQ test set (Liu et al. 2015; Karras et al. 2017) for free-form image editing. **a** Original image. **b** Masked input image. **c** Output of our *PICNet*. In the first two columns, we erased eyeglasses. Wrinkles and facial hair were removed in the next two columns. Finally, we freely changed mouth expressions. Note that due to the provision of multiple and diverse results, the users can easily select their favorite result. We refer readers to our online demo for testing

Table 5 2-alternative-forced-choice (2AFCs) score on CelebA-HQ (Liu et al. 2015; Karras et al. 2017) testing set

	GL	CA	EC	PICNet	Real
2AFC (%)	15.1 ± 1.8	17.8 ± 2.0	44.2 ± 3.7	57.0 ± 4.5	90.44 ± 1.5

All testing images were degraded by center masks. Here, the participants were required to judge whether a randomly displayed image was real *or* fake. The reported values are the percentages of images generated by each method that were judged "real"

Table 6 Visual fidelity and perceived quality (VFPQ) score on Places2 (Zhou et al. 2018) test set

	VFPQ(%)			
	[0.01, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]
GL	23.3 ± 4.3	9.8 ± 1.6	6.4 ± 1.0	4.1 ± 0.4
CA	11.4 ± 2.1	9.7 ± 1.3	7.6 ± 0.9	8.6 ± 0.9
PConv	27.8 ± 4.0	13.5 ± 1.6	11.0 ± 1.4	5.3 ± 0.7
EC	42.3 ± 6.0	38.8 ± 4.9	33.6 ± 3.2	26.7 ± 3.3
PICNet	57.5 ± 3.6	63.0 ± 4.0	69.9 ± 3.8	71.4 ± 3.4

All testing images were degraded by free-form masks provided in PConv (Liu et al. 2018). Participants selected the most realistic image from among blinded methods for the same masked input, with multiple selections allowed. Headers are ranges of mask sizes (as fraction of image). For each method, we report the percentage of trials for which it was selected, and the 95% margin of error

ing the evaluation, showing that they made conscientious discerning judgement. Our model achieved better realism scores than existing state-of-the-art methods. Table 6 shows the VFPQ evaluation results. We found that the participants strongly favored our completed results for all mask ratios, and especially so on the challenging large mask ratios. This suggests that once the visible regions do not impose strong constraints, our multiple and diverse results were naturally varied but mostly realistic and reasonable.

4.3 Additional Results

We show additional results of our proposed PICNet in Figs. 15, 16 and 17. Our approach is suitable for a wide applications, e.g. face editing, scene recomposition, object removal and outpainting.

Face Editing We first show free-form image editing on face images in Fig. 15. Our model works well for conventional object removal, e.g. removing eyeglasses in the first two columns. Next, we smoothed faces by removing wrinkles and facial hair. Finally, we changed mouth expressions by selecting an example among our multiple and diverse completed results.

High Resolution Natural Image Editing The original PICNet did not handle high resolution (HR) image completion,

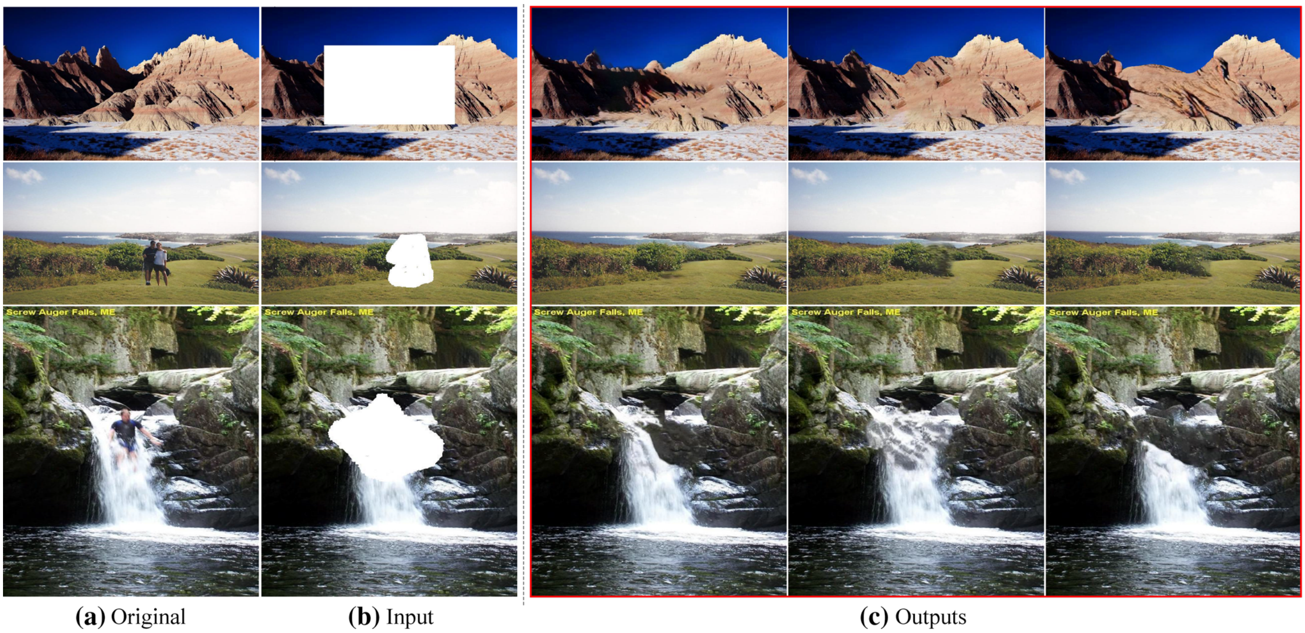


Fig. 16 Additional results of *PICNet* on the Places2 test set (Zhou et al. 2018) for free-form image editing. **a** Original image. **b** Input masked image. **c** Multiple and diverse outputs of our *PICNet*. Here, we show examples of reshaping the mountain ridge and subject removal, but, unlike conventional inpainting, we can provide multiple and diverse choices and on high resolution images

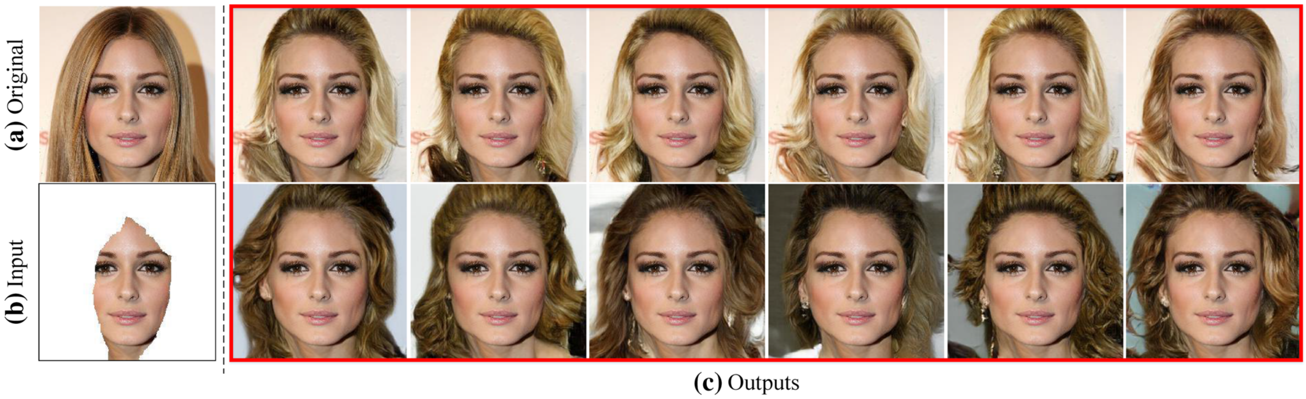


Fig. 17 Outpainting examples of our models. **a** Original image. **b** Masked input. **c** Multiple and diverse results of our *PICNet*. Note that, it provides different hair styles for the users

because the generation from a random vector $\mathbf{z}$ only works for a fixed feature size (Karras et al. 2020). However, following the *two-stage* image completion approaches (Yang et al. 2017; Yu et al. 2018; Song et al. 2018a; Yu et al. 2019; Nazeri et al. 2019; Yi et al. 2020; Zeng et al. 2020), we trained another encoder-decoder framework to refine the fixed resolution output of our PICNet. Since this work does *not* focus on HR images, we used a simple design for the refinement network by directly reapplying the PICNet framework in the second refinement stage, but without the sampling process. Note that the multiple and diverse solutions were seeded by the first content generation stage.

As can be seen in Fig. 16, our approach produces diverse results as well as visually realistic appearance for HR natural image editing, e.g. reshaping the mountain ridge and generating various mountain streams. This demonstrates that our model works well for HR images.

Outpainting In our dual pipeline framework, the masked image $\mathbf{I}_m$ and its corresponding complement image $\mathbf{I}_c$ can be easily swapped. Therefore, we randomly reversed the input mask during training on Celeba-HQ. Figure 17 shows examples where information is missing from the image border regions. This "outpainting" is a challenging task as these regions have much larger uncertainty (Iizuka et al. 2017). Note that the subject's hair can be significantly varied dur-

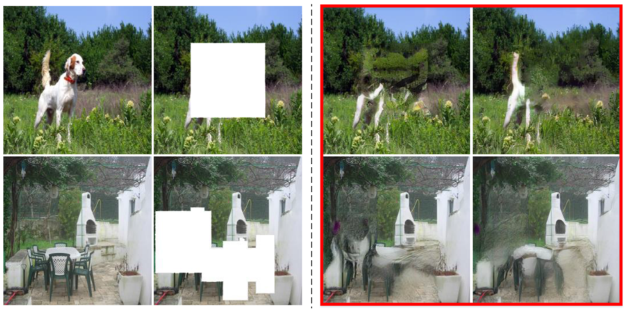


Fig. 18 Failure cases of our *PICNet*. **a** Original image. **b** Masked input. **c** Failure results of our *PICNet*, where the semantic information is heavily masked, e.g. only four legs are visible of the dogs

ing completion, suggesting that our model is applicable to style editing. Our structure has been extended to other related tasks, such as spherical image generation (Hara and Harada 2020).

4.4 Limitations

Although our model achieved better results than existing methods on various datasets by selecting images from the number of diverse sampling results, the model does not cope well with heavily structured objects with important information missing, as shown in Fig. 18. As semantic image completion is as yet an immature task that builds upon conventional image inpainting, a full understanding of semantic image content remains a challenge. In Fig. 18 top, we can see that although the four legs of dog are visible, the model cannot generate a complete dog even after multiple sampling. In the bottom image, if the content is not correctly generated, our attention model fails to provide high-quality visual results.

5 Conclusion

In this paper we have presented a novel solution to the image completion task. Unlike existing methods (Pathak et al. 2016; Iizuka et al. 2017; Yu et al. 2018, 2019; Nazeri et al. 2019; Yi et al. 2020), our probabilistically principled framework can generate multiple and diverse solutions with plausible content for a given masked image. The resulting *PICNet* shows that prior-conditional lower bound coupling is significant for conditional image generation, leading to a more reasonable two-branch training than the current deterministic structure. We also introduce an enhanced short+long term patch attention layer which improves realism by automatically attending to both high quality visible features and semantically correct generated features.

Experiments on a variety of datasets demonstrated that the multiple solutions were diverse and of high quality. On the latest learning based feature-level metrics and traditional pixel- and patch-level metrics, we demonstrated that PICNet outperformed the single-solution approaches (Iizuka et al. 2017; Yu et al. 2018; Liu et al. 2018; Nazeri et al. 2019), especially for large mask ratios with large uncertainty. We further showed in studies that users strongly favored our completed results when compared to the results in existing approaches. We additionally demonstrated that our PICNet is suitable for many interesting free-form image editing, e.g. object removal, expression changing, and scene recomposition. These multiple and diverse results can also be easily extended to HR image editing.

Acknowledgements This study is supported under the RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAFICP) Funding Initiative, as well as cash and in-kind contribution from Singapore Telecommunications Limited (Singtel), through Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU). This research is also supported by the Monash FIT Start-up Grant.

References

Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., & Verdera, J. (2001). Filling-in by joint interpolation of vector fields and gray levels. *IEEE Transactions on Image Processing*, *10*(8), 1200–1211.

Bao, J., Chen, D., Wen, F., Li, H., & Hua, G. (2017). Cvae-gan: Fine-grained image generation through asymmetric training. In *2017 IEEE international conference on computer vision (ICCV)* (pp. 2764–2773). IEEE.

Barnes, C., Shechtman, E., Finkelstein, A., & Goldman, D. B. (2009). Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (ToG)*, *28*, 24.

Bertalmio, M, Sapiro, G., Caselles, V., & Ballester. C. (2000). Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques* (pp. 417–424). ACM Press/Addison-Wesley Publishing Co.

Bertalmio, M., Vese, L., Sapiro, G., & Osher, S. (2003). Simultaneous structure and texture image inpainting. *IEEE Transactions on Image Processing*, *12*(8), 882–889.

Chen, Z., Nie, S., Wu, T., & Healey, C. G. (2018). High resolution face completion with multiple controllable attributes via fully end-to-end progressive generative adversarial networks. ArXiv preprint arXiv:180107632.

Criminisi, A., Perez, P., & Toyama, K. (2003). Object removal by exemplar-based inpainting. In *Computer vision and pattern recognition, 2003. Proceedings. 2003 IEEE computer society conference on* (Vol. 2, pp. II–II). IEEE.

Criminisi, A., Pérez, P., & Toyama, K. (2004). Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, *13*(9), 1200–1212.

Deng, Y., & Wang, J. (2020). Image inpainting using parallel network. In *2020 IEEE international conference on image processing (ICIP)* (pp. 1088–1092). IEEE.

Doersch, C, Singh, S, Gupta, A, Sivic, J, & Efros, A. (2012). What makes paris look like paris? *ACM Transactions on Graphics*, *31*(4), 1–9.

Eslami, S. M. A., Jimenez Rezende, D., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., et al. (2018). Neural scene representation and rendering. *Science*, *360*(6394), 1204–1210.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Hara, T., & Harada, T. (2020). Spherical image generation from a single normal field of view image by considering scene symmetry. ArXiv preprint arXiv:200102993.

Hays, J., & Efros, A. A. (2007). Scene completion using millions of photographs. In *ACM Transactions on Graphics (TOG)* (Vol. 26, p. 4). ACM.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems* (pp. 6626–6637).

Liu, H., Jiang, B., Song, Y., Huang, W., & Yang, C. (2020). Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Proceedings of the European conference on computer vision*.

Iizuka, S., Simo-Serra, E., & Ishikawa, H. (2017). Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, *36*(4), 107.

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5967–5976). IEEE.

Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. In *Advances in neural information processing systems* (pp. 2017–2025).

Jia, J., & Tang, C. K. (2004). Inference of segmented color and texture description by tensor voting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(6), 771–786.

Jo, Y., & Park, J. (2019). Sc-fegan: Face editing generative adversarial network with user's sketch and color. ArXiv preprint arXiv:190206838.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of Gans for improved quality, stability, and variation. ArXiv preprint arXiv:1710.10196.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8110–8119).

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. ArXiv preprint arXiv:1312.6114.

Köhler, R., Schuler, C., Schölkopf, B., & Harmeling, S. (2014). Mask-specific inpainting with deep neural networks. In *German conference on pattern recognition* (pp. 523–534). Springer.

Lee, H. Y., Tseng, H. Y., Huang, J. B., Singh, M., & Yang, M. H. (2018). Diverse image-to-image translation via disentangled representations. In *European conference on computer vision (ECCV)*.

Levin, A., Zomet, A., & Weiss, Y. (2003). Learning how to inpaint from global image statistics. In *Null* (p. 305). IEEE.

Li, Y., Liu, S., Yang, J., & Yang, M. H. (2017). Generative face completion. In *Computer vision and pattern recognition (CVPR), 2017 IEEE conference on* (pp. 5892–5900). IEEE.

Liu, G., Reda, F. A., Shih, K. J., Wang, T. C., Tao, A., & Catanzaro, B. (2018). Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision* (pp. 3730–3738).

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Smolley, S. P. (2017). Least squares generative adversarial networks. In *Computer vision (ICCV), 2017 IEEE international conference on* (pp. 2813–2821). IEEE.

Mathieu, M., Couprie, C., & LeCun, Y. (2015). Deep multi-scale video prediction beyond mean square error. ArXiv preprint arXiv:151105440.

Nazeri, K., Ng, E., Joseph, T., Qureshi, F., & Ebrahimi, M. (2019). Edgeconnect: generative image inpainting with adversarial edge learning. ArXiv preprint arXiv:190100212.

Park, E., Yang, J., Yumer, E., Ceylan, D., & Berg, A. C. (2017). Transformation-grounded image generation network for novel 3D view synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 702–711). IEEE.

Park, T., Liu, M. Y., Wang, T. C., & Zhu, J. Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2337–2346).

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A. A. (2016). Context encoders: feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2536–2544).

Peng, J., Liu, D., Xu, S., & Li, H. (2021). Generating diverse structure for image inpainting with hierarchical VQ-VAE. ArXiv preprint arXiv:210310022.

Portenier, T., Hu, Q., Szabo, A., Bigdeli, S. A., Favaro, P., & Zwicker, M. (2018). Faceshop: deep sketch-based face image editing. *ACM Transactions on Graphics (TOG)*, *37*(4), 99.

Ren, J. S., Xu, L., Yan, Q., & Sun, W. (2015). Shepard convolutional neural networks. In *Advances in neural information processing systems* (pp. 901–909).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems* (pp. 2234–2242).

Shaham, T. R., Dekel, T., & Michaeli, T. (2019). Singan: learning a generative model from a single natural image. In *Proceedings of the IEEE international conference on computer vision* (pp. 4570–4580).

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., & Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2107–2116).

Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems* (pp. 3483–3491).

Song, Y., Yang, C., Lin, Z., Liu, X., Huang, Q., Li, H., & Jay, C. (2018a). Contextual-based image inpainting: infer, match, and translate. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3–19).

Song, Y., Yang, C., Shen, Y., Wang, P., Huang, Q., & Kuo, C. C. J. (2018b). Spg-net: Segmentation prediction and guidance network for image inpainting. ArXiv preprint arXiv:1805.03356.

Walker, J., Doersch, C., Gupta, A., & Hebert, M. (2016). An uncertain future: forecasting from static images using variational autoencoders. In *European conference on computer vision (ECCV)*.

Wang, Y., Tao, X., Qi, X., Shen, X., & Jia, J. (2018). Image inpainting via generative multi-column convolutional neural networks. In *Advances in neural information processing systems* (pp. 331–340).

Yan, Z., Li, X., Li, M., Zuo, W., & Shan, S. (2018). Shift-net: image inpainting via deep feature rearrangement. In *The European conference on computer vision (ECCV)*.

Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., & Li, H. (2017). High-resolution image inpainting using multi-scale neural patch synthesis. In *The IEEE conference on computer vision and pattern recognition (CVPR)* (Vol. 1, p. 3).

Yeh, R. A., Chen, C., Lim, T. Y., Schwing, A. G., Hasegawa-Johnson, M., & Do, M. N. (2017). Semantic image inpainting with deep generative models. In *Computer vision and pattern recognition (CVPR), 2017 IEEE conference on* (pp. 6882–6890). IEEE.

Yi, Z., Tang, Q., Azizi, S., Jang, D., & Xu, Z. (2020). Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7508–7517).

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative image inpainting with contextual attention. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5505–5514).

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2019). Free-form image inpainting with gated convolution. In *Proceedings of the IEEE international conference on computer vision* (pp. 4471–4480).

Zeng, Y., Lin, Z., Yang, J., Zhang, J., Shechtman, E., & Lu, H. (2020). High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *European conference on computer vision* (pp. 1–17). Springer.

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2018a). Self-attention generative adversarial networks. ArXiv preprint arXiv:180508318.

Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *European conference on computer vision* (pp. 649–666). Springer.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018b). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhao, L., Mo, Q., Lin, S., Wang, Z., Zuo, Z., Chen, H., Xing, W., & Lu, D. (2020). Uctgan: diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5741–5750).

Zheng, C., Cham, T. J., & Cai, J. (2018). T2net: synthetic-to-realistic translation for solving single-image depth estimation tasks. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 767–783).

Zheng, C., Cham, T. J., & Cai, J. (2019). Pluralistic image completion. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: a 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464.

Zhou, T., Tulsiani, S., Sun, W., Malik, J., & Efros, A. A. (2016). View synthesis by appearance flow. In *European conference on computer vision* (pp. 286–301). Springer.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017a). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232).

Zhu, J. Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., & Shechtman, E. (2017b). Toward multimodal image-to-image translation. In *Advances in neural information processing systems* (pp. 465–476).