Pluralistic Image Completion with Gaussian Mixture Models

Xiaobo Xia[1*] **Wenhao Yang**[2*] **Jie Ren**[3]
Yewen Li[4] **Yibing Zhan**[5] **Bo Han**[6] **Tongliang Liu**[1†]
[1]TML Lab, University of Sydney [2]Nanjing University [3]University of Edinburgh
[4]Nanyang Technological University [5]JD Explore Academy [6]Hong Kong Baptist University

Abstract

Pluralistic image completion focuses on generating both *visually realistic* and *diverse* results for image completion. Prior methods enjoy the empirical successes of this task. However, their used constraints for pluralistic image completion are argued to be *not well interpretable* and *unsatisfactory* from two aspects. First, the constraints for visual reality can be *weakly correlated* to the objective of image completion or even *redundant*. Second, the constraints for diversity are designed to be *task-agnostic*, which causes the constraints to not work well. In this paper, to address the issues, we propose an end-to-end probabilistic method. Specifically, we introduce a unified *probabilistic graph model* that represents the complex interactions in image completion. The entire procedure of image completion is then mathematically divided into several sub-procedures, which helps efficient enforcement of constraints. The sub-procedure directly related to pluralistic results is identified, where the interaction is established by a *Gaussian mixture model* (GMM). The inherent parameters of GMM are *task-related*, which are *optimized adaptively* during training, while the number of its primitives can control the diversity of results conveniently. We formally establish the effectiveness of our method and demonstrate it with comprehensive experiments. The implementation is available at https://github.com/tmllab/PICMM.

1 Introduction

Pluralistic image completion refers to the task of filling in the missing region of an incomplete image, so as to produce *visually realistic* and *diverse* image completion solutions [59–61]. Different from single image completion [39, 12, 55, 50] that learns a *deterministic* mapping from an incomplete image to a complete image, and produces a *unique* result, pluralistic image completion can generate *various* results with visually realistic contents. Pluralistic image completion follows the fact that image completion is a *highly subjective* process and benefits a series of applications such as photo restoration [27], object removal [13], and transmission error concealment [37].

Pluralistic image completion is complicated and challenging. The state-of-the-art methods encode images to latent features. To constrain the visual reality of results, *heuristic constraints* are introduced, *e.g.*, the perceptual loss [27] and attention module [59, 44]. To constrain the diversity of results, the distribution of latent features is assumed to be a *unimodal* Gaussian distribution with *predefined parameters* [59]. Image completion is performed by *sampling* from the unimodal Gaussian distribution.

Although the above paradigm enjoys empirical successes of pluralistic image completion, it is *not well interpretable* and *unsatisfactory* in practice. We detail the issues from two aspects. First, the

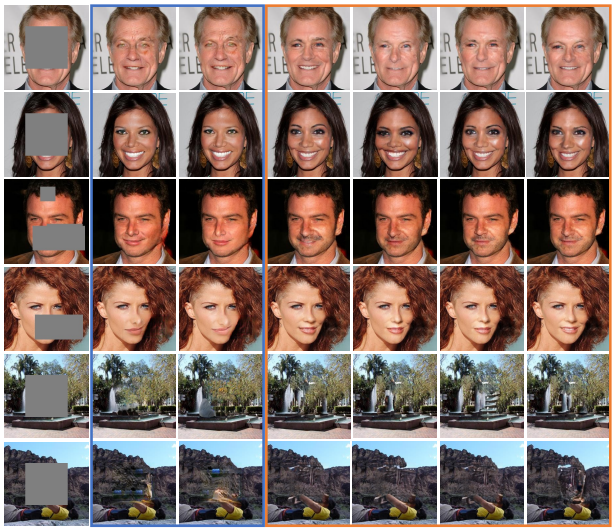
[*]The first two authors made equal contributions.

[†]Corresponding author: Tongliang Liu (tongliang.liu@sydney.edu.au).

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

added constraints for the visual reality of results are based on *general purposes*. These constraints may work well in some tasks. Unfortunately, there is no clear understanding of what role they play in pluralistic image completion. Inappropriate constraints are likely to have side effects. It is not easy to finish the determination of added constraints as our desideratum, especially in complex image completion tasks. Second, the diversity of image completion results is hard to be constrained in reason. For a specific task, the parameters of the unimodal Gaussian distribution are directly related to the diversity. Nevertheless, these parameters are designed to be *prior knowledge* and *task-agnostic*, which cause *unreasonable pluralistic results*.

In this paper, to address the above issues, we present an end-to-end probabilistic method. The method is constructed with a unified *probabilistic graph model* (PGM) [17, 22] and translates the problem of pluralistic image completion into a structured mathematical representation. Specifically, we suppose that pluralistic image completion can be represented with a directed acyclic graph that is designed reasonably. The nodes of the graph denote different kinds of images and corresponding latent features. The edges of the graph represent the interactions between nodes. Based on this directed acyclic graph, the entire procedure of pluralistic image completion is divided into several sub-procedures. Then, we perform different constraints for different sub-procedures as our desideratum.



(a) Input (b) PIC (c) Ours

Figure 1: **Example completion results of the proposed method on the images of faces and natural sceneries with various masks** (missing regions are shown in gray). The baseline PIC refers to [59]. For each group, the masked input image is shown left, followed by diverse and plausible completion results from our method without any post-processing. (Zoom in to see the details.)

In particular, the diversity of image completion results is implemented by modeling the interaction between the latent features of the missing region and incomplete image with a *Gaussian mixture model* (GMM) [40, 64]. Compared with the unimodal Gaussian distribution, GMM has a larger capacity and is more competitive to meet the output diversity [56, 38]. For our method, the inherent parameters of GMM are *task-related* and *optimized adaptively* during training. The different primitives of GMM represent the outputs with different patterns. Additionally, the number of primitives can be chosen artificially with the needs for diversity in real-world applications.

Before delving into details, we clearly emphasize our contributions as follows:

- To our best knowledge, this paper is the first one that uses a PGM for pluralistic image completion. Besides, the PGM is specially designed with justifications, rather than borrowing existing models. With the designed PGM, the added constraints for visual reality and diversity are explainable.
- We propose to use GMM to increase result diversity. One remarkable advantage is that the inherent parameters for diversity are task-related, leading to more reasonable results.
- We conduct a series of experiments on benchmark datasets to support our claims. In both qualitative and quantitative comparisons with the state-of-the-art methods, our method achieves superior performance. The generated contents for image completion are both visually realistic and diverse, such as those shown in Figure 1. Codes will be released for the reproducibility of results.

The rest of the paper is organized as follows. In Section 2, we briefly review pluralistic image completion. In Section 3, we discuss the proposed method step by step. Experimental results are provided in Section 4. Finally, we conclude the paper in Section 5.

2 Preliminaries

Problem Setting. We first discuss image completion. For the task of image completion, suppose we have an image, originally $\mathbf{I}_o$, but degraded by some missing pixels to become a masked partial image $\mathbf{I}_m$. The image that comprises the original missing pixels is called a complement partial image, which can be denoted by $\mathbf{I}_c$. The goal of the image completion task is to reconstruct $\mathbf{I}_o$ by using $\mathbf{I}_m$. The illustrations for different kinds of images are provided in Figure 2. Traditional single image completion reconstructs the original image in a *deterministic* fashion, which results in only a single solution [12, 9].

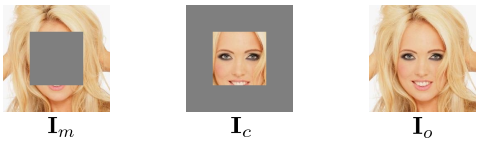


Figure 2: The illustrations for different kinds of images in image completion.

Pluralistic image completion [59] is motivated by the fact that image completion is a highly subjective process. Also, completed images by different experts may agree on high-level semantics, but have substantially different details [59]. Therefore, in pluralistic image completion, we need to generate *multiple* and *diverse* plausible results when presented with a masked partial image. The problem of pluralistic image completion is challenging since we need to take care of the diversity and visual authenticity of results at the same time.

Prior Work. Traditional methods on single image completion, *e.g.,* diffusion-based methods [3, 1, 23] and patch-based methods [2, 4, 5], assume that image holes share similar content to visible regions. Therefore, they choose to directly propagate the contextual appearances or realign the background patches to complete the image holes. Benefiting from the strong power of deep neural networks, deep learning based single image completion methods achieve promising completion performance, *e.g.,* employing shepard convolutional neural networks [39], restraining global and local consistency between image holes and visible regions [16], performing deep feature rearrangement [47], and learning a pyramid-context encoder network [53], *etc*. In addition to these methods, multiple single image completion methods involve the generative adversarial networks [10] to learn the semantic of images [48, 49], *e.g.,* [25, 51, 8, 6, 57, 12, 9]. Although these single-solution methods achieve outstanding performance in predicting deterministic result for image holes, they cannot generate various semantically meaningful results.

Existing works for pluralistic image completion usually rely on generative models [24], *e.g.*, CVAE [43, 13, 59], hierarchical VQ-VAE [37], PD-GAN [27], and BicycleGAN [63]. Generative pluralistic image completion methods can diversify meaningful completion results with different kinds of constraints. However, such methods are not well interpretable as discussed. The issues largely limit their applications in the real world.

3 Methodology

This section presents a probabilistic method for pluralistic image completion. We first design a probabilistic graph model for our task (Section 3.1). Then, based on the probabilistic graph model, we show how to divide the entire procedure of pluralistic image completion into several sub-procedures (Section 3.2). Afterward, we discuss how to use GMM to diversify image completion results (Section 3.3). Finally, we add reconstruction and adversarial losses to strengthen image completion, and summarize all algorithm flows (Section 3.4).

3.1 Probabilistic Graph Model Construction

Recall the notations in image completion, *i.e.*, $\mathbf{I}_o$, $\mathbf{I}_m$, and $\mathbf{I}_c$, they correspond to the latent features $\mathbf{z}_o$, $\mathbf{z}_m$, and $\mathbf{z}_c$ respectively. We suppose a probabilistic graph model for the image completion task, which is shown in Figure 3. The probabilistic graph model is designed reasonably. The graphical illustration of overall computational paths of the probabilistic graph model is presented in Figure 3(d). Specifically, the images $\mathbf{I}_m$ and $\mathbf{I}_c$ form $\mathbf{I}_o$. Their latent features $\mathbf{z}_m$ and $\mathbf{z}_c$ can be exploited to form $\mathbf{I}_o$. Given $\mathbf{z}_m$, we can surmise $\mathbf{z}_c$. We use the latent features here, since they are low-dimensional and informative [11, 21, 45, 46]. We further details Figure 3 as follows.

(a) Inference (b) Dynamic Inference (c) Generation (d) Overall

Figure 3: The illustrations of the probabilistic graph model for the pluralistic image completion task.

Inference. We do not directly map $\mathbf{I}_m$ and $\mathbf{I}_c$ to $\mathbf{I}_o$, since they are too high-dimensional and have information redundancy. We exploit generative models. The processes $\mathbf{I}_m \rightarrow \mathbf{z}_m$ and $\mathbf{I}_c \rightarrow \mathbf{z}_c$ denote that we *encode* images into corresponding latent features. During training, we learn the encoder f to finish the inference.

Dynamic Inference. Given $\mathbf{I}_m$ (*resp.* $\mathbf{I}_c$), we can infer the latent feature $\mathbf{z}_m$ (*resp.* $\mathbf{z}_c$). Afterward, $\mathbf{z}_m \dashrightarrow \mathbf{z}_c$ means that we use a variational distribution $p(\mathbf{z}_c|\mathbf{z}_m)$ to approximate the distribution $p(\mathbf{z}_c|\mathbf{I}_c)$. Here we denote this relationship using a dashed line, which is designed to be dynamic to diversify outputs.

Generation. In image completion, the images $\mathbf{I}_m$ and $\mathbf{I}_c$ can compose the image $\mathbf{I}_o$. Accordingly, for their latent features, we splice $\mathbf{z}_m$ and $\mathbf{z}_c$ to output $\mathbf{I}_o$. During training, we learn the decoder g to finish the generation.

3.2 Image Completion Objective Decomposition

We cannot observe the underlying $\mathbf{I}_c$ in the test procedure. Based on the probabilistic graph model, we perform the following variational inference with the Kullback-Leibler (KL) divergence:

$$L_{\mathrm{P}} = \mathrm{KL}\left[q_\psi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m) \| p_\phi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m, \mathbf{I}_c)\right], \tag{1}$$

where $q_\psi(\cdot|\cdot)$ and $p_\phi(\cdot|\cdot)$ are two conditioned distributions, with ψ and ϕ being the parameters of their corresponding functions. The divergence (1) is minimized with respect to all parameters. It should be noted that, in fact, the overall objective should be $\mathrm{KL}[q_\psi(\mathbf{I}_o|\mathbf{I}_m)\|p_\phi(\mathbf{I}_o|\mathbf{I}_m, \mathbf{I}_c)] = \mathrm{KL}[\mathbb{E}_{\mathbf{z}_m,\mathbf{z}_c}[q_\psi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m)]\|\mathbb{E}_{\mathbf{z}_m,\mathbf{z}_c}[p_\phi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m, \mathbf{I}_c)]]$. As the expectation in this formula is *intractable*, we need to deal with the 1-step Monte Carlo sampling objective, which is why the divergence (1) is formed. Another advantage of using the divergence (1) is that constraining each sampling point can make the overall constraint tighter. We further decompose the divergence (1) based on the probabilistic graph model as follows.

Proposition 1 *Regarding the KL divergence (1), we show that the divergence can be decomposed as:*

$$\begin{aligned} L_{\mathrm{P}} &= \mathrm{KL}\left[q_\psi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m) \| p_\phi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m, \mathbf{I}_c)\right] \\ &= \underbrace{\mathbb{E}_{(\mathbf{z}_m,\mathbf{z}_c)\sim q_\psi(\mathbf{z}_m,\mathbf{z}_c|\mathbf{I}_m)} \mathrm{KL}\left[q_\psi(\mathbf{I}_o|\mathbf{z}_m, \mathbf{z}_c) \| p_\phi(\mathbf{I}_o|\mathbf{I}_m, \mathbf{I}_c)\right]}_{ⓐ} \\ &+ \underbrace{\mathbb{E}_{\mathbf{z}_m\sim q_\psi(\mathbf{z}_m|\mathbf{I}_m)} \mathrm{KL}\left[q_\theta(\mathbf{z}_c|\mathbf{z}_m) \| p_\phi(\mathbf{z}_c|\mathbf{I}_c)\right]}_{ⓑ} + \underbrace{\mathrm{KL}\left[q_\psi(\mathbf{z}_m|\mathbf{I}_m) \| p_\phi(\mathbf{z}_m|\mathbf{I}_m)\right]}_{ⓒ}, \end{aligned} \tag{2}$$

where q_θ is the variational distribution, and the parameters θ are involved in divergence minimization.

The proof can be found in Appendix A.1. We analyze three terms in Eq. (2) combining the probabilistic graph model and our purpose:

- For ⓐ, as discussed, we can generate $\mathbf{I}_o$ with $\mathbf{I}_m$ and $\mathbf{I}_c$, or with their latent features $\mathbf{z}_m$ and $\mathbf{z}_c$. We minimize the distance about the generation of $\mathbf{I}_o$ in the two ways. The minimization can ensure that the generation from the images and latent features is *consistent*.
- For ⓑ, we do not limit the inference from $\mathbf{I}_c$ to $\mathbf{z}_c$ as a deterministic function. Instead, to diversify outputs, we use the variational distribution to restrain $q_\theta(\mathbf{z}_c|\mathbf{z}_m)$. The choice and discussion about the variational distribution will be provided in Section 3.3.

- For ⓒ, it aims to keep that the inference using the variational distribution and the inference using the posterior is *close*, which also guarantees the reliability of sampling. Besides, the variational distribution does not include $\mathbf{I}_c$ to enable latent features to be used in test. We employ the variational evidence lower bound (ELBO) L_{ELBO} [21, 15] to constrain ⓒ. It is because, suppose that $p(\mathbf{I}_m)$ is a constant given $\mathbf{I}_m$, it is equivalent for maximizing the L_{ELBO} and minimizing ⓒ. The details are provided in Appendix A.4.

3.3 Gaussian Mixture Model

As discussed, we need a suitable choice for the variational distribution $q_\theta(\mathbf{z}_c|\mathbf{z}_m)$ to output pluralistic results. In this work, we employ the Gaussian Mixture Model (GMM) [30, 31] for the dynamic mode $q_\theta(\mathbf{z}_c|\mathbf{z}_m)$, which is more flexible than a unimodal and can diversify outputs.

We denote the latent features of $\mathbf{I}_c$ that are inferred from $\mathbf{z}_m$ by $\hat{\mathbf{z}}_c$, which is distinguished from $\mathbf{z}_c$ achieved by the encoder. For the latent features $\mathbf{z}$, we denote the mean and covariance by $\boldsymbol{\mu}^{\mathbf{z}}$ and $\boldsymbol{\Sigma}^{\mathbf{z}}$. Mathematically, we have

$$q_\theta(\mathbf{z}_c|\mathbf{z}_m) = \sum_{i=1}^{k} \alpha_i \mathcal{N}\left(\mathbf{z}_c|\boldsymbol{\mu}_i^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_i^{\hat{\mathbf{z}}_c}\right), \tag{3}$$

where k denotes the number of primitives, α_i denotes the mixing coefficient of the i-th primitive, and $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ denotes the distribution parameters of the i-th primitive. Note that the mixing coefficient α_i is the parameter of the Categorical distribution [31]. Intuitively, in Eq. (3), we model $\mathbf{z}_m$ with GMM, where the i-th primitive can be used to represent $\hat{\mathbf{z}}_c^{(i)}$. In this way, we can obtain different $\hat{\mathbf{z}}_c$ to meet diversity requirements.

Besides, among all the parameters of GMM, except for k, the other parameters are called the *inherent parameters* of GMM. In GMM (3), the number of primitives k can be determined artificially according to the need for output diversity. The inherent parameters are optimized adaptively during training. Specifically, for the optimization of the mixing coefficients $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k]$, we exploit the the *frequency loss* [38]. The objective is

$$\begin{aligned} &L_{\text{F}} = (\mathbf{v} - \boldsymbol{\alpha})(\mathbf{v} - \boldsymbol{\alpha})^\top, \mathbf{v} = [v_1, \ldots, v_k], \text{and} \\ &v_j = \begin{cases} 1, j = \underset{i}{\operatorname{argmin}}\, \text{KL}\left[\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_i^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_i^{\hat{\mathbf{z}}_c}) \| \mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}^{\mathbf{z}_c}, \boldsymbol{\Sigma}^{\mathbf{z}_c})\right]; \\ 0, \text{ otherwise.} \end{cases} \end{aligned} \tag{4}$$

For our task, with the frequency loss (4), the accumulated frequency approximate gradient is an *asymptotically unbiased estimation* of the true gradient. We provide the detailed derivation and analysis in Appendix A.2. The reason we mention the frequency loss is that, the metric for evaluating the performance of each primitive is adapted to KL divergence in our task, which is different from the original metric in [38]. For the optimization of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of $\mathbf{z}_m$, we utilize the *back-propogate-max-operation* to reserve the distinguishable property of GMM (3). For our problem, the objective of the back-propogate-max-operation is shown in the following proposition.

Proposition 2 *By modeling the variational distribution with GMM (3), we have*

$$L_{\text{BM}} = -\frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}|}{|\boldsymbol{\Sigma}^{\mathbf{z}_c}|} + \frac{1}{2}\text{tr}\left((\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1}\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}\right) - \frac{1}{2}(\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c} - \boldsymbol{\mu}^{\mathbf{z}_c})^\top(\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1}(\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c} - \boldsymbol{\mu}^{\mathbf{z}_c}), \tag{5}$$

where $j = \text{argmin}_i\, \text{KL}\left[\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_i^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_i^{\hat{\mathbf{z}}_c}) \| \mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}^{\mathbf{z}_c}, \boldsymbol{\Sigma}^{\mathbf{z}_c})\right]$, and $\text{tr}(\cdot)$ denotes the trace of a matrix. The derivation of Proposition 2 can be found in Appendix A.3. For Proposition 2, if we minimize L_{BM}, we optimize $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of $\mathbf{z}_m$, which reduces the distance between $\hat{\mathbf{z}}_c^{(j)}$ and $\mathbf{z}_c$. For the term ⓑ in Eq. (2), we then have $L_{\text{GMM}} = \mathbb{E}_{\mathbf{z}_m \sim q_\psi(\mathbf{z}_m|\mathbf{I}_m)}\text{KL}\left[q_\theta(\mathbf{z}_c|\mathbf{z}_m) \| p_\phi(\mathbf{z}_c|\mathbf{I}_c)\right] \approx L_{\text{F}} + L_{\text{BM}}$.

Discussion. Note that the diversity of results is directly related to $\mathbf{z}_m$. When GMM for $\mathbf{z}_m$ is *dominated* by the i-th primitive ($\alpha_i \approx 1$), the diversity of image completion results is reduced. Although this phenomenon did not appear in our experiments, it may exist in practice. Actually, this phenomenon is reasonable. From human cognition, if $\mathbf{I}_m$ shows *only one pattern*, the inference for $\mathbf{I}_c$ will be deterministic [3]. For example, if the size of the missing region of an incomplete image is

very small, we can infer the missing region both deterministically and reasonably, where the diversity of results is reduced. Prior methods with task-agnostic constraints do not consider this fact, which may create *unreasonable results*. For our method, the constraints for diversity are task-related, which can take this phenomenon into account.

Algorithm 1 Training procedure

Input: images $\mathbf{I}_o$, $\mathbf{I}_m$, and $\mathbf{I}_c$, the number of primitives of GMM k, the initialized encoder f, and decoder g.

1: **Encode** $\mathbf{I}_m$ (*resp.* $\mathbf{I}_c$) to $\mathbf{z}_m$ (*resp.* $\mathbf{z}_c$) with f;
2: **Model** $\mathbf{z}_m$ with GMM and $\mathbf{z}_c$ with a unimodal Gaussian distribution;
3: **Calculate** the loss L_{GMM} as discussed in Section 3.3;
4: **Update** the parameters of GMM with L_{GMM};
5: **Infer** $\hat{\mathbf{z}}_c^{(j)}$ from $\mathbf{z}_m$ with Eq. (3) and Eq. (4), $j = 1, \ldots, k$;
6: **Generate** images by g with $\hat{\mathbf{I}}_o = g(\mathbf{z}_m, \mathbf{z}_c)$ and $\hat{\mathbf{I}}_o^{(j)} = g(\mathbf{z}_m, \hat{\mathbf{z}}_c^{(j)})$, $j = 1, \ldots, k$;
7: **Calculate** the loss L_{ELBO} and L_{C} as discussed in Section 3.2 and 3.4;
8: **Update** all the parameters *w.r.t.* L_{ELBO} and L_{C}.

Output: the trained encoder f^* and decoder g^*.

Algorithm 2 Test procedure

Input: the image $\mathbf{I}_m$, the number of primitives of GMM k, the trained encoder f^*, and decoder g^*.

1: **Encode** $\mathbf{I}_m$ to $\mathbf{z}_m$ with f^*;
2: **Model** $\mathbf{z}_m$ with GMM;
3: **for** $j = 1$ to k **do**
4: **Sample** i with $i \sim \texttt{Categorical}(\alpha_1, \ldots, \alpha_k)$;
5: **Infer** $\hat{\mathbf{z}}_c^{(j)}$ with $\hat{\mathbf{z}}_c^{(j)} = \mathcal{N}(\boldsymbol{\mu}_i^{\mathbf{z}_m}, \boldsymbol{\Sigma}_i^{\mathbf{z}_m})$;
6: **Generate** images $\hat{\mathbf{I}}_o^{(j)}$ by g^* with $\hat{\mathbf{I}}_o^{(j)} = g^*(\mathbf{z}_m, \hat{\mathbf{z}}_c^{(j)})$.
7: **end for**

Output: pluralistic image completion results $\{\hat{\mathbf{I}}_o^{(j)}\}_{j=1}^k$.

3.4 Reconstruction and Adversarial Losses

To ensure the minimization of the term ⓐ in Eq. (2), we incorporate the use of reconstruction and adversarial losses [10]. The reconstruction loss is formulated as:

$$L_{\text{R}} = \mathbb{E}\left[\|\hat{\mathbf{I}}_o - \mathbf{I}_o\|_1\right] + \mathbb{E}\left[\|\hat{\mathbf{I}}_m^{(j)} - \mathbf{I}_m\|_1\right], \tag{6}$$

where $\hat{\mathbf{I}}_o$ is generated from $\mathbf{z}_m$ and $\mathbf{z}_c$, and $\hat{\mathbf{I}}_m^{(j)}$ is generated from $\hat{\mathbf{z}}_c^{(j)}$. Here, $\hat{\mathbf{z}}_c^{(j)}$ can be inferred from $\mathbf{z}_m$ by using Eq. (3) and Eq. (4). The image $\hat{\mathbf{I}}_m^{(j)}$ is obtained from $\hat{\mathbf{I}}_o^{(j)}$ with the image mask that can degenerate $\mathbf{I}_o$ to $\mathbf{I}_m$. The image $\hat{\mathbf{I}}_o^{(j)}$ is generated from $\mathbf{z}_m$ and $\hat{\mathbf{z}}_c^{(j)}$. The reconstruction loss (6) controls the visual rationality of image completion results. Also, we efficiently avoid the deterministic fashion with GMM. Furthermore, we leverage adversarial training [10] to make the generated images more realistic. The adversarial loss L_{A} is formulated as

$$L_{\text{A}} = \mathbb{E}\left[\|\mathcal{D}(\hat{\mathbf{I}}_o) - 1\|_2^2\right] + \mathbb{E}\left[\|\mathcal{D}(\hat{\mathbf{I}}_o^{(j)}) - \mathcal{D}(\mathbf{I}_o)\|_2^2\right], \tag{7}$$

where $\mathcal{D}$ is the discriminator optimized by the discriminator loss based on LSGAN [29]. We jointly train our encoder f and decoder g through the following combined loss:

$$L_{\text{C}} = L_{\text{R}} + \lambda_{\text{A}} L_{\text{A}}, \tag{8}$$

where the weight λ_{A} is set to 0.05 in all experiments. Then, the final objective of the loss function is the sum of three losses, *i.e.*, $L_{\text{FINAL}} = L_{\text{GMM}} + L_{\text{ELBO}} + L_{\text{C}}$. The final loss is to maximize the log-likelihood of the conditional data distribution. Specifically, the adversarial loss uses the Wasserstein distance to make generated images realistic given latent features. The ELBO loss uses KL divergence and distribution log-likelihood to make variational posterior approximate real posterior.

Algorithm Flows. For the convenience of following technical details, we provide the algorithm flows. The algorithm flows of training and test stages can be found in Algorithm 1 and Algorithm 2. The illustration is shown in Appendix B.

It should be noted that the number of primitives k does not mean that we are limited to generate only k diverse results given $\mathbf{I}_m$ (in Steps 3 and 4 of Algorithm 2). In fact, we can sample from the k primitives to obtain lots of image completion results. Moreover, as GMM has a much larger capacity than the unimodal Gaussian distribution [31], we can achieve greater diversity. We provide empirical observations in Section 4.3 and Appendix C.

4 Experiments

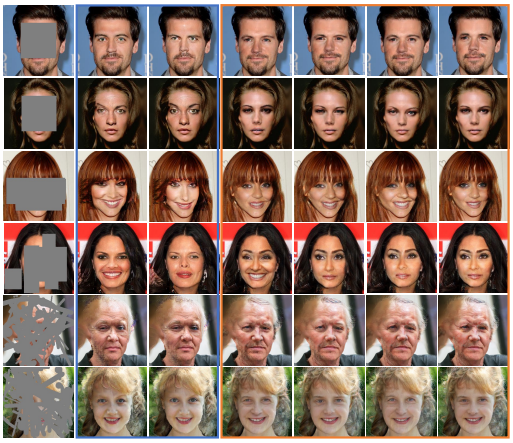
In this section, we conduct a series of experiments to justify our claims. We first introduce the implementation of our method (Section 4.1). The comprehensive experimental results and comparison with advanced methods are then provided and discussed (Section 4.2). Finally, we conduct an analysis study to present and discuss our method in more detail (Section 4.3).

4.1 Implementation Details

Datasets. We evaluated our proposed model on five popularly used datasets, *i.e.*, CelebA-HQ [18, 28], FFHQ [19], Paris StreetView [7], Places2 [62], and ImageNet [41]. We verify the effectivness of our method with different types of mask regions, including both center and random masks [59, 44].

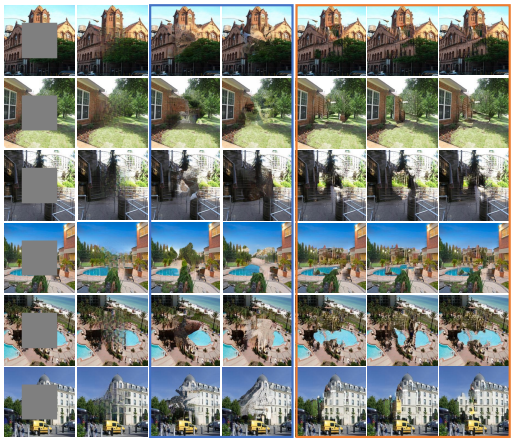
Network and Optimization. The proposed method can be implemented efficiently. The encoder and decoder of our pipeline are inspired by PIC [59] for comparison. We apply average pooling and interpolation with convolutional layers to implement downsampling and upsampling respectively. During optimization, we use the Adam optimizer [20]. The learning rate is fixed to 10^{-4} during the training procedure.

Baselines. We compare the proposed method with the following state-of-the-art methods, which include (1) single image completion methods: DFv2 [52], EC [33], and MED [26]; (2) pluralistic image completion methods: PIC [59] and ICT [44]. We abbreviate our method (**P**luralistic **I**mage **C**ompletion with Gaussian **M**ixture **M**odels) as PICMM. The methods are implemented by PyTorch and evaluated on NVIDIA Tesla A100 GPUs.



(a) Input (b) PIC (c) Ours

Figure 4: Qualitative comparison of our method with PIC on CelebA-HQ (**first 4 rows**) and FFHQ (**last 2 rows**). Best viewed by zooming in.



(a) Input (b) EC (c) ICT (d) Ours

Figure 6: Qualitative comparison of our method with EC and ICT on Places2. Best viewed by zooming in.

(a) Input (b) MED (c) PIC (d) Ours

Figure 5: Qualitative comparison of our method with MED and PIC on Paris StreetView. Best viewed by zooming in.

(a) Input (b) ICT (c) Ours

Figure 7: Qualitative comparison of our method with ICT on ImageNet. Best viewed by zooming in.

Measurement. For measurement, we provide both qualitative and quantitative results. For the metrics of quantitative results, we use common evaluation metrics such as the peak signal-to-noise

ratio (PSNR), structural similarity (SSIM), mean absolute error (MAE), and fréchet inception distance (FID) [14] to measure the similarity between the image completion result and ground truth. Furthermore, pluralistic image completion is supposed to focus on generating diverse realistic results rather than merely approximating ground-truth ones [58]. To *measure the result diversity*, we add two perceptual quality metrics to quantitative results, which include LPIPS [59] and DIV-FID [44].

4.2 Comparison with Prior Methods

4.2.1 Qualitative Comparisons.

We provide extensive qualitative comparison results to justify our claims. First, we show the results on CelebA-HQ and FFHQ in Figure 4, which are images related to human faces. As can be seen, the image completion results achieved by our method are realistic, following high image quality. Also, compared with the baseline PIC, our completion results are more diverse, *e.g.*, see the images in the second and last rows.

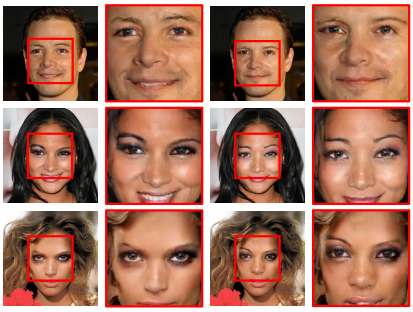


Figure 8: Refined result analysis of our method. The face images come from CelebA-HQ.

Second, we provide the results on Paris StreetView and Places in Figures 5 and 6 respectively. For the results on Paris StreetView, we can see that image completion by the baselines MED and PIC are somewhat distorted. By contrast, our method can finish the image completion task better. The details can be checked in the complements to walls. Then, we turn the attention to the image completion results on Places2. Our method still achieves superior performance compared with baselines.

Lastly, we provide the results on ImageNet in Figure 7. Note that the sources of the data in ImageNet are very complicated. The image completion task is rather challenging on this dataset, even for single image completion [59]. Different from some prior work [36, 16] that were trained on a 100k subset of training images of ImageNet, we directly the network on the original ImageNet training dataset with all images. The results mean that our method can infer the content effectively.

Refined Result Analysis. We provide some refined results in Figure 8 to analyze the image completion details of our method. For the face images, we can find that the image completion results have very different facial expressions.

Dataset		FFHQ				Places2			
Method	Mask	PSNR ↑	SSIM ↑	MAE ↓	FID ↓	PSNR ↑	SSIM ↑	MAE ↓	FID ↓
DFv2 [52]	Center	25.868	0.922	0.0231	16.278	26.533	0.881	0.0215	24.763
EC [33]		26.901	0.938	0.0209	14.276	26.520	0.880	0.0220	25.642
MED [26]		26.325	0.922	0.0230	14.791	26.469	0.877	0.0224	26.977
PIC [59]		26.781	0.933	0.0215	14.513	26.099	0.865	0.0236	26.393
ICT [44]		**27.922**	**0.948**	0.0208	10.995	26.503	0.880	0.0244	**21.598**
PICMM†		27.551	0.937	**0.0203**	**10.604**	**26.554**	**0.886**	**0.0208**	24.373
DFv2 [52]	Random	24.962	0.882	0.0310	19.506	25.692	**0.834**	0.0280	29.981
EC [33]		25.908	0.882	0.0301	17.039	25.510	0.831	0.0293	30.130
MED [26]		25.118	0.867	0.0349	19.644	25.632	0.827	0.0291	31.395
PIC [59]		25.580	0.889	0.0303	17.364	25.035	0.806	0.0315	33.472
ICT [44]		**26.681**	0.910	0.0292	**14.529**	**25.788**	0.832	0.0267	25.420
PICMM†		25.950	**0.912**	**0.0289**	17.014	25.310	0.829	**0.0236**	**25.025**

Table 1: Quantitative results on FFHQ and Places2 datasets with different mask settings. The best results are in **bold**.

4.2.2 Quantitative Comparisons

Common Metrics. Quantitative evaluation with common metrics is difficult for pluralistic image completion, as our goal is to obtain diverse but reasonable solutions for a masked image [59]. To make a feasible comparison, as did in [59], we sample one completed image many times from k images to compare with the original image. The results are provided in Tables 1 and 2. As can be seen, for FFHQ and Places2, on both the settings of center and random masking, our method achieves superior performance. On ImageNet, our method performs the best with the center masking consistently, while it is competitive with the random masking.

Dataset		ImageNet			
Method	Mask	PSNR ↑	SSIM ↑	MAE ↓	FID ↓
PIC [59]	Center	24.010	0.867	0.0319	47.750
ICT [44]		24.757	0.888	0.0263	28.818
PICMM†		**24.932**	**0.897**	**0.0260**	**23.718**
PIC [59]	Random	22.711	0.791	0.0462	59.428
ICT [44]		**23.775**	0.835	**0.0358**	**35.842**
PICMM†		23.322	**0.846**	0.0415	39.742

Table 2: Quantitative results on ImageNet with different mask settings. The best results are in **bold**.

It should be noted that ICT is a very strong baseline and works better than our method in some cases. It is because ICT is built with the image transformer [35], which is powerful in vision tasks. Compared with ICT, the advantages of our method lies in not only better interpretability as discussed, but also faster inference speeds. We will demonstrate this in inference time.

Diversity Metrics. For pluralistic image completion, the diversity evaluations are significant. We generate five completion images in each task and report the mean of the diversity metric. The results are shown in Table 3, which represents that our method can generate more diverse completion images than baselines. The results with common and diversity metrics support our claim very well. That is to say, the image completion results of our method are both high-quality and pluralistic.

Dataset	FFHQ		Places2	
Method	LPIPS ↑	DIV-FID ↑	LPIPS ↑	DIV-FID ↑
PIC [59]	0.029	9.130	0.047	17.742
ICT [44]	0.065	13.909	0.089	25.253
PICMM†	**0.071**	**17.361**	**0.092**	**27.194**

Table 3: The diversity result comparison on FFHQ and Places2 datasets with the center mask setting. The best results are in **bold**.

Inference Time. We report the inference time of our method, compared with PIC and ICT. All inference runs on one NVIDIA Tesla A100 GPU for fairness. The size of images is 256×256. We perform pluralistic image completion for 100 different images. For each image, we generate 6 completion results for it. We report the total inference time for these images. The results are shown in Table 4. Note that as ICT is a transformer-based method, the calculated consumption of its inference is much heavier. What's worse, ICT needs *iterative Gibbs sampling* during inference. The two issues make ICT have to face heavy computational consumption, which is also mentioned in [44]. In contrast, our method is more inference-efficiency, which is **more than 750 times faster than ICT**. The merit could make our method easier to use in the real world.

Method	Inference time (minutes) ↓
PIC [59]	0.387
ICT [44]	179.126
PICMM†	**0.235**

Table 4: The inference time comparison with PIC and ICT in the center mask setting. The best result is in **bold**.

4.3 More Analyses and Justifications

GMM Primitive Visualization. We further stress the diversity of the completion results achieved by our method. Specifically, we visualize the $\hat{\mathbf{z}}_c$ which is inferred from $\mathbf{z}_m$ as discussed. The visualization results are obtained with t-SNE [42]. Different $\hat{\mathbf{z}}_c$ are presented by the 2D-vectors in different colors, which are shown in Figure 9. As can be seen, the vectors are *scattered*, which clearly demonstrates that the completion results of our method are diverse.

Refined Diversity Comparison. We argue that, benefiting from that GMM has larger capacities than a unimodal Gaussian distribution, the completion results of our method would be more diverse. To justify our claims, we provide the refined diversity comparison of our method with PIC and ICT, as shown in Figure 10. Note that the six image completion results of PIC are obtained by sampling from

Figure 9: GMM primitive visualization of pluralistic completion results by our method. The face image comes from CelebA-HQ. The scene image comes from Places2.

Figure 10: Refined diversity comparison with PIC (**the first row**) and ICT (**the second row**). The original image in this figure comes from CelebA-HQ.

one unimodal Gaussian distribution. The results of our method are obtained by sampling from six primitives. Clearly, our results are more diverse than PIC's. Besides, for ICT, it does not explicitly control the output diversity. The diversity of pluralistic image completion is thus *less interpretable* than our method.

In addition, we state that our method is not limited to only generating k completion results before. In fact, we can sample from k primitives of GMM to obtain lots of completion results. Due to the limited page, we present the results and more analyses in Appendix C. Moreover, the richer completion result comparison is presented in Appendix D.

5 Conclusion

In this paper, we focus on the complicated and challenging problem of pluralistic image completion. We propose a novel end-to-end probabilistic method for this challenging problem. Based on the probabilistic graph model, our method divides the entire procedure of pluralistic image completion into several sub-procedures, where GMM is used to diversify outputs. Experiments on a variety of datasets show that the image completion results by our method are both high-quality and pluralistic. In future work, investigating the feasibility of the proposed method for other diverse decision-making scenarios might prove important.

Acknowledgements

Yibing Zhan was partially supported by the Major Science and Technology Innovation 2030 "New Generation Artificial Intelligence" key project (No. 2021ZD0111700) and the National Natural Science Foundation of China (Grant No. 62002090). Xiaobo Xia was supported by Australian Research Council Projects DE-190101473 and Google PhD Fellowship. Bo Han was supported by the RGC Early Career Scheme No. 22200720, NSFC Young Scientists Fund No. 62006202, and Guangdong Basic and Applied Basic Research Foundation No. 2022A1515011652. Tongliang Liu was partially supported by Australian Research Council Projects DP180103424, DE-190101473, IC-190100031, DP-220102121, and FT-220100318. The authors would give special thanks to Mingrui Zhu (Xidian University), Zihan Ding (Princeton University), and Chenlai Qian (Southeast University) for helpful discussions.

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

[3] 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, 2000.

[4] 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.

[5] 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.

[6] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *CVPR*, pages 7093–7102, 2018.

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

[8] Brian Dolhansky and Cristian Canton Ferrer. Eye in-painting with exemplar generative adversarial networks. In *CVPR*, pages 7902–7911, 2018.

[9] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Xiaohui Shen, Zhenyu Xie, Bowen Wu, and Jian Yin. Fashion editing with adversarial parsing learning. In *CVPR*, pages 8120–8128, 2020.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[11] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

[12] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *CVPR*, 2019.

[13] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R Scott, and Larry S Davis. Finet: Compatible and diverse fashion image inpainting. In *ICCV*, pages 4481–4491, 2019.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

[15] Zhuo Huang, Xiaobo Xia, Li Shen, Bo Han, Mingming Gong, Chen Gong, and Tongliang Liu. Harnessing out-of-distribution examples via augmenting content and style. *arXiv preprint arXiv:2207.03162*, 2022.

[16] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4):1–14, 2017.

[17] Michael I Jordan. An introduction to probabilistic graphical models, 2003.

[18] 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.

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019.

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

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

[22] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.

[23] Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics. In *ICCV*, pages 305–312, 2003.

[24] Yewen Li, Chaojie Wang, Xiaobo Xia, Tongliang Liu, Xin Miao, and Bo An. Out-of-distribution detection with an adaptive likelihood ratio on informative hierarchical vae. In *NeurIPS*, 2022.

[25] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *CVPR*, pages 3911–3919, 2017.

[26] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *ECCV*, pages 725–741, 2020.

[27] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting. In *CVPR*, pages 9371–9381, 2021.

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.

[29] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *CVPR*, pages 2794–2802, 2017.

[30] Geoffrey J McLachlan and Suren Rathnayake. On the number of components in a gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(5): 341–355, 2014.

[31] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT press, 2018.

[32] Daniel Mas Montserrat, Hanxiang Hao, Sri K Yarlagadda, Sriram Baireddy, Ruiting Shao, János Horváth, Emily Bartusiak, Justin Yang, David Guera, Fengqing Zhu, et al. Deepfakes detection with automatic face weighting. In *CVPR Workshops*, pages 668–669, 2020.

[33] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.

[34] Thanh Thi Nguyen, Cuong M Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. Deep learning for deepfakes creation and detection. *arXiv preprint arXiv:1909.11573*, 2019.

[35] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, pages 4055–4064, 2018.

[36] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016.

[37] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *CVPR*, pages 10775–10784, 2021.

[38] Jie Ren, Yewen Li, Zihan Ding, Wei Pan, and Hao Dong. Probabilistic mixture-of-experts for efficient deep reinforcement learning. *arXiv preprint arXiv:2104.09122*, 2021.

[39] Jimmy S Ren, Li Xu, Qiong Yan, and Wenxiu Sun. Shepard convolutional neural networks. In *NeurIPS*, 2015.

[40] Douglas A Reynolds. Gaussian mixture models. *Encyclopedia of Biometrics*, 741:659–663, 2009.

[41] 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.

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

[43] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *ECCV*, pages 835–851, 2016.

[44] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *ICCV*, 2021.

[45] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In *NeurIPS*, 2020.

[46] Xiaobo Xia, Bo Han, Nannan Wang, Jiankang Deng, Jiatong Li, Yinian Mao, and Tongliang Liu. Extended T: Learning with mixed closed-set and open-set noisy labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[47] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *ECCV*, pages 1–17, 2018.

[48] Shuo Yang, Wei Yu, Ying Zheng, Hongxun Yao, and Tao Mei. Adaptive semantic-visual tree for hierarchical embeddings. In *ACM MM*, pages 2097–2105, 2019.

[49] Shuo Yang, Peize Sun, Yi Jiang, Xiaobo Xia, Ruiheng Zhang, Zehuan Yuan, Changhu Wang, Ping Luo, and Min Xu. Objects in semantic topology. In *ICLR*, 2022.

[50] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *CVPR*, pages 7508–7517, 2020.

[51] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, pages 5505–5514, 2018.

[52] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, pages 4471–4480, 2019.

[53] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *CVPR*, pages 1486–1494, 2019.

[54] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[55] Yu Zeng, Zhe Lin, Huchuan Lu, and Vishal M Patel. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In *ICCV*, pages 14164–14173, 2021.

[56] Haizheng Zhang, C Lee Giles, Henry C Foley, and John Yen. Probabilistic community discovery using hierarchical latent gaussian mixture model. In *AAAI*, pages 663–668, 2007.

[57] Haoran Zhang, Zhenzhen Hu, Changzhi Luo, Wangmeng Zuo, and Meng Wang. Semantic image inpainting with progressive generative networks. In *ACM MM*, pages 1939–1947, 2018.

[58] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *CVPR*, pages 5741–5750, 2020.

[59] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *CVPR*, pages 1438–1447, 2019.

[60] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic free-form image completion. *International Journal of Computer Vision*, 129(10):2786–2805, 2021.

[61] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Tfill: Image completion via a transformer-based architecture. *arXiv preprint arXiv:2104.00845*, 2021.

[62] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.

[63] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Multimodal image-to-image translation by enforcing bi-cycle consistency. In *NeurIPS*, pages 465–476, 2017.

[64] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] ,
   (b) Did you describe the limitations of your work? [Yes] ,
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] ,
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplementary Material of "Pluralistic Image Completion with Gaussian Mixture Models"

The appendix is organized as follows. We detail theoretical analyses in Section A. The related work is reviewed in Section B. Supplementary experimental results are provided in Section C and Section D.

A The Details of Theoretical Analyses

A.1 The Proof of Proposition 1

As we do not have the access to the underlying $\mathbf{I}_c$, we can preform the following variational approximation with the Kullback-Leibler (KL) divergence:

$$
\begin{aligned}
\min \quad & \mathrm{KL}\left[q_\psi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m) \| p_\phi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m, \mathbf{I}_c)\right] \\
&= \int_{\mathbf{I}_o}\int_{\mathbf{z}_m}\int_{\mathbf{z}_c} q_\psi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m) \log \frac{q_\psi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m)}{p_\phi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m, \mathbf{I}_c)} \mathrm{d}\mathbf{I}_o \mathrm{d}\mathbf{z}_m \mathrm{d}\mathbf{z}_c \\
&= \int_{\mathbf{I}_o}\int_{\mathbf{z}_m}\int_{\mathbf{z}_c} q_\psi(\mathbf{I}_o|\mathbf{z}_m, \mathbf{z}_c) q_\psi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m) \log \frac{q_\psi(\mathbf{I}_o|\mathbf{z}_m, \mathbf{z}_c) q_\psi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m)}{p_\phi(\mathbf{I}_o|\mathbf{I}_m, \mathbf{I}_c) p_\phi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m, \mathbf{I}_c)} \mathrm{d}\mathbf{I}_o \mathrm{d}\mathbf{z}_m \mathrm{d}\mathbf{z}_c \\
&= \int_{\mathbf{I}_o}\int_{\mathbf{z}_m}\int_{\mathbf{z}_c} q_\psi(\mathbf{I}_o|\mathbf{z}_m, \mathbf{z}_c) q_\psi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m) \log \frac{q_\psi(\mathbf{I}_o|\mathbf{z}_m, \mathbf{z}_c)}{p_\phi(\mathbf{I}_o|\mathbf{I}_m, \mathbf{I}_c)} \mathrm{d}\mathbf{I}_o \mathrm{d}\mathbf{z}_m \mathrm{d}\mathbf{z}_c \\
&+ \int_{\mathbf{I}_o}\int_{\mathbf{z}_m}\int_{\mathbf{z}_c} q_\psi(\mathbf{I}_o|\mathbf{z}_m, \mathbf{z}_c) q_\psi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m) \log \frac{q_\psi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m)}{p_\phi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m, \mathbf{I}_c)} \mathrm{d}\mathbf{I}_o \mathrm{d}\mathbf{z}_m \mathrm{d}\mathbf{z}_c \\
&= \int_{\mathbf{z}_m}\int_{\mathbf{z}_c} q_\psi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m) \mathrm{KL}\left[q_\psi(\mathbf{I}_o|\mathbf{z}_m, \mathbf{z}_c) \| p_\phi(\mathbf{I}_o|\mathbf{I}_m, \mathbf{I}_c)\right] \mathrm{d}\mathbf{z}_m \mathrm{d}\mathbf{z}_c \\
&+ \int_{\mathbf{I}_o} q_\psi(\mathbf{I}_o|\mathbf{z}_m, \mathbf{z}_c) \mathrm{KL}\left[q_\psi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m) \| p_\phi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m, \mathbf{I}_c)\right] \mathrm{d}\mathbf{I}_o \\
&= \mathbb{E}_{(\mathbf{z}_m, \mathbf{z}_c) \sim q_\psi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m)} \mathrm{KL}\left[q_\psi(\mathbf{I}_o|\mathbf{z}_m, \mathbf{z}_c) \| p_\phi(\mathbf{I}_o|\mathbf{I}_m, \mathbf{I}_c)\right] + \mathrm{KL}\left[q_\psi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m) \| p_\phi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m, \mathbf{I}_c)\right].
\end{aligned} \tag{9}
$$

The second term in Eq. (9) is

$$
\begin{aligned}
& \mathrm{KL}\left[q_\psi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m) \| p_\phi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m, \mathbf{I}_c)\right] \\
&= \int_{\mathbf{z}_m}\int_{\mathbf{z}_c} q_\theta(\mathbf{z}_c|\mathbf{z}_m) q_\psi(\mathbf{z}_m|\mathbf{I}_m) \log \frac{q_\theta(\mathbf{z}_c|\mathbf{z}_m) q_\psi(\mathbf{z}_m|\mathbf{I}_m)}{p_\phi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m, \mathbf{I}_c)} \mathrm{d}\mathbf{z}_m \mathrm{d}\mathbf{z}_c \\
&= \int_{\mathbf{z}_m}\int_{\mathbf{z}_c} q_\theta(\mathbf{z}_c|\mathbf{z}_m) q_\psi(\mathbf{z}_m|\mathbf{I}_m) \log \frac{q_\theta(\mathbf{z}_c|\mathbf{z}_m)}{p_\phi(\mathbf{z}_c|\mathbf{I}_c)} \mathrm{d}\mathbf{z}_m \mathrm{d}\mathbf{z}_c + q_\theta(\mathbf{z}_c|\mathbf{z}_m) q_\psi(\mathbf{z}_m|\mathbf{I}_m) \log \frac{q_\psi(\mathbf{z}_m|\mathbf{I}_m)}{p_\phi(\mathbf{z}_m|\mathbf{I}_m)} \mathrm{d}\mathbf{z}_m \mathrm{d}\mathbf{z}_c \\
&= \mathbb{E}_{\mathbf{z}_m \sim q_\psi(\mathbf{z}_m|\mathbf{I}_m)} \mathrm{KL}\left[q_\theta(\mathbf{z}_c|\mathbf{z}_m) \| p_\phi(\mathbf{z}_c|\mathbf{I}_c)\right] + \mathrm{KL}\left[q_\psi(\mathbf{z}_m|\mathbf{I}_m) \| p_\phi(\mathbf{z}_m|\mathbf{I}_m)\right].
\end{aligned} \tag{10}
$$

Combining Eq. (9) and Eq. (10), we complete the proof. ■

Note that the latent variables $\mathbf{z}_m$ and $\mathbf{z}_c$ are *not codependent*. If $\mathbf{z}_m$ and $\mathbf{z}_c$ are codependent, the inference of $\mathbf{z}_m$ (*resp.* $\mathbf{z}_c$) *must need* $\mathbf{z}_c$ (*resp.* $\mathbf{z}_m$). Then, in Eq. (10), there should be $q_\psi(\mathbf{z}_m|\mathbf{I}_m, \mathbf{z}_c)$ and $p_\phi(\mathbf{z}_c|\mathbf{I}_c, \mathbf{z}_m)$, rather than $q_\psi(\mathbf{z}_m|\mathbf{I}_m)$ and $p_\phi(\mathbf{z}_c|\mathbf{I}_c)$. As shown in the our Figure 3, $\mathbf{z}_m$ and $\mathbf{z}_c$ can be inferred without each other, but using $\mathbf{I}_m$ and $\mathbf{I}_c$. Therefore, $q_\psi(\mathbf{z}_m|\mathbf{I}_m)$ and $p_\phi(\mathbf{z}_c|\mathbf{I}_c)$ hold in Eq. (10).

A.2 The Derivation of the Frequency Loss

As discussed, we exploit the frequency loss [38] for the Gaussian mixture model. Here, we detail The derivation of the frequency loss for our task, *i.e.*, pluralistic image completion. Specifically, we use the KL divergence between each primitive $\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_i^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_i^{\hat{\mathbf{z}}_c})$ and $\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}^{\mathbf{z}_c}, \boldsymbol{\Sigma}^{\mathbf{z}_c})$ as the metric of the performance:

$$
\mathrm{KL}\left[\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_i^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_i^{\hat{\mathbf{z}}_c}) \| \mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}^{\mathbf{z}_c}, \boldsymbol{\Sigma}^{\mathbf{z}_c})\right], i = 1, ..., k. \tag{11}
$$

Then, by comparing the KL divergence of each primitive, we can get the best primitive's index j. That is

$$j = \underset{i}{\operatorname{argmin}} \operatorname{KL}\left[\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_i^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_i^{\hat{\mathbf{z}}_c}) \| \mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}^{\mathbf{z}_c}, \boldsymbol{\Sigma}^{\mathbf{z}_c})\right]. \tag{12}$$

We now can give a detailed derivation for Eq. (4) below. Specifically, for stochastic Gaussian mixture models, the gradient value of a single-instance sampling process from each primitive's performance under the metric Eq. (11), to parameters $\theta_{\boldsymbol{\alpha}}$ that output the mixing coefficients $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k]$, can be estimated with a *frequency approximate gradient*. Mathematically, we have

$$\text{grad}_i = \delta_i \nabla_{\theta_i} \alpha_i \text{ and } \delta_i = -\mathbb{1}_i^{\text{best}} + \alpha_i, \tag{13}$$

where δ_i is the gradient of the performance metric Eq. (11) for α_i, $\nabla_{\theta_i}\alpha_i$ is the gradient of α_i for parameters θ_i, and $\mathbb{1}_i^{\text{best}}$ is the indicator function, where $\mathbb{1}_i^{\text{best}} = 1$ if $i = j$ and $\mathbb{1}_i^{\text{best}} = 0$ otherwise.

The accumulated frequency approximate gradient is an *asymptotically unbiased estimation* of the true gradient for the sampling process from a categorical distribution $\boldsymbol{\alpha}$, with a batch of $n \to \infty$ examples. For δ_i, we can detail it below:

$$\begin{aligned}
\delta_i &= -\mathbb{1}_i^{\text{best}} + \alpha_i \\
&= \mathbb{1}_i^{\text{best}}\alpha_i - \mathbb{1}_i^{\text{best}}\alpha_i - \mathbb{1}_i^{\text{best}} + \alpha_i \\
&= \frac{1}{2}\mathbb{1}_i^{\text{best}}\nabla_{\alpha_i}(1-\alpha_i)^2 + \frac{1}{2}(1-\mathbb{1}_i^{\text{best}})\nabla_{\alpha_i}\alpha_i^2.
\end{aligned} \tag{14}$$

Suppose a batch of examples with a number of n is applied, and the true probability of the primitive i to be the best primitive is p_i. The batch accumulated gradient will be

$$\overline{\text{grad}} = \frac{1}{n}\sum_{l=1}^{n}\text{grad}_i^l = \frac{n_t}{2n}\nabla_{\alpha_i}(1-\alpha_i)^2\nabla_{\theta_i}\alpha_i + \frac{(n-n_t)}{2n}\nabla_{\alpha_i}\alpha_i^2\nabla_{\theta_i}\alpha_i \overset{n\to\infty}{=} (\alpha_i - p_i)\nabla_{\theta_i}\alpha_i, \tag{15}$$

where the last formula indicates that $n_t = np_i$ when $n \to \infty$, since the true probability can be approximated by $\frac{n_t}{n}$ in the limit case. Also, as $\nabla_{\theta_i}\alpha_i$ is not always equal to 0, the gradient equals to 0 if and only if $\alpha_i = p_i$. Optimising with the above Eq. (15) is the same as minimising the distance between α_i and p_i, with the optimal situation as $\alpha_i = p_i$ when letting the last formula of Eq. (15) be zero. Based the above analyses, we therefore can build a frequency loss in this paper.

The derivation is completed. ■

A.3 The Proof of Proposition 2

We employ the *mixture of Gaussian* for the *dynamic* model $q_\theta(\mathbf{z}_c|\mathbf{z}_m)$, which is more flexible than a unimodal and can diversify the output. More formally, we have

$$q_\theta(\mathbf{z}_c|\mathbf{z}_m) = \sum_{i=1}^{k}\alpha_i\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_i^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_i^{\hat{\mathbf{z}}_c}), \tag{16}$$

where k denotes the number of primitives, and α_i denotes the weight of the i-th primitive. Then the KL divergence between $q_\theta(\mathbf{z}_c|\mathbf{z}_m)$ and $p_\phi(\mathbf{z}_c|\mathbf{I}_c)$ becomes

$$\begin{aligned}
&\operatorname{KL}\left[q_\theta(\mathbf{z}_c|\mathbf{z}_m)\|p_\phi(\mathbf{z}_c|\mathbf{I}_c)\right] \\
&= \int_{\mathbf{z}_c} q_\theta(\mathbf{z}_c|\mathbf{z}_m)\log\frac{q_\theta(\mathbf{z}_c|\mathbf{z}_m)}{p_\phi(\mathbf{z}_c|\mathbf{I}_c)}\mathrm{d}\mathbf{z}_c \\
&= \int_{\mathbf{z}_c}\sum_{i=1}^{k}\alpha_i\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_i^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})\log\frac{\sum_{i=1}^{k}\alpha_i\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_i^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_i^{\hat{\mathbf{z}}_c})}{\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}^{\mathbf{z}_c}, \boldsymbol{\Sigma}^{\mathbf{z}_c})}\mathrm{d}\mathbf{z}_c.
\end{aligned} \tag{17}$$

We further approximate Eq. (17) by the back-propogate-max-operation with the frequency loss. We will have

$$\operatorname{KL}\left[q_\theta(\mathbf{z}_c|\mathbf{z}_m)\|p_\phi(\mathbf{z}_c|\mathbf{I}_c)\right] \approx \int_{\mathbf{z}_c}\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})\log\frac{\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})}{\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}^{\mathbf{z}_c}, \boldsymbol{\Sigma}^{\mathbf{z}_c})}\mathrm{d}\mathbf{z}_c, \tag{18}$$

where $j = \text{argmin}_i \text{KL}\left[\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_i^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_i^{\hat{\mathbf{z}}_c}) \| \mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}^{\mathbf{z}_c}, \boldsymbol{\Sigma}^{\mathbf{z}_c})\right]$. We then have

$$
\begin{aligned}
&\int_{\mathbf{z}_c} \mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}) \log \frac{\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})}{\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}^{\mathbf{z}_c}, \boldsymbol{\Sigma}^{\mathbf{z}_c})} \mathrm{d}\mathbf{z}_c \\
&= \int_{\mathbf{z}_c} \mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}) \log \frac{(2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{z}_c - \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c})^\top (\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})^{-1} (\mathbf{z}_c - \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c})\right]}{(2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}^{\mathbf{z}_c}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{z}_c - \boldsymbol{\mu}^{\mathbf{z}_c})^\top (\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1} (\mathbf{z}_c - \boldsymbol{\mu}^{\mathbf{z}_c})\right]} \mathrm{d}\mathbf{z}_c \\
&= \int_{\mathbf{z}_c} \mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}) \log \frac{|\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{z}_c - \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c})^\top (\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})^{-1} (\mathbf{z}_c - \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c})\right]}{|\boldsymbol{\Sigma}^{\mathbf{z}_c}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{z}_c - \boldsymbol{\mu}^{\mathbf{z}_c})^\top (\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1} (\mathbf{z}_c - \boldsymbol{\mu}^{\mathbf{z}_c})\right]} \mathrm{d}\mathbf{z}_c \\
&= \int_{\mathbf{z}_c} \mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}) \{-\frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}|}{|\boldsymbol{\Sigma}^{\mathbf{z}_c}|} + \frac{1}{2}[(\mathbf{z}_c - \boldsymbol{\mu}^{\mathbf{z}_c})^\top (\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1} (\mathbf{z}_c - \boldsymbol{\mu}^{\mathbf{z}_c}) - (\mathbf{z}_c - \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c})^\top (\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})^{-1} (\mathbf{z}_c - \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c})]\} \mathrm{d}\mathbf{z}_c \\
&= -\frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}|}{|\boldsymbol{\Sigma}^{\mathbf{z}_c}|} + \frac{1}{2} \int_{\mathbf{z}_c} \mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}) \Big[{\mathbf{z}_c}^\top (\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1} \mathbf{z}_c - {\mathbf{z}_c}^\top (\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1} \boldsymbol{\mu}^{\mathbf{z}_c} - {\boldsymbol{\mu}^{\mathbf{z}_c}}^\top (\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1} \mathbf{z}_c + {\boldsymbol{\mu}^{\mathbf{z}_c}}^\top (\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1} \boldsymbol{\mu}^{\mathbf{z}_c} \\
&- ({\mathbf{z}_c}^\top (\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})^{-1} \mathbf{z}_c - {\mathbf{z}_c}^\top (\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})^{-1} \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c} - {\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}}^\top (\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})^{-1} \mathbf{z}_c + {\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}}^\top (\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})^{-1} \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c})\Big] \mathrm{d}\mathbf{z}_c.
\end{aligned}
\tag{19}
$$

Suppose the covariance is a diagonal matrix [21], we have

$$
\int_{\mathbf{z}_c} \mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}) {\mathbf{z}_c}^\top \mathbf{z}_c \mathrm{d}\mathbf{z}_c = \text{tr}(\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}) + {\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}}^\top \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c},
\tag{20}
$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix. Then, Eq. (19) would be:

$$
\begin{aligned}
&\int_{\mathbf{z}_c} \mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}) \log \frac{\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}, \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})}{\mathcal{N}(\mathbf{z}_c|\boldsymbol{\mu}^{\mathbf{z}_c}, \boldsymbol{\Sigma}^{\mathbf{z}_c})} \mathrm{d}\mathbf{z}_c \\
&= -\frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}|}{|\boldsymbol{\Sigma}^{\mathbf{z}_c}|} - \frac{1}{2}\Big[\text{tr}\left((\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1} \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}\right) + {\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}}^\top (\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1} \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c} - {\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}}^\top (\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1} \boldsymbol{\mu}^{\mathbf{z}_c} - {\boldsymbol{\mu}^{\mathbf{z}_c}}^\top (\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1} \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c} \\
&+ {\boldsymbol{\mu}^{\mathbf{z}_c}}^\top (\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1} \boldsymbol{\mu}^{\mathbf{z}_c} + \text{tr}\left((\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})^{-1} \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}\right) + {\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}}^\top (\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})^{-1} \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c} - {\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}}^\top (\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})^{-1} \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c} - {\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}}^\top (\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})^{-1} \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c} \\
&+ {\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}}^\top (\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c})^{-1} \boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c}\Big] \\
&= -\frac{1}{2} \log \frac{|\boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}|}{|\boldsymbol{\Sigma}^{\mathbf{z}_c}|} + \frac{1}{2}\text{tr}\left((\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1} \boldsymbol{\Sigma}_j^{\hat{\mathbf{z}}_c}\right) - \frac{1}{2}(\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c} - \boldsymbol{\mu}^{\mathbf{z}_c})^\top (\boldsymbol{\Sigma}^{\mathbf{z}_c})^{-1} (\boldsymbol{\mu}_j^{\hat{\mathbf{z}}_c} - \boldsymbol{\mu}^{\mathbf{z}_c}).
\end{aligned}
\tag{21}
$$

The proof is completed. ■

A.4 The Fully Optimization Expression of Proposition 1

As discussed in the main paper, we have

$$
\begin{aligned}
L_{\text{P}} &= \text{KL}\left[q_\psi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m) \| p_\phi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m, \mathbf{I}_c)\right] \\
&= \underbrace{\mathbb{E}_{(\mathbf{z}_m, \mathbf{z}_c) \sim q_\psi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m)} \text{KL}\left[q_\psi(\mathbf{I}_o|\mathbf{z}_m, \mathbf{z}_c) \| p_\phi(\mathbf{I}_o|\mathbf{I}_m, \mathbf{I}_c)\right]}_{ⓐ} \\
&+ \underbrace{\mathbb{E}_{\mathbf{z}_m \sim q_\psi(\mathbf{z}_m|\mathbf{I}_m)} \text{KL}\left[q_\theta(\mathbf{z}_c|\mathbf{z}_m) \| p_\phi(\mathbf{z}_c|\mathbf{I}_c)\right]}_{ⓑ} \\
&+ \underbrace{\text{KL}\left[q_\psi(\mathbf{z}_m|\mathbf{I}_m) \| p_\phi(\mathbf{z}_m|\mathbf{I}_m)\right]}_{ⓒ}.
\end{aligned}
\tag{22}
$$

As Eq. (22) shows, the item ⓒ may be hard to implement, because we cannot know the real $p_\phi(\mathbf{z}_m|\mathbf{I}_m)$, which is used to restrain $q_\psi(\mathbf{z}_m|\mathbf{I}_m)$ outputted by a neural network.

However, inspired by [21], we can employ a variational evidence lower bound (ELBO) to implement the item ⓒ. Firstly, we assume that there exists a ground truth model $p(\mathbf{I}_m)$, which is a constant given $\mathbf{I}_m$ and does not depend on $\mathbf{z}_m$. Then, we can derive the relationship between $p(\mathbf{I}_m)$ and ⓒ:

$$
\begin{aligned}
\log p(\mathbf{I}_m) &= \mathbb{E}_{\mathbf{z}_m \sim q_\psi(\mathbf{z}_m|\mathbf{I}_m)}[\log p(\mathbf{I}_m)] \\
&= \mathbb{E}_{\mathbf{z}_m \sim q_\psi(\mathbf{z}_m|\mathbf{I}_m)}[\log \frac{p(\mathbf{I}_m|\mathbf{z}_m)p(\mathbf{z}_m)}{p_\phi(\mathbf{z}_m|\mathbf{I}_m)}] \\
&= \mathbb{E}_{\mathbf{z}_m \sim q_\psi(\mathbf{z}_m|\mathbf{I}_m)}[\log \frac{p(\mathbf{I}_m|\mathbf{z}_m)p(\mathbf{z}_m)}{p_\phi(\mathbf{z}_m|\mathbf{I}_m)} \frac{q_\psi(\mathbf{z}_m|\mathbf{I}_m)}{q_\psi(\mathbf{z}_m|\mathbf{I}_m)}] \\
&= \mathbb{E}_{\mathbf{z}_m}[\log p(\mathbf{I}_m|\mathbf{z}_m)] - \mathbb{E}_{\mathbf{z}_m}[\log \frac{q_\psi(\mathbf{z}_m|\mathbf{I}_m)}{p(\mathbf{z}_m)}] + \mathbb{E}_{\mathbf{z}_m}[\log \frac{q_\psi(\mathbf{z}_m|\mathbf{I}_m)}{p_\phi(\mathbf{z}_m|\mathbf{I}_m)}] \\
&= \mathbb{E}_{\mathbf{z}_m}[\log p(\mathbf{I}_m|\mathbf{z}_m)] - \mathrm{KL}[q_\psi(\mathbf{z}_m|\mathbf{I}_m)||p(\mathbf{z}_m)] + \mathrm{KL}[q_\psi(\mathbf{z}_m|\mathbf{I}_m)||p_\phi(\mathbf{z}_m|\mathbf{I}_m)] \\
&= \underbrace{\mathbb{E}_{\mathbf{z}_m}[\log p(\mathbf{I}_m|\mathbf{z}_m)] - \mathrm{KL}[q_\psi(\mathbf{z}_m|\mathbf{I}_m)||p(\mathbf{z}_m)]}_{\text{ELBO}} + ⓒ
\end{aligned}
\tag{23}
$$

Thus, ⓒ is equal to $\log p(\mathbf{I}_m) - \text{ELBO}$. Since $\log p(\mathbf{I}_m)$ is a constant, minimizing ⓒ is equal to maximize the ELBO.

Accordingly, the fully optimization expression of Theorem 1 is below:

$$
\begin{aligned}
L_{\mathrm{P}} &= \mathrm{KL}\left[q_\psi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m) \| p_\phi(\mathbf{I}_o, \mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m, \mathbf{I}_c)\right] \\
&= \underbrace{\mathbb{E}_{(\mathbf{z}_m, \mathbf{z}_c) \sim q_\psi(\mathbf{z}_m, \mathbf{z}_c|\mathbf{I}_m)} \mathrm{KL}\left[q_\psi(\mathbf{I}_o|\mathbf{z}_m, \mathbf{z}_c) \| p_\phi(\mathbf{I}_o|\mathbf{I}_m, \mathbf{I}_c)\right]}_{ⓐ} \\
&+ \underbrace{\mathbb{E}_{\mathbf{z}_m \sim q_\psi(\mathbf{z}_m|\mathbf{I}_m)} \mathrm{KL}\left[q_\theta(\mathbf{z}_c|\mathbf{z}_m) \| p_\phi(\mathbf{z}_c|\mathbf{I}_c)\right]}_{ⓑ} \\
&- \underbrace{\mathbb{E}_{\mathbf{z}_m \sim q_\psi(\mathbf{z}_m|\mathbf{I}_m)}[\log p(\mathbf{I}_m|\mathbf{z}_m)] + \mathrm{KL}[q_\psi(\mathbf{z}_m|\mathbf{I}_m)||p(\mathbf{z}_m)]}_{ⓒ},
\end{aligned}
\tag{24}
$$

■

B Advance of PICMM Over PIC

The illustration of our method is shown in Figure 11.

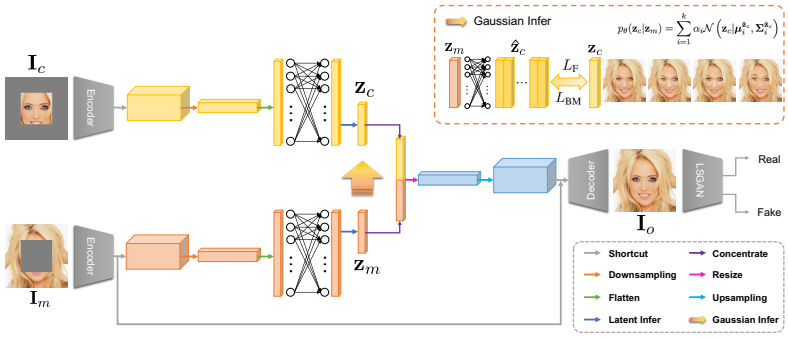


Figure 11: The algorithm framework of the proposed method.

We also summarize the differences between PIC and our method as follows:

(1) Our method has *better interpretability* in pluralistic image completion than PIC, which is highlighted in this paper.

(2) GMM is used in our method, which better promotes diversity than PIC. More importantly, the inherent parameters for diversity are *task-related*, rather than *task-agnostic* in PIC.

(3) Our method uses KL divergence for the whole objective. We decompose the divergence and then minimize different decomposition terms, which supports our theory. While, PIC only uses KL divergence as a single loss in its objective.

(4) Our method exploits the same CNN backbone as PIC. In all quantitative comparisons, our method outperforms PIC.

C Supplementary Experimental Results on Diversity Analyses

In the main paper, we state that our method is not limited to only generate k image completion results. Instead, we can sample from the k primitives of GMM to generate a different number of images. The results are provided in Figure 12 and 13. The images in the i-th row of Figure 12 and 13 are obtained by sampling from the i-th primitive of GMM. As can be seen, the images in the same rows are less diverse than the images in the different rows, since one primitive has a limited capacity. Multiple primitives can better meet the diversity needs. We are able to sample from different primitives to achieve the better diversity in pluralistic image completion.

Figure 12: Pluralistic image completion results of our method by sampling from k primitives ($k = 6$). With each primitive, six diverse images are generated.

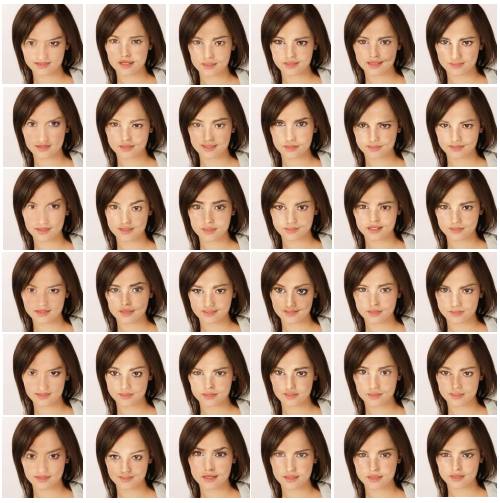


Figure 13: Pluralistic image completion results of our method by sampling from k primitives ($k = 6$). With each primitive, six diverse images are generated.

D Supplementary Pluralistic Completion Result Comparisons

In the main paper, we provide some comparisons of pluralistic completion results with state-of-the-art methods. We provide more comparisons here. The comparison methods include DFv2[3], EC[4], MED[5], PIC[6], and ICT[7]. The experimental results on CelebA-HQ are provided in Figures 14 and 15. The experimental results on FFHQ, Paris StreetView, and Places2 are shown in Figures 16, 17, and 18 respectively. In addition, the experimental results on ImageNet are presented in Figure 19.

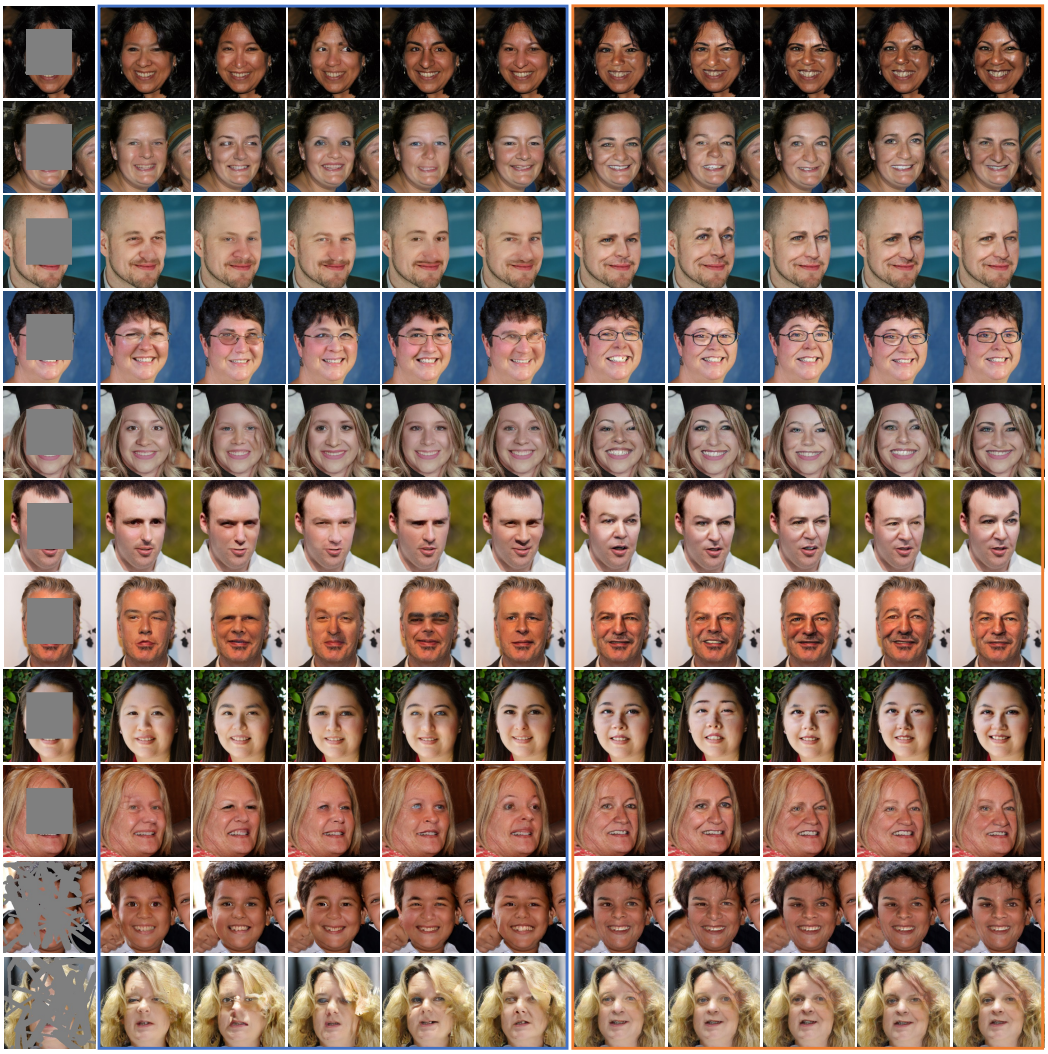
(a) Input (b) DFv2 (c) EC (d) MED (e) PIC (f) Ours

Figure 14: Pluralistic image completion results comparison with baselines. The original images come from CelebA-HQ. Best viewed by zooming in.

[3]https://github.com/JiahuiYu/generative_inpainting
[4]https://github.com/knazeri/edge-connect
[5]https://github.com/KumapowerLIU/Rethinking-Inpainting-MEDFE
[6]https://github.com/lyndonzheng/Pluralistic-Inpainting
[7]https://github.com/raywzy/ICT

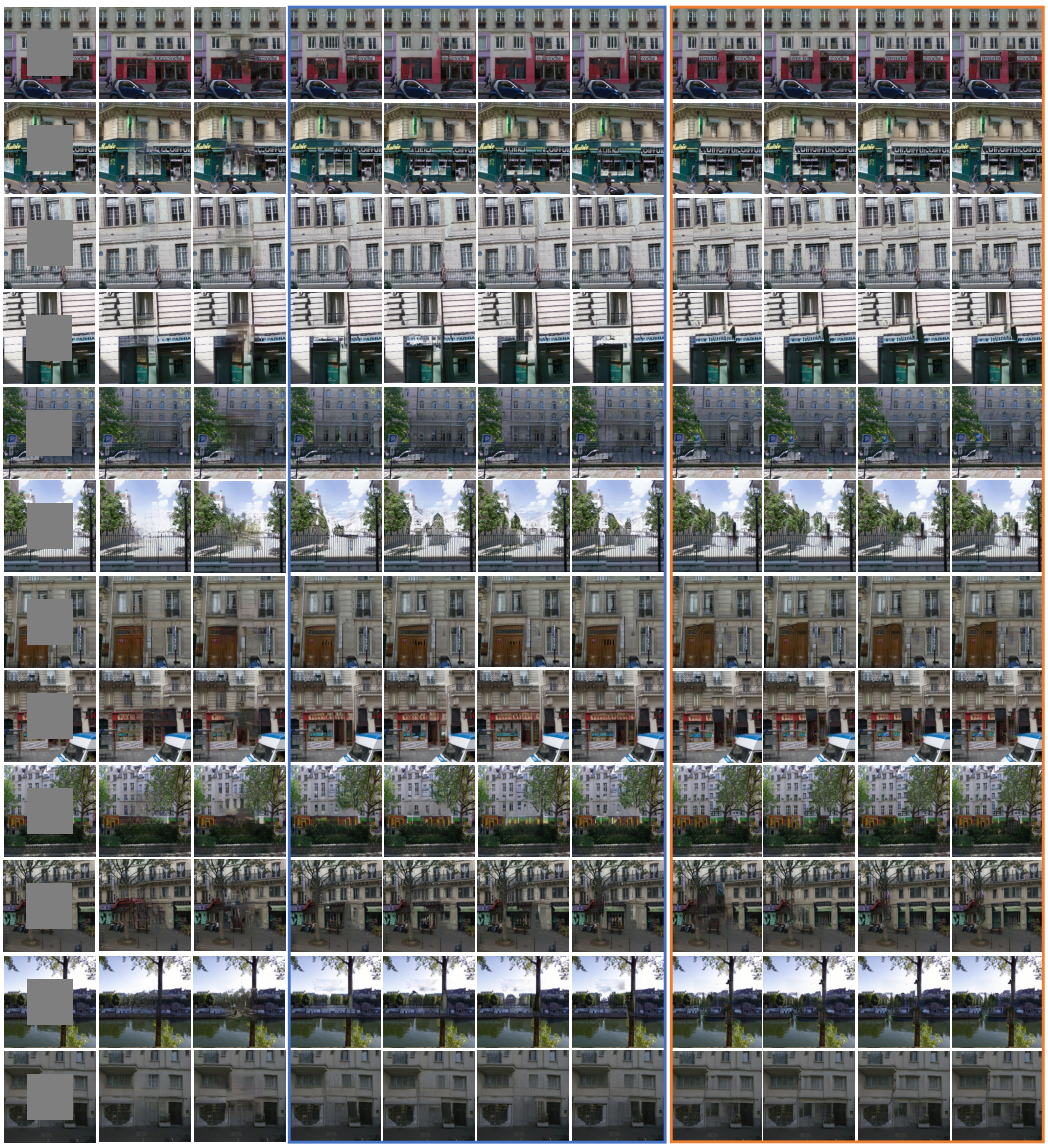
(a) Input (b) DFv2 (c) EC (d) MED (e) PIC (f) Ours

Figure 15: Pluralistic image completion results comparison with baselines. The original images come from CelebA-HQ. Best viewed by zooming in.



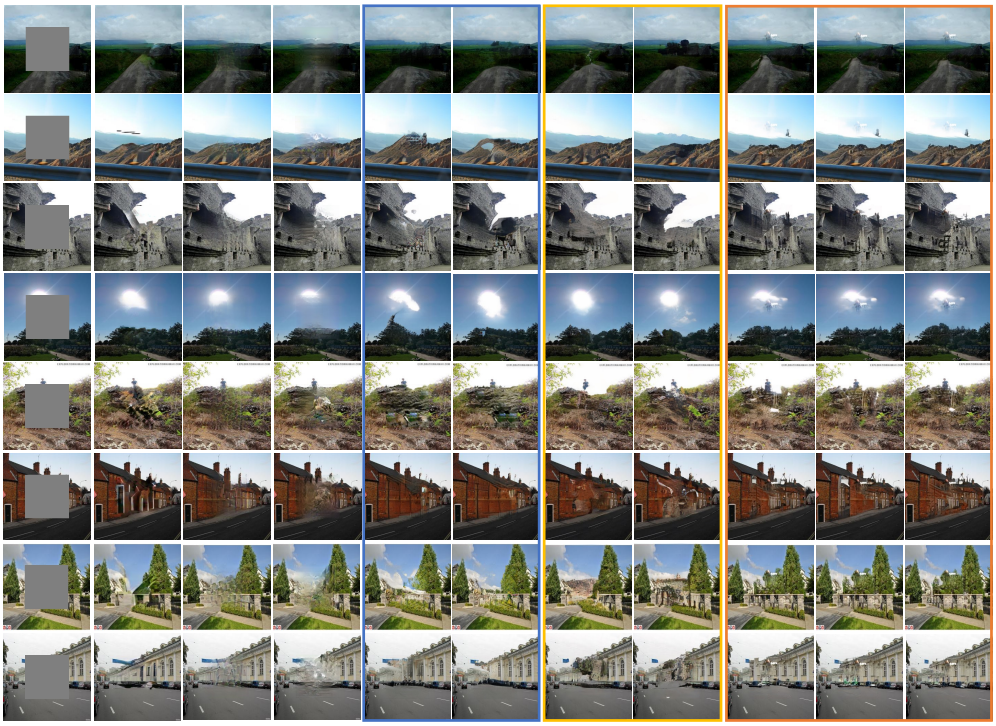
(a) Input (b) ICT (c) Ours

Figure 16: Pluralistic image completion results comparison with baselines. The original images come from FFHQ. Best viewed by zooming in.



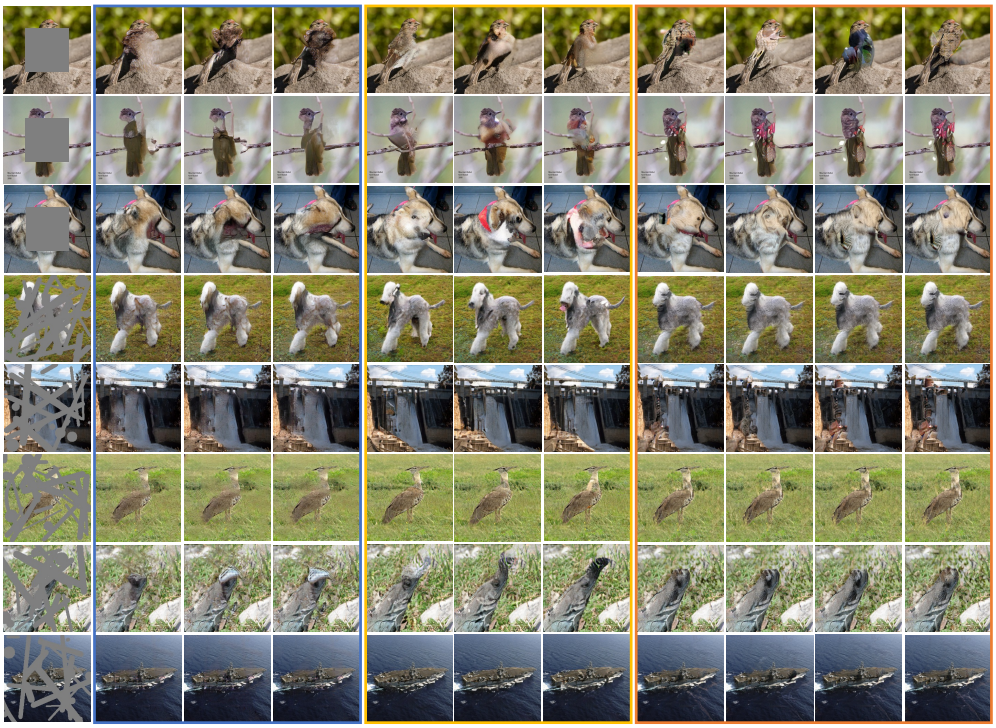
(a) Input (b) EC (c) MED (d) PIC (e) Ours

Figure 17: Pluralistic image completion results comparison with baselines. The original images come from Paris StreetView. Best viewed by zooming in.



(a) Input (b) DFv2 (c) EC (d) MED (e) PIC (f) ICT (g) Ours

Figure 18: Pluralistic image completion results comparison with baselines. The original images come from Places2. Best viewed by zooming in.



(a) Input (b) PIC (c) ICT (d) Ours

Figure 19: Pluralistic image completion results comparison with baselines. The original images come from ImageNet. Best viewed by zooming in.

E Ablation Study

To help understand the contribution of the different components in the proposed method, we provide ablation studies here. Specifically, experiments are conducted on the dataset CelebA-HQ with the setting of center masking. The number of primitives is set to $k = 6$. It should be noted that, without the loss L_{R}, the image completion is unsupervised, resulting in rather unrealistic completed images. The loss L_{R} is hence added by default. The results of ablation studies are provided in Table 5, which demonstrate the effectiveness of different components.

L_{F}	L_{BM}	L_{ELBO}	L_{A}	PSNR ↑	SSIM ↑	MAE ↓	FID ↓	LPIPS ↑	DIV-FID ↑
✓	✓	✓	✓	27.098	0.964	0.0196	11.021	0.098	27.923
✓	✓	✓	-	27.034	0.961	0.0194	11.193	0.095	27.534
✓	✓	-	-	27.035	0.961	0.0198	11.202	0.094	27.068
✓	-	-	-	26.251	0.922	0.0232	11.589	0.086	26.238

Table 5: Ablation study of PICMM.

F Results of Comparison with More Image Completion Methods

To make the experimental comparison more comprehensive and more convincing, we add the baseline VQ-VAE [37] and AOT-GAN [54]. The results of comparing our PICMM with the two baselines are shown in Table 6, which verify the effectiveness of our method clearly.

Method	PSNR ↑	SSIM ↑	MAE ↓	FID ↓	LPIPS ↑	DIV-FID ↑
VQ-VAE [37]	24.031	0.819	0.0309	30.747	0.083	23.676
AOT-GAN [54]	25.455	0.831	0.0233	26.427	0.089	25.843
PICMM†	**25.467**	**0.838**	**0.0223**	**24.334**	**0.091**	**28.033**

Table 6: The results of comparing our PICMM with VQ-VAE and AOT-GAN. The experiments are conducted on Places2 with random masking.

G Evaluations on the Number of Primitives

We supplement the experiments with different number of primitives, *i.e.*, $k = 1, 2, 3, 4, 5$. The experiments are conducted on CelebA-HQ using center masks. The implementation details keep the same as the details in the main paper. We provide experimental results in Table 7. The results show that the increase of k can diversify the completed images, without degenerating the performance with respect to common evaluation metrics.

k	PSNR ↑	SSIM ↑	MAE ↓	FID ↓	LPIPS ↑	DIV-FID ↑
1	26.792	0.961	0.0201	12.011	0.084	24.223
2	26.979	0.965	0.0198	11.579	0.089	26.001
3	27.152	0.960	0.0197	11.362	0.092	27.314
4	26.943	0.964	0.0197	11.638	0.096	27.540
5	27.121	0.966	0.0195	11.157	0.098	27.756

Table 7: Evaluations on the number of primitives k.

H Higher-Resolution Image Completion Results

Before this section, we resize the image to 256×256 for image completion. Here, to show that our method can generate higher-resolution images, we set the image size to 512×512. Experiments are performed on the datasets FFHQ. The results are provided in Figure 20.

I Possible Negative Impacts and Limitation

For possible negative impacts, since our PICMM is a method based on generative models, like a large number of methods of this kind, generated images with our method may danger to the security of AI systems, with the form of deep fakes. There are some advanced works targeting the issue, *e.g.*, [32, 34], which are helpful to handle the possible negative impacts.

For the limitation, our PICMM does not outperform ICT sometimes with respect to common metrics, since ICT uses a much larger transformer-based structure and iterative Gibbs sampling, which make ICT's inference slow. For the people preferring the performance on common metrics, this would be a limitation of our PICMM.

Figure 20: Pluralistic image completion results by our method. The original images come from FFHQ. The image size is 512×512. Best viewed by zooming in.