Don't Look into the Dark: Latent Codes for Pluralistic Image Inpainting

Haiwei Chen, Yajie Zhao
University of Southern California
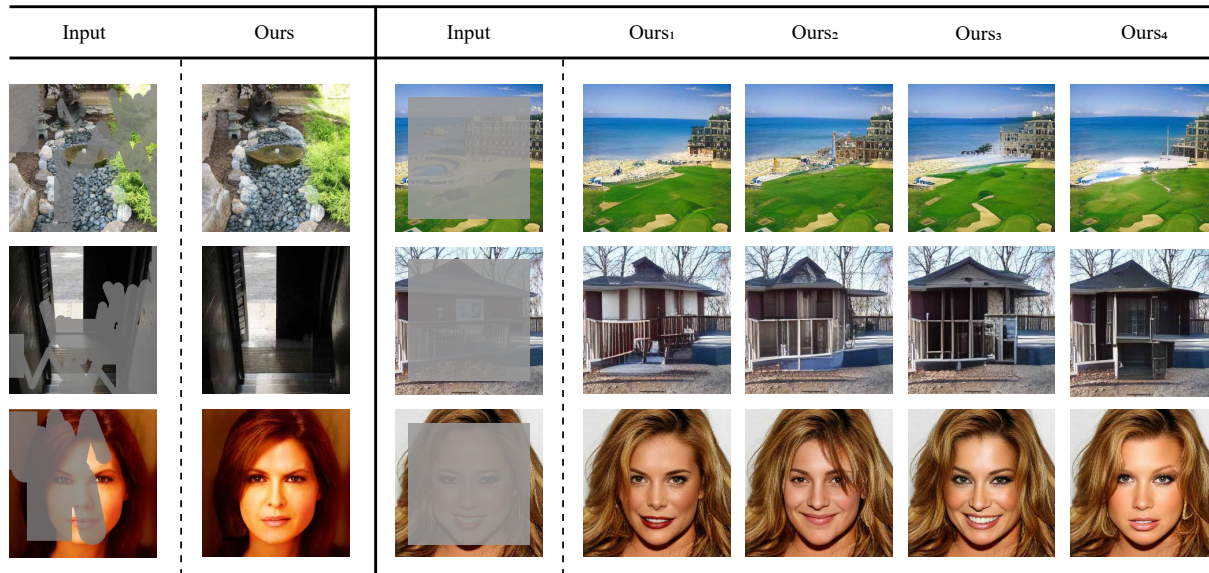USC Institute for Creative Technologies
{chenh, zhao}@ict.usc.edu

Figure 1. Inpainting results on the Places Dataset [50] (first two rows) and the CelebA-HQ Dataset [18] (third row). Our method is able to diversely complete partial image with free-form, large holes with state-of-the-art visual quality.

Abstract

We present a method for large-mask pluralistic image inpainting based on the generative framework of discrete latent codes. Our method learns latent priors, discretized as tokens, by only performing computations at the visible locations of the image. This is realized by a restrictive partial encoder that predicts the token label for each visible block, a bidirectional transformer that infers the missing labels by only looking at these tokens, and a dedicated synthesis network that couples the tokens with the partial image priors to generate coherent and pluralistic complete image even under extreme mask settings. Experiments on public benchmarks validate our design choices as the proposed method outperforms strong baselines in both visual quality and diversity metrics.

1. Introduction

Image inpainting is the task of filling the missing pixels of a masked image with appropriate contents that are coherent to its visible regions. As a long-studied topic in computer vision, image inpainting has evolved from a restoration technique solely relying on existing information from the input image (e.g. [3]) to data-driven generative methods (e.g. [23, 27, 36, 41, 44, 48]) that hallucinates detailed contents from not only the observable pixels but also learned, rich image priors.

Pluralistic inpainting refers to the ability of a model to generate multiple plausible results that complete a partial image. It offers a view of image inpainting as a generative method that models the smooth distributions of the complete images given the partial image as prior information [48]. However, modeling such distributions is challeng-

ing in the typical encoder-decoder network structures. In order to synthesize missing contents that both respect the partial image and maintain sample diversity, the decoder in this setting takes as input two types of information: 1) features propagated from the visible regions and 2) random noise vectors sampled from a prior distribution. If the training objective is to reconstruct a ground-truth image from a partial image, the objective itself may discourage conditioning on the random variable. Moreover, as the training dataset contains numerous examples that only require low-level information to complete an image (e.g. smoothly interpolating a wall texture), the model may choose to ignore the latent priors when the available image cues are strong enough to provide an answer. The phenomenon has been found in image translation networks [16], where adding noise to generate a conditional image does little to create pluralistic results.

In this paper, we investigate a unique approach to pluralistic image inpainting, following a branch of recently developed synthesis methods known as the generative transformers [4, 10, 29]. These approaches synthesize images by procedurally predicting latent codes, termed as "tokens", that semantically encode information of an image, analogous to sentence generation in natural language processing [8, 38]. As tokens drawn in each step are based on a learned posterior distribution of the previous step, generative transformers offer fine-grained control over diversifying the synthesized contents. Our paper is the first to adapt the generative transformer framework to target the image inpainting task. To this end, we design a three-stage pipeline to 1) encode a partial image into discrete latent codes, 2) predict the missing tokens with a bidirectional transformer, and 3) couple the predicted tokens with the partial image priors and decode them into a complete image. The resulted method design fundamentally differentiates itself from typical inpainting approaches in the past: instead of modeling the complex interaction between the missing and observed regions, our method barely looks at the missing regions during the encoding and token prediction stage: the designed mask-aware encoder utilizes restrictive convolutions to only operate on the visible and near-visible regions, and the bidirectional transformer only attends to the visible tokens to make a prediction. This separation design leads to a paradigm that divides feature reasoning and generative modeling into two separate stages, which we found beneficial for large-mask pluralistic image inpainting.

Through experiments, we have validated that our design choices lead to a robust, high-quality, and pluralistic solution to challenging image inpainting settings. Our method has achieved state-of-the-art performance both in terms of visual quality and sample diversity in two public benchmarks Places [50] and CelebA-HQ [18].

2. Related Works

The term "image inpainting" was originally introduced in [3] as a propagation algorithm that fills in the missing regions following lines near the boundaries. Following works are methodologically divided into two branches: the diffusion-based methods [1, 11, 21, 32, 35] that aim at interpolating information in the hole areas under local or global optimization constraints, and the patch-based methods [2, 6, 7, 9, 20] that composite candidate patches from selected regions into the hole areas. The early works are characterized by analyzing the inner statistics of the single image provided, many of which have shown satisfactory results for filling small holes, restoring relatively simple salient structures, and reconstructing textures.

Image inpainting has been advanced greatly by recent works that utilize artificial neural networks (ANN), as they introduce deep image priors to the synthesis of missing contents in a more robust and versatile way. Notably, the perceptual loss [17, 51] and the adversarial loss [12] have been found to drastically improve the visual quality achieved by the ANN methods. Many works have proposed designs that incorporate both local coherence and global consistency through multi-scale discriminators [15], feature rearrangement [34, 40], spectral convolution [36], attention layers [23, 25, 31, 43, 49] and GAN inversion [42]. Mask-aware operations have also been explored in various forms, notably with partial convolution [24], gated convolution [44] and continuously masked transformer [19]. Another relevant direction that shares similarity to our generative method is the progressive approach [13, 22, 45, 46], where inpainting is done step-wise via various forms of residual networks to gradually update both the mask and the feature maps of interest.

While the majority of inpainting works (e.g. [15, 24, 25, 28, 34, 36, 39–41, 43, 44, 49]) produce deterministic results given a fixed partial image input, various techniques have been proposed for pluralistic inpainting. [48] first studies the coupling of smooth latent priors with pixels generated for a masked image, where the authors suggest the undesirable side effect of reconstructive inpainting on sample diversity and propose a dual-path pipeline to better condition the generated results on both the observed regions and the sampled priors. [26] modulates a latent vector in the image space based on a coarse inpainting prediction. More recent works [23, 27] have shown major improvements in the visual quality of pluralistic inpainting. MAT [23] achieves state-of-the-art inpainting quality with a unifying design of normalized transformer, shifting attention and style code modulation. RePaint [27] leverages the denoising diffusion probabilistic model to image inpainting. The diffusion model can create realistic and diverse inpainting results through iteratively denoising resampled pixels in many steps. The limitation of it is the very slow sampling speed,
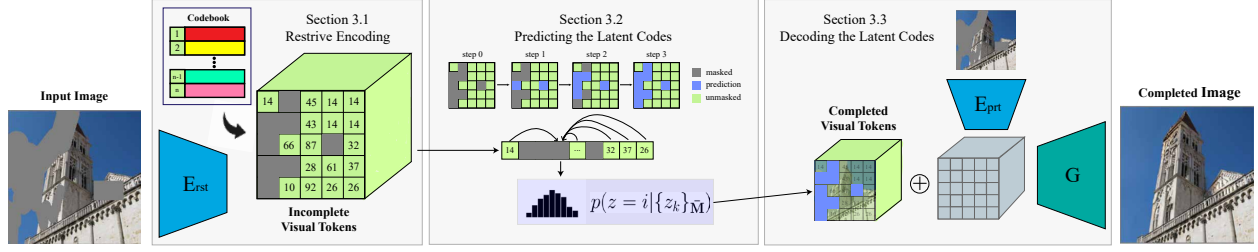
Figure 2. Overall pipeline of our method. E_{rst} denotes our proposed restrictive encoder that predicts partial tokens from the source image (see Section 3.1). The grey square space in the figure denotes missing tokens, which are iteratively predicted by a bidirectional transformer (see Section 3.2). E_{prt} denotes an encoder with partial convolution layers, which processes the source image into complementary features to the predicted tokens. The coupled features are decoded into a complete image by a generate G (see Section 3.3).

which may be hindering for some real-world applications.

Our method extends from a family of methods characterized by learning priors from discrete latent codes that are obtained from a vector-quantized autoencoder [37]. Past research in this direction has only focused on image synthesis, from directly predicting pixels as word tokens [5], to predicting tokens encoding visual features of larger receptive fields [10, 37]. While the pioneering works infer latent codes autoregressively, MaskGIT [4] finds it beneficial to synthesize an image in a scattered manner with a bidirectional transformer: in every iteration, a number of new codes are predicted in parallel and inserted into scattered locations of the code map until the entire grid is filled. While [4] has partially adapted its bidirectional framework to the image inpainting setting, our method design addresses several unanswered aspects of this adaptation: how partial images can be robustly masked into latent codes, and how the latent codes should be decoded into synthesized pixels that respect the observable area.

3. Method

Our method is divided into three stages to complete an input partial image. The neural network model takes as input a partial image X_M and a mask image M specifying the area to complete. The first stage encodes the partial image into a set of discrete tokens, referred to as latent codes, at a lower resolution and specifies the masked tokens that need to be predicted (Section 3.1); the second stage utilizes a bidirectional transformer to predict the missing tokens iteratively (Section 3.2); and the third stage couples the predicted tokens with features from the partial image and decodes them into a completed image (Section 3.3). Figure 2 provides a visualization of the overall pipeline of our method.

3.1. Encoding with Restrictive Convolutions

Our latent codes are represented by a discrete codebook $\mathbf{C}$ of learned tokens $\mathbf{C} = \{\mathbf{z}_k\}_{k=1}^K \in \mathbb{R}^{n_z}$, where k is the number of tokens and n_z is the number of channels

of each token feature. The tokens are given a set of labels $\mathbf{Y} = \{\mathbf{y}\}_{k=1}^K \in \mathbb{Z}$. We employ the setup of VQGAN [10] to learn the codebook by training on full images, where the token grid is $1/16$ of the resolution of the image. Ideally, encoding a partial image amounts to extracting valid tokens for the observed parts and invalid tokens, which are given a special [MASK] token, for the masked parts. However, the convolutional nature of the VQGAN network leads individual tokens to encode not just local information, but also information of its proximity. Directly encoding partial images with a VQGAN encoder thus leads to degradation of the tokens, as the masked region inevitably affects how an encoder chooses to extract tokens. Another concern is the fact that pixel-level masking (e.g. on a 256x256 image) does not directly translate to token-level masking (on a 16x16 token grid): small regions of masked pixels may still contain rich information in their neighborhood. However, directly down-sampling the mask image may lead many observable pixels to be masked out at a lower resolution, therefore discarding a fair amount of useful information. A good approach is thus needed to determine when a token should be considered masked in the encoding step.

Separating Masked Pixels The principle idea behind our encoding method is to prevent the participation of large areas of masked pixels in each convolutional network layer, controlled by a hyperparameter ratio α. To simplify the explanations, let's consider an encoder network with only non-strided convolutions and down-sampling layers. The **standard partial convolution** [24] widely used in past image inpainting works are characterized by scaling the matrix multiplication in the convolution operation and an update rule for the mask:

$$x' = \begin{cases} W^T(\mathbf{X} \odot \mathbf{M})\frac{1}{\text{sum}(\mathbf{M})} + b, & \text{if sum}(\mathbf{M}) > 0 \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

$$\mathbf{M}(x)' = \begin{cases} 1, & \text{if sum}(\mathbf{M}) > 0 \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where $\mathbf{X}$ denotes a $n \times n$ neighborhood in the feature map under a convolution kernel size of n, and $\mathbf{M}$ denotes the $n \times n$ mask in that area. $\odot$ denotes element-wise multiplication. Effectively, the standard partial convolution mitigates the impact of masked pixels on the features' signal strength with an adaptive scaling and propagates new features into the masked pixels as long as there are visible pixels surrounding them. In contrast, we choose to separate the image prior learning step and the synthesis step aggressively in two different stages. In the encoding stage, we thus propose the **restrictive partial convolution** that only considers regions surrounded by a certain proportion of visible pixels:

$$x' = \begin{cases} W^T(\mathbf{X} \odot \mathbf{M}) \frac{1}{\text{sum}(M)} + b, & \text{if } \frac{\text{sum}(\mathbf{M})}{\text{sum}(\mathbf{1})} >= \alpha \\ 0, & \text{otherwise}, \end{cases} \quad (3)$$

and at each down-sampling layer, we update the mask by:

$$\mathbf{M}(x)' = \begin{cases} 1, & \text{if } \frac{\text{sum}(\mathbf{M})}{\text{sum}(\mathbf{1})} >= \alpha \\ 0, & \text{otherwise}, \end{cases} \quad (4)$$

where $\mathbf{1}$ denotes a $n \times n$ constant tensor of value 1. Different from the standard partial convolution, the algorithm does not update the mask at each restrictive partial convolution layer. The changes made here prevent feature propagation to densely unmasked regions subject to the α value.

Besides separating the masked regions from the observed features, this particular convolution design also addresses the inevitable mismatch between the input pixel-level mask and the updated mask seen in the much-lower-resolution feature space. By using a small α value, the encoder is designated to fill in tokens for small regions of unseen pixels while leaving the larger regions to be predicted in the next stage (see Section 3.2). Figure 3 provides a visual illustration of this process: the smaller the α, the more likely that the encoder would predict a token label for local areas that are partially masked (marked by the red-colored grid locations in the figure). When α is set to be larger than 0.5, more observable pixels are "blocked out" and left to be predicted in the next stage. We have empirically found that setting $\alpha = 0.5$ produces the best inpainting results.

Encoder Design Let X denotes an input image, and M denotes a mask of the same size, given a pre-trained codebook $\mathbf{C}$ and VQGAN encoder E_{VQ} of a dataset, our encoder $E(X_M)$ learns to predict the probability of token labels in each visible region of a partial image $X_M = X \odot M$, supervised by the "ground-truth" token labels at those locations from encoding the complete image with $E_{VQ}(X)$. Our encoder is constructed by the restrictive partial convolutions and self-attention layers. It processes a partial image into probability estimations on the token labels $\mathbf{Y}_{\hat{M}} = \{\mathbf{y}_i\}$, given the down-sampled mask $\hat{M}$. The
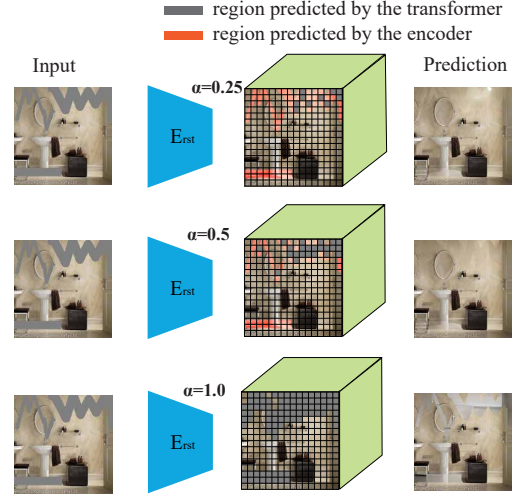

Figure 3. A visualization of mask down-sampling, shown on a 16x16 grid on the third column, from different α values following Equation 4. Smaller α values (top two rows) lead the restrictive encoder to predict tokens for more small mask areas (marked by the red pixels). Larger α is undesirable (bottom two rows) as it unnecessarily discards useful information from the image, leading to more inconsistent inpainting results.

training objective is hence minimizing the negative log-likelihood:

$$\mathcal{L}_{encoder} = -\mathbb{E}_{\mathbf{y} \in \mathbf{Y}}[\sum_{\forall i \in \hat{M}} \log p(y_i | X_M)], \quad (5)$$

where $\mathbf{y}$ are the target tokens, and X_M is the partial image.

3.2. Predicting the latent codes

The restrictive encoder has thus far encoded the input image into two distinctive regions of tokens: the visible region labeled by valid tokens $D = \{\mathbf{y}_i\}_{\hat{M}}$, and the unseen region $\bar{D} = \{\mathbf{m}\}_{1-\hat{M}}$ that contains a set of [MASK] tokens $\mathbf{m}$ (visualized by grey blocks in Figure 3). A bidirectional transformer based on the BERT model [8] is used to predict the token indices for each masked location in $\bar{D}$ based on the visible set of tokens D. The transformer retrieves visual features of the visible labels from the codebook, augments them with positional encoding, and processes them with attention layers to make independent label predictions on each masked location.

Training the transformer Training the generative transformer is simply by the maximum likelihood estimation - learning to predict the missing labels from the available ones:

$$\mathcal{L}_{\text{transformer}} = -\mathbb{E}_{\mathbf{y} \in \mathbf{Y}}[\sum_{\forall i \in \bar{D}} \log p(y_i | \mathbf{C}(D))], \quad (6)$$

where the only differences from Eq.5 are that the labels are predicted for the set of unseen locations $\bar{D}$ and the input to the transformer is a flattened list of visual tokens retrieved from the codebook $\mathbf{C}(D)$. We train the transformer with full images and only mask the down-sampled token map during the training. Specifically, during training, full images are encoded by the VQGAN encoder E_{VQ} to obtain a complete list of tokens. The list of tokens are then randomly masked by a ratio between 15% and 75%.

Sampling with the transformer During inference, the missing tokens are predicted iteratively through a parallel decoding algorithm [4]. Given a sampling step $k = 5$, and a cosine scheduling function f, the algorithm predicts labels for all missing tokens $\bar{D}$ at each step i, while only choosing to keep $n = f(i)$ predicted tokens with top prediction scores given by the transformer. The cosine scheduling function is chosen to ensure that $\sum_{i=0}^{k} f(i) = |\bar{D}|$. As pluralistic results are desired, we sample each token by drawing from its predicted probability distribution $p(y_i|\mathbf{C}(D))$.

An important modification we add to the sampling algorithm from [4] is the inclusion of an adaptive temperature t which scales the logits prior to the softmax function with $p_i = \frac{\exp(tz_i)}{\sum_n \exp(tz_i)}$. The temperature controls the confidence level of the sampler: the lower the temperature, the more likely that labels with higher confidence scores would be sampled. Empirically, we found it beneficial to start with a high temperature and gradually anneal the temperature in each step. This encourages the sampler to introduce more diverse tokens early on and draw with more certainty when more evidence is present. In our model, we use a starting temperature of 1, and set the annealing factor $s = 0.9$, which scales the temperature value in each sampling step. The chosen starting temperature has substantial impacts on the visual quality and diversity of the inpainting results (see the ablation study section in Section 4).

3.3. Decoding the latent codes

Due to the discrete, quantized nature of the codebook representation, the visual tokens learned to encode an image usually do not fully recover the original image. Stochastically sampled tokens may further alter the global appearance of the fill-in area if they are decoded into an image directly (see Figure. 4.A). Therefore, compositing the generated pixels with the partial image oftentimes results in noticeable discontinuities at the mask boundaries (see Figure 4.B).

In order to synthesize pixels that coherently complete the partial images, we find it necessary to couple the quantized latent codes with smooth image priors encoded from the input partial image. Since the smooth features are responsible for locally bridging the synthesized contents and the existing contents, the encoder, denoted as E_{prt}, utilizes layers of the standard partial convolution (Eq.1) to extract local features propagated to the masked regions. The features in the
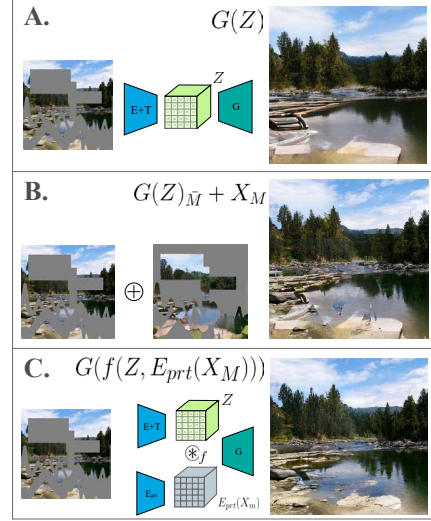

Figure 4. A visual comparison between the decoder designs. **A**. Directly decoding the predicted latent codes Z with the restrictive encoder E and transformer T, and **B**. its composition with the source image X_M. **C**. Our proposed decoding design, where partial image priors $E_{prt}(X_M)$ are composed with Z through a composition function f described in Equation.7-8.

masked region are combined with the latent codes via an averaging operation:

$$h1 = (1 - \hat{M})(Z + E_{prt}(X_M))/2, \qquad (7)$$

where Z is a feature map of the tokens predicted by the transformer. The recomposed features in the empty space are then combined with features extracted from the visible area and decoded by a convolutional generator G:

$$X' = G(h1 + \hat{M}E_{prt}(X_M)). \qquad (8)$$

The design is visualized in Figure 4.C. During training, the network learns to recover the ground truth image X given the partial image X_M and a set of chosen latent codes. As we train the network with a reconstruction objective, we use the set of latent codes obtained from encoding the ground truth image X, where $Z = E_{VQ}(X)$. The encoder E_{prt} and the generator G are optimized by a combination of the adversarial loss [12] $\mathcal{L}_{adv}$ with R1 regularization [30] and the perceptual reconstruction loss (LPIPS) [47] $\mathcal{L}_P$ with:

$$\mathcal{L}_{decode} = \mathcal{L}_{adv} + 0.1R_1 + 0.1\mathcal{L}_P. \qquad (9)$$

4. Experiments

Dataset Our experiments are conducted on the Places365-Standard [50] and the CelebA-HQ [18], two benchmarks widely evaluated by past image inpainting methods. The
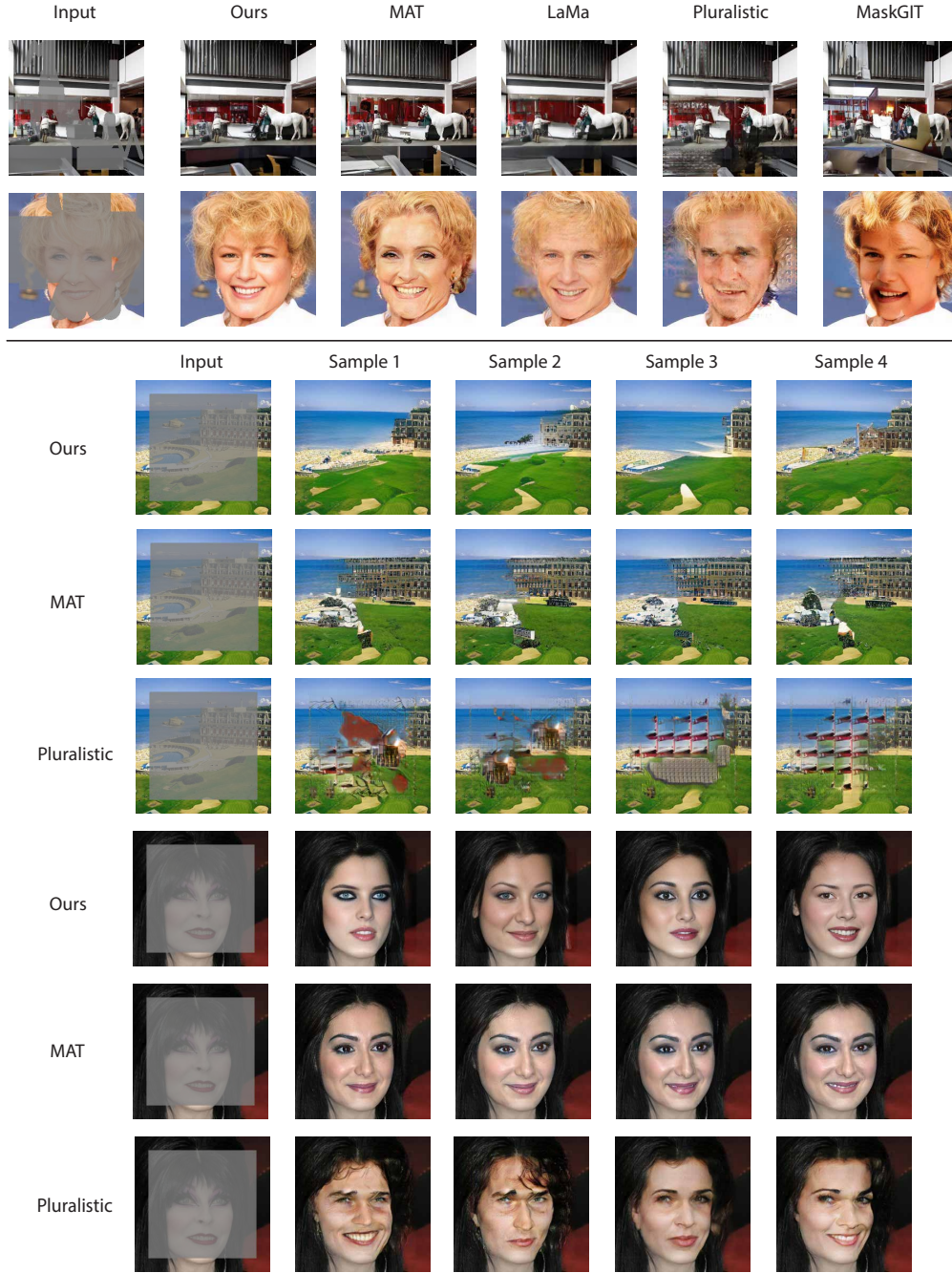
Figure 5. Visual examples on inpainting with both the random masks (upper half) and the challenging large box mask (lower half), compared to the selected baseline methods.

Places365-Standard dataset contains 1.8 million images for training and 36.5 thousand images for evaluation across over 205 scene categories. The CelebA-HQ dataset is split into 24,183 training images and 2,993 test images. For both datasets, we use an image resolution of 256×256. We further set three different mask settings for our experiments:

1) small random hole, 2) large random hole, and 3) large box hole. The first two settings are directly adopted from MAT [23] (free-form holes with strokes and boxes). The third, challenging setting uses a very large box mask centered in the image with its width and length equal to 80% of the image size. As such setting leaves a majority of the

Methods	Places (256 × 256)				CelebA-HQ (256 × 256)			
	FID↓			Diversity↑	FID↓			Diversity↑
	Small Mask	Large Mask	Box	Box	Small Mask	Large Mask	Box	Box
Ours	**1.02**	**2.82**	**13.30**	0.29±0.06	**2.70**	**5.04**	**12.79**	**0.28±0.05**
MAT[23]	1.19	3.32	17.5	0.26 ± 0.04	2.94	5.16	15.18	0.10 ± 0.02
LaMa[36]	1.22	3.78	19.48	-	3.98	8.75	23.24	-
MaskGIT[4]	19.84	36.38	52.71	**0.39±0.05**	19.68	40.76	30.87	0.25±0.04
Pluralistic[48]	4.83	16.26	86.57	-	9.7	28.89	43.08	0.18±0.02

Table 1. Comparisons of FID and diversity scores to the baseline methods. **Bold** text denotes the best, and blue text denotes the second. Since LaMa [36] does not generate pluralistic results, and Pluralistic [48] produces degenerate results in the Places Box setting, we omit their diversity scores in the table.

pixels empty, we find it suitable for the evaluation of pluralistic inpainting, as well as on whether the inpainting method extends to the extreme cases.

Evaluation metric Our main objectives are to evaluate both the visual quality and the sample diversity of the inpainted image. To this end, we opt for the perceptual metric FID [14] and the LPIPS-based diversity score [51]. Specifically, to compute the diversity score for each dataset, we use a smaller subset of both Places-Standard and CelebA-HQ with 1000 images. For each image, 100 inpainting samples are drawn under the large box setting, and the diversity score for each sample is computed as the average of the pair-wise LPIPS distances between the drawn samples. The final score shows the average of the individual scores and their standard deviation. In addition, we provide visual examples for a qualitative evaluation (see Figure 5,6 and more in the Supplements).

4.1. Comparisons to The State of Arts

Our results are compared to four recent baseline image inpainting methods. The baseline methods are evaluated directly with the provided pre-trained models and their public source codes: 1) **MAT** [23] is the current state-of-the-art method in image inpainting that is able to produce high-quality and pluralistic inpainting results in challenging settings. 2) **LaMa** [36] is another state-of-the-art method that is characterized by the use of Fourier convolution. However, it does not generate pluralistic results; 3) **MaskGIT** [4] is a latent-code-based image synthesis method that has strongly motivated our work and has been shown to be adaptable to image inpainting; and 4) **Pluralistic** [48] is a seminal work in exploring pluralistic image inpainting by coupling smooth latent priors with the partial image features.

Table 1 lists the quantitative evaluation results: ours outperforms the baseline methods in all settings in terms of FID and diversity score, except for the diversity score on Places, where we rank the second best. While MaskGIT produces more diverse results on the Places dataset, it does so at the cost of visual quality, as the method is not designed to coherently compose the synthesized contents with

Type	Model	FID↓
	Full Model	**12.81**
Temperature	$t = 0.1$	14.62
	$t = 0.5$	13.61
	$t = 2$	13.96
Restrictive Conv	$\alpha = 0.25$	12.9
	$\alpha = 0.75$	14.12
	$\alpha = 1.0$	15.04
Network Design	Vanilla Encoder	15.32
	Vanilla Decoder	18.50

Table 2. Quantitative ablation study. "Temperature" adjustments change the temperature value in the sampling procedure. "Restrictive Conv" adjustments change the mask update rule in the restrictive encoder. "Network Design" adjustments replace our designed network structures with the vanilla ones: for the "Vanilla Encoder" setting, an encoder network with the regular convolution layers are used; for the "Vanilla Decoder" setting, the predicted latent codes are directly decoded into an image. In the "Full Model", we set $t = 1.0$ and $\alpha = 0.5$.

the existing contents. Figure 5 shows visual examples of the inpainting results under the three different mask settings. While our results can be seen slightly better in inpainting consistency compared to the baselines, the method truly shines in the pluralistic comparisons (lower half of the figure): while our model synthesizes pluralistic inpainting results that are high-quality and varied in both local details and global structures, the baseline methods either only produce locally diversifies results (MAT [23], see 4th,7th rows in Figure 5) or fail to generate visually coherent results in the challenging box setting (Pluralistic [48], see 5th,8th rows in Figure 5).

4.2. Ablation Study

We validate several design choices of our model with an ablation study on a smaller evaluation subset of the Places dataset with 3000 images. Table 2 shows the quantitative results of the ablation study based on the the FID score metric.

Effectiveness of the Restrictive Design One main question we ask when designing the restrictive encoder is how
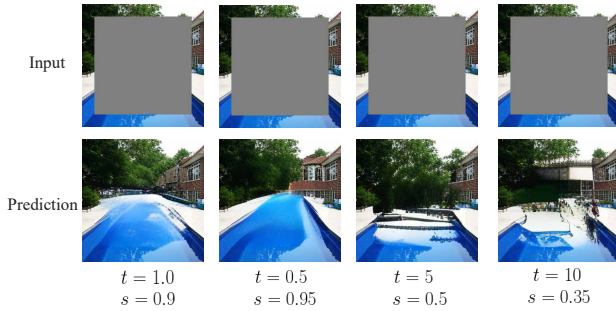
Figure 6. Comparisons of inpainting results with regard to different sampling temperature t and annealing factors s.

Figure 7. Failure cases in our results.

much our model needs to separate small mask regions. In the extreme case, a one-pixel mask can be turned into a masked token in the down-sampled token grid, thus discarding 93.74% of information around that pixel. Obviously, this behavior is not desirable and we have experimentally found that larger α values lead to decreased performance (see Table 2 and Figure 3). Our final choice of $\alpha = 0.5$ leads our encoder to complete the inpainting by itself for local regions that are less than 50% masked. On average, the restrictive encoder achieves a label classification accuracy of 23% for labels in small mask regions and near mask boundaries in the large mask setting (tested on the Places dataset), whereas an encoder with the regular convolution only achieves an accuracy of 9%. Furthermore, due to the inductive bias learned in the training process, a "wrongly" classified label does not necessarily translate to poor inpainting results. The result could be a completed image that is different from the original one, while still visually plausible. We provide some further visual comparisons with regard to this observation in the supplementary material.

Effectiveness of the Sampling Function As described in Section 3.2, we have found the adaptive temperature a key factor in controlling the visual quality and diversity outcome of the inpainting. Figure 6 provides a visual comparison between different configurations of the temperature and annealing factor. We have observed that a larger starting temperature creates more diverse output, though at the risk of breaking the coherence of the inpainting. Overly small temperature, on the other hand, led our model to mainly interpolate patterns in the masked area, resulting in large areas of homogeneous textures in the inpainting results. More visual comparisons are included in the supplementary material.

4.3. Limitations of our model

Our model shares limitations with previous inpainting methods in its limited ability to complete semantically salient objects such as shops and furniture, as well as people and animals, when trained on the Places dataset (see Fig-
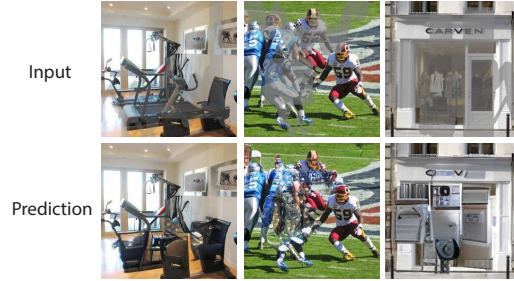
ure 7). The inference speed of our method is also relatively slower than the end-to-end, single-pass inpainting methods, as the majority of the computation time is spent on the iterative sampling steps (see supplementary material for an inference time comparison). We also have left some areas unexplored. For instance: whether the aforementioned synthesis problem can be mitigated by training with semantic labels, or how the model extends to higher resolution input. We believe that these extensions are well within reach as past latent code synthesis methods (e.g. [4, 10]) have demonstrated such capabilities.

5. Conclusion

In this paper, we present a pluralistic image inpainting method that first analyzes only the visible and near-visible regions through latent code prediction, and synthesizes the missing contents through a versatile bidirectional transformer and a reconstruction network that composes the code prediction with partial image priors. We have validated our design choices through comparative experiments on public benchmarks and an ablation study, as our method achieves state-of-the-art performance in both visual quality and sample diversity.

Acknowledgements

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

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph., 28(3):24, 2009. 2

[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. 1, 2

[4] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11315–11325, 2022. 2, 3, 5, 7, 8

[5] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In International conference on machine learning, pages 1691–1703. PMLR, 2020. 3

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

[7] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. Image melding: Combining inconsistent images using patch-based synthesis. ACM Transactions on graphics (TOG), 31(4):1–10, 2012. 2

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018. 2, 4

[9] Iddo Drori, Daniel Cohen-Or, and Hezy Yeshurun. Fragment-based image completion. In ACM SIGGRAPH 2003 Papers, pages 303–312. 2003. 2

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12873–12883, 2021. 2, 3, 8

[11] Mohamed-Jalal Fadili, J-L Starck, and Fionn Murtagh. Inpainting and zooming using sparse representations. The Computer Journal, 52(1):64–79, 2009. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014. 2, 5, 1

[13] Zongyu Guo, Zhibo Chen, Tao Yu, Jiale Chen, and Sen Liu. Progressive image inpainting with full-resolution residual network. In Proceedings of the 27th acm international conference on multimedia, pages 2496–2504, 2019. 2

[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. Advances in neural information processing systems, 30, 2017. 7

[15] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. ACM Transactions on Graphics (ToG), 36(4):1–14, 2017. 2

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1125–1134, 2017. 2

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and superresolution. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pages 694–711. Springer, 2016. 2

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In International Conference on Learning Representations, 2018. 1, 2, 5, 6

[19] Keunsoo Ko and Chang-Su Kim. Continuously masked transformer for image inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 13169–13178, 2023. 2

[20] Joo Ho Lee, Inchang Choi, and Min H Kim. Laplacian patch-based image synthesis. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2727–2735, 2016. 2

[21] Levin and Zomet. Learning how to inpaint from global image statistics. In Proceedings Ninth IEEE international conference on computer vision, pages 305–312. IEEE, 2003. 2

[22] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7760–7768, 2020. 2

[23] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10758–10768, 2022. 1, 2, 6, 7

[24] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European conference on computer vision (ECCV), pages 85–100, 2018. 2, 3

[25] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4170–4179, 2019. 2

[26] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9371–9381, 2021. 2

[27] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11461–11471, 2022. 1, 2

[28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2536–2544, 2016. 2

[29] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems, 32, 2019. 2

[30] Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In Proceedings of the AAAI conference on artificial intelligence, 2018. 5, 1

[31] Pourya Shamsolmoali, Masoumeh Zareapoor, and Eric Granger. Transinpaint: Transformer-based image inpainting with context adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 849–858, 2023. 2

[32] Jianhong Shen and Tony F Chan. Mathematical models for local nontexture inpaintings. SIAM Journal on Applied Mathematics, 62(3):1019–1043, 2002. 2

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 2

[34] Yuhang Song, Chao Yang, Zhe Lin, Xiaofeng Liu, Qin Huang, Hao Li, and C-C Jay Kuo. Contextual-based image inpainting: Infer, match, and translate. In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018. 2

[35] Jian Sun, Lu Yuan, Jiaya Jia, and Heung-Yeung Shum. Image completion with structure propagation. In ACM SIGGRAPH 2005 Papers, pages 861–868. 2005. 2

[36] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. arXiv preprint arXiv:2109.07161, 2021. 1, 2, 7

[37] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. Advances in neural information processing systems, 30, 2017. 3

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017. 2

[39] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5840–5848, 2019. 2

[40] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In Proceedings of the European conference on computer vision (ECCV), pages 1–17, 2018. 2

[41] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7508–7517, 2020. 1, 2

[42] Ahmet Burak Yildirim, Hamza Pehlivan, Bahri Batuhan Bilecen, and Aysegul Dundar. Diverse inpainting and editing with gan inversion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 23120–23130, 2023. 2

[43] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5505–5514, 2018. 2

[44] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4471–4480, 2019. 1, 2

[45] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16, pages 1–17. Springer, 2020. 2

[46] Haoran Zhang, Zhenzhen Hu, Changzhi Luo, Wangmeng Zuo, and Meng Wang. Semantic image inpainting with progressive generative networks. In Proceedings of the 26th ACM international conference on Multimedia, pages 1939–1947, 2018. 2

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018. 5, 2

[48] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1438–1447, 2019. 1, 2, 7

[49] Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai, and Dinh Phung. Bridging global context interactions for high-fidelity image completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11512–11522, 2022. 2

[50] 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. 1, 2, 5, 4

[51] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. Advances in neural information processing systems, 30, 2017. 2, 7