Continuously Masked Transformer for Image Inpainting

Keunsoo Ko
The Catholic University of Korea
ksko@catholic.ac.kr

Chang-Su Kim*
Korea University
changsukim@korea.ac.kr

Abstract

A novel continuous-mask-aware transformer for image inpainting, called CMT, is proposed in this paper, which uses a continuous mask to represent the amounts of errors in tokens. First, we initialize a mask and use it during the self-attention. To facilitate the masked self-attention, we also introduce the notion of overlapping tokens. Second, we update the mask by modeling the error propagation during the masked self-attention. Through several masked self-attention and mask update (MSAU) layers, we predict initial inpainting results. Finally, we refine the initial results to reconstruct a more faithful image. Experimental results on multiple datasets show that the proposed CMT algorithm outperforms existing algorithms significantly. The source codes are available at https://github.com/keunsoo-ko/CMT.

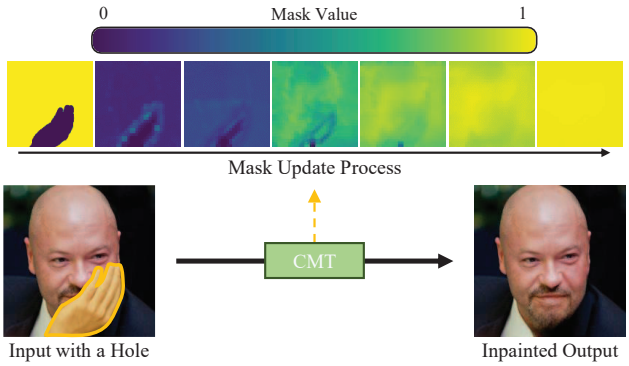


Figure 1: Illustration of the mask update process in the proposed CMT algorithm: The man's hand is designated as a hole. Initially, in the mask, the hole pixels are assigned 0, while the others 1. The mask is updated gradually through several MSAU layers, and the inpainted image is obtained.

1. Introduction

The objective of image inpainting is to reconstruct visually plausible images by filling in holes or defects in input images, such as old photos with scratches and flawed photos with distracting objects. In inpainting, it is necessary to predict the contents inside holes based on intact regions. Early inpainting algorithms [3, 15, 25, 9] fill in a hole patch using similar patches in the same image or from an external database, but they may fail to generate detailed patterns.

Recently, convolutional neural networks (CNNs) have been developed for inpainting [24, 30, 12, 32, 36], achieving promising performances. However, they suffer from visual artifacts, such as blurriness, color discrepancies, and artificial edges, since ordinary convolutional layers are applied to all pixels — both hole and non-hole pixels — in the same manner. To alleviate such artifacts, mask-aware inpainting algorithms based on CNNs have been proposed [17, 33, 21, 31], in which adaptive filtering is performed to process each pixel according to its state.

With the success of the transformers [5, 20] in vision tasks, transformer-based inpainting algorithms [35, 28, 4, 19] also have been proposed. They provide decent inpainting results, for the global attention facilitates to inpaint holes using the information in distant regions. In particular, Li *et al*. [16] developed a mask-aware transformer, which classifies tokens as either valid or invalid using a binary mask. Their algorithm declares an output token as valid if it depends on at least one valid input token. However, this binary masking has a limitation that a valid token can be still erroneous.

In this paper, we propose a novel transformer for image inpainting, called continuously masked transformer (CMT), which uses a continuous mask to represent the amounts of errors in tokens. First, we initialize a continuous mask and use it during the self-attention process. To facilitate the masked self-attention, we employ overlapping tokens, which consist of ordinary and shifted tokens. Then, we update the mask by modeling the error propagation during the masked self-attention as illustrated in Figure 1. We generate initial inpainting results through several masked self-attention and mask update (MSAU) layers. Finally, we develop the refinement network to refine the initial results to reconstruct a more faithful image. Extensive experiments demonstrate that the proposed CMT algorithm provides excellent inpainting performance.

*Corresponding author

This work has the following major contributions:

- To the best of our knowledge, CMT is the first continuous-mask-aware transformer for image inpainting.
- We also develop a novel mask update scheme by formulating the error propagation during the forward pass in the network.
- We introduce the notion of overlapping tokens to facilitate more communication among tokens during the masked self-attention.
- The proposed CMT algorithm significantly outperforms conventional image inpainting algorithms on the Places2 [40], CelebA-HQ [14], and DTD [2] datasets.

2. Related Work

2.1. Deep inpainting

Early inpainting neural networks [24, 30] were trained to fill in the square hole at the center of an image, where the hole-to-image (H2I) area ratio is fixed at $\frac{1}{4}$. This training strategy, however, may not be suitable for practical applications, in which holes of various shapes and sizes should be inpainted. To overcome this issue, one or more rectangular holes of different sizes are placed at random positions in [12, 32, 36], or complex irregular patterns are generated with numerous H2I ratios around 0.3 within $(0.01, 0.6]$ in [17, 33, 37]. The latter irregular patterns have been adopted to train and test many inpainting algorithms [23, 29, 18, 34, 8, 35]. Moreover, with the advances in generative models [7, 22], several pluralistic inpainting algorithms [39, 28, 38, 27, 16] have been developed. Given a damaged image containing huge holes, they can generate multiple inpainted images. They focus on the generation or synthesis of plausible contents constrained on intact regions, rather than on the faithful reconstruction of original contents. In this paper, we aim at reconstructing an image when hole regions cover a moderate portion of the image, as done in [17, 33, 37, 23, 29, 18, 34, 8, 35].

2.2. Mask-aware inpainting

Ordinary convolutional layers are used for inpainting in [24, 12, 30, 32], which are applied to all pixels including missing ones inside holes, but they often cause visual artifacts such as blurriness and artificial edges. To reduce such artifacts, mask-aware convolutional layers have been proposed [17, 33, 31, 29], which exploit masks to process each pixel adaptively according to its state. Liu *et al*. [17] proposed the partial convolutional layer using a binary mask, which assigns 1 to valid (or error-free) pixels and 0 to invalid (or erroneous) ones. Their algorithm excludes invalid pixels from the convolution and then updates the mask: an output pixel is declared valid if it depends on at least one valid input pixel. This hard binary masking, however, has a limitation that it cannot express the amounts of errors in a soft manner. For example, in a deeper layer, more pixels are affected by erroneous input pixels, but fewer pixels are declared invalid by the binary masking. To overcome this limitation, Yu *et al*. [33] proposed the gated convolutional layer using a continuous mask, which is learned by the network through end-to-end training. Yi *et al*. [31] developed a lightweight gated convolutional layer based on separable convolution [11]. Xie *et al*. [29] proposed the bidirectional mask updating, which employs a reverse mask backwardly updated from the last layer, as well as an ordinary mask. Also, Ma *et al*. [21] proposed the region-wise convolutions, which process existing and missing regions separately.

With the success of the transformer [5] in various vision tasks including image inpainting [35, 28], Li *et al*. [16] proposed a mask-aware transformer for inpainting. Similarly to the partial convolution layer [17], they adopted binary masking, which classifies tokens as either valid or invalid. However, this binary masking also suffers from the aforementioned limitation. Even when a token is declared valid, it may be still erroneous. Furthermore, different valid tokens may contain different amounts of errors. It hence may not be the best strategy to regard valid tokens as error-free during the self-attention process in the transformer. Therefore, in this paper, we propose CMT to exploit a continuous mask in the self-attention and also design a mask update scheme to represent the error propagation during the masked self-attention. Note that the proposed mask update scheme is fundamentally different from Yu *et al*.'s learning-based scheme [33]; we formulate the error propagation during the forward pass in the network explicitly.

3. Proposed Algorithm: CMT

Given an image, we generate tokens and conduct self-attention using a continuous mask. Meanwhile, we update the mask by modeling the error propagation. It is recommended to watch the video in the supplement.

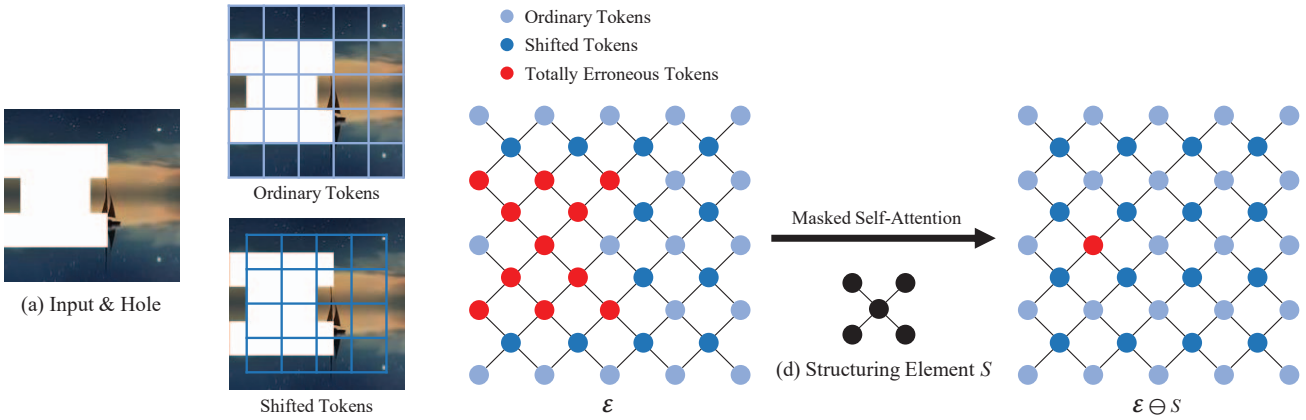
Tokenization & mask initialization: We partition an image into N patches and flatten them to yield $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^t$. We then generate a binary mask $B = [\mathbf{b}_1, \ldots, \mathbf{b}_N]^t$ of the same size, each element of which is set to 1 if the corresponding element in X is error-free and 0 otherwise.

Then, we obtain tokens $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^t$ by employing a projection matrix U_z, given by

$$\mathbf{z}_i = \eta(U_z \mathbf{x}_i + \mathbf{p}_i) \tag{1}$$

where $\mathbf{p}_i$ is a learnable position bias, and η denotes the layer normalization [1]. We then initialize continuous token masks $M = [\mathbf{m}_1, \ldots, \mathbf{m}_N]^t$ by

$$\mathbf{m}_i = \phi(U_z, \mathbf{b}_i) \tag{2}$$



(a) Input & Hole (b) Overlapping Tokens (c) Token Graph before Attention (e) Token Graph after Attention

Figure 2: Illustration of overlapping tokens and the masked self-attention process: An image in (a) is decomposed into overlapping tokens in (b). In (c), the tokens are represented by a graph. Note that the set $\mathcal{E}$ of totally erroneous tokens in (c) is eroded into $\mathcal{E} \ominus \mathcal{S}$ in (e) by the masked self-attention, where $\mathcal{S}$ is a structuring element in (d).

where the error propagator ϕ is defined as

$$\mathbf{m}_{\text{out}} = \phi(U, \mathbf{m}_{\text{in}}) \triangleq \text{abs}(U)\mathbf{m}_{\text{in}} \oslash \text{abs}(U)\mathbf{1}. \quad (3)$$

Here, $\oslash$ denotes the element-wise division, and $\mathbf{1}$ is a column vector consisting of all 1's. In (3), $\text{abs}(U)$ is used instead of U to prevent positive and negative coefficients in U from canceling out nonnegative values in the input mask $\mathbf{m}_{\text{in}}$. Also, by dividing by $\text{abs}(U)\mathbf{1}$, all values in the output mask $\mathbf{m}_{\text{out}}$ are also normalized to $[0, 1]$: 0 indicates a totally erroneous element, while 1 does an error-free one. Consequently, $\phi(U_z, \mathbf{b}_i)$ in (2) predicts approximately how much errors propagate during the tokenization in (1). In other words, the token mask $\mathbf{m}_i$ in (2) represents the amounts of element-wise errors in the token $\mathbf{z}_i$ in (1).

Masked self-attention & mask update: We extract query Q, key K, and value V from the token matrix Z by

$$Q = [\mathbf{q}_1, \cdots, \mathbf{q}_N]^t = ZU_q^t, \quad (4)$$
$$K = [\mathbf{k}_1, \cdots, \mathbf{k}_N]^t = ZU_k^t, \quad (5)$$
$$V = [\mathbf{v}_1, \cdots, \mathbf{v}_N]^t = ZU_v^t, \quad (6)$$

where U_q, U_k, U_v are projection matrices. Equivalently, we have

$$\mathbf{q}_i = U_q\mathbf{z}_i, \quad \mathbf{k}_i = U_k\mathbf{z}_i, \quad \mathbf{v}_i = U_v\mathbf{z}_i. \quad (7)$$

Hence, similarly to (2), we obtain the masks for these query, key, and value vectors by

$$\bar{\mathbf{q}}_i = \phi(U_q, \mathbf{m}_i),\ \bar{\mathbf{k}}_i = \phi(U_k, \mathbf{m}_i),\ \bar{\mathbf{v}}_i = \phi(U_v, \mathbf{m}_i). \quad (8)$$

Then, we compute the attended output

$$Z^* = \begin{bmatrix} (\mathbf{z}_1^*)^t \\ \vdots \\ (\mathbf{z}_N^*)^t \end{bmatrix} = A \begin{bmatrix} (\bar{\mathbf{v}}_1 \otimes \mathbf{v}_1)^t \\ \vdots \\ (\bar{\mathbf{v}}_N \otimes \mathbf{v}_N)^t \end{bmatrix} \quad (9)$$

where $\otimes$ denotes the element-wise multiplication. Also, the attention matrix $A = [\mathbf{a}_1, \ldots, \mathbf{a}_N]^t$ is determined by

$$\mathbf{a}_i^t = \text{softmax}([\mu(\mathbf{q}_i, \mathbf{k}_1), \ldots, \mu(\mathbf{q}_i, \mathbf{k}_N)]) \quad (10)$$

where

$$\mu(\mathbf{q}_i, \mathbf{k}_j) \triangleq (\bar{\mathbf{q}}_i \otimes \mathbf{q}_i)^t(\bar{\mathbf{k}}_j \otimes \mathbf{k}_j). \quad (11)$$

Note that, in the masked self-attention in (9)$\sim$(11), we use the masked vectors $\bar{\mathbf{q}}_i \otimes \mathbf{q}_i$, $\bar{\mathbf{k}}_i \otimes \mathbf{k}_i$, and $\bar{\mathbf{v}}_i \otimes \mathbf{v}_i$ instead of $\mathbf{q}_i$, $\mathbf{k}_i$, and $\mathbf{v}_i$ in the standard $\mathbf{qkv}$ self-attention [5].

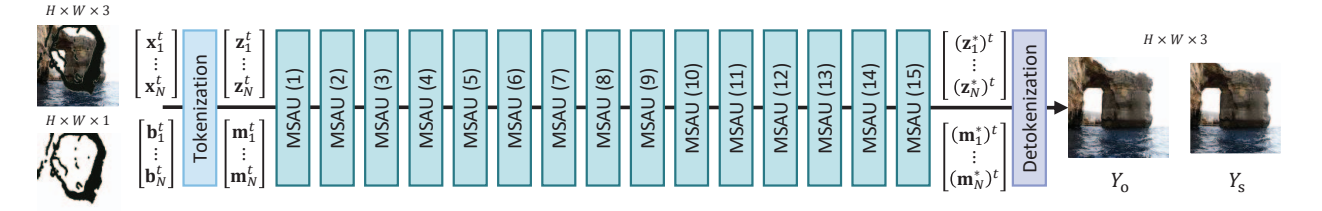
Using the attention matrix A, we also update the masks for the attended tokens in Z^* in (9) by

$$M^* = \begin{bmatrix} (\mathbf{m}_1^*)^t \\ \vdots \\ (\mathbf{m}_N^*)^t \end{bmatrix} = A \begin{bmatrix} \bar{\mathbf{v}}_1^t \\ \vdots \\ \bar{\mathbf{v}}_N^t \end{bmatrix}. \quad (12)$$

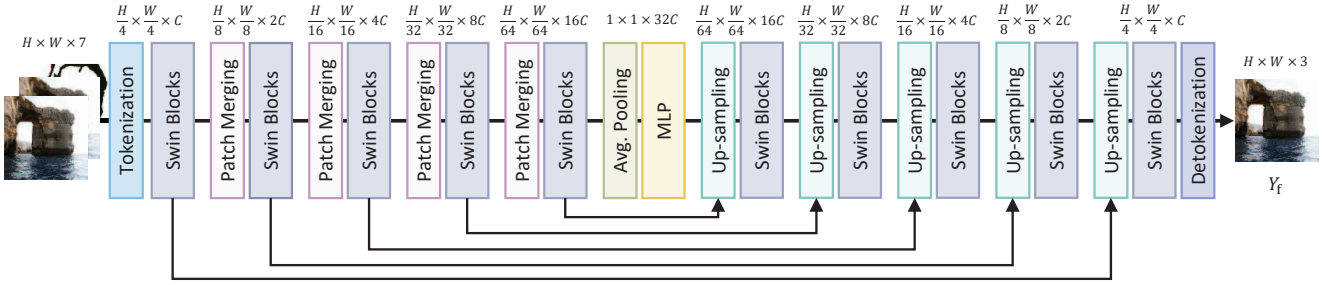
Thus, each updated mask $\mathbf{m}_i^*$ represents the amounts of errors in the attended token $\mathbf{z}_i^*$.

Overlapping tokens: Suppose that token $\mathbf{z}_i$ is totally erroneous, *i.e.*, $\mathbf{m}_i = \mathbf{0}$. Then, $\bar{\mathbf{q}}_i = \bar{\mathbf{k}}_i = \bar{\mathbf{v}}_i = \mathbf{0}$, and $\mu(\mathbf{q}_i, \mathbf{k}_j) = 0$ for all j. In such a case, the attended token $\mathbf{z}_i^*$ and its mask $\mathbf{m}_i^*$ are both $\mathbf{0}$, so $\mathbf{z}_i^*$ remains totally erroneous. To solve this problem, as well as to facilitate more communication among tokens, we introduce the notion of overlapping tokens.

As shown in Figure 2(b), the set of overlapping tokens consists of ordinary and shifted ones. We partition an image into ordinary tokens, as done in the existing vision transformers [5, 20]. In addition, by moving the token windows by half the width both horizontally and vertically, we extract shifted tokens. Thus, each shifted token overlaps with four ordinary tokens, and vice versa. Note that the notion of



(a) CMT



(b) Refinement Network

Figure 3: The proposed algorithm performs inpainting in two stages by employing CMT and the refinement network sequentially.

shifted tokens is unrelated to shifted windows in the Swin transformer [20].

By employing the center of each token as a node, we can draw a graph in Figure 2(c). Here, two nodes are connected by an edge if they overlap each other. In this example, red nodes depict totally erroneous tokens.

We propose a simple update rule for $\mathbf{z}_i^*$ when $\mathbf{z}_i$ is totally erroneous:

$$\mathbf{z}_i^* = \frac{1}{4} \sum_{\mathbf{z}_j^* \in \mathcal{N}} \mathbf{z}_j^* \tag{13}$$

where $\mathcal{N}$ is the set of the four overlapping neighbors of $\mathbf{z}_i^*$. We also update its mask by

$$\mathbf{m}_i^* = \frac{1}{4} \sum_{\mathbf{m}_j^* \in \mathcal{N}} \mathbf{m}_j^*. \tag{14}$$

In other words, a totally erroneous token is replaced with the self-attended results of its four neighbors.

Consequently, a totally erroneous token remains so, only if its four neighbors are also totally erroneous. Therefore, after the proposed masked self-attention, the set $\mathcal{E}$ of totally erroneous tokens is reduced. More specifically, the reduced set is given by $\mathcal{E} \ominus \mathcal{S}$, where $\ominus$ is the erosion operator [6], and $\mathcal{S}$ is the structuring element in Figure 2(d).

Two-stage inpainting: In Figure 3, the proposed algorithm consists of CMT and the refinement network; it performs inpainting in two stages, as done in [32, 33, 21, 28, 35, 16].

First, CMT in Figure 3(a) converts an input image into overlapping tokens $\mathbf{z}_1, \ldots, \mathbf{z}_N$ and initializes the corresponding masks $\mathbf{m}_1, \ldots, \mathbf{m}_N$. Then, it performs the masked self-attention and mask update (MSAU) 15 times. The output tokens $\mathbf{z}_1^*, \ldots, \mathbf{z}_N^*$ are then detokenized through two linear layers. As a result, the ordinary and shifted tokens generate inpainted images Y_o and Y_s, respectively.

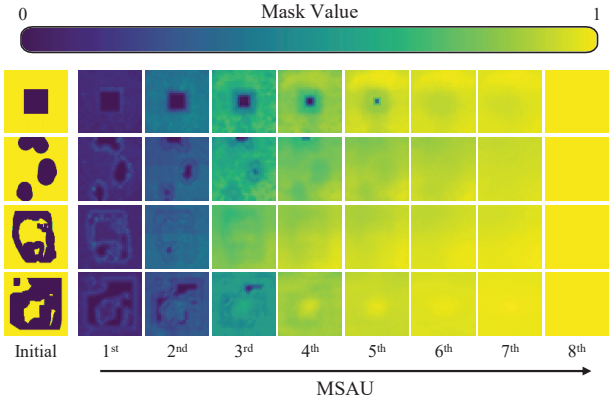


Figure 4: Illustration of the mask update processes for four images with holes. In these cases, all mask values are almost 1 after the eighth MSAU layer.

As mentioned previously, each MSAU layer erodes the set of totally erroneous tokens using the structuring element in Figure 2(d). Hence, roughly speaking, each MSAU layer reduces the width of a totally erroneous region by cP, $1 \leq c \leq \sqrt{2}$, where P is the patch size for the tokenization. In CMT, we set $P = 16$. Thus, composed of 15 MSAU layers, CMT can fill in a huge hole of width bigger than 200. Figure 4 illustrates how masks for four images with holes are updated by the proposed CMT network.

Second, the refinement network in Figure 3(b) improves the details in the initial inpainting results Y_o and Y_s, by employing smaller 4×4 patches for tokens. Specifically, we replace non-hole pixels Y_o and Y_s with their original values

Table 1: Quantitative comparison on the Places2 dataset [40] according to the hole-to-image (H2I) area ratios.

	H2I $\in (0.01, 0.1]$			H2I $\in (0.1, 0.2]$			H2I $\in (0.2, 0.3]$		
	PSNR(↑)	SSIM(↑)	FID(↓)	PSNR(↑)	SSIM(↑)	FID(↓)	PSNR(↑)	SSIM(↑)	FID(↓)
EdgeConnect [23]	33.80	0.9811	3.41	28.41	0.9524	7.98	25.29	0.9148	13.56
RN [34]	33.66	0.9804	5.59	28.63	0.9530	10.76	25.57	0.9174	16.44
MEDFE [18]	33.09	0.9784	4.62	27.34	0.9434	12.26	24.10	0.8977	22.45
HiFill [31]	30.34	0.9669	6.22	25.09	0.9189	14.65	22.26	0.8633	25.22
ICT [28]	31.66	0.9771	3.98	26.45	0.9446	9.29	23.23	0.9012	15.82
BAT [35]	34.54	0.9839	2.49	28.15	0.9557	6.36	24.47	0.9149	11.27
MAT [16]	34.43	0.9838	2.41	28.08	0.9549	6.19	24.62	0.9155	10.79
CMT (Proposed)	**35.43**	**0.9854**	**2.29**	**29.28**	**0.9596**	**5.93**	**25.88**	**0.9240**	**10.36**

	H2I $\in (0.3, 0.4]$			H2I $\in (0.4, 0.5]$			H2I $\in (0.5, 0.6]$		
	PSNR(↑)	SSIM(↑)	FID(↓)	PSNR(↑)	SSIM(↑)	FID(↓)	PSNR(↑)	SSIM(↑)	FID(↓)
EdgeConnect [23]	23.12	0.8744	18.90	21.33	0.8289	26.04	18.99	0.7605	36.83
RN [34]	23.32	0.8774	22.90	21.37	0.8283	32.01	18.69	0.7448	52.96
MEDFE [18]	21.85	0.8490	33.75	20.09	0.7954	47.78	17.84	0.7204	65.52
HiFill [31]	20.27	0.8064	38.26	18.52	0.7394	60.56	16.47	0.6530	93.44
ICT [28]	21.01	0.8547	22.90	19.20	0.8028	32.47	17.06	0.7309	47.40
BAT [35]	21.85	0.8688	17.39	19.78	0.8154	25.70	17.27	0.7362	40.34
MAT [16]	22.20	0.8721	15.31	20.26	0.8232	20.12	17.65	0.7472	27.53
CMT (Proposed)	**23.56**	**0.8850**	**14.69**	**21.70**	**0.8408**	**19.36**	**19.23**	**0.7723**	**27.29**

to yield Y_{o}' and Y_{s}'. Next, the refinement network takes the binary hole mask, as well as Y_{o}' and Y_{s}', as input. It adopts the encoder-decoder architecture of the Swin transformer [20], which has been successfully used in dense prediction tasks. Finally, it yields an inpainted result Y_{f}. Again, we replace non-hole pixels in Y_{f} with their original values.

Loss functions: CMT and the refinement network are trained sequentially. Both use the loss function

$$\mathcal{L} = w_{\mathrm{MR}}\mathcal{L}_{\mathrm{MR}} + w_{\mathrm{P}}\mathcal{L}_{\mathrm{P}} + w_{\mathrm{G}}\mathcal{L}_{\mathrm{G}} \tag{15}$$

where $\mathcal{L}_{\mathrm{MR}}$ is the mask-based reconstruction loss, $\mathcal{L}_{\mathrm{P}}$ is the perceptual loss, and $\mathcal{L}_{\mathrm{G}}$ is the adversarial loss. We set the weights $w_{\mathrm{MR}} = 1$, $w_{\mathrm{P}} = 10$, and $w_{\mathrm{G}} = 0.001$.

We use the mask-based reconstruction loss [17]

$$\mathcal{L}_{\mathrm{MR}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\|\mathbf{b}_i \otimes (\mathbf{y}_i - \hat{\mathbf{y}}_i)\|_1}{10\|\mathbf{b}_i\|_1} + \frac{\|(\mathbf{1} - \mathbf{b}_i) \otimes (\mathbf{y}_i - \hat{\mathbf{y}}_i)\|_1}{\|\mathbf{1} - \mathbf{b}_i\|_1} \tag{16}$$

where $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ are a predicted patch and its ground-truth, respectively. Note that, using the hole mask $\mathbf{b}_i$, prediction errors for hole pixels are weighted 10 times more than those for non-hole pixels.

The perceptual loss [13], given by

$$\mathcal{L}_{\mathrm{P}} = \|\mathbf{f}_Y - \mathbf{f}_{\hat{Y}}\|_2, \tag{17}$$

minimizes the distance between deep features $\mathbf{f}_Y$ and $\mathbf{f}_{\hat{Y}}$ extracted from a predicted image Y and its ground-truth $\hat{Y}$, respectively, through VGG16 [26].

The adversarial loss [7], $\mathcal{L}_{\mathrm{G}}$, is used to enhance the subjective quality of an inpainted image by emphasizing high-frequency information. We adopt the discriminator based on spectral normalization [22] to stabilize the training.

4. Experiments

We present the implementation details in the supplement.

4.1. Datasets and metrics

Places2 [40]: It contains 1.8 million training and 36,500 validation images, from which we randomly select 100,000 and 12,000 images for training and evaluation, respectively.

CelebA-HQ [14]: It has about 30,000 facial images, which we split into 26,000 training and 4,000 evaluation images.

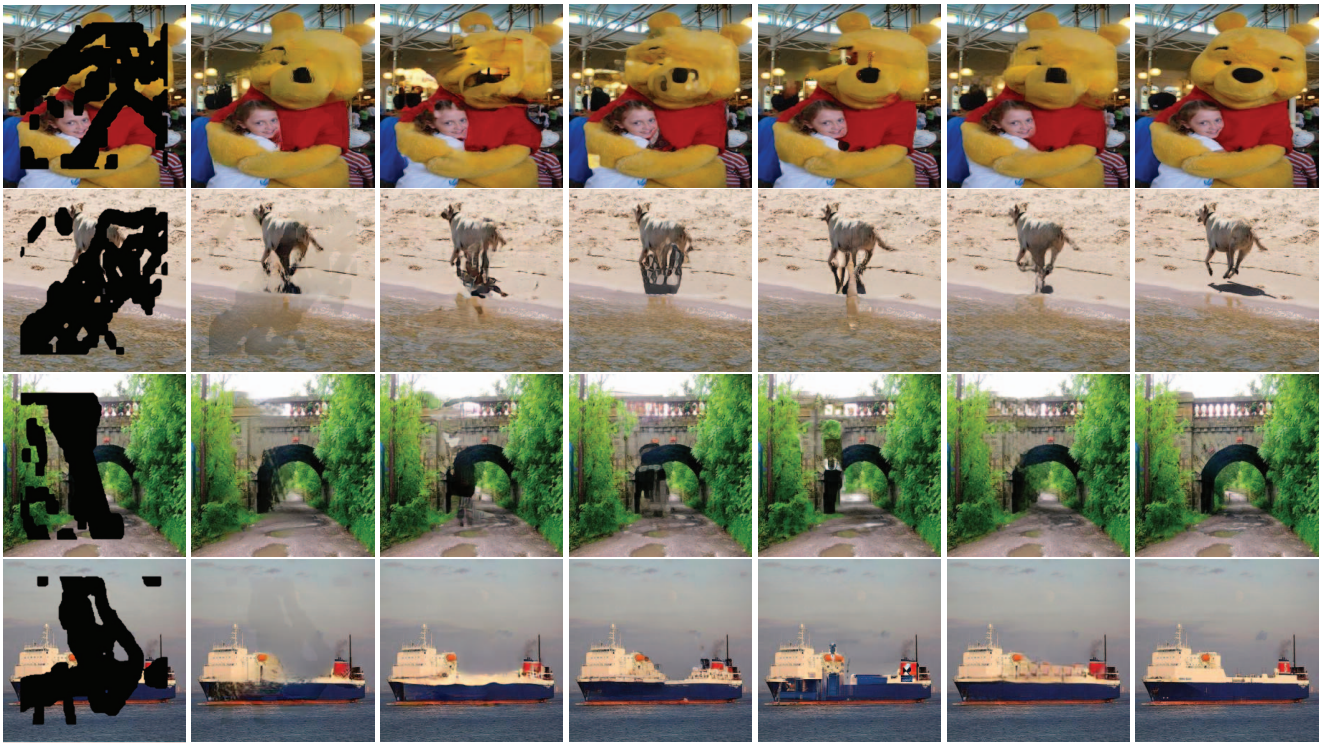
DTD [2]: It consists of 5,640 texture images. We select 2,000 images randomly and use them for evaluation only.

Irregular mask dataset [17]: It provides six mask sets, each of which contains 2,000 irregular hole patterns. They have patterns with H2I ratios within $(0.01, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.4]$, $(0.4, 0.5]$, and $(0.5, 0.6]$, respectively. We use these mask sets in all evaluations.

Evaluation metrics: For quantitative comparisons, we use the peak signal-to-noise ratio (PSNR), the structural similarity index measure (SSIM), and the Fréchet inception distance (FID) [10] measuring a perceptual similarity of a generated image to the real one.

4.2. Comparative assessment

The publicly available source codes of conventional algorithms are used in all comparative experiments.



(a) Input (b) RN [34] (c) ICT [28] (d) BAT [35] (e) MAT [16] (f) CMT (Proposed) (g) Ground-Truth

Figure 5: Qualitative comparison of inpainted images on the Places2 dataset [40].

Table 2: Quantitative comparison on the CelebA-HQ [14] dataset.

		PIC [39]	ICT [28]	BAT [35]	MAT [16]	CMT (Proposed)
(0.01, 0.2]	PSNR(↑)	33.67	33.27	34.63	35.31	**35.92**
	SSIM(↑)	0.9783	0.9793	0.9830	0.9842	**0.9859**
	FID(↓)	2.34	1.87	1.06	0.90	**0.84**
(0.2, 0.4]	PSNR(↑)	26.48	26.40	26.91	27.67	**28.24**
	SSIM(↑)	0.9342	0.9389	0.9440	0.9461	**0.9515**
	FID(↓)	6.43	5.61	3.75	2.55	**2.54**
(0.4, 0.6]	PSNR(↑)	21.58	21.84	22.26	23.22	**23.78**
	SSIM(↑)	0.8650	0.8765	0.8831	0.8884	**0.8997**
	FID(↓)	14.22	12.42	7.30	**4.60**	5.23

Table 3: Quantitative comparison on the DTD dataset [2].

	PSNR(↑)	SSIM(↑)	FID(↓)
EdgeConnect [23]	22.86	0.8642	30.54
RN [34]	21.06	0.8379	70.39
MEDFE [18]	21.77	0.8413	60.24
HiFill [31]	20.87	0.8163	48.34
ICT [28]	21.65	0.8533	33.29
BAT [35]	22.62	0.8706	19.53
MAT [16]	22.57	0.8665	21.89
CMT (Proposed)	**24.05**	**0.8798**	**19.31**

Comparison on Places2: In Table 1, we compare the PSNR, SSIM, and FID scores of the proposed CMT algorithm with those of recent inpainting algorithms: EdgeConnect [23], RN [34], MEDFE [18], HiFill [31], ICT [28], BAT [35], and MAT [16]. Note that ICT, BAT, and MAT are transformer-based. In this test, the inpainting results are obtained on 256×256 images, except for HiFill and MAT trained on 512×512 images. For comparison, we downsample the inpainted images of HiFill and MAT by a factor of 2. The proposed CMT performs the best in all tests for all H2I ratio ranges with no exception. Especially, CMT outperforms the state-of-the-art mask-aware transformer MAT by significant margins in terms of PSNR and SSIM. The comparison on 512×512 inpainted images is also available in the supplement.

Figure 5 shows qualitative results, in which RN suffers from color discrepancies, ICT and BAT yield undesirable artifacts, and MAT generates unnatural details. In contrast, CMT provides more plausible results with fewer artifacts.

Comparison on CelebA-HQ: Table 2 compares the proposed CMT algorithm with PIC [39], ICT [28], BAT [35], and MAT [16]. Here, all networks are trained for the spatial resolution of 256×256. Except for FID on the high H2I range of $(0.4, 0.6]$, CMT surpasses the conventional algorithms. As shown in Figure 6, CMT generates less severe artifacts and restores facial details more successfully, espe-

(a) Input (b) PIC [39] (c) ICT [28] (d) BAT [35] (e) MAT [16] (f) CMT (Proposed) (g) Ground-truth

Figure 6: Qualitative comparison of inpainted images on the CelebA-HQ dataset [14].

(a) Input (b) RN [34] (c) ICT [28] (d) BAT [35] (e) MAT [16] (f) CMT (Proposed) (g) Ground-truth

Figure 7: Qualitative comparison of inpainted images on the DTD dataset [2].

cially around the eyes.

Comparison on DTD: Next, we fill in holes in texture images in the DTD dataset, by employing the inpainting networks trained on Place2. Holes with H2I ratios in $(0.3, 0.4]$ are used. In Table 3, CMT again outperforms the existing algorithms in every metric. In Figure 7, CMT restores repetitive patterns successfully.

Figure 8 compares CMT with MAT [16] when a texture image is degraded by different holes. We see that the proposed CMT using a continuous mask restores the holes more faithfully than MAT using a binary mask.

4.3. Ablation and analysis

Table 4 compares several ablated methods to analyze the efficacy of the proposed algorithm and its components. In this test, the Places2 dataset [40] is used with the H2I ratios in (0.3,0.4].

Overlapping tokens: Method I employs ordinary tokens only as in [5], while II uses overlapping tokens. Both I and II perform the standard self-attention in [5] with no mask update scheme. By comparing I and II, we see that overlapping tokens improve inpainting performance by facilitating

(a) Input (b) MAT [16] (c) CMT

Figure 8: Comparison on DTD [2].

(a) Input (b) Scribbling (c) Output (d) Input (e) Scribbling (f) Output

Figure 9: Examples of old photo restoration and distraction removal.

Table 4: Ablation studies of the proposed CMT algorithm.

Method	Tokens	MSAU	Refine	PSNR	SSIM	FID
I	Ordinary			22.46	0.8534	22.39
II	Overlapping			22.64	0.8592	19.83
III	Ordinary	(Binary)		22.51	0.8547	21.71
IV	Ordinary	✓		22.83	0.8612	19.62
V	Overlapping	(Binary)		22.95	0.8618	18.83
VI	Overlapping	✓		23.22	0.8771	15.69
VII	Ordinary	✓	✓	23.12	0.8744	17.79
VIII	Overlapping	✓	✓	**23.56**	**0.8850**	**14.69**

more communication among tokens.

MSAU & mask type: In methods III and IV, the MSAU layers are employed with ordinary tokens. Here, a totally erroneous token is replaced with the self-attended results of its four neighbors: the top, bottom, left, and right tokens. Methods III and V replace the continuous masking in the MSAU layers with the binary masking; a token is declared valid when it is conditioned on at least one valid input token, as done in [16]. The performance gaps between I and IV or between II and VI indicate that the proposed masked self-attention is effective for image inpainting. Also, from comparisons between III and V or between IV and VI, we see that overlapping tokens make the masked self-attention more effective. Furthermore, the gaps between III and IV or between V and VI confirm the efficacy of the proposed continuous masking in comparison with the binary masking.

Refinement: Finally, the refinement network in methods VI and VII enhance the initial results of CMT. Note that the proposed CMT algorithm (method VII) provides the best PSNR, SSIM, and FID.

4.4. Real applications

We apply the proposed CMT algorithm to restore old photos with scratches and enhance flawed photos with distracting objects in Figure 9 (a) and (d), respectively. We draw scribbles to cover the defects, as in Figure 9 (b) and (e). Then, we employ the network trained on Places2 [40] to fill in the defects, as in Figure 9 (c) and (f). The proposed algorithm removes the scratches and the distracting objects successfully to reconstruct faithful images.

5. Conclusions

We proposed the novel continuous-mask-aware transformer, referred to as CMT, to exploit a continuous mask representing the amounts of errors for image inpainting. First, a continuous mask is initialized and used during the self-attention. Here, overlapping tokens are employed to facilitate the masked self-attention. Next, the mask is updated by modeling the error propagation during the masked self-attention. Through several MSAU layers, initial inpainting results are obtained. Then, the initial results are refined to yield a final inpainted image. Extensive experiments showed that the proposed CMT algorithm provides significantly better inpainting results than existing algorithms.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. NRF-2021R1A4A1031864 and No. NRF-2022R1A2B5B03002310)

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. In *NIPS Deep Learning Symposium*, 2016. 2

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 2, 5, 6, 7, 8

[3] Antonio Criminisi, Patrick Perez, and Kentaro Toyama. Object removal by exemplar-based inpainting. In *CVPR*, 2003. 1

[4] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *CVPR*, 2022. 1

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 1, 2, 3, 7

[6] Rafael C Gonzalez and Richard E. Woods. *Digital image processing*. Pearson, 2018. 4

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial Nets. In *NIPS*, 2014. 2, 5

[8] Qing Guo, Xiaoguang Li, Felix Juefei-Xu, Hongkai Yu, Yang Liu, and Song Wang. JPGNet: Joint predictive filtering and generative network for image inpainting. In *ACM Int. Conf. Multimedia*, 2021. 2

[9] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Trans. Graph.*, 26(3):1–7, 2007. 1

[10] 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 *NIPS*, 2017. 5

[11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2

[12] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Trans. Graph.*, 36(4):1–14, 2017. 1, 2

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2017. 2, 5, 6, 7

[15] Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics. In *ICCV*, 2003. 1

[16] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. MAT: Mask-aware transformer for large hole image inpainting. In *CVPR*, 2022. 1, 2, 4, 5, 6, 7, 8

[17] 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 *ECCV*, 2018. 1, 2, 5

[18] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *ECCV*, 2020. 2, 5, 6

[19] Qiankun Liu, Zhentao Tan, Dongdong Chen, Qi Chu, Xiyang Dai, Yinpeng Chen, Mengchen Liu, Lu Yuan, and Nenghai Yu. Reduce information loss in transformers for pluralistic image inpainting. In *CVPR*, 2022. 1

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1, 3, 4, 5

[21] Yuqing Ma, Xianglong Liu, Shihao Bai, Lei Wang, Dailan He, and Aishan Liu. Coarse-to-fine image inpainting via region-wise convolutions and non-local correlation. In *IJCAI*, 2019. 1, 2, 4

[22] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 2, 5

[23] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. EdgeConnect: Structure guided image inpainting using edge prediction. In *ICCVW.*, 2019. 2, 5, 6

[24] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context Encoders: Feature learning by inpainting. In *CVPR*, 2016. 1, 2

[25] Stefan Roth and Michael J Black. Fields of experts: A framework for learning image priors. In *CVPR*, 2005. 1

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2014. 5

[27] 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. In *WACV*, 2022. 2

[28] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *ICCV*, 2021. 1, 2, 4, 5, 6, 7

[29] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. Image inpainting with learnable bidirectional attention maps. In *ICCV*, 2019. 2

[30] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *CVPR*, 2017. 1, 2

[31] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *CVPR*, 2020. 1, 2, 5, 6

[32] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 1, 2, 4

[33] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 1, 2, 4

[34] Tao Yu, Zongyu Guo, Xin Jin, Shilin Wu, Zhibo Chen, Weiping Li, Zhizheng Zhang, and Sen Liu. Region normalization for image inpainting. In *AAAI*, 2020. 2, 5, 6, 7

[35] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. In *ACM Int. Conf. Multimedia*, 2021. 1, 2, 4, 5, 6, 7

[36] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *CVPR*, 2019. 1, 2

[37] 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 *ECCV*, 2020. 2

[38] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *ICLR*, 2021. 2

[39] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *CVPR*, 2019. 2, 6, 7

[40] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2017. 2, 5, 6, 7, 8