Improving Cross-modal Alignment for Text-Guided Image Inpainting

Yucheng Zhou, Guodong Long[*]

Australian AI Institute, School of Computer Science, FEIT, University of Technology Sydney
yucheng.zhou-1@student.uts.edu.au, guodong.long@uts.edu.au

Abstract

Text-guided image inpainting (TGII) aims to restore missing regions based on a given text in a damaged image. Existing methods are based on a strong vision encoder and a cross-modal fusion model to integrate cross-modal features. However, these methods allocate most of the computation to visual encoding, while light computation on modeling modality interactions. Moreover, they take cross-modal fusion for depth features, which ignores a fine-grained alignment between text and image. Recently, vision-language pre-trained models (VLPM), encapsulating rich cross-modal alignment knowledge, have advanced in most multimodal tasks. In this work, we propose a novel model for TGII by improving cross-modal alignment (CMA). CMA model consists of a VLPM as a vision-language encoder, an image generator and global-local discriminators. To explore cross-modal alignment knowledge for image restoration, we introduce cross-modal alignment distillation and in-sample distribution distillation. In addition, we employ adversarial training to enhance the model to fill the missing region in complicated structures effectively. Experiments are conducted on two popular vision-language datasets. Results show that our model achieves state-of-the-art performance compared with other strong competitors.

1 Introduction

Text-guided image inpainting (TGII), involving computer vision (CV) and natural language processing (NLP), aims to restore visual content for a missing area in a damaged image based on a given text (Zhang et al., 2020a). With the development of CV and NLP, it plays an essential role in many real-world applications, such as image editing (Zhu et al., 2020), damaged image restoration (Liu et al., 2019), and image rendering (Kirillov et al., 2020). Therefore, it has become one of the most crucial areas in CV and NLP.
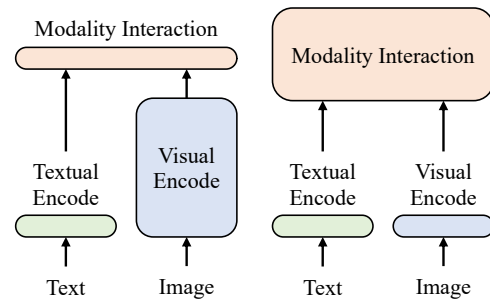

Figure 1: Different categories of vision-and-language models. *left*: most of the computation on visual encoding; *right*: most of the computation on modeling modality interactions.

Existing methods (Zhang et al., 2020b; Lin et al., 2020; Wu et al., 2021) adopt an encoder-decoder framework as their backbone with a vision-language fusion module to introduce textual information. These methods use separate encoders for images and texts, heavier on the former. Then, the vision-language fusion module is used to integrate the features from the two modalities through a simple similarity calculation of features, or shallow attention layers (Vaswani et al., 2017), as shown on the left of Figure 1. Most of the computation of these methods on visual encoding, while textual features only serve as a complement to deep visual features, which ignores the importance of deep interaction of multimodal information. Moreover, these methods share a common drawback: they do not perform well for natural image datasets with a wide variety of objects (e.g., MSCOCO (Lin et al., 2014)). The reason is that these methods lack fine-grained alignment knowledge of texts and images to guide the fusion of cross-modal information in multimodal interactions. Besides, their fusion modules lack powerful cross-modal reasoning capabilities.

Recently, providing the success of pre-trained language/vision Transformer (Devlin et al., 2019; Dosovitskiy et al., 2021; Zhou et al., 2022a),

---

[*]Corresponding author.

some works (e.g., ViLT (Kim et al., 2021) and SimVLM (Wang et al., 2021)) pre-train a vision-and-language Transformer on a large-scale image-text dataset. These Vision-and-Language Pre-trained Models (VLPM) achieve exciting performance on many multimodal downstream tasks. The reason is that VLPM encapsulates rich image and text alignment knowledge and has a strong cross-modal reasoning capability (Chen et al., 2020; Kim et al., 2021). To enhance the cross-modal interaction between images and texts, Kim et al. (2021) propose ViLT, which focuses most of the computation on modeling the multimodal interaction, as shown on the right side of Figure 1.

Motivated by Kim et al. (2021), we propose a novel model enhanced by cross-modal alignment (CMA) for text-guided image inpainting, comprising a vision-and-language encoder, an image generator and global-local discriminators. Different from previous works (Zhang et al., 2020b; Lin et al., 2020; Wu et al., 2021), we employ a vision-and-language encoder based on VLPM to encode images and texts instead of separate vision and language encoders. The vision-and-language encoder can encode images and texts in a cross-modal interaction manner to implement visual priors reconstruction. Then, the visual features integrating textual information (i.e., reconstructed visual priors) obtained by VLPM are passed to an image generator to generate a restored image. To improve cross-modal alignment for image inpainting, we introduce cross-modal alignment distillation to guide a fine-grained fusion of cross-modal knowledge. Moreover, to further strengthen the visual priors reconstruction, we utilize in-sample distillation to enhance the model's cross-modal reasoning capability for the content of missing regions. Besides, we employ adversarial learning to improve the quality of generated images through global-local discriminators.

Experiments are conducted on two popular image-text datasets with a wide variety of objects (i.e., MSCOCO (Lin et al., 2014) and Flicker30K (Plummer et al., 2017)). Experimental results demonstrate that our method achieves state-of-the-art performance compared to other strong competitors. In addition, we analyze the effectiveness of each module of our method and the impact of cross-modal alignment. Moreover, we also conduct extensive analysis to verify the effectiveness of our method.

2 Method

In this section, we will introduce our CMA model as shown in Figure 2. CMA model consists of a vision-and-language encoder, an image generator and global-local discriminators. In addition, details about training and inference are elaborated.

2.1 Vision-and-Language Encoding

In previous works (Zhang et al., 2020b; Lin et al., 2020; Wu et al., 2021), images and texts are encoded by separate encoders. Specifically, a CNN-based image encoder is used to extract visual features for images, while a RNN-based text encoder is used to encode text to obtain textual features. Next, the image and text features are integrated by a multimodal fusion module to obtain multimodal representations (i.e., reconstructed visual priors). In this work, we employ a novel Transformer-based cross-modal encoder for image and text encoding, a vision-and-language encoder instead of two separate encoders. For language encoding, we employ a word embedding matrix to embed text $T \in \mathbb{R}^{L \times |V|}$ into $\bar{T} \in \mathbb{R}^{L \times e}$, where L, $|V|$, and e denote the length of text, size of vocabulary and size of embedding, respectively. For visual encoding, an image $I \in \mathbb{R}^{C \times H \times W}$ is sliced into patches and flattened to $V \in \mathbb{R}^{N \times (P^2 \times C)}$, where P is the size of patch and N equal to $(H \times W)/P^2$. Following Kim et al. (2021), V is embedded into $\bar{V} \in \mathbb{R}^{N \times e}$:

$$\bar{V} = \text{Linear-Projection}(V) \qquad (1)$$

where the details of linear projection can be found in (Kim et al., 2021). Next, images and texts are encoded in a cross-modal interaction manner:

$$[\hat{T}; \hat{V}] = \text{Trans-Enc}([\bar{T}; \bar{V}]) \qquad (2)$$

where $[;]$ denote a concatenation operation, and outputs of the vision-and-language encoder can be represented as $\hat{T} \in \mathbb{R}^{L \times e}$ for textual representations and $\hat{V} \in \mathbb{R}^{N \times e}$ for reconstructed visual priors. Compared to CNN constrained by its inherent properties (e.g., spatial-invariant kernels), which are not conducive to understanding global features (Wan et al., 2021), Transformer-based encoders have a natural advantage in encoding global features across modalities.

2.2 Image Generation

The process of image generation includes two stages. The first stage is to downsample the visual priors to extract deep visual representations.
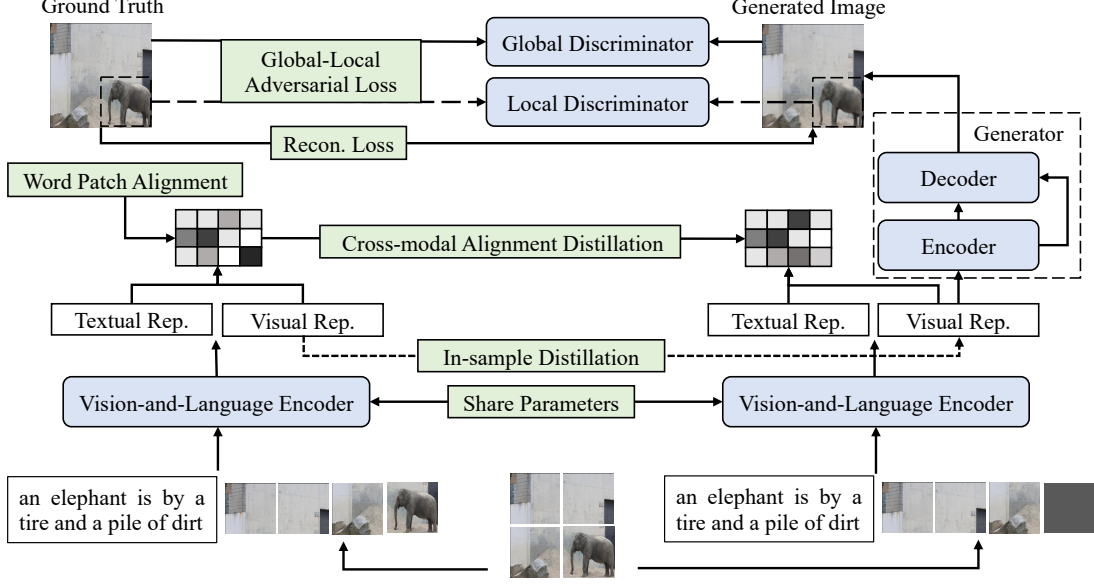
Figure 2: Overview of our model. Blue rounded rectangles denote trainable modules. Green right rectangles indicate training objectives or operations.

Next, performing upsampling for deep visual representations to generate a restored image. Different from the previous works (Zhang et al., 2020b; Lin et al., 2020; Wu et al., 2021), we do not integrate a fusion module to introduce textual features in the image generation process because the texts are introduced into the vision-and-language encoders. Although transformers demonstrate their effectiveness in long-term relations and the advantage of understanding global features, their computational complexity is quadratic with the input length, which hinders image generation (Wan et al., 2021). Therefore, our image generator consists of two CNN-based components, an encoder for downsampling and a decoder for upsampling. Similar to (Zhang et al., 2020a), we employ a 5-layer ResNet (He et al., 2016) as the downsampling encoder, and the visual priors are fed into it to obtain deep visual representations:

$$\boldsymbol{v} = \text{Down-Sampling}(\hat{\boldsymbol{V}}) \quad (3)$$

In addition, we pass $\boldsymbol{v}$ into the upsampling decoder to perform image reconstruction to generate a restored image, and the upsampling decoder consists of a 5-layer residual generator network.

$$\boldsymbol{I}_r = \text{Up-Sampling}(\boldsymbol{v}, \hat{\boldsymbol{V}}) \quad (4)$$

The visual priors obtained by the vision-and-language encoder are input to the upsampling decoder through a skip connection to provide detailed information that forgets in the downsampling stage.

2.3 Training

For model training, we construct five training objectives, including cross-modal alignment distillation, in-sample distillation, word patch alignment, reconstruction loss and global-local adversarial loss.

Cross-Modal Alignment Distillation. To guide image inpainting through cross-modal alignment knowledge, we pass the original image and text to the vision-and-language encoder to obtain a text representation $\hat{T}_o$ and a visual priors $\hat{V}_o$. Then, we obtain the correlation map M_o between each token of text and each image patch by calculating the similarity between them. The correlation map corresponding to the corrupted image and text is denoted as M_r. Next, we compute the pair-wise similarity distillation loss between the two correlation maps:

$$\ell_{\text{CMAD}} = \frac{1}{N \times L} \sum_{i=1}^{L} \sum_{j=1}^{N} (a_{ij}^r - a_{ij}^o)^2 \quad (5)$$

$$a_{ij}^r = \frac{\boldsymbol{t}_i^{r\top} \boldsymbol{v}_j^r}{\|\boldsymbol{t}_i^r\|_2 \|\boldsymbol{v}_j^r\|_2}, a_{ij}^r \in \boldsymbol{M}_r \quad (6)$$

$$a_{ij}^o = \frac{\boldsymbol{t}_i^{o\top} \boldsymbol{v}_j^o}{\|\boldsymbol{t}_i^o\|_2 \|\boldsymbol{v}_j^o\|_2}, a_{ij}^o \in \boldsymbol{M}_o \quad (7)$$

where $\boldsymbol{t}_i^r \in \hat{\boldsymbol{T}}$, $\boldsymbol{v}_j^r \in \hat{\boldsymbol{V}}$, $\boldsymbol{t}_i^o \in \hat{\boldsymbol{T}}_o$ and $\boldsymbol{v}_j^o \in \hat{\boldsymbol{V}}_o$, respectively.

In-Sample Distillation. Besides using cross-modal alignment knowledge to guide image inpaint-

ing, we propose in-sample distillation. The purpose of this objective is to guide the damaged image to infer visual priors closer to that from the original image based on the text and known regions, i.e.,

$$\ell_{\text{ISD}} = \frac{1}{N} \sum_{i=1}^{N} \text{KL} \left(\boldsymbol{v}_i^o \| \boldsymbol{v}_i^r \right). \qquad (8)$$

where KL denotes Kullback–Leibler divergence.

Word Patch Alignment. To maintain that the cross-modal alignment knowledge of the model is not degraded, we use the word patch alignment objective to preserve the cross-modal alignment knowledge integrated into the model.

$$\ell_{\text{WPA}} = \min_{\mathbf{T} \in \Pi(\mathbf{a}, \mathbf{b})} \sum_{i=1}^{N} \sum_{j=1}^{L} \mathbf{T}_{ij} \cdot c \left(\boldsymbol{v}_i^o, \boldsymbol{t}_j^o \right) \quad (9)$$

$$c(\boldsymbol{v}_i^o, \boldsymbol{t}_j^o) = 1 - \frac{\boldsymbol{v}_i^{o\top} \boldsymbol{t}_j^o}{\|\boldsymbol{v}_i^o\|_2 \|\boldsymbol{t}_j^o\|_2} \qquad (10)$$

$$\Pi(\mathbf{a}, \mathbf{b}) = \{\mathbf{T} \in \mathbb{R}^{N \times L} \mid \mathbf{T} \mathbf{1}_m = \mathbf{a},$$
$$\mathbf{T}^\top \mathbf{1}_n = \mathbf{b}\} \qquad (11)$$

where $\mathbf{1}_n$ denotes an n-dimensional all-one vector; $\mathbf{T}$ is a transport plan. The details can be found in (Kim et al., 2021).

Reconstruction Loss. We adopt the ℓ_1-norm error as the loss function between the restored image and its corresponding ground-truth image $\boldsymbol{I}_{gt}$:

$$\ell_1 = \|\boldsymbol{I}_r - \boldsymbol{I}_{gt}\|_1, \qquad (12)$$

Global Adversarial Loss. In image generation, the adversarial loss (Goodfellow et al., 2014) effectively improves the quality of the generated image. However, adversarial training is unstable since keeping the balance between the generator and discriminator is difficult. To tackle this problem, Arjovsky et al. (2017) propose the Wasserstein generative adversarial nets (WGAN) that can improve the stability of adversarial learning. We employ a WGAN hinge loss as the adversarial loss, and it can be formulated as follows:

$$\ell_{\text{G-adv,G}} = \mathbb{E}_{\hat{\boldsymbol{x}} \sim P_{data}(\boldsymbol{I}_r)}[-D(\hat{\boldsymbol{x}})] \qquad (13)$$

$$\ell_{\text{G-adv,D}} = \mathbb{E}_{\boldsymbol{x} \sim P_{data}(\boldsymbol{I}_{gt})}[\text{ReLU}(1 - D(\boldsymbol{x}))]$$
$$+ \mathbb{E}_{\hat{\boldsymbol{x}} \sim P_{data}(\boldsymbol{I}_r)}[\text{ReLU}(1 + D(\hat{\boldsymbol{x}}))] \qquad (14)$$

where $D(\cdot)$ denotes a global discriminator; ReLU is the rectified linear unit function; $\mathcal{L}_{adv,G}$ and

$\mathcal{L}_{adv,D}$ are loss function for inpainting model and discriminator, respectively.

The discriminator $D(\cdot)$ consists of 5-layer ResNet, followed by a fully connected layer. Each convolutional layer in ResNet applies spectral normalization to satisfy the Lipschitz constraints of Wasserstein GANs.

Local Adversarial Loss. For the local discriminator, we use the same settings as the global discriminator, and the loss is defined as:

$$\ell_{\text{L-adv,G}} = \mathbb{E}_{\hat{\boldsymbol{x}}_l \sim P_{data}(\boldsymbol{I}_r)}[-D_l(\hat{\boldsymbol{x}}_l)] \qquad (15)$$

$$\ell_{\text{L-adv,D}} = \mathbb{E}_{\boldsymbol{x}_l \sim P_{data}(\boldsymbol{I}_{gt})}[\text{ReLU}(1 - D_l(\boldsymbol{x}_l))]$$
$$+ \mathbb{E}_{\hat{\boldsymbol{x}}_l \sim P_{data}(\boldsymbol{I}_r)}[\text{ReLU}(1 + D_l(\hat{\boldsymbol{x}}_l))] \qquad (16)$$

where x_l denotes the restored image or ground truth corresponding to the missing region; $D_l(\cdot)$ denotes a local discriminator.

Considering the loss functions above, the objective function of our model in the generation stage can be defined as:

$$\ell_G = \lambda \ell_{\text{CMAD}} + \lambda \ell_{\text{ISD}} + \alpha \ell_{\text{WPA}} + \beta \ell_1$$
$$+ \gamma \ell_{\text{G-adv,G}} + \gamma \ell_{\text{L-adv,G}} \qquad (17)$$

where the λ, α, β and γ are the hyper-parameters used to balance the objective function.

In addition, the objective function of our model in the discriminative stage can be defined as:

$$\ell_D = \gamma \ell_{\text{G-adv,D}} + \gamma \ell_{\text{L-adv,D}} \qquad (18)$$

2.4 Inference

During inference, we first pass image $\boldsymbol{I}$ and text $\boldsymbol{T}$ into the cross-modal encoder (i.e., Equ.1 and Equ.2) to obtain the visual representation $\hat{\boldsymbol{V}}$. Next, we deliver the visual representation $\hat{\boldsymbol{V}}$ into the generator (i.e., Equ.3 and Equ.4) to generate a restored image $\boldsymbol{I}_r$.

3 Experiments

3.1 Dataset and Evaluation Metrics

We conduct the experiment on two datasets: MSCOCO (Lin et al., 2014) and Flicker30K (Plummer et al., 2017). For the MSCOCO and Flicker30K datasets, we split them following their original training, validation and test set. Following Zhang et al. (2020a), we also set the mask for an image in two types: center mask and object mask. A center mask refers to a square mask taking 50%

Method	ℓ_1 (%) ↓	FID ↓	KID ↓	TV loss (%) ↓	PSNR ↑	SSIM (%) ↑
CSA (Liu et al., 2019)	5.14	50.88	2.85	4.32	19.87	82.53
PICNet (Zheng et al., 2019)	5.63	53.57	3.19	4.55	19.58	81.87
CTSDG (Guo et al., 2021)	5.03	48.95	2.56	4.31	19.98	82.69
MMFL (Lin et al., 2020)	4.36	44.33	2.22	4.28	20.73	82.92
TGII (Zhang et al., 2020b)	4.22	44.02	1.87	4.39	20.87	83.07
TDANet (Zhang et al., 2020a)	4.13	42.38	1.67	4.59	20.91	83.34
ALMR (Wu et al., 2021)	4.17	43.43	1.79	4.27	20.81	83.15
CMA (ours)	**3.78**	**39.52**	**1.34**	**4.17**	**22.07**	**85.18**

Table 1: Results on MSCOCO with the center mask.

Method	ℓ_1 (%) ↓	FID ↓	KID ↓	TV loss (%) ↓	PSNR ↑	SSIM (%) ↑
CSA (Liu et al., 2019)	8.79	53.49	3.65	4.85	18.99	75.50
PICNet (Zheng et al., 2019)	9.15	56.80	3.72	5.09	18.78	74.98
CTSDG (Guo et al., 2021)	8.69	51.64	3.36	4.82	19.13	75.84
MMFL (Lin et al., 2020)	7.59	47.15	2.91	4.53	20.07	76.16
TGII (Zhang et al., 2020b)	7.53	46.73	2.82	4.59	20.27	76.46
TDANet (Zhang et al., 2020a)	7.48	45.30	2.45	4.70	20.55	76.93
ALMR (Wu et al., 2021)	7.54	46.01	2.61	4.46	20.35	76.50
CMA (ours)	**7.00**	**42.23**	**2.01**	**4.30**	**21.75**	**78.67**

Table 2: Results on MSCOCO with the object mask.

Method	ℓ_1 (%) ↓	FID ↓	KID ↓	TV loss (%) ↓	PSNR ↑	SSIM (%) ↑
CSA (Liu et al., 2019)	4.99	50.76	2.78	4.14	19.93	83.97
PICNet (Zheng et al., 2019)	5.29	52.25	3.14	4.39	19.70	82.21
CTSDG (Guo et al., 2021)	4.78	48.66	2.54	4.10	20.39	84.21
MMFL (Lin et al., 2020)	4.13	43.92	2.12	4.26	20.78	83.96
TGII (Zhang et al., 2020b)	4.11	43.34	1.73	4.29	21.48	84.49
TDANet (Zhang et al., 2020a)	3.92	41.46	1.54	4.16	21.36	84.17
ALMR (Wu et al., 2021)	4.05	42.43	1.60	4.12	21.22	83.42
CMA (ours)	**3.61**	**38.30**	**1.29**	**4.00**	**22.55**	**86.33**

Table 3: Results on Flickr30K with the center mask.

area in the center of an image. An object mask indicates masking an image based on the object boxes provided by every image. To evaluate the performance of our model and other methods on these datasets, we utilize the ℓ_1 loss, Frechet Inception Distance (FID) (Heusel et al., 2017), Kernel Inception Distance (KID) (Heusel et al., 2017), total variation (TV) (Rudin et al., 1992), peak signal-to-noise ratio (PSNR) (Fardo et al., 2016) and structural similarity index (SSIM) (Wang et al., 2004) as metrics to report the results. ℓ_1 measure the ℓ_1 distance between restored and original images. FID and KID measure the quality of restored images based on human perception. PSNR and SSIM measure structural similarity between restored and original images and the ℓ_2 distance.

3.2 Experimental Settings

For training images in MSCOCO and Flicker30K, we resize them to make their minimal height/width 256 and crop them based on size 256×256 at the center. During training, we set the λ as 2, α as 1, β as 1 and γ as 0.1 in the objective function, and the model is trained by an Adam optimizer (Kingma and Ba, 2015) with the learning rate of 1×10^{-4}. In the vision-and-language encoder, the patch size, intermediate size and hidden size are 32, 3072 and 768. For a masked patch, we utilize a special token [Vmask] as input. We employ ViLT (Kim et al., 2021) to initialize our vision-and-language encoder. The weight decay and gradient clipping are set to 0.01 and 1.0. The maximum training epoch and batch size are 200 and 128. Warmup steps and maximum sequence length are set to 2000 and 40. Our experiments are conducted on $2 \times$ V100 GPUs. We choose models with the best result on the validation set and report the results on the test set based on the models.

3.3 Main Results

The experimental results of our method and previous works on MSCOCO and Flicker30K are shown

Method	ℓ_1 (%) ↓	FID ↓	KID ↓	TV loss (%) ↓	PSNR ↑	SSIM (%) ↑
CSA (Liu et al., 2019)	8.60	53.18	3.57	4.75	19.21	75.82
PICNet (Zheng et al., 2019)	9.01	56.59	3.64	4.99	19.16	75.95
CTSDG (Guo et al., 2021)	8.45	51.35	3.30	4.69	19.22	76.73
MMFL (Lin et al., 2020)	7.58	46.53	2.86	4.53	20.34	76.42
TGII (Zhang et al., 2020b)	7.39	46.52	2.73	4.56	20.82	76.82
TDANet (Zhang et al., 2020a)	7.37	44.72	2.42	4.55	21.04	77.43
ALMR (Wu et al., 2021)	7.40	45.74	2.52	4.36	20.56	76.57
CMA (ours)	**6.86**	**41.28**	**1.91**	**4.25**	**22.07**	**79.93**

Table 4: Results on Flickr30K with the object mask.

Method	ℓ_1 (%) ↓	FID ↓	KID ↓	TV loss (%) ↓	PSNR ↑	SSIM (%) ↑
CSA (Liu et al., 2019)	4.04	56.88	2.74	3.70	20.03	81.12
PICNet (Zheng et al., 2019)	3.78	47.33	2.46	3.74	20.16	82.04
CTSDG (Guo et al., 2021)	3.63	38.05	1.98	3.71	20.73	82.64
MMFL (Lin et al., 2020)	3.42	29.79	1.39	3.59	20.68	82.36
TGII (Zhang et al., 2020b)	3.54	32.57	1.51	3.63	20.55	81.89
TDANet (Zhang et al., 2020a)	3.57	30.82	1.49	3.61	20.79	82.68
ALMR (Wu et al., 2021)	3.32	15.78	0.52	3.52	20.66	80.61
CMA (ours)	**3.07**	**14.21**	**0.41**	**3.51**	**20.84**	**82.68**

Table 5: Results on CUB with the center mask.

Method	ℓ_1 (%) ↓	FID ↓	KID ↓	TV loss (%) ↓	PSNR ↑	SSIM (%) ↑
CSA (Liu et al., 2019)	6.76	59.91	3.00	4.21	19.03	70.97
PICNet (Zheng et al., 2019)	6.42	50.16	2.82	4.32	18.96	70.27
CTSDG (Guo et al., 2021)	5.83	40.19	2.23	4.13	19.38	72.16
MMFL (Lin et al., 2020)	4.63	32.00	2.13	3.79	20.32	78.24
TGII (Zhang et al., 2020b)	4.72	34.39	2.07	3.93	20.10	78.65
TDANet (Zhang et al., 2020a)	4.71	33.10	2.03	3.72	20.40	77.77
ALMR (Wu et al., 2021)	4.51	26.33	1.66	3.71	20.25	75.69
CMA (ours)	**4.24**	**18.96**	**1.49**	**3.64**	**20.48**	**78.75**

Table 6: Results on CUB with the object mask.

in Table 1, Table 2, Table 3 and Table 4. From the tables, we can see that our model achieves state-of-the-art results compared with other strong competitors. In addition, we can make two observations: Firstly, using text to guide image inpainting can significantly improve the performance. For example, the text-guided image inpainting methods (i.e., MMFL, TGII, TDANet, ALMR and CMA) outperform standard image inpainting methods (i.e., CSA, PICNet and CTSDG). Next, for the gap between results on center masks and object masks, it shows that recovering completely removed objects is more difficult. Besides, we show results on the CUB (Wah et al., 2011) dataset without a wide variety of objects in Tabel 5 and Tabel 6. Results show our method outperforms other methods.

3.4 Ablation Study

We conduct an ablation study to investigate the effectiveness of each component of our approach and the results are reported in Table 7. We first in-

Method	FID ↓	KID ↓	PSNR ↑	SSIM (%) ↑
CMA	39.52	1.34	22.07	85.18
w/o G-Adv.	43.72	2.17	21.39	84.12
w/o L-Adv.	42.59	1.84	21.74	84.75
w/o Adv.	50.94	2.91	19.68	82.49
w/o Recon.	47.57	2.48	20.49	83.56
w/o CMAD	53.39	3.12	19.55	81.94
w/o ISD	52.84	3.06	19.67	82.23
w/o WPA	52.33	3.10	19.63	82.58

Table 7: Ablation study. 'G-Adv.', 'L-Adv.' and 'Adv.' denote global, local and both adversarial losses, respectively. 'Recon.', 'CMAD', 'ISD' and 'WPA' indicate reconstruction loss, cross-modal alignment distillation, in-sample distillation objectives and word patch alignment, respectively.

vestigate the impact of the adversarial learning by removing global adversarial loss, local adversarial loss and both adversarial losses and find that the performance drops. The reason is that adversarial objectives focus on high-level features, which can
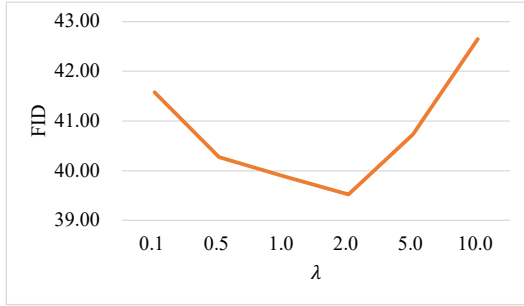
Figure 3: Performance of our model with different trade-off parameters λ.
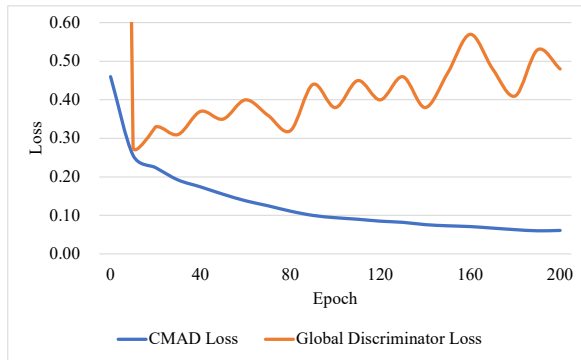

Figure 4: The adversarial loss and cross-modal alignment distillation loss on the training set w.r.t different epochs in the training phase.

effectively fill the missing region on complicated structures. Next, we test our method without reconstruction loss, which also decreases scores on all metrics. It demonstrates that image reconstruction plays an essential role in model training. Finally, we compare our method with the baseline without cross-modal alignment distillation and in-sample distillation, and the performance drops significantly. It indicates that the CMAD and ISD objectives can enhance the cross-modal alignment for text-guided image inpainting.

3.5 Analysis

3.5.1 Impact of Cross-Modal Alignment

To assess the impact of cross-modal alignment for our method, we set different trade-off parameters λ. Specifically, during model training, we set a different λ for the loss function in Eq.17. The performance of our method in different λ is reported in Figure 3. It is observed that with the λ increased, FID first drops and then rises, indicating that our method's performance first rises and then drops. The results demonstrate that the cross-modal alignment objective is crucial for our method, but

Method	Naturalness	Semantic Consistency
CMA	1.44	1.46
TDANet	2.39	2.52
ALMR	2.77	2.64
CSA	3.40	3.38

Table 8: Numerical ranking score of user study. The lower the score, the better the performance.
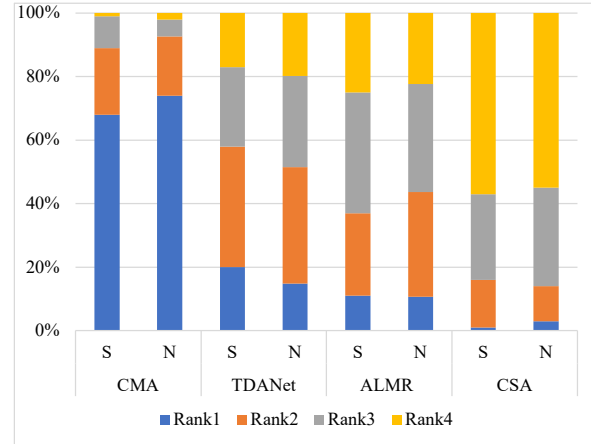

Figure 5: Ranking score distribution of the user study. "S" means semantic consistency score, "N" means naturalness.

overemphasizing the objective may hinder model learning.

3.5.2 Deep Dive into Cross-Modal Alignment

We take a deep dive into the impact of cross-modal alignment. In our method, adversarial learning involves a global discriminator to distinguish whether the generated image is original or generated, and the cross-modal alignment objective is to guide visual priors reconstruction. Their loss values, i.e., $\ell_{\text{G}-\text{adv,D}}$ and ℓ_{CMAD}, are plotted in Figure 4. We can see that the loss of cross-modal alignment objective quickly drops and then slowly drops, indicating that the model gets good performance on visual priors reconstruction. Meanwhile, we can observe that the adversarial loss of the discriminator quickly drops and then slowly goes up, demonstrating that we can see that the classification loss of the discriminator gets good performance and then is fooled later by the image generated from the image generator. Therefore, it shows that the cross-modal alignment objective supports the adversarial learning very well.

3.5.3 User Study

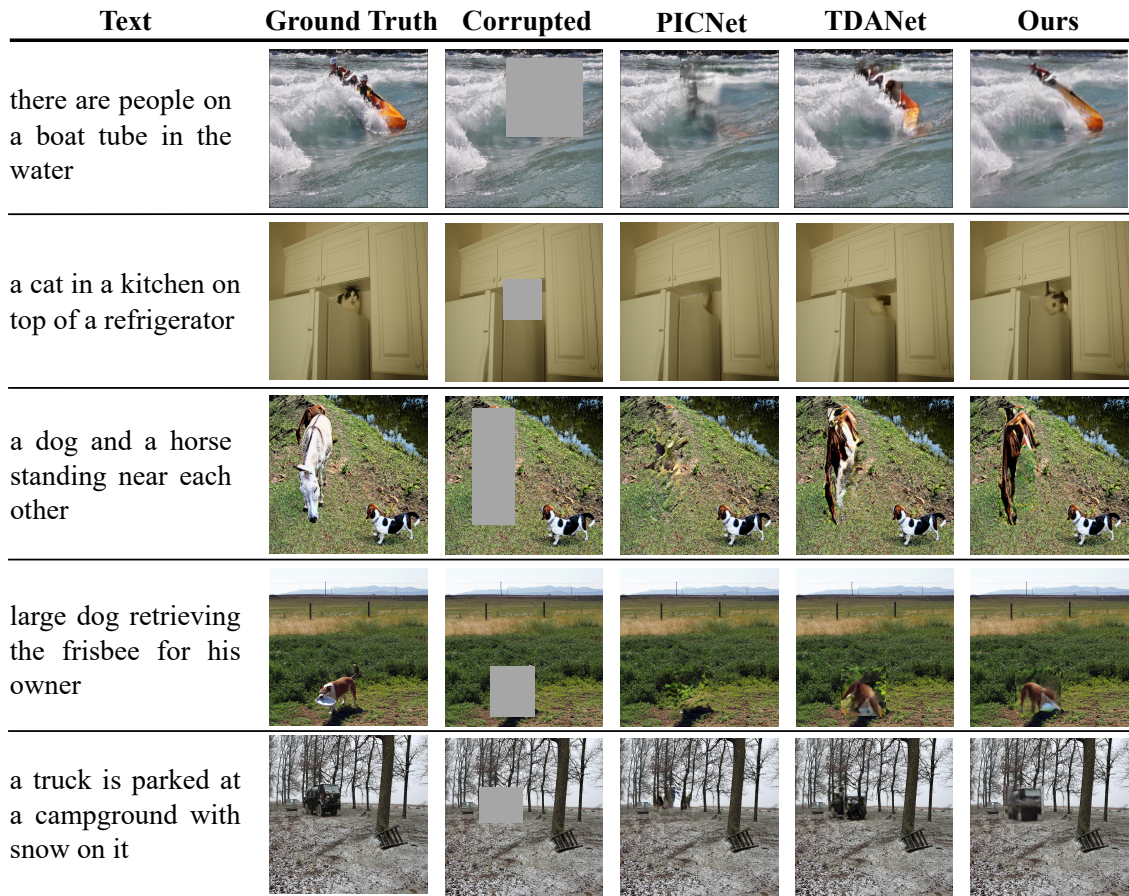Following Zhang et al. (2020a), we take a user study to quantify the qualitative comparison from

3451

Text	Ground Truth	Corrupted	PICNet	TDANet	Ours
there are people on a boat tube in the water					
a cat in a kitchen on top of a refrigerator					
a dog and a horse standing near each other					
large dog retrieving the frisbee for his owner					
a truck is parked at a campground with snow on it					

Figure 6: Qualitative results randomly sampled from MSCOCO test set.

the human perspective. We randomly collected 100 images in center masks from the MSCOCO test dataset. Each sample includes four generated images from CMA, PICNet, ALMR and CSA, respectively. These four images are randomly shuffled for five volunteers who rank images according to naturalness and the semantic consistency with the text description. Next, we computed the average ranking score as shown in Table 8 and show their distribution in Figure 5. From the results, we can observe that our model outperforms other competitors in naturalness and semantic consistency. It demonstrates the effectiveness of our method from a qualitative perspective.

3.6 Qualitative Results

As demonstrated in Figure 6, we give qualitative comparison examples of our method and other methods. We can find that our model can generate more semantically plausible objects in the missing region. Firstly, comparing PICNet with TDANet and our method, extracting the semantic information from the text can improve the repair effect of

the model in the missing area of the image. Secondly, we can observe that TDANet can generate preliminary results, such as the outlines and part content of "cat", "truck" and "people" in examples, but many details of the object are very unnatural. In contrast, our method can further generate the details of the object based on generating the outline and content of the object. It demonstrates that cross-modal alignment can effectively supplement missing object information.

4 Related Work

4.1 Image Inpainting

Image inpainting (Bertalmío et al., 2000) aims to restore a damaged image, whose categories of approaches mainly are patch-based and deep learning-based. The patch-based methods (Barnes et al., 2009; Huang et al., 2014) fill the holes through searching and pasting patches based on image known regions. Huang et al. (2014) propose a method using the mid-level structural cues to automatically guide the image inpainting. However,

these methods are not effective to fill in the missing region on complicated structures, due to the focus on low-level features. To address the limitation of existing patch-based methods, there has been growing interest in deep learning-based methods (Pathak et al., 2016; Iizuka et al., 2017; Ren et al., 2019). The Context Encoder (CE) is proposed by (Pathak et al., 2016), which uses the encoder-decoder architecture and the Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) to learn image features. Although the CE improves the inpainting by the image features learning, it is not effective to tackle the visual artifacts and exhibits blurriness in the image recovered regions. For solving the aforementioned problems, Iizuka et al. (2017) introduce the local and global discriminator for the image inpainting of arbitrary missing regions, which improves the local and global consistency of generated image. The StructureFlow (Ren et al., 2019) consisted of a structure reconstructor and a texture generator, which can focus on recovering global structures and synthesizing high-frequency details. Although the existing image inpainting methods can fill in the holes in the image, generating specific content in the missing region remains challenging, without any known information.

4.2 Text-Guided Image Inpainting

To address this problem, many text-guided image inpainting works are proposed (Zhang et al., 2020b; Lin et al., 2020). In these works, the specific content in the missing area can be restored based on the given descriptive text. Existing text-guided image inpainting methods include two processes: semantic information extraction and multimodal fusion. The semantic information extraction aims to obtain semantic information which does not match the image, such as the dual multimodal attention mechanism (Zhang et al., 2020a). However, these methods are hard to work well in the image set with a variety of different objects. The reason is that these methods lack fine-grained alignment knowledge of texts and images to guide the fusion of cross-modal information in multimodal interactions. Besides, their fusion modules lack powerful cross-modal reasoning capabilities.

4.3 Vision-Language Pre-training (VLP)

Motivated by the success of the language/vision pre-trained model (Devlin et al., 2019; Dosovitskiy et al., 2021; Zhou et al., 2022b), there is a surging interest in developing a pre-trained model for mul-

tiple modalities (e.g., vision and language) (Chen et al., 2020; Radford et al., 2021; Kim et al., 2021; Zhou, 2022). For example, a pioneering work CLIP (Radford et al., 2021) employs contrastive learning to predict whether matching between image and text and shows its powerful capability in many downstream tasks. UNITER (Chen et al., 2020) and UNIMO (Chen et al., 2020) employ an object detector (e.g., Faster R-CNN (Ren et al., 2017)) to capture vision features, and a multi-layer transformer (Vaswani et al., 2017) is used to joint learn vision features and text features. Kim et al. (2021) discuss different taxonomy of vision-and-language models and propose ViLT, a pre-trained model more focused on modeling modality interactions. In addition, ViLT totally discards convolutional visual features and adopts vision transformers.

5 Conclusion

In this work, we explore a novel CMA model for text-guided image inpainting. In the CMA model, we integrate a vision encoder and a text encoder into a vision-and-language encoder, which is different from previous works. The vision-and-language encoder allocates more computation on modeling modality interactions instead of visual encoding. In addition, we introduce two objectives to improve cross-modal alignment, dubbed cross-modal alignment and in-sample distillation. The cross-modal alignment objective guides the model to fuse cross-modal features. Experimental results demonstrate that the proposed model delivers new state-of-the-art performance, followed by further analyses to provide comprehensive insights.

Limitations

Compared with previous text-guided image inpainting methods (Zhang et al., 2020b; Lin et al., 2020; Wu et al., 2021), our method performs well for natural image datasets with a wide variety of objects. However, the recovery of our method for completely missing objects in images is not perfect. The reason is that the size of our model limits its capabilities. The large image generation models (e.g., DALL·E (Ramesh et al., 2021)) show their powerful capability to generate a plausible image. Due to the limitation of computational resources, we could not train a large model of similar size to DALL·E, which hinders verifying the effectiveness of our method on a large model.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223.

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

Marcelo Bertalmío, Guillermo Sapiro, Vicent Caselles, and Coloma Ballester. 2000. Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2000, New Orleans, LA, USA, July 23-28, 2000*, pages 417–424. ACM.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Fernando A. Fardo, Victor H. Conforto, Francisco C. de Oliveira, and Paulo S. Rodrigues. 2016. A formal evaluation of PSNR as quality measurement parameter for image segmentation algorithms. *CoRR*, abs/1605.07116.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680.

Xiefan Guo, Hongyu Yang, and Di Huang. 2021. Image inpainting via conditional texture and structure dual generation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 14114–14123. IEEE.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637.

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

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

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick. 2020. Pointrend: Image segmentation as rendering. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9796–9805. Computer Vision Foundation / IEEE.

Qing Lin, Bo Yan, Jichun Li, and Weimin Tan. 2020. MMFL: multimodal fusion learning for text-guided image inpainting. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1094–1102. ACM.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision -*

ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. 2019. Coherent semantic attention for image inpainting. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4169–4178. IEEE.

Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. 2016. Context encoders: Feature learning by inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2536–2544. IEEE Computer Society.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Int. J. Comput. Vis.*, 123(1):74–93.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149.

Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H. Li, Shan Liu, and Ge Li. 2019. Structureflow: Image inpainting via structure-aware appearance flow. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 181–190. IEEE.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. 1992. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset.

Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. 2021. High-fidelity pluralistic image completion with transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 4672–4681. IEEE.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904.

Xingcai Wu, Yucheng Xie, Jiaqi Zeng, Zhenguo Yang, Yi Yu, Qing Li, and Wenyin Liu. 2021. Adversarial learning with mask reconstruction for text-guided image inpainting. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 3464–3472. ACM.

Lisai Zhang, Qingcai Chen, Baotian Hu, and Shuoran Jiang. 2020a. Text-guided neural image inpainting. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1302–1310. ACM.

Zijian Zhang, Zhou Zhao, Zhu Zhang, Baoxing Huai, and Jing Yuan. 2020b. Text-guided image inpainting. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 4079–4087. ACM.

Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2019. Pluralistic image completion. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1438–1447. Computer Vision Foundation / IEEE.

Yucheng Zhou. 2022. Sketch storytelling. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 4748–4752. IEEE.

Yucheng Zhou, Xiubo Geng, Tao Shen, Guodong Long, and Daxin Jiang. 2022a. Eventbert: A pre-trained model for event correlation reasoning. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 850–859. ACM.

Yucheng Zhou, Tao Shen, Xiubo Geng, Guodong Long, and Daxin Jiang. 2022b. Claret: Pre-training a correlation-aware context-to-event transformer for

event-centric generation and classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2559–2575. Association for Computational Linguistics.

Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020. In-domain GAN inversion for real image editing. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*, volume 12362 of *Lecture Notes in Computer Science*, pages 592–608. Springer.