An Unified Framework for Language Guided Image Completion

Jihyun Kim*[1], Seong-Hun Jeong*[2], Kyeongbo Kong*[2], and Suk-Ju Kang[1]
[1]Sogang University, {jhkim5950, sjkang}@sogang.ac.kr
[2]Pukyong National University, tlqwkrk915@pukyong.ac.kr, kbkong@pknu.ac.kr

Abstract

Image completion is a research field which aims to generate visual contents for unknown regions of an image. Image outpainting and wide-range image blending, which we refer to as extensive painting, are considered challenging because compared to the large unknown regions, relatively less context is provided. Some recent studies have tried to decrease the complexity of extensive painting by generating image hints for the missing regions. In this paper, we introduce a novel modality of hints, the natural language. Moreover, we propose a Captioning-based Extensive Painting (CEP) module, which combines models for two different multi-modal tasks: image captioning and text-guided image completion. In order to generate appropriate captions for masked images, the image captioning model is optimized using self-critical sequence training (SCST) method with random masks. The biggest benefit of our methodology is the accessibility to well-designed image captioning and text-guided image manipulation models such as OFA and GLIDE without the need for additional architectural changes. In evaluation, our model demonstrates remarkable performance even with complicated image datasets both quantitatively and qualitatively.

1. Introduction

Everyone should be familiar with Mona Lisa, the masterpiece by Leonardo da Vinci, but have you ever imagined what kind of skirts and shoes Mona Lisa is wearing? If so, you have performed a type of image completion, known more specifically as image outpainting. Image completion is a research field which aims to generate plausible visual contents for unknown regions in given images. This research area includes not only conventional tasks, such as image inpainting [1, 2, 3] and outpainting [4, 5, 6], but also a newly suggested task, wide-range image blending [7]. Image inpainting is an image restoration task which fills in missing/masked regions of an input image. By contrast, image outpainting is a task which aims to extend a given image beyond its original boundaries. Lastly, wide-range image blending aims to generate an intermediate image between two different images so that they form a single panoramic image. Image outpainting and wide-range image blending have been proven to be more difficult compared to image inpainting because they create novel contents for a larger area and have less contextual information to take reference from. As such, we have decided that tackling image outpainting and wide-range image blending would be an excellent way to demonstrate the competency and efficiency of our method (Fig. 1). For convenience, we refer to the two challenging tasks as *extensive painting*.

Recently, there have been efforts to solve the lack of information problem by generating image hints for the missing regions. One of the techniques that has been instituted to resolve this issue is Bidirectional Rearrangement (BR) [8], which utilizes the opposite parts of the input image as hints by switching the left and rights sides, and fills in the gap between the two split images. In addition, the Mirrored Rearrangement (MR) [9] predicts missing region by using the mirror flipped input image as hints. While the image hints made by the BR and the MR methods have only been applicable to symmetrical images due to structural reasons, a more recent study, Image-Adaptive Hint (IAH) [10], expands upon the hint-based method to asymmetrical images. IAH generates image hints via Vision Transformer in an image-adaptive way. Although various hint-generation methods have been implemented, the output hints are limited to the image format.

In this paper, we introduce a novel modality of hints for extensive painting tasks, the natural language. As shown in Fig. 2, to incorporate language hints to extensive painting, our unified framework named Captioning-based Extensive Painting (CEP) module consists of two multi-modal tasks: an image captioning task and a text-guided image manipulation task. Firstly, the image captioning module generates language hints by capturing semantic and textual information of the given images. Secondly, with the guidance of the language hints, the text-guided image manipulation module generates visual contents for masked regions that are in line

*equal contribution

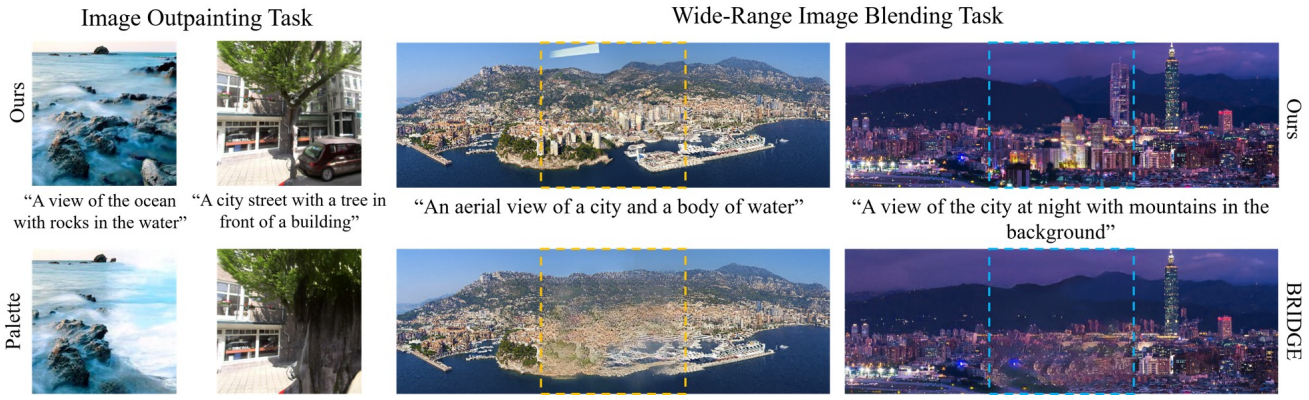

Figure 1. **Illustration of image outpainting and wide-range image blending.** Image outpainting is a task which aims to extend a given image beyond its original boundaries. Wide-range image blending aims to generate an intermediate image between two different images so that they form a single panoramic image. Compared to the state-of-the-art algorithm of each task, our method has outstanding performance. Red boxes indicate generated regions.

with the content of the unmasked regions. For each image outpainting task, hint generation and hint-guided image editing are each performed once. On the contrary, for the wide-range image blending task, image outpainting on the two input images is repeated until the gap between the two images is completely filled. In the end, to ensure smooth transition, our framework performs image inpainting on the very middle part of the generated intermediate image.

The most important part of our CEP module is image captioning. No matter excellent the text-guided image manipulation models are, the final results of our unified framework heavily depend on the captions. Nonetheless, existing captioning models are incapable of generating proper captions for the masked images because they have been trained only on complete images. For instance, during our experiment, the captioning model repeatedly recognized the masked image as a combination of two different images, and generated inappropriate prefixes such as 'two pictures of', resulting in generation of semantically awkward images. To solve this problem, we added to our CEP module an optimizing process using self-critical sequence training (SCST) method with randomly masked images and their matching captions. As a result, our optimized captioning model is able to predict an appropriate caption regardless of the mask shapes. Most importantly, it can generate language hints for various image completion tasks, which was never possible for prior hint-based methods [8, 9, 10] due to their dependency on image structures.

Another benefit of our module is the accessibility to various pretrained models. Since both image captioning and text-guided image manipulation have been popular even among many multi-modal tasks, various algorithms are available. For example, popular algorithms such as One For All (OFA) [11] and ClipCap [12] are for the image captioning task, while GLIDE [13] and a text-drive blended diffusion [14] are for the text-guided image manipulation task. All of the algorithms mentioned above have achieved a remarkable level of performance. In particular, trained on hundreds of millions of data, large-scale text-guided image synthesis models are capable of generating photo-realistic images across most of the image domains [15]. By taking advantage of these multi-modal models, our module is able to generate semantically plausible and realistic images for image outpainting, as well as for wide-range image blending.

In essence, our main contributions include:

- We introduce natural language hints for image completion tasks, specifically for image outpainting and wide-range image blending. We refer to the two challenging tasks as *extensive painting* in reference to their large unknown regions. Although there have been efforts to utilize image hints for image outpainting tasks, we are the first to generate hints in a language format.

- We optimized an image captioning model with a large-scale caption dataset, using random masks. As a result, our image captioning model generates appropriate language hints for various image completion tasks, regardless of the mask shapes.

- Our framework can utilize various algorithms for the image captioning task and the text-guided image manipulation task, and links them together for extensive painting. Thanks to the recent progress in multi-modal tasks, especially in the text-guided image synthesis task, our module achieves remarkable performance and high level of generalization in image domains.

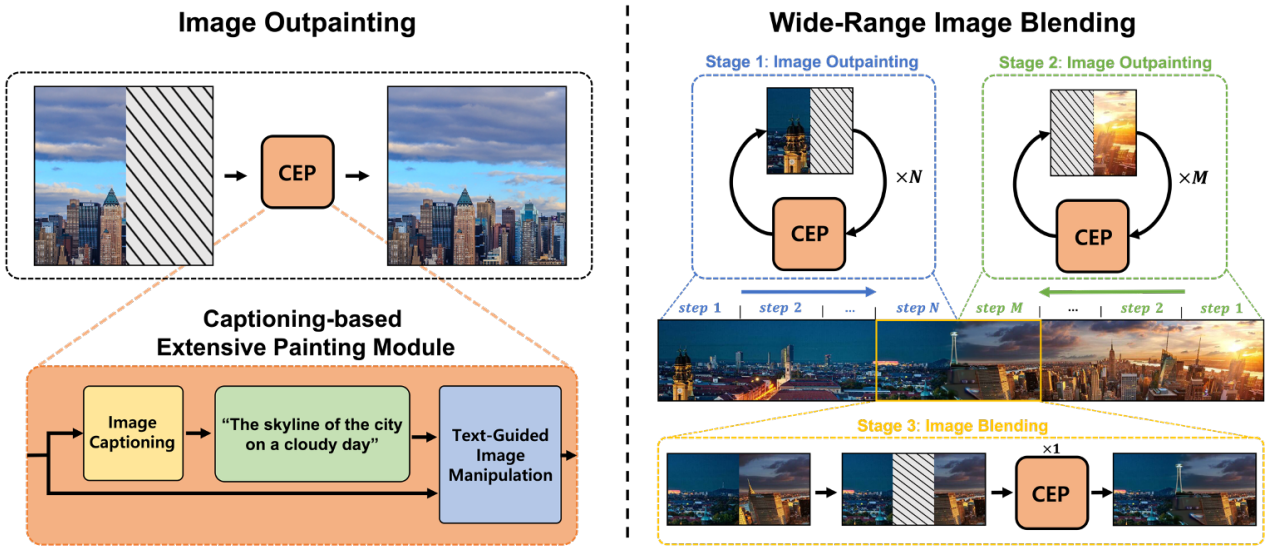

Figure 2. **Overall process of extensive painting with the proposed CEP module.** ***Left***: Image outpainting is performed through our CEP module which consists of image captioning and text-guided image manipulation networks. A masked image is passed through the image captioning network to generate the language hint. Then, with the language hint, the missing regions are predicted by an text-guided image manipulation network. ***Right***: Wide-range Image blending consists of three stages. First two stages repeat image outpainting for N, M times on different images with the reverse direction. For a natural connection, the predicted output from the previous step is used as the input for the next step. Then, to blend the resulted images, in the image blending stage, we mask the disconnected region and pass it through our CEP module once.

2. Related Work

2.1. Image Completion

The two major tasks of image completion include image inpainting and image outpainting. Classic approaches for both tasks were mostly patch-based [16, 17, 18, 19, 20], which find patches from known regions for each masked region. However, these methods often failed to learn semantic structures and had limited abilities to generate new contents. On the contrary, learning-based methods [1, 2, 3] were relatively better at capturing semantics. As for image outpainting, due to the lack of adjacent information and larger unknown area, there have been efforts to generate visual hints for the missing regions [8, 9, 10]. Recently, [7] proposed a new image completion task, wide-range image blending, which aims to merge two different images into one by generating appropriate contents in between the two images. For this task, [7] presents an encoder-decoder architecture, which sequentially predicts the feature representations for the intermediate region.

2.2. Image captioning

Modern image captioning models typically employ an encoder-decoder architecture [21, 22, 23], in which the encoder extracts visual features from an image and the decoder generates a sequence of words from the extracted visual features. To train an image captioning model, cross entropy loss followed by reinforcement learning [24] is commonly used. Such training method enables the usage of non-differentiable caption metrics as optimization objectives. Recently, there were efforts to adopt vision-language pretraining on large image-text corpus to image captioning tasks [25, 26].

2.3. Text-guided Image Manipulation

Initially text-guided image manipulation models focused on editing an image according to a given text prompt and the model decided which part of the image to edit [27]. [28] introduces a text-guided image manipulation model which generates contents only for desired regions using a dual attention mechanism. Recent works further improve the performance of text-guided image manipulation models by training them on large scale datasets [13, 14].

3. Proposed Method

In this section, we provide an overview of our framework, which is shown in Fig. 2. Our framework, named Captioning-based Extensive Painting (CEP) module, comprises two types of networks 1) an image captioning network G_{CAP} for language hints generation and 2) a text-guided image manipulation network, G_{IM}. Basically, when a masked image is given as an input, G_{CAP} generates language hints by captioning the masked image. Then, using the language hint, G_{IM} generates images for the masked re-

gion. The two parts of CEP module operate in a pair to conduct either image outpainting or wide-range image blending, the two most challenging image completion tasks. In the following sections, we describe the process of the CEP module in details. Then, we will explain how to apply the CEP module to extensive painting tasks.

3.1. Captioning-based Extensive Painting Module

Typically, image completion for large unknown area is challenging due to the limited amount of neighbouring information. Therefore, the goal of generating hints during extensive painting is to provide as detailed as possible information about the input image [12]. In order to meet the goal, we adopt an image captioning model, which aims to describe the contents of an image with natural language. Then, using this caption as a hint, we fill in the missing region of the image using a text-guided image manipulation network. Overall process is as follows:

Let I_{GT} be the ground-truth image and M be the binary image (1 for the missing region and 0 for the background). Then, an incomplete image can be represented as:

$$I_{IC} = I_{GT} \odot (1 - M), \tag{1}$$

where $\odot$ denotes the Hadamard product. The image captioning network generates the text caption used for the language hint

$$T_{hint} = G_{CAP}(I_{IC}). \tag{2}$$

After obtaining T_{hint}, text-guided image manipulation network G_{IM} generates a complete image. The masked image I_{IC} and the language hint T_{hint} were used as inputs as follow:

$$I_{pred} = G_{IM}\left(I_{IC}, T_{hint}\right), \tag{3}$$

where I_{pred} denotes the final output result.

Image Captioning Network Image captioning is the most popular task in the multi-modal field. However, existing image captioning models have been trained only on complete images. Therefore, they are incapable of generating proper captions of the masked images. To overcome this problem, we perform additional training with randomly masked dataset $I_{rand} = I_{GT} \odot (1 - \tilde{M})$, where $\tilde{M}$ is the random mask.

For image captioning, we optimized a language model, which was pre-trained over a large dataset, with self-critical sequence training (SCST) approach [24, 29, 30]. This approach is based on the REINFORCE algorithm [31], where the reward is set to the metric used at the test time. Given the sentence T^s_{hint} sampled from the policy during training and a captioning model with parameters θ, we minimize the negative expected reward:

$$L_R(\theta) = -\mathbf{E}_{T^s_{hint} \sim p_\theta}\left[r(T^s_{hint})\right], \tag{4}$$

where r is the reward computed by an evaluation metric (e.g. CIDEr), by comparing the generated sequence to the corresponding ground-truth sequences. Then, the gradient of (4) can be approximated by:

$$\nabla_\theta L_R(\theta) \approx -\left(r(T^s_{hint}) - r(\hat{T}_{hint})\right) \nabla_\theta \log p_\theta(T^s_{hint}), \tag{5}$$

where $r(\hat{T}_{hint})$ is the baseline reward obtained by greedily decoding the model. This gradient tends to increase the probability of the caption sampling from the policy during training than the reward from the current model [32]. Through this process, our newly optimized captioning model is able to predict the appropriate caption, regardless of the mask shapes.

One of the advantages of language hints over image hints is that it works well regardless of the image completion tasks. Prior image hint-based methods [8, 9, 10] have large dependency on image structure. More specifically, BR [8] and MR [9] are applicable only to symmetric images, since they either rearrange the input image by switching the left and right side, or mirror the input image next to the masked region. Moreover, IAH [10] can generate only fixed size hints. On the contrary, since the CEP uses other modality as the hint, it can be applied to any type of image completion without any restrictions.

Text-guided Image Manipulation Network Recently, thanks to the works on massive text-image paired datasets [15, 33, 34, 35], large-scale language-vision models such as GLIDE [13], Imagen [36], DALL-E [36], and text-drive blended diffusion [14] have achieved an unprecedented level of generalization. In other words, these large-scale models trained on large-scale datasets are competent in zero-shot image generation on a wide range of domains, even compared to models trained on the specific dataset. Therefore, we simply utilize these models without any change. We expect these benefits to enrich various real-life applications of extensive painting. Next, we will apply our CEP module to the most difficult image completion tasks, image outpainting and wide-range image blending.

3.2. Image Outpainting

The goal of image outpainting is to generate the outside of the images when the given information is only the inside of the images. Generally, the outpainting task predicts a single side or both sides of images in horizontal direction. Since we optimized the image captioning network of the CEP module for random masked images, we can perform the outpainting task by putting only the mask corresponding to the missing region and passing it through the CEP module.

Table 1. Quantitative results for the image outpaitning task on Landscape [6], Landmarks [37], and AmsterTime [38] dataset, using metrics on Ŕenchet Inception Distance (FID) [39] (the lower, the better) and Kernel Inception Distance (KID) [40] (the lower, the better). The best score for each column is depicted in bold letters.

Outpainting	Landscape		Landmarks		AmsterTime	
Method	FID	KID	FID	KID	FID	KID
CEP (ours)	**25.46**	**0.001**	**13.48**	**0.001**	**26.98**	**0.002**
Boundless [4]	44.19	0.011	22.38	0.011	59.12	0.026
Image Outpainting [41]	75.79	0.038	62.67	0.055	84.91	0.086
SRN [5]	37.07	0.016	37.49	0.030	63.12	0.034
NS-OUT [6]	52.28	0.072	43.87	0.030	66.84	0.038
Palette [42]	64.66	0.038	15.31	0.004	39.22	0.009

Table 2. Quantitative results for wide-range image blending task on Landmarks [37], Scenery [6], and AmsterTime [38] datasets using metrics on on FID [39] (the lower, the better), and KID [40] (the lower, the better). The best score for each column is depicted in bold letters.

Wide-Range Image Blending	Landmarks		Scenery		AmsterTime		Image 4K	
model	FID	KID	FID	KID	FID	KID	FID	KID
CEP (ours)	**20.39**	**0.004**	38.05	0.017	**29.43**	**0.003**	**29.41**	**0.003**
BRIDGE [7]	36.72	0.022	**36.31**	**0.011**	71.95	0.047	40.65	0.009
CA [3]	52.49	0.016	91.87	0.074	43.46	0.021	61.46	0.028
PEN-Net [43]	79.57	0.035	159.70	0.115	74.67	0.047	86.71	0.049
Hifill [44]	70.90	0.041	139.39	0.123	74.82	0.058	76.57	0.042
SRN [5]	94.35	0.041	70.94	0.039	81.21	0.048	123.58	0.109
NS-OUT [6]	103.77	0.090	82.69	0.044	135.63	0.144	116.64	0.069

3.3. Wide-range Image Blending

To perform the wide-range image blending task, the overall process consists of three stages. Stage 1 is the multi-step prediction, which repeats outpainting using the predicted output from the previous step as the input for the next step. In stage 2, outpainting is repeated in the opposite direction to stage 1 until the two extrapolated images become connected. Finally, in stage 3, we apply a mask to the disconnected region and pass the masked image through our CEP module once.

4. Experiments

4.1. Baseline methods

Our proposed CEP module is model-agnostic for both image captioning and text-guided image manipulation tasks. For the image captioning task, we implemented OFA [11] and ClipCap [12], and for the text-guided image manipulation task, we utilized GLIDE [13] and the text-driven blended diffusion model [14].

4.2. Datasets

We evaluate our module on conventional datasets for extensive painting, Scenery6000 [6] and Beach datasets [41]. We further conduct experiments on complicated datasets such as AmsterTime [38] and Landmarks [37] datasets.

Beach dataset [41] This dataset was generated by selecting images from 'beach' category of place365 [45]. It consists of 9,465 train images and 1,050 test images, each with 256×256 pixel resolution.

Scenery6000 dataset [6] This dataset contains scenery images of varying sizes. Among the 6,040 images in total, 1,000 are for testing, and 5,040 images are for training.

AmsterTime dataset [38] This dataset includes 1,231 images of Amsterdam's urbanscape, with 800×600 pixel resolution.

Landmarks Dataset [37] This dataset contains 26,397 train images and 3,103 test images of landmarks around the world, ranging from natural landscapes to architectures.

4K dataset This is a Kaggle provided dataset composed of various scenery images and object images in 4K resolution. All of the 2,056 images were used for testing.

4.3. System Set-up

To optimize OFA [11] via SCST optimization using CIDEr evaluation metric, we utilize the most commonly used MS COCO Caption dataset [46]. Opimization is conducted for 5,000 steps with a batch size of 2 and a learning rate of 5e-6. As for the other image captioning model [12] and text-guided image manipulation models [13, 14], the original setup is used.

For image outpainting, input images are cropped and resized into the resolution of 256×256, and for wide-range image blending, input images are first cropped and resized into the resolution of 768×256. Then for self-

Figure 3. Qualitative results of image outpainting task on AmsterTime dataset.

reconstruction, the middle 256×256 region is cropped, and the remaining left 256×256 and right 256×256 regions serve as the two input images. The mask shape for image outpainting is 128×256. For wide-range image blending, the mask shape for extrapolation step is 128×256 and for inpainting step, 64×256. All models are implemented on 4 NVIDA GeForce RTX 3090 GPUs.

4.4. Quantitative Results

We used Frenchet Inception Distance (FID) [39] and Kernel Inception Distance (KID) [40] as our evaluation metrics. Note that we have utilized weights for Landmarks dataset to test on AmsterTime dataset, for their visual similarities and the lack of training images for AmsterTime (AmsterTime dataset contains 1,231 images in total, whereas Landmarks dataset possesses 26,397 images just for training).

Image Outpainting The results for image outpainting task on Landscape, Landmarks, and AmsterTime datasets are summarized in Table 1. Our CEP module outperforms all baselines, including the state-of-the-art image completion model Palette [42]. Moreover, while the Palette model scores particularly high in Landmarks dataset, in which the number of training images is large, our module constantly achieves high scores in all datasets.

Wide-Range Image Blending The quantitative results are shown in Table 2. In all datasets except for Scenery dataset, our method beats every baseline, even the BRIDGE model [7], which was particularly designed for this task.

Figure 4. Qualitative results of wide-range image blending task on Image 4K dataset.

4.5. Qualitative Results

Fig. 3 shows qualitative results of conventional outpainting algorithms and the CEP module on AmsterTime dataset. Previous methods tend to produce structurally unnatural images, but our method produces structurally coherent and content-preserving images. This characteristic also maintains for wide-range image blending task. As shown in Fig. 4, our method produces clean and semantically meaningful images while BRIDGE, the state-of-the-art method in wide-range image blending, produces repetitive structures. Additional results are to be found in the appendix.

4.6. Ablation Studies

Effect of optimizing an image captioning model Image captioning models created inaccurate prefixes when given masked images, such as "two photos of" and its variants. When text-guided image manipulation models were provided with such inaccurate captions, they generated two divided images during wide-range image blending as shown in Fig. 5. Thus, we measured the effect of optimizing an image captioning model, OFA, using SCST method with random masks. Then, we counted the number of false prefixes. As summarized in Table 3, SCST optimization successfully reduces the number of error cases.

Effect of captions for extensive painting In Fig. 6, we closely look at the effect of captions for extensive painting.

Figure 5. Effect of optimizing an image captioning model on wide-range image blending; Withoutthe SCST optimization process, the text-guided image manipulation model generated two divided images, whereas with the optimization process, we obtained smoothly blended images

In the image outpainting task, a canal was included in the generated image, because of the word "canal" in the caption. Note that this was not included in the initial image. Also in the image blending task, a yellow beak of an eagle was generated through the caption, "eagle with yellow beak". This indicates that image captions for missing regions generate diverse and natural images.

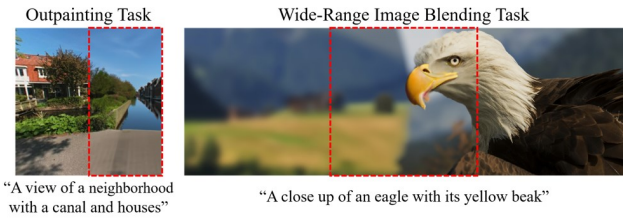

Figure 6. Effect of captions for extensive painting; Image captions for missing region provides diverse and natural images. Red boxes indicate generated regions.

Table 3. Effect of SCST optimization on the image captioning model; Image captioning model OFA [11] was optimized via SCST method using CIDEr with randomly masked COCO caption dataset [46]. GLDIE [13] was used for image text-guided image manipulation, and evaluation was conducted on image 4K dataset.

OFA training	# of False prefixes	FID	KID
w/o training	1055	33.81	0.003
w/ training	**3**	**32.55**	**0.002**

Table 4. Effect of mask sizes; Image inpainting was conducted at the last stage of wide-range image blending on Landmarks dataset [37], using OFA [11] and GLIDE [13] with varying mask widths.

mask size	FID	KID
32	21.15	0.004
64	**20.85**	**0.003**
96	20.90	0.003
128	22.12	0.004

Table 5. Comparison of hint-based image outpainting models on beach dataset.

Method	Hint Format	FID
CEP (ours)	Language	**24.95**
BR [8]	Image	37.19
MR [9]	Image	36.65
IAH [10]	Image	31.81

Effect of Mask Sizes In the final step of wide-image blending, we masked the very center of the panoramic image and performed image inpainting for a smooth transition. From Table 4, we can conclude that performing image inpainting using a mask with a width of 64, yields the best results.

Comparison on hint-based methods We compared our method with conventional hint-based methods on beach dataset. As shown in Table 5, our algorithm was superior to the existing ones.

Comparison on captioning models Finally, we compared the results of our method using different captioning models. In Table 6, the CEP module with OFA outperformed the CEP module with ClipCap.

Table 6. Comparison of captioning models on Landmarks dataset.

Captioning Model	FID	KID
OFA [25]	**13.48**	**0.002**
ClipCap [12]	16.35	0.003

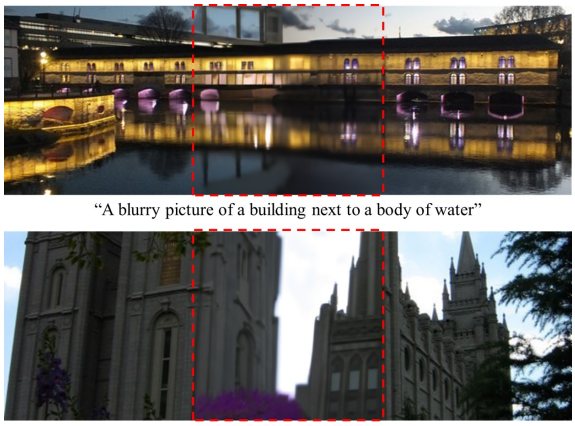

Figure 7. Failure cases; The word "blurry" in captions generated blurry images. Red boxes indicate generated regions.

4.7. Limitations

Our method fails in certain cases. As shown in Fig. 7 the image captioning model generates a caption containing a word "blurry" for blurry images. As a result, the image generation model generates exaggeratedly blurry images when given captions with a word "blurry". Another limitation of our CEP module employing GLIDE [13] is that the resolution of the output at each step is limited to 256×256.

5. Conclusion

In this paper, we propose a novel modality of hints, the natural language, and incorporate it to image outpainting, and even to wide-range image blending tasks. Since the proposed captioning-based Extensive Painting (CEP) module can adopt any image captioning and text-guided image generation algorithms for each hint generation and image generation network, we take advantage of the well-designed pretrained models. For both tasks, our module outperforms the baseline models most of the times, and generates photo-realistic images even for semantically complicated images.

6. Acknowledgement

We acknowledge fundings from Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A1A01051225), and from the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1004208).

References

[1] 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] 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.

[3] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. *In The IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[4] Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10521–10530, 2019.

[5] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1399–1408, 2019.

[6] Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. Very long natural scenery image prediction by outpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10561–10570, 2019.

[7] Chia-Ni Lu, Ya-Chu Chang, and Wei-Chen Chiu. Bridging the visual gap: Wide-range image blending. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–851, 2021.

[8] Kyunghun Kim, Yeohun Yun, Keon-Woo Kang, Kyeongbo Kong, Siyeong Lee, and Suk-Ju Kang. Painting outside as inside: Edge guided image outpainting via bidirectional rearrangement with progressive step learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2122–2130, 2021.

[9] Naofumi Akimoto, Daiki Ito, and Yoshimitsu Aoki. Scenery image extension via inpainting with a mirrored input. *IEEE Access*, 9:59286–59300, 2021.

[10] Daehyeon Kong, Kyeongbo Kong, Kyunghun Kim, Sung-Jun Min, and Suk-Ju Kang. Image-adaptive hint generation via vision transformer for outpainting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3572–3581, 2022.

[11] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *International Conference on Machine Learning*, 2022.

[12] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[13] Thomas A Halgren, Robert B Murphy, Richard A Friesner, Hege S Beard, Leah L Frye, W Thomas Pollard, and Jay L Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *Journal of medicinal chemistry*, 47(7):1750–1759, 2004.

[14] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.

[15] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[16] 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.

[17] 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.

[18] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1033–1038. IEEE, 1999.

[19] Josef Sivic, Biliana Kaneva, Antonio Torralba, Shai Avidan, and William T Freeman. Creating and exploring a large photorealistic virtual space. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008.

[20] Yinda Zhang, Jianxiong Xiao, James Hays, and Ping Tan. Framebreak: Dramatic image extrapolation by guided shift-maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1171–1178, 2013.

[21] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.

[22] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[23] An-An Liu, Yingchen Zhai, Ning Xu, Weizhi Nie, Wenhui Li, and Yongdong Zhang. Region-aware image captioning via interaction learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[24] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017.

[25] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.

[26] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[27] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020.

[28] Lisai Zhang, Qingcai Chen, Baotian Hu, and Shuoran Jiang. Text-guided neural image inpainting. In *ACM International Conference on Multimedia*, 2020.

[29] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*, 2017.

[30] Junlong Gao, Shiqi Wang, Shanshe Wang, Siwei Ma, and Wen Gao. Self-critical n-step training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6300–6308, 2019.

[31] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[32] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[34] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[35] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[37] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2575–2584, 2020.

[38] Burak Yildiz, Seyran Khademi, Ronald Maria Siebes, and Jan van Gemert. Amstertime: A visual place recognition benchmark dataset for severe domain shift. *arXiv preprint arXiv:2203.16291*, 2022.

[39] 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.

[40] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[41] Mark Sabini and Gili Rusak. Painting outside the box: Image outpainting with gans. *arXiv preprint arXiv:1808.08483*, 2018.

[42] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, pages 1–10, 2022.

[43] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 1486–1494, 2019.

[44] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting, 2020.

[45] Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós, and Álvaro García-Martín. Semantic-aware scene recognition. *Pattern Recognition*, 102:107256, 2020.

[46] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.