Conditional Image Repainting via Semantic Bridge and Piecewise Value Function

Shuchen Weng[1], Wenbo Li[2], Dawei Li[2], Hongxia Jin[2], and Boxin Shi*[1,3]

[1] NELVT, Department of Computer Science and Technology, Peking University
[2] Samsung Research America AI Center
[3] Institute for Artificial Intelligence, Peking University
{shuchenweng,shiboxin}@pku.edu.cn
{wenbo.li1,dawei.l,hongxia.jin}@samsung.com

Abstract. We study conditional image repainting where a model is trained to generate visual content conditioned on user inputs, and composite the generated content seamlessly onto a user provided image while preserving the semantics of users' inputs. The content generation community has been pursuing to lower the skill barriers. The usage of human language is the rose among thorns for this purpose, because the language is friendly to users but poses great difficulties for the model in associating relevant words with the semantically ambiguous regions. To resolve this issue, we propose a delicate mechanism which bridges the semantic chasm between the language input and the generated visual content. The state-of-the-art image compositing techniques pose a latent ceiling of fidelity for the composited content during the adversarial training process. In this work, we improve the compositing by breaking through the latent ceiling using a novel piecewise value function. We demonstrate on two datasets that the proposed techniques can better assist tackling conditional image repainting compared to the existing ones.

Keywords: Image generation, semantic, compositing, adversarial

1 Introduction

The advanced image editing techniques lower the skill barriers and simplify the required user inputs. For example, FaceApp [1] simplifies the user inputs to just one click for various face editing tasks including the alternations of smile, age and style. The research community also witnessed the great efforts along this direction, *e.g.*, GauGAN [13] is trained to synthesize images collaboratively with users, *i.e.*, users draw contours of objects, and GauGAN fills the object textures.

Aiming to further push the frontier of the practicability of the image editing techniques, we study a practical use case of image editing, *i.e.*, *conditional image repainting*, which is achieved through the collaboration between users and the trained model. By "repainting", we mean that the model is trained to repaint an

* Corresponding author.

Fig. 1. From left to right, we show the input image, input semantic parsing mask (the white indicates the unaltered regions), generated content, and composited image. The input color description is "The grass is green and yellow, the pavement is gray, and the sky is white".

area of an existing image with some visual content. By "conditional", we mean that the visual content to be repainted is generated by the model conditioned on several user inputs. As such, the conditional image repainting can be formulated into two sub-tasks, *i.e.*, conditional content *generation* and content *compositing*. Figure 1 illustrates our targeted conditional image repainting problem.

Conditional content generation refers to visual synthesis tasks conditioned on user inputs. The user inputs cover *three aspects*, *i.e.*, geometry (shape, pose and semantic labels), colors, and gray-scale textures, and can be roughly divided into two categories, *i.e.*, visually-concrete reference (*e.g.*, reference images and semantic parsing masks) and visually-abstract description (*e.g.*, language and latent code). For example, in [13,22], the geometry is provided by users as semantic parsing masks, and the gray-scale textures altogether with colors as a whole is squeezed into a latent code. To enable higher control flexibility, Li *et al.* [12] attempt to use a separate reference image as user input for each aspect. However, it is sometimes cumbersome to find a desirable reference image, or very expensive to modify a reference image if the off-the-shelf one is not satisfying.

This motivates us to study the conditional content generation based on inputs that are more user-friendly. For the geometry, we follow [13,22] to use semantic parsing masks as input which can be easily manipulated by users. The gray-scale textures are highly correlated with the content class to be generated, so we use the latent code as input for the sake of reducing users' burden. A naive solution of providing the color input is to ask users to select different colors from a palette, and associate them with different parts of the geometry. However, this solution may require extensive and precise operations on the user interface, which is impractical on mobile devices with relatively small screens. Natural language is a user-friendly option for summarizing colors and their distributions. In order to enforce the generated content at each region to reflect its relevant words, the generation model needs to associate the word features with the relevant regions on the image plane. The word features correspond to semantically meaningful words, but as an embryo of the generated content, the region features are semantically ambiguous, thus causing a great challenge for

the cross-modality association. We name such a challenge as the *semantic chasm*. In order to bridge the chasm, we disambiguate the semantics of region features through the semantic parsing mark which can be considered as a cross-modality mediator. By overlaying a semantic parsing mask on the image plane, regions covered by object masks can be associated with words which are relevant to the object class names. Guided by this philosophy, we propose a delicate and plug-n-play SEmantic-BridgE (SEBE) attention mechanism for assisting using language as the input color condition.

To complete the repainting task, the model needs to adjust the contrast and brightness of the generated visual content according to a user-provide image while preserving the semantics of the user-input conditions, and then composites the adjusted content at the user-indicated location of the user-provided image. The assumption is that the composited content should visually be indistinguishable from the innate content of the provided image. Based on this assumption, state-of-the-art image compositing techniques [3,9] train an adversarial discriminator for segmenting the composited content so as to supervise the compositing model (which plays a similar role to a generator). Such an adversarial training poses a latent ceiling of fidelity for the composited content with a rigid value function by restricting that when training the compositing model, under no circumstances, the discriminator should identify the innate content as the composited one. However, there might be cases being mistakenly penalized where the fidelity of the composited content is high enough to confuse the discriminator. Therefore, we propose a *piecewise value function* for the discriminator which applies the proper penalization opportunistically. In order to pave the way for the piecewise value function, we preprocess the input to the discriminator so as to impede the convergence of the discriminator.

We conduct extensive experiments on CUB-200-2011 dataset [21] and COCO-Stuff dataset [2], and show that the proposed SEBE attention mechanism and piecewise value function are beneficial for conditional image repainting.

2 Related Works

Cross-modality attention. Most existing methods [4,15,17,25,26,29] addressed the semantic chasm by attentively estimating the relevance between words and regions. Such an attentive model is trained in a data-driven fashion and the training is partially supervised by a cross-modality retrieval [25] or reconstruction [15] loss. Let $e[i]$ and $h[j]$ denote features of the i-th word and the j-th region, respectively. Their relevance is estimated as an attention weight $\beta[j,i]$:

$$\beta[j,i] = \frac{\exp(s[j,i])}{\sum_{k=1}^{K}\exp(s[j,k])}, \quad s[j,i] = (h[j])^T\phi(e[i]), \tag{1}$$

where $s[j,i]$ is the similarity between $h[j]$ and $e[i]$ which is computed by the dot product, and K is the number of words. $\phi(\cdot)$ is a linear layer for mapping the word features to the domain of the region features. With the estimated

attention weights, in most existing methods [4,15,17,25,29], the words' features are aggregated by their relevance to each region on the image plane: $c[j] = \sum_{i=1}^{N} \beta[j,i]\phi(e[i])$, where $c[j]$ is the aggregated word features (or named context feature vector) for the j-th region. For each region, $c[j]$ is concatenated with $h[j]$, so as to enforce the generated content at each region to reflect its relevant words. Despite great improvements achieved, the cross-modality attention estimation is still challenging because of the semantic ambiguity of regions.

Content compositing has been tackled from different angles, *e.g.*, handcrafted feature matching [20,24], fusion of semantic information [19], image reconstruction [6], *etc.* Recent progress in content compositing derives from the compositing models [3,5,9,18] based on the adversarial training. As a concurrent work, Cong *et al.* [5] train a U-Net [16] based model to fuse the composited content and the innate content of the provided image, and also propose a domain verification discriminator for supervising the compositing model. Compared to [5], [3,9,18] address the compositing problem in a more parametric fashion, which train a model to infer a set of contrast and brightness affine parameters for adjusting the color tone of the composited content. The compositing model is supervised by a segmentation [3,9] or detection [18] based adversarial discriminator.

Conditional normalization is proposed to alleviate the "condition dilution" problem by performing an affine transformation after each normalization operation. The affine parameters, which are inferred through a network from the input condition, are responsible for modulating the activations either element-by-element [13] or channel-by-channel [10]. The element-wise transformation is tailored for the input condition with spatial dimensions, *e.g.*, parsing mask, while the channel-wise one is much more general and not limited to spatial-explicit condition, and thus should be suitable for our gray-scale texture condition, *i.e.*, Gaussian noise vector.

3 Preliminaries

This section revisits recent techniques for addressing our targeted problems, so as to analyze their limitations in detail, and clarify our motivations technically.

3.1 Object-driven attention for content generation

To further resolve the semantic chasm challenge, Li *et al.* [11] proposed the object-driven attention in which with the help of objects' names, the cross-modality attention estimation can be converted to the attention estimation within the same modality. The intuition of the object-driven attention (2) is as follows: *(i)* The attention of a region to a word can be estimated by comparing the name embedding of the object that covers this region and the embedding of this word. If the name of an object to be generated in the image matches with one in a sentence, the word embedding of these two names should be similar thus leading to high and reliable attention weight. *(ii)* The descriptive words of an object should reside around the object name in a sentence, so the feature

of the object name should contain the information of its descriptive words due to the property of the bi-directional LSTM based text encoder [25]. *(iii)* Given the reliable attention weight and the meaningful feature of the object name, it should be more effective in enforcing the reflection of the relevant words on image regions. Similar to (1), the object-driven attention is formulated as:

$$\beta^{\text{obj}}[t,i] = \frac{\exp(s^{\text{obj}}[t,i])}{\sum_{k=1}^{K}\exp(s^{\text{obj}}[t,k])}, \quad s^{\text{obj}}[t,i] = (\hat{e}_{+}[t])^{T}\hat{e}[i], \tag{2}$$

where $\hat{e}[i]$ and $\hat{e}_{+}[t]$ denote the GloVe embedding [14] of the i-th word and the name of the t-th object, respectively. For the j-th region, if it is covered by the t-th object, its object-driven context feature $c^{\text{obj}}[j]$ can be computed similarly to computing $c[j]$ using the cross-modality attention: $c^{\text{obj}}[j] = \sum_{i=1}^{N}\beta^{\text{obj}}[t,i]\phi(e[i])$. If a region is not covered by any object, its object-driven context feature is set to all-zero. For each region, $c[j]$ and $c^{\text{obj}}[j]$ are concatenated with $h[j]$ to enforce the reflection of the relevant words.

Limitation 1 *Despite the rationality of the object-driven attention, it has three weaknesses that we cannot ignore.* (i) *The dot product is not suitable for computing the similarity in the object-driven attention (2), because its output depends on the magnitude of the word embedding vectors. For example, the similarity between any pair of identical word embedding vectors should be high and constant, because their similarity should indicate the "perfect-match". But, the dot product cannot guarantee this.* (ii) *The context feature vectors driven by these two attentions are concatenated with the region features without partiality. When these two attentions are not in consensus for a particular region, it causes extra burdens for the generation model learning to figure how out to use these two attentions in the training stage.* (iii) *The concatenation of two context feature vectors causes large overhead of the runtime memory.*

3.2 Segmentation-based adversarial training for compositing

We study [3,9] to design our compositing model, which employ the segmentation-based adversarial discriminator for training. During training, the adversarial discriminator D learns to identify the composited foreground content by maximizing the value function V_1:

$$\begin{aligned}\max_{D} V_1(D,G) = \mathbb{E}_{y\sim p_{\text{data}}(y)}\xi\big(\log(\mathbf{1}-D(y|\bar{y}))\big) + \mathbb{E}_{y\sim p_{\text{g}}(y)}\xi\big(\log D(y|\bar{y})\big) + \\ \mathbb{E}_{y\sim p_{\text{data}}(y)}\xi\big(\log(\mathbf{1}-D(\bar{y}|y))\big) + \underline{\mathbb{E}_{y\sim p_{\text{g}}(y)}\xi\big(\log(\mathbf{1}-D(\bar{y}|y))\big)},\end{aligned} \tag{3}$$

where p_{g} is a probability distribution defined by the compositing model G, and y *and* $\bar{y}$ *represent the foreground and background content, respectively.* $D(y|\bar{y})$ outputs the probability of y being the composited foreground content conditioned on $\bar{y}$. ξ represents the mean-reduction function for pixels inside a content.

D has two directions with the shared parameters: $D(y|\bar{y})$ and $D(\bar{y}|y)$. Given a fixed G, the optimality of $D(y|\bar{y})$ is proved in [7] to be $D^{*}_{G}(y|\bar{y}) = \frac{p_{\text{data}}(y)}{p_{\text{data}}(y)+p_{\text{g}}(y)}$.

In the supplementary, we prove the optimality of $D(\bar{y}|y)$ to be $D^*_G(\bar{y}|y) = 0$ which has nothing to do with p_{g}, and thus can be regarded as a posterior-collapse state. This is understandable because the amount of real images for training is limited, given enough training steps, $D(\bar{y}|y)$ should be able to memorize the data. At that time, the loss terms related to $D(\bar{y}|y)$ should be invalid.

During the training of G, G minimizes the value function V_2:

$$\min_G V_2(G, D) = \mathbb{E}_{y\sim p_{\mathrm{g}}(y)}\xi\big(\log D(y|\bar{y})\big) + \underline{\mathbb{E}_{y\sim p_{\mathrm{g}}(y)}\xi\big(\log D(\bar{y}|y)\big)}. \tag{4}$$

The *intuition of* V_2(4) is that both the composited foreground content and the innate background content should be identified as the innate content by D, *i.e.*, $D(y|\bar{y}) = 0$ and $D(\bar{y}|y) = 0$, and thus these two contents should be indistinguishable. However, the minimax game shown in V_1(3) and V_2(4) is atypical from the perspective of adversarial training, because the underlined terms in V_1(3) and V_2(4) should be identical for each player in a typical minimax game. The typical value function V_2' can be formed by substituting the underlined term in V_1(3) for that in V_2(4):

$$\min_G V_2'(G, D) = \mathbb{E}_{y\sim p_{\mathrm{g}}(y)}\xi\big(\log D(y|\bar{y})\big) + \underline{\mathbb{E}_{y\sim p_{\mathrm{g}}(y)}\xi\big(\log(\mathbf{1} - D(\bar{y}|y))\big)}. \tag{5}$$

Limitation 2 *Minimizing* V_2'(5) *pushes* G *to evolve toward confusing* D *to identify the innate content as the composited one, i.e.,* $D(\bar{y}|y) = 1$*, which contradicts the intuition of* V_2(4)*. Moreover, as discussed above, it should not be difficult for* $D(\bar{y}|y)$ *to reach its optimality. Thus,* $D(\bar{y}|y)$ *could be reliable for most of the training time, so the gradients deriving from the underlined term in* V_2'(5) *would keep steady no matter how* G *evolves, thus bringing the potential harm to the training. Considering the high reliability of* $D(\bar{y}|y)$*, the underlined term in* V_2(4) *would be kept minimal during the training period no matter how* G *evolves, so this term makes little sense in supervising* G*. Here we come to understand that [3,9] abandon the harm of* V_2'(5) *and embrace the limitation of* V_2(4)*. The limitation of* V_2(4) *stems from the convenience for* $D(\bar{y}|y)$ *in reaching the convergence. Supposing that we can impede the convergence of* $D(\bar{y}|y)$*, the reliability of* $D(\bar{y}|y)$ *should be weakened. In this fashion, the aforementioned intuition of* V_2(4) *may be too strict for the weakened* $D(\bar{y}|y)$*, because there are higher chances that the fidelity of the composited content is high enough to confuse* $D(\bar{y}|y)$ *to mistakenly identified the innate content as the composited one. So,* G *could be excessively penalized by* V_2(4) *given the weakened* $D(\bar{y}|y)$*. In other words,* V_2(4) *poses a latent ceiling of fidelity for the composited content by constraining that the fidelity of the composited content should not be high enough to confuse* $D(\bar{y}|y)$*.*

4 Conditional Image Repainting

In this work, the conditional image repainting is formulated as a generation-compositing setting. In the generation phase, the content generation model G^{cg} accepts three inputs: *(i)* a semantic parsing mask $x^{\mathrm{g}} \in \mathbb{L}^{N_{\mathrm{g}}\times H_1\times W_1}$ for defining

the content geometry, where $\mathbb{L} \in \{0, 1\}$, and N_g, H_1 and W_1 represent the number of object classes, image height and width, respectively; *(ii)* a sentence x^c describing the colors and their distributions on the geometry; *(iii)* a Gaussian noise vector $x^t \sim \mathcal{N}(0, 1)$ encoding the gray-scale textures. Then, G^{cg} maps these inputs to a visual content $\dot{y}$, concluding the generation phase.

$\dot{y}$ can be composited onto a user-provided image $\bar{y}$ at a user-indicated location. In order to make $\dot{y}$ and $\bar{y}$ more harmonious, the content compositing model G^{cc} infers a set of contrast and brightness affine parameters for $\dot{y}$ to be adjusted.

4.1 Semantic-bridge attention for content generation

In order to resolve Limitation 1, we propose the SEmantic-BridgE (SEBE) attention mechanism for content generation. The intuition of SEBE attention is the same as that of the object-driven attention in §3, *i.e.*, bridging the semantic chasm between the word features and the region features through the semantics of the geometry that covers those regions. SEBE achieves improvements over the object-driven attention in three aspects, *i.e.*, attention estimation, attention selection, and computational overhead.

Trustier attention estimation. Let $\hat{e}[i]$ and $\hat{e}_+[t]$ denote the GloVe embedding of the i-th word in the input sentence, and the name of the t-th object to be generated on the image plane. For any region, if it is covered by the geometry of the t-th object on the image plane, then its region feature h_j can be represented by $\hat{e}_+[t]$. Given the GloVe embedding of any word $\hat{e}[i]$, the cross-modality attention estimation between words and regions can thus be formulated as the attention estimation of the same modality (viz. the space of GloVe embedding).

The object-driven attention computes the similarity between $\hat{e}[i]$ and $\hat{e}_+[t]$ as their dot product in (2). However, as discussed in §3, the dot product operation is not suitable for computing the similarity for embedding of the same modality, because its output depends on the magnitude of the word embedding vectors. Therefore, supposing that the j-th region is covered by the t-th object, we formulate the attention estimation of SEBE as

$$\beta^{\text{SEBE}}[j, i] = \frac{s^{\text{SEBE}}[j, i] + 1}{\sum_{k=1}^{K}(s^{\text{SEBE}}[j, k] + 1)}, \quad s^{\text{SEBE}}[j, i] = \frac{(\hat{e}_+[t])^T \hat{e}[i]}{\|\hat{e}_+[t]\| \|\hat{e}[i]\|}, \tag{6}$$

where $s^{\text{SEBE}}[j, k] \in [-1, 1]$ is the cosine similarity between $\hat{e}_+[t]$ and $\hat{e}[i]$, viz. the dot product operation normalized the magnitude of two vectors. The attention weight $\beta^{\text{SEBE}}[j, i] \in [0, 1]$ is computed by shifting $s^{\text{SEBE}}[j, i]$ to be non-negative and normalizing the shifted value by L1 Norm of attention weights of the t-th object for all words. Thus, if $\hat{e}_+[t]$ and $\hat{e}[i]$ are embedding vectors of the same word (viz. object name), $s^{\text{SEBE}}[j, i]$ is able to stay constantly as 1. The coverage relationship between regions and objects is specified in the input semantic parsing mask x^g. If a region is not covered by any objects, we set its SEBE attention weight to be zero.

Smarter attention selection. As discussed in §3, the object-driven attention and the cross-modality are used to compute their respective context feature

vector for each region, and these two types of context feature vectors are concatenated with the region features for model's further processing. Such an impartial treatment of these two attentions shift the duty of selecting which attention to trust from the input end to model, which causes extra learning burdens for model. Therefore, in SEBE, we address the attention selection at the input end with the philosophy of "loudness is persuasive". For the j-th region, we formulate its context feature vector computation based on two types of attentions as

$$c^{\text{SEBE}}[j] = \sum_{i=1}^{N} \max(\beta^{\text{SEBE}}[j,i], \beta[j,i])\phi(e[i]). \tag{7}$$

If the j-th region is covered by an object, $\beta^{\text{SEBE}}[j,i]$ is computed as in (6), and otherwise is set to zero. $\beta[j,i]$ is a cross-modality attention weight which is computed as in (1). $\phi(\cdot)$ is a linear layer as introduced below (1). Consequently, there is only one context feature vector, *i.e.*, $c^{\text{SEBE}}[j]$, for each region.

Lighter computational overhead. Reducing a half of the context feature vectors reduce the runtime memory overhead significantly because the concatenated features are supposed to be fed in a series of residual blocks which need to keep the feature dimensions the same throughout the process.

4.2 Piecewise value function for content compositing

In order to resolve Limitation 2, we propose a piecewise value function for content compositing. Specifically, we modify V_2(4) by replacing its rigid underlined term with a piecewise term:

$$\mathcal{S}(D(\bar{y}|y)) = \begin{cases} \mathbb{E}_{y\sim p_{\text{g}}(y)}\xi\big(\log D(\bar{y}|y)\big), & \text{if } D(\bar{y}|y) < 0.5 \\ \mathbb{E}_{y\sim p_{\text{g}}(y)}\xi\big(\log(1 - D(\bar{y}|y))\big), & \text{otherwise} \end{cases} \tag{8}$$

The philosophies behind (8) are two-fold: *(i)* when $D(\bar{y}|y) < 0.5$, it retains the intuition of V_2(4) in §3.2 that the composited content should be indistinguishable from the innate one, *i.e.*, $D(y|\bar{y}) = 0$ and $D(\bar{y}|y) = 0$; *(ii)* otherwise, we consider that $y \sim p_{\text{g}}(y)$ has successfully confused $D(\bar{y}|y)$, so we urge G^{cc} to evolve along the direction of making $D(\bar{y}|y)$ more confused (viz. producing composited content of higher fidelity) by encouraging $D(\bar{y}|y)$ to approximate 1.

Considering the gradient ineffectiveness issue of the underlined term in V_2'(5) as justified in §3.2, combining these two philosophies in (8) help further improve G^{cc} while circumventing the weakness of V_2'(5). By replacing the underlined term in V_2(4) with (8), we have the piecewise value function V_3 as

$$\min_G V_3(G, D) = \mathbb{E}_{y\sim p_{\text{g}}(y)}\xi\big(\log D(y|\bar{y})\big) + \mathcal{S}(D(\bar{y}|y)). \tag{9}$$

V_1(3) and V_3(9) compose a two-player minimax game for G^{cc} and its adversarial discriminator D^{cc}. As discussed in §3.2, it is convenient for $D^{\text{cc}}(y|\bar{y})$ to reach the convergence. Thus, in order to impede the convergence, we propose a novel but delicate strategy to improve D^{cc} based on the design in [9], which will be introduced in the network architecture design §4.3.

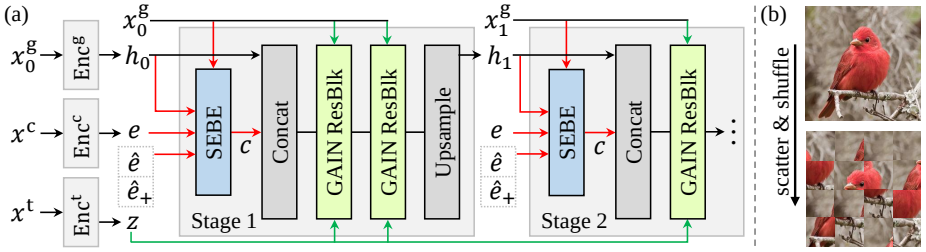

Fig. 2. *(a)* The multistage conditional content generator G^{cg} takes as input three conditions, *i.e.*, geometry x_0^{g}, color x^{c}, and gray-scale texture x^{t} for the first stage. x^{c} and x^{t} are encoded by $\mathrm{Enc}^{\mathrm{c}}$ and $\mathrm{Enc}^{\mathrm{t}}$ to form e and z for condition injection. x_0^{g} is encoded by $\mathrm{Enc}^{\mathrm{g}}$ to form the initial region features h_0. The resolution of x_i^{g} is doubled with the increment of i, while the specified objects remain the same. $\hat{e}$ and $\hat{e}_+$ denote the GloVe embedding of words in x^{c} and names of objects specified in x_i^{g}. Red arrows indicate that regions are associated with their relevant words through SEBE, and the word features e are aggregated by these associations to form the context features c. Green arrows indicate the injection of z under the guidance of x_i^{g}. *(b)* The input preparation process, "scatter & shuffle", for the compositing discriminator D^{cc}.

4.3 Network architecture design

Conditional content generator G^{cg} is multistage, in which every stage shares the architecture that stacks two residual blocks as shown in Fig. 2(a). In each block, we employ the Gated Adaptive Instance Normalization (GAIN) [23] for injecting the texture code (*i.e.*, , Gaussian noise), which is proved to overcome the shortcomings of AdaIN in injecting the texture code for the non-rigid geometry. h_{i+1} represents the intermediate features from the previous stage, which can be fed into a conv-tanh block to generate an image $\dot{y}_{i+1}$ (omitted in Fig. 2(a)) The resolution of $\dot{y}_{i+1}$ is doubled with the increment of i. Red arrows indicate the inputs to and output from SEBE, which supplement the introduction in §4.1.

Compositing model G^{cc}. Its design follows the recently proposed pixel transformation method [3] using a neural network to infer the contrast and brightness transformation parameters given both the composited and the innate contents.

Compositing discriminator D^{cc} is not designed following the segmentation-based discriminator in [3], because as discussed in §3.2, it is not difficult for the segmentation-based discriminator to reach convergence because the discriminator can memorize data after some epochs. This could weaken the effectiveness of the proposed piecewise value function in §4.2. Therefore, in order to impede the convergence, we exponentially increase the amount of real training images by reorganizing images using a simple "scatter & shuffle" strategy (which is also applied to the composited images) as shown in Fig. 2(b). This makes D^{cc} very hard to go through all reorganized images multiple times to memorize data during training. Then, the real/fake labels are no longer distributed by pixels but by patches. Therefore, we build D^{cc} as a simple CNN for patch-wise classification.

4.4 Learning

We train the proposed generation-compositing framework by solving a minimax optimization problem given by

$$\min_{D^{\mathrm{cg}},D^{\mathrm{cc}}} \max_{G^{\mathrm{cg}},G^{\mathrm{cc}}} \mathcal{L}_{\mathrm{cg}}(D^{\mathrm{cg}},G^{\mathrm{cg}},G^{\mathrm{cc}}) + \lambda_1\mathcal{L}_{\mathrm{cm}}(G^{\mathrm{cg}}) + \lambda_2\mathcal{L}_{\mathrm{cc}}(D^{\mathrm{cc}},G^{\mathrm{cc}}) + \lambda_3\mathcal{L}_{\mathrm{r}}(G^{\mathrm{cc}}), \tag{10}$$

where $\mathcal{L}_{\mathrm{cg}}$, $\mathcal{L}_{\mathrm{cm}}$, $\mathcal{L}_{\mathrm{cc}}$, and $\mathcal{L}_{\mathrm{r}}$ are the GAN loss for the overall image quality, DAMSM loss [25] for the color condition, GAN loss and a regularization loss [3] for the compositing performance, respectively. D^{cg} is a set of joint-conditional-unconditional patch discriminators [11] for each stage of G^{cg}. $\mathcal{L}_{\mathrm{cg}}$, $\mathcal{L}_{\mathrm{cm}}$, and $\mathcal{L}_{\mathrm{r}}$ are borrowed from [23]. $\mathcal{L}_{\mathrm{cc}}$ is defined by $V_1(3)$ for training D^{cc} and by $V_3(9)$ for training G^{cc}. Let y denote a composited image. Applying D^{cc} to y, we have $\mathbf{p} = D^{\mathrm{cc}}(y)$, where $\mathbf{p} = \{p_1, \ldots, p_i, \ldots, p_N\}$ is a set of probabilities with each indicating how likely a patch belongs to the composited content. We define two index sets for each y, *i.e.*, the composited patch index set $\mathbb{I}^{\mathrm{ci}}$ and the innate patch index set $\mathbb{I}^{\mathrm{ii}}$. For training D^{cc}, $\mathcal{L}_{\mathrm{cc}}$ is a typical classification loss. For training G^{cc}, $\mathcal{L}_{\mathrm{cc}}$ is defined as

$$\mathcal{L}_{\mathrm{cc}}(G^{\mathrm{cc}},D^{\mathrm{cc}}) = -\frac{1}{N}\Big(\sum_{i\in\mathbb{I}^{\mathrm{ci}}}\log(1-p_i) + \sum_{i\in\mathbb{I}^{\mathrm{ii}}}\log\psi(p_i)\Big), \tag{11}$$

where ψ corresponds to the piecewise term in (8), which is defined as $\psi(p_i) = 1 - p_i$, if $p_i < 0.5$; otherwise, $\psi(p_i) = p_i$.

Based on the experiments on a held-out validation set, we set the hyperparameters in this section as: $\lambda_1 = 20$, $\lambda_2 = 0.03$, and $\lambda_3 = 1.0$.

5 Experiments

Datasets. We use CUB-200-2011 [21] and COCO-Stuff [2] for evaluation. For CUB, we annotate bird images with parsing masks, and follow [25] for data processing. For COCO, we select 9 most common stuff classes to use, including sky, grass, road, clouds, pavement, dirt, sand, bush, and sea. We annotate 10 captions per image, and use 6.2K images for training and 1.4K for test.

Quantitative evaluation metrics. Three evaluation metrics are used: *(i)* we use the *Fréchet inception distance* (FID) [8] score to evaluate the general image quality. *(ii)* Following [25], we use R-precision to evaluate whether the generated image is well conditioned on the input color description. More specifically, given a generated image y conditioned on the input sentence x^{c} and 9 randomly sampled sentences, we rank these 10 sentences by the pre-trained DAMSM model. If the ground truth sentence x^{c} is ranked the highest, we count this a success retrieval. We perform this retrieval task on all generated images and calculate the percentage of success retrievals as the R-precision score. *(iii)* For measuring the compositing quality, we follow [18] to use the M-score which is the output by a manipulation detection model [28]. The higher M-score, the higher possibility that an image has been manipulated. For each compared method, we randomly pick 500 generated images to calculate the average M-score.

Table 1. The quantitative experiments. ↑ (↓) means the higher (lower), the better. The best performances are highlighted in **bold**. The compared baselines are divided into six categories: Rows 1-2 for generation, and Rows 3-4 for compositing.

Category	Methods	CUB-200-2011			COCO-Stuff		
		FID ↓	R-prcn (%) ↑	M-score ↓	FID ↓	R-prcn (%) ↑	M-score ↓
Attn Est	SEBE w/ DotPrdct	12.6	98.72	32.86	19.3	59.14	81.62
	SEBE w/o CrsMod	12.68	98.75	35.68	19.43	58.73	72.11
	CrsMod	12.21	98.7	31.38	19.06	60.48	78.54
Attn Sel	SEBE w/o SAS	12.31	98.91	34.18	20.03	61.13	81.14
Seg	Seg V_2(4)	12.12	98.74	28.1	19.11	60.36	75.16
	Seg V_3(9)	12.25	98.99	33.53	19.25	57.34	76.35
Cls	Cls V_2(4)	12.39	98.81	27.17	19.23	57.4	74.74
	Cls V_2'(5)	12.44	**99.18**	34.95	19	57.65	81.22
	Cls V_3(9) w/o Pwise	12.62	98.99	26.65	18.96	58.53	76.21
Ours	SEBE-GAIN-Cls-V_3(9)	**12.08**	98.94	**24.6**	**18.91**	**65.43**	**67.96**

5.1 Content generation

We evaluate two aspects of our method for content generation, *i.e.*, attention estimation (abbr. Attn Est) and attention selection (abbr. Attn Sel) in §4.1. For each aspect, we create some baselines either by disabling modules of our model or adapting the existing techniques to our task. The quantitative and qualitative comparison are shown in Table 1 and left side of Fig. 3, respectively. Note that our full-version method outperforms the compared baselines in most metrics on both datasets, which demonstrates the effectiveness of our proposed modules quantitatively, so we focus on analyzing the qualitative results in the following.
Attention estimation. We create three baselines for this aspect: *(i) SEBE w/ DotPrdct* using the object-driven attention (2) [11] for estimation, and keeping the attention selection (7); *(ii) SEBE w/o CrsMod* by disabling the cross-modality attention in (7); *(iii) CrsMod* using the cross-modality attention estimation (1) [25], and by disabling the SEBE attention in (7). Figure 3 shows that SEBE is effective in controlling the color artifacts such as Column 3 and 5 for *SEBE w/ DotPrdct*, Column 2 and 4 for *SEBE w/o CrsMod*, and Column 4 for *CrsMod*. Please refer to §2 and Limitation 1 for shortcomings of these baselines.
Attention selection. We create *SEBE w/o SAS* by disabling the attention selection (7). As justified in §4.1, this module should be able to shift model's burden in selecting attention to the input end. The texture artifacts in Column 1 and the color artifacts in Column 4 of Fig. 3 are obvious for *SEBE w/o SAS*.

5.2 Content compositing

For content compositing, we evaluate the influences of discriminator design, *i.e.*, the full-image and segmentation based discriminator [3] (abbr. Seg) vs. the shuffled-patches and classification based one (abbr. Cls), and the influences for different value functions including V_2(4), V_2'(5), V_3(9) w/o the piecewise (abbr. Pwise) term (8), and V_3(9).
Seg vs. Cls. When the Seg and Cls discriminators are evaluated with the same value functions, *i.e.*, V_2(4) and V_3(9), Cls outperforms Seg in terms of M-score on

Fig. 3. Qualitative comparison for content generation (left) and compositing (right). Best viewed on the computer, in color and zoomed-in. Input color descriptions: (1) "This bird is black and yellow in color, and has an orange beak." (2) "The bird has a white belly and chest with gray wings and tail and black striped head." (3) "This bird has a white belly and breast, with a long orange hooked bill." (4) "The pavement is brown and gray." (5) "The grass in the picture is brown and green."

both datasets, which demonstrate the effectiveness of our discriminator design in §4.3. In the supplementary, we show that the Seg discriminator reaches the convergence much faster than the Cls one. This implies that Seg discriminator and the compositing model reach the Nash equilibrium faster, which prevents further improving the compositing model in the training.

Value functions. In Table 1, *Cls* $V_3(9)$ outperforms *Cls* $V_2(4)$ significantly in terms of M-score, while this is not the case for the Seg. This phenomenon echos our analysis in Limitation 2 that we need to first impede the early convergence of the discriminator before the value function is improved. It also proves the necessity of modifying the discriminator as in §4.3. In addition, *Cls* $V_2'(5)$ yields much worse M-score than our method, which provides some evidence for our discussions in Limitation 2 about the weakness of $V_2'(5)$. Here we come to know that both $V_2(4)$ and $V_2'(5)$ cannot achieve prominent compositing performance alone. In fact, our proposed $V_3(9)$ implements a mechanism for choosing to apply $V_2(4)$ or $V_2'(5)$ at the right time. To further study the impact of the proposed piecewise term (8), we simply remove it from $V_3(9)$ to see the results. This means that the discriminator only cares about the composited content but ignores the innate one when training the compositing model. From Table 1, we see that the influences are more obvious on COCO-Stuff than on CUB. This might be because that COCO-Stuff is more challenging than CUB.

Assumption in Limitation 2 is that the fidelity of the composited content is high enough to confuse $D(\bar{y}|y)$ to mistakenly identified the innate content as the

Fig. 4. Composited images in which the innate content is correctly identified or misidentified by D^{cc} are shown on the left and right, respectively.

Fig. 5. Comparison with [27] for mask based object removal. Masks are placed in the lower-left corner of the edited images, where the gray indicates regions to be filled. From left to right, we show the real image, the result of [27], and our result.

composited one, which is also the motivation for us to improve the value function in §4.2. Therefore, we select a discriminator in-between the training process, and visualize in Fig. 4 the randomly-sampled composited images in which the innate content is correctly identified or misidentified by our compositing discriminator D^{cc}. We can see that the compositing fidelity of the misidentified images is generally higher than that of the correctly identified images, which provides evidence supporting the assumption.

5.3 Qualitative study

Object removal is always considered as a task for image inpainting. However, as shown in Fig. 5, the recently proposed [27] cannot handle cases with complicated background well. Surprisingly, our method can successfully remove the objects despite the cost of substituting the generated content for a large portion of content in the original images, *e.g.*, sky and lawn in Fig. 5. In the supplementary, we provide more analyses about this task and the limitations of our work in handling this task, and indicate the future direction of our research.
Iterative image editing in the wild is shown on the right of Fig. 6. From the real image in Column 5 to the final editing results in Column 8, the whole scenes look quite different, which demonstrates the robustness and flexibility of our method.

6 Conclusion

Targeting at a relatively new and practical task, conditional image repainting, we propose two novel and delicate modules for addressing the weaknesses of the

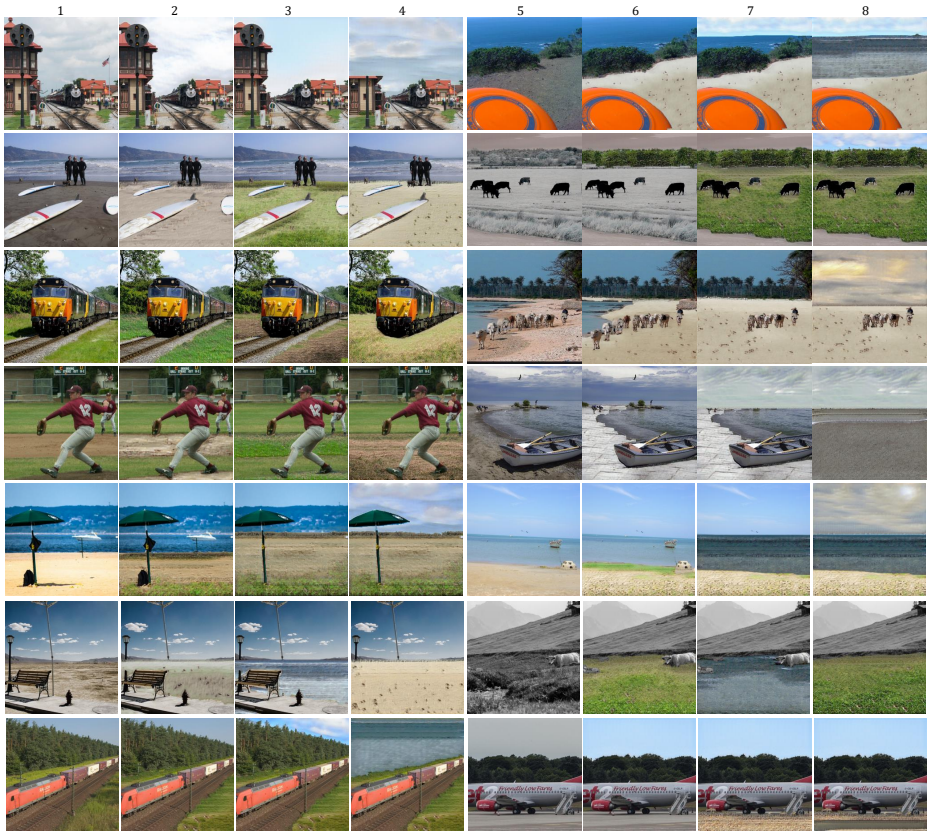

Fig. 6. Image editing in the wild. Column 1 and Column 5 show the real images. Columns 2-4 show the alternation of input conditions. Columns 6-8 show the iterative editing. See the supplementary for detailed input conditions for producing these images.

existing component technologies, *i.e.*, semantic-bridge attention mechanism for assisting using languages as conditional input, and a piecewise value function to improve the adversarial training of the compositing model. We observe favorable performance with both quantitative and qualitative results, and also explore several interesting potential application scenarios of the proposed techniques.

Acknowledgements

PKU affiliated authors are supported by National Natural Science Foundation of China under Grant No. 61872012, National Key R&D Program of China (2019YFF0302902), and Beijing Academy of Artificial Intelligence (BAAI).

References

1. FaceApp, https://www.faceapp.com/ 1
2. Caesar, H., Uijlings, J.R.R., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: CVPR (2018) 3, 10
3. Chen, B., Kae, A.: Toward realistic image compositing with adversarial learning. In: CVPR (2019) 3, 4, 5, 6, 9, 10, 11
4. Chen, X., Qing, L., He, X., Luo, X., Xu, Y.: FTGAN: A fully-trained generative adversarial networks for text to face generation. CoRR **abs/1904.05729** (2019) 3, 4
5. Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W., Zhang, L.: Deep image harmonization via domain verification. CoRR **abs/1911.13239** (2019) 4
6. Cun, X., Pun, C.: Improving the harmony of the composite image by spatial-separated attention module. CoRR **abs/1907.06406** (2019) 4
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: NIPS (2014) 5
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a nash equilibrium. NIPS (2017) 10
9. Huang, H., Xu, S., Cai, J., Liu, W., Hu, S.: Temporally coherent video harmonization using adversarial networks. TIP (2020) 3, 4, 5, 6, 8
10. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019) 4, 17, 18
11. Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., Gao, J.: Object-driven text-to-image synthesis via adversarial training. In: CVPR (2019) 4, 10, 11, 19
12. Li, Y., Singh, K.K., Ojha, U., Lee, Y.J.: Mixnmatch: Multifactor disentanglement and encoding for conditional image generation. CoRR **abs/1911.11758** (2019) 2
13. Park, T., Liu, M., Wang, T., Zhu, J.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019) 1, 2, 4, 17, 18, 21, 24
14. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP (2014) 5
15. Qiao, T., Zhang, J., Xu, D., Tao, D.: Mirrorgan: Learning text-to-image generation by redescription. In: CVPR (2019) 3, 4
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015) 4
17. Tan, H., Liu, X., Li, X., Zhang, Y., Yin, B.: Semantics-enhanced adversarial nets for text-to-image synthesis. In: ICCV (2019) 3, 4
18. Tripathi, S., Chandra, S., Agrawal, A., Tyagi, A., Rehg, J.M., Chari, V.: Learning to generate synthetic data via compositing. In: CVPR (2019) 4, 10
19. Tsai, Y., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.: Deep image harmonization. In: CVPR (2017) 4
20. Tsai, Y., Shen, X., Lin, Z., Sunkavalli, K., Yang, M.: Sky is not the limit: semantic-aware sky replacement. TOG (2016) 4
21. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 3, 10
22. Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR (2018) 2

23. Weng, S., Li, W., Li, D., Jin, H., Shi, B.: Misc: Multi-condition injection and spatially-adaptive compositing for conditional person image synthesis. In: CVPR (2020) 9, 10
24. Wu, H., Zheng, S., Zhang, J., Huang, K.: GP-GAN: towards realistic high-resolution image blending. In: ACM MM (2019) 4
25. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: CVPR (2018) 3, 4, 5, 10, 11, 17, 19
26. Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., Shao, J.: Semantics disentangling for text-to-image generation. In: CVPR (2019) 3
27. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: ICCV (2019) 13, 21, 30, 31
28. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Learning rich features for image manipulation detection. In: CVPR (2018) 10
29. Zhu, M., Pan, P., Chen, W., Yang, Y.: DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In: CVPR (2019) 3, 4