Grounded, Controllable and Debiased Image Completion with Lexical Semantics

Shengyu Zhang[1*], Tan Jiang[1*], Qinghao Huang[2], Ziqi Tan[1], Kun Kuang[1†], Zhou Zhao[1†], Siliang Tang[1], Jin Yu[2], Hongxia Yang[2], Yi Yang[1], Fei Wu[1†]

[1] Zhejiang University [2] Alibaba Group

{sy_zhang, jiangtan, tanziqi, kunkuang, zhaozhou, siliang, yangyics, wufei}@zju.edu.cn

{kola.yu, yang.yhx}@alibaba-inc.com, wfnuser@alumni.sjtu.edu.cn

Abstract

In this paper, we present an approach, namely Lexical Semantic Image Completion (LSIC) [1], that may have potential applications in art, design, and heritage conservation, among several others. Existing image completion procedure is highly subjective by considering only visual context, which may trigger unpredictable results which are plausible but not faithful to a grounded knowledge. To permit both grounded and controllable completion process, we advocate generating results faithful to both visual and lexical semantic context, i.e.*, the description of leaving holes or blank regions in the image (*e.g.*, hole description). One major challenge for LSIC comes from modeling and aligning the structure of visual-semantic context and translating across different modalities. We devise multi-grained reasoning blocks to address this challenge. Another challenge relates to the unimodal biases, which occurs when the model generates plausible results without using the textual description. We devise an unsupervised unpaired-creation learning path that explicitly performs counterfactual thinking,* i.e.*, what the complete image would be if given an unpaired text description to the incomplete image. A cycle consistency loss is devised to guarantee counterfactual faithfulness. We conduct extensive quantitative and qualitative experiments that reveal the strengths of LSIC in being grounded, controllable, and debiased.*

1. Introduction

The recent progress in deep neural networks has shown high capability in image completion[10]. However, on the one hand, these techniques tend to generate blurry regions and artifacts [8], especially when the hole is rather large, due to lack of information of foreground objects[7]. On the other hand, the subjective nature [10] of image completion may lead to results that are visually authentic but not faithful to a grounded truth (*i.e.*, factual cues or attribute information). The *grounded* and *controllable* image completion can be a fundamental requirement in many real-world scenarios. To bridge these gaps, we propose an approach named *Lexical Semantic Image Completion (LSIC)*. The completion results are conditioned not only on the structural continuity and visual semantic but also on the lexical semantic concepts within natural language descriptions.

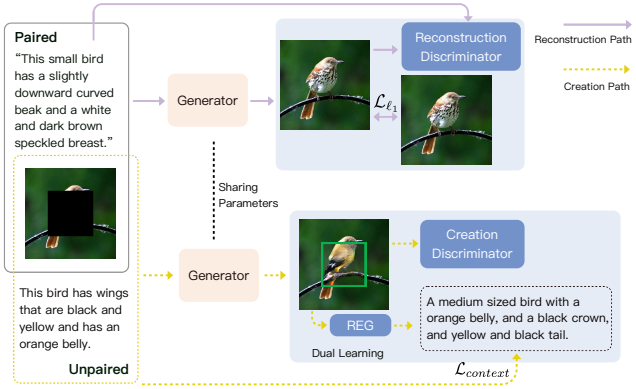

Figure 1. Overall schema of our model, which mainly comprises a progressive reasoning generator and two learning paths.

One major challenge of LSIC is the sheer difficulty to model both the visual semantic structure within the unmasked image and the lexical semantic structure within the sentence and to learn the aligned relationship between them. To address this challenge, we propose first to perform coarse-grained reasoning to depict rough shapes and colors and refine it *progressively* by performing fine-grained reasoning, which is realized by coarse-grained reasoning block (CGR) and fine-grained reasoning block (FGR) in our model. Another challenge regards collecting dataset containing multiple text conditions per masked image, which is often prohibitively expensive to acquire or even unavailable. The annotated sentences for one image are often se-

[1] Please refer to the full version of this paper [9] for better clarity.
[*] These authors contributed equally to this work.
[†] Corresponding Authors.

mantically equivalent in existing datasets. The only-one-text may lead to *unimodal biases*, which is common in the VQA task [4]. In other words, the completion result may be mostly conditioned on the unmasked regions while disregarding the text information, resulting in the loss of controllable generation ability or suffering a performance drop on the test dataset. To this end, we consolidate the idea of Dual Learning [2] and devise an *unpaired-creation* training path to guarantee the faithfulness of counterfactual thinking, *i.e.*, what the complete image would be if given an unpaired text description to the incomplete image. As a result, we aim to achieve grounded, controllable, and debiased image completion. Extensive experiments on multiple datasets with SOTA text-guided image manipulation methods demonstrate the effectiveness of LSIC.

2. Generator

Let I, I_m, and $\hat{I}$ denote the original image, the partially masked image, and the generated image, respectively. t, $\bar{t}$ and $\hat{t}$ denote the paired text to I, an unpaired text which is randomly sampled and the re-generated text by REG, separately. The generator takes text t or $\bar{t}$ and the masked image I_m as inputs. We propose to explicitly model the lexical semantic structure. by transforming text t or $\bar{t}$ into graph representation g or $\bar{g}$ by their grammar dependencies. Since the semantic relation between nodes cannot be simply evaluated by the similarity of word embedding vectors, we initialize all the edge weights by default 1. The masked image is encoded by two standard resnet blocks into an image code $r_0 = \{r_i\}_{i=1,...,N_{0,r}}$, where $N_{0,r}$ is the number of initial visual region features. To perform structure alignment and translation both globally and locally, we devise two novel reasoning blocks, termed coarse-grained reasoning block (CGR) and fine-grained reasoning block (FGR).

Coarse-Grained Reasoning block Starting from the initial image code r_0 and the initial semantic graph $g_0 = (V_0, E_0)$. The CGR firstly performs visual structure reasoning by the resnet block and obtains $c_1 = \{c_i\}_{i=1,...,N_{1,r}}$. For lexical semantic structure modeling, CGR employs the Graph Convolution Network (GCN) to reason along the grammar connection between words and thus generates features $V_1 = \{v_{1,j}\}_{j=1,...,N_v}$ with the semantic relationship. Then, we obtain the high-level semantic concepts by pooling the graph into a global representation $v_{1,*}$. This process can be formulated as:

$$V_1 = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}V_0\mathbf{\Theta}, \tag{1}$$

$$v_{1,*} = Pool(V_1) = 1/N_v \sum_j v_{1,j} \tag{2}$$

where $\mathbf{\Theta}$ is the weight of graph convolutions and $\mathbf{A}$ is the adjacency matrix with inserted self-connections. $\mathbf{D}$ denotes the diagonal degree matrix for $\mathbf{A}$ The $Pool$ function used in our model is mean-pooling. We then filter relevant content using a *gated fusion* function. Given the image region feature $r_{1,i}$ and graph representation $v_{1,*}$, the gated fusion function performs the following operations:

$$\alpha_i = \sigma(W_{1,a}[c_{1,i}, v_{1,*}]) \tag{3}$$

$$r_{1,i} = \alpha_i * W_{1,r}c_{1,i} + (1-\alpha_i) * W_{1,g}v_{1,*} \tag{4}$$

where σ is the sigmoid function. $[.,.]$ denotes the concatenate operation. $W_{1,a}$ and $W_{1,r}$ and $W_{1,g}$ are linear transformations.

Fine-Grained Reasoning block To further capture fine-grained detail like texture and patterns, we design the fine-grained reasoning block using node-level attention. In the τth iteration, FGR takes the nodes features $V_{\tau-1}$ and image features $r_{\tau-1}$ from previous $(\tau-1)th$ reasoning block as input. Similarly, FGR performs $c_\tau = ResBlock(r_{\tau-1})$ and $V_\tau = GCN(V_{\tau-1}, E_{\tau-1})$ for visual-semantic structure modeling. Different from CGR, FGR builds the gated fusion function as follows:

$$c_{\tau,*} = 1/N_{\tau,r} \sum_i c_{\tau,i} \tag{5}$$

$$\beta_{\tau,j} = \sigma(W_{\tau,a}[c_{\tau,*}, v_{\tau,j}]) \tag{6}$$

$$o_{\tau,j} = \beta_{\tau,j} * W_{\tau,r}c_{\tau,*} + (1-\beta_{\tau,j}) * W_{\tau,g}v_{\tau,j} \tag{7}$$

Given the fused features $o_\tau = \{o_{\tau,j}\}_{j=1,...,N_v}$ and image features c_τ, we apply attention mechanism to perform local visual-semantic reasoning on salient and reusable visual patterns as well as meaningful semantic concepts. For i_{th} image region feature, we compute the lexical-semantic-aware visual features as:

$$\epsilon_{\tau,i,j} = \frac{\exp(f(c_{\tau,i}, o_{\tau,j}))}{\sum_k \exp(f(c_{\tau,i}, o_{\tau,k}))} \tag{8}$$

$$r_{\tau,i} = c_{\tau,i} + \sum_{j=1}^{N_v} \epsilon_{\tau,i,j} W_{\tau,m} o_{\tau,j} \tag{9}$$

where function f computes the joint-space similarity of $c_{\tau,i}$ and $o_{\tau,j}$ by $f(c_{\tau,i}, o_{\tau,j}) = (W_{\tau,l}c_{\tau,i})^T(W_{\tau,n}o_{\tau,j})$. $W_{\tau,m}$, $W_{\tau,l}$ and $W_{\tau,n}$ are linear transformations at step τ. The first CGR and following $T-1$ FGRs are stacked sequentially and form the multi-grained progressive generation process. Our generator incorporates multiple output layers to generate multi-scale images hierarchically.

3. Discriminator and Two Learning Paths

As shown in Fig. 1, our framework comprises two parallel training paths, *i.e.*, the supervised *reconstruction* path, and the unsupervised *creation* path.

Reconstruction Path The reconstruction path follows the conventional pipeline, which takes the masked image I_m and the paired textual description t as input. We train it adversarially using a conditional discriminator D_R. We add hierarchical ℓ_1 losses in different scales. During training, the loss function introduced by this path can be defined as:

$$\mathcal{L}_G^R = \underbrace{-\lambda_{adv}\mathbb{E}_{\hat{I}\sim p_G}\log D_R(\hat{I}, v_{0,*})}_{\text{conditional adversarial loss}} + \underbrace{\lambda_{l_1}\mathbb{E}_{\hat{I}\sim p_G}||I-\hat{I}||_1}_{\ell_1 \text{ loss}} \tag{10}$$

$$\mathcal{L}_D^R = -\mathbb{E}_{I\sim p_{data}}\log D_R(I, v_{0,*}) - \mathbb{E}_{\hat{I}\sim p_G}\log\left(1 - D_R(\hat{I}, v_{0,*})\right) \tag{11}$$

where $v_{0,*}$ denotes the initial global representation of semantic graph, which is obtained by $v_{0,*} = Pool(V_0)$. The pooling method used in our paper is mean pooling. We here omit the multi-scale ℓ_1 losses for brevity.

Creation Path To reduce unimodal biases and enhance controllable image completion, we propose an unsupervised creation path via dual learning. The creation path takes the unpaired textual description and masked image as input. Since there is no ground-truth image, we employ an unconditional discriminator to guarantee the visual plausibility. We incorporate a referring expression generator to re-generate the description with the unmasked area as context. The cross-entropy loss between the re-generated words and the input tokens penalizes the inconsistency between the completion area and semantic-visual context, *i.e*. the input text and unmasked area. Therefore, we name it context loss. The loss function introduced by creation path can be formulated as:

$$\mathcal{L}_G^C = \underbrace{-\lambda_{adv}\mathbb{E}_{\hat{I}\sim p_G}\log D_C(\hat{I})}_{\text{adversarial loss}} \underbrace{-\lambda_{ce}\mathbb{E}_{\hat{I}\sim p_G}\sum\nolimits_{\kappa}\log P(\hat{t}_\kappa)}_{\text{context loss}} \tag{12}$$

$$\mathcal{L}_D^C = -\mathbb{E}_{I\sim p_{data}}\log D_C(I) - \mathbb{E}_{\hat{I}\sim p_G}\log\left(1 - D_C(\hat{I})\right) \tag{13}$$

where $\hat{t} = \{\hat{t}_\kappa\}_{\kappa=1,\dots,N_e} = REG(\hat{I})$ is the re-generated sentence and N_e is the length of the sentence.

4. Experiments

We mainly carry out experiments on two fine-grained caption-annotated dataset, CUB and Oxford-102. Since there is no research precisely comparable when this work is conducted, we adopt three state-of-the-art semantic image manipulation methods and make proper adjustments to them. Specifically, we incorporate Dong *et al*. [1], MC-GAN [5], and TAGAN [3].

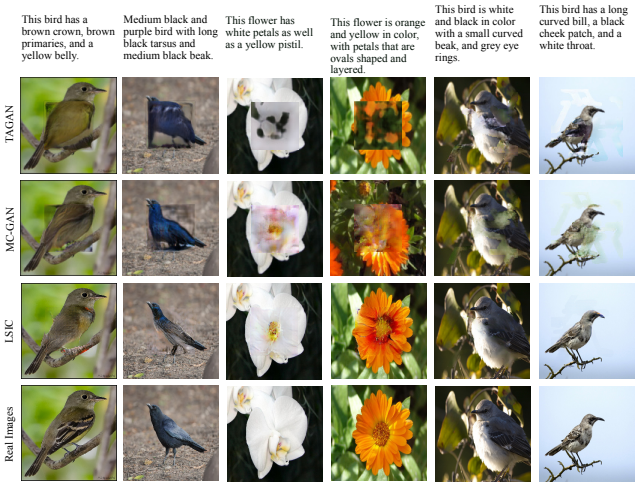

Figure 2. Qualitative comparison of three methods on the CUB and Flower test set with centered-square/irregular masks.

4.1. Quantitative Evaluation

Following the image completion convention, we choose to evaluate the generation results with three numeric metrics, *i.e*., Peak Signal-to-Noise Ratio (PSNR), Total Variation (TV) loss, and Structural Similarity Index (SSIM). We also employ an image generation metric named Inception Score (IS) [6], which measures both the visual quality and generation diversity. We consider both center-square mask and irregular ones in the experiments. Overall, the results (see Table 1) verify the visual authenticity, global consistency of our results as well as the completion variety. We attribute these substantial improvements to the multi-grained reasoning blocks and progressive generation process.

4.2. Qualitative Evaluation

Subjective Analysis Figure 2 displays the completion results produced by our proposed method and three modified comparison models concerning quality assessment. These samples are conditioned on text descriptions and center-masked images on the test dataset. Figure 2 also shows the free-form completion results on the CUB test set. Our method produces images with a coherent structure and vivid grounded details (*i.e*., factual attributes) in most cases, comparing to the Dong *et al*., MC-GAN, and TAGAN.

Generation Variety Figure 3 shows the controllable completion results. We deliberately change the factual attribute (*e.g*. colors and sizes) within the input text. The results show that our model is able to capture the fine-grained semantic concepts and generate completions with corresponding details. These results indicate that the proposed Creation path is a promising direction for better leveraging limited annotations and and reducing unimodal biases in image completion.

Table 1. Quantitative results on the CUB test set and Oxford-102 dataset.

Method	CUB (Center)				CUB (Free-form)				Oxford-102 (Center)			
	PSNR ↑	TV loss ↓	SSIM ↑	IS ↑	PSNR ↑	TV loss ↓	SSIM ↑	IS ↑	PSNR ↑	TV loss ↓	SSIM ↑	IS ↑
Dong *et al.*	14.63	15.08	0.70	2.71	16.75	14.19	0.75	2.96	13.89	16.04	0.71	3.07
TAGAN	19.10	13.42	0.76	4.04	23.96	13.05	0.83	4.11	19.50	13.92	0.78	3.89
MC-GAN	18.23	14.88	0.75	3.98	24.30	13.27	0.82	4.20	19.50	15.69	0.76	4.31
Ours	**19.68**	**10.73**	**0.82**	**4.34**	**26.19**	**10.87**	**0.90**	**5.82**	**19.83**	**10.99**	**0.81**	**5.28**

Table 2. Ablation test of different architectures.

Models	PSNR ↑	TV loss ↓	SSIM ↑	IS ↑
B(aseline)	19.20	12.53	0.801	3.71
+R(easoning blocks)	19.31	11.84	0.810	3.84
+C(reation path)	19.41	11.89	0.813	4.09
+D(ual learning)	**19.68**	**10.73**	**0.819**	**4.34**
1 FGR	19.66	11.24	0.817	4.21
0 FGR	19.41	11.54	0.811	3.86

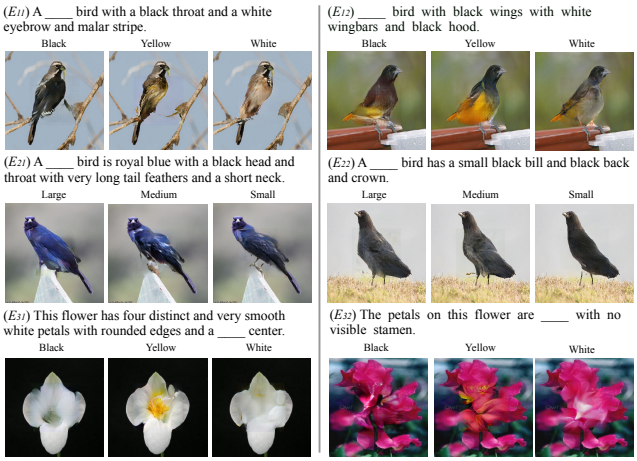

Figure 3. Controllable completion results.

4.3. Ablation Study

To obtain a better understanding of different modules in our model, we surgically remove some components and construct different architectures (see Table 2). B denotes the baseline method, which only takes the masked image as input. R stands for the group of reasoning blocks, which includes CGR and FGR. C is the creation path without referring expression generator, which is named as D, *i.e.* Dual Learning. The results indicate that the elimination of any component would result in a decrease in efficiency. To investigate whether the hierarchically stacked FGR blocks is beneficial, we gradually replace the last FGR block in our model with a plain resnet block, which takes only the visual features from the previous reasoning block as the input, *i.e.*, without considering the semantic concepts. Our model includes two FGR. Therefore, 1 FGR indicates that the last FGR is replaced, and 0 FGR indicates that all FGRs are replaced by plain resnet blocks. The results verify the merit of our hierarchical architecture.

5. Conclusion

In this paper, we propose a framework for the challenging Lexical Semantic Image Completion task, which aims to generate *grounded* results faithful to the textual description and *controllable* results by changing the attributes within the text. Our architecture encapsulates the *coarse-grained reasoning block* and the *fine-grained reasoning block* to progressively complement the broken image. Besides conventional *paired-reconstruction* generation, we incorporate the idea of Dual Learning and devise an *unpaired-creation* path to mitigate the *unimodal biases* problem with counterfactual thinking. The consistent quantitative improvement across various metrics and substantial qualitative results on two fine-grained datasets reveal the efficacy of our proposed method.

References

[1] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic Image Synthesis via Adversarial Learning. In *ICCV*, 2017.

[2] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *NIPS*, 2016.

[3] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. In *NIPS*, 2018.

[4] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. *CoRR*, 2020.

[5] Hyojin Park, Youngjoon Yoo, and Nojun Kwak. MC-GAN: Multi-conditional Generative Adversarial Network for Image Synthesis. In *British Machine Vision Conference*, page 76, 2018.

[6] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016.

[7] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *CVPR*, 2019.

[8] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, pages 5505–5514, 2018.

[9] Shengyu Zhang, Tan Jiang, Qinghao Huang, Ziqi Tan, Zhou Zhao, Siliang Tang, Jin Yu, Hongxia Yang, Yi Yang, and Fei Wu. Grounded and controllable image completion by incorporating lexical semantics. *CoRR*, 2020.

[10] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic Image Completion. In *CVPR*, 2019.