# Language-Guided Global Image Editing via Cross-Modal Cyclic Mechanism

Wentao Jiang [1] Ning Xu [2] Jiayun Wang [1] Chen Gao [1] Jing Shi [3] Zhe Lin [2] Si Liu [1*]
[1] Beihang University [2] Adobe Research [3] University of Rochester
{jiangwentao,wangjiayun12,gaochen,liusi}@buaa.edu.cn, {xnu,zlin}@adobe.com, j.shi@rochester.edu

## Abstract

*Editing an image automatically via a linguistic request can significantly save laborious manual work and is friendly to photography novice. In this paper, we focus on the task of language-guided global image editing. Existing works suffer from imbalanced and insufficient data distribution of real-world datasets and thus fail to understand language requests well. To handle this issue, we propose to create a cycle with our image generator by creating a novel model called Editing Description Network (EDNet) which predicts an editing embedding given a pair of images. Given the cycle, we propose several free augmentation strategies to help our model understand various editing requests given the imbalanced dataset. In addition, two other novel ideas are proposed: an Image-Request Attention (IRA) module which allows our method to edit an image spatial-adaptively when the image requires different editing degree at different regions, as well as a new evaluation metric for this task which is more semantic and reasonable than conventional pixel losses (*e.g. *L1). Extensive experiments on two benchmark datasets demonstrate the effectiveness of our method over existing approaches.*

## 1. Introduction

Image editing has a wide range of applications in many scenarios. With the growth of social media such as Instagram and Facebook, more and more users like to edit their photos before they post them. People would like to use specific photo editing software like Photoshop, but using this kind of professional software is not easy. It may cost novice users a lot of time to learn, the process of editing is also time-consuming. Moreover, as smartphones have become the main user terminal, a way that can automatically edit the images using the voice of users (like Siri or Cortana) will be more user-friendly.

In this paper, we focus on the task of global image editing via linguistic requests: given an input image and a linguistic editing request, the model is required to produce a target image that matches the request, as shown in the last row of Figure 1. Global image editing is to retouch the input image by adjusting the brightness, hue, saturation, contrast, tint, etc. Among the vision and language area, text-guided image manipulation [11, 16, 5] may seem similar to our task, but they are actually different. First, current text-guided image manipulation methods [11, 16] are designed for domain-specific datasets, which are either constrained to images of a single salient object (e.g., bird) or virtual dataset with simple objects, as shown in the first two rows of Figure 1. While in our task, we edit the images from the real-world environment that contains various objects and scenes. Second, text-guided image manipulation methods are designed for template text inputs, which summarize the attributes of the target image (e.g., "A bird with black eye rings and a black bill") instead of describing the editing request. The linguistic requests on virtual datasets are also generated au-

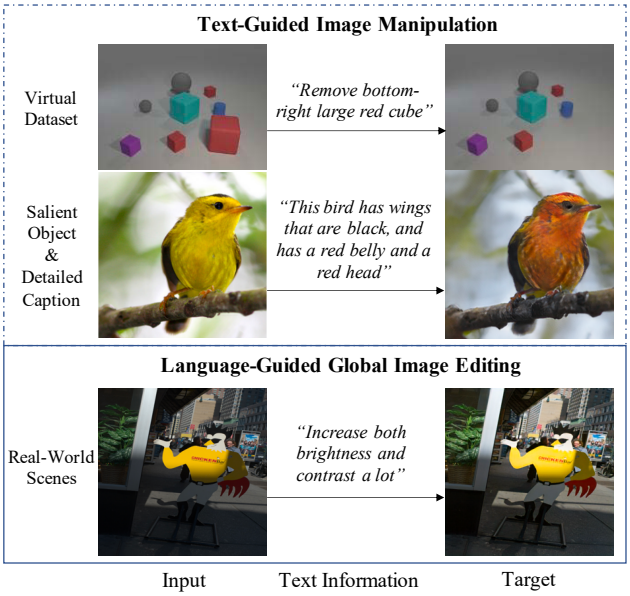


Figure 1. The example of language-guided global image editing, compared to existing tasks. Our task work on real-world scenarios with editing requests.

*Corresponding author

tomatically using the templates, which is hard to generalize to the requests from users. In our task, we receive images from the real-world environment and various linguistic requests from users that describe the editing process.

To tackle the language-guided global image editing, some methods [14, 22] try to map the language request into a sequence of executable editing operations. In this way, they require predefined editing operations and need additional annotations of editing operations. [26] is a GAN-based method, which models each global operation as a convolutional kernel and uses a neural generator that directly outputs the edited image. However, existing methods mainly have two limitations. First, current methods suffer from the problem of insufficient and imbalanced data of real-world datasets. For example, The GIER dataset proposed by [22] has $7,000$ input-target-request triplets collected from online image editing websites through the time period from 2009 to 2020. Among the data, editing requests related to hue adjustment only account for less than $10\%$ of the entire dataset, which indicates the imbalanced distribution of different editing operations. In addition, more than $80\%$ of the editing requests related to brightness operation is to *increase* the brightness, indicating that the imbalance even exists inside a specific operation. Such severe data imbalance issue will make methods fail to understand the language input well, resulting in simply brightening the input images when the received requests are related to brightness operation but actually to decrease the brightness, and also failing to respond to requests related to hue operation.

Second, even for global editing requests, it is often desirable to have different editing degrees on different image regions. For example, for an input image with a dark background and a bright foreground, given a vague user request like "brighten the image", it is more reasonable to brighten the background a lot but the foreground a little instead of brightening the whole image uniformly. However, previous methods can only apply editing operations globally, which results in unsatisfactory results.

To tackle the aforementioned limitations, we propose the Cycle Augmentation GAN (CAGAN) for language-guided global image editing. First, we propose a new cross-modal cyclic mechanism and data augmentation strategy to address the problem of insufficient and imbalanced data. Since directly collecting the training triplet examples (input-target-request) is expensive and laborious, we design a cross-modal cyclic mechanism to augment the data. Specifically, we devise an Editing Description Network (EDNet) to take in the input image and the edited image obtained from the generator, then produce the editing embedding that specifies the image transformation applied on the input image. With EDNet and generator, we can apply swapping augmentation (i.e., swap the input and target image) and random augmentations (i.e., adjusting the brightness, hue of the image) and then reconstruct the input image without leveraging linguistic requests. As we learn a better EDNet, we can use EDNet to boost the performance of the generator by maximizing the cosine similarity between the editing embeddings obtained from EDNet and the condition embedding obtained requests.

Second, we propose an Image-Request Attention (IRA) to adaptively edit the input image in different spatial locations. The IRA calculates the attention between embeddings of linguistic requests and patches on visual feature maps. By leveraging ground truth target images that are spatial-adaptive retouched as supervision, IRA learns to assign an appropriate degree for each location. For example, a very light place will receive a low attention degree when the request is "brighten the image" since the real intention is likely to only brighten the dark places. Finally, we propose a new evaluation metric for language-guided image editing called Redescription Similarity Score (RSS). To calculate the RSS, we leverage a pre-trained speaker model [24] to generate requests given the input image and the generated image, then calculate commonly used sentence similarity metrics between generated requests and ground truth requests. Higher similarity indicates better performance.

To sum up, we make the following contributions:

- We propose the CAGAN with a newly designed cross-modal cyclic mechanism and augmentation strategy for language-guided global image editing, which mitigates the problem of insufficient and unbalanced data.
- We propose IRA that calculates the degree of editing in the spatial dimension, which produces reasonable and interpretable editing results.
- We propose a new metric (RSS) to evaluate the performance of editing, which uses a speaker model to redescribe the input-output image pair.
- Experiments on both GIER [22] dataset and MA5k-Req [23] dataset demonstrate the effectiveness of our method.

## 2. Related Work

### 2.1. Image Editing

Image Editing has been studied a lot these years. Some works [9, 18] are proposed for global image editing but they do not use linguistic requests. Text-guided image manipulation [11, 16, 5] is a task that use language for image editing. But this task is designed for constrained domain-specific data, which are limited to a single salient object (e.g., bird) or virtual dataset domain. The text information is more like templates, which summarizing the attributes of target images. Recently, [22, 26, 23] are proposed to realize

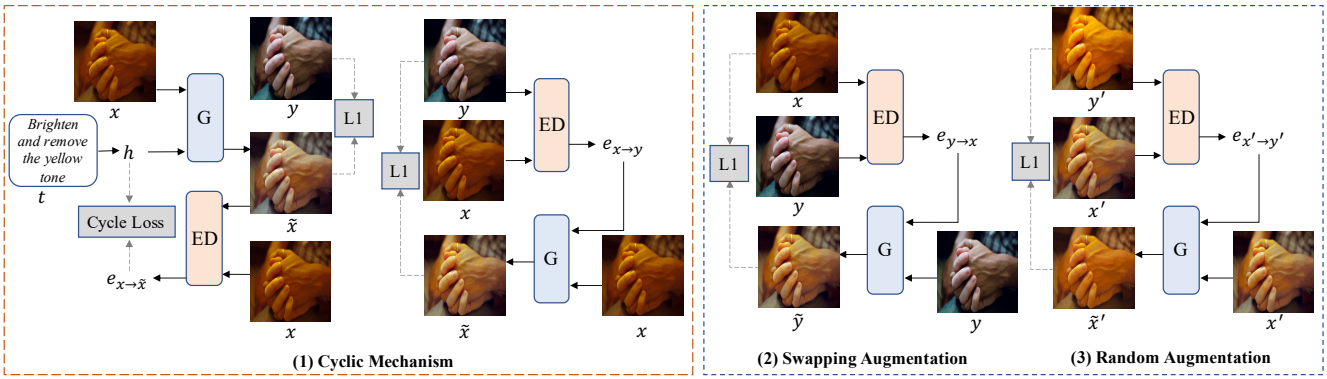


Figure 2. The left side indicates the cross-modal cyclic mechanism. Our model uses an RNN encoder to produce language embedding $h$ given the request $t$. The generator receives input image $x$ and $h$ to generate the edited image $\tilde{x}$. Then, the EDNet use $x$ and $\tilde{x}$ to produce editing embedding $e_{x\to\tilde{x}}$. Here, we apply reconstruction loss and cycle consistency loss for the cross-modal cyclic mechanism. For the augmentation strategy, we swap the input and target images or augment them with random adjustment. With the swapped/augmented images, EDNet produces the editing embedding $e_{y\to x}$ and $e_{x'\to y'}$. The CAGAN is required to reconstruct $x$ and $y'$. The discriminator is omitted here for clarity.

language-guided global image editing that works on real-world scenarios and requests from real users. [22] introduces a new language-guided image editing dataset GIER, and edits images by using predefined editing operations, but its training requires the annotation of the operation. [23] proposed a text-to-operation model to map the vague editing language request into a series of editing operations. [26] leverage an augmented image-to-image translation framework [10] to learn editing operators as a convolutional kernel. However, existing methods both suffer from the insufficient and imbalance of training data and can not adaptively edit the image. Our CAGAN is proposed to solve the aforementioned problem by a newly designed cross-modal cyclic mechanism and IRA.

### 2.2. Conditional GAN

Generative adversarial networks (GANs) [15, 7, 6] have made rapid progress on image generation in recent years. Built on the basis of GANs, the conditional GAN aims to synthesize the image according to the conditional signal. The input conditional signal can be images [10, 30], human poses [31], or semantic segmentation masks [19]. Text-to-image synthesis [21, 28, 29, 27, 12] which learns a mapping from textual descriptions to images propose to embed text information as condition for GAN. Based on the framework of text-to-image, SISGAN [4], TAGAN [16] and ManiGAN [11] are proposed to manipulate the input image through textual descriptions. But the textual descriptions used in the aforementioned methods are more like summarizing the attributes of the target image instead of describing the editing request. Recently, [5] propose the GeNeVA task and the GeNeVA-GAN for iterative image generation, where a new object is added one-by-one following the linguistic requests. But GeNeVA-GAN focuses on iterative editing and only tests on virtual toy datasets. We evaluated the aforementioned methods on real-world datasets but the performance is unsatisfactory, which could prove the difficulty of our task. Different from the previous methods, CAGAN is proposed to model linguistic requests instead of image captions and tested on real-world scenarios.

### 2.3. Cyclic Mechanism

Cycle consistency is widely researched in unsupervised and semi-supervised representation learning, where a transformation and its inverse operation are applied sequentially on input data, the consistency requires that the output representation should be close to the original input data in feature space. CycleGAN [30] is a popular method that leverages cyclic training for image-to-image transformation. StarGAN [3] augments CycleGAN using the conditional labels for multi-domain image-to-image transformation. Recently, MirrorGAN [20] propose to utilize an additional image captioning (Image-to-Text) network to describe the image that generates from a Text-to-Image network. In this paper, we build a new EDNet that takes in the input-output image pairs for producing editing embeddings for cycle consistency. Besides, indispensable data augmentations are applied in training EDNet, which protects our model from suffering the insufficient and imbalance of data.

## 3. Method

Our task is to edit a given image according to the modification specified in the input linguistic request. We first describe our basic generator in Sec. 3.1, which however has the difficulty to understand language requests well given the insufficient and unbalanced training data distribution. Therefore, we introduce our EDNet which can predict an editing embedding given a pair of images to create a cyclic loop with the generator (Sec. 3.2). Such cyclic mechanism

is useful to help our generator learn editing requests better even without real request annotation given swapping the order of image pairs and randomly augmented image pairs. In addition, another module IRA is equipped with the basic generator to predict the degree of editing which is useful for spatial-adaptive editing (Sec. 3.3).

### 3.1. Generator

Given the linguistic request $t$, we first use a BiLSTM encoder to encode the linguistic request to obtain the sentence embedding $h \in \mathbb{R}^{C_h}$ that represents the request. The generator $G$ takes in the input image $x$ and language embedding $h$ for generating the edited image $\tilde{x}$, as shown in Figure 2. With the target image $y$ and the discriminator $D$, $L1$ loss and adversarial loss can be used to supervise the generator $G$:

$$\mathcal{L}_G^{L1} = |\tilde{x} - y| \,. \tag{1}$$

$$\begin{aligned} \mathcal{L}_G^{adv} &= -\mathbb{E}_{\tilde{x}\sim\mathcal{P}_{model}} \left[\log D(\tilde{x})\right], \\ \mathcal{L}_D^{adv} &= -\mathbb{E}_{y\sim\mathcal{P}_{data}} \left[\log D(y)\right] - \mathbb{E}_{\tilde{x}\sim\mathcal{P}_{model}} \left[\log(1 - D(\tilde{x}))\right]. \end{aligned} \tag{2}$$

### 3.2. Cross-Modal Cyclic Mechanism

However, real-world datasets for language-guided global image editing are usually insufficient and unbalanced, e.g., over $80\%$ of the editing requests related to adjusting the brightness is to increase the brightness and less than $10\%$ of the requests in the entire dataset are related to hue operation. The learning methods are consequently biased by the unbalanced data distribution and tend to increase the brightness for every input image. However, directly augmenting the images is not enough since we need to annotate additional linguistic requests. Thus we develop a cross-modal cyclic mechanism to augment the training image pairs without requiring additional request annotations.

To create a cyclic loop with the generator $G$, we devise another network called EDNet which receives the input image $x$ and the generated edited image $\tilde{x}$ to produce the editing embedding $e_{x\to\tilde{x}}$ that indicates the editing operation. With the text conditional generator $G$ and the proposed EDNet, we have complete the cross-modal cyclic mechanism. Intuitively, we can train the model by reconstructing the target image $y$. Suppose that we learn a better EDNet, we can use EDNet to boost $G$ by maximizing the similarity of $e_{x\to\tilde{x}}$ and $h$. The cross-modal cyclic mechanism can be illustrated with the following equations:

$$\begin{aligned} \tilde{x} &= G(x, h), \\ e_{x\to\tilde{x}} &= ED(x, \tilde{x}), \\ \mathcal{L}_{cyc} &= 1 - \frac{e_{x\to\tilde{x}} \cdot h}{\|e_{x\to\tilde{x}}\|\|h\|}, \\ \mathcal{L}_{rec} &= \left|G\left(x, ED\left(x, y\right)\right) - y\right|, \end{aligned} \tag{3}$$

where $\mathcal{L}_{cyc}$ is the cycle consistency loss which is to maximize the cosine similarity between $e_{x\to\tilde{x}}$ and $h$, $\mathcal{L}_{rec}$ is the reconstruction loss. Next, we will explain how to build the EDNet and how to learn it better using data augmentations without additional ground truth editing requests.

**Editing Description Network.** To model the editing operation of image pairs, we devise the EDNet. We first use ResNet-101 [8] as the feature extractor to encode the input image $x$ and the target image $y$:

$$\begin{aligned} F_x &= \mathrm{ResNet}\left(x\right) \\ F_y &= \mathrm{ResNet}\left(y\right), \end{aligned} \tag{4}$$

where $F_x, F_y \in \mathbb{R}^{C_f\times N\times N}$ are feature maps of $x$ and $y$. $C_f$, $N$ indicates the number of channel and height or width of the feature maps. Then, we use attention to model the editing operation between images. For each feature $F_x^i$ in the input feature map $F_x$, the input-to-target attention computes its alignment with the feature $F_y^j$ in the target feature map $F_y$. The input feature maps and the attended target feature are then flattened and merged together with a fully-connected layer to obtain the editing embedding $e_{x\to y} \in \mathbb{R}^{C_h}$:

$$\begin{aligned} \alpha_{x\to y}^{i,j} &= \mathrm{softmax}_j\left(\left(W_1 F_x^i\right)^\top \left(W_2 F_y^j\right)\right), \\ F_{x\to y}^i &= \sum_j \alpha_{x\to y}^{i,j} F_y^j, \\ e_{x\to y} &= \tanh\left(W_3[F_x; F_{x\to y}] + b_3\right). \end{aligned} \tag{5}$$

We decompose the attention weight into two small matrices $W_1$ and $W_2$ so as to reduce the number of parameters since the dimension of the image feature is usually large.

**Data Augmentation on Image Pairs.** In the vanilla cross-modal cyclic mechanism, the data we used to train EDNet still suffer from insufficient and unbalanced data, which constrain the ability of EDNet. Thus we apply two kinds of data augmentation on the image pairs to learn EDNet better by leveraging the cyclic mechanism, as shown in Figure 2. First, we can swap the input image and target image, which means the CAGAN is asked to reconstruct the input image. This kind of augmentation mitigates the imbalance of requests, e.g., by converting brightening operations into darkening operations. Second, we apply random image transformation on both the input image $x$ and the target image $y$ to construct the new image pair $x'$ and $y'$. The random transformation includes adjusting the brightness, contrast, hue, sharpness, and saturation of the images, which are the combinations of frequent-used global editing operations. The EDNet and $G$ are also required to reconstruct the $y'$. This kind of augmentation mitigates the lack of data by increasing the training image pairs. Mathematically, the

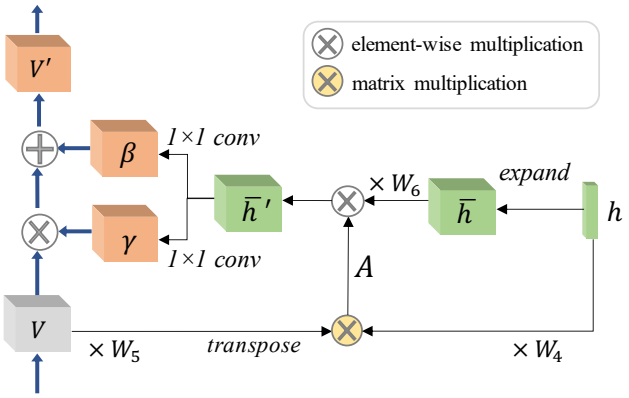


Figure 3. IRA calculates the attention between language embedding $h$ and patches on image feature map $V$. Given the Hadamard product of attention matrix $A$ and the expanded linguistic embeddings $\bar{h}$, the model learns to generate modulation parameters that are used to modify the visual feature $V$.

augmentation loss $\mathcal{L}_{aug}$ is calculated by:

$$\begin{aligned} x' &= \text{random\_adjust}(x) \\ y' &= \text{random\_adjust}(y) \\ \mathcal{L}_{aug} &= |G(y, ED(y, x)) - x| \\ &\quad + |G(x', ED(x', y')) - y'| \end{aligned} \tag{6}$$

### 3.3. Generator with Image-Request Attention

After obtaining the language embedding $h$ extracted from the request, we leverage a conditional generator $G$ to edit the input image $x$. However, even for global editing requests, sometimes spatial-adaptive editing is more preferred while previous methods do not have such capability. To solve this issue, we propose an Image-Request Attention (IRA) module in the generator to adaptively edit the input image in different spatial locations. IRA predict an attention map based on the correlation between the input image and request, as shown in Figure 3. We leverage a CNN to encoder the input image $x$ to the visual featuer map $V \in \mathbb{R}^{C_v \times H \times W}$. The proposed IRA embeds visual features $V$ and the language embedding $h \in \mathbb{R}^{C_h \times 1}$ into the same space, then calculate the attention matrix $A \in \mathbb{R}^{H \times W}$ by:

$$A = \text{Sigmoid}\left((W_4 h)^T (W_5 V)\right), \tag{7}$$

where $W_4 \in \mathbb{R}^{C_v \times C_h}$ and $W_5 \in \mathbb{R}^{C_v \times C_v}$ are learnable parameters. We also use the sigmoid function to normalize the weight of degree into $[0, 1]$. The attention calculates the multi-modal similarity which indicates that the larger the value in $A$, the greater the degree of editing. We expand and repeat the language embedding $h$ in spatial dimension to be $\bar{h} \in \mathbb{R}^{C_h \times H \times W}$. Then, we reweight the elements in $\bar{h}$ using attention matrix $A$ to obtain $\bar{h}' \in \mathbb{R}^{C_v \times H \times W}$:

$$\bar{h}' = W_6 \bar{h} \odot A, \tag{8}$$

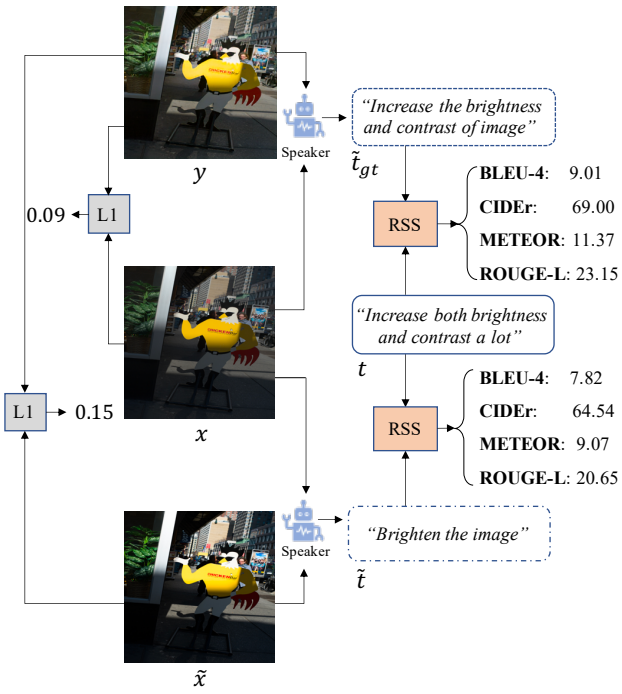


Figure 4. The overview of the proposed Redescription Similarity Score and why it is better than $L1$ distance.

where $W_6 \in \mathbb{R}^{C_v \times C_h}$ is a learnable parameter, $\odot$ denotes the Hadamard product. The language embedding after reweighting $\bar{h}'$ is then fed to the generator to serve as the conditional signal for editing.

Unlike some of the previous methods that rely on limited predefined operations, our generator directly models the linguistic request as modulation parameters that are used to modify the visual feature maps. With the weitghted condition embedding $\bar{h}'$, we generate the modulation parameters $\gamma \in \mathbb{R}^{C_v \times H \times W}$ and $\beta \in \mathbb{R}^{C_v \times H \times W}$ by:

$$\begin{aligned} \gamma &= W_7 \bar{h}'; \\ \beta &= W_8 \bar{h}', \end{aligned} \tag{9}$$

where $W_7 \in \mathbb{R}^{C_v \times C_v}$ and $W_8 \in \mathbb{R}^{C_v \times C_v}$ are learnable convolutional filters. The produced modulation parameters $\gamma$ and $\beta$ achieve image editing through scaling and shifting the visual feature map $V \in \mathbb{R}^{C_v \times H \times W}$ for only once:

$$V' = \gamma \odot V + \beta, \tag{10}$$

where $V'$ indicates the edited visual feature map. The edited visual feature map $V'$ is then fed to the subsequent decoder of $G$ for generating the result $\tilde{x}$.

## 4. Experiments

### 4.1. New Metric: Redescription Similarity Score

The existing metrics used by related works for evaluating language-guided global image editing are $L1$ distance

| Method | MA5k-Req | | | | | | | GIER | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | IS↑ | FID↓ | RSS↑ | | | | User ↑ | IS↑ | FID↓ | RSS↑ | | | | User ↑ |
| | | | BLEU-4 | CIDEr | METEOR | ROUGE-L | | | | BLEU-4 | CIDEr | METEOR | ROUGE-L | |
| TAGAN[16] | 12.67 | 60.21 | 8.25 | 63.06 | 10.62 | 21.92 | 1.98 | 8.91 | 79.54 | 2.83 | 26.91 | 8.42 | 21.14 | 2.61 |
| SISGAN [4] | 13.76 | 53.43 | 8.33 | 64.13 | 10.81 | 21.99 | 2.06 | 6.57 | 144.61 | 2.43 | 24.02 | 7.60 | 18.43 | 1.79 |
| GeNeVA [5] | 15.04 | 33.73 | 8.39 | 64.14 | 10.90 | 22.20 | 2.69 | 9.70 | 67.70 | 2.76 | 25.49 | 8.34 | 20.94 | 2.55 |
| PixAug [26] | 15.22 | 34.13 | 8.37 | 64.17 | 10.83 | 22.12 | 2.88 | 7.90 | 96.83 | 2.85 | 25.46 | 8.38 | 20.94 | 2.47 |
| OMN [22] | - | - | - | - | - | - | - | 9.63 | 65.99 | 3.67 | 30.16 | 8.90 | 22.64 | 2.82 |
| w.o. EDNet | 16.23 | 10.01 | 8.42 | 65.84 | 11.03 | 22.17 | 3.02 | 9.83 | 44.88 | 3.79 | 35.73 | 9.15 | 23.14 | 2.93 |
| w.o. Aug | 16.29 | 10.09 | 8.39 | 65.89 | 11.01 | 22.20 | 3.05 | 9.91 | 44.70 | 3.81 | 35.81 | 9.16 | 23.23 | 2.91 |
| w.o. IRA | 16.93 | 10.12 | 8.53 | 66.13 | 11.07 | 22.65 | 3.12 | 10.17 | 42.36 | 3.78 | 36.28 | 9.21 | 23.40 | 3.02 |
| Ours | **17.16** | **9.95** | **8.66** | **66.18** | **11.13** | **22.83** | **3.29** | **10.35** | **42.01** | **4.09** | **37.03** | **9.45** | **23.60** | **3.07** |

Table 1. Quantitative comparison with existing methods on MA5k-Req and GIER datasets.

and user studies. However, the only quantitative metric $L1$ distance can not reflect the quality of editing. It is because we only have one target image for each input-request pair, while the editing results could be actually diverse as long as the results match the request. We also find that a simple baseline to beat all the comparison methods is to not edit the images at all as shown on the left of Figure 4, which is unreasonable and reflects that the $L1$ distance is not suitable as a metric. Some other metrics used in text-to-image or image manipulation [11] like manipulative precision (MP) are also not suitable for our problem since the language information of our problem is describing the editing request instead of summarizing the attributes of the target image.

Therefore, we propose a new evaluation metric called Redescription Similarity Score (RSS). To calculate the RSS, we leverage a difference-speaker [24] that is trained on our datasets. As shown on the right side of Figure 4, we use the trained speaker to generate the editing request $\tilde{t}$ given the input image $x$ and the synthesized image $\tilde{x}$. Then, RSS is calculated by evaluating the similarity between the generated request $\tilde{t}$ and the ground truth requests $t$. Our RSS consists of four well-known metrics BLEU-4 [17], CIDEr [25], METEOR [1], ROUGE-L [13] which are used to evaluate the semantic similarity between sentences. Higher RSS indicates that the generated requests are more semantically similar to ground truth requests, which means the method for global editing is better.

## 4.2. Experimental Settings

**Datasets.** We adopt GIER [22] dataset and MA5k-Req [23] dataset in our experiments. GIER dataset consists of 7k samples, where each sample contains an input image, a linguistic request, a target image, and a list of applied operations. The original MA5k dataset [2] consists of 25k examples, where each sample is an input-target image pair. But it is used for image retouching and does not have language annotations. [23] followed the annotating procedure of [22] to collect the linguistic request for each input-target image pair to form the MA5k-Req dataset.

**Implementation details.** We resize the images to $256 \times 256$ for all the experiments. As for the generator, we adopt an encoder-bottleneck-decoder structure that consists of two downsample blocks, three bottleneck blocks, and two upsample blocks. The scaling and shifting applied on the feature map are only adopted in the first bottleneck. We train our model and baselines for $50$ epochs for every experiment.

**Baseslines.** We compare our method with existing language-guided global image editing methods PixAug and OMN. *PixAug* [26] is a GAN-based model following the language-augmented pix2pix [10] model that uses predefined operations for image retouching. *OMN* [22] is Operation Modular Network that comprises submodules of the predefined global operations. The parameters of each operation are also predicted by the modular network. Note that OMN relies on the annotations of operations and thus can not train on the MA5k-Req dataset. Besides, we also adapt the existing text-guided image manipulation methods to our task. *GeNeVA* [5] learns to generate the image step-by-step according to the text description. To fit our task, we use it for a single-step generation. *TAGAN* [16] and *SISGAN* [4] are two approaches for text-guided image manipulation. For a fair comparison, we add $L1$ loss for the baseline models which are originally train on unpaired data. We also conduct component analyses on IRA and our cyclic mechanism and augmentation. *w.o. IRA* means the CAGAN without using IRA. *w.o. EDNet* means the CAGAN without using cyclic mechanism and data augmentations. *w.o. Aug* means the CAGAN without using data augmentations.

## 4.3. Comparison

**Quantitative Results.** We conduct quantitative comparisons with the baselines using the Inception Score (IS), Fréchet Inception Distance (FID), and the newly proposed Redescription Similarity Score on both GIER and FiveK datasets, as shown in Table 1. Note that IS and FID only evaluate the results between generated image set and the target image set, only RSS evaluates the results in example level. We also conduct user studies for comparison. We randomly selected $100$ examples from the test set of two

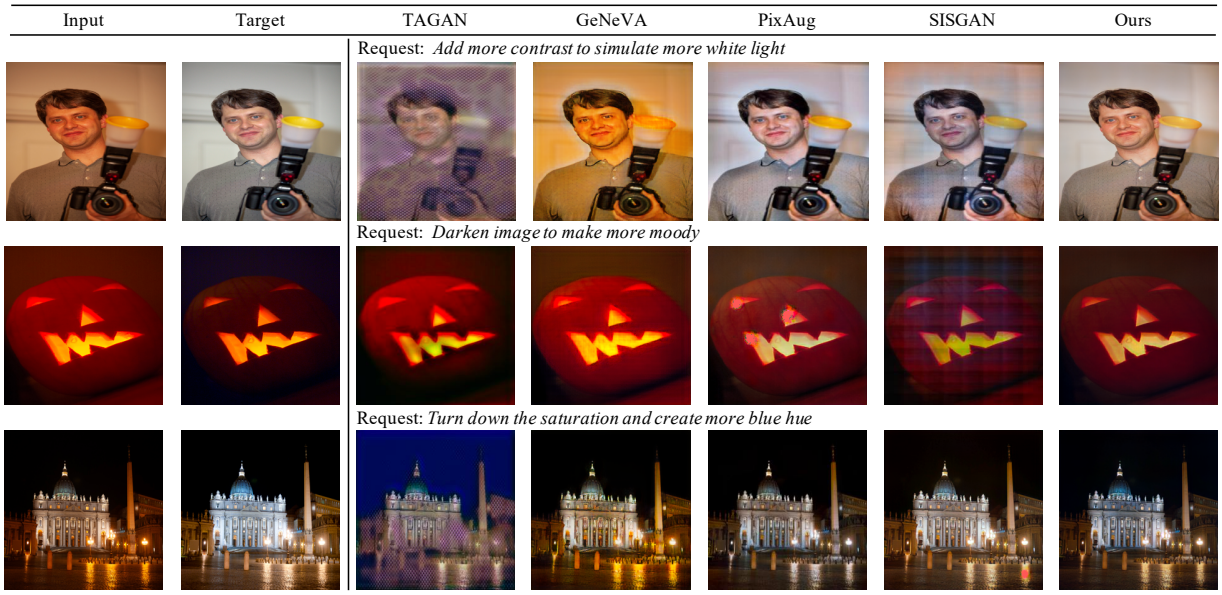


Figure 5. Qualitative comparison with baseline models on MA5k-Req dataset. Best viewed in color.

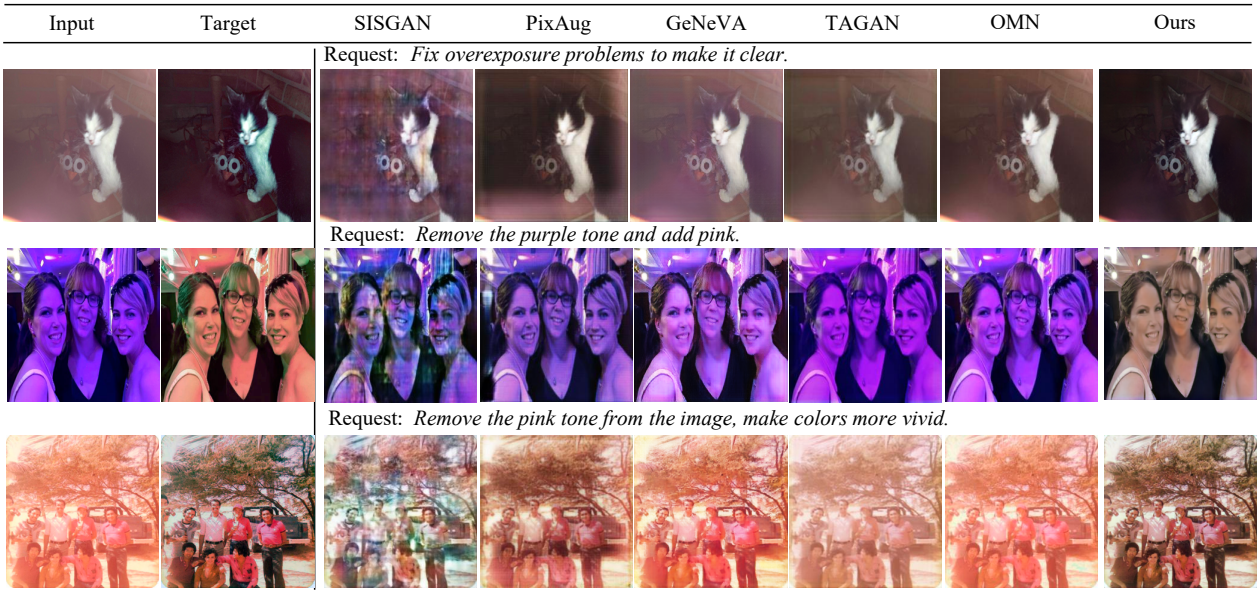


Figure 6. Qualitative comparison with baseline models on GIER dataset. Best viewed in color.

datasets respectively, and obtain 100 edited images for each method. Given the input image and the request, the user is asked to evaluate the image by considering the realism and the consistency between the image and its corresponding request. 10 volunteers are invited to rate scores from 1 (worst) to 5 (best). The average scores are reported in the "User" column of Table 1.

We can see that our full model is leading in all the evaluation metrics. Another observation is that the ranking of our user studies is quite similar to the ranking of RSS, which indicates that the newly proposed evaluation metric could reflect the intuitive feeling of users. The performance of our model drops without using the IRA or EDNet, which demonstrates the effectiveness of the proposed components. The results from user studies also support our view.

**Qualitative Results.** Qualitative comparison with baselines on the FiveK dataset and GIER dataset are shown in Figure 5 and Figure 6. As shown in the first two rows of Figure 5 and the first row of Figure 6, our CAGAN can adjust the lightness and contrast of the image according to the linguistic requests accurately. While the images generated by baseline models are abnormal, which do not match the editing request well or even fail to preserve the content of images. The much worse performance of SISGAN, PixAug


Figure 7. Visualization of the IRA. Warmer colors indicate higher values.

and TAGAN might because their task is different from ours and their moel ability is limited when testing on complex images. In the last row of the Figure 5 and the last two rows of Figure 6, our model shows the performance on adjusting the hue of images. Since the requests related to hue are quite few in the dataset ($< 10\%$), the baseline models often fail in these cases. By leveraging the cross-modal cyclic mechanism and data augmentations, our method is able to adjust the hue according to the requests.

### 4.4. Ablation Studies

**Image-Request Attention.** In CAGAN, our Image-Request Attention calculates the attention weight for the embedding of linguistic requests and patches on visual feature maps. The calculated weight matrix indicates the degree of editing for each location. This design works well in the situation that the request is vague and simple. As shown in the first row of Figure 7, for the underexposure input image with vague requests "brighten the photo", our IRA learns to assign appropriate editing degree in spatial dimension by increasing the lightness in dark places a lot but bright places a little. While without using IRA, the image is only brightened slightly in all the locations. When we have an overexposed image with a request that needs to reduce the lightness, as shown in the second row of Figure 7, IRA can also decrease the brightness of the light and its circular halo accurately. By leveraging the IRA, our model can provide spatial-adaptive editing.

**Cross-Modal Cyclic Mechanism and Data Augmentation.** In CAGAN, we leverage the EDNet to achieve cross-modal cyclic mechanism and data augmentation. The augmented EDNet is used to supervise the generator, which mitigates the insufficient and imbalance of data. The first three columns of Figure 8 show examples of adjusting the brightness. Since most of the requests in GIER and FiveK datasets related to the brightness adjusting are to increase the brightness, the model without EDNet or augmentation


Figure 8. Ablation study on data augmentation for EDNet.

fails to darken the image well and is not sensitive to the requests with subtle differences. By leveraging the EDNet and data augmentation strategy, the augmented generator is able to darken the image and edit the image that matches the requests well. In the last column of Figure 8, the model without EDNet or augmentation simply brighten the image and fail in removing the brown tone well since the examples include adjusting the hue are less than $10\%$ in the entire dataset. With the cross-modal cyclic mechanism and data augmentation, CAGAN can remove the brown tone well that matches the request.

## 5. Conclusion

In this paper, we study the language-guided global image editing problem that takes an input image and a linguistic request as input and then outputs the edited image that matches the request. To mitigate the insufficient and unbalanced data distribution, we build the EDNet and develop a cyclic mechanism for data augmentations. To produce reasonable and spatial-adaptive results when the requests are vague, we devise the IRA to assign an appropriate editing degree for each location. Both the design of EDNet and IPA can improve the performance of editing. Lastly, to tackle the lack of evaluation metric of the current problem, we propose the RSS by using a speaker model to redescribe the requests.

**Acknowledgements** Wentao Jiang, Jiayun Wang, Chen Gao, Si Liu from Beihang University are supported in part by National Natural Science Foundation of China (Grant 61876177), Beijing Natural Science Foundation (4202034), Fundamental Research Funds for the Central Universities.

## References


[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[3] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[4] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017.

[5] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10304–10312, 2019.

[6] Chen Gao, Yunpeng Chen, Si Liu, Zhenxiong Tan, and Shuicheng Yan. Adversarialnas: Adversarial neural architecture search for gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5680–5689, 2020.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics (TOG)*, 37(2):1–17, 2018.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[11] 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.

[12] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019.

[13] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[14] Ramesh Manuvinakurike, Jacqueline Brixey, Trung Bui, Walter Chang, Doo Soon Kim, Ron Artstein, and Kallirroi Georgila. Edit me: A corpus and a framework for understanding natural language image editing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[15] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[16] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. In *Advances in neural information processing systems*, pages 42–51, 2018.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[18] Jongchan Park, Joon-Young Lee, Donggeun Yoo, and In So Kweon. Distort-and-recover: Color enhancement using deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5928–5936, 2018.

[19] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[20] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019.

[21] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[22] Jing Shi, Ning Xu, Trung Bui, Franck Dernoncourt, Zheng Wen, and Chenliang Xu. A benchmark and baseline for language-driven image editing. *arXiv preprint arXiv:2010.02330*, 2020.

[23] Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dernoncourt, and Chenliang Xu. Learning by planning: Language-guided global image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13590–13599, 2021.

[24] Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. Expressing visual relationships via language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1873–1883, 2019.

[25] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evalua-

tion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[26] Hai Wang, Jason D Williams, and SingBing Kang. Learning to globally edit images with textual description. *arXiv preprint arXiv:1810.05786*, 2018.

[27] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[28] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.

[29] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[31] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019.