Sequential Attention GAN for Interactive Image Editing via Dialogue

Yu Cheng [1] Zhe Gan [1] Yitong Li [2] Jingjing Liu [1] Jianfeng Gao [3]

[1] Microsoft Dynamics 365 AI Research [2]Duke University
[3]Microsoft Research

Abstract

In this paper, we introduce a new task - interactive image editing via conversational language, where users can guide an agent to edit images via multi-turn dialogue in natural language. In each dialogue turn, the agent takes a source image and a natural language description from the user as the input, and generates a target image following the textual description. Two new datasets are created for this task, Zap-Seq and DeepFashion-Seq, collected via crowdsourcing. For this task, we propose a new Sequential Attention Genrative Adversarial Network (SeqAttnGAN) framework, which applies a neural state tracker to encode both source image and textual descriptions, and generates high quality images in each dialogue turn. To achieve better region specific text-to-image generation, we also introduce an attention mechanism into the model. Experiments on the two datasets, including quantitative evaluation and user study, show that our model outperforms state-of-the-art approaches in both image quality and text-to-image consistency.

1. Introduction

The volume of visual media has grown tremendously in recent years, which has intensified the need for professional image editing tools (e.g., Adobe Photoshop, Microsoft Photos). However, visual editing still remains a challenging task relying heavily on manual efforts [6], which is time-consuming and requires artistic creativity as well as iterative experimentation. A natural approach to automating the process is to provide an interactive environment, where a system can generate images automatically following users' verbal command; and where users can provide feedback on intermediate results, which in turn allows the system to further refine the generated images.

In this work, we propose a new task - interactive image editing via conversational language, where a system can generate new images by interacting with users in multi-turn dialogue. Figure 1 shows an example system, which supports natural language communication between the agent and an user for image editing (e.g., customizing shape, color, size, texture, etc.). Potential applications of such a system range from dialogue-based visual design to language-guided visual assistance.

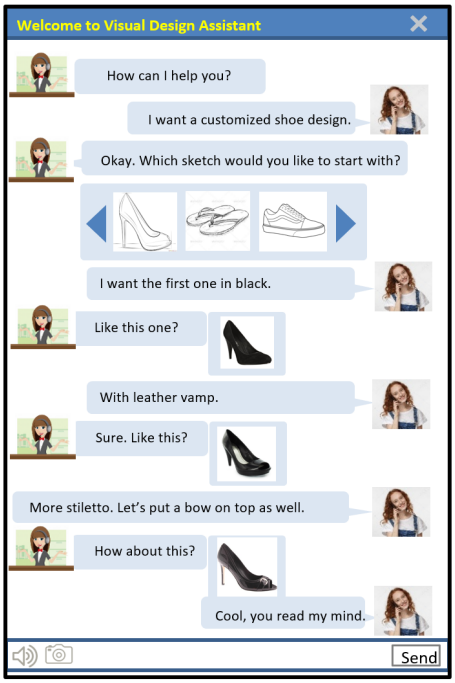


Figure 1. Example of a visual design assistant powered by interactive image editing. In each dialogue turn, the user provides natural language feedback to guide the system to modify the design. The system refines the images iteratively based on the user's feedback.

Current approaches useful for this task are limited to either keyword-based (e.g., object attributes) [3, 11] or single-turn setting [6, 33, 25]. While these paradigms are effective to some degree, the restriction on the format of user feedback inevitably constrains the information that a user can convey to the agent to influence the image generation process. To solve these challenges, we propose a new Conditional Generative Adversarial Network (GAN) framework, which includes an image generator for generating interme-

diate results, and a neural state tracker for encoding dialogue context. In each dialogue turn, the generator generates a new image by taking into input the dialogue history and previously generated images. To fully utilize the sequential information, the proposed model performs end-to-end training on the full dialogue sequence. Moreover, we adopt a regularizer based on an image-text matching loss as well as an attention mechanism, for better fine-grained text-to-image generation and multi-turn refinement.

As this is a newly proposed task, we also introduce two new datasets, namely Zap-Seq and DeepFashion-Seq, which were collected via crowdsourcing in a real-world application scenario. In total, there are 8,734 dialogues in Zap-Seq and 4,820 in DeepFashion-Seq. Each dialogue consists of a sequences of images, with slight variation in design, accompanied by a sequence of textual descriptions on the differences between each pair of images, similar to the examples in Figure 1.

Experiments on these two datasets show that the proposed SeqAttnGAN framework achieves better performance than state-of-the-art techniques. In particular, by incorporating dialogue history information, SeqAttnGAN is able to generate high quality images, beating all baseline models in terms of contextual relevance and consistency. Results also show that allowing natural language feedback is more effective than only taking keywords or visual attributes as input, as used in previous approaches.

Our contributions are summarized as follows:

- We propose a new task for visual editing - interactive image editing via dialogue, which allows users to provide natural language command and feedback for image editing, via multi-turn dialogue.
- We introduce two new datasets for this task, Zap-Seq and DeepFashion-Seq, collected through crowdsourcing. With free-formed descriptions and diverse vocabularies, the two datasets provide reliable benchmarks for the interactive image editing task.
- We propose a new conditional GAN framework - SeqAttnGAN, which can fully utilize dialogue history to synthesize images that conform to user's iterative feedback.

2. Related Work

Image Generation and Editing Language-based image editing [6, 21] is a task designed for minimizing labor work while helping users create visual data. Specifically, systems that can perform automatic image editing should be able to understand which part of the image that the user is referring to. This is a very challenging task, which requires comprehensive understanding of both natural language and visual information. Following this thread, several studies have explored the task. Hu *et al.* [18] tackled the language-based image segmentation task, taking phrase as the input. Ramesh *et al.* [21] developed a system using simple language to modify the image, where a classification model is utilized to understand user intent. Wang *et al.* [34] proposed a neural model for global image editing.

Since the introduction of GANs [14], there has been a surge of interest in image generation tasks. In the conditional GAN space, there have been some studies on generating images from images [19], captions [28]/ attributes [11], and object-patch [27]. There were also studies on how to parameterize the models and training framework [23] beyond the vanilla GAN [26]. Zhang *et al.* [40] stacked several GANs for text-to-image synthesis, with different GANs to generate images of different sizes. In these studies, the image is synthesized on the context level but is not region-specific.

AttnGAN [38] proposed by Xu *et al.* embedded attention mechanism into the generator to focus on fine-grained word level information. Chen *et al.* [6] presented a framework targeting image segmentation and colorization with a recurrent attentive model. The FashionGAN work [33] generated new clothing on a person based on textual descriptions. The TAGAN (text-adaptive generative adversarial network) [25] proposed a method for manipulating images with natural language description. While these paradigms are effective, the restrictions on specific user inputs (either pre-defined attributes or single-turn interaction) limit their impact.

Dialogue-based Vision Tasks AI tasks that lie in the intersection between computer vision and natural language processing have drawn much attention in the research community, benefiting from the latest deep learning techniques and GANs. Such tasks include visual question-answering [2], visual-semantic embeddings [36], grounding phrases in image regions [30], and image-grounded conversation [24].

Most approaches have focused on end-to-end neural models based on the encoder-decoder architectures and sequence-to-sequence learning [13, 32, 4, 8]. Das *et al.* [1] proposed the task of visual dialogue, where the agent can answer questions about images in an interactive dialogue. De Vries *et al.* [9] introduced the GuessWhat?! game, where a series of questions is asked to pinpoint a specic object in an image, with yes/no/NA answers. However, these dialogue settings are purely text-based, where visual feature only plays a supportive role. DeVault *et al.* [22] investigated building dialogue systems that can help users efficiently explore data through visualizations. Guo *et al.* [15] introduced an agent presenting candidate images to the user and retrieving new images based on user's feedback. Another piece of related work is [3] for interactive image generation by encoding history information. Different from them, text information is used to guide the image generation/editing in

our work.

3. Datasets: Zap-Seq and DeepFashion-Seq

The interactive image editing task is defined as follows: a user can interact with the system via iterative dialogue turns to edit an image. In the t-th dialogue turn, the system presents a reference image generated by the system $\hat{x}_t$ to the user. The user then gives a natural language feedback o_t, to describe the difference between the reference image and the desired image he/she wants. This process continues iteratively until the user is satisfied with the result rendered by the system, or the maximum number of dialogue turns has reached.

Existing image generation datasets are mostly single-turned, thus not suitable for this new multi-turn task. Therefore, we developed two new datasets for the proposed task - Zap-Seq and DeepFashion-Seq, which were derived from two existing datasets (UT-Zap50K [39] and DeepFashion [20]).

First, we retrieve sequences of images from the two datasets, with each sequence containing 3 to 5 images. Every pair of consecutive images are slightly different in one or two attributes [41]. As a result, a total of 8,734 image sequences were extracted from UT-Zap50K and 4,820 sequences from DeepFashion. After collecting these image sequences, the second step is to collect natural language descriptions that can capture the difference between each image pair. We resorted to crowdsourcing via Amazon Mechanical Turk [5] for this data collection task. Specifically, each human annotator was asked to provide a free-formed sentence to describe the difference between any two given images. Figure 2 provides some examples. The interface of the data collection task is provided in Appendix. To provide a robust dataset, we also randomly select images from the two original datasets to form additional sequences, which makes up 10% of the total datasets.

After manually removing wrong and duplicate annotations, we collected a total of 18,497 descriptions collected for the Zap-Seq dataset and 12,765 for DeepFashion-Seq. The statistics on the two datasets are shown in Table 1.

Most descriptions are very concise (between 4 to 8 words), yet the vocabulary is highly diverse (943 unique words in the Zap-Seq dataset and 687 in DeepFashion-Seq). Compared with pre-defined attributes, most descriptions often include fine-grained spatial or structural details. More information about the data collection procedure and the datasets can be found in Appendix.

4. Sequential Attention GAN (SeqAttnGAN)

For the new task, we develop a model to generate new images given current dialogue description, while preserving coherency to the natural language description, visual quality, and naturalness. Our framework is inspired by the Generative Adversarial Networks (GANs) [14, 23] and consists of three components: an image generator, a neural state tracker and a context encoder. As shown in Figure 3, in the t-th dialogue turn, the context encoder encodes the user response o_t and passes it to the state tracker. The state tracker then aggregates this representation with the dialogue history from previous turns. Base on the joint representation of user response $\{o_1, o_2, ..., o_t\}$ and previous intermediate images $\{\hat{x}_1, \hat{x}_2, ..., \hat{x}_{t-1}\}$, the image generator produces a new image $\hat{x}_t$.

Figure 2. Examples of the data collection. Each annotator is asked to provide a natural language sentence describing the difference between two design images. The images and collected descriptions are used to form "dialogue sequences" for the task.

Specifically, the generator G takes the hidden state h_t of t-th step as input and generates an image $\hat{x}_t$, defined as,

$$\hat{x}_t = G(h(t)) \tag{1}$$

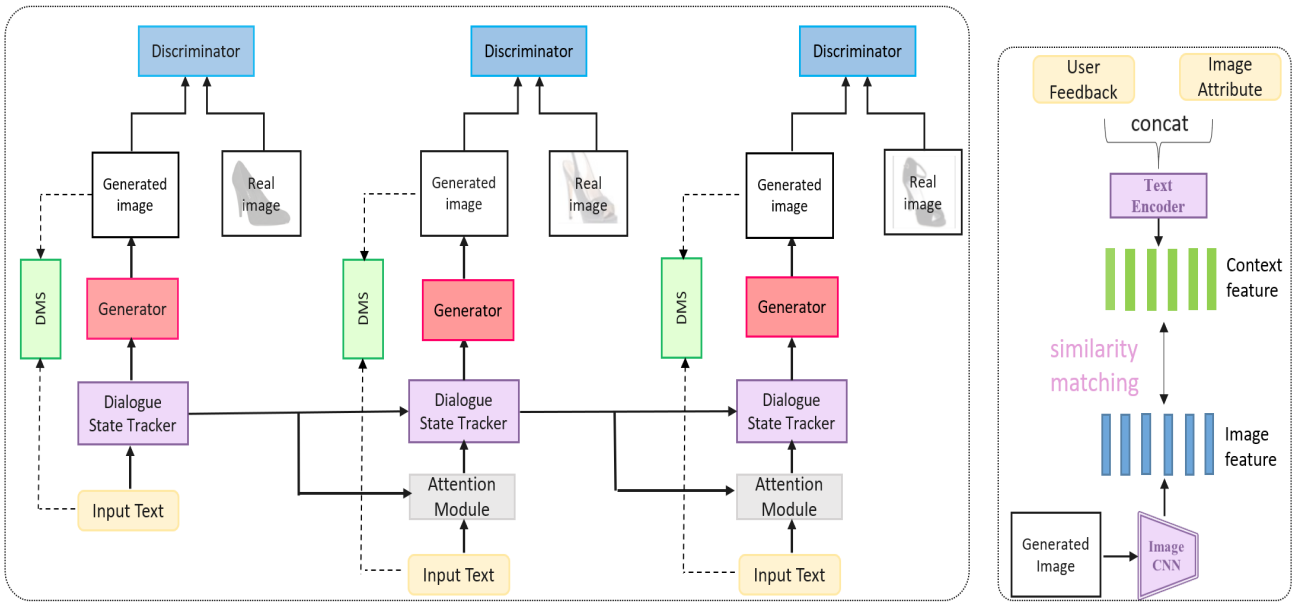
where $h(t)$ is produced from the state tracker based on a GRU unit that fuses the representation h_t with the dialogue history representation from previous dialogue turns, and outputs the aggregated feature vector h_{t+1}:

$$\begin{aligned} h_1 &= F^{enc}(x_0, z_1, o_1) \\ h_t &= F_t(h_{t-1}, z_t, F_t^{attn}(o_t, h_{t-1})) \end{aligned} \tag{2}$$

where z_t is a noise vector sampled in each step from a standard normal distribution. x_0 is the "initial image". The neural state tracker F_t is based on a gated recurrent unit (GRU) [7]. $F^{enc}()$ is concatenation and embedding through a linear transformation, to obtain the final response representation of x_0, z_1, and o_1. $F^{attn}()$ represents the proposed attention model, which will be discussed in the following subsection.

Dataset	#dialogues	#turns per dialogue	#descriptions	#words per description	#unique words
Zap-Seq	8,734	3.41	18,497	6.83	973
DeepFashion-Seq	4,820	3.25	12,765	5.79	687

Table 1. Statistics on the Zap-Seq and DeepFashion-Seq datasets.



(a) the overall framework of Sequential Attention GAN

(b) deep multimodal similarity regularize (DMS)

Figure 3. The architectures of SeqAttnGAN and deep multimodal similarity regularizer (DMS). For simplicity, we only illustrate three steps as an example. The attention module automatically retrieves the dialogue context for generating different sub-regions of the image. The DMS function provides a fine-grained image-text matching loss. During test, the target image is generated through the model with a forward pass.

Attention Module To perform compositional mapping [37, 33, 38], i.e., enforcing the model to produce regions and associated features that conform to the textual description, we introduce an attention module $F_t^{attn}()$ into the framework. $F_t^{attn}()$ takes user feedback o_t and image features from the previous hidden layer h_{t-1} as inputs. First, the user feedback o_t is converted to the common semantic space via a transform layer. Then, a word-context vector is computed for each sub-region of the image based on its hidden features h_{t-1}. For the i-th sub-region of the image (the i column of h_{t-1}), a word-context vector c_i can be obtained by learning the attention weights of every word in o_t given the i-th sub-region of the image. Finally, $F_t^{attn}(o_t, h_{t-1})$ produces a word-context matrix $\{c_0, c_1, ..., c_i, ...\}$, which is passed to the neural tracker F_t to generate an image in the t-th step.

Compared with AttnGAN [38], our framework deploys the attention component in a dialogue sequence. All the dialogue turn share the same generator, while AttnGAN has disjoint generators for different scales. Hence we name our model as **Sequential Attention GAN** (SeqAttnGAN).

Following [40], the objective of SeqAttnGAN is the joint conditional-unconditional losses over the discriminator and generator. With the supervision of x_t in the t-th turn, the loss of the generator G is defined as:

$$\mathcal{L}_G = -\frac{1}{2}\mathbb{E}_{\hat{x}_t \sim P_G}[\log D_t(\hat{x}_t)] - \frac{1}{2}\mathbb{E}_{\hat{x}_t \sim P_G}[\log D_t(\hat{x}_t, h_t)] \tag{3}$$

where the loss of the discriminator D is calculated by:

$$\mathcal{L}_D = -\frac{1}{2}\mathbb{E}_{x_t \sim P_{data}}[\log D(x_t)] - \frac{1}{2}\mathbb{E}_{\hat{x}_t \sim P_G}[\log(1 - D(\hat{x}_t))]$$
$$-\frac{1}{2}\mathbb{E}_{x_t \sim P_{data}}[\log D(x_t, h_t)] - \frac{1}{2}\mathbb{E}_{\hat{x}_t \sim P_G}[\log(1 - D(\hat{x}_t, h_t)] \tag{4}$$

where x_t is from the true data distribution P_{data} and $\hat{x}_t$ is from the model distribution P_G.

Deep Multimodal Similarity Regularizer In addition to the GAN loss, we bring in another term to SeqAttnGAN. We adopt the deep attentional multimodal similarity model (DAMSM) used in [12, 38] to: 1) maximize the utility of all

the input information (such as attributes) to boost the model performance; 2) regularize the model in order to stabilize the image generator. DAMSM is to match the similarity between the synthesized images and user input sentences, acting as an effective regularizer. For simplicity, we call it DMS regularizer in our paper.

The DMS function is pre-trained using dialogue data. Specifically, for any dialogue sample $\{x_1, o_1, ..., x_t, o_t, ...\}$, we first retrieve an initial image x_t as I_i, then concatenate the attribute value of x_{t-1} (denoted as a_{t-1}) and the annotated description o_t to have a new text D_i. Note that here we combine attributes and reference feedback as the text, which is different from [38, 40]. After selecting N image-description pairs, we have the image-description corpus as $\{I_i, D_i\}_{i=1}^{N}$. Following [17, 38], the posterior probability of description D_i matching the image I_i is defined as:

$$P(D_i|I_i) = \frac{\exp(\gamma R(I_i, D_i))}{\sum_{j=1}^{M} \exp(\gamma R(I_i, D_j))} \tag{5}$$

where γ is a smoothing factor. $R()$ is the word-level attention-driven image-text matching score [38] (i.e., the attention weights are calculated between the sub-region of an image and each word of its corresponding text) in word level. Given a batch of M pairs, the loss function for matching the images with their corresponding descriptions is:

$$\mathcal{L}_{DMS}^{i \to d} = -\sum_{i=1}^{M} \log P(D_i|I_i) \tag{6}$$

Symmetrically, we can also define the loss function for matching textual descriptions with their corresponding images (by switching D_i and I_i). Combining these two, the pre-trained DMS function is computed as:

$$\mathcal{L}_{DMS} = \mathcal{L}_{DMS}^{i \to d} + \mathcal{L}_{DMS}^{d \to i} \tag{7}$$

In summary, we use a similar idea to DAMSM in [38] to form the DMS regularizer. The training pairs are created by concatenating attributes and user description. The image-description matching score is calculated in each step. By bringing in the discriminator power of DMS, the model can generate region-specific features that align well with the descriptions as well as improving the visual diversity.

Adding all terms together, given a sequence of full supervisions $\{x_1, o_1, ..., x_T, o_T\}$, the overall objective of SeqAttnGAN is defined as:

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_G + \mathcal{L}_D + \lambda \mathcal{L}_{DMS} \tag{8}$$

where λ is the hyperparameter to balance the two loss functions. $\mathcal{L}_G$, $\mathcal{L}_D$ and $\mathcal{L}_{DMS}$ are computed in each step and aggregated back to update the gradients, similar to the training of regular dialogue systems.

Implementation Details The text encoder used in the model follows [29], which can be jointly tuned in each module in our framework. For the image encoder, we use a Convolutional Neural Network (CNN) with ResNet-101 [16] pre-trained on ImageNet [10] with fixed parameters. For the DMS function, we use the Bi-directional Long Short-Term Memory (BiLSTM) [31] to encode the text. The batch size M is set to 50. The hyper-parameters λ and γ are set via the experiments.

5. Experiments

We validate the effectiveness of our model through both quantitative and qualitative evaluations. Given the subjective nature of dialogue with visual synthesis, we also conduct a user study via crowdsourcing to compare our approach with state-of-the-art methods.

5.1. Datasets and Baselines

All experiments are performed on the Zap-Seq and DeepFashion-Seq datasets with the same splits: 90% images were used for training, and the model was evaluated on a held-out test set from the rest 10%. The training process used image pairs sampled from the training set that has no overlap with the test set.

We compare our approach with several baselines:

(1) **StackGAN**. The first baseline is StackGAN v1 [40] (the resolution of images on Zap-Seq and DeepFashion-Seq is low). A generator is trained to generate target images with a resolution of 64×64 pixels, based on the reference attributes and the descriptions. In other components of the StackGAN architecture, all hyper-parameters and training epochs remain the same as the original.

(2) **AttnGAN**. AttnGAN [38] currently achieves the state-of-the-art Inception Score on MS-COCO. Similar to StackGAN, we utilize AttnGAN1, generating images at a resolution of 64×64 pixels. The discriminator and all the hyper-parameters stay unchanged.

(3) **LIBE**. We also used the recurrent attentive model employed in [6] for image coloring and segmentation task as another baseline. Like AttnGAN and StackGAN, LIBE utilizes image-caption pairs to train the model. The hyper-parameter setting and training details remain the same as in the original paper.

For training, we use bounding box information for images. Data augmentation is also performed in both datasets. Specifically, images are cropped to 64×64 and augmented with horizontal flips. To perform a fair comparison, all models share the same structure of the generator and discriminator. The text encoder is also shared. The baseline model training follows standard conditional GAN training procedure, using Adam with the default batch size of 32.

Model	Zap-Seq		DeepFashion-Seq	
	IS	FID	IS	FID
StackGAN	7.88	60.62	6.24	65.62
AttnGAN	**9.79**	**48.58**	8.28	55.76
LIBE	4.73	76.52	3.89	79.04
SeqAttnGAN	9.58	50.31	**8.41**	**53.18**

Table 2. Comparison of Inception Score (IS) and Frechet Inception Distance (FID) between our model and the baselines on the two datasets. Note that AttnGAN is a strong baseline for image generation in terms of IS and FID.

Dataset	StackGAN	AttnGAN	LIBE	SeqAttnGAN
Zap-Seq	0.437	0.527	0.159	**0.651**
DF-Seq	0.316	0.405	0.112	**0.498**

Table 3. Comparison of SSIM score from our model and the baselines on the two datastes. Here DF-Seq is the DeepFashion-Seq dataset.

5.2. Quantitative Evaluation

In this section, we provide the quantitative evaluation and analysis on the two datasets. In each step of a dialogue from the test set, we randomly sampled one image from each model, then calculated the IS and FID scores comparing each of the selected sample with the ground-truth image. The averaged numbers are presented in Table 2. Our SeqAttnGAN model outperforms StackGAN and LIBE on the Zap-Seq dataset, with slightly worse performance than AttnGAN. On the DeepFashion-Seq dataset, our model achieves the best results among all the models.

Next, to evaluate whether the generated images are coherent with the input text, we also measure the Structural Similarity Index (SSIM) score between generated images and the ground-truth images. Table 3 summarizes the results, which shows that the generated images yielded by our model are more consistent with the ground-truth than all the baselines. This indicates that the proposed model can generate images with better contextual coherency.

Figure 4 and Figure 5 present a few examples comparing all the approaches with the ground-truth. For each image, we can observe that our model generates images consistent with the ground-truth images and the reference descriptions. Specifically, SeqAttnGAN can generate images with good visual quality, while incorporating changes described in the text. Even for some fine-grained features ("kitten heel", "leather", "button"), the generated images can well satisfy the requirement. AttnGAN is able to synthesize visually sharp/diverse images, but not as good as our method in terms of context relevance. StackGAN does not perform as well as our model and AttnGAN, in terms of both visual quality and content consistency. This observation is consistent with the quantitative study.

Model	Zap-Seq	DeepFashion-Seq
StackGAN	2.68 ± 0.24	2.53 ± 0.27
AttnGAN	2.37 ± 0.27	2.46 ± 0.29
SeqAttnGAN	$\mathbf{1.84} \pm 0.23$	$\mathbf{1.88} \pm 0.25$

Table 4. Results from the user study. A lower number indicates a higher rank.

Some examples from LIBE are shown in Figure 6. Visually, LIBE cannot generate good quality samples and can only capture the color information to some degree.

5.3. Human Evaluation

We perform a human evaluation on Amazon Mechanical Turk. For each dataset, we randomly sampled 100 image sequences generated by all the models, each assigned to 3 workers to label. The source model of each image is hidden from the annotators for fair comparison. The participants were asked to rank the quality of the generated image sequences with respect to: 1) consistence to the description and the source image, 2) visual quality and naturalness.

Table 4 provides the results from this evaluation. For each approach, we computed the average ranking (where 1 is the best and 3 is the worst) and standard deviation. Results show that our approach achieves the best rank. This human study indicates that our solution achieves the best visual quality and image-text consistency among all the models.

Besides the crowdsouring human evaluation, we also recruited real users to interact with the proposed system. Figure 7 shows examples of several dialogue interactions with real users. We can observe that users often start the dialogue with a high-level description of main attributes (e,g., color, category). As the dialogue progresses, users give more specific feedback on fine-grained descriptions. The benefit of using free-formed dialogue can be demonstrated by the flexible usage of fine-grained attribute words (white shoelace), as well as comparative descriptions (thinner, more open). Our model is able to capture both coarse and fine-grained differences between images through multi-turn refinement. Overall, these results show that the proposed SeqAttnGAN model exhibits promising potential on generalizing to real-world applications.

5.4. Ablation Study

We conducted an ablation study to verify the effectiveness of two main components in the proposed SeqAttnGAN model: attention module and DMS regularizer. We first compare the IS, FID and SSIM scores of SeqAttnGAN with/without attention and DMS. Table 5 shows the ablation results on Zap-Seq and DeepFashion-Seq.

We observe that both attention and DMS can improve the model with a large margin. We also show some exam-

Figure 4. Examples of images generated from the given descriptions in the Zap-Seq dataset. The first row shows the ground-truth images and its reference descriptions, followed by images generated by the three approaches: SeqAttnGAN, AttnGAN and StackGAN. To save space, we only display some key phrases of each description.

Figure 5. Examples of images generated by different methods on the DeepFashions-Seq dataset.

ples generated with different variations in Figure 8. Results show that the original model outperforms the variation settings, as DMS helps stabilizing the training while the attention module improves image-and-text consistency.. Similar observation can been found in the DeepFashion-Seq dataset.

6. Conclusion

In this paper, we present interactive image editing via dialogue, a novel task that resides in the intersection of computer vision and language. We demonstrate the value of this task as well as its many challenges. To provide benchmarks for this new task, we release two datasets, with image se-

Model	Zap-Seq			DeepFashion-Seq		
	IS	FID	SSIM	IS	FID	SSIM
SeqAttnGAN	**9.58**	**50.31**	**0.651**	**8.41**	**53.18**	**0.498**
w/o Attn	8.52	57.19	0.548	7.58	58.15	0.433
w/o DSM	8.21	58.07	0.478	7.24	60.22	0.412

Table 5. Ablation study on using different variations of SeqAttnGAN, measured by IS, FID and SSIM.

Figure 6. Examples generated using LIBE. In each row, the left three images are generated from the model and the right three are ground-truth images.

Figure 7. Examples of user interacting with our image editing system using SeqAttnGAN. Each row represents an interactive dialogue between the user and our system.

quences accompanied by textual descriptions. To solve this task, we propose the SeqAttnGAN model, which can jointly model user's description and dialogue history to iteratively generate images.

Experimental results demonstrate the effectiveness of SeqAttnGAN. In both quantitative and human evolution, our approach with sequential training outperforms baseline methods that rely on pre-dened attributes or trained in a single-turn paradigm, while offering a more expressive and natural human-machine communication. Particularly, the proposed attention technique can enforce the networks to focus on specific areas of the image, and the DMS function can regularize the model to boost the rendering power.

Figure 8. Examples generated by different variations of our model. The first row is SeqAttnGAN, the second is without attention and the last is without DMS.

The results are limited by the current fashion data we adopted. In future work, we plan to build a generic system for other types of images (e.g., face [35]). Currently the framework still needs associated attributes to train the regularizer, which is not easy to be generalized. We would also like to investigate other ways to avoid using attribute data. Finally, we plan to investigate models to support more robust natural language interactions, which requires techniques such as user intent understanding, co-reference resolution, etc.

References

[1] *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[3] R. Y. Benmalek, C. Cardie, S. J. Belongie, X. He, and J. Gao. The neural painter: Multi-turn image generation. *CoRR*, abs/1806.06183, 2018.

[4] A. Bordes and J. Weston. Learning end-to-end goal-oriented dialog. *CoRR*, abs/1605.07683, 2016.

[5] M. Buhrmester, T. Kwang, and S. Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6:3–5, 8 2011.

[6] J. Chen, Y. Shen, J. Gao, J. Liu, and X. Liu. Language-based image editing with recurrent attentive models. *arXiv preprint arXiv:1711.06288*, 2017.

[7] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[8] A. Das, S. Kottur, J. M. F. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *ICCV*, pages 2970–2979. IEEE Computer Society, 2017.

[9] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. C. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, pages 4466–4475. IEEE Computer Society, 2017.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[11] M. Dixit, R. Kwitt, M. Niethammer, and N. Vasconcelos. Aga: Attribute guided augmentation. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jan 2017.

[12] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

[13] J. Gao, M. Galley, and L. Li. Neural approaches to conversational ai. *arXiv preprint arXiv:1809.08267*, 2018.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[15] X. Guo, H. Wu, Y. Cheng, S. Rennie, and R. S. Feris. Dialog-based interactive image retrieval. *CoRR*, abs/1805.00145, 2018.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[17] X. He, L. Deng, and W. Chou. Discriminative learning in sequential pattern recognition. volume 25, pages 14–36. Institute of Electrical and Electronics Engineers, Inc., September 2008.

[18] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017.

[20] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] R. Manuvinakurike, T. Bui, W. Chang, and K. Georgila. Conversational Image Editing: Incremental Intent Identication in a New Dialogue Task. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 284–295, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[22] R. Manuvinakurike, D. DeVault, and K. Georgila. Using Reinforcement Learning to Model Incrementality in a Fast-Paced Dialogue Game. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, Saarbruecken Germany, Aug. 2017. SIGDIAL.

[23] M. Mirza and S. Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[24] N. Mostafazadeh, C. Brockett, B. Dolan, M. Galley, J. Gao, G. Spithourakis, and L. Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. In *IJCNLP*, 2017.

[25] S. Nam, Y. Kim, and S. J. Kim. Text-adaptive generative adversarial networks: Manipulating images with natural language. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 42–51. Curran Associates, Inc., 2018.

[26] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2642–2651, 2017.

[27] X. Ouyang, Y. Cheng, Y. Jiang, C.-L. Li, and P. Zhou. Pedestrian-synthesis-gan: Generating pedestrian data in real scene and beyond. *arXiv preprint arXiv:1804.02047*, 2018.

[28] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1060–1069, 2016.

[29] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1060–1069. JMLR.org, 2016.

[30] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[31] M. Schuster and K. Paliwal. Bidirectional recurrent neural networks. volume 45, Nov. 1997.

[32] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3776–3783, 2016.

[33] R. U. D. L. C. C. L. Shizhan Zhu, Sanja Fidler. Be your own prada: Fashion synthesis with structural coherence. In *International Conference on Computer Vision (ICCV)*, 2017.

[34] H. Wang, J. D. Williams, and S. Kang. Learning to globally edit images with textual description. *CoRR*, abs/1810.05786, 2018.

[35] J. Wang, Y. Cheng, and R. S. Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *CVPR*, pages 2295–2304. IEEE Computer Society, 2016.

[36] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5005–5013, 2016.

[37] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016.

[38] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *CoRR*, abs/1711.10485, 2017.

[39] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 192–199, 2014.

[40] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

[41] B. Zhao, J. Feng, X. Wu, and S. Yan. Memory-augmented attribute manipulation networks for interactive fashion search. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6156–6164, 2017.