IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

An Interactive Image Editing System using an Uncertainty-based Confirmation Strategy

SEITARO SHINAGAWA[1,2], KOICHIRO YOSHINO[1,2,3], (Member, IEEE), SEYED HOSSEIN ALAVI[4], KALLIRROI GEORGILA[4†], DAVID TRAUM[4†], (Member, IEEE), SAKRIANI SAKTI[1,2], (Member, IEEE), and SATOSHI NAKAMURA[1,2], (Fellow, IEEE)

[1]Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192, Japan
[2]RIKEN, Center for Advanced Intelligence Project AIP, Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan
[3]PRESTO, Japan Science and Technology Agency, Honcho 4-1-8, Kawaguchi-shi, Saitama, 332-0012, Japan.
[4]Institute for Creative Technologies, University of Southern California, 12015 E Waterfront Dr, Los Angeles, CA 90094, USA

Corresponding author: Seitaro Shinagawa (e-mail: shinagawa.seitaro.si8@is.naist.jp).

† These authors were supported in part by the U.S. Army.

ABSTRACT We propose an interactive image editing system that has a confirmation dialogue strategy using an entropy-based uncertainty calculation on its generated images with Deep Convolutional Generative Adversarial Networks (DCGAN). DCGAN is an image generative model that learns an image manifold of a given dataset and enables continuous change of an image. Our proposed image editing system combines DCGAN with a natural language interface that accepts image editing requests in natural language. Although such a system is helpful for human users, it often faces uncertain requests to generate acceptable images. A promising approach to solve this problem is introducing a dialogue process that shows multiple candidates and confirms the user's intention. However, confirming every editing request creates redundant dialogues. To achieve more efficient dialogues, we propose an entropy-based dialogue strategy that decides when the system should confirm, and enables effective image editing through a dialogue that reduces redundant confirmations. We conducted image editing dialogue experiments using an avatar face illustration dataset for editing by natural language requests. Through quantitative and qualitative analysis, our results show that our entropy-based confirmation strategy achieved an effective dialogue by generating images desired by users.

INDEX TERMS confirmation, generative adversarial networks, image editing, natural language interface

I. INTRODUCTION

TIMELY and appropriately assisting human users is critical in intelligent systems. Image generation or editing systems help users create desired images through interaction [1], [2]. The capability of natural language interaction on such systems would be useful because a natural language interface does not require any special skills; it only requires the ability for natural language communication. For example, image editing systems that accept natural language requests have a natural language interface. It allows users to input requests via voice or chat. The system provides a new image according to the user request.

Such image editing systems often face ambiguities caused by natural language. Unlike general image-to-image translation tasks [3], such editing systems must be able to handle vague, under-specified, and ambiguous natural language requests. For example, the following natural language request, "make this avatar's hair short," lacks a specific objective image or criterion for creating the image desired by the user. It should be "make this avatar's hair short by her ears" in the less ambiguous case. However, such lack of specificity often occurs in a real situation. This is one challenging obstacle that must be overcome to generate images based on some given text. Asking the user about the ambiguity is one way to solve the problem. This solution is one of our motivations for introducing an interactive process in image editing. A trade-off also exists between the generated image quality and the constraints on the image generation system. For example, a masking mechanism is an efficient way to improve the quality of generated images in image-to-image translation tasks [4],

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997012, IEEE Access

[5], [6]. Masking denotes an element-wise multiplication of a mask, which consists of binary values, with the input image. Even in image editing with natural language, such generation systems based on masking constraints generate more accurate images than a system without them because they can identify the parts of the image mentioned in the user's request and perform image editing on those parts of the image only [7]. However, such a strong constraint limits large changes to the image. For example, in interactive image editing, it is difficult for systems with a strong constraint to work on such a request as "make the current portrait's hair longer" because the request will greatly change the image. In such cases, using a generation system without any constraints can create more relevant images to the user's intention.

Considering a problematic case where the system cannot decide which generated image is better as an editing result for users, one possible solution is direct confirmation with them. However, asking users to choose a single image for every request is completely unreasonable. Thus, the system is expected to ask them when it is unsure which is the best image to present.

In this paper, we assume two different types of interactive image editing systems: a system with a strong constraint and one without a constraint on their generative processes. We tackle this problem to find a better dialogue strategy using these two systems and introduce an uncertainty score based on the entropy of the generated masks to decide on the best system to a given image editing request. We call the system with the strong constraint based on the masking mechanism "w/ mask" and the system without a constraint "w/o mask." The system confirms with the user when it is tentative about selecting a better image to match the user's editing intent using uncertainty scores.

Section II describes the image editing task. Section III shows the interactive image editing system and its dialogue strategy that we use in our experiments. Section IV presents the experimental setting, and Section V shows our results. Related works are mentioned in Section VI, and we conclude in Section VII.

II. INTERACTIVE IMAGE EDITING DIALOGUE

In this section, we describe the interactive image editing dialogue task. Its overview is shown in Figure 1. It has a human user and a system. The dialogue's purpose is to generate goal image X^g, which is the user's desired image, through a dialogue. The user makes requests in natural language to change the current image closer to the goal. The system generates a new image based on the previous image when the user makes a request for a change.

Step 1 First source image X_0^s and goal image X^g are given to the user.

Step 2 At the i-th turn interaction, the user makes a natural language request I_i to edit the previous image X_{i-1}^s.

Step 3 The system generates a new image X_i based on the request I_i and the previous image X_{i-1}^s.

Step 4 The system resets X_i as the new source image X_i^s, and the user chooses whether to continue the dialogue. If the user decides to continue, they go to the next turn (go to 2 with $i \mathrel{+}= 1$). If the user decides to stop the dialogue, the dialogue is finished, and image X_i^s is compared with goal image X^g.

Note that since the goal image is invisible to the system, it cannot be optimized directly to generate the goal image.

If we have several image generators on Step 3, the system must choose one image as the new image X_i. When the system cannot choose between images, one solution is to seek confirmation from the user about which image is better. We assume that the system has multiple image candidates $X_{i,1}, X_{i,2}, \ldots, X_{i,n}$ in Step 3 and two choices: $\{confirm, not\ confirm\}$. If it selects $confirm$, the following sub-steps of the confirmation procedure are inserted before Step 4:

3-c1) The system shows image candidates to the user to confirm which image is relevant to the request.

3-c2) The user selects the most relevant image. The system sets the selected image as its generated image X_i.

Figure 1 summarizes the steps of a single turn to decide on the next source image X_i^s from Step 2 to Step 4. Since the confirmation steps lengthen the interaction, the system has to reduce the number of confirmations. Criteria exist upon which the system selects $confirm$ or $not\ confirm$ (see Section III-D).

III. DCGAN-BASED IMAGE EDITING MODELS AND DIALOGUE STRATEGY

In this section, we describe the internal architecture of our interactive image editing system, shown in Figure 1 (right), composed of image editing models based on Deep Convolutional Generative Adversarial Networks (DCGAN) [8]. We use two image editing models: a model without a generation constraint and a model with a generation constraint. We first describe DCGAN's general idea and then describe its extension to image editing tasks. We also describe dialogue strategies to use these models in an interactive process.

A. DEEP CONVOLUTIONAL GENERATIVE ADVERSARIAL NETWORKS (DCGAN)

A Deep Convolutional Generative Adversarial Network (DCGAN) [8] is a commonly used generative model for image generation. DCGAN is composed of generator G and discriminator D for adversarial learning [9]. The generator is defined:

$$\hat{X} = G(z). \tag{1}$$

It generates image $\hat{X}$ from given noise z (e.g., Gaussian: $z \sim N(0, I)$). The discriminator is defined:

$$\hat{y} = D(x). \quad (x \in \{X, \hat{X}\}) \tag{2}$$

It classifies a given image into two classes: original target image X from the training data (real) or generated target

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997012, IEEE Access

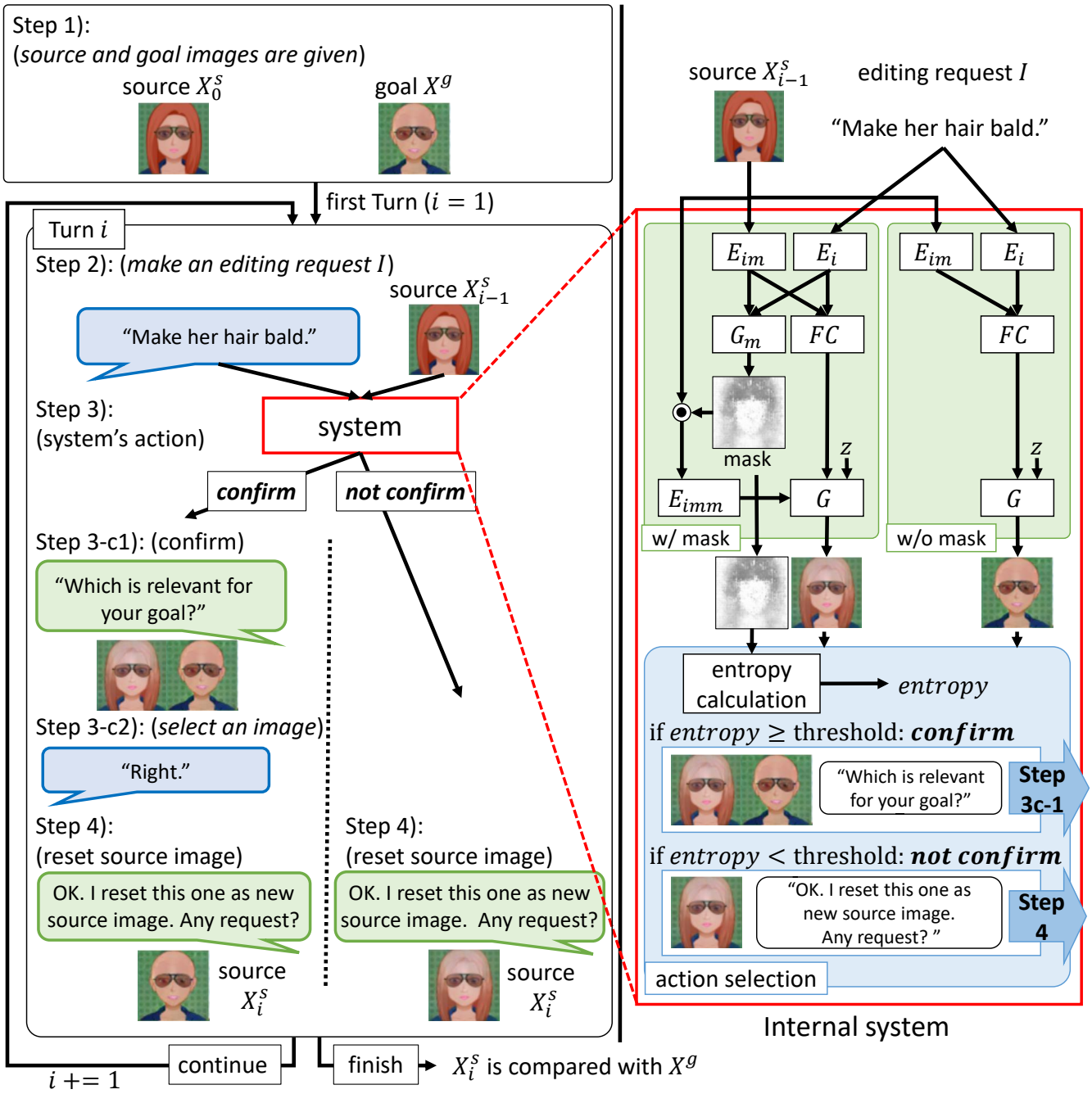

FIGURE 1. The left figure represents an overview of the interactive image editing dialogue and the right figure represents the internal architecture of the system. In the left figure, the user's utterance is blue and the system's one is green. The system decides to $confirm$ or $not\ confirm$ based on the user's editing request I and the current source image X_{i-1}^s. The right figure shows the whole system which consists of DCGAN-based image editing models and an entropy-based confirmation mechanism. w/o mask model described in Section III-B generates an image and w/ mask model described in Section III-C generates a mask and an image. Our proposed confirmation method (blue box), action selection module described in Section III-D, can select $confirm$ or $not\ confirm$ based on the entropy calculation of the mask.

image $\hat{X}$ by generator G (fake). The discrimination result will be used to train the generator.

DCGAN is optimized by the following objective:

$$\min_{\theta_G} \max_{\theta_D} V(G,D) = \mathbb{E}_{X\sim p_{data}}[\log D(X)] + \mathbb{E}_{z\sim p_z}[\log(1-D(G(z)))]. \quad (3)$$

θ_G and θ_D are the trainable parameters of the generator and the discriminator. p_{data} and p_z denote the data and noise distributions. Adversarial learning resembles a mini-max game between the generator and the discriminator. The discriminator is optimized to correctly classify generated images from the generator (fake) and training examples (real). On the other hand, the generator is optimized to trick

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997012, IEEE Access

the discriminator into predicting the generated images as training examples. This competitive training improves the image modeling performance [8]. To stabilize the training, we rewrite (3) and get the following training objectives as shown in [9]:

$$\min_{\theta_D} \mathcal{L}_D = -\mathbb{E}_{X\sim p_{data}}[\log D(X)] - \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))], \tag{4}$$

$$\min_{\theta_G} \mathcal{L}_G = -\mathbb{E}_{z\sim p_z}[\log D(G(z))]. \tag{5}$$

B. DCGAN FOR IMAGE EDITING WITHOUT CONSTRAINT

The original DCGAN was an unconditional image generation model; however, image editing tasks require conditional generation because the system has to control generated images based on the given pair of the original image (source image) and the editing request as a generation condition. To achieve this conditional generation, we introduce an extension of the DCGAN model that has an encoder part for extracting conditional information from the given pair of the source image and the editing request [7].

The encoder part learns function $\phi = f(X^s, I)$ by estimating target image feature ϕ from the unified representation of source image X^s and its editing request I. The encoder part consists of source image encoder E_{im}, instruction encoder E_i, and a 1-layer fully-connected layer FC. Function $\phi = f(X^s, I)$ is defined:

$$\phi^{im} = E_{im}(X^s), \tag{6}$$

$$\phi^{i} = E_{i}(I), \tag{7}$$

$$\begin{aligned}\phi &= f(X^s, I)\\ &= FC(\phi^{im}, \phi^i)\\ &= sigmoid(W_{im}\phi^{im} + W_i\phi^i).\end{aligned} \tag{8}$$

We use 4-layer convolutional neural networks [10] for E_{im} and 1-layer long short-term memory neural networks [11] for E_i. Assuming I consists of word tokens $I = (w_1, w_2, \cdots, w_T)$, (7) is achieved by $\phi^i_t = LSTM(w_t, \phi^i_{t-1})$, where $\phi^i_0 = 0$ and $\phi^i = \phi^i_T$.

Then we rewrite (1) and (2):

$$\hat{X} = G(z, \phi), \tag{9}$$

$$\hat{y} = D(x, \phi). \quad (x \in \{X, \hat{X}\}). \tag{10}$$

Condition ϕ is fed into both the generator and the discriminator. This formulation is necessary for training a conditional DCGAN by a matching aware method [12]. This formulation enables the discriminator to classify whether the input image corresponds to the input condition, and the generator to learn the mapping between the generated image and the condition.

The objective function of the discriminator (defined in (4)) is extended by the following three functions:

$$\mathcal{L}_{D_{Xr}} = -\mathbb{E}_{X\sim p_{data}}[\log D(X, c_r)], \tag{11}$$

$$\mathcal{L}_{D_{Xw}} = -\mathbb{E}_{X\sim p_{data}}[\log(1 - D(X, c_w))], \tag{12}$$

$$\mathcal{L}_{D_{\hat{X}r}} = -\mathbb{E}_{X\sim p_{data}}\mathbb{E}_{z\sim p_z}[\log(1 - D(G(z, c_r), c_r))]. \tag{13}$$

The objective function of the generator (defined in (5)) is also rewritten:

$$\mathcal{L}_{G_{\hat{X}r}} = -\mathbb{E}_{X\sim p_{data}}\mathbb{E}_{z\sim p_z}[\log D(G(z, c_r), c_r)]. \tag{14}$$

The notations c_r and c_w are used for a condition that corresponds to a training example X and for a condition that does not correspond to a training example X, respectively. Objective (11) encourages the discriminator to classify the matched pair of the training example and the condition as real. Objective (12) encourages the discriminator to classify the mismatched pair of the training example to the condition as fake. Objective (13) encourages the discriminator to classify the matched pair of the generated image and the condition as fake. Objective (14) encourages the generator to trick the discriminator into classifying the matched pair of the generated image and the condition as real. In summary, the discriminator not only learns to correctly classify the input image itself as real or fake but also to classify between input images that correspond and do not correspond to the conditions.

The model requires triplet (c_r, c_w, X) in training. We have to select c_r to be the target image feature and c_w to be far away from the target image feature. Suppose that the training examples are composed of triplets (X^s, X^t, I), where X^s indicates the source image and X^t represents the target image that corresponds to the given input pair of X^s and editing request I. One choice of triplet could be $(c_r, c_w, X) = (f(X^s, I), f(X^s, 0), X^t)$, where $I = 0$ represents $\phi^i = 0$ in practice. We suppose that $f(X^s, 0)$ is editing with a meaningless editing request. To ensure that $f(X^s, 0)$ results in a value far from target image feature c_r, we use additional patterns of triplets $(c_r, c_w, X) \in \{(f(X^s, 0), f(X^t, 0), X^s), (f(X^t, 0), f(X^s, 0), X^t)\}$. This step encourages the model to learn an identity mapping between the source and target images if the given editing request is meaningless ($I = 0$).

Therefore, we define the overall objectives:

$$\min_{\theta_D, \theta_{Enc}} \mathcal{L}_D = \lambda_{Xr}\mathcal{L}_{D_{Xr}} + \lambda_{Xw}\mathcal{L}_{D_{Xw}} + \lambda_{\hat{X}r}\mathcal{L}_{D_{\hat{X}r}}, \tag{15}$$

$$\min_{\theta_G, \theta_{Enc}} \mathcal{L}_G = \lambda_{g\hat{X}r}\mathcal{L}_{G_{\hat{X}r}} + \lambda_f\mathcal{L}_{fmatch}. \tag{16}$$

θ_D, θ_G, and θ_{Enc} are the trainable parameters of D, G, and the encoder part, respectively. L_D and L_G are the objectives for training D and G. In each iteration, the model uses (15) if $L_D > L_G$, and otherwise it uses (16). Note that $\mathcal{L}_{fmatch}$ represents the objective of the feature matching [13] to stabilize the training of G and D. It is achieved by the sum of the layer-wise mean squared errors between the latent

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997012, IEEE Access

features in D extracted from real image X and that from generated image $\hat{X}$. λ_{Xr}, λ_{Xw}, $\lambda_{\hat{X}r}$, $\lambda_{g\hat{X}fr}$, and λ_f are the coefficients of each objective. We use 1.0 for each coefficient.

C. DCGAN FOR IMAGE EDITING WITH A CONSTRAINT

The image editing model based on DCGAN sometimes offers drastic changes to the source image, which are inappropriate for a cooperative process with users. To prevent this problem, we introduce an additional module called Source Image Masking (SIM) [7], which functions as a constraint on DCGAN for image editing. The SIM idea is to explicitly indicate the editing points on the source image with masking. SIM is composed of two parts, mask generator G_m and image encoder with mask E_{imm}. We next define the procedure for generating and forwarding a mask :

$$m_{mono} = G_m(\phi^{im}, \phi^i), \tag{17}$$

$$\phi^{imm} = E_{imm}(X^s \odot m_{color}). \tag{18}$$

m_{color} is a channel-wise copied mask from mono-channel mask m_{mono}. We utilized m_{mono} for the entropy calculation, which decides on the system's dialogue strategy in Section III-D. $\odot$ indicates the Hadamard product. ϕ^{imm} is fed into G as additional input. Rewriting (9), we get

$$\hat{X} = G(z, \phi, \phi^{imm}). \tag{19}$$

D. SYSTEM'S CONFIRMATION OF ACTION DECISIONS

Confirmation, which shows multiple editing results to users from multiple models, is a safe action described in Section I. However, the user must pay additional cost for responding to the confirmation. When a confirmation action must be selected, basing it on some uncertainty scores of image generation will smooth the dialogue. We use the entropy scores of the generated image as the uncertainty scores and calculate the entropy:

$$entropy = -\frac{1}{WH}\Sigma_i^W \Sigma_j^H \{m_{ij}\log(m_{ij}) + (1-m_{ij})\log(1-m_{ij})\} \leq -\log 0.5. \tag{20}$$

We define m_{ij} as the value of the predicted mask at the (i,j)-th position with width W and height H. $-\alpha \log 0.5$ $(0 \leq \alpha \leq 1)$ is our confirmation threshold. The mixed model selects $confirm$ if $entropy \geq -\alpha \log 0.5$. We tried several α in our experiment.

IV. EXPERIMENTAL SETTINGS

We conducted experimental dialogues to investigate the effectiveness of our proposed dialogue strategy. In this section, we describe the dataset for the image editing dialogues, the training details of each model, and the user evaluation settings.

A. DATASET

For training w/ and w/o mask models and evaluation, we utilized the Avatar Image Manipulation with an Instruction dataset [7]. The task is portrait image editing based on instructions, which involve natural language editing requests. The data consist of 22 types of editing, e.g., changing a beard, eyebrows, and hair. Each sample is composed of a triplet of {source image, target image, instruction (editing request)}. We split the dataset into $train : validation : test = 4,296 : 230 : 230$ according to existing work. We also used 161,065 examples composed of one image sample to improve the generator's image modeling.

B. TRAINING MODELS

During training, we alternatively repeated the training of the image generator and the image editing. In the image generator training phase, we trained the model as an auto-encoder to generate the same image to the given source image for stabilizing the generator. We also set the instruction vector to zero in this training phase. This process enhances the generator's ability to generate clear images. In the image editing training phase, we utilized full triplets of {source image, target image, instruction}. The dataset consists of the editing requests that represent only one attribute change such as hair change; thus, we can prepare the ground truth of the mask by comparing a pair of source and target images to improve SIM's mask generator G_m training. We used the ground truth mask in the training, whose pixels were set to zero where the pixels in the same position of the source and target images are different, or otherwise they are set to one. We also provided a mask loss function as mean squared error between the generated mask and the ground truth one to improve the SIM model. We trained the models using $Adam$ [14] ($\alpha = 2.0 \times 10^{-4}, \beta = 0.5$) until $5,000$ phases. The images were resized to 64×64. The following are the hidden sizes: 128 for ϕ^i and ϕ, 1024 for ϕ^{im}, and $512 \times 4 \times 4$ for ϕ^{imm}. The batch size is 64, and the vocabulary size is 1892.

C. EVALUATION METRIC FOR IMAGE QUALITY

We utilized Structured Similarity (SSIM) [15] to evaluate the improvement of the image quality that represents the similarity between generated image X and goal image Y. We calculated SSIM between images X and Y as follows:

$$SSIM_{ch}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{21}$$

$$SSIM(X,Y) = \frac{1}{CN}\sum_{ch=1}^{C}\sum_{i=1}^{N} SSIM_{ch}(x_{i,ch}, y_{i,ch}). \tag{22}$$

$x_{i,ch}$ and $y_{i,ch}$ are the i-th local patches of each RGB channel ch of image X and Y. Whole patches are derived by vertically and horizontally sliding a squared window with width L one-by-one. μ_x, μ_y are their mean, and σ_x^2, σ_y^2, and σ_{xy} are their variance and co-variance. C_1, C_2 are constant values. For the whole experiment, we adopted commonly used parameters: $L = 7$, $C_1 = (255 \cdot 0.01)^2$, $C_2 = (255 \cdot 0.03)^2$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997012, IEEE Access

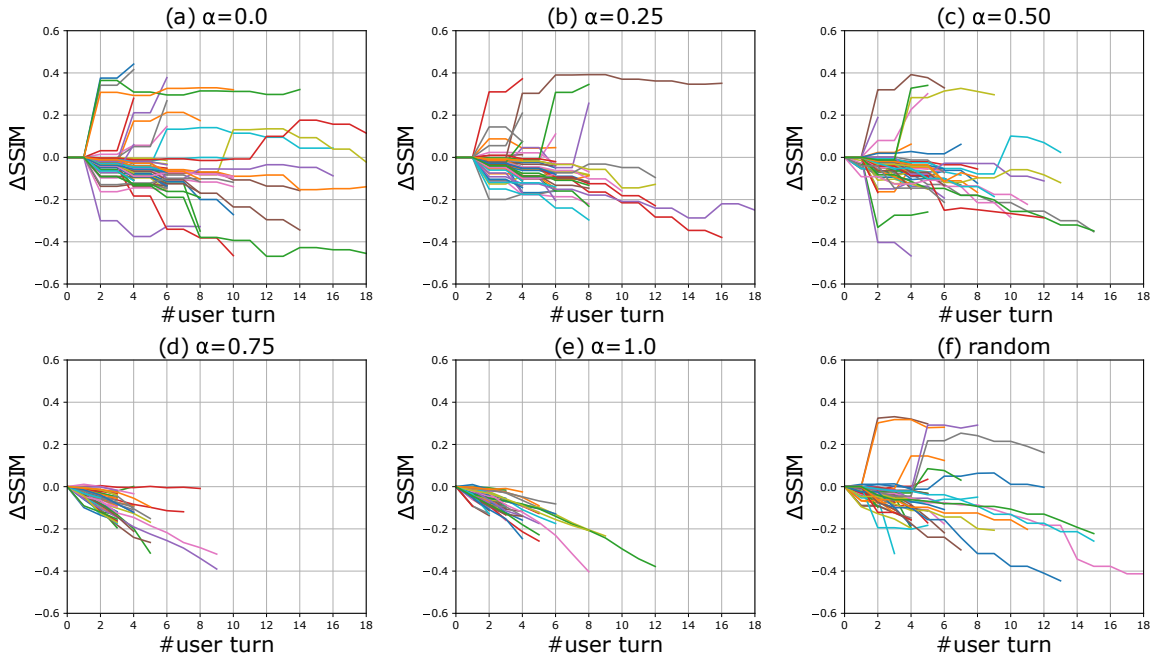

FIGURE 2. Experimental results of image editing dialogue between 18 evaluators (users) and the system: $\#user\ turn$ denotes total number of user actions (*making an editing request* and *selecting an image*); (smaller is better). $\Delta SSIM$ denotes current source-goal SSIM, subtracted by first source-goal SSIM (higher is better). Each plot in figures represents each dialogue sample. α indicates threshold for system to select confirmation: (a) $\alpha = 0.0$, (b) $\alpha = 0.25$, (c) $\alpha = 0.50$, (d) $\alpha = 0.75$, (e) $\alpha = 1.0$, and (f) random: system randomly selects confirmation. If α becomes smaller, system tends to select confirmation with a lower uncertainty score. Note that every $\Delta SSIM$ is calculated after the user's action. Therefore, when the system selects confirmation after the user makes an editing request, $\Delta SSIM$ keeps the same value. Degradation as dialogues progress is caused by image editing models.

D. USER EVALUATION OF IMAGE EDITING DIALOGUE

In a pilot study, we found that the w/ mask model tends to successfully edit a small region in a single turn, such as changing eye color or adding a mustache or glasses. However, the w/ mask model often fails to edit a large region of the source image, such as changing hairstyle. Therefore, we focused on hair editing to evaluate the image editing dialogue. We evaluated our proposed confirmation strategy on two aspects. First, we evaluated the necessity of confirmation by comparing between the strategy without confirmation using the w/ mask model and strategy with confirmation using both the w/o and w/ mask models. Second, we evaluated the effectiveness of the confirmation strategy by comparing the strategy without confirmation or a random strategy with the others. We used 21 patterns (9 for male portraits and 12 for female portraits) as pairs of source and goal images, and conducted image editing dialogue experiments with human evaluators. The evaluators were 18 people whose TOEIC scores exceeded 730 and could use English for daily use. At the task's beginning, the evaluators looked at the source and goal images and talked with our interactive image editing system, which has different dialogue strategies. Each pattern was evaluated by three evaluators over the following six strategies: the system selected $confirm$ with thresholds $\alpha = 0.0, 0.25, 0.50, 0.75, 1.0$ (as described in Section III-D) and randomly selected $confirm$. We compared these different strategies to identify the effectiveness of our proposed method on the problem of interactive image editing. Note that α represents proactiveness for confirmation: when $\alpha = 0.0$, the system selects $confirm$ every time; and $\alpha = 1.0$, it selects $not\ confirm$ every time. In other words, $\alpha = 1.0$ corresponds to the case where the system uses the w/ mask model every time.

1) Necessity of confirmation (limitation of a single model)

Confirmation is useful when the system needs to deal with multiple editing results from multiple models. It is difficult for a single editing model to accept every editing request because a trade-off exists between editing flexibility and the model constraints. We first investigated how the single w/ mask model works on an interactive image editing task. We compared models with different confirmation strategy settings for the improvement of image quality through dialogues (higher is better).

2) Effectiveness of confirmation strategy

Second, we investigated the effectiveness of our proposed confirmation strategy. If our confirmation method works with appropriate timing, it will improve performance (higher image quality with shorter dialogue length).

V. RESULTS

Next we describe and discuss our experimental results in two parts in Sections IV-D1 and IV-D2.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997012, IEEE Access

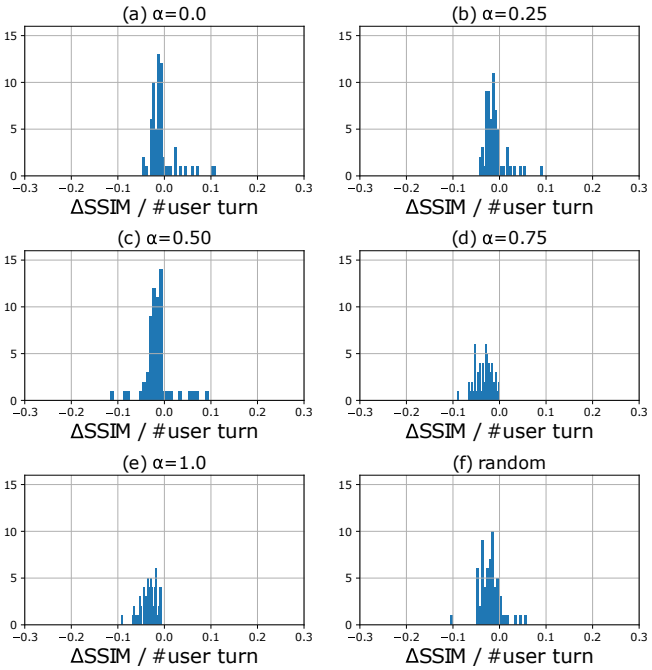

FIGURE 3. Histogram of $\Delta SSIM/\#user\ turn$ at end of dialogues on each strategy: higher $\Delta SSIM/\#user\ turn$ dialogue created more similar images to goal and more efficient dialogues.

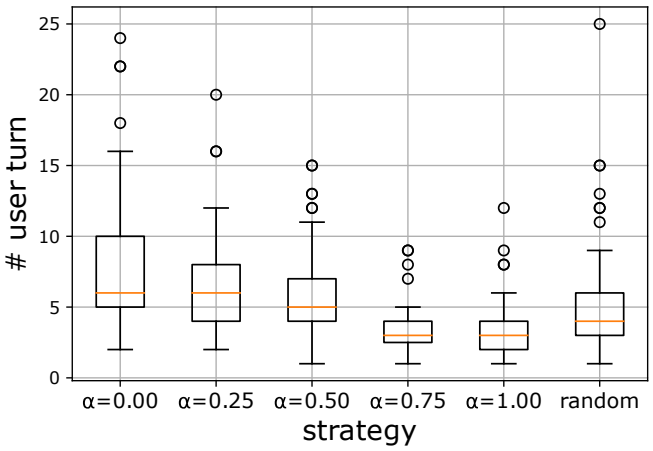

FIGURE 4. Distribution of $\#user\ turn$ on each strategy: smaller $\#user\ turn$ dialogue represents more efficient dialogue.

1) Necessity of confirmation (limitation of a single model)

Figure 2 indicates the relative changes of SSIM from the current image to the goal image and plots the overall dialogue on each setting as the dialogue progressed. $\#user\ turn$ denotes the total number of the user actions of making an editing request and selecting an image (smaller is better). $\Delta SSIM$ denotes relative SSIM, which is subtracted from the first source-goal's SSIM. i-th turn's $\Delta SSIM$ is defined as $\Delta SSIM = SSIM(X_i^s, X^g) - SSIM(X_0^s, X^g)$ (higher is better). The result with a higher α, such as $\alpha = 0.75$, indicates almost the same behavior to $\alpha = 1.0$, which corresponds to just using the w/ mask model. $\Delta SSIM$ worsened as the dialogue progressed due to the image editing models, which were trained with single turn editing triplets of {source image, target image, editing request}. In other words, the models were inadequately generalized to the degraded source images. Thus, degradation, which occurred in a turn, tended to be gradually strengthened in the next turn. On the other hand, the results with lower α, such as $\alpha = 0.0$, $\alpha = 0.25$, and $\alpha = 0.50$, indicate some dialogue examples achieved a better SSIM than before the dialogue. This indicates that the w/o mask model is necessary to get better SSIM scores to change a larger region, such as a woman's hair.

2) Effectiveness of confirmation strategy

An effective dialogue strategy satisfies not only the improvement of the image quality but also the efficiency of image editing dialogue; a shorter dialogue is better. To evaluate the whole dialogue performance in these two aspects, we visualized the histogram of $\Delta SSIM/\#user\ turn$ collected from the end of the dialogues (Figure 3). We applied Mann–Whitney U test [16] to compare (a) $\alpha = 0.0$, (b) $\alpha = 0.25$, (c) $\alpha = 0.50$, and (d) $\alpha = 0.75$ with (e) $\alpha = 1.0$ and (f) $random$, and found significance of p-value < 0.001 on the following: (a) $\alpha = 0.0$ and (e) $\alpha = 1.0$, (b) $\alpha = 0.25$ and (e) $\alpha = 1.0$, (c) $\alpha = 0.50$ and (e) $\alpha = 1.0$, and (a) $\alpha = 0.0$ and (f) $random$. This result indicates that the strategies with (a) $\alpha = 0.0$ and (b) $\alpha = 0.25$ realized a better SSIM with fewer dialogue turns than the strategies with rarely confirming strategies ((d) $\alpha = 0.75$ and (e) $\alpha = 1.0$) or random confirmation strategy.

However, (a) $\alpha = 0.0$ and (b) $\alpha = 0.25$ were confirmed in most cases. When we compared all combinations of the two strategies in {(a) $\alpha = 0.0$ and (b) $\alpha = 0.25$, and (b) $\alpha = 0.50$}, they were comparable. We showed the distribution of $\#user\ turn$ for each strategy in Figure 4 to compare their effectiveness. We found a significance of p-value < 0.001 between (c) $\alpha = 0.50$ and (a) $\alpha = 0.0$, indicating that (c) $\alpha = 0.50$ was a more efficient dialogue.

Although (c) $\alpha = 0.50$ was not significant compared with (f) $random$, we found some interesting cases where the system used $confirm$ and $not\ confirm$ more properly than random. Figure 5 shows a dialogue example where the user discovered a good strategy. First, they tried to change the hair to a ponytail. The system successfully generated a ponytail image, but unintentionally changed the eyes to green. The user asked the system to change the eyes back to blue, and it successfully obeyed without any redundant confirmation on this turn. On the other hand, with the random confirmation strategy, the system occasionally confirmed with inappropriate timing. For example, Figure 6 indicates an inefficient case. The system should have used $confirm$ for the editing request on $i = 2$, which indicate requests for changing to a smaller part. The user cannot fundamentally avoid such cases with the random confirmation strategy.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.2997012, IEEE Access

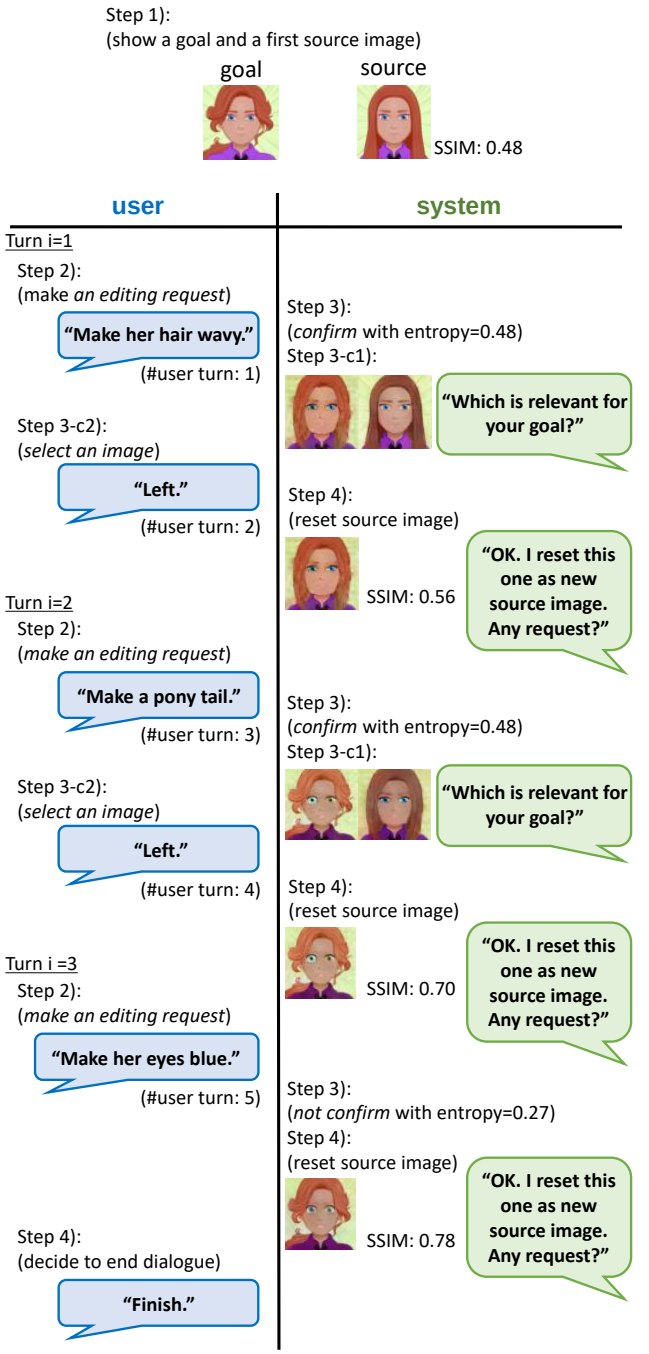

FIGURE 5. Dialogue example with $\alpha = 0.50$ (confirmation threshold $-\alpha \log 0.5 = 0.35$). i indicates turn index defined in Section II. $\#user\ turn$ denotes number of user actions, which represents total number of *making an editing request* and *selecting an image*. We put the source-goal SSIM next to each source image when the system decides on a generated image for each turn.

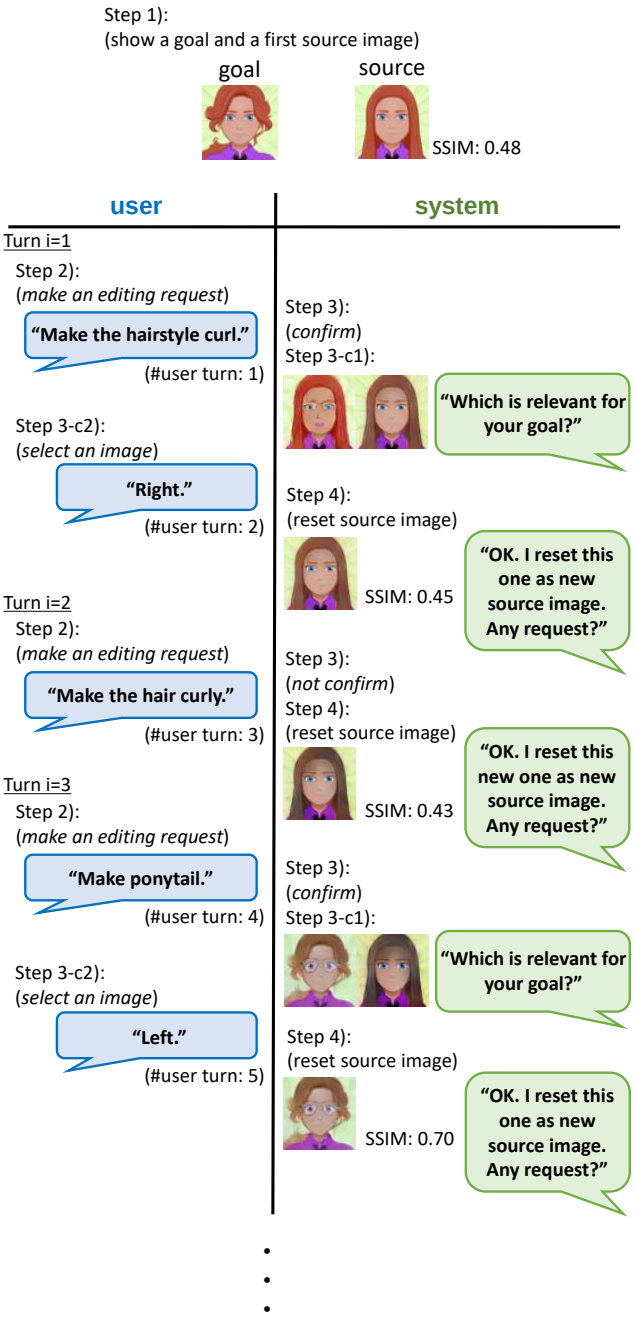

FIGURE 6. Inefficient dialogue example with random confirmation: i indicates turn index defined in Section II. $\#user\ turn$ denotes number of user actions, which represents total number of *making an editing request* and *selecting an image*.

VI. RELATED WORKS

A. VISION AND DIALOGUE

Vision and dialogue is an emerging topic of intersection field between computer vision and natural language processing. Conversational image editing system research [17], [18] attempts to understand the user utterance and identify the user's intention in an interactive image editing task using existing image editing software such as Adobe Photoshop and OpenCV. Our proposed method has the same motivation to identify the user's intention; however, our editing system is based on image generative models. Image generative models potentially enable the system to edit images more flexibly

but they have difficulty to handle their generated images. Our proposed confirmation method provides a means to handle the generated images.

B. CONFIRMATION STRATEGY IN DIALOGUE

Confirmation strategy has mainly been investigated in the spoken dialogue system research field [19], [20]. Such spoken dialogue systems need to consider the mistakes of speech recognition or natural language understanding. In this situation, confirmation effectively manages the dialogue process. The confirmation method, based on confidence measures [19], calculates the confidence score of each content word in the speech recognition candidates. The system asks the user for confirmation when the confidence for the content word existence in the user utterance is uncertain. Similarly, our proposed confirmation method provides a confidence score for confirmation. However, the calculation of confidence scores is based on the entropy of the image editing model. The confirmation method for a document retrieval dialogue task is based on minimizing the Bayes risk [20]. It requires a classification model to calculate the Bayes risk. In contrast, our entropy-based method does not require any additional model or dialogue data for training the model.

C. UNCERTAINTY DETECTION FOR GAN-BASED IMAGE GENERATION

Controlling generated images is an essential problem in GAN-based image generation because of the instability of the generated image quality. To stabilize the image quality, the truncation trick, which restricts the acceptable sample on latent space z, performs well in conditional image generation [21]. However, it does not provide any information about the uncertainty. Our entropy-based method provides uncertainty scores for the generated images.

Uncertainty detection for GAN-based models has been scrutinized in anomaly detection [22], [23]. However, it measures the distances between the generated images and the samples in a training dataset without indicating their suitability for the given condition. Our entropy-based method is based on a mask, which is made from the given condition, that can provide a confidence score that represents the suitability of the generated image for the given condition.

VII. CONCLUSION

We proposed an entropy-based confirmation method using a masking mechanism for interactive image editing. The mask mechanism is useful for dealing with such complicated conditions as natural language, but such a strong constraint limits the acceptable language requests. In an avatar image editing task with natural language editing requests, changing such vast regions as hair is restricted in the w/ mask constraint model. The system's capability to confirm an action provides a chance to select a relevant image generated from both the w/o and w/ mask models. We demonstrated that our proposed strategy led to more similar images with fewer dialogue turns during human evaluations. We also showed an interesting case where our confirmation method achieved an efficient dialogue strategy. It first changed a large part and then fine-tuned a small part. In future work for more effective dialogues, we will collect dialogue data and enable our system to learn adaptive strategies using reinforcement learning, for example. Another future direction is applying our method to more natural/photo-realistic image datasets. Masking mechanisms are effective in image-to-image translation tasks with these datasets [4], [5], [6]; thus, we expect our method also works well with them.

REFERENCES

[1] A. Brock, T. Lim, J. Ritchie, and N. Weston, "Neural photo editing with introspective adversarial networks," in The International Conference on Learning Representations, 2017.

[2] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in Proceedings of European Conference on Computer Vision, 2016.

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

[4] X. Liang, H. Zhang, L. Lin, and E. Xing, "Generative semantic manipulation with mask-contrasting gan," in The European Conference on Computer Vision, pp. 558–573, 2018.

[5] Y. A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim, "Unsupervised attention-guided image-to-image translation," in Advances in Neural Information Processing Systems, pp. 3693–3703, 2018.

[6] S. Mo, M. Cho, and J. Shin, "Instagan: Instance-aware image-to-image translation," in The International Conference on Learning Representations (ICLR), 2019.

[7] S. Shinagawa, K. Yoshino, S. Sakti, Y. Suzuki, and S. Nakamura, "Image manipulation system with natural language instruction," IEICE Transactions on Information and Systems, Vol.J102-D, No.8, pp.514–529, 2019.

[8] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in The International Conference on Learning Representations, 2016.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems, pp. 2672–2680, 2014.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text-to-image synthesis," in The International Conference on Machine Learning, 2016.

[13] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in Advances in Neural Information Processing Systems, pp. 2226–2234, 2016.

[14] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in The International Conference on Learning Representations, 2015.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

[16] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," The annals of mathematical statistics, pp. 50–60, 1947.

[17] R. Manuvinakurike, T. Bui, W. Chang, and K. Georgila, "Conversational image editing: Incremental intent identification in a new dialogue task," in Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, (Melbourne, Australia), pp. 284–295, Association for Computational Linguistics, July 2018.

[18] T.-H. Lin, T. Bui, D. S. Kim, and J. Oh, "A multimodal dialogue system for conversational image editing," in Proceedings of The Second Workshop on Conversational AI at the Thirty-second Conference on Neural Information Processing Systems (NeurIPS 2018), November 2018.

[19] K. Komatani and T. Kawahara, "Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recog-

nizer output," in Proceedings of the 18th conference on Computational linguistics-Volume 1, pp. 467–473, Association for Computational Linguistics, 2000.

[20] T. Misu and T. Kawahara, "Bayes risk-based dialogue management for document retrieval system with speech interface," Speech Communication, vol. 52, no. 1, pp. 61–71, 2010.

[21] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in The International Conference on Learning Representations, 2019.

[22] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in International conference on information processing in medical imaging, pp. 146–157, Springer, 2017.

[23] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in Asian Conference on Computer Vision, pp. 622–637, Springer, 2018.

SEITARO SHINAGAWA received his B.E degree and M.S. degree in information science in 2013 and 2015 from Tohoku University, respectively. From 2015 to 2020, he was a Ph.D. student of the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST). From 2020, he is a researcher in NAIST. His working area is visually-grounded dialogue systems, especially interactive image generation system with dialogue. He is a member of JSAI.

KOICHIRO YOSHINO received his B.A. degree in 2009 from Keio University, M.S. and Ph.D. degrees in informatics in 2011 and 2014 from Kyoto University, respectively. From 2014 to 2015, he was a research fellow (PD) of Japan Society for Promotion of Science. From 2015, he is an assistant professor of the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST). He is also a researcher of PRESTO, JST, concurrently. He is working on areas of spoken and natural language processing, especially on spoken dialogue systems. Dr. Koichiro Yoshino received the JSAI SIG-research award in 2013, the ANLP outstanding paper award in 2018, and the ACL-NLP4ConvAI workshop best paper award in 2019. He is a member of IEEE, ACL, ISCA, IPSJ, ANLP and RSJ.

SEYED HOSSEIN ALAVI, also known as Soheil, is a Ph.D. student in Computer Science at University of Southern California since 2016. He received his B.Sc. in Computer Science in 2016 from Sharif University of Technology. Concurrently, he also received his second B.Sc. in Aerospace Engineering from the same university. Currently, Soheil is a member of Natural Language Dialogue group at Institute for Creative technologies working under the supervision of Dr. David Traum. He is a member of New York Academy of Sciences (NYAS). Soheil's research interests include spoken dialogue systems, multimodal systems, response generation and retrieval, deep learning, and natural language processing.

KALLIRROI GEORGILA is a Research Associate Professor in the Department of Computer Science at the University of Southern California (USC) and at the USC Institute for Creative Technologies. Before joining USC she was a Research Scientist at the Educational Testing Service in Princeton, USA, and before that a Research Fellow at the School of Informatics of the University of Edinburgh, in the United Kingdom. Her research interests include all aspects of natural language dialogue processing with a focus on machine learning, particularly reinforcement learning of dialogue policies, speech recognition, and expressive conversational speech synthesis. She has served on the organizing, senior, and program committees of many conferences and workshops. She has also served as Vice President of SIGdial (the Special Interest Group on Discourse and Dialogue). Currently she is an Associate Editor of the Dialogue and Discourse journal and on the Editorial Board of the Computational Linguistics journal.

DAVID TRAUM is the Director for Natural Language Research at the Institute for Creative Technologies (ICT) and Research Professor in the Department of Computer Science at the University of Southern California (USC). He leads the Natural Language Dialogue Group at ICT. More information about the group can be found here: http://nld.ict.usc.edu/group/ Traum's research focuses on Dialogue Communication between Human and Artificial Agents. He has engaged in theoretical, implementational and empirical approaches to the problem, studying human-human natural language and multi-modal dialogue, as well as building a number of dialogue systems to communicate with human users. Traum has authored over 250 refereed technical articles, is a founding editor of the Journal Dialogue and Discourse, has chaired and served on many conference program committees, and is a past President of SIGDIAL, the international special interest group in discourse and dialogue. Traum earned his Ph.D. in Computer Science at the University of Rochester in 1994.

SAKRIANI SAKTI received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with the Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003-2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006-2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005-2008) with Dialog Systems Group University of Ulm, Germany, and received her Ph.D. degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003-2007), A-STAR, and U-STAR (2006-2011). In 2009-2011, she served as a visiting professor of the Computer Science Department, University of Indonesia (UI), Indonesia. In 2011-2017, she was an assistant professor at the Augmented Human Communication Laboratory, NAIST, Japan. She also served as a visiting scientific researcher of INRIA Paris-Rocquencourt, France, in 2015-2016, under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation". Currently, she is a research associate professor at NAIST, as well as a research scientist at RIKEN, the Center of for Advanced Intelligent Project AIP, Japan. She is a member of JNS, SFN, ASJ, ISCA, IEICE, and IEEE. She is also the officer of ELRA/ISCA Special Interest Group on Under-resourced Languages (SIGUL) and a board Member of Spoken Language Technologies for Under-Resourced Languages (SLTU). Her research interests include statistical pattern recognition, graphical modeling framework, deep learning, multilingual speech recognition & synthesis, spoken language translation, affective dialog system, and cognitive communication.

SATOSHI NAKAMURA is Professor of Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan, Project Leader of Tourism Information Analytics Team of RIKEN, Center for Advanced Intelligence Project AIP, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008 and Vice president of ATR in 2007-2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009-2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT, and A-STAR. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received the LREC Antonio Zampolli Award 2012. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011, IEEE Signal Processing Magazine Editorial Board Member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016, and ISCA Fellow since 2020.