Interactive Subspace Exploration on Generative Image Modelling

TOBY CHONG LONG HIN, The University of Tokyo
I-CHAO SHEN, National Taiwan University
ISSEI SATO, The University of Tokyo
TAKEO IGARASHI, The University of Tokyo

Generative image modeling techniques such as GAN demonstrate highly convincing image generation result. However, user interaction is often necessary to obtain desired results. Existing attempts add interactivity but require either tailored architectures or extra data. We present a human-in-the-optimization method that allows users to directly explore and search the latent vector space of generative image modeling. Our system provides multiple candidates by sampling the latent vector space, and the user selects the best blending weights within the subspace using multiple sliders. The system samples latent vectors based on inputs and presents new candidates to the user iteratively. An advantage of our formulation is that one can apply our method to arbitrary pre-trained model without developing specialized architecture or data. We demonstrate our method with various generative image modeling applications, and show superior performance in a comparative user study with prior art iGAN [Zhu et al. 2016].

Additional Key Words and Phrases: Human-in-the-loop Optimization, Bayesian Optimization, Generative Image Modeling

ACM Reference Format:
Toby Chong Long Hin, I-Chao Shen, Issei Sato, and Takeo Igarashi. 2019. Interactive Subspace Exploration on Generative Image Modelling. *ACM Trans. Graph.* 1, 1 (June 2019), 9 pages. https://doi.org/10.1145/nnnnnn.nn

1 INTRODUCTION

With the growing maturity of generative image modeling techniques such as generative adversarial networks [Goodfellow et al. 2014] (GANs) and auto-encoder architecture [Kingma and Welling 2013], many applications have been proposed, including high quality face generation [Karras et al. 2018], image inpainting [Iizuka et al. 2017], image generation from text [Tao Xu 2018], and image style transfer and translation [Huang et al. 2018; Isola et al. 2017; Zhu et al. 2017]. These applications produce images from a random latent vector and additional inputs (e.g. a sketch or a sentence) that exhibit characteristics observed in the training data.

A common problem with these approaches is the lack of controllability of the generated results. When the user detects a defect in the generated result (e.g. artifacts, unwanted image attributes), it is often difficult to fix such problem. This is particularly relevant to people not familiar with machine learning techniques. For example, some artists create art using GAN such as face portrait, abstract art, garment design and more (Figure 2). With more publicly released machine learning resources, such as source codes and pretrained network model, practicalists can simply download them and generate images with it without understanding all the theory of behind the algorithm. However, interfacing with generative image modeling to obtain images with high user satisfaction has remain a challenge due to the lack of an unified user interface.

Fig. 1. Our system provides additional control to a pretrained and frozen generative image modeling network without modifying the network itself. The user starts with an input signal (a). They explore the latent space with sliders and pick the best result within slider space (b). They can optionally provide a guidance image (c). The system provides image candidates that align with the user input (d).

Fig. 2. Artists create different art forms with the help of generative image modeling : (a) printed art based on GAN, and (b) fashion design using GAN.
Some existing approaches allow interactive control of the image generation process, e.g., using color point cues for image colorization [Zhang et al. 2017], or sketches as structural cues for image translation [Isola et al. 2017] and face image editing [Portenier et al. 2018]. However, it either requires new network architecture usually tailored to allow only specific controls. Second, the tailored network must be trained using additional data such as labels.

iGAN [Zhu et al. 2016] addresses these challenges by building a natural image manifold and allowing the user to explore the manifold by drawing a desired image. They directly optimize for images that matches the user stroke. Although their approach is straightforward, we observed that drawing of users are usually too poor to properly guide the generation process (Fig 9).

In this paper, we address these challenges by allowing the user to directly explore the latent vector space using human-in-the-loop optimization. The user provides feedback to candidates presented by the system by manipulating sliders and drawing images. Slider control allows the user to explore the subspace around the desired point in the latent vector space. Image editing tools allow the user to directly indicate desired changes to the system, such as to the colors used, and to identify regions that should be preserved and problematic regions that should be removed. The system takes the feedback and provide new candidates that matches the user input. This process repeats until the user obtains a satisfying image.

Our method considers pre-train networks as black-box functions and adds interactivity regardless of network architecture. Hence, it is applicable to a board range of existing works as well as future architectures, which is critical as image generative model is evolving in an immense pace.

Our work builds on human-in-the-loop parameter tweaking for image editing [Koyama et al. 2017]. A notable difference is that our target space is much larger (512d) than theirs (10d).

Exploring high dimensional space with Bayesian Optimization could be impractical due to the high sample count requirement. Therefore, we propose two novel user interface elements together with appropriate algorithms to assist the optimization. First, we use multiple sliders rather than a single slider to allow the user to explore a larger subspace than the 1-D subspace of the original method. This allows the system to attain an optimum more efficiently. Second, we introduce a content-aware sampling strategy that favors results aligns with user edits. We achieve this by formulating different image operations as content-aware bias term, and adding it to the acquisition function in Bayesian optimization which provides the next “best” candidates when optimized.

We demonstrate the effectiveness and versatility of our framework to three image generation applications: image generation, image translation, and text-to-image generation, together with other user studies and ablation studies.

To summarize, the key contributions of this paper are:

- Introducing a human-in-the-loop optimization framework for guiding generative image modelling, where the user directly explores and search the latent space assisted by the system.
- Imparting controllability to generative image modeling without requiring tailoring of the network architecture and additional training.
- Two extensions to human-in-the-loop Bayesian optimization [Koyama et al. 2017]: (i) sequential subspace search effectively exploring the high-dimensional latent vector space, and (ii) the content-aware sampling strategy that favours images aligning with user edits.

2 RELATED WORK

2.1 Interactive Generative Image Modeling

One promising approach to incorporate different user inputs is to use conditional GAN to enable inputs such as labels or aerial images, to solve image-image translation problem [Isola et al. 2017] and sketching of terrain authoring system [Guérin et al. 2017]. Zhang et al. [2017] created additional local and global hint networks to incorporate local and global user inputs. Portenier et al. developed FaceShop [2018], a novel network architecture combining both image completion and translation in a single framework. The user draws strokes, using both geometry and color constraints to guide face editing. However, these methods require tailored network architectures and training data to endow specific applications with controllability. To many practitioner such as artists, providing meaningful label data is either financially impractical or impossible due to the abstract nature of their work. We address this problem by developing a generic framework that introduces additional controllability to pretrained models.

iGAN [Zhu et al. 2016] is closely related to our method. They provide a blank canvas on which the user can draw line sketches, paint colors, and warp the image content. However, as the user provides more input, the system continuously optimizes some randomly initialized latent vectors, which generate images that matches the user’s edits.

Another problem with iGAN is that it relies solely on user drawn inputs, but casual users often fail to express their intentions accurately by drawing (Figure 9). Our system therefore allows the user to directly explore the latent vector space by using sliders. The user only “evaluates” images presented by the system and “selects” the best ones rather than “drawing” images accurately. Another difference is that iGAN initializes candidates by random sampling and tends to converge to similar ones as user add more input. In contrast, Bayesian optimization in our work provides candidates that searches the space efficiently.

GAN Dissection [Bau et al. 2018] provides a framework for visualizing and understanding the structures learned by generative networks and provides users with an intuitive painting interface so that they can manipulate generated images by painting objects directly. Unlike our method, their method requires a separate network and segmentation masks to identify the function of specific neurons inside the generative network. Neural Collage [Suzuki et al. 2018] enables users to change the attributes of an image (e.g. change the breed of dog, color of petals). It requires both specific architecture and additional image labels, together with a manually created
mask, to combine the processed result with an original image. GAN-Breeder\textsuperscript{3} is a community project in which users can simply keep selecting the most interesting image to discover totally new images. Yet, an ideal system “understands” user even with minimal user interaction.

2.2 Bayesian Optimization with Gaussian Process

Bayesian optimization (BO) is a framework designed to optimize expensive-to-evaluate black-box functions with minimal number of evaluations:

$$\max_x f(x),$$

where $f(x)$ is a black-box function with unknown derivatives and convexity properties.

During each trial, BO provides a new sample candidates based on prior observations; more specifically, it optimizes an acquisition function using a predefined prior; in this work we focused on BO with a Gaussian Process (GP) as the prior, which is commonly used for such task [Koyama et al. 2017]. The acquisition function seeks the next sample candidate that maximizes a criterion such as expected improvement (EI) [Brochu et al. 2010], knowledge gradient (KG) [Scott et al. 2011], or variations thereof; please refer to [Shahriari et al. 2016] for details. As BO effectively approximates arbitrary functions, it is applicable to high-dimensional parameter spaces exploration, such as hyperparameter tuning systems for machine learning algorithms [Bergstra et al. 2011; Koyama et al. 2017], photo enhancement [Koyama et al. 2017], material BRDF design [Brochu et al. 2007; Koyama et al. 2017].

Bayesian optimization with inequality constraints has been proposed [Gardner et al. 2014; Gelbart et al. 2014], for scenarios where the feasibility cannot be determined in advance. In this problem setting, they incorporate inequality constraints into a black-box optimization:

$$\max_x f(x) \text{ s.t. } c(x) \leq \epsilon,$$

where $f(x)$ and $c(x)$ are expensive-to-evaluate black-box functions. Instead, we formulated the inequality constraints by incorporating a feasibility indicator function into the the acquisition function instead.

3 OVERVIEW

Figure 3 shows the workflow of our framework. The system first shows an image $I_0^t$ (Figure 3(a)) generated with a random vector $z_0^t$ using frozen pretrained generator network $G$, i.e. $I_0^t = G(z_0^t)$, and $I_j^t$ refers to the $i$-th candidate in the $j$-th iteration. However, the user may not be satisfied for its defects or personal preference on $I_0^t$, such as hair blending in with background and unshaven beard respectively. Assume that the user prefers a face similar to the current image $I_0^t$, but without the defect and the beard. The user first adjusts the multi-way slider to obtain an image $I_0^t = G(z_0^t)$ (Figure 3(b)) without defects by blending generated candidate images $\{I_c^t = G(z_c^t)\}_{c=1}^C$, where $c$ is the number of candidates ($c = 4$ in our current implementation). Furthermore, our framework provides several image-editing tools as in iGAN allowing the user to engage in additional guidance (Section 4).

The editing tools allow the user to directly edit $I_0^t$ by painting over it, or sourcing external images guidance image $I_I^0$ (Figure 3(c)).

The user presses the “next” button and the system provides the next set of candidate images. We use $z_0^t = z_1^t$ and the new latent vectors $\{z_c^t\}_{c=1}^C$ to generate the candidate images $\{I_c^t\}_{c=1}^C$ for the next user input iteration. This iterative process continues until the user obtains the image that matches his/her preferences.

4 USER INTERFACE

Our user interface consists of a main viewing window and multiple candidate images as shown in Figure 4. For each candidate image, we provide an associated user slider. Adjusting the slider enables the user to explore and compose a new image $I$ that is shown in the principal window. Moreover, we provide several local image editing tools to enable additional user guidance. At each iteration, the user specifies preferences by manipulating the sliders and optionally performing local edits. The user can drag the several sliders and “blend” images, and paint on the blended image locally. Finally, the user hits the “next” button to request the system to update the internal model and present new candidates.

4.1 Multi-way slider

At iteration $t$, the user manipulates the sliders to compose a new image $I_c^t$ that is the closest to his/her preferences within the slider space. The slider values correspond to blending weights of candidate images; the user explores the convex subspace of the latent vector space bounded by the candidate images. If the user is not satisfied with the blended image $I_c^t$, he/she can use the image editing tools to create a guidance image $I_I^t$ (Section 4.2). Otherwise, he/she can directly use the blended image as the guidance image, i.e. $I_I^t = I_c^t$.

4.2 Image editing tools

The user creates a guidance image by using the image editing tools on the blended image $I_c^t$ and the system will provide new candidates that matches the guidance image. 

\begin{itemize}
  \item **Color painting** assigns colors to any region on the blended image.
  \item **Eraser** removes content locally. The user uses eraser when they are not satisfied with certain region and would like to simply see change in that region.
\end{itemize}

\textsuperscript{3}https://ganbreeder.app/
Keep allows the user to specify a region to remain unchanged. Copy & Paste The user finds external images from any source, e.g. from a web search or image databases, that match his/her preferences for a specific region. The user then copies and pastes that region of the external image to $I_1$.

Note that our method allows the users to use any external image editor to provide image-based guidelines.

5 ALGORITHM

Our method focuses on adjusting the input latent vector $z$ to control image generation process $G(z)$ of generative image modeling. We consider this latent vector adjusting process to be a numerical optimization, and model the user preference as our objective function. Since evaluating such user preference function is expensive-and-hard, we design our method based on Bayesian optimization that aims for optimizing it with minimal number of observations. We include pseudo-code to assist user understanding in the supplemental material.

5.1 Sequential Subspace Search

Given an initial latent vector $z_0$, which generates an image $I_0 = G(z_0)$. We seek to maximise $g(I)$, a user preference function describing how much a user prefers an image $I$.

Koyama et al. [2017] extends BO and proposes to use slider interface which allows the users to consistently express their own preferences to a large number of options. In this work, we propose multiple sliders user interface to accommodate the large search space of generative image methonds.

User interaction and its mathematical formulation. Our user interface consists of $c$ sliders, each represents a latent vector $z_1^{1...c}$ and a corresponding image $I_1^{1...c}$. The user explores the entire subspace by steering through it with the provided sliders. Given slider values $s^i$, blending weights $a^i$ for each latent vector $z^i$ are given as $a^i = s^i / \sum_{j=1}^{c} s^j$. Note that we use normalized slider values, i.e. slider values $(1,1,1,1)$ is equivalent to $(0.5,0.5,0.5,0.5)$. The blended image, $I_b^i = G(z_b^i)$, generated by the blended latent vector $z_b^i = \sum_{j=1}^{c} a^j z^j$, is updated and displayed on the left side of the user interface as the user manipulates the sliders 4. After the user finishes manipulating the sliders, we assume that they arrive to the global minimum of the user preference function within the subspace, i.e. they choose the best image within the subspace. Below we refer the underlying perceptual response $g(I)$ of user observing an image $I = G(z)$ generated by a fixed generator $G$ as “user preference at latent space”. To avoid complicated notation, we use $g(z)$ as a short hand for $g(G(z))$ since we assume a fixed $G$. And we refer to $g$ as the goodness values of $z$ and $g(z)$ as the function itself. The latent space refers to a search space which is defined during the training of $G$.

Modelling user preference. We formulate the estimation of the goodness values $g(z)$ from slider interaction as a maximum a posterior MAP estimation. Following sequential line search [Koyama et al. 2017], slider manipulation is modeled with BTL model [Tsukida and Gupta 2011] and user preference at latent space with a Gaussian process prior. This allows targeting general generative model inference with no domain specific knowledge assumption. Concretely, the objective function that we want to maximize using MAP estimation is the following:

$$
(g^\text{MAP}, \theta^\text{MAP}) = \arg\min_{g, \theta} p(g, \theta|\mathcal{P}, z^b) = \arg\min_{g, \theta} p(\mathcal{P}, z^b|g, \theta) p(g|\theta) p(\theta),
$$

where $\theta$ is the parameters of the Gaussian Process, $\mathcal{P}$ is the latent vectors sampled with user interaction, $g^\text{MAP}$ and $\theta^\text{MAP}$ are the maximum likelihood estimate of $g$ and $\theta$ respectively. This optimization serves two purposes, to extract numerical value of user preference from slider manipulation (slider manipulation modelling) and to estimate user preference for any unobserved latent variable $z$ (user preference modelling).

5.1.1 Slider manipulation modelling. Given a set of $c$ multidimensional latent variables $\mathcal{P}_t = \{z_t^i\}_{i=1}^c$, and the variable corresponding to $z_t^b$ that is chosen by the user, we describe this situation as $z_t^b > \mathcal{P}_t$.

and it’s likelihood using BTL model as

$$
p(z_t^b > \mathcal{P}_t | \{g(z_t^i)\}_{i=1}^c) = \frac{\exp(g(z_t^b)/s)}{\sum_{i=1}^{c} \exp(g(z_t^i)/s)},
$$

where $s$ is a hyperparameter to adjust the sensitivity of the model. We set $s = 1$ through all our experiments.

5.1.2 User preference modelling. We also model the underlying user preference toward the latent space as a Gaussian Process $GP$. We assume $GP$ to follow a multivariate Gaussian distribution, parameterized with $\theta$, describing the kernel used in the function. In all of our experiments, we follow Koyama et al. [2017] and use RBF kernel.

As $D$ and $\theta$ are conditionally independent given $g$, at iteration $t$, we have

$$
p(\mathcal{P}, z^b, g, \theta) = p(\mathcal{P}, z^b|g) = \prod_{j=0}^{t} p(z_j^b > \{z_j^i\}_{i=1}^c|g).
$$
We define the following prior.

\[ p(g|\theta) = N(g_0; 0, K), \quad (7) \]
\[ p(\theta_i) = \mathcal{L}N(\ln 0.5, 0.1), \quad (8) \]
\[ p(\theta) = \prod_i p(\theta_i), \quad (9) \]

where \( p(g|\theta) \) is defined to be the \( \mathcal{GP} \) prior, and \( K \) and \( \theta \) are the covariance matrix and the parameters of the kernel (in our experiment we use RBF kernel) and the latent variables \( \{P, z^b\} \). We also construct \( g' \), a Gaussian Process regressor that approximates \( g \) using \( K \) and \( \theta \).

For further explanation and implementation detail of Gaussian Process with sequential line search, we advice the reader to read the sequential line search paper [Koyama et al. 2017]. Figure 7 illustrates an example optimization sequence of our subspace search with multi-way sliders.

5.2 Preference learning by Bayesian optimization

For iteration \( t > 1 \), we use the chosen latent variable in the last iteration to be the starting latent variable, such that \( z^0 = z^\text{chosen}_{t-1} \). Therefore at iteration \( t > 1 \), we have \( m = t \cdot c + 1 \) samples. We collect an observation \( O_m \) at iteration \( t_m \):

\[ O_m = \{P_m; g'(P_m)\}. \quad (10) \]

The next observations \( \{z^t_{m+1}\}_{t=z} \) should be “the ones most worth observing” based on the all previous observed data \( O = \{O_t\}_{t=1}^m \). We define an acquisition function \( a_G(z) \) to measure the “worthiness” of the next sampling candidate \( z_{m+1} \). For each iteration, the system maximizes the acquisition function to determine the next sampling point:

\[ z_{m+1} = \arg\max_{z \in \mathcal{Z}} a_G(z). \quad (11) \]

In order to choose the next sampling point which is most worthy to sample from, the expected improvement (EI) criterion is often used. Let \( g'^+ \) be the maximum value among the observation \( O \), the acquisition function is defined as

\[ a^\text{EI}_G(z) = E[\max\{g'(z) - g'^+, 0\}], \quad (12) \]

where \( E[X] \) means the expectation value of \( X \).

Selection of multiple candidates. We combine expected improvement with constant liar strategy [Ginsbourger et al. 2010] to acquire multiple points at each iteration. We first obtain the first candidate through maximizing the current acquisition function. Then, we assign maximum score to this sample point and update the acquisition (e.g. we assume the new candidate is as good as the best candidate we have seen so far). And we pick the second candidate that maximizes the update acquisition function. We repeat this process to obtain a set of \( c - 1 \) candidates.

5.3 Content-aware sampling strategy

To incorporate the user guidances by image editing (Section 4.2), we extend the original acquisition function (Eq. 12) into the following form:

\[ a^\text{EI}_G(z) = E[\max\{g'(z) - g'^+, 0\}] - \sigma_1 C(G(z)) - \sigma_2 \mathcal{R}(z), \quad (13) \]

where \( C(I) \) is the content-aware bias term, \( \mathcal{R}(z) \) is a regularization term, and \( \sigma_1 \) and \( \sigma_2 \) are the balance weights.

Regional guidance term. We handle all the guidance from the image editing operation with the following term:

\[ C(I) = \sum_{x, y} \left( I_{x,y} - I_{x,y}^* \right)^2 + M_{x,y}. \quad (14) \]

where \( M \) is a mask initialized as all 0.2. We use \( i, j \) as the 2D pixel coordinate of the image. For color painting and copy & paste, we set \( M = 1 \) inside ROI. It encourages the optimized image to match the user’s edit inside ROI. When preserving the region inside ROI by setting \( M = 1 \) inside ROI. Eraser sets \( M = 0 \) inside ROI to encourage the optimization to alter region outside ROI.

Regularization term. We introduce a regularization term to incorporate our priority knowledge of \( z \), i.e. the latent variables are usually sampled from a known distribution \( p \), e.g. Normal distribution during training. Therefore we design a regularization term \( \mathcal{R} \) to prevent the estimation from deviating too far from \( p \): \( \mathcal{R}(z) = \log(p(z)) \).

6 RESULTS

We applied our framework on face generation application using PGGAN [Karras et al. 2018] to control the appearance of human faces. We used the pretrained generator model provided by the authors4. Our method can be used to control different parts of the face; to change hair colors and styles (change hair color and style in Figure 1(a), and remove hair in Figure 7(c)); to add or remove beards (Figure 7(a)); to remove accessories such as earrings and glasses (Figure 7(b)); and to add or remove smiles (Figure 7(d)). In addition, we applied our framework to image-to-image translation and text-to-image synthesis. Representative results are shown in Figure 1. We used either pretrained generator networks or networks trained as described in the original paper. During user interactions, the generator network parameters were frozen.

Image Translation. We applied our method to MUNIT [Huang et al. 2018], which performs image-to-image translation (e.g. hangbag sketches into photographs of hangbags of different colors). We

\[ \text{https://github.com/tkarras/progressive-growing_of_gans} \]
used the pretrained models provided by the authors\(^5\). MUNIT separates content and style into separate latent vectors, and we applied our method to the 8-d style vector (the blue latent vector part of Figure 6(b)). Our method could be used to design new heels (check teaser figure (b)), boots (Figure 10(b)), and hangbags (Figure 10(c)) given sketch inputs by assigning different colors for different parts.

**Image Synthesis from Text.** Finally we applied our method to AttnGAN [Tao Xu 2018], which allows text-to-image synthesis. We choose AttnGAN as a state-of-the-art text-to-image synthesis method to show that our framework could handle various network architectures. AttnGAN encodes input text into a sentence vector, accompanied by a random latent vector drawn from a normal distribution \(N(0, 1)\), and synthesizes an image. We applied our method to the 100-d random noise vector (the blue latent vector part in Figure 6(c)). We showed that our method can be used to improve the synthesized bird images (check teaser figure (c) and Figure 11(b)), and repair artifacts (Figure 11(a)).

6.1 User study

We conducted a user study to clarify how our method compares with iGAN [Zhu et al. 2016] with respect to steering the image generation process. We measure the perceptual quality of image and the time required to generate it.

**Procedure.** We used face image editing with a reference image as the target task. We randomly sampled four images from a subset of the CelebA dataset [Liu et al. 2015] and excluded the images used to train PGGAN [Karras et al. 2018]. Each participant was shown a reference image, and asked to generate an image as similar as possible to the reference using either our method or iGAN.

Participants were judged to have finished the task when either (1) they were satisfied with the result, (2) they found it hard to improve the results, or (3) the 10 minute time limit was reached. We recruited 8 people for the user study, 3 female and 5 male. We divided the participants randomly into two groups, each containing 4 people. Our user study proceeded as follows: (tutorial A \(\rightarrow\) A1 \(\rightarrow\) A2 \(\rightarrow\) tutorial B \(\rightarrow\) B1 \(\rightarrow\) B2), where A1 denotes as using method A on image 1. The order of the methods and images are fully balanced. Before the participants started to edit the test image, we provided a 10 minute tutorial session on both methods. From both methods, we removed import functions (“copy and paste” in our method and “import” in iGAN), to prevent the participants from directly using reference image as input. This provides a better simulation of the actual use case, in which users have no access to any picture of the desired goal, but only a vague idea in their minds.

**Crowdsourced Evaluation.** To evaluate the perceptual quality of images generated in the user study, we carried out a crowd-sourced comparison study to evaluate the visual quality of the resulting images. For each query, we showed crowdworkers three images: a reference image and the edited result using both iGAN and our method. We then asked them to select the result that better matched the reference image. We composed a survey of 16 queries. We used Amazon Mechanical Turk interface to conduct the survey, in which each participant was shown all 16 queries, with each query shown twice and the order of the candidates switched. For each crowdworker, we discarded inconsistent answers, where a duplicated query was answered differently, and discarded all answers from participants who answered over 25% of queries inconsistently. In total, we report the results obtained from 50 crowdworkers that passed our consistency checks.

**Study Result.** We compared the user interaction time for both methods, and the results are shown in Figure 8(a). The user interaction time was significantly shorter using our method than using iGAN. Figure 8(b) shows the results of the crowd-sourced comparison user study. As shown, the results obtained using our method were voted higher than those obtained using iGAN across all four reference images. In total, 77% of the crowdworkers preferred our results to the results obtained using iGAN. The edited images obtained from the user study are shown in Figure 9.

We attribute these differences to the fact that novice users often fail to render their mental images accurately using drawing tools. In addition, preprocessing is often applied to image during the training (e.g. cropping, scaling) and user might not be able to do drawing following those rules. This makes direct optimization, such as iGAN, perform poorly in practical tasks. In contrast, the same users find it easier to compare and evaluate images.

**Additional study.** Besides the above user study, we performed additional studies to analyse the theoretical and practical advantage of multiple sliders instead of one. We showed that the user explores the latent space more efficiently with 4-slider than 1-slider in the Slider-time complexity study and the theoretical advantage of multiple sliders in the ablation study. Detail of both studies are located in the supplemental material. We show that multiple sliders allow the users to explore the latent space more efficiently.

\(^5\)https://github.com/NVlabs/MUNIT
Fig. 7. We use our method to control the synthesis results using PGGAN [Karras et al. 2018]. The user is able (a) to add beard, (b) remove glasses, (c) make a man bald and old, and (d) make a woman smile.

(a) Average interaction time (sec) for all participants, referring to the time the user interact with the system, excluding wait time for computation.

(b) The percentages of crowdworker preference to our method (green) and iGAN (red) on different reference images (shown in Figure 9).

Fig. 8. Results of the user study.

7 CONCLUSION

We introduced a human-in-the-loop optimization method that allows the user to explore the latent vector space in generative image modeling with sliders and image editing. Our method only requires a pretrained generative model, so it provides a plug-and-play option for various generative image modeling methods. Sliders allow sampling large latent hyper volume efficiently. In addition, we model the user preference using indirect observation (e.g., slider interaction) rather than direct input (i.e., user strokes). Our method is easier and encourages the user to explore the latent space instead of drawing, which they tend to fail. The focus on exploration also provides artistic freedom to the user by not requiring a set goal in mind. It supports a wide range of applications outside those in Section 6. We encourage readers to apply our method to generate artwork with their own architecture and datasets.

REFERENCES

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. 2018. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. arXiv preprint arXiv:1811.10597 (2018).

James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. 2015. Hyperopt: a Python library for model selection and hyperparameter optimization. Computational Science & Discovery 8, 1 (2015), 014008.
Fig. 9. Here we show several edited results from user study. Grey lines in iGAN edit represent contour sketching and the rest strokes represent colour painting (Please find more in supplemental material).

Fig. 10. The control sequences of our framework running on MUNIT [Huang et al. 2018]. The users are able to control the slider and use the color paint tool to design (a) a new boot, and (b) a new handbag.
### Input sentence (simplified)

| Input sentence (simplified) | Initial result | Candidates | Slider result | Edit result | Intermediate operations (if any) | Final result |
|----------------------------|----------------|-------------|--------------|-------------|----------------------------------|-------------|
| “Red bird with yellow belly” | I₀            | (I₀, I₁, I₂, I₃) | I₀          |             |                                  |             |
| “Black wing with white belly” |               |             |              |             |                                  |             |

### Diagram

![Diagram showing the process of editing a bird image](image-url)

**Fig. 11.** We use our method to control the synthesis results from text-to-image application (using AttnGAN [Tao Xu 2018]). The user is able to (a) add a larger beak, and (b) improve the synthesized bird in only few iterations.

---

Eric Brochu, Tyson Brochu, and Nando de Freitas. 2010. A Bayesian interactive optimization approach to procedural animation design. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 103–112.

Eric Brochu, Nando de Freitas, and Abhijit Ghosh. 2007. Active Preference Learning with Discrete Choice Data. In *Advances in Neural Information Processing Systems*. 3201393.

Jacob R. Gardner, Matt J. Kusner, Zhixiang Xu, Kilian Q. Weinberger, and John P. Cunningham. 2014. Bayesian Optimization with Inequality Constraints. In *Proceedings of the 31st International Conference on Machine Learning*. 937–945.

Michael A. Gelbart, Jasper Snoek, and Ryan P. Adams. 2014. Bayesian Optimization with Unknown Constraints. In *Proceedings of the 30th Uncertainty in Artificial Intelligence*. 427–436.

David Ginsbourger, Rodolphe Le Riche, and Laurent Carraro. 2010. Kriging is well-suited to parallelize optimization. In *Computational Intelligence in Expensive Optimization Problems*. Springer, 131–162.

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

Éric Guérin, Julie Digne, Éric Galin, Adrien Peytavie, Christian Wolf, Bedrich Benes, and Benoît Martinez. 2017. Interactive Example-based Terrain Authoring with Conditional Generative Adversarial Networks. *ACM Trans. Graph.* 36, 6, Article 228 (Nov. 2017), 13 pages. https://doi.org/10.1145/3130800.3130804

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Consistent Image Completion. *ACM Transactions on Graphics* (Proc. of SIGGRAPH 2017) 36, 4, Article 107 (2017), 107:1–107:14 pages.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017 IEEE Conference on.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive growing of GANs for improved quality, stability, and variation. In *Proc. International Conference on Learning Representations (ICLR)*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

Yuki Koyama, Isser Sato, Daisuke Sakamoto, and Takeshi Igashita. 2017. Sequential Line Search for Efficient Visual Design Optimization by Crowds. *ACM Trans. Graph.* 36, 4, Article 48 (July 2017), 11 pages. https://doi.org/10.1145/3072959.3073598

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ECCV)*.

Pushparaja Murugan. 2017. Hyperparameter Optimization in Deep Convolutional Neural Network / Bayesian Approach with Gaussian Process Prior. *CoRR abs/1712.07233* (2017). arXiv:1712.07233. http://arxiv.org/abs/1712.07233

Tiziano Portenier, Qiyan Hu, Attila Szabó, Sivashan Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. 2018. FaceShop: Deep Sketch-based Face Image Editing. *ACM Trans. Graph.* 37, 4, Article 99 (July 2018), 13 pages. https://doi.org/10.1145/3197517.3201593

Warren Scott, Peter Frazier, and Warren Powell. 2011. The correlated knowledge gradient for simulation optimization of continuous parameters using gaussian process regression. *SIAM Journal on Optimization* 21, 3 (2011), 996–1026.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando De Freitas. 2015. Taking the human out of the loop: A review of bayesian optimization. *Proc. IEEE* 104, 1 (2016), 148–175.

Ryohei Suzuki, Masanori Koyama, Takeru Miyato, and Taizan Yonetsuji. 2018. Collaging on Internal Representations: An Intuitive Approach for Semantic Transfiguration. *CoRR abs/1811.10153* (2018). arXiv:1811.10153 http://arxiv.org/abs/1811.10153

Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, Xiaodong He, Tao Xu, Pengchuan Zhang. 2018. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. (2018).

Kristi Tsukida and Maya R. Gupta. 2011. *How to Analyze Paired Comparison Data*. Technical Report. University of Washington, Department of Electrical Engineering.

Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. 2017. Real-Time User-Guided Image Colorization with Learned Deep Priors. *ACM Transactions on Graphics (TOG)* 9, 4 (2017).

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. 2016. Generative Visual Manipulation on the Natural Image Manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ECCV), 2017 IEEE International Conference on*.