Semantic Manga Character Sketch Generation

Kittinun Aukkapinyo and Seiji Hotta
Department of Computer and Information Science, Tokyo University of Agriculture and Technology, Tokyo, Japan
Email: kittinun.auk@gmail.com, s-hotta@cc.tuat.ac.jp

***Abstract*—A comic is one of the approaches in information and knowledge transfer. Manga is a unique and attractive comic style that originated in Japan. One of the important components in each manga is a character. It requires experience and knowledge in illustrating a manga character. Several research studies employed Generative Adversarial Networks (GANs) to illustrate a character. Although unconditional GANs could produce a high-quality image, it still lacks controllability over synthesized character. This research proposed an approach to employ conditional GANs with a semantic mask to control posture, anatomy, and basic dressing style during the synthesizing process. It also introduces an approach to systematically specify desired character's style and parse a character into a semantic mask. As a result of a human-based evaluation, this research can competently employ a semantic mask as a condition to synthesize a decent sketch version of a manga character.**

***Index Terms*—image generation, comic computing, deep learning**

I. Introduction

A. Background

Manga is a comic that originated in Japan, which can contribute to information transfer. It can improve recognition of particular information, such as academic knowledge and marketing content, through its attractiveness. It can also serve as a tool to transfer complicated knowledge to others. A character is one of the most important components in each manga creation. Although character drawing is an easy task for professional human artists, it is considerably difficult for inexperienced humans. In addition, it is also a challenging task for a machine to creatively synthesize a manga character.

A generative adversarial network [1] is one of the unsupervised machine learning techniques that can be used to train a computer to synthesize new contents from training data such as images, videos, and sounds. It consists of two deep neural networks architecture - a generator and a discriminator. A manga character can be represented as an image composed of strokes and color information, such features allow a generator to learn to illustrate a character. However, a vanilla GAN has an issue of controllability of a generated image. So, a conditional GAN (cGAN) is introduced to provide some conditions along with input during training. It is to control output with a certain condition such as a less information version of the final image.

There are many applications of generative adversarial networks in rendering and translating real-world images, such as black-and-white image colorization. It is an uncomplicated task when compared to manga characters due to their surrealism. In addition, there is an extremely high variance in definitions of real-world things in each manga. For example, a character's anatomy and fantasy clothing style. It is more challenging for a generative model to learn the extreme differences in features as to what defines a particular object.

Several research studies employed a Generative Adversarial Network (GAN) to generate a character ranging from coarse-to-fine solutions where users provide some conditions, such as a sketch or real-world image [2]-[8]. Chen *et al.* [2], introduced a framework, CartoonGAN, to train a computer to transform an input image of real-world scenery into a comic style. Another one is Auto-painter [3]. This paper proposes a methodology to train a computer to paint and generate comic drawings from its sketch version using a GAN. In addition, Li *et al.* [8] generated a character from a random noise vector. Although it can generate a good quality image, it still has a problem of controllability in a generation.

The existing methods focused on generating colorful manga characters from a high-level condition, for example, black-and-white images. Also, there are no specific selection criteria for training illustrations such as posture, dressing style, and gender. It could affect the generation of a character with a body due to extreme high variance during training. Our proposed method uses a semantic mask as a condition to control a synthesized output. Each pixel represents a component of a character. In addition, there are procedures to reduce the variance of training examples. An illustration is also preprocessed into a sketch version to control the variation of colorization techniques. This work will focus on sketching a woman in her plain cloth with a simple posture. A colorization might be done manually by a user or another colorization model in the future.

This research aims to use a semantic mask of a character as a condition to develop a generative model for specific character style creation. A Self Correction for Human Parsing framework [9] is employed and retrained with our manga character datasets to extract each semantic component of a character such as the upper torso, legs, and hair. Then, pix2pix with SPatially-Adaptive (DE)normalization (SPADE) [10] is used to train a generator to generate images based on their semantic

Manuscript received May 6, 2022; revised September 20, 2022.

doi: 10.18178/joig.10.4.145-150

masks. As a result, a generator can aesthetically generate a character from both seen and unseen semantic masks.

B. *Contribution*

This paper proposed an approach to generate a manga character with its body. There are challenges in developing a generative model for synthesizing manga characters. They have extreme variations in drawing and dressing styles from different artists, such as anatomy, dressing styles. Moreover, each object's appearance is also varied when compared to a real-world object. So, it mainly focuses on reducing the variance of training examples. Our proposed method defines a set of selection criteria to deal with a high variation of illustrations. It is also a procedure to specify the desired style of a generated character. Our criteria desire a woman with a plain dressing style, which looks similar to real-world practice. There will be only one character in an image with a simple or transparent background. A generator can learn useful features on a single character regardless of error in semantic segmentation. Also, the control of variance allows us to reduce the number of training samples.

In addition, an illustration is preprocessed into its sketch version to reduce variation on colorization styles. It discards color and shadowing information. Also, a network will have much fewer features to learn to reconstruct a new sketch character. The variation control reduces training samples and time to develop a generative model for a manga character with its body. A sketch version allows a user to apply their desired artistic technique, such as colorization.

Apart from variation control, this paper considers this task as a mask-to-image translation rather than a general image-to-image translation. A pix2pix with a SPADE generator can serve as a mask-to-image translation framework that efficiently utilizes semantic information in a source image. Unlike a general image-to-image translation framework such as pix2pix, a SPADE layer will preserve semantic information through a generator training process. It will be robust against translation at the pixel level of a source mask.

II. METHODOLOGY

A. *Overview of the Proposed Framework*

The proposed framework consists of 4 iterative processes to develop a generative model to synthesize a manga character as shown in Fig. 1. The first process is to create datasets of illustrations. Next, a selection model is trained to reduce the variance of training examples and constructed filtered training datasets. A character parsing framework is then trained to extract semantic masks from an illustration. Then, an image translation framework is trained to synthesize illustrations based on semantic masks. Since an evaluation of an aesthetic of synthesized images is subjective, a model is evaluated with real humans. These procedures can be repeated to improve the performance generative model.

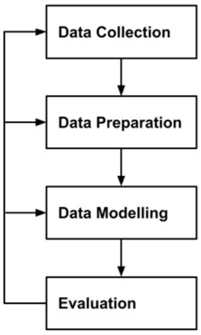


Figure 1. An overview of the proposed framework.

B. *Data Collection*

Danbooru is an online database of Japanese-style illustrations from all over the world. Its users regularly contribute to this database by uploading images along with their metadata [11]. For example, a set of tags, source, artist. Tags are considerably valuable metadata about an illustration. They describe an illustration in both physical appearance and story, such as a character's name, dressing, background. The metadata of all illustrations is stored in millions of records of the Danbooru database. It provides an application programming interface (API) for retrieving images with their corresponding metadata. Hence, it is more feasible and systematic to gather illustrations than crawling over the Internet.

C. *Data Preparation*

1) *Illustration selector*

The variance of training examples affects the performance of the machine learning model. Illustrations on the Danbooru database have a wide range of drawing styles. There is an extremely high variance in anatomy, posture, facial expression, and dressing style as shown in Fig. 2. As an initial study, human-like illustrations with a casual dressing style are used as a criterion to reduce the variance from the enormous image database. There are two processes to select our desired images.

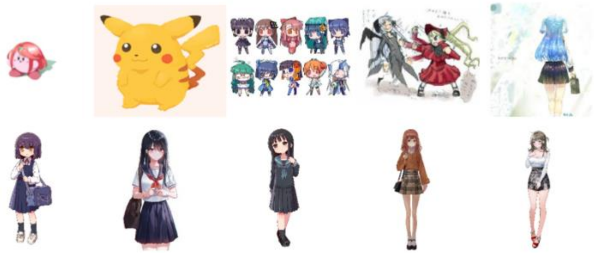


Figure 2. An example of illustration available on the Danbooru database.

The first process is to use a set of user-generated tags to search a database for matched images. Tags can describe the whole illustration such as its background, number of people, clothing style. The desired tag should be simple and real-world clothing styles, such as school uniforms or casual outfits. Several datasets are created using a different combination of tags. However, tags that indicate that images have simple or transparent backgrounds are always included when creating datasets. This is to make a model focus on the learning feature of a character. After that, all images in the dataset are downloaded into our storage.

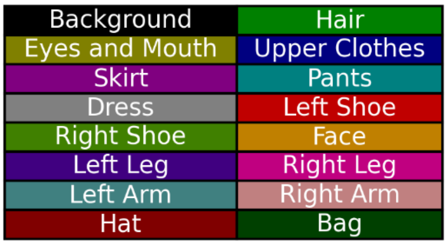
Tags are not good enough information to reduce the variance since their definitions are varied regarding individual artists' drawing styles. Also, they can't accurately describe the posture and anatomy of a character. Some illustrations contain full character bodies but some are only close-ups. Since the number of illustrations in each dataset is considerably high, it is inefficient to manually select the one that matches our criteria. Hence, a Convolutional Neural Network (CNN) is fine-tuned to serve as a selection model.

The training dataset is created to train a binary image classification. It is not required to have a large number of training examples since it's fine-tuning. A set of selection criteria is defined. First, an image must contain only a single woman character on the frontal side. She must have a real-world dressing style such as plain or casual cloth. Next, she must not hold any tools or equipment. Lastly, an image must have a simple or transparent background. An image that matches these criteria is annotated as positive and should be included in the dataset.

This research employs EfficientNetB0 [12], which is a state-of-the-art architecture to serve as a framework to train a selection model. Its source code and pre-trained weights on ImageNet are provided on the TensorFlow library [13]. The last layer is changed to a binary softmax classifier to classify if an image meets our criteria. Since it is fine-tuning, the rest of the layers are frozen. A dataset for the selection model is created with a manual selection of illustrations. After that, a selection model decides whether to include a particular illustration in a dataset. So, it serves as a data selector for use in this research.

2) *Semantic mask parsing*

Since a character has an appearance similar to humans, the existing human parsing framework can develop a character parsing model to extract its semantic information. Images from the selection model are sampled and used as training examples to train a character parsing model. Each character is annotated into 16 classes at pixel-level as shown in Fig 3. So, each pixel represents a part of a character such as cloth, facial component, skirt.



Background	Hair
Eyes and Mouth	Upper Clothes
Skirt	Pants
Dress	Left Shoe
Right Shoe	Face
Left Leg	Right Leg
Left Arm	Right Arm
Hat	Bag

Figure 3. A class label and its representative color.

This research uses the Self-Correction for Human Parsing (SCHP) human parsing framework to train a character parsing model. With the self-correction method from initial outputs of the Augmented-CE2P framework, it can precisely learn to extract the semantic mask of a character. A retrained character parsing model can serve as a semantic mask extractor.

3) *Sketch extraction*

This is a procedure to process an image into its sketch version. Since a color image contains high details of both strokes and color information, it might affect the feature learning of a generator. It is a process to remove redundant information from an image. The artistic details, such as shadows, will be removed from an illustration. An extended difference-of-Gaussians (XDoG) is employed to extract sketch from an image [14]. It has a parameter γ, which controls the level of the details of the sketch. A greater γ preserves more details and produces a detailed sketch, as shown in Fig. 4. A sketch version of an illustration is then obtained through this extraction process.

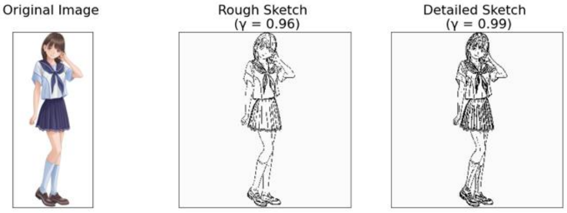


Figure 4. An example results from the sketch extraction at different levels of details.

4) *Mask-to-Image dataset creation*

Images that meet our selection criteria are combined into a dataset with each combination of tags. Their semantic masks are extracted using a character parsing model. After that, they are resized to have an equal size regarding their long edge. Padding is done to convert them into a square aspect ratio while preserving the original ratio for a character. A semantic mask becomes a source image and its corresponding image becomes a target image. In other words, a model will learn to reconstruct an original image based on its semantic mask. Then, they are combined into pairs of images consisting of source and target images as shown in Fig. 5. In each dataset, processed images are separated into training and validation datasets with a ratio of 90:10.

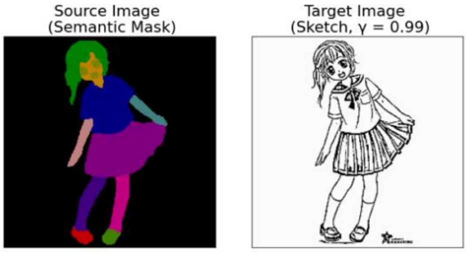


Figure 5. An example results from the sketch extraction at different levels of details.

D. *Data Modelling*

The main image translation model is based on the pix2pix framework. The framework receives a pair of an image with a semantic mask as an input to generate a new image. However, a vanilla pix2pix is not mainly for reconstructing an image from a semantic mask. This is because the semantic information in a mask is diminishing after batch normalization in a generator. Many studies attempted to solve this problem for mask-to-image generation, such as pix2pixHD [15].

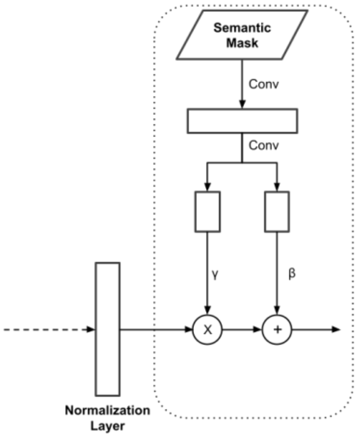


Figure 6. A denormalization process in a SPADE generator.

This research employs a SPADE generator to preserve semantic information after the normalization. It modulates normalized features with semantic segmentation maps as shown in Fig. 6. So, a generator in pix2pix can efficiently learn useful features based on semantic annotation. In addition, a SPADE generator allows multi-modal synthesis by using an additional noise vector to control a style of generated images.

E. Evaluation

The evaluation of synthesized images from generative models is subjective since it involves personal perception and preference. Human-based evaluation is more appropriate than numerical metrics such as Fréchet Inception Distance (FID) [16] or Inception Score (IS) [17]. Our proposed method will be evaluated against existing image-to-image translation frameworks such as pix2pixHD using the same dataset. They will be trained to generate an illustration from its semantic mask. Then, generated images on the testing dataset, which is unseen to the model during training, will be presented to human participants. They will vote for the most preferred generated illustration for each semantic mask. The voting score S for each framework will be computed to find the best framework to synthesize an illustration based on the semantic mask. The voting equation is shown as follows:

$$S_i = \sum_{i=1}^{N} v_{ij} \tag{1}$$

where S_i represents the voting score for the i^{th} method, N represents the total number of testing images, v_{ij} represents the vote from participant j for the i^{th} method (0 or 1).

Another statistical metric is the variance of the voting scores for each method. It is to measure the stability of the quality of generated images.

III. Experiments and Discussion

First, a query was developed to retrieve images' metadata from the Danbooru database with 1girl, full_body, solo, transparent_background, and white_background tags. Also, there were additional conditions to exclude inappropriate images. It was to retrieve images that contain a single woman character with a plain background. 69731 images were matched with our query. Then, an illustration selector was created by fine-tuning the EfficientNetB0 with ImageNet pre-trained weights. The computation resource was the single NVIDIA GeForce RTX 3070 on a standalone computer. A dataset for fine-tuning contains 3808 sampled images from retrieved images. The proportion of training and testing datasets was 80:20. The weights updating was enabled only at the final softmax layers during training. A trained illustration selector filtered complicated images out and there were 9668 images remaining for further procedures.

Next, Self-Correction for Human Parsing (SCHP) is trained to extract semantic masks from an image. An image-to-mask dataset is created to train SCHP to parse a character into a mask. It contains 866 images with their corresponding manual annotated label in pixel-level. They were split into training and testing datasets with a ratio of 90:10. There were 16 annotated classes to describe a character as shown in Fig. 3.

After that, images were processed into sketches with two different γ parameters, which were 0.96 and 0.99. More details could be preserved with a greater γ. Mask-to-image datasets were created by pairing sketches with their corresponding semantic mask. pix2pixHD and pix2pix with SPADE generator were employed to train a mask-to-image translation model. Therefore, there were four different configurations of our data modeling. NVIDIA provided implementation for both variants of pix2pix [10], [15]. An image was resized to 256x256 pixels and a batch size of 2. The learning rate is 0.0002 and the maximum epoch is 200. An example of generated images from each configuration is shown in Fig. 7. A generator can synthesize a sketch based on its semantic information.

A questionnaire was used to conduct a human-based evaluation. It asked participants to vote for the most preferred one from all generated characters in each scenario. A consists of 15 items, which were semantic masks and their generated characters. In each item, generated images were shuffled, so participants wouldn't know the configuration of each one. Semantic masks were sampled from a testing dataset to measure the generalization of our models. 38 participants participated in this evaluation. They were 20-29 years old and often read Japanese comics. Some of the participants were illustrators who regularly got commissions to illustrate Japanese-style characters.

TABLE I. The Result of an Evaluation

Configuration	# Votes	Variance
pix2pixHD (γ = 0.96)	42	0.068
pix2pixHD (γ = 0.99)	213	0.234
pix2pix with SPADE (γ = 0.96)	35	0.058
pix2pix with SPADE (γ = 0.99)	280	0.250

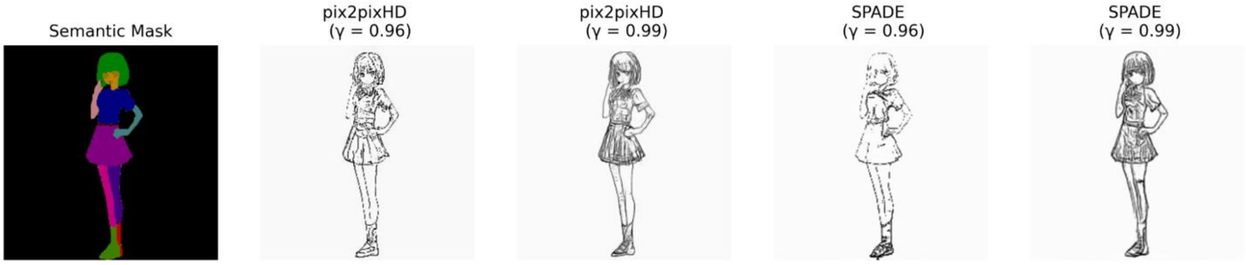


Figure 7. An example of generated images from different scenarios.

The result of an evaluation is shown in Table I. As a result, detailed sketches were significantly preferable to rough sketches according to the number of votes. Both frameworks can steadily synthesize images since they have a low variance in voting. Although there is a difference in the drawing's style of a generated image, they can synthesize a sketch based on its semantic mask.

The best configuration is to extract detailed sketches, and train a mask-to-image translation model with pix2pix with SPADE. It is the most preferred setting for synthesizing a sketch from voters. The pix2pixHD might be less popular but it can also produce sensible sketches according to their semantic information. Therefore, semantic masks could be used as a condition to control an illustration of a character.

There is some limitation due to a variance control from our selection criteria. It allows us to control specific styles for generated sketches, but it also reduces the number of training samples. There were only 8701 training images remaining for training a mask-to-image translation model. However, since Danbooru is an online database, the number of training samples is increased in the future. Also, an error from the character parsing process produces an incorrect semantic mask, as shown in Fig. 8. It might happen due to a high variance in drawing style and few training samples for a character parsing model. It affects the quality of generated images during the data modelling process. Therefore, an improvement in the character parsing process might also improve a synthesized sketch.

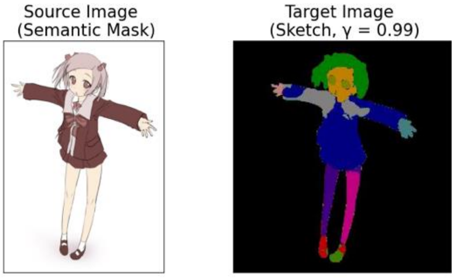


Figure 8. An example of semantic masks with segmentation errors.

In addition, a sketch extraction could eradicate information on some plain illustrations, as shown in Fig. 9. Therefore, it might affect the training of a generator to synthesize a sketch. Since its ground truth is incomplete, it can't learn to generate a proper semantic sketch. A current approach converts all images into a sketch with the same value of the γ parameter. Each image might have a different appropriate level of detail preservation.

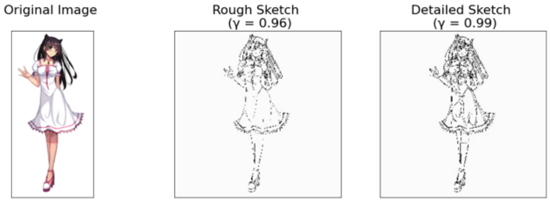


Figure 9. An example of plain illustration with sketch extraction.

IV. Conclusion

This research introduces a semantic mask as a condition to illustrate a new character. The process starts with retrieving desired illustrations from the Danbooru database based on their metadata. Then, an illustration selector is created to remove undesired images from their visual feature. The remaining images are processed into sketches with different levels of detail preservation. SCHP is trained to extract a semantic mask from an illustration. After that, a mask-to-image dataset is created for training pix2pix in varied configurations to generate a character based on its semantic mask. Next, a human-based evaluation is conducted and pix2pix with SPADE is the most preferred configuration.

In the future, an increase in the number of training samples could help in improving the quality of generated sketches. Also, a refinement of the extracted semantic mask might help to improve the generative model since it would accurately learn from more accurate annotation. In addition, more conditions might be embedded as a noise vector to control a generated image.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Kittinun Aukkapinyo conducted the research and experiment, and composed a manuscript. Seiji Hotta supervised and provided suggestions during the research; all authors had approved the final version.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672-2680, 2014.

[2] Y. Chen, Y. K. Lai, and Y. J. Liu, "Cartoongan: Generative adversarial networks for photo cartoonization," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9465-9474.

[3] Y. Liu, Z. Qin, T. Wan, and Z. Luo, "Autopainter: Cartoon image generation from sketch by using conditional wasserstein generative adversarial networks," *Neurocomputing*, vol. 311, pp. 78-87, 2018.

[4] H. Chen, N. N. Zheng, L. Liang, Y. Li, Y. Q. Xu, and H. Y. Shum, "Pictoon: A personalized image-based cartoon system," in *Proc. the Tenth ACM International Conference on Multimedia*, 2002, p. 171-178.

[5] Z. Wang, "Generating anime sketches with c-gan," *Journal of Physics: Conference Series*, vol. 1827, no. 1, p. 012157, 2021.

[6] Z. Cui, Y. Ito, K. Nakano, and A. Kasagi, "Anime-style image generation using gan," *Bulletin of Networking, Computing, Systems, and Software*, vol. 11, no. 1, pp. 18-24, 2022.

[7] B. Li, Y. Zhu, Y. Wang, C. W. Lin, B. Ghanem, and L. Shen, "Anigan: Style-Guided generative adversarial networks for unsupervised anime face generation," *IEEE Transactions on Multimedia*, 2021.

[8] Y. Jin, J. Zhang, M. Li, Y. Tian, H. Zhu, and Z. Fang, "Towards the automatic anime characters creation with generative adversarial networks," arXiv preprint arXiv:1708.05509, 2017.

[9] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-Correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1, 2020.

[10] T. Park, M. Y. Liu, T. C. Wang, and J. Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[11] G. Branwen. (2022). Danbooru2021: A large-scale crowdsourced and tagged anime illustration dataset. [Online]. Available: https://www.gwern.net/Danbooru2021

[12] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. the 36th International Conference on Machine Learning*, Jun. 2019, pp. 6105-6114.

[13] M. Abadi, *et al.* (2015). TensorFlow: Large-Scale machine learning on heterogeneous systems. [Online]. Available: https://www.tensorflow.org/

[14] H. Winnemoller, J. E. Kyprianidis, and S. C. Olsen, "Xdog: An extended difference-of-Gaussians compendium including advanced image stylization," *Computers & Graphics*, vol. 36, no. 6, pp. 740-753, 2012.

[15] T. C. Wang, M. Y. Liu, J. Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-Resolution image synthesis and semantic manipulation with conditional gans," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[17] S. Barratt and R. Sharma, "A note on the inception score," arXiv preprint arXiv:1801.01973, 2018.

Kittinun Aukkapinyo received the B.Sc. degree in information and communication technology from Faculty of Information and Communication Technology, Mahidol University, Nakhon Pathom, Thailand. He was a Data Scientist with Wongnai Media Co., Ltd, Bangkok, Thailand. He is currently pursing M.Eng. degree in computer and information science at Tokyo University of Agriculture and Technology, Tokyo, Japan. His current research interests include pattern recognition, computer vision, multimedia information retrieval, and generative model.

Seiji Hotta received his B.S., M.S., and Ph.D. from Kyushu Institute of Design (merged with Kyushu University in Oct. 2003), Japan, in 1998, 1999, and 2002, respectively. He joined Nagasaki University, in 2002, as Research Associate. From 2007, he is Associate Professor of Tokyo University of Agriculture and Technology (TUAT), Japan. His research interest includes Design of algorithms for Pattern Recognition, Subspace Classifiers, Image and Video Processing.