IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol. 003, LIIR068
DOI: https://doi.org/10.52731/liir.v003.068

Façade Design Support System with Control of Image Generation using GAN

Shosuke Haji [*], Kazuki Yamaji [*], Tomohiro Takagi [*],
So Takahashi [†], Yukihiko Hayase [‡], Yasuko Ebihara [‡],
Hiroshi Ito [‡], Yoshiyuki Sakai [‡], Tomoyuki Furukawa [‡]

Abstract

Designing the façade, which is the front side of a building, is a crucial yet time consuming part of the architectural design process. Advancements in image generation have led to generative models capable of producing creative, high-quality images. However, it is difficult to apply existing image generative models to façade design as the generated images should provide the architect with inspiration while also reflecting the designer's knowledge and building requirements. The existing models are inadequate for controlling image generation. Thus, we propose a system that supports designers in coming up with ideas for façade design by enabling them to intervene in image generation. The proposed system first determines the base image by using text-to-image retrieval. Next, the system generates diverse images using adversarial generation networks and the user selects images in alternation. This allows for repeated divergence and convergence of ideas and provides the user with inspiration. Our experiments demonstrated that the proposed system is able to arrive at the target idea while providing a variety of ideas through controlled generation.

Keywords: StyleGAN, façade generation, multimodal retrieval, image editing.

1 Introduction

Designing the façade, i.e., the front side of the building, is an important process in building design. When designing a façade, the designer creates a rendering (image perspective) consisting of multiple designs and selects the desired one. However, creating multiple designs is often time consuming. The size and purpose of the building are often determined before the idea generation process begins, and designs must be generated within those constraints. In addition, in order to create a façade with CAD, it is necessary to create the outline of the building and the window frames, and then color them on the basis of the material. An office building has many windows, and the overall appearance changes depending on their arrangement, so it is necessary to create each window in detail. Thus, we propose a system that uses image generation technology to support designers in coming up with façade ideas.

Adversarial generative networks (GANs)[1] are one of the leading approaches to image generation. A GAN-based approach can be used for various tasks, such as high-quality image generation, image-to-image translation, super-resolution, and image editing[2][3].

[*] Department of Computer Science, Meiji University, Kanagawa, Japan
[†] University of Electro-Communications, Tokyo, Japan
[‡] Kume Sekkei Co., Ltd., Tokyo, Japan

StyleGAN[4] is an architecture that is used by many GAN-based image generation methods and has achieved high-quality image generation. StyleGAN explicitly generates a style vector that represents the features of an image, and by using AdaIN[5], a type of style transformation, it generates an image containing the features represented by the style vector. StyleGAN has been used with this architecture in generating images with the features of two images, editing images on the basis of certain attributes (e.g., age and gender), and other studies. GAN inversion, which encodes a real image into a latent representation of pretrained GAN, has also been actively studied to utilize these methods[6].

Existing methods that use GAN to generate buildings include models that use segment labels as input to generate façade images[7][8][9][10] and models that generate cityscapes[11][12][13]. Many of these models generate images appropriate for their inputs. However, the optimal inputs are limited, and suitable input images must be prepared by the user, which is time consuming. In addition, in order to support idea generation, it is important for the model to present various designs that the designer would not have come up with. Such image diversity has not been realized with the existing models.

HouseGAN[14] and HouseGAN++[15] are models for floor plan generation that can support idea generation. These models are input with a graph that represents room types as nodes and connections between rooms as edges, and they generate a variety of floor plans that satisfy the conditions of the graph. Designers are offered a variety of room layout possibilities from the model subject to requirements such as a predetermined number of rooms. The designer can create a floor plan by selecting an appropriate one from them or modify it using the designer's own knowledge.

However, to our knowledge, there is no practical image generation system for façade design. One reason may be that there are not enough datasets available. LSUN tower and LSUN church_outdoor are available as façade image datasets[16], but they contain buildings such as the Eiffel Tower and historic churches, which differ in domain from contemporary façades. Another is the difficulty of the task. Unlike other tasks such as face generation, in façade design, part of the target building is obscured by people or trees, or surrounding buildings may be included. In addition, the structure, material, and size of the target building can vary greatly.

To overcome these problems, our proposed system does not have a domain-specific model structure for façade generation. Influenced by the generalization of large-scale language models and other models, we use a general-purpose and highly expressive model.

To support idea generation, we break down the process into divergence and convergence. Divergence presents many façade images to generate new ideas, and in convergence, one idea is selected from the ideas generated by divergence. By repeating these steps, the designer can choose their desired design from the various options presented.

In the proposed system, the designer is the main idea generator, so the designer always has control over the idea generation process. First, the designer searches the database for images of existing façades and selects the image that will serve as the basis for the idea. In the next step, the system generates various images based on the base image using the latent space of StyleGAN2. By diffusing the latent representation of the base image in meaningful directions, semantically diverse images can be generated. The designer selects one of the generated images, and the system further alters the selected image. This allows for the divergence and convergence of ideas.

Our contributions are as follows:

(1) A system that supports idea generation for façade design. In contrast to conventional image generative models, our system is aimed at assisting designers and repeats divergence and convergence with the user controlling the generation process.
(2) A consistent system design with limited data. Existing image generative models utilize large amounts of data and pretrained models for specific domains, making it difficult to apply them to domains for which data are not available. We overcame this challenge by combining techniques suitable for the domain of contemporary façades.

2 Related Work

2.1 Image Generation

Research in image generation has grown rapidly, and GANs have emerged as one of the mainstream approaches. With the improvement of the architecture and training stability, GANs have achieved high-quality image generation and have been applied to various applications.[2][3] The diffusion model is an alternative approach that has received considerable attention. Stable Diffusion[17] can generate creative, high-quality images from the conditions set by the input text. However, with these text-to-image generation approaches, the generated images cannot be edited and the generation cannot be controlled. They also require fairly large-scale training and are difficult to reproduce.

2.2 Image Generation

StyleGAN[4] and its improved version, StyleGAN2[18], control the generated image by incorporating a style vector representing image features into the synthesis network, which is part of the generator, using style transformations. The synthesis network applies a style vector to each PGGAN[19] resolution. The style transformation in the low-resolution layer controls the global features of the generated image, and the high-resolution layer controls the local features. StyleGAN2-ADA[20] uses data augmentation to generate high-quality images with significantly less training data than StyleGAN2.

2.3 GAN Inversion

GAN inversion is the task of embedding an image into the latent space of a pretrained GAN to obtain a latent representation that reproduces the original image[6]. There are two approaches to GAN inversion: encoder-based and optimization-based. In the optimization-based approach, one random latent representation is prepared and iteratively updated so that the distance between the image generated by the latent representation and the target image becomes smaller. This approach does not require pretraining as in the encoder-based approach. The latent representation is adjusted so that the distance between the target image and the reproduced image is small, so the image reproducibility is high.

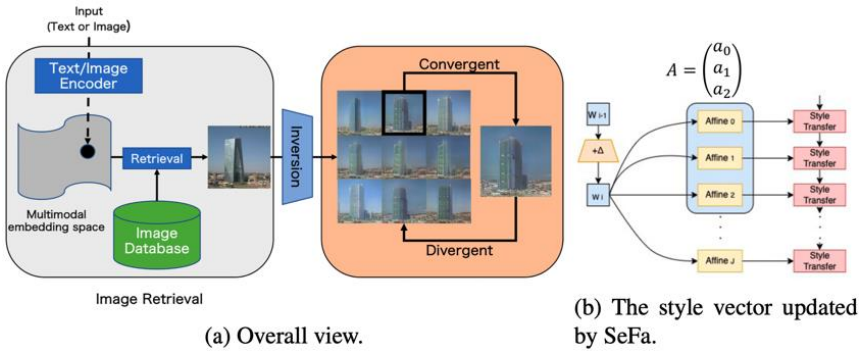
2.4 Image Generation Control in GANs

The most basic method of controlling the generated image is by inputting conditions[2][3]. Prior GAN studies used categories, images, text, etc., as input conditions. However, with these generative models, it is unclear to the user what image will be generated for what input. To our knowledge, when an unintended image is generated, there is no method of changing it to the

user's target image. A subsequent approach was developed which outputs the generated image and its segmentation image simultaneously and edits the generated image by editing its segmentation[21][22]. While these methods provide intuitive and flexible control, they may limit the diversity of the generated images. Another approach to controlling generation is by editing the latent representation of the generated images, i.e., by selecting a meaningful direction in the latent space and changing the latent representation in that direction. Methods include finding a semantic boundary with supervised learning[23][24], finding a boundary with unsupervised leaning, [25][26] and using a network to change latent expressions[27].

2.5 Multi-Modal Representations by Contrastive Learning

CLIP[28] uses contrastive learning to acquire image and text representations. CLIP learns to embed images and text into a common latent space. This allows images and text to be processed in a unified manner within the same latent space. In addition, the input of the text encoder is learned in natural language, which allows for flexible text input, and acquires a semantically rich latent space at the same time. By training with large-scale data, CLIP performs with high accuracy on tasks such as zero-shot text-to-image retrieval without additional training. The SIMAT Dataset[29] uses the latent space of CLIP to perform semantic operations on images and text.



(a) Overall view. (b) The style vector updated by SeFa.

Figure 1: System Overview.

3 Proposed Approach

In our proposed system, idea generation begins with text-to-image retrieval, and the generation and selection of images are repeated under the user's control. An overview of the proposed system is shown in Figure 1a. The system is divided into three steps.

(1) Text input is used to retrieve images from the database, and the user selects an image. This image is the basis for subsequent ideas.

(2) The selected image is inverted into a style vector.

(3) Diverse images are repeatedly generated by image transformation using the StyleGAN2 latent space and one image is selected from them. This encourages the designer to diverge and converge ideas.

By selecting the base image by image retrieval before inverting to a style vector, the image generated in step 3 will be close to the base image, preventing irrelevant images from being generated (e.g., an image of a house is generated when the target is an office building).

3.1 Multi-Modal Image Retrieval

When designing a façade, its purpose and size are typically determined in advance. Thus, in our method, images of base buildings are retrieved from a database and selected.

In the proposed system, the idea generation process begins with image retrieval rather than generation so that the base image for idea generation is more similar to the designer's target façade. When the image generated by GAN is used as the base, it is difficult to learn to generate façade images on the basis of the conditions, and images of sufficient quality cannot be generated. Façades have various characteristics, such as the outer shape (e.g., skyscrapers, houses), material (e.g., wood, glass, bricks), and style (e.g., modern, colorful), some of which are abstract and cannot be clearly defined. It is difficult to train StyleGAN2, which inputs these features as conditions. In addition, when using random façade image generation with noise as input, many unwanted images are presented, which hinders idea generation. However, in image retrieval, it is easy for the user to control the images that are displayed by adjusting the search query, and the images are photographs, so the quality is consistent.

The base building image is selected from a text-to-image search. We used CLIP to treat images and text as the same feature.

Here, the input text T_{input} is embedded in the latent space of CLIP.

$$x_{input} = E_{txt}(T_{input}) \tag{1}$$

where E_{txt} is the CLIP text encoder.

Meanwhile, the database is prepared with the collected façade images $I_{database} \in C_{database}$, and these images are embedded in the latent space using CLIP's image encoder E_{img}.

$$x_{database} = E_{img}(I_{database}) \tag{2}$$

In this step, the cosine similarity between $I_{database} \in C_{database}$ and each $x_{database}$ is measured, and the the top-k images with the highest similarity in the database is are displayed. The purpose of the proposed system is to support idea generation and to present many ideas. Selecting from the top-k images can display many images to the user and provide many ideas. Hereafter, the image selected here is denoted as $I^{selected}$.

By using CLIP, the actual input can be either text T_{input} or image I_{input}. If the retrieval is based on a specific building, the image of that building can be input into encoder E_{img}.

Furthermore, due to the semantically rich latent space of CLIP, the transformation of the displayed images by semantic operations is carried out as follows:

$$x_{input} = E_{txt}(T_{input}) \pm E_{txt}(T_{edit}) \tag{3}$$

This allows for additional text T_{edit} (which can be an image) to be used to display an image closer to the target, even if the initial search query did not retrieve the target image.

While the purpose of this step is to prepare a base image, an image prepared from outside the database can also be used.

3.2 StyleGAN Inversion

The image selected in the previous step is inverted into the StyleGAN2 latent space. In this study, we used the optimization-based GAN inversion method. First, we find w^* that satisfies the following equation:

$$w^* = argmax_w l\left(I^{selected}, G_{Synthesis}(w)\right) \tag{4}$$

where $G_{Synthesis}$ is the synthesis network of StyleGAN2 and l is the distance between the target image and the generated image. We used NVIDIA's StyleGAN2-ADA implementation[§] to perform inversion, where the distance l is the LPIPS distance using VGG-16.

3.3 Generation and Selection of Diverse Images

The generation of images by StyleGAN2 and the selection of images from within those images are iterated to assist in the divergence and convergence of the designer's ideas. In the i-th cycle, the procedure for generating the diverse images is as follows.

First, the latent representation corresponding to the image selected in the i-1-th cycle is $w_{i-1}^{selected}$, and the latent representation are is diffused by the following equation.

$$w_i = w_{i-1}^{selected} + \Delta \tag{5}$$

where Δ represents the minute change to the generated image and $w_0^{selected} = w^*$, which was inverted. Next, the image I_i is generated from the changed latent representation.

$$I_i = G_{Synthesis}(w_i) \tag{6}$$

In the proposed method, multiple images are generated simultaneously in each cycle. This encourages the divergence of ideas by displaying images that have undergone various changes from the one previously selected. The user selects the desired image, $I^{selected}$, and the latent representation corresponding to the selected image is $w_i^{selected}$, which is used as the base for the next cycle. Here, $w_i^{selected}$ is known because the image is generated by the Equation (6). By repeating this cycle of generation and selection, the user's ideas can be expanded by through the generation by StyleGAN2, while still allowing the ideas to converge through image selection in each cycle.

Note that Equation (5) can control diversity by varying the amount of change for each resolution in the synthesis network. The existing method, Style Mixing[4], controls the generated image by replacing the style vector of some resolutions with the style of other face images in face generation. In particular, when the resolution is low, the shape of the face changes, and when the resolution is high, details such as the hairstyle and skin color change. Similar properties are observed in the StyleGAN2 trained on buildings, with the external shape and structure of the building changing when only the low-resolution latent representation is changed, and the color and other detailed features changing when only the high-resolution latent representation is changed.

We believe that semantic diversity is important when generating images in each cycle in order to support the generation of ideas. Furthermore, if the latent representation is diffused completely at random, the style vector may change unintentionally, producing an image that does not resemble a building. Therefore, we use SeFa[25] as a method to make meaningful changes in the latent representation. SeFa is an unsupervised method for discovering meaningful directions in the latent space of a GAN.

In most GANs, when generating an image I_{gen} from a latent representation, it can be formulated as an affine transformation:

$$I_{gen} = G(z) = \boldsymbol{Az} + \boldsymbol{b} \tag{7}$$

[§] https://github.com/NVlabs/stylegan2-ada-pytorch

where G is the generator and z is the latent representation. At this time, the semantically meaningful direction n determined by SeFa is obtained by the following:

$$\boldsymbol{n}^* = argmax_{\{n\in\mathbb{R}^d:\boldsymbol{n}^t\boldsymbol{n}=1\}}\|\boldsymbol{An}\|_2^2 \tag{8}$$

where d is the number of dimensions of the latent space z and $\|\cdot\|_2^2$ denotes the l_2 norm. Note that $\boldsymbol{n}^*$, which maximizes the norm, is an eigenvector of $\boldsymbol{A^tA}$.

In StyleGAN2, affine transformation is performed before each style transformation.

$$y_j = \boldsymbol{a_j w} + \boldsymbol{b_j}, \qquad \boldsymbol{J} \in [\boldsymbol{0},\boldsymbol{J}] \tag{9}$$

where j represents the j-th style transformation and J corresponds to the number of style transformations in the synthesis network. A smaller j represents a style transformation at a lower resolution. At this time, SeFa is applied as follows.

$$\boldsymbol{A} = \left(\boldsymbol{a}_{\boldsymbol{j_{start}}}, \dots, \boldsymbol{a}_{\boldsymbol{j_{end}}}\right)^{\boldsymbol{t}}, \qquad \boldsymbol{0} \le \boldsymbol{j_{start}} \le \boldsymbol{j_{end}} \le \boldsymbol{J} \tag{10}$$

From $\boldsymbol{j_{start}}$ to $\boldsymbol{j_{end}}$ represents the style transformation whose representation the user want to change. Figure 1b. shows an example where $\boldsymbol{j_{start}} = \boldsymbol{0}, \boldsymbol{j_{end}} = \boldsymbol{2}$. We used P eigenvectors $n_1^*, \dots, n_p^*$ randomly selected from the direction vectors obtained by SeFa to change the latent representation as follows:

$$\Delta = \sum_{p=1}^{P} \alpha_p n_p^* \tag{11}$$

where α_p is the rate of change. Diversity can be controlled by setting j_{start}, j_{end} to small values to obtain ideas for structural features or to larger values for detailed feature ideas such as color.

4 Experiments

4.1 Experimental Settings

CLIP in image retrieval used the pretrained model by OpenAI[**]. The text encoder was Transformer and the image encoder was ViT-B/32. StyleGAN2 used a model pretrained on the LSUN church dataset and fine-tuned with 5052 contemporary building façade images collected from the Internet. StyleGAN2-ADA was used for fine-tuning. The image size was 256 x 256. In the experiment, we set $P = 5, 10 \le |\alpha_p| \le 20, j_{start} = 0, j_{end} = 13$. The image data in the database used in this paper are prepared only for validating the operation of the system. Some of the data includes images purchased from Pixta[††], which is mainly comprises contemporary photographs of the façades of office buildings, houses, and commercial facilities, as well a small number of indoor and non-building photographs.

4.2 Image Retrieval

In the first step, image retrieval, it is important to be able to find the user's target image. We examined the results of the text search and performed semantic operations when the target image was not displayed. The results are shown in Figure 2.

[**] https://github.com/openai/CLIP

[††] https://pixta.jp/

Figure 2a shows that the input text and semantically appropriate images can be found by image retrieval.

In Figure 2b, semantic operations are performed on the latent space of CLIP. Compared with the results of $E_{txt}("skyscraper")$ in Figure 2a, it can be observed that $E_{txt}("skyscraper") - E_{txt}("glass")$ does not display the blue building with glass walls, while $E_{txt}("skyscraper") + E_{txt}("green")$ shows the green skyscraper.

Note that this is the result of zero-shot retrievals by CLIP, which has not been trained to adapt to our data. Therefore, this result is domain- and database-independent.

(a) Nine images displayed from search query. (b) Semantic image transformation in latent space.

Figure 2: A text search yields nine semantically similar images.

4.3 Generation and Selection of Diverse Images

Next, we verify that a variety of images can be generated from a single base image while maintaining the characteristics of the building. Specifically, we determine that the system can generate various images and control the generation on the basis of the user's intention.

The results of images generated by SeFa from a single image are shown in Figure 3 The latent representation corresponding to the image on the left was diffused in the semantic direction to generate six images. The semantic changes made by SeFa altered the features of the building, resulting in a variety of generated images. Façade images with slightly different shapes and colors were generated without making extreme changes to the structure of the building.

The results of repeated image generation and selection are shown in Figure 4. From the inverted image, which was the base image for the idea generation, we selected an image with the intention of generating a vertical Façade, which emphasizes the vertical direction. Subsequently, a variety of vertical façades were generated. Furthermore, the system also provided façade designs that had different features from the base image. Although the base façade had a rounded shape, by selecting a rectangular image in the second and subsequent cycles, a variety of rectangular buildings were generated. Because our proposed system generates the next images on the basis of the selected image, the variety of generated images can be controlled by user selection.

In this way, multiple cycles of generation and selection provide an opportunity to discover new ideas that were not initially considered and expand the target design.

Figure 3: Diverse images generated by SeFa based on the left image. The resulting images vary in shape, color, and structure.

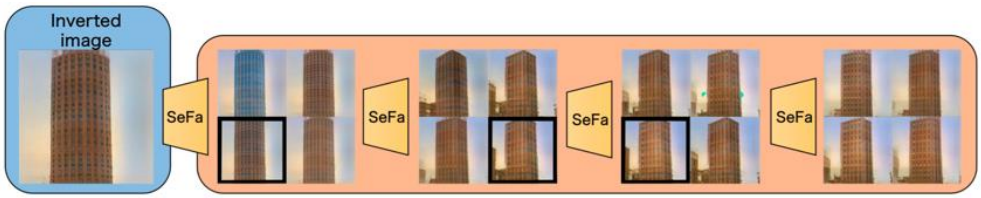


Figure 4: The area is the black box is repeatedly selected and generated to transform the inverted image into a rectangular building.

4.4 Flow of System

Finally, Figure 5 shows the result of repeated generation and selection after selecting the base image by image search. To determine the flexibility of image retrieval with CLIP, both images and text are used here. A glass-walled skyscraper, which is a search result of $E_{img}(I_{query}) + E_{txt}("glass-walled\ building")$, is entered in the searched. We select one of the results and obtain the corresponding style vector. After five iterations of generation and selection, a façade image is obtained with an outer shape that differs greatly from the base image selected by the retrieval.

In our system, the user only needs to prepare text or images as input. After inputting it into the system, a variety of designs can be collected by selecting the desired one from the images displayed by the system.

As shown in Figure 5, the image becomes slightly collapsed after five cycles. This is a limitation of the generation capability of StyleGAN2, and a future task is to ensure that StyleGAN2 can continue to generate a variety of images while maintaining quality. Nevertheless, this system is for supporting idea generation, and the quality of the generated images does not necessarily have to be high, as long as it leads to inspiration for the designer.

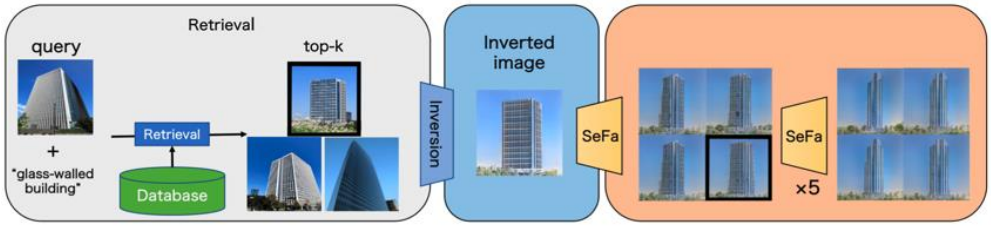


Figure 5: Beginning with image retrieval, after five iterations of generation and convergence, the images have changed significantly from the first selection.

4.5 Evaluation of Practicality

To verify the practicality of the proposed system, an evaluation was also conducted by the designers. The results indicated that the text-to-image retrieval displayed the correct target image, and the buildings generated were diverse. Thus, our system should be feasible for supporting designers.

However, it was pointed out that the "green" in Figure 2b generally refers to greenery as in "plants" rather than the "color" green in the designer's idea generation, so the result was different from the original intent. This is largely due to the fact that there were few images of green skyscrapers in the verification data. However, this is important for the purpose of supporting experts. However, we overcame the problem with limited number of data by using a generic model(CLIP), the model may not cover specialized experience. In addition, the designers suggested that further quality improvements in the future would provide more detailed design ideas.

5 Conclusion

We proposed an image generation system for generating design ideas for building façades. The proposed system generates a variety of images by diffusing latent representations in meaningful directions while moderately controlling their generation through selection. This encourages the divergence and convergence of ideas while also utilizing the designer's knowledge and experience. In addition, the base image for idea generation is selected through a flexible retrieval process using CLIP. The qualitative evaluation showed that the proposed system provides the user with ideas through the diversity of generated images, while also enabling the user to control the generation depending on their target. Future research directions include higher quality image generation and more flexible and intuitive control of generation based on expertise.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014..

[2] Zhengwei Wang, Qi She, and Tomas E Ward. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys (CSUR)*, 54(2):1– 38, 2021.

[3] Eoin Brophy, Zhengwei Wang, Qi She, and Tomas Ward. Generative adversarial networks in time series: A survey and taxonomy. *arXiv preprint arXiv:2107.11098*, 2021.

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[5] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*,

pages 1501–1510, 2017.

[6] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

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

[8] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

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

[10] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1429– 1437, 2019.

[11] Mohamed R Ibrahim, James Haworth, and Nicola Christie. Re-designing cities with conditional adversarial networks. *arXiv preprint arXiv:2104.04013*, 2021.

[12] MaximilianBachlandDanielCFerreira.City-gan:Learningarchitecturalstylesusing a custom conditional gan architecture. *arXiv preprint arXiv:1907.05280*, 2019.

[13] Sagar Joglekar, Daniele Quercia, Miriam Redi, Luca Maria Aiello, Tobias Kauer, and Nishanth Sastry. Facelift: a transparent deep learning framework to beautify urban scenes. *Royal Society open science*, 7(1):190987, 2020.

[14] Nelson Nauata, Kai-Hung Chang, Chin-Yi Cheng, Greg Mori, and Yasutaka Furukawa. House-gan: Relational generative adversarial networks for graph-constrained house layout geeration. In *European Conference on Computer Vision*, pages 162– 177. Springer, 2020.

[15] Nelson Nauata, Sepidehsadat Hosseini, Kai-Hung Chang, Hang Chu, Chin-Yi Cheng, and Yasutaka Furukawa. House-gan++: Generative adversarial layout refinement networks. *arXiv preprint arXiv:2103.02574*, 2021.

[16] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjo¨rn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[20] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020.

[21] Yichun Shi, Xiao Yang, Yangyue Wan, and Xiaohui Shen. Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11254–11264, 2022.

[22] HuanLing,KarstenKreis, DaiqingLi, SeungWookKim, AntonioTorralba, and Sanja Fidler. Editgan: High-precision semantic image editing. *Advances in Neural Information Processing Systems*, 34:16331–16345, 2021.

[23] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020.

[24] Erik Ha¨rko¨nen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020.

[25] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1532–1540, 2021.

[26] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning*, pages 9786–9796. PMLR, 2020.

[27] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[29] Guillaume Couairon, Matthieu Cord, Matthijs Douze, and Holger Schwenk. Embedding arithmetic for text-driven image transformation. *arXiv preprint arXiv:2112.03162*, 2021.