Background Image Editing with HyperStyle and Semantic Segmentation

Syuusuke Ishihata, Ryohei Orihara[a], Yuichi Sei[b], Yasuyuki Tahara[c] and Akihiko Ohsuga[d]
The University of Electro-Communications, Tokyo, Japan

Keywords: StyleGAN,GAN Inversion, Image Editing.

Abstract: Recently, research has been conducted on applying StyleGAN to image editing tasks. Although the technique can be applied to editing background images, because they are more diverse than foreground images such as face images, specifying an object in background images to be edited is difficult. For example, because natural language instructions can be ambiguous, edited images become undesirable for the user. It is challenging to resolve style and content dependencies in image editing. In our study, we propose an editing method that adapts Style Transformer, the latest GAN inversion encoder approach, to HyperStyle by introducing semantic segmentation to maintain the reconstruction quality and separate the style and the content of the background image. The content is edited while keeping the original style by manipulating the coarse part of latent variables and the residual parameters obtained by HyperStyle, and the style is edited without changing the content by manipulating the medium and fine part of latent vectors as in the conventional StyleGAN. As a result, the qualitative evaluation confirms that our model enabled the editing of image content and style separately, and the quantitative evaluation validates that the reconstruction quality is comparable to the conventional method.

1 INTRODUCTION

StyleGAN (Karras et al., 2019), one of the Generative Adversarial Networks (GANs) models with unsupervised learning, is capable of generating high quality images and has excellent interpolation performance between images. Several research projects use the ability of StyleGAN for image editing tasks. For example, StyleCLIP (Patashnik et al., 2021) edits the input image to match the content of the text. Latent codes are edited to produce images that match the content of the natural language using Contrastive LanguageImage Pre-training (CLIP) (Radford et al., 2021), a model used to classify natural language and images, as the loss function. In addition, a task called GAN Inversion estimates latent variables such that the Generator reconstructs the input image from them, and the estimated latent variables can be manipulated to edit the target image.

For the background images, it is useful to edit style and content separately, as shown in Figure 1. For example, it can reduce the time required to create a photo book or video work, from the point of view of generating various images from a single photo. However, background images include outdoors, such as mountains or forests, and indoors, such as rooms, or stadiums. Since background images are more diverse than foreground images such as face data, the quality of GANs' generated images is compromised. Editability is reduced accordingly.

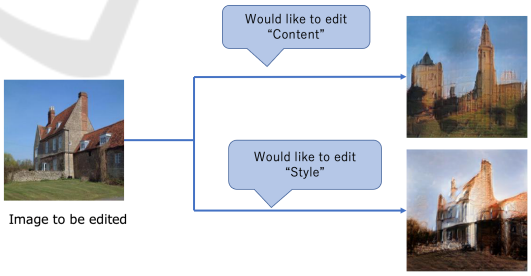

Figure 1: Overview of our research on image editing. We aim to edit either the content, style or both in the image.

Another problem is the difficulty of editing content. Although it is possible to edit the style of an image using StyleCLIP, it is difficult to edit the image intuitively when the image content is specified using text only. The results may differ from the editor's intention, appeared in the images as disappointing effects such as a slight misalignment or a different front-back relationship between objects. We checked

[a] https://orcid.org/0000-0002-9039-7704
[b] https://orcid.org/0000-0002-2552-6717
[c] https://orcid.org/0000-0002-1939-4455
[d] https://orcid.org/0000-0001-6717-7028

Ishihata, S., Orihara, R., Sei, Y., Tahara, Y. and Ohsuga, A.
Background Image Editing with HyperStyle and Semantic Segmentation.
DOI: 10.5220/0011661500003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 293-300
ISBN: 978-989-758-623-1; ISSN: 2184-433X

whether StyleCLIP could edit images in consideration of the content. In Figure 2, although the input image should be edited so that the tree is placed on the left side, it has been edited into an image where the whole image has been covered by a tree. The 'left' part of the text prompt was ignored. Editing with consideration for the contents was insufficient. On the other hand, a semantic segmentation mask provides a visual representation of the editor's intended content. Therefore, a semantic segmentation mask might be convenient for editing content. The GAN Inversion task has an approach using HyperNetworks that modifies the parameters of the Generator to achieve both reconstruction quality and editability of the generated image. In the case of background images, an encoder-based approach such as pixel2style2pixel (pSp) (Richardson et al., 2021) results in lower reconstruction quality. The quality of the Generator's performance is degraded due to the diversity of background images. Because HyperNetworks improves the performance of the Generator, it can solve the problem. In the GAN Inversion task, the performance is measured in a space defined by two axes, namely, reconstruction quality and editability. While both axes are important, our study focuses particularly on editability.

Therefore, we propose a framework for manipulating content with a semantic segmentation mask while maintaining the style editability of StyleGAN. The GAN Inversion approach called HyperStyle (Alaluf et al., 2021) uses HyperNetworks's outputs to adjust the Generator weights to improve the reconstruction quality. In the background image, the effect of the Generator adjustment causes a change in overall shape. We hypothesize that using a semantic segmentation mask as input to HyperNetworks could control the content of the image. We introduce two HyperStyle networks, one with the same inputs as the conventional method and the other set up for semantic segmentation mask and the input image, to achieve better control of the content.

This paper is organized as follows. We describe related work in Section 2, the proposed methodology in Section 3, the experiments in Section 4, and a discussion of the results in Section 5. We summarize this paper and discuss future works in Section 6.

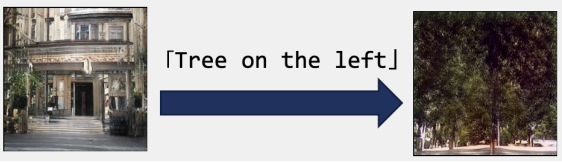

Figure 2: Example of editing a background image in StyleCLIP. The text prompt is 'Tree on the left'.

2 RELATED WORK

2.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are generative models that use two neural networks, Generator and Discriminator. The Generator fools the Discriminator to recognize the generated data as training data, and the Discriminator counters the Generator by correctly recognizing the generated data as fake data. The alternate learning approach improves the quality of the generated data. Based on the approach, various image generation and transformation approaches have been proposed, including DCGAN, which uses a Convolutional Neural Network (CNN) to improve GAN's performance, Pix2pix (Isola et al., 2017) and Pix2pixHD (Wang et al., 2018) for image transformation, PGGAN (Karras et al., 2018) for higher resolution, and AttnGAN (Xu et al., 2018) for generating images from text. Examples of image transformations include converting black-and-white images or line drawings to color images, however, there are also approaches to synthesizing from a semantic segmentation mask. Unlike U-Net(Ronneberger et al., 2015)-based approaches such as pix2pix and pix2pixHD, semantic segmentation is incorporated into normalization approaches such as SPatially-Adaptive (DE)normalization (SPADE) (Park et al., 2019) and semantic region-adaptive normalization (SEAN) (Zhu et al., 2020) to synthesize images according to their segmentation-labeled shapes. While these image-to-image approaches allow rough editing of images, it is difficult to control the style in such a way that StyleGAN does.

StyleGAN is the generative model that enables the generation of higher-resolution images. Latent code w which is transformed from stochastically generated variable z in an MLP-based Mapping Network affects image style. It is possible to control the representation of the coarse to the fine style of an image by w. For example, in the case of face images, face orientation and age can be changed by adding vectors in the latent variable. StyleGAN2 is an improved version of StyleGAN, which eliminates Adaptive Instance Normalization (AdaIN) and uses weight demodulation to normalize and convolve the weight demodulation. Since GAN Inversion estimates latent variables from images, it is easier to edit images for it than StyleGAN, which transforms noise into latent variables with a Mapping Network. We use GAN Inversion in our method, which employs the Generator of StyleGAN2.

In the Generator of a GAN (e.g., StyleGAN), because the latent variable used in the input determines

the image to be generated, manipulation of latent variables can be used to edit the image in the desired way. For example, StyleCLIP (Patashnik et al., 2021) is an approach for editing images with text that takes advantage of the expressive ability of StyleGAN. In addition to StyleGAN, it uses CLIP (Radford et al., 2021), a multimodal image classification model that learns the relationship between natural language text and images, as a loss function. Using CLIP as the loss function, image editing can be performed by the text content.

2.2 GAN Inversion

GAN Inversion is the estimation of latent variables from a real image such that the GAN generator can reproduce the image. Such methods include those that (1) directly optimize latent variables, where reconstruction quality is high, and optimization takes time such as Pivotal Tuning Inversion (PTI) (Roich et al., 2021), and (2) encoder-based approach, encoding images into latent vectors such as pixel2style2pixel (pSp) (Richardson et al., 2021) and Style Transformer (Hu et al., 2022). The encoder estimates a latent variable such that the Generator produces the same image as the input, as shown in Figure 3. Although encoder-based approaches have a faster estimation time, reconstruction quality tends to be lower. Since Generator is often a pre-trained model, an encoder is trained only in many approaches.

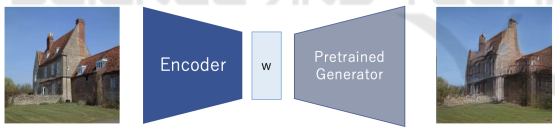

Figure 3: Overview of General type encoder GAN Inversion pipeline. First, the encoder estimates latent codes to be input to the pre-trained Generator from images. Then, Generator creates the same images as the encoder's input.

For GAN Inversion, especially encoder-based approaches, there are also approaches to improve the reconstruction quality of the generated images by updating Generator parameters with HyperNetworks (Ha et al., 2017) which is a model to learn Neural Network (NN) parameters, such as HyperStyle (Alaluf et al., 2021). In HyperStyle the parameter $\hat{\theta}$ is given by modifying Generator's parameter θ as the following equation.

$$\hat{\theta}_l^{i,j} = \theta_l^{i,j}(1+\Delta_l^{i,j}) \quad (1)$$

$\theta_l^{i,j}$ is the weights for the j-th channel of the i-th filter in the convolution layer for the l-th Generator. Another approach similar to HyperStyle is HyperInverter (Dinh et al., 2022). Our study used HyperStyle, one of the HyperNetworks-based approaches.

3 METHOD

The goal of our study is to enable flexible editing by separating style and content without compromising reconstruction quality. Therefore, we focused on the GAN Inversion method. In GAN Inversion, it is said that the relationship between reconstruction quality and editability is a trade-off (Tov et al., 2021). Many approaches have been devised to solve the problem. One of the approaches is HyperStyle which aims to eliminate the trade-off. First, we began by analyzing HyperStyle, then the architectural details and loss functions are described.

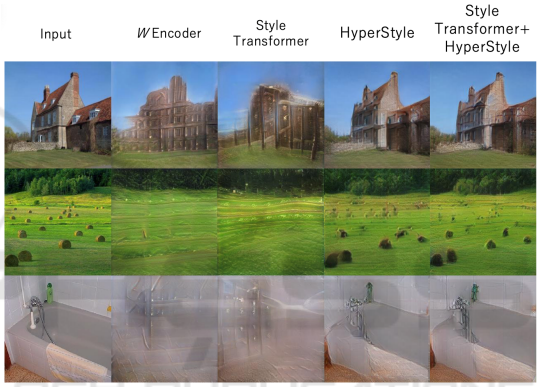

Figure 4: Example of reconstruction quality of GAN Inversion. *W* Encoder and Style Transformer are the encoder network in GAN Inversion. In both cases, the output image is blurred, but HyperStyle improves the reconstruction quality and clarifies the shape of objects.

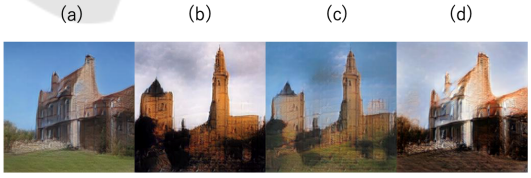

Figure 5: Confirmation of content information by residue parameters obtained from HyperStyle. Input images are (a) and (b). GAN Inversion (c) of (a) using StyleGAN2 Generator adjusted by the residual parameters according to (b). GAN Inversion (d) of (b) in the Generator of StyleGAN2 adjusted by the residual parameters according to (a).

3.1 HyperStyle

The encoder-based approaches lose the information of the input image when encoding it into latent variables. Since background images are a more diverse

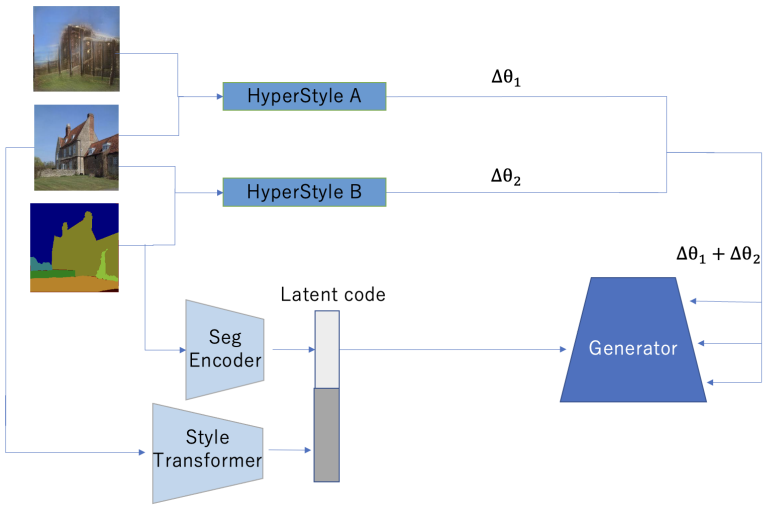

Figure 6: Our Network Architecture. Seg Encoder and GAN Inversion encoder, Style Transformer estimate latent variables. Then, the semantic segmentation mask, initial reconstructed images, and real images are input to two HyperStyle networks. The input to HyperStyleA is a pair of the initial reconstructed image and a real image, and the input to HyperStyleB is a pair of semantic segmentation mask and a real image. The outputs of each HyperStyle, $\Delta\theta_1$ and $\Delta\theta_2$, are added together and are used to modify Generator's weights.

data set than foreground images such as face images, it is difficult to reconstruct an image with encoder-only GAN Inversion like W Encoder and Style Transformer. W Encoder is a model for W space defined in the ablation study in pSp, and HyperStyle also uses the encoder to improve editability. Although Style Transformer also originally gets latent codes in $W+$ space, HyperStyle uses W space. We adjust the output of Style Transformer to be a point in W space. As shown in columns 2 and 3 in Figure 4, the reconstructed image results are quite different from the input image, and the entire image is blurred. The style of the whole image remains, however, the information on the content in the image is lost. The content information of the input image is lost when encoding it to latent variables. HyperStyle has recovered it. When the residual parameters taken from the HyperStyle's HyperNetworks were replaced with the residual parameters obtained from another image, the content of the image becomes that of another image. For example, (c) in Figure 5 shows the GAN Inversion in (a), where the residual parameters for the Generator are replaced with those obtained from (b). The content is that of (b) and the style remains the same as in (a). We wonder if the residual parameter might play an important role in content editing.

3.2 Architecture

Figure 6 shows our architecture. Two HyperStyles are prepared, and their outputs, the residual parameters, are added to the trained Generator parameters. The input data to the two HyperStyle networks are semantic segmentation masks, initially reconstructed images, and real images.

The input to HyperStyleA is a pair of the initial reconstructed image and a real image, and the input to HyperStyleB is a pair of semantic segmentation mask and a real image. The Generator, modified by the residual parameters, generates a reconstructed image that reproduces the real image from the latent variables. In latent variables, it would be possible to control the shape of the image using information related to the content of the image, such as the output of the semantic segmentation mask encoder, Seg Encoder, in the lower convolution layer of the Generator. Therefore, the low to medium resolution of latent variables obtained from Style Transformer are replaced with the outputs of Seg Encoder. The output of the Style Transformer is essentially a point in $W+$ space with different latent variables input to each layer. However, to achieve both editability and reconstruction quality, HyperStyle assumes the latent variables are distributed in W space, which is considered to have higher editability although its reconstruction quality is lower than that of $W+$ space. In this paper, we adjust the latent variable for the output of Style Transformer to be $w \in \mathbb{R}^{1\times 512}$ to avoid impairing editability.

Due to the nature of StyleGAN, the input latent variables to the high-resolution layer can control the representation of the fine parts of the image. There-

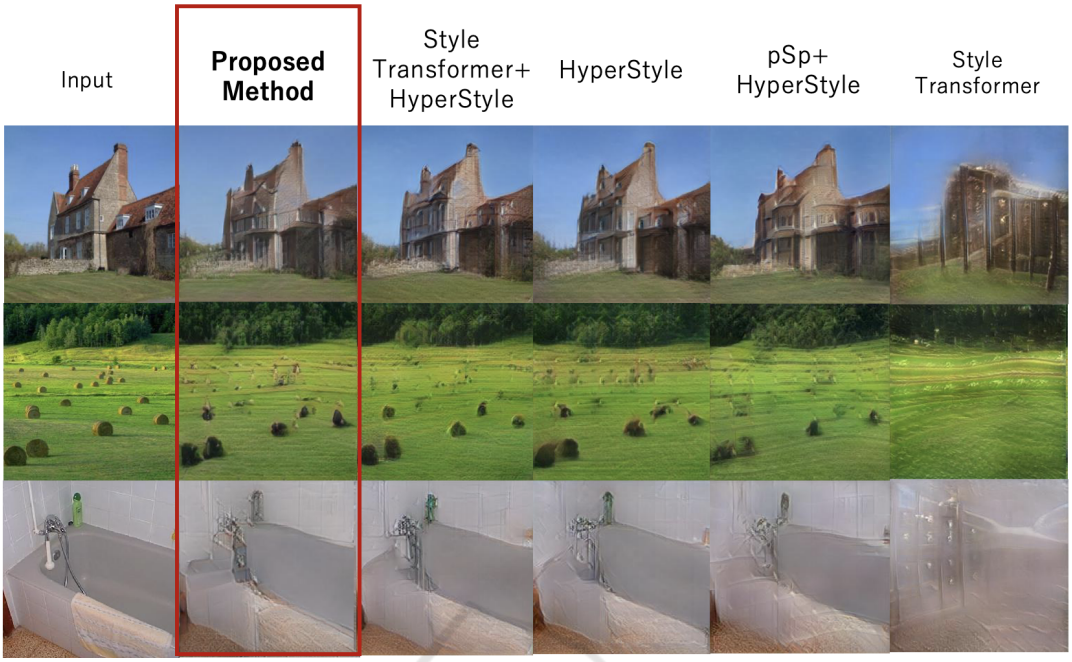

Figure 7: Result of reconstructed images. The proposed method is visually equivalent to other HyperStyle methods, and the reconstruction results are significantly better than the Style Transformer-only encoder method.

fore, when mixing the style without changing the content, the medium to high resolution portion of the latent variables are replaced. From Section 3.1, the residual parameters by HyperNetworks contribute significantly to recovering the missing content information. In addition, the input latent variables to the low-resolution layer can control the representation of the rough part of the image. To mix the content while preserving the style of the original image, the residual parameters and the output of the Seg Encoder from the original image are replaced with the residual parameters and the output of the Seg Encoder from another image.

3.3 Loss Function

The model of Style Transformer has been trained in advance, and the HyperStyle network is trained in the model. The loss function of the model is the same as that of HyperStyle (Alaluf et al., 2021), as shown in the following formula.

$$\lambda_2 L_2(x,\hat{y}) + \lambda_{sim} L_{sim}(x,y,\hat{y}) + \lambda_{perc} L_{LPIPS}(x,\hat{y}) \quad (2)$$

x and y are identical in the task and both are images from the original data set. $\hat{y}$ is the output of StyleGAN with adjusted parameters. For similarity loss, an identity-based face recognition model is often used for tasks that generate and edit face images (Richardson et al., 2021). In our study, because the focus is on editing background images, MoCo-based (Tov et al., 2021) similarity loss is used.

4 EXPERIMENTS

4.1 Implementation Details

It is necessary to train the model on a dataset with as many scenes as possible. We use the ADE20K dataset (Zhou et al., 2017). The dataset has 20,210 training data and 2,000 validation data for background scenes. With the dataset, we use the pre-trained StyleGAN2 (Karras et al., 2020) Generator with 200,000 iterations and Style Transformer(Hu et al., 2022) with 100,000 iterations. Its output is an image whose resolution of 256×256, and the same is true for both the input and semantic segmentation images. Our method is implemented in Pytorch. We train the models of the method, namely HyperStyleA, B, and Seg Encoder, for 200,000 iterations. We set $\lambda_2 = 1.0$, $\lambda_{perc} = 0.8$, and $\lambda_{sim} = 0.5$ in the loss function. Ranger (Wright, 2019) is employed as the optimization approach with a learning rate of $lr = 0.0001$.

4.2 Reconstruction Quality Evaluation

The reconstruction quality is evaluated using qualitative and quantitative evaluation.

Table 1: Quantitative results of background image reconstruction quality, FID, and KID scores on ADE20k (Zhou et al., 2017) validation data.

	Quality of Image Reconstruction				Fidelity	
method	L2(↓)	LPIPS(↓)	PSNR(↑)	MS-SSIM(↑)	FID(↓)	KID($\times 10^3$)(↓)
Proposed method	0.06412	0.27514	18.25542	0.57358	**44.97**	**18.370**
Style Transformer+HyperStyle	**0.05276**	**0.22515**	**19.11787**	**0.64569**	47.50	19.990
pSp + HyperStyle	0.05982	0.28317	18.54680	0.59826	58.06	26.121
HyperStyle	0.05547	0.23650	18.88020	0.62725	48.10	19.907
Style Transformer	0.10120	0.45700	16.20895	0.36557	159.99	113.650
pSp	0.08000	0.36000	17.22778	0.47954	78.22	39.871
W Encoder	0.10000	0.44000	16.14369	0.36623	169.00	110.610

The evaluation of reconstruction quality shows that the results of our method are visually comparable to the approaches using HyperStyle in columns 3 to 5 in Figure 7. In particular, the proposed method reproduces the shape of the image better than the approach using pSp as an encoder (column 5). In addition, the entire image is blurred for Style Transformer (column 6), which is a simple encoder. The proposed method produced the reconstruction image closer to the original input image.

In the experiment, a quantitative evaluation is conducted using the same indicators as for HyperInverter (Dinh et al., 2022), a method similar to HyperStyle. We evaluated the quality of image reconstruction using L2 distance, Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), Peak Signal to Noise Ratio (PSNR) and multi-scale structural similarity (MS-SSIM) (Wang et al., 2003), and the realism of images using Frchet inception distance (FID) (Heusel et al., 2017) and Kernel Inception Distance (KID) (Bikowski et al., 2018) metrics, which are often used in GAN methods. The FID and KID metrics measure the fidelity between real and generated images. The lower scores are for these, the closer the generated and real images are. The results are shown in Table 1. It compares (1) the three encoder methods, *W* Encoder, pSp, and Style Transformer used in our method, and (2) the approaches that adapt HyperStyle to them. HyperStyle in row 6 of Table 1 is the approach when the encoder is *W* Encoder. The results show that our method is the best and more realistic in terms of FID and KID metrics, although the reconstruction quality is slightly inferior to other methods using HyperStyle.

4.3 Editability Evaluation

The evaluation is based on two criteria: The result of mixing the style only while preserving the content of the original image (Figure 8), and the result of mixing the content only while preserving the original image style (Figure 9).

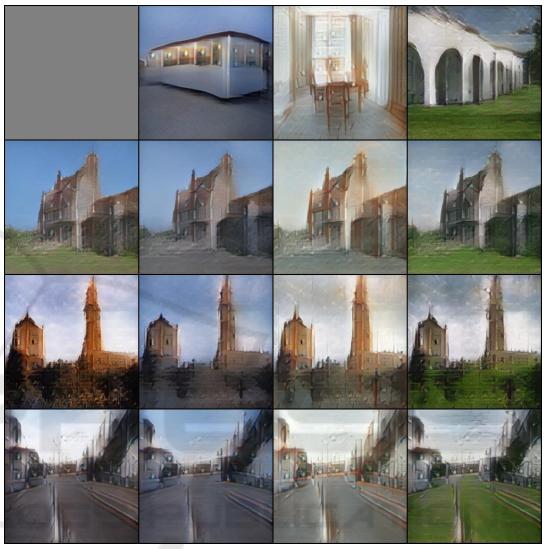

Figure 8: Result of style editing. It is a style-only mixing and the content component is unchanged. The first row and the first column are the input images. The style of the image in each column are edited to the style of the image in the first row.

In Figure 8, the style is transferred from the image in the first row onto the image in each column. It can be seen that the global style of the image in the first column has been changed and the content of the original image has been retained. For example, as shown in the second row, the sky has changed from clear to cloudy. Similarly, in the third row, the sky has changed, but the shape of the cloud remains the same. On the other hand, Figure 9 shows that the image content is that of the reference image while the source image's style is retained. It can be seen that the shape of the buildings and the ground has changed without any color change. Because these results show that images can be edited separately for style and content, our method has high editability.

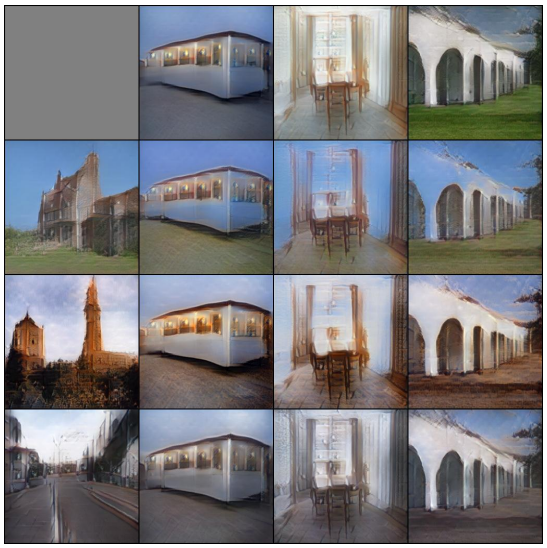

Figure 9: Result of content editing. The first row and the first column are the input images. It is a content-only mixing and the image's style is fixed. The content of the image in each column is edited to the content of the image in the first row.

5 DISCUSSION

In previous research, it was difficult to edit image style and content separately for highly diverse domains such as background images. The traditional image editing approaches by latent variable manipulation, as in pSp, also slightly change the style of the image when trying to edit the content of the image. Figures 8 and 9 show that our method can edit the image style and content independently. The advantage of our method is that either the style or the content can be edited without style and content dependencies in image editing.

Table 1 shows that other HyperStyle-based approaches without semantic segmentation tend to have better metrics on reconstruction quality than the proposed method. We assume that the reason is how those two HyperStyles were trained. The amount of parameter changes to modify the Generator has increased and the number of HyperNetworks to be trained has also increased, which may have prevented fine-tuning of the parameters. However, the evaluation by FID and KID scores shows better results than the other methods, indicating its superiority in terms of fidelity. These results suggest that improving the quality of reconstructed images is an issue.

In the method, latent variables in W space were employed for editability, however, $W+$ space such as the one used by pSp is said to provide higher reconstruction quality than it. Looking at the results in Table 1, from W Encoder and pSp results, it certainly appears that $W+$ space has an advantage in reconstruction quality. However, when HyperStyle is applied, the reconstruction quality is better with the encoder in W space. We assume that the reason is that the improvement in reconstruction quality in W space is more dependent on the residual parameters related to Generator performance than in $W+$ space. Therefore, W space is more appropriate for the method using HyperStyle. It is assumed that the reconstruction quality can be improved by devising factors other than latent space, such as a GAN Inversion encoder that predicts the correct latent variable in W space and a residual parameter by HyperNetworks.

6 CONCLUSIONS

In this paper, the problem of poor reconstruction quality of GAN Inversion due to the diversity of background images is solved by HyperStyle, a method to update the parameters of the Generator using HyperNetworks. In addition, we can confirm that editing content, which was impossible with text, is feasible using HyperNetworks' residual parameters. Our method allows flexible editing of background images with style and content separately while the quality of reconstruction images is comparable to existing approaches, such as HyperStyle.

There are three future works.

The first is to improve content editability. Mixing of image content could be achieved using the output of Seg Encoder and HyperNetworks' residual parameters. However, there is a limitation in editing because an image other than the image to be edited is required. In the future, we would like to explore methods like SPADE (Park et al., 2019) and SEAN (Zhu et al., 2020) that control content editing using semantic segmentation masks only. We aim to realize an intuitive method of editing background images that considers usability.

The next step is to improve reconstruction quality. The residual parameters by HyperNetworks are important for the purpose. We need to explore a GAN Inversion encoder that can improve the quality of reconstruction while considering it.

Lastly, we will use text-based style control. In the case of a text-based image editing approach such as StyleCLIP, style control is achieved by manipulating latent vectors. Our method would be able to do the same by manipulating latent variables according to the textual content. Initially, we will apply StyleCLIP to our method.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP21H03496, JP22K12157.

REFERENCES

Alaluf, Y., Tov, O., Mokady, R., Gal, R., and Bermano, A. H. (2021). Hyperstyle: Stylegan inversion with hypernetworks for real image editing.

Bikowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying mmd gans.

Dinh, T. M., Tran, A. T., Nguyen, R., and Hua, B.-S. (2022). Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ha, D., Dai, A. M., and Le, Q. V. (2017). Hypernetworks. In *International Conference on Learning Representations*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hu, X., Huang, Q., Shi, Z., Li, S., Gao, C., Sun, L., and Li, Q. (2022). Style transformer for image inversion and editing. *arXiv preprint arXiv:2203.07932*.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *CVPR*.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D. (2021). Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., and Cohen-Or, D. (2021). Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Roich, D., Mokady, R., Bermano, A. H., and Cohen-Or, D. (2021). Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., and Cohen-Or, D. (2021). Designing an encoder for stylegan image manipulation. *arXiv preprint arXiv:2102.02766*.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, pages 1398–1402.

Wright, L. (2019). Ranger - a synergistic optimizer. https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhu, P., Abdal, R., Qin, Y., and Wonka, P. (2020). Sean: Image synthesis with semantic region-adaptive normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.