기하적 변형이 포함된 이미지의 GAN Inversion*

강경국[0,1], 김성태[2], 조성현[1,2]
포항공과대학교 [1]컴퓨터공학과, [2]인공지능대학원
{kkang831, seongtae0205, s.cho}@postech.ac.kr

GAN Inversion for Out-of-Range Images with Geometric Transformations

Kyoungkook Kang[0,1], Seongtae Kim[2], Sunghyun Cho[1,2]
{[1]Dept. of Computer Science and Engineering, [2]Graduate School of Artificial Intelligence} POSTECH

Abstract

In this paper, we propose a novel GAN inversion approach to semantic manipulation of out-of-range images that are geometrically unaligned with the training images of a GAN model. To find a latent code that is semantically editable, our approach inverts an input out-of-range image into an alternative latent space than the original latent space. We also propose a regularized inversion method to find a proper solution that supports semantic manipulation in the alternative space. Our experiments show that our approach effectively supports semantic manipulation of out-of-range images with geometric transformations.

1. Introduction

Recently, it has been shown that rich semantic information is encoded in the latent space of GANs, and furthermore, that images can be effectively manipulated in a semantically meaningful way by modifying latent code. To enable such semantic manipulation for real images, GAN inversion has recently attracted much attention [3, 4, 5, 6]. GAN inversion maps a real image into the latent space of a pre-trained GAN model. As shown in [3], for successful semantic manipulation of real images, it is critical to find an in-domain latent code of a pre-trained GAN model.

Unfortunately, such in-domain latent codes can be found only for a small fraction of real images that align with the training images of a pre-trained GAN model. For example, most GAN models use geometrically aligned face images as their training data for ease of training. As a result, images with a small amount of translation or other geometric transforms are out of their ranges, i.e. images without the align method used when creating the training set, and the previous GAN inversion methods cannot find in-domain latent codes for such out-of-range images. This severely limits the applicability of semantic editing of real images using GAN inversion.

In this paper, we propose a novel GAN inversion approach to semantic manipulation of out-of-range images. Specifically, Our approach inverts an image that is geometrically unaligned with the training images for the StyleGAN [1] and StyleGAN2 [2] frameworks. To this end, we propose to invert an image into another space $F\,/\,W^{+}$ space. To find a in-domain latent code in the $F\,/\,W^{+}$ space that faithfully reconstructs the input image and supports semantic manipulation, we also propose a regularization approach for the $F\,/\,W^{+}$ space based on an encoder network.

2. Our Approach

StyleGAN frameworks [1, 2] have a unique intermediate latent space, and $w \in W$ is fed to multiple layers of different scales of the generator to control the style of each scale. For StyleGAN inversion, more extended latent space W^{+} is generally used to enhance the reconstruction accuracy [4]. A $w^{+} \in W^{+}$ is a set of w vectors $\{w_1, w_2, \ldots, w_N\}$, where each w is the input of each layer.

Nonetheless, GAN inversion to the extended latent space W^{+} still fails to find an in-domain latent code for out-of-range images as discussed in Sec. 1. To overcome this limitation, we propose another latent space $F\,/\,W^{+}$ where each element w^{*} in the space is defined as $w^{*} = (f, w_{M+})$. w_{M+} is a subset of w^{+}, which is the inputs of the layers larger than a specific

* 구두발표논문
* 본 논문은 요약논문 (Extended Abstract) 으로서, 본 논문의 원본 논문은 현재 타 학술대회에 제출 중
* 본 연구는 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2019-0-01906, 인공지능대학원 지원(포항공과대학교))과 한국연구재단의 지원(No. 2020R1C1C1014863)을 받아 수행되었음.

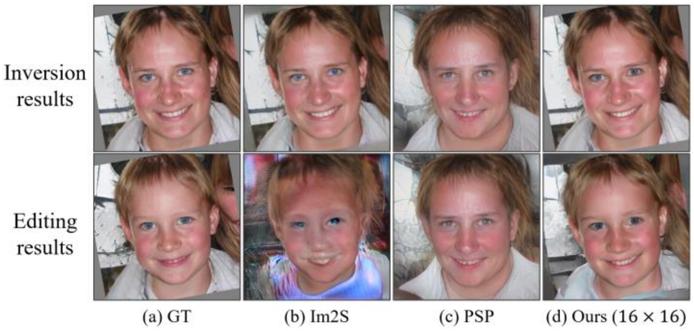

Figure 1. Qualitative comparison of the reconstruction quality and editing quality of different methods on geometrically transformed images. We sample 50 images from the CelebA-HQ dataset and applied different degrees of translation, rotation, and scaling. In this example, an input image is rotated by 20 degrees, and aging editing is applied.

Mth layer for the fine scales of the generator. f is a coarse-scale feature map of the generator before the Mth layer. In our experiments, we test two scales, 8×8 and 16×16, for f.

The $F\,/\,W^{+}$ space provides a couple of nice properties that enable semantic editing of out-of-range images. First, compared to $\{w_1, \ldots, w_{M-1}\}$, f can represent a wider range of images including images with geometric transformations, as f has a greater degree of freedom. Second, w_{M+} is invariant to translations of images, as it is the input of the spatially global operation of the StyleGAN frameworks.

To enable semantic manipulation for out-of-range images, both f and w_{M+} must be in proper domains. To this end, we adopt a regularized optimization scheme both on f and w_{M+}. For w_{M+}, we adopt the $P-norm^{+}$ space-based regularization proposed by Zhu et al. [5]. For f, we first find an initial latent code f^{o} that lies in the extended domain of f using an encoder E and find a latent code f that is close to f^{o}.

For the training of the encoder, we sample latent codes $\left(f^{gt}, w_{M+}^{gt}\right)$ and its image I and train our encoder E with a loss function defined as:

$$L_{enc} = \left\|G\left(E(I), w_{M+}^{gt}\right) - I\right\|^2 + \lambda \left\|F\left(G\left(E(I), w_{M+}^{gt}\right)\right) - F(I)\right\|^2$$

where λ is a weight to balance two terms, and F is a LPIPS [7] network to calculate perceptual similarity. The encoder has a VGG-like architecture consisting of 11 convolution blocks and three pooling layers. Our training procedure does not use geometrically transformed images. Nevertheless, our encoder still performs effectively for geometrically transformed images thanks to the spatially-invariant property of CNNs.

3. Experimental Results

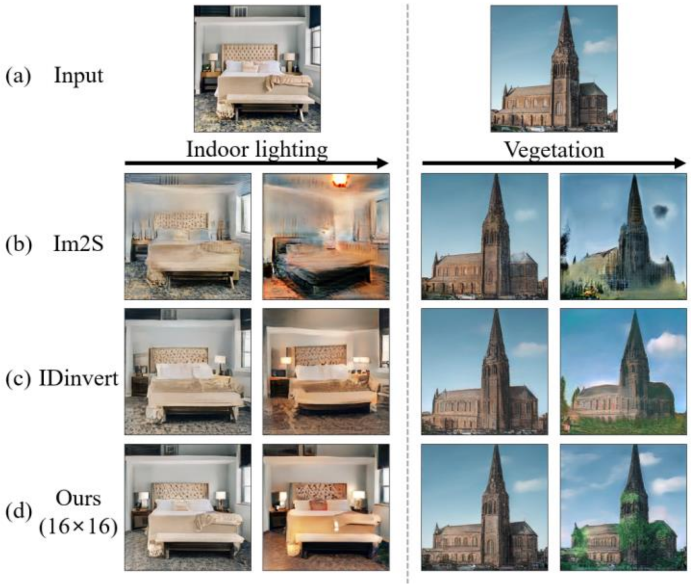

Figure 2. Qualitative comparison of the reconstruction quality and editing quality of different methods on natural images. The input images on the top row are collected from the internet. We use StyleGAN [1] models pre-trained on the LSUN bedroom and tower datasets.

We compare our method with recent state-of-the-art approaches: IDinvert[3], Im2S [4], and PSP [6]. Figure 1 shows a reconstruction and editing quality comparison. The figure shows that our 16×16 version can reconstruct input image and successfully edit inversion result. Only Im2S shows high-quality reconstruction results. However, due to the lacks in-domain constraints, Im2S tends to produce out-of-domain latent codes that are not semantically editable.

Due to the large diversity of natural images, it is difficult to accurately reconstruct and edit a natural image using previous GAN inversion approaches. On the other hand, thanks to the high degree-of-freedom of the $F\,/\,W^{+}$ space, our approach is especially effective in handling such natural images. Figure 2 shows reconstruction and editing quality comparisons.

Reference

[1] Karras, T et al. A style-based generator architecture for generative adversarial networks. In CVPR, 2019.

[2] Karras, T et al. Analyzing and improving the image quality of stylegan. In CVPR, 2020.

[3] Zhu, J et al. In-domain gan inversion for real image editing. In ECCV, 2020.

[4] Abdal, R et al. Image2stylegan: How to embed images into the stylegan latent space? In CVPR, 2019.

[5] Zhu, P et al. Improved stylegan embedding: Where are the good latents? arXiv preprint arXiv:2012.09036, 2020.

[6] Richardson, E et al. Encoding in style: a stylegan encoder for image-to-image translation. In CVPR, 2021.

[7] Zhang, R et al. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018.