Image Processing Using Multi-Code GAN Prior

Jinjin Gu\textsuperscript{1,2}, Yujun Shen\textsuperscript{1}, Bolei Zhou\textsuperscript{1}
\textsuperscript{1}The Chinese University of Hong Kong \hspace{0.5cm} \textsuperscript{2}The Chinese University of Hong Kong, Shenzhen
jinjingu@link.cuhk.edu.cn, \{sy116, bzhou\}@ie.cuhk.edu.hk

Abstract

Despite the success of Generative Adversarial Networks (GANs) in image synthesis, applying trained GAN models to real image processing remains challenging. Previous methods typically invert a target image back to the latent space either by back-propagation or by learning an additional encoder. However, the reconstructions from both of the methods are far from ideal. In this work, we propose a novel approach, called mGANprior, to incorporate the well-trained GANs as effective prior to a variety of image processing tasks. In particular, we employ multiple latent codes to generate multiple feature maps at some intermediate layer of the generator, then compose them with adaptive channel importance to recover the input image. Such an over-parameterization of the latent space significantly improves the image reconstruction quality, outperforming existing competitors. The resulting high-fidelity image reconstruction enables the trained GAN models as prior to many real-world applications, such as image colorization, super-resolution, image inpainting, and semantic manipulation. We further analyze the properties of the layer-wise representation learned by GAN models and shed light on what knowledge each layer is capable of representing.\footnote{Code is available at this link.}

1. Introduction

Recently, Generative Adversarial Networks (GANs) [16] have advanced image generation by improving the synthesis quality [23, 8, 24] and stabilizing the training process [1, 7, 17]. The capability to produce high-quality images makes GANs applicable to many image processing tasks, such as semantic face editing [27, 36], super-resolution [28, 42], image-to-image translation [53, 11, 31], etc. However, most of these GAN-based approaches require special design of network structures [27, 53] or loss functions [36, 28] for a particular task, limiting their generalization ability. On the other hand, the large-scale GAN models, like StyleGAN [24] and BigGAN [8], can synthesize photo-realistic images after being trained with millions of diverse images. Their neural representations are shown to contain various levels of semantics underlying the observed data [21, 15, 35, 44]. Reusing these models as prior to real image processing with minor effort could potentially lead to wider applications but remains much less explored.

The main challenge towards this goal is that the standard GAN model is initially designed for synthesizing images from random noises, thus is unable to take real images for any post-processing. A common practice is to invert a given image back to a latent code such that it can be reconstructed by the generator. In this way, the inverted code can be...
used for further processing. To reverse the generation process, existing approaches fall into two types. One is to directly optimize the latent code by minimizing the reconstruction error through back-propagation [30, 12, 32]. The other is to train an extra encoder to learn the mapping from the image space to the latent space [34, 52, 6, 5]. However, the reconstructions achieved by both methods are far from ideal, especially when the given image is with high resolution. Consequently, the reconstructed image with low quality is unable to be used for image processing tasks.

In principle, it is impossible to recover every detail of any arbitrary real image using a single latent code, otherwise, we would have an unbeatable image compression method. In other words, the expressiveness of the latent code is limited due to its finite dimensionality. Therefore, to faithfully recover a target image, we propose to employ multiple latent codes and compose their corresponding feature maps at some intermediate layer of the generator. Utilizing multiple latent codes allows the generator to recover the target image using all the possible composition knowledge learned in the deep generative representation. The experiments show that our approach significantly improves the image reconstruction quality. More importantly, being able to better reconstruct the input image, our approach facilitates various real image processing applications by using pre-trained GAN models as prior without retraining or modification, which is shown in Fig.1. We summarize our contributions as follows:

- We propose mGANprior, shorted for multi-code GAN prior, as an effective GAN inversion method by using multiple latent codes and adaptive channel importance. The method faithfully reconstructs the given real image, surpassing existing approaches.
- We apply the proposed mGANprior to a range of real-world applications, such as image colorization, super-resolution, image inpainting, semantic manipulation, etc., demonstrating its potential in real image processing.
- We further analyze the internal representation of different layers in a GAN generator by composing the features from the inverted latent codes at each layer respectively.

2. Related Work

**GAN Inversion.** The task of GAN inversion targets at reversing a given image back to a latent code with a pre-trained GAN model. As an important step for applying GANs to real-world applications, it has attracted increasing attention recently. To invert a fixed generator in GAN, existing methods either optimized the latent code based on gradient descent [30, 12, 32] or learned an extra encoder to project the image space back to the latent space [34, 52, 6, 5]. Bau et al. [3] proposed to use encoder to provide better initialization for optimization. There are also some models taking invertibility into account at the training stage [14, 13, 26]. However, all the above methods only consider using a single latent code to recover the input image and the reconstruction quality is far from ideal, especially when the test image shows a huge domain gap to training data. That is because the input image may not lie in the synthesis space of the generator, in which case the perfect inversion with a single latent code does not exist. By contrast, we propose to increase the number of latent codes, which significantly improve the inversion quality no matter whether the target image is in-domain or out-of-domain.

**Image Processing with GANs.** GANs have been widely used for real image processing due to its great power of synthesizing photo-realistic images. These applications include image denoising [9, 25], image inpainting [45, 47], super-resolution [28, 42], image colorization [38, 20], style mixing [19, 10], semantic image manipulation [41, 29], etc. However, current GAN-based models are usually designed for a particular task with specialized architectures [19, 41] or loss functions [28, 10], and trained with paired data by taking one image as input and the other as supervision [45, 20]. Differently, our approach can reuse the knowledge contained in a well-trained GAN model and further enable a single GAN model as prior to all the aforementioned tasks without retraining or modification. It is worth noticing that our method can achieve similar or even better results than existing GAN-based methods that are particularly trained for a certain task.

**Deep Model Prior.** Generally, the impressive performance of the deep convolutional model can be attributed to its capacity of capturing statistical information from large-scale data as prior. Such prior can be inversely used for image generation and image reconstruction [40, 39, 2]. Upchurch et al. [40] inverted a discriminative model, starting from deep convolutional features, to achieve semantic image transformation. Ulyanov et al. [39] reconstructed the target image with a U-Net structure to show that the structure of a generator network is sufficient to capture the low-level image statistics prior to any learning. Athar et al. [2] learned a universal image prior for a variety of image restoration tasks. Some work theoretically explored the prior provided by deep generative models [32, 18], but the results using GAN prior to real image processing are still unsatisfying. A recent work [3] applied generative image prior to semantic photo manipulation, but it can only edit some partial regions of the input image yet fails to apply to other tasks like colorization or super-resolution. That is because it only inverts the GAN model to some intermediate feature space instead of the earliest hidden space. By contrast, our method reverses the entire generative process, i.e., from the image space to the initial latent space, which supports more flexible image processing tasks.
A recent work \[1\] pointed out that inverting a generative model from the image space to some intermediate feature maps. More concretely, the generator \(G(\cdot)\) is divided into two sub-networks, \(G_1^\ell(\cdot)\) and \(G_2^\ell(\cdot)\). Here, \(\ell\) is the index of the intermediate layer to perform feature composition. With such a separation, for any \(z_n\), we can extract the corresponding spatial feature \(F_n^\ell = G_1^\ell(z_n)\) for further composition.

**Adaptive Channel Importance.** Recall that we would like each \(z_n\) to recover some particular regions of the target image. Bau et al. \[4\] observed that different units (i.e., channels) of the generator in GAN are responsible for generating different visual concepts such as objects and textures. Based on this observation, we introduce the adaptive channel importance \(\alpha_n\) for each \(z_n\) to help them align with different semantics. Here, \(\alpha_n \in \mathbb{R}^C\) is a \(C\)-dimensional vector and \(C\) is the number of channels in the \(\ell\)-th layer of \(G(\cdot)\). We expect each entry of \(\alpha_n\) to represent how important the corresponding channel of the feature map \(F_n^\ell\) is. With such composition, the reconstructed image can be generated with

\[
x^{inv} = G_2^\ell\left(\sum_{n=1}^{N} F_n^\ell \odot \alpha_n\right),
\]

where \(\odot\) denotes the channel-wise multiplication as

\[
\{F_n^\ell \odot \alpha_n\}_{i,j,c} = \{F_n^\ell\}_{i,j,c} \times \{\alpha_n\}_{c}.
\]

Here, \(i\) and \(j\) indicate the spatial location, while \(c\) stands for the channel index.

**Optimization Objective.** After introducing the feature composition technique together with the introduced adaptive channel importance to integrate multiple latent codes, there are \(2^N\) sets of parameters to be optimized in total. Accordingly we reformulate Eq.\((1)\) as

\[
\{z_n\}_{n=1}^{N}, \{\alpha_n\}_{n=1}^{N} = \arg\min_{\{z_n\}_{n=1}^{N}, \{\alpha_n\}_{n=1}^{N}} \mathcal{L}(x^{inv}, x).
\]
To improve the reconstruction quality, we define the objective function by leveraging both low-level and high-level information. In particular, we use pixel-wise reconstruction error as well as the $l_1$ distance between the perceptual features extracted from the two images\(^2\). Therefore, the objective function is as follows:

$$
L(x_1, x_2) = ||x_1 - x_2||^2 + ||\phi(x_1), \phi(x_2)||_1,
$$

(5)

where $\phi(\cdot)$ denotes the perceptual feature extractor. We use the gradient descent algorithm to find the optimal latent codes as well as the corresponding channel importance scores.

3.2. Multi-Code GAN Prior for Image Processing

After inversion, we apply the reconstruction result as multi-code GAN prior to a variety of image processing tasks. Each task requires an image as a reference, which is the input image for processing. For example, image colorization task deals with grayscale images and image inpainting task restores images with missing holes. Given an input, we apply the proposed multi-code GAN inversion method to reconstruct it and then post-process the reconstructed image to approximate the input. When the approximation is close enough to the input, we assume the reconstruction before post-processing is what we want. Here, to adapt mGANprior to a specific task, we modify Eq.(5) based on the post-processing function:

- For image colorization task, with a grayscale image $I_{gray}$ as the input, we expect the inversion result to have the same gray channel as $I_{gray}$ with

$$
L_{color} = L(\text{gray}(x^{inv}), I_{gray}),
$$

(6)

where $\text{gray}(\cdot)$ stands for the operation to take the gray channel of an image.

- For image super-resolution task, with a low-resolution image $I_{LR}$ as the input, we downsample the inversion result to approximate $I_{LR}$ with

$$
L_{SR} = L(\text{down}(x^{inv}), I_{LR}),
$$

(7)

where $\text{down}(\cdot)$ stands for the downsampling operation.

- For image inpainting task, with an intact image $I_{ori}$ and a binary mask $m$ indicating known pixels, we only reconstruct the incorrupt parts and let the GAN model fill in the missing pixels automatically with

$$
L_{inp} = L(x^{inv} \circ m, I_{ori} \circ m),
$$

(8)

where $\circ$ denotes the element-wise product.

\(^2\)In this experiment, we use pre-trained VGG-16 model \([37]\) as the feature extractor, and the output of layer conv$_{43}$ is used.

4. Experiments

We conduct extensive experiments on state-of-the-art GAN models, *i.e.*, PGGAN \([23]\) and StyleGAN \([24]\), to verify the effectiveness of mGANprior. These models are trained on various datasets, including CelebA-HQ \([23]\) and FFHQ \([24]\) for faces as well as LSUN \([46]\) for scenes.

4.1. Comparison with Other Inversion Methods

There are many attempts on GAN inversion in the literature. In this section, we compare our multi-code inversion approach with the following baseline methods: (a) optimizing a single latent code $z$ as in Eq.(1) \([32]\), (b) learning an encoder to reverse the generator \([52]\), (c) using the encoder as initialization for optimization \([5]\), and (d) our proposed mGANprior.
Table 1: Quantitative comparison of different GAN inversion methods: including (a) optimizing a single latent code [32], (b) learning an encoder [52], (c) using the encoder as initialization for optimization [5], and (d) our proposed mGANprior. ↑ means the higher the better while ↓ means the lower the better.

<table>
<thead>
<tr>
<th>Method</th>
<th>Bedroom PSNR ↑</th>
<th>Church PSNR ↑</th>
<th>Face PSNR ↑</th>
<th>Bedroom LPIPS ↓</th>
<th>Church LPIPS ↓</th>
<th>Face LPIPS ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>17.19 0.5897</td>
<td>17.15 0.5339</td>
<td>19.17 0.5797</td>
<td>17.15 0.5339</td>
<td>19.17 0.5797</td>
<td>17.15 0.5339</td>
</tr>
<tr>
<td>(b)</td>
<td>11.59 0.6247</td>
<td>11.58 0.5961</td>
<td>11.18 0.6992</td>
<td>11.58 0.5961</td>
<td>11.18 0.6992</td>
<td>11.58 0.5961</td>
</tr>
<tr>
<td>(c)</td>
<td>18.34 0.5201</td>
<td>17.81 0.4789</td>
<td>20.33 0.5321</td>
<td>17.81 0.4789</td>
<td>20.33 0.5321</td>
<td>17.81 0.4789</td>
</tr>
<tr>
<td>(d)</td>
<td>25.13 0.1578</td>
<td>22.76 0.1799</td>
<td>23.59 0.4432</td>
<td>22.76 0.1799</td>
<td>23.59 0.4432</td>
<td>22.76 0.1799</td>
</tr>
</tbody>
</table>

To quantitatively evaluate the inversion results, we introduce the Peak Signal-to-Noise Ratio (PSNR) to measure the similarity between the original input and the reconstruction result from pixel level, as well as the LPIPS metric [49] which is known to align with human perception. We make comparisons on three PGGAN [23] models that are trained on LSUN bedroom (indoor scene), LSUN church (outdoor scene), and CelebA-HQ (human face) respectively. For each model, we invert 300 real images for testing.

Tab.1 and Fig.3 show the quantitative and qualitative comparisons respectively. From Tab.1, we can tell that mGANprior beats other competitors on all three models from both pixel level (PSNR) and perception level (LPIPS). We also observe in Fig.3 that existing methods fail to recover the details of the target image, which is due to the limited representation capability of a single latent code. By contrast, our method achieves much more satisfying reconstructions with most details, benefitting from multiple latent codes. We even recover an eastern face with a model trained on western data (CelebA-HQ [23]).

4.2. Analysis on Inverted Codes

As described in Sec.3, our method achieves high-fidelity GAN inversion with \( N \) latent codes and \( N \) importance factors. Taking PGGAN as an example, if we choose the 6th layer (i.e., with 512 channels) as the composition layer with \( N = 10 \), the number of parameters to optimize is \( 10 \times (512 + 512) \), which is 20 times the dimension of the original latent space. In this section, we perform detailed analysis on the inverted codes.

Number of Codes. Obviously, there is a trade-off between the dimension of the optimization space and the inversion quality. To better analysis such trade-off, we evaluate our method by varying the number of latent codes to optimize. Fig.4 shows that the more latent codes used, the better reconstruction we are able to obtain. However, it does not imply that the performance can be infinitely improved by increasing the number of latent codes. From Fig.4, we can see that after the number reaches 20, there is no significant improvement via involving more latent codes.

Different Composition Layers. On which layer to perform feature composition also affects the performance of the proposed mGANprior. We thus compose the latent codes on various layers of PGGAN (i.e., from 1st to 8th) and compare the inversion quality, as shown in Fig.4. In general, a higher composition layer could lead to a better inversion effect. However, as revealed in [4], higher layers contain the information of local pixel patterns such as edges and colors rather than the high-level semantics. Composing features at higher layers is hard to reuse of the semantic knowledge learned by GANs. This will be discussed more in Sec.4.4.

Role of Each Latent Code. We employ multiple latent codes by expecting each of them to take charge of inverting a particular region and hence complement with each other. In this part, we visualize the roles that different latent codes play in the inversion process. As pointed out by [4], for a particular layer in a GAN model, different units (channels) control different semantic concepts. Recall that mGANprior uses adaptive channel importance to help determine what kind of semantics a particular \( z \) should focus on. Therefore, for each \( z_n \), we set the elements in \( \alpha_n \) that are larger than 0.2 as 0, getting \( \alpha'_n \). Then we compute the difference map between the reconstructions using \( \alpha_n \) and \( \alpha'_n \). With the help of a segmentation model [51], we can also get the segmentation maps for various visual concepts, such as tower and tree. We finally annotate each latent code based on the Intersection-over-Union (IoU) metric between the corresponding difference map and all candidate segmentation maps. Fig.5 shows the segmentation result.
and the IoU maps of some chosen latent codes. It turns out that the latent codes are specialized to invert different meaningful image regions to compose the whole image. This is also a huge advantage of using multiple latent codes over using a single code.

4.3. Image Processing Applications

With the high-fidelity image reconstruction, our multi-code inversion method facilitates many image processing tasks with pre-trained GANs as prior. In this section, we apply the proposed mGANprior to a variety of real-world applications to demonstrate its effectiveness, including image colorization, image super-resolution, image inpainting and denoising, as well as semantic manipulation and style mixing. For each application, the GAN model is fixed.

**Image Colorization.** Given a grayscale image as input, we can colorize it with mGANprior as described in Sec. 3.2. We compare our inversion method with optimizing the intermediate feature maps [3]. We also compare with DIP [39], which uses a discriminative model as prior, and Zhang et al. [48], which is specially designed for colorization task. We do experiments on PGGAN models trained for bedroom and church synthesis, and use the area under the curve of the cumulative error distribution over ab color space as the evaluation metric, following [48]. Tab.2 and Fig.6 show the quantitative and qualitative comparisons respectively. It turns out that using the discriminative model as prior fails to colorize the image adequately. That is because discriminative models focus on learning high-level representation which are not suitable for low-level tasks. On the contrary, using the generative model as prior leads to much more satisfying colorful images. We also achieve comparable results as the model whose primary goal is image colorization (Fig.6 (c) and (d)). This benefits from the rich knowledge learned by GANs. Note that Zhang et al. [48] is proposed for general image colorization, while our approach can be only applied to a certain image category corresponding to the given GAN model. A larger GAN model trained on a more diverse dataset should improve its generalization ability.

**Image Super-Resolution.** We also evaluate our approach on the image super-resolution (SR) task. We do experiments on the PGGAN model trained for face synthesis and set the SR factor as 16. Such a large factor is very challenging for the SR task. We compare with DIP [39] as well as the state-of-the-art SR methods, RCAN [50] and ESRGAN [42]. Besides PSNR and LPIPS, we introduce Naturalness Image Quality Evaluator (NIQE) [33] as an extra metric. Tab.3 shows the quantitative comparison. We can con-
Corrupted Image (a) Single Latent Code (b) Optimizing Feature Maps (c) DIP (d) Ours Ground Truth

Figure 8: Qualitative comparison of different inpainting methods, including (a) inversion by optimizing a single latent code [30, 32], (b) inversion by optimizing feature maps [3], (c) DIP [39], and (d) our mGANprior.

Target Image Inversion Expression Laugh Target Image Inversion Female Gender Male

Target Image Inversion Age Old Target Image Inversion Pose Right

Figure 9: Real face manipulation with respect to four various attributes. In each four-element tuple, from left to right are: input face, inversion result, and manipulation results by making a particular semantic more negative and more positive.

Table 3: Quantitative comparison of different super-resolution methods with SR factor 16. Competitors include DIP [39], RCAN [50], and ESRGAN [42]. ↑ means the higher the better while ↓ means the lower the better.

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR↑</th>
<th>LPIPS↓</th>
<th>NIQE↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) DIP [39]</td>
<td>26.87</td>
<td>0.4236</td>
<td>4.66</td>
</tr>
<tr>
<td>(b) RCAN [50]</td>
<td>28.82</td>
<td>0.4579</td>
<td>5.70</td>
</tr>
<tr>
<td>(c) ESRGAN [42]</td>
<td>25.26</td>
<td>0.3862</td>
<td>3.27</td>
</tr>
<tr>
<td>(d) Ours</td>
<td>26.93</td>
<td>0.3584</td>
<td>3.19</td>
</tr>
</tbody>
</table>

Table 4: Quantitative comparison of different inpainting methods. We do test with both centrally cropping a 64 × 64 box and randomly cropping 80% pixels. ↑ means higher score is better.

<table>
<thead>
<tr>
<th>Method</th>
<th>Center Crop</th>
<th>Random Crop</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Single latent code [30, 32]</td>
<td>10.37</td>
<td>0.1672</td>
</tr>
<tr>
<td>(b) Optimizing feature maps [3]</td>
<td>14.75</td>
<td>0.4563</td>
</tr>
<tr>
<td>(c) DIP [39]</td>
<td>17.92</td>
<td>0.4327</td>
</tr>
<tr>
<td>(d) Ours</td>
<td><strong>21.43</strong></td>
<td><strong>0.5320</strong></td>
</tr>
</tbody>
</table>

Include that our approach achieves comparable or even better performance than the advanced learning-based competitors. A visualization example is also shown in Fig.7, where our method reconstructs the human eye with more details. Compared to existing learning-based models, like RCAN and ESRGAN, our mGANprior is more flexible to the SR factor. This suggests that the freely-trained PGGAN model has spontaneously learned rich knowledge such that it can be used as prior to enhance a low-resolution (LR) image.

**Image Inpainting and Denoising.** We further extend our approach to image restoration tasks, like image inpainting and image denoising. We first corrupt the image contents by randomly cropping or adding noises, and then use different algorithms to restore them. Experiments are conducted on PGGAN models and we compare with several baseline inversion methods as well as DIP [39]. PSNR and Structural SIMilarity (SSIM) [43] are used as evaluation metrics.
shows the quantitative comparison, where our approach achieves the best performances on both settings of center crop and random crop. Fig. 8 includes some examples of restoring corrupted images. It is obvious that both existing inversion methods and DIP fail to adequately fill in the missing pixels or completely remove the added noises. By contrast, our method is able to use well-trained GANs as prior to convincingly repair the corrupted images with meaningful filled content.

Semantic Manipulation. Besides the aforementioned low-level applications, we also test our approach with some high-level tasks, like semantic manipulation and style mixing. As pointed out by prior work [21, 15, 35], GANs have already encoded some interpretable semantics inside the latent space. From this point, our inversion method provides a feasible way to utilize these learned semantics for real image manipulation. We apply the manipulation framework based on latent code proposed in [35] to achieve semantic facial attribute editing. Fig. 9 shows the manipulation results. We see that mGANprior can provide rich enough information for semantic manipulation.

4.4. Knowledge Representation in GANs

As discussed above, the major limitation of using single latent code is its limited expressiveness, especially when the test image presents domain gap to the training data. Here we verify whether using multiple codes can help alleviate this problem. In particular, we try to use GAN models trained for synthesizing face, church, conference room, and bedroom, to invert a bedroom image. As shown in Fig. 10, when using a single latent code, the reconstructed image still lies in the original training domain (e.g., the inversion with PGGAN CelebA-HQ model looks like a face instead of a bedroom). On the contrary, our approach is able to compose a bedroom image no matter what data the GAN generator is trained with.

We further analyze the layer-wise knowledge of a well-trained GAN model by performing feature composition at different layers. Fig. 10 suggests that the higher layer is used, the better the reconstruction will be. That is because reconstruction focuses on recovering low-level pixel values, and GANs tend to represent abstract semantics at bottom layers while represent content details at top layers. We also observe that the 4th layer is good enough for the bedroom model to invert a bedroom image, but the other three models need the 8th layer for satisfying inversion. The reason is that bedroom shares different semantics from face, church, and conference room, therefore the high-level knowledge (contained in bottom layers) from these models cannot be reused. We further make per-layer analysis by applying our approach to image colorization and image inpainting tasks, as shown in Fig. 11. The colorization task gets the best result at the 8th layer while the inpainting task at the 4th layer. That is because colorization is more like a low-level rendering task while inpainting requires the GAN prior to fill in the missing content with meaningful objects. This is consistent with the analysis from Fig. 10, which is that low-level knowledge from GAN prior can be reused at higher layers while high-level knowledge at lower layers.

5. Conclusion

We present mGANprior that employs multiple latent codes for reconstructing real images with a pre-trained GAN model. It enables these GAN models as powerful prior to a variety of image processing tasks.

Acknowledgement: This work is supported in part by the Early Career Scheme (ECS) through the Research Grants Council of Hong Kong under Grant No.24206219 and in part by SenseTime Collaborative Grant.
References


[38] Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Infrared image colorization based on a triplet dcgan architecture. In CVPR Workshop, 2017.


