High-Resolution GAN Inversion for Degraded Images in Large Diverse Datasets

Yanbo Wang[1*], **Chuming Lin**[2], **Donghao Luo**[2], **Ying Tai**[2], **Zhizhong Zhang**[1†], **Yuan Xie**[1]

[1]School of Computer Science and Technology, East China Normal University
[2]Tencent Youtu Lab
51205901021@stu.ecnu.edu.cn, {chuminglin, michaelluo, yingtai}@tencent.com, {zzzhang, yxie}@cs.ecnu.edu.cn

Abstract

The last decades are marked by massive and diverse image data, which shows increasingly high resolution and quality. However, some images we obtained may be corrupted, affecting the perception and the application of downstream tasks. A generic method for generating a high-quality image from the degraded one is in demand. In this paper, we present a novel GAN inversion framework that utilizes the powerful generative ability of StyleGAN-XL for this problem. To ease the inversion challenge with StyleGAN-XL, Clustering & Regularize Inversion (CRI) is proposed. Specifically, the latent space is firstly divided into finer-grained sub-spaces by clustering. Instead of initializing the inversion with the average latent vector, we approximate a centroid latent vector from the clusters, which generates an image close to the input image. Then, an offset with a regularization term is introduced to keep the inverted latent vector within a certain range. We validate our CRI scheme on multiple restoration tasks (*i.e.*, inpainting, colorization, and super-resolution) of complex natural images, and show preferable quantitative and qualitative results. We further demonstrate our technique is robust in terms of data and different GAN models. To our best knowledge, we are the first to adopt StyleGAN-XL for generating high-quality natural images from diverse degraded inputs. *Code is available at https://github.com/Booooooooooo/CRI.*

Introduction

With the mushroom development of imaging devices, images are now characterized by increasingly high resolution and diverse scenes. However, the quality of images on the Internet is uneven. For example, we may get images with low-resolution, images with missing parts, or gray-scale images. Therefore we hope to generate high-quality images directly from the degraded raw natural images.

One feasible approach is to leverage the generative ability of generative adversarial networks (GAN) and conduct GAN inversion. The main idea is to "invert" the corrupted image back to the latent space of the pre-trained GAN and then reconstruct a restored image from the optimal latent (Xia

et al. 2022). For a given degraded function $D(\cdot)$ and degraded image I_d, the latent vector w_{re} is iterated through back-propagation by minimizing the reconstruction loss of I_d and the degraded synthesis image $D(\phi_{Synthesis}(w_{re}))$, where $\phi_{Synthesis}$ is the synthesis layers of GAN. For example, PTI (Roich et al. 2021) inverts high-quality face images to the latent space and uses the result for image attributes manipulation. Nevertheless, PTI cannot be generalized to degraded input images from natural scenes. On the other hand, DGP (Pan et al. 2021) provides compelling results to generate missing semantics, *e.g.*, color, patch, resolution, for various images. But the scheme is limited to vanilla GAN models (*i.e.*, BigGAN (Brock, Donahue, and Simonyan 2019)) which only generate results with relatively low resolution (*e.g.*, 128^2, 256^2).

In this paper, we present an effective framework to generate high-resolution natural images from degraded inputs. We utilize the powerful capability of StyleGAN-XL (Sauer, Schwarz, and Geiger 2022), the state-of-the-art model for large-scale image synthesis, to generate high-quality images from degraded images. However, simply applying existing inversion methods (Pan et al. 2021; Roich et al. 2021; Menon et al. 2020) on StyleGAN-XL pre-trained on ImageNet (Krizhevsky, Sutskever, and Hinton 2012) obtains unsatisfactory results and faces two challenges:

(1) The latent space of StyleGAN-XL pre-trained on ImageNet is much larger and more complex compared to other popular GAN models like BigGAN and StyleGAN (Karras, Laine, and Aila 2019). Due to the diversity of data, the scale of the generator increases. For instance, for face images (*i.e.*, FFHQ (Karras, Laine, and Aila 2019)) StyleGAN2 (Karras et al. 2020) at resolution 1024^2 only uses 18 latent vectors, while StyleGAN-XL trained on ImageNet at resolution 512^2 needs 37 latent vectors since the data domain is more complicated. In the meantime, StyleGAN-XL employs an intermediate latent space compared to BigGAN which is the essence of its flexible editability but also adds to the inversion difficulty. *For such a complex latent space, finding a good initialization of latent vector is quite important.*

(2) The visual quality of the result is not explicitly imposed by the existing inversion methods. For example, images I_{re} and I_t generated by w_{re} and w_t in Figure 1(a) result in the same degraded images ($D(I_{re})$ and $D(I_t)$). Minimizing the traditional pixel or feature loss as existing meth-
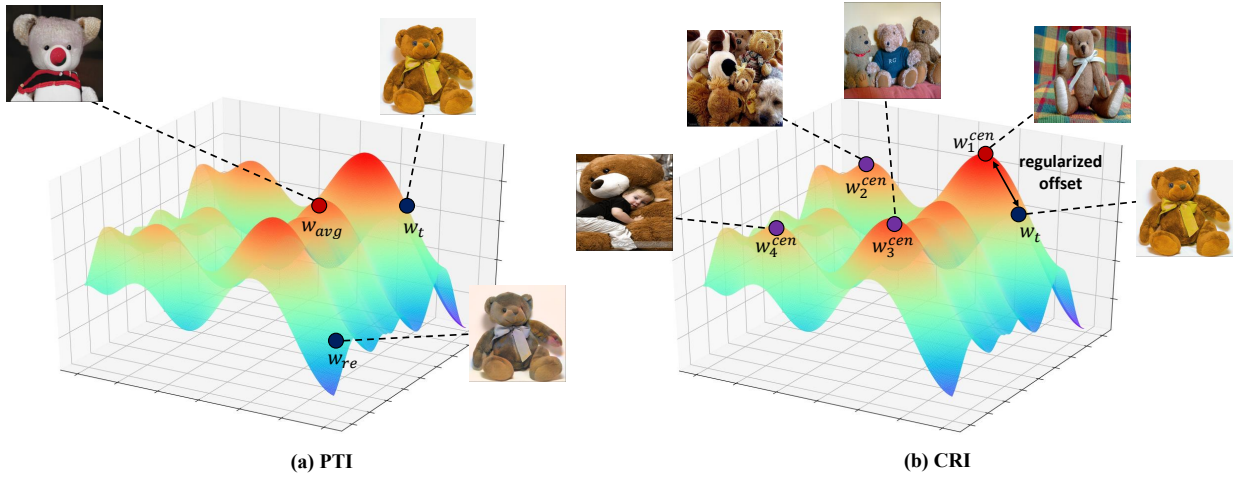

(a) PTI   **(b) CRI**

Figure 1: Comparison of existing GAN inversion methods and our proposed CRI in the latent space (*e.g.*, $\mathcal{W}$ and $\mathcal{W}+$ space). (a) Existing inversion methods (*e.g.*, PTI) begin the optimization from an average latent vector w_{avg} of the class, and minimize the reconstruction loss. Due to the ill-posedness of image restoration, the resulted image $\phi_{Synthesis}(w_{re})$ is similar to the target image $\phi_{Synthesis}(w_t)$ but is visually unpleasant as it falls into the distribution margin. (b) Our CRI starts the optimization by finding the "nearest" centroid as the starting latent vector. $w_1^{cen}, w_2^{cen}, w_3^{cen}, w_4^{cen}$ are candidate centroid and w_1^{cen} is the selected centroid. By bounding the scope of latent vectors with the regularized offset, the perceptual quality is guaranteed.

ods (Roich et al. 2021) can achieve low distortion (indicated by metrics such as LPIPS (Zhang et al. 2018)) but neglect the perception (indicated by metrics such as NIQE (Mittal, Soundararajan, and Bovik 2012) and FID (Heusel et al. 2017)). Consequently, some latent vectors may fall into the distribution margins of the $\mathcal{W}$ latent space of StyleGAN-XL (*i.e.*, blue regions in Figure 1), bringing poor and unpleasant results. *Hence, a constraint specially designed from the perception perspective is in need.*

To tackle the aforementioned challenges, we proposed Clustering & Regularize Inversion (CRI). We show diverse datasets such as ImageNet exhibit multi-modal characteristics, as shown in Figure 1(b). For instance, images of the teddy bear class in ImageNet may contain one to many bears (*i.e.*, $w_1^{cen}, w_2^{cen}, w_3^{cen}$) or even a person holding a teddy bear (*i.e.*, w_4^{cen}). Therefore, instead of starting the optimization from an average latent vector like most existing methods, we choose a "nearest" centroid by clustering. The large and complex latent space is first clustered into finer-grained sub-spaces. The latent vector w_k^{cen} (*i.e.*, centroid) which generates the "nearest" image is selected as the starting point for inversion. Meanwhile, inspired by the inherent distortion-perception trade-off in the latent space of Style-GAN (Tov et al. 2021), we improve the perceptual quality of results by constraining the scope of latent vectors. Specifically, an offset term w^{off} with a regularization term is introduced. During the optimization, we keep the w_k^{cen} frozen and iterate w^{off} instead of optimizing the whole latent vector $w_k^{cen} + w^{off}$. By doing so, we explicitly bound the latent vectors to the high perception areas in the latent space. In short, our main contributions can be summarized as:

- We fully explore the potential of GAN inversion for degraded natural images in diverse scenes. To our best

knowledge, we are the first to adopt StyleGAN-XL to generate high-quality images from degraded inputs.

- We propose a simple yet effective GAN inversion method to address the challenge of inversion with StyleGAN-XL. The key idea is to find a better starting latent vector and introduce a constraint on the image quality for optimization, summarized as Clustering & Regularize.

- We conduct extensive experiments on restoration tasks with different datasets and GAN models. The experimental results demonstrate the superiority of our method against other inversion methods.

Related Works

Style-based Generator

The Style-based generator is first proposed by (Karras, Laine, and Aila 2019). A mapping module is introduced to map the random values (denoted by $z \in \mathcal{Z}$) to an intermediate latent space named $\mathcal{W}$ space. By feeding latent vectors $w \in \mathcal{W}$ to each synthesis layer, StyleGAN can control the attributes of the generated image. It quickly surpassed previous image generative models (Brock, Donahue, and Simonyan 2019; Karras et al. 2017), showing superior perceptual quality and variety. StyleGAN2 (Karras et al. 2020) introduced path length regularization and weight demodulation, which further improves the image quality. Recently, StyleGAN3 (Karras et al. 2021) addressed the aliasing problem and proposed a new architecture that boosts the generator to be fully equivariant to translation and rotation.

GAN Inversion

GAN inversion aims at mapping a given image back into the latent space of a pre-trained GAN model, which can be

viewed as the inverse problem of image synthesis. It was first introduced by (Zhu et al. 2016). Generally, inversion methods can be divided into optimization-based (Creswell and Bharath 2018; Abdal, Qin, and Wonka 2020; Roich et al. 2021) and learning-based (Richardson et al. 2021; Dinh et al. 2022; Tov et al. 2021; Dong et al. 2021). The optimization-based approaches directly update the latent vector by minimizing the reconstruction loss. Learning-based methods employ an encoder to learn the mapping from the given image to its corresponding latent. Recently, (Dong et al. 2021) studies the invertibility of an arbitrary pre-trained DNN using learning-based inversion. However, despite the gain in inference speed, the reconstruction quality of learning-based methods is often worse than that of optimization-based approaches. Also, the encoder cannot be adapted to out-of-domain data. Our proposed CRI is based on optimization as we find that the relatively easy optimization-based approaches still do not work well for complex scenes.

Due to the many desirable properties of StyleGANs, recent works focus on conducting inversion with StyleGANs. For StyleGANs, inversion is usually conducted in the $\mathcal{W}$ space. It has been shown that the extended $\mathcal{W}+$ where different latent vectors are fed into each of the generator's layers is much more expressive and enables better image preservation (Abdal, Qin, and Wonka 2019). As shown in (Tov et al. 2021), there exist trade-offs between distortion, perceptual quality, and editability within the $\mathcal{W}+$ space. Recently, PTI (Roich et al. 2021) proposed pivotal tuning to mitigate the distortion-editability trade-off for out-of-distribution images. In comparison, we consider GAN inversion as a tool for exploiting the GAN priors for restoration tasks. Finding a "sweet spot" in the distortion-perception trade-off rather than the distortion-editability trade-off is the primary purpose and challenge of our work.

Deep Priors in Image Restoration

Image restoration is the task of recovering a clean image given its degraded version. Due to the powerful ability to generate photo-realistic images, many works exploit pre-trained GANs as priors to improve the quality (Geonung Kim 2022; Wei et al. 2022; Bartz et al. 2020). Leveraging a pre-trained StyleGAN is very common in face restoration (Menon et al. 2020; Yang et al. 2021). As one representative work, PULSE (Menon et al. 2020) propose to solve face image super-resolution task with inversion on Style-GAN2. As for more complex real-world images, attempts are still limited to vanilla GANs (*e.g.*, BigGAN). DGP (Pan et al. 2021) presents a relaxed formulation for mining the priors in BigGAN. The generator is jointly finetuned with the latent vectors. (Wu et al. 2021) 'retrieves' color information encapsulated in BigGAN via an encoder and then incorporates these features with feature modulations. In this paper, we explore the potential of StyleGAN-XL for degraded images from natural complex scenes, achieving better results than existing GAN inversion methods.

Method

Suppose the given input image I_d is obtained via $I_d = D(I)$ where I is the original high-quality image, $D(\cdot)$ is the degra-

dation function. Our aim is to generate high-quality images. By utilizing StyleGAN-XL, the problem can be translated as finding a latent vector w that satisfies:

$$w = \arg\min_w L(D(\phi_{Synthesis}(w)), I_d), \qquad (1)$$

where $\phi_{Synthesis}(\cdot)$ is the synthesis network of the pre-trained generator, $L(\cdot)$ is a distance metric in the image or feature space.

Although various attempt has been made (Menon et al. 2020; Pan et al. 2021) to seek generative priors, generating high-resolution natural images remains challenging. Since the latent space is much more complex for StyleGAN-XL pre-trained on ImageNet, it is essential to provide additional help for the optimization process. We proposed a Clustering & Regularize strategy to break through this problem. Clustering offers a better initial starting point for the optimization, while Regularize constrains the scope of w during optimization. In this section, we will first give details about this strategy and then describe the overall pipeline.

Start from a Centroid

One main challenge for restoring natural images with StyleGAN-XL lies in the complexity of the latent space. The latent vectors for StyleGAN-ImageNet present a multi-modal characteristic. Simply starting from an average latent vector becomes sub-optimal. Thus, we begin the optimization by finding a proper centroid as shown in Figure 1(b).

For a given class, we first randomly sample M latent vectors $\{w_j\}_{j=1}^M$ using the pre-trained mapping module $\phi_{Mapping}$ of StyleGAN-XL:

$$\{w_j\}_{j=1}^M = \phi_{Mapping}(\{z_j\}_{j=1}^M, c), \qquad (2)$$

where $\{z_j\}_{j=1}^M \sim \mathcal{N}(0,1)$ and c is a one-hot vector of the given class. $\{w_j\}_{j=1}^M$ are clustered into N clusters, obtaining N centroids $\{w_i^{cen}\}_{i=1}^N$ corresponding to N center images:

$$\{I_i^{cen}\}_{i=1}^N = \phi_{Synthesis}(\{w_i^{cen}\}_{i=1}^N). \qquad (3)$$

Given degraded input image I_d, we measure the distance between I_d and $\{I_i^{cen}\}_{i=1}^N$ in the feature space of a pre-trained model (e.g. VGG (Simonyan and Zisserman 2014)). The corresponding w_k^{cen} of the "nearest" center image is selected as the starting point for the subsequent optimization.

Regularized Offset

Due to the ill-posedness of image restoration, we may find multiple latent vectors that satisfy Eq. 1. However, some of the generated images may be beyond the manifold of natural images. Applying the traditional loss on I_d and $D(\phi_{synthesis}(w))$ does not explicitly guarantee the perception for inversion. We propose to improve the perceptual quality of the result images by constraining the scope of the latent vectors during optimization. Instead of optimizing w directly, we introduce an offset term w^{off} and keep the centroid w_k^{cen} frozen. The output image is defined as:

$$I_{syn} = \phi_{synthesis}(w_k^{cen} + w^{off}). \qquad (4)$$

Then we add a regularization term for w^{off}

$$reg = ||w^{off}||_2. \tag{5}$$

This regularization term limits the range of the output latent $w = w_k^{cen} + w^{off}$ to a certain extent so that the output image does not exceed the natural manifold. By doing so, the latent is allowed to be gradually rectified towards the target while remaining high perceptual quality.

Loss Function

The overall pipeline of CRI can be divided into two stages: the optimization stage and the finetune stage. As noted in DGP (Pan et al. 2021), a fixed generator is perhaps a crucial limitation for faithfully reconstructing unseen and complex images. However, we found that finetuning the generator together with the latent vector at the same time like DGP is sub-optimal for more sophisticated GANs. For StyleGANs with disentangled latent vectors, dividing optimization and finetune into two stages helps reduce the inversion difficulty.

In the first stage, only the w^{off} is optimized and the weights of the generator are frozen. The overall loss for the optimization stage can be defined as:

$$L_{op} = L_{LPIPS}(I_d, D(I_{syn})) + \lambda_1 L_2(I_d, D(I_{syn})) + \lambda_2||w^{off}||_2, \tag{6}$$

where $D(\cdot)$ is the degradation function, λ_1 and λ_2 are hyperparameters. As the noise input is removed in StyleGAN3, we do not optimize this term. However, the noise vector can be considered to improve the visual details.

We use the pivotal tuning technique (Roich et al. 2021) in the second stage, where the pivotal latent $w = w_c^{cen} + w^{off}$ is frozen. The generator is finetuned with the following loss:

$$L_{ft} = L_{LPIPS}(I_d, D(I_{syn})) + \lambda_{L2}L_2(I_d, D(I_{syn})) + \lambda_R L_R, \tag{7}$$

where λ_{L2} and λ_R are hyperparameters, L_R is the locality regularization term defined as:

$$L_R = L_{LPIPS}(x_r, x_r^*) + \lambda_{L2}^R L_2(x_r, x_r^*), \tag{8}$$

where λ_{L2}^R is a hyperparameter, $x_r = \phi_{synthesis}(w_r; \theta)$ is generated with the original weights of the generator and $x_r^* = \phi_{synthesis}(w_r; \theta^*)$ is generated using the currently tuned ones, w_r is the interpolated code between a random latent vector and the pivotal latent vector. Pseudocode of CRI is summarized in Algorithm 1.

Experiments

In this section, we present comprehensive experiments to evaluate our method. We experiment with image inpainting, image colorization and super-resolution. We start by comparing with current inversion methods on StyleGAN-XL (Sauer, Schwarz, and Geiger 2022) pre-trained on ImageNet (Krizhevsky, Sutskever, and Hinton 2012). We also compare the generating results of our method on StyleGAN-XL and DGP on BigGAN. For face images, we conduct experiments with StyleGAN2 (Karras et al. 2020) pre-trained on FFHQ (Karras, Laine, and Aila 2019), which shows the generalizability of our method. Then, we extend our method to out-of-domain images. Finally, we conduct an ablation

---

Algorithm 1 Pseudocode of CRI

```
# G: pre−trained GAN; Id: degraded input image
# ci: the class index of the input image
# loss_fn_w , loss_fn_g : Equation 6, 7

# Clustering centroids
z_samples = randn(M, G.z_dim)
c_samples = one_hot(M, ci , G.c_dim)
w_samples = G.mapping(z_samples, c_samples)
km = KMeans(cluster_n). fit (w_samples)
centroids = km. cluster_centers_

# Estimating a centroid
center_images = G. synthesis ( centroids )
center_features = VGG(center_images)
target_features = VGG(Id)
dis = L2( target_features , center_features )
w_st = centroids [ dis .argmin(0)]

# Optimizing regularized offset
w_offset = zeros (w_st .shape)
w_offset . requires_grad_ (True)
w_st . requires_grad_ (False )
image = G. synthesis (w_st+w_offset )
loss = loss_fn_w (degrade(image), Id)
loss .backward()

# Finetune Generator
w_offset . requires_grad_ (False )
G. requires_grad_ (True)
image = G. synthesis (w_st+w_offset )
loss = loss_fn_g (degrade(image), Id)
loss .backward()
```

---

study to justify the effectiveness of our inversion method. All the methods are inverted to $\mathcal{W}+$ space except PTI-colorization as we found $\mathcal{W}$ space generates better results.

Restoration on ImageNet

We first compare our method with other GAN inversion methods (Menon et al. 2020; Pan et al. 2021; Roich et al. 2021) on ImageNet. We use StyleGAN-XL (Sauer, Schwarz, and Geiger 2022) as it is state-of-the-art on large-scale image synthesis. The resolution of the generated images is 512^2. We invert 1000 images from the validation set of ImageNet, each from different classes, to quantitatively evaluate the methods. The scale factor of the SR task is set to 4.

Qualitative Results Figure 2 presents a qualitative comparison against other GAN-inversion methods (Menon et al. 2020; Pan et al. 2021; Roich et al. 2021) on ImageNet dataset. Our method achieves superior results for all tasks, i.e. image inpainting, image colorization and image super-resolution. Due to the lack of finetuning stage, PULSE fails to generalize to complex real-world images. It can only produce the general shape, but could not fill in the missing information. At the same time, since the latent vector and the generator are simultaneously optimized and finetuned, it is
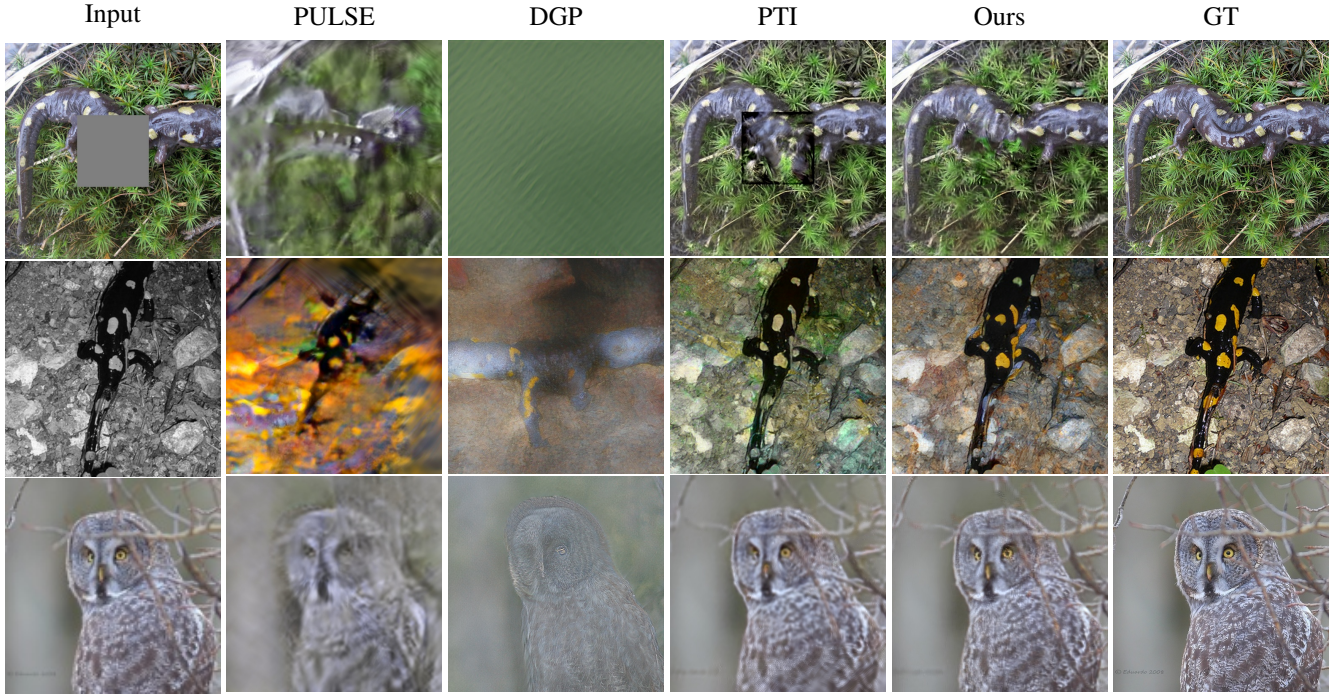
Figure 2: Image Restoration with images from the ImageNet valid set. StyleGAN-XL pre-trained on ImageNet is used. Our method achieves more natural transitions, more realistic colors and sharper details. Zoom in for better visualization.

Task	Measure	DGP	PTI	Ours
Inpainting	LPIPS ↓	0.2612	0.1219	**0.1164**
	FID ↓	166.24	54.62	**44.73**
Colorization	LPIPS ↓	0.3105	0.1994	**0.1914**
	FID ↓	78.32	47.51	**45.90**
SR	LPIPS ↓	0.2726	0.2429	**0.2390**
	NIQE ↓	8.4044	5.5659	**5.3065**

Table 1: Quantitative comparison against DGP and PTI on image inpainting, image colorization and image super-resolution. We use StyleGAN-XL pre-trained on ImageNet, and invert images from the ImageNet valid set.

Task	Measure	DGP	Ours
Inpainting	LPIPS ↓	0.1316	**0.1211**
	FID ↓	116.12	**80.20**
Colorization	LPIPS ↓	0.2256	**0.1842**
	FID ↓	189.42	**92.48**
SR	LPIPS ↓	0.2226	**0.2224**
	NIQE ↓	9.2102	**5.5686**

Table 2: Quantitative comparison between DGP with Big-GAN pre-trained on ImageNet and CRI with StyleGAN-XL pre-trained on ImageNet.

difficult for DGP to find a faithful output. Finetuning with incorrect latent vectors disrupts the generative capacity of the pre-trained generator which result in poor visual quality. Compared to PTI, our method is capable of recovering realistic fine details (*e.g.* smoother contours for inpainting, truer colors for colorization and sharper details for SR).

Quantitative Results   Following (Wu et al. 2021; Zeng et al. 2021), we use LPIPS (Zhang et al. 2018), FID (Heusel et al. 2017) for image inpainting and image colorization. Following (Pan et al. 2021), LPIPS and NIQE (Mittal, Soundararajan, and Bovik 2012) are used for SR. Also, we find these metrics are close to visual perception. As shown in Table 1, the quantitative results align with our qualitative results as we achieve the best score for all metrics.

Comparison with DGP-BigGAN   We compare the results of DGP with BigGAN-ImageNet and the results of our methods. Quantitative and Qualitative results are shown in Table 2 and Figure 3. Since BigGAN only supports the resolution of 256^2, we resize the output of DGP+BigGAN to 512^2 for comparison. For SR, the input of both DGP and CRI are set to 128^2. Our CRI+StyleGAN-XL outperforms DGP+BigGAN on image inpainting, colorization and SR.

Restoration on Face

To illustrate our method is not limited by the GAN model or data, we conduct experiments on StyleGAN2 (Karras et al. 2020) pre-trained on FFHQ (Karras, Laine, and Aila 2019). The resolution of the generated images is 1024^2. We invert 500 images from CelebA-HQ (Karras et al. 2017). The scale factor of SR is set to 16.
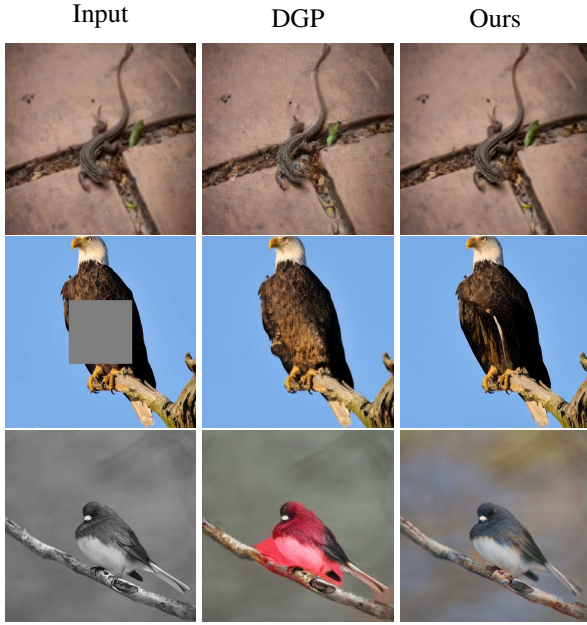
Figure 3: Qualitative comparison between DGP with Big-GAN pre-trained on ImageNet and CRI with StyleGAN-XL pre-trained on ImageNet. Zoom in for better visualization.

(a) BSD100   (b) DIV2K

Figure 4: Evaluation of CRI on out-of-domain datasets, including (a) image from BSD100, (b) image from DIV2K.

Quantitative Results For face restoration, we employ LPIPS and FID as evaluation metrics shown in Table 3. The FID score is calculated between the inverted 500 images and the CelebA-HQ dataset (30000 images in total). Our method obtains the lowest LPIPS and FID for all tasks, indicating that our results are perceptually close to the ground truth and have a close distance to the real face distribution.

Qualitative Results Qualitative results of each task are shown in Figure 6. Our method outperforms other methods in terms of image quality and identity preservation. PULSE and DGP can produce results of good visual quality, but the similarity is relatively low. For instance, DGP-SR fails to preserve the hairstyle. Images generated by PTI are similar to the input in terms of known information. For example, the background of the image in the inpainting task and the structure in the colorization task are quite authentic. However, the overall view of the generated images is not natural since the inverted latent vectors beyond the manifold of natural images. Compared to PTI, our results are more natural and generate sharper details.

Task	Measure	DGP	PULSE	PTI	Ours
Inpainting	LPIPS ↓	0.1311	0.1387	*0.1293*	**0.1279**
	FID ↓	68.32	59.20	53.48	**48.00**
Colorization	LPIPS ↓	0.1621	0.1924	0.1990	**0.1505**
	FID ↓	74.24	116.06	56.01	**46.54**
SR	LPIPS ↓	0.1481	0.1385	*0.1317*	**0.1297**
	FID ↓	68.56	60.01	60.76	**58.88**

Table 3: Quantitative comparison against PULSE, DGP and PTI on image inpainting, image colorization and image super-resolution. We use StyleGANv2 pre-trained on FFHQ, and invert images from the CelebA-HQ.
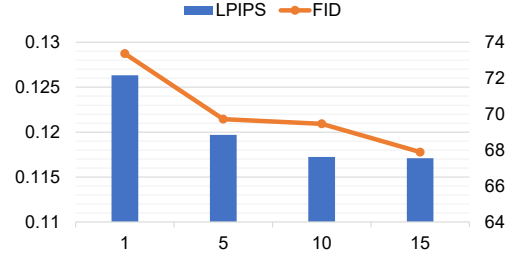

Figure 5: Comparison of numbers of the cluster.

Robustness on Out-of-domain Images

We also experiments with images from BSD100 (Martin et al. 2001) dataset and DIV2K (Agustsson and Timofte 2017) dataset. Deit-M (Touvron et al. 2021) is used as a classifier to provide the class information (*i.e.*, c in Eq. 2) for the input. As shown in Figure 4, our CRI can reconstruct the missing part of out-of-domain images.

Ablation Study

Effect of Centroids We show the effectiveness of the clustering centroids by adopting different cluster numbers. To exclude the effect of the random sampling of $\{z_j\}_{j=1}^M$, we cluster the same $\{z_j\}_{j=1}^M$ into $N = \{1, 5, 10, 15\}$ clusters respectively using KMeans. The results of colorization are presented in Figure 5. The performance becomes progressively better as the number of centroids increases. However, the clustering time also becomes longer. Considering the performance and the clustering time trade-off, we choose $N = 10$ as the default.

Effect of Regularized Offset To validate the effectiveness of our regularized offset, we conduct an ablation study on face colorization with StyleGAN2. 60 images from CelebA-HQ are used. We first remove the regularization term and then replace the L_2 regularization with L_1. As shown in Table 4, using L_2 regularized offset is beneficial for both LPIPS and the ID Similarity compared to L_1 regularization. Optimizing without a regularization term (w/o Reg) fails to generate authentic colors (see Figure 7) but achieves high ID similarity, which also demonstrates the inherent trade-off between distortion and perception. In the meanwhile, our
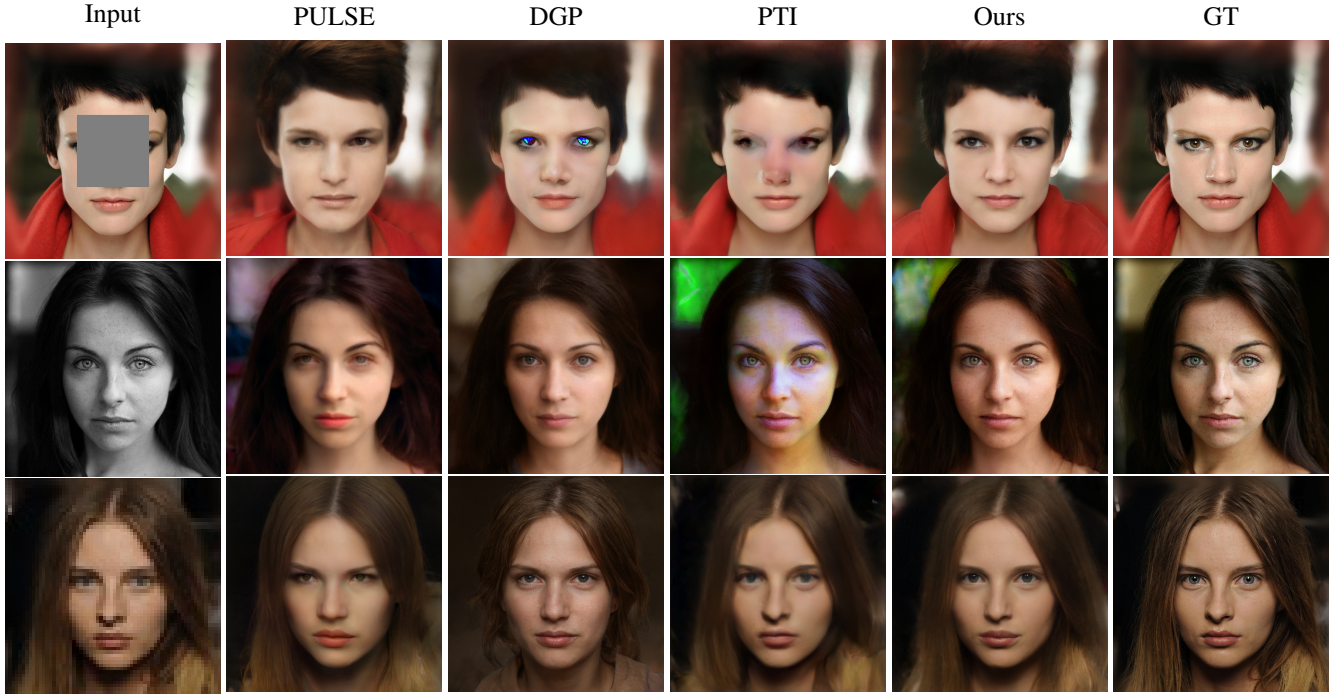
Figure 6: Image Restoration with images from the CelebA-HQ dataset. StyleGAN2 pre-trained on FFHQ is used for all methods. Our method guaranteed similarity while generating high quality face images. Zoom in for better visualization.
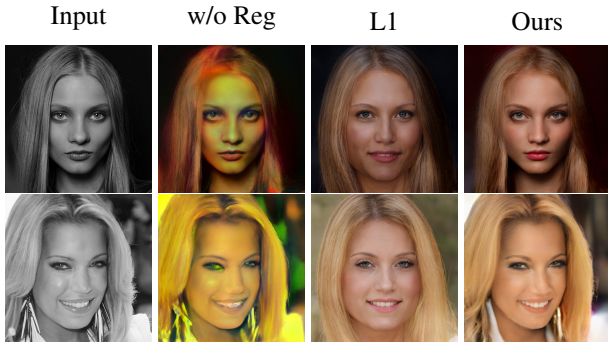

Figure 7: Qualitative results of different regularization strategy. w/o Reg: no regularization term, L1: regularize with L1 norm, Ours: regularize with L2 norm.

proposed CRI can achieve a "sweet spot" in the distortion-perception trade-off by introducing the regularized offset.

Implementation Details

For StyleGAN-XL with ImageNet, we use 1000 iterations for the optimization stage and use 350 iterations for the fine-tune stage as in (Sauer, Schwarz, and Geiger 2022). For StyleGAN2 with FFHQ, we use 500 iterations for the first stage and 20 iterations for the second stage as inversion with a face is simpler. For more detailed implementations, please refer to the supplementary materials.

Metric	w/o Reg	L_1	Ours
LPIPS↓	0.1996	0.1694	**0.1560**
ID Similarity↑	**0.7251**	0.1955	0.5756

Table 4: Comparison of different regularization strategy. w/o Reg: no regularization term, L1: regularize with L1 norm, Ours: regularize with L2 norm.

Conclusion

We have presented a general framework for generating high-resolution images from degraded images in diverse natural scenes. This is achieved by leveraging the generative power of StyleGAN-XL. Furthermore, we propose a simple yet effective technique, term as Clustering & Regularize, to ease the inversion difficulty of StyleGAN-XL pre-trained on massive natural images. Clustering is introduced to solve the difficulty caused by the large and complex latent space of StyleGAN-XL. It divides the latent space into sub-spaces and provides the inversion with a better initialization. While Regularize is derived from the perceptual aspect. By introducing an offset term and constraining it with regularization we can bind inversion for better visual quality. CRI allows us to achieve good perception and can effectively provide the results with rich image semantics. Extensive experiments on image restoration tasks illustrate the effectiveness of CRI. By designing the degradation function $D(\cdot)$, we believe our framework can be applied to other degradations, *e.g.*, noise and blur, which will be left for future work.

Acknowledgments

References

Abdal, R.; Qin, Y.; and Wonka, P. 2019. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 4432–4441.

Abdal, R.; Qin, Y.; and Wonka, P. 2020. Image2stylegan++: How to edit the embedded images? In *CVPR*, 8296–8305.

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 126–135.

Bartz, C.; Bethge, J.; Yang, H.; and Meinel, C. 2020. One Model to Reconstruct Them All: A Novel Way to Use the Stochastic Noise in StyleGAN. arXiv:2010.11113.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large scale GAN training for high fidelity natural image synthesis. *ICLR*.

Creswell, A.; and Bharath, A. A. 2018. Inverting the generator of a generative adversarial network. *TNNLS*, 30(7): 1967–1974.

Dinh, T. M.; Tran, A. T.; Nguyen, R.; and Hua, B.-S. 2022. Hyperinverter: Improving stylegan inversion via hypernetwork. In *CVPR*, 11389–11398.

Dong, X.; Yin, H.; Alvarez, J. M.; Kautz, J.; Molchanov, P.; and Kung, H. T. 2021. Privacy Vulnerability of Split Computing to Data-Free Model Inversion Attacks.

Geonung Kim, S. K. H. L. S. K. J. K. S.-H. B. S. C., Kyoungkook Kang. 2022. BigColor: Colorization using a Generative Color Prior for Natural Images. In *European Conference on Computer Vision (ECCV)*.

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. *NIPS*, 30.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-free generative adversarial networks. *NIPS*, 34: 852–863.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *CVPR*, 8110–8119.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *NIPS*, 25.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*. IEEE.

Menon, S.; Damian, A.; Hu, S.; Ravi, N.; and Rudin, C. 2020. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2437–2445.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3): 209–212.

Pan, X.; Zhan, X.; Dai, B.; Lin, D.; Loy, C. C.; and Luo, P. 2021. Exploiting deep generative prior for versatile image restoration and manipulation. *TPAMI*.

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 *CVPR*, 2287–2296.

Roich, D.; Mokady, R.; Bermano, A. H.; and Cohen-Or, D. 2021. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*.

Sauer, A.; Schwarz, K.; and Geiger, A. 2022. Stylegan-xl: Scaling stylegan to large diverse datasets. In *SIGGRAPH*, 1–10.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*, 10347–10357. PMLR.

Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; and Cohen-Or, D. 2021. Designing an encoder for stylegan image manipulation. *TOG*, 40(4): 1–14.

Wei, T.; Chen, D.; Zhou, W.; Liao, J.; Zhang, W.; Yuan, L.; Hua, G.; and Yu, N. 2022. E2Style: Improve the Efficiency and Effectiveness of StyleGAN Inversion. *IEEE Transactions on Image Processing*.

Wu, Y.; Wang, X.; Li, Y.; Zhang, H.; Zhao, X.; and Shan, Y. 2021. Towards vivid and diverse image colorization with generative color prior. In *ICCV*, 14377–14386.

Xia, W.; Zhang, Y.; Yang, Y.; Xue, J.-H.; Zhou, B.; and Yang, M.-H. 2022. Gan inversion: A survey. *TPAMI*.

Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2021. Gan prior embedded network for blind face restoration in the wild. In *CVPR*, 672–681.

Zeng, Y.; Lin, Z.; Lu, H.; and Patel, V. M. 2021. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In *ICCV*, 14164–14173.

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*, 586–595.

Zhu, J.-Y.; Krähenbühl, P.; Shechtman, E.; and Efros, A. A. 2016. Generative visual manipulation on the natural image manifold. In *ECCV*, 597–613. Springer.