CODE: Confident Ordinary Differential Editing

Bastien van Delft, Tommaso Martorella, Alexandre Alahi

École Polytechnique Fédérale de Lausanne (EPFL)
firstname.lastname@epfl.ch

Abstract

Conditioning image generation facilitates seamless editing and the creation of photorealistic images. However, conditioning on noisy or Out-of-Distribution (OoD) images poses significant challenges, particularly in balancing fidelity to the input and realism of the output. We introduce Confident Ordinary Differential Editing (CODE), a novel approach for image synthesis that effectively handles OoD guidance images. Utilizing a diffusion model as a generative prior, CODE enhances images through score-based updates along the probability-flow Ordinary Differential Equation (ODE) trajectory. This method requires no task-specific training, no handcrafted modules, and no assumptions regarding the corruptions affecting the conditioning image. Our method is compatible with any diffusion model. Positioned at the intersection of conditional image generation and blind image restoration, CODE operates in a fully blind manner, relying solely on a pre-trained generative model. Our method introduces an alternative approach to blind restoration: instead of targeting a specific ground truth image based on assumptions about the underlying corruption, CODE aims to increase the likelihood of the input image while maintaining fidelity. This results in the most probable in-distribution image around the input. Our contributions are twofold. First, CODE introduces a novel editing method based on ODE, providing enhanced control, realism, and fidelity compared to its SDE-based counterpart. Second, we introduce a confidence interval-based clipping method, which improves CODE's effectiveness by allowing it to disregard certain pixels or information, thus enhancing the restoration process in a blind manner. Experimental results demonstrate CODE's effectiveness over existing methods, particularly in scenarios involving severe degradation or OoD inputs.

Website — https://vita-epfl.github.io/CODE/
Code — https://github.com/vita-epfl/CODE/
Main and Appendix — https://arxiv.org/pdf/2408.12418

1 Introduction

Conditional image generation consists of guiding the creation of content using different sorts of conditioning, such as text, images, or segmentation maps. Our research focuses on scenarios where the guidance is an Out-of-Distribution (OoD) image relative to the training data distribution. This is especially relevant for handling corrupted images, similar to denoising or restoration methods. The main challenge in these scenarios is balancing fidelity to the input with realism in the generated images. Traditional methods for restoring corrupted images, such as Image-to-Image Translation or Style Transfer, are limited by the need for distinct datasets per style or per noise. Another approach models the corruption function as an inverse problem, requiring detailed knowledge of each possible corruption, making it impractical for most real unknown OoD scenarios. Guided image synthesis for OoD inputs aims to rectify corrupted images without prior knowledge of the corruption, positioning it as Blind Image Restoration (BIR). Despite recent advancements, achieving human-level generalization remains challenging.

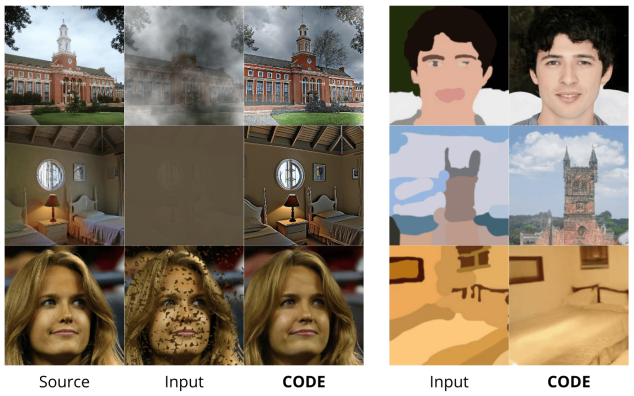


Figure 1: CODE: a conditional image generation framework for robust Out-of-Distribution image guidance.

Our work aims to generate realistic and plausible images from potentially corrupted inputs using only a pre-trained generative model, without additional data augmentation or finetuning on corrupted data, ***and without any specific assumption about the corruptions.*** Unlike other BIR methods that strive to reconstruct a ground-truth image relying on specific guidance or human-based assumptions, our approach is fully blind, seeking to maximize the input image's likelihood while minimizing modifications to the input image. As such, we differ from traditional BIR approaches.

BIR is inherently ill-posed due to the loss of information

from unknown degradation, necessitating auxiliary information to enhance restoration quality. Previous approaches have incorporated domain-specific priors such as facial heatmaps or landmarks (Chen et al. 2018, 2021; Yu et al. 2018), but these degrade with increased degradation and lack versatility. Generative priors from pre-trained models like GANs (Chan et al. 2021; Zhou et al. 2022; Yang et al. 2021; Pan et al. 2021; Menon et al. 2020) become unstable with severe degradation, leading to unrealistic reconstructions. Methods like (Wang et al. 2021a) combine facial priors with generative priors to improve fidelity but fail under extreme degradation. In (Meng et al. 2021), the authors replace GANs with diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2020) as generative priors. However, as the degradation increases, the method forces a choice between realism and fidelity.

BIR still fails to achieve faithful and realistic reconstruction for a wide range of corruptions on a wide range of images. Dealing with various unknown corruptions prevents inverse methods from being easily applicable while dealing with a wide range of images prevents relying on carefully designed domain priors. We introduce Confident Ordinary Differential Editing (CODE), an unsupervised method that generates faithful and realistic image reconstructions from a single degraded image without information on the degradation type, even under severe conditions. CODE leverages the generative prior of a pre-trained diffusion model without requiring additional training or finetuning. Consequently, it is compatible with any pre-trained Diffusion Model and any dataset. CODE optimizes the likelihood of the generated image while constraining the distance to the input image, framing restoration as an optimization problem. Similar to GAN-inversion methods (Tov et al. 2021; Abdal, Qin, and Wonka 2020, 2019; Zhu et al. 2020; Menon et al. 2020; Pan et al. 2021), CODE inverts the observation to a latent space before optimization but similar to SDEdit (Meng et al. 2021) we propose to replace GANs with diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2020) as generative priors. Unlike GAN inversion, which relies on an auxiliary trained encoder, diffusion model inversion uses differential equations. In SDEdit, random noise is injected into the degraded observation to partially invert it in order to subsequently revert the process using a stochastic differential equation (SDE). As more noise is injected, a higher degree of realism is ensured, but at the expense of fidelity due to the additional loss of information caused by the noise randomness and the non-deterministic sampling from the SDE. We found that in some cases, as the degree of degradation increases, the method requires too high a degree of noise injection to work, forcing a choice between realism or fidelity. CODE refines SDEdit (Meng et al. 2021) by leveraging the probability-flow Ordinary Differential Equation (ODE) (Song et al. 2020), ensuring bijective correspondence with latent spaces. We use Langevin dynamics with score-based updates for correction, followed by the probability-flow ODE to project the adjusted latent representation back into the image space. This decouples noise injection levels, correction levels, and latent spaces, enhancing control over the editing process. Furthermore, CODE introduces a confidence-based clipping method that relies on the marginal distribution of each latent space. This method allows for the disregard of certain image information based on probability, which synergizes with our editing method. Our experimental results show CODE's superiority over SDEdit in realism and fidelity, especially in challenging scenarios.

2 Background

Related Works

A detailed comparison of the requirements of state-of-the-art methods is provided in the Appendix A.

Inverse Problems In the inverse problem setup, methods are designed to leverage sensible *assumptions on the degradation operators*. When combined with powerful generative models such as diffusion models, these approaches have achieved outstanding results, setting new benchmarks in the field (Saharia et al. 2022; Liang et al. 2021; Kawar et al. 2022; Murata et al. 2023; Zhu et al. 2023; Chung et al. 2023; Wang, Yu, and Zhang 2022). Several subcategories of the inverse problem setting, like blind linear problems and non-blind non-linear problems, drop some assumptions about the degradation operators and, therefore, extend their applicability. However, while producing exceptional results in controlled applications like deblurring and super-resolution, the necessity for assumptions on the degradation operator makes them often impractical for unknown corruptions or in real-world scenarios. Consequently, these methods are not directly applicable to our context, where such exact information is typically unavailable. DDRM (Kawar et al. 2022), DDNM (Wang, Yu, and Zhang 2022), GibbsDDRM (Murata et al. 2023), DPS (Chung et al. 2023), and DiffPIR (Zhu et al. 2023) belong to this category.

Conditional Generative Models with Paired Data A parallel approach involves conditioning a generative model on a degraded image. Most methods in this category *require training with pairs of degraded and natural images* (Mirza and Osindero 2014; Isola et al. 2017; Batzolis et al. 2021; Xia et al. 2023; Li et al. 2023; Liu et al. 2023; Chung, Kim, and Ye 2023). Additionally, these methods often depend on carefully designed loss functions and guidance mechanisms to enhance performance, as demonstrated by (Song et al. 2020). Conditional Generative Adversarial Networks, as explored in (Isola et al. 2017), exemplify this approach, where generative models are trained to regenerate the original sample when conditioned on its version in another domain. However, when the degradation process is unknown or varies widely, these models struggle to generalize effectively, rendering them less applicable and not comparable to our method, which operates without such constraints. DDB (Chung, Kim, and Ye 2023) and I^2SB (Liu et al. 2023) fall into this category.

Unsupervised Bridge Problem with Unpaired Datasets In scenarios where two distinct datasets of *clean and degraded data are available without direct paired data*, methodologies based on principles like cycle consistency and realism have been developed, as evidenced by the works of (Zhu et al. 2017) using GANs (Goodfellow et al. 2014) and (Su et al. 2023) using Diffusion Models (Ho, Jain, and Abbeel 2020),(Sohl-Dickstein et al. 2015). A direct application of such methods to our scenario is not feasible due to the need

for datasets of degraded images, which would hamper the ability to generalize to unseen corruptions.

Blind Image Restoration with task-specific or domain-specific information Blind Image Restoration methods aim to handle a variety of degradations without restricting themselves to specific types. A recent trend in this field is the transposition of the problem into a latent space where corrections are made based on prior distributions. Notable works in this area include (Abdal, Qin, and Wonka 2019, 2020; Chan et al. 2021; Zhu et al. 2020; Poirier-Ginter and Lalonde 2023) have explored various aspects of this approach, utilizing GAN inversion, or VAE/VQVAE encoding (Kingma and Welling 2013; Oord, Vinyals, and Kavukcuoglu 2017), and have achieved significant advancements, particularly in scenarios involving light but diverse degradations. Usually, methods for blind image restoration *incorporate domain-specific information* (Zhou et al. 2022; Wang et al. 2021a; Gu et al. 2022) or *task-specific guidances* (Fei et al. 2023).

Moreover, several methods (Lin et al. 2023; Yang et al. 2023) rely on the *combination of many blocks trained separately* (such as Real-ESRGAN (Wang et al. 2021b)) and *incorporate different task-specific information* (e.g., different restorers), making it even harder to ensure resilience to diverse degradations.

SDEdit and ILVR The works by Meng *et al.* in (Meng et al. 2021) and Choi *et al.* in (Choi et al. 2021) inspired the formulation of CODE. SDEdit relies on the stochastic exploration of a given input's neighborhood to yield realistic and faithful outputs within a limited editing range. This method, tested for robustness by (Gao et al. 2022), bears similarities to the proposed gradient updates within the latent space of GAN models but is grounded in more solid theoretical foundations. ILVR is an iterative conditioning method designed to generate diverse images that share semantics with a downsampled guidance image. However, it requires a clean image for downsampling, which is not feasible in our scenario where the input guidance is already corrupted. Downsampling in this context would exacerbate information loss, making ILVR unsuitable for our application.

Preliminary - Diffusion Models

Denoising Diffusion Probabilistic Models We denote $\mathbf{x}_0$ the data from the data distribution, in our case natural images, and $\mathbf{x}_1$, ..., $\mathbf{x}_T$ the latent variables. The forward process is in DDPM (Ho, Jain, and Abbeel 2020) then defined by:

$$\mathbf{x}_{t+1} = \alpha_t \cdot \mathbf{x}_t + (1 - \alpha_t) \cdot \epsilon, \text{with } \epsilon \sim \mathcal{N}(0, \mathbf{I}),$$

Where α_t is a schedule predefined as an hyperparameter. The diffusion model ϵ_θ is then trained to minimize $\mathbb{E}_{\mathbf{x}_t, t} \left\| \epsilon_\theta(\mathbf{x}_t, t) - \epsilon \right\|$.

Score-based Generative Models In the case of Score-based Generative Models (Song and Ermon 2019; Song et al. 2020; Song and Ermon 2020), the model s_θ learns to approximate the score function, $\nabla_{\mathbf{x}} \log p(\mathbf{x})$, by minimizing:

$$\mathbb{E}_{p(\mathbf{x})} \left\| s_\theta(\mathbf{x}) - \nabla_{\mathbf{x}} \log p(\mathbf{x}) \right\|.$$

The most common approach to solve this is denoising score matching (Vincent 2011), which is further described in the Appendix C.

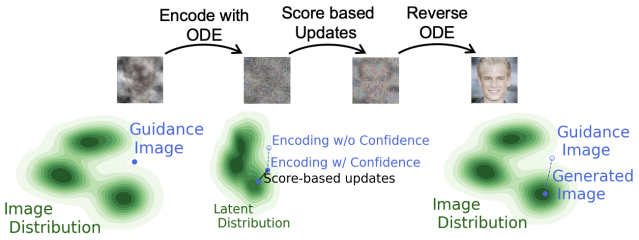


Figure 2: Editing corrupted images with ODE. The green contour plot represents the distribution of images. Given a corrupted image, we encode it into a latent space using the probability-flow ODE and our Confidence-Based Clipping. We use Langevin Dynamics in the latent space to correct the encoded image. Finally, we project the updated latent back into the visual domain.

Crucially, one can sample from $p(x_t)$ while using only the score function through Langevin dynamics (Langevin 1908) sampling by repeating the following update step:

$$x_{t+1} = x_t + \epsilon \cdot s_\theta(x_t, t) + \sqrt{2\epsilon} \cdot \eta \text{ , with } \eta \sim \mathcal{N}(0, \sigma^2). \quad (1)$$

3 Method: Confident Ordinary Differential Editing

Editing with Ordinary Differential Equations

Our approach, described in Figure 2, formulates a theoretically grounded method for mapping OoD samples to in-distribution ones.

From Gaussian Perturbation to Ordinary Inversion Our method draws inspiration from SDEdit (Meng et al. 2021) but introduces significant enhancements. SDEdit inverts the diffusion process by injecting Gaussian noise into the input image and then using this noisy image as a starting point to generate an image using DDPM (Ho, Jain, and Abbeel 2020). This process involves a trade-off between fidelity and realism: more noise results in more realistic images but less fidelity to the original input.

In contrast, we propose inverting the Probability-Flow Ordinary Differential Equation (ODE) as a superior alternative to noise injection. This approach maintains the fidelity of the reconstructed image by avoiding extra noise. The inversion process and its reverse operation ensure precise image reconstruction, limited only by approximation errors (Su et al. 2023). Unlike SDEdit, which requires increasing noise levels to revert to deeper latent spaces, our method allows inversion to any latent space along the ODE trajectory while preserving image integrity. This decouples the noise injection level from the depth of inversion. We use the ODE solver from DDIMs (Song, Meng, and Ermon 2020) in our experiments.

The primary motivation for inverting the degraded image is the model's ability to process out-of-distribution images. Direct estimation of the score on degraded images is impractical due to the poor performance of the score estimation on OoD data. By mapping the corrupted input back to the latent

Algorithm 1: CODE Simple - Confidence-based Clipping

Require: N *(Langevin iterations)*, ϵ *(step-size)*, x_0 *(Observation)*, L *(L-th latent-space)*, η *(size of the confidence interval.)*

$x_{L,0} = ODE\_SOLVER_{forward_{0\to L}}(Clip_{CBC_\eta}(x_0))$

for $k = 0$ **to** $N-1$ **do**

$x_{L,k+1} = x_{L,k} - \epsilon \cdot s_\theta(x_{L,k}, L) + \sqrt{2\epsilon} \cdot \eta$, where $\eta \sim \mathcal{N}(0, \mathbf{I})$

end for

$\tilde{x}_0 = ODE\_SOLVER_{backward_{L\to 0}}(x_{L,N})$

space, we obtain more accurate estimates within a distribution closely resembling a multivariate Gaussian. This concept was foundational to SDEdit; however, their reliance on noise injection prevented full inversion of the diffusion process without losing information from the observation.

Langevin Dynamics in Latent Spaces There exists a direct correspondence between DDPM, ϵ_θ, and Noise Conditional Score Network, s_θ, such that $s_\theta(\mathbf{x}_t, t) = -\frac{\epsilon_\theta(\mathbf{x}_t, t)}{\sigma_t}$.

Building upon that, we propose to perform gradient-update in our latent spaces utilizing Langevin dynamics as in equation (1) to increase the likelihood of our latent representation. The method is described in the Appendix D, Algorithm 2. Analogous to SDEdit and contrasting with alternative methods, our editing method can be tailored to prioritize either realism or fidelity by selecting the step size in the Langevin dynamics and the latent spaces where to optimize.

Our editing technique, relying on updates within a designated latent space, facilitates an extensive array of editing possibilities on the input image, as optimizing in one latent space yields distinct outcomes compared to optimizing in another. Whereas SDEdit provides a singular hyper-parameter to govern the editing process, our method bifurcates this control mechanism into two distinct parameters: the step size in the updates and the choice of latent space for optimization. This dual-parameter approach enables our editing method to equal SDEdit's performance on tasks where the latter is effective and to outperform in tasks that are unattainable for SDEdit.

Confidence Based Clipping (CBC)

Here, we present a clipping method for the latent codes applied during the encoding process that does not depend on the prediction or the original sample. The proof is available in Appendix B.

Proposition 1. *Let Φ be the cumulative distribution function of $\mathcal{N}(0, \mathcal{I})$ and let $x_0 \in [-1, 1]$. For $\alpha_t \in [0, 1]$, $\forall t \in [0, 1]$, assume that $x_t \sim \mathcal{N}(\sqrt{\alpha_t} \cdot \alpha_0, \sqrt{1-\alpha_t} \cdot \mathcal{I})$. Then, for all η:*

$$\mathcal{P}(x_t \in [-\sqrt{\alpha_t} - \eta \cdot \sqrt{1-\alpha_t}, \sqrt{\alpha_t} + \eta \cdot \sqrt{1-\alpha_t}]) \geq \Phi(\eta) - \Phi(-\eta).$$

Specifically, for $\eta = 2$:

$$\mathcal{P}(x_t \in [-\sqrt{\alpha_t} - 2 \cdot \sqrt{1-\alpha_t}, \sqrt{\alpha_t} + 2 \cdot \sqrt{1-\alpha_t}]) \geq 0.95.$$

During the encoding process, we propose to clip the latent codes using a confidence interval derived from Proposition 1.

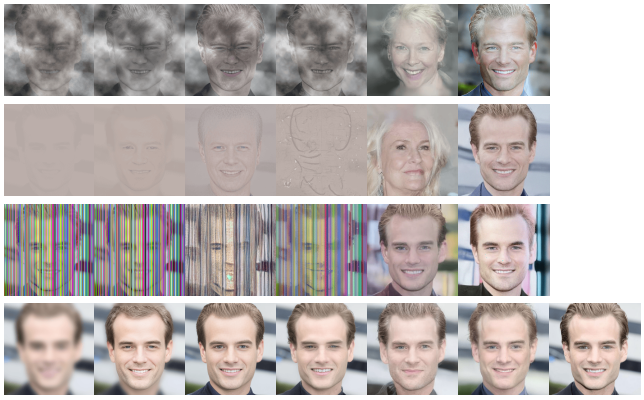


Figure 3: Visual comparison on CelebAHQ with various corruption types. CODE is the only method performing on all corruptions types, significantly improving over SDEdit on two complex corruptions, Fog and Contrast. Other baselines demonstrate lower versatility while requiring extra training.

Confidence-based clipping is performed as follows:

$$x_t^{clipped} = \text{Clip}(x_t, \text{min} = -\sqrt{\alpha_t} - \eta \cdot \sqrt{1-\alpha_t}, \text{max} = \sqrt{\alpha_t} + \eta \cdot \sqrt{1-\alpha_t}),$$

where t is the timestep, α_t is the predefined schedule of the DM, and η is the chosen confidence parameter.

Similar to our editing method, CBC is agnostic to the input and suitable for blind restoration scenarios. We combine CBC with our ODE editing method to form our complete method, CODE, detailed in Algorithm 1. As shown in Figure 9, the two methods synergize efficiently. It is crucial to note that CBC cannot be used in combination with SDEdit.

4 Experiments

Setup We use open-source pre-trained DDPM models (Ho, Jain, and Abbeel 2020) from HuggingFace, specifically the EMA checkpoints of DDPM models trained on CelebA-HQ (Karras et al. 2018), LSUN-Bedroom, and LSUN-Church (Yu et al. 2016), all at 256x256 resolution. For all experiments, DDIM inversion (Song and Ermon 2020) with 200 steps is utilized. Enhancement follows the complete Algorithm 3 described in Appendix D. It is used with $N = 200$ Langevin iteration steps, a step size ϵ of $[10^{-2}, 10^{-3}]$ for shallow latent spaces (up to $L = 40$), and $[10^{-5}, 10^{-6}]$ for deeper latent spaces ($L > 100$). We use $K = 4$ annealing steps and $\alpha = 0.8$ as the annealing coefficient. When activating CBC, we use $\eta = 1.7$. A full description of the setup being used to automatically compute the metrics is provided in Appendix E. For SDEdit, samples are generated with L in $[300, 500, 700]$ steps.

We tested our approach on 47 corruption types, including 17 from (Hendrycks and Dietterich 2019) (noise, blur, weather, and digital artifacts) and 28 from (Mintun, Kirillov, and Xie 2021). The corruption codebases are publicly

Figure 4: Visual comparison of general image restoration on various corruptions - LSUN

available[1] [2]. Additionally, we introduced two masking types: masking entire vertical lines and random pixels with random colors. Unlike traditional masking in masked autoencoders (He et al. 2022), our method does not assume knowledge of masked pixels' positions, posing a more realistic recovery task. CODE operates completely blind to the corruption type, with no knowledge of the specific task or affected pixels.

For each corruption type, we test on at least 500 corrupted images. For each image, we kept the best 4 samples generated based on PSNR with respect to the original non-degraded images.

Baselines Our main baseline is the domain-agnostic method SDEdit (Meng et al. 2021), the only one comparable to ours in terms of requirements and assumptions. On CelebAHQ, the performance is also qualitatively benchmarked against domain-specific SOTA models, namely CodeFormer (Zhou et al. 2022), GFPGAN (Wang et al. 2021a), and DiffBIR (Lin et al. 2023). We also conducted visual experiments on LSUN-Bedroom and LSUN-Church to demonstrate the efficacy of CODE over diverse domains similar to SDEdit in (Meng et al. 2021).

Evaluation Metrics We evaluate our results using PSNR, SSIM, LPIPS, and FID. PSNR and SSIM are measured against the corrupted image (input) to assess fidelity to the guidance. FID is used to evaluate the quality of our generated images. Given the absence of assumptions about the input and corruptions, a key metric is the trade-off between realism and fidelity—specifically, the gain in realism relative to a given loss in fidelity. To quantify this, we use L2 distance in the pixel space as a measure of fidelity and FID as a measure of realism, plotting them against each other in Figure 5. Additionally, we report LPIPS with respect to the original, non-degraded image (source) to assess reconstruction quality. This metric is particularly informative for evaluating each corruption individually, as it also reflects the complexity of the corruption, with detailed results provided in Appendix G.

Results We present a brief qualitative comparison of results in Figure 3 to showcase that most methods, without further assumptions, cannot perform properly. In the vast majority of scenarios involving severe corruption like contrast, random pixel masking, or fog, only SDEdit and CODE can generate convincing images. For less intensive corruptions, which typically include erasing fine details or introducing minor noise, most baseline models tend to perform well. We provide extensive results in Appendix E. For quantitative metrics, we focus on SDEdit and CODE and compare them using the same pre-trained model on CelebA-HQ. Consequently, the differences come only from the way the pre-trained diffusion model is leveraged. Average metrics across the employed corruptions are detailed in Table 1. **CODE outperforms SDEdit by 36% in FID-score** while maintaining **a fidelity to the input (PSNR-Input) 5% higher than SDEdit**. Moreover, the standard deviation of the FID score highlights that SDEdit fails in certain cases while CODE is more stable. Finally, we report in Figure 5 the trade-off curves between fidelity and realism for both CODE and SDEdit. We performed a polynomial regression on CODE and SDEdit results to obtain

[1]https://github.com/hendrycks/robustness

[2]https://github.com/facebookresearch/augmentation-corruption/blob/fbr_main/imagenet_c_bar/corrupt.py

	LPIPS-Source ↓	SSIM-Input ↑	PSNR-Input ↑	FID ↓
Inputs	0.48 (0.35)	-	-	143.49 (96.31)
SDEdit	0.32 (0.13)	0.46 (0.21)	18.74 (3.92)	47.84 (42.29)
CODE	**0.30** (0.12)	**0.49** (0.22)	**19.61** (4.66)	**30.66** (16.21)

Table 1: Average values of different metrics across the 47 considered corruptions, along with the standard deviations. CODE outperforms SDEdit on all metrics. CODE preserves a higher degree of fidelity while reaching a higher degree of realism using the same pre-trained model.

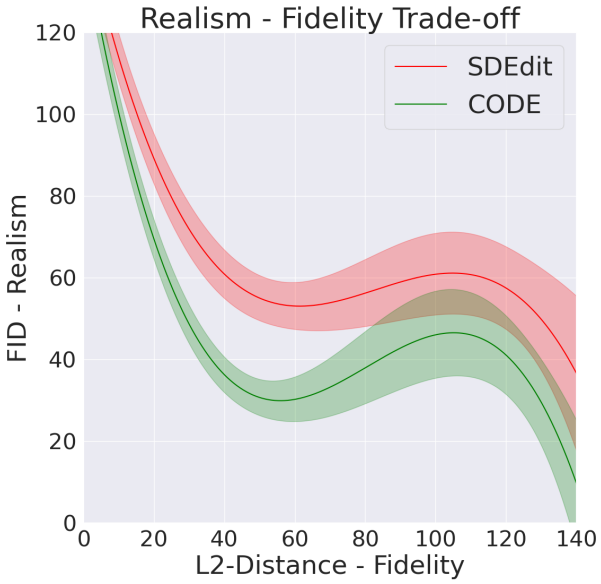


Figure 5: Comparison of realism-fidelity trade-off between SDEdit and CODE. Polynomial regression curves with shaded areas show one standard deviation. CODE produces more realistic images at the same fidelity. Both methods converge when the input distance is large, as they use the same pre-trained model.

Figure 6: Analysis of Langevin Dynamics convergence. As the number of updates increases, the output realism increases until convergence and then stabilizes.

such a curve. Both methods offer hyper-parameters to control such trade-offs. However, we highlight that CODE offers a better possibility. Overall, CODE generates more realistic outputs for a given degree of fidelity.

5 Ablation Study

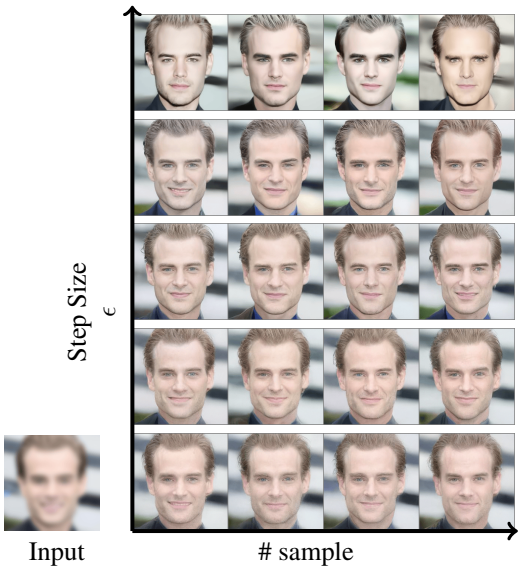
Analysis of Hyperparameters

Number of Updates. As shown in Figure 6, the number of update iterations conducted in a latent space is pivotal for ensuring convergence and reducing variability. In practice, we employed 300 steps in all our experiments.

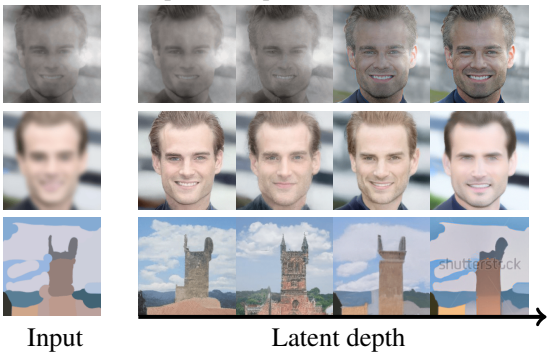
Step Size. The step size emerges as a critical parameter. A smaller step size results in high fidelity to the input and low variability among generated samples, albeit compromising realism. Conversely, an increased step size enhances realism and variability, as depicted in Figure 7a. As the number of updates is fixed in our experiments, the step size is what governs the size of the explored neighborhood around the input image. As a result, its impact is related to the amount of noise injected in SDEdit.

Latent Space Choice. The choice of latent space significantly influences the type of changes made during updates. As shown in Figure 7b, updates in a shallow latent space lead to minor but detailed and realistic modifications. In contrast, updates in deeper latent spaces can cause more significant or complex changes. Interestingly, regarding stroke guidance, optimization in the deepest latent space led to the addition of text and lines to the image. This suggests that the training set likely contained numerous images with these text and lines, implying that their inclusion by the model significantly enhanced the image's likelihood. Empirically, we found that the deeper the latent space, the less the notion of distance is close to an L2 pixel-based distance.
The optimal latent space is not one-size-fits-all but depends on the specific input being processed. For complex corruptions, using a mix of updates in different latent spaces proves most effective. On the other hand, shallow latent spaces are best for addressing simple corruptions like blur. This ability to independently select the latent space without affecting other parameters, such as the level of noise injection, is a key strength of our editing method. We disentangle what was previously a single parameter into multiple, allowing for tailored optimization on a per-sample basis.

Fidelity. Our editing method is anchored in the corrupted sample, hence the generation is very impacted by the variations in the corruption. As shown in Figure 8, the outputs are



(a) Impact of step size selection



(b) Impact of latent space selection

Figure 7: (a) Impact of step size on sample diversity and realism using CODE: Larger steps increase diversity but reduce fidelity. (b) Impact of latent space choice on the quality and characteristics of generated images across different corruption types.

faithful to the corrupted image and do not map to a single ground-truth image.

Ablation Study of Confidence-based clipping

Impact of Confidence-Based Clipping In this section, we study the impact of the confidence parameter η in CBC. As we reduce η, the interval shrinks, keeping only the most likely pixel values. We propose to encode an image using DDIM with different values of η and decode it back to see the result. We study this in the case of in-distribution images and of corrupted images. Results can be seen in Figure 9b. A smaller η results in a loss of fine-grained details, a shift of the average tone, and the removal of unlikely pixels such as masked pixels. When applied to corrupted samples, CBC proves efficient in removing part of the noisy artifact while keeping most of the image structure. Interestingly, CBC also

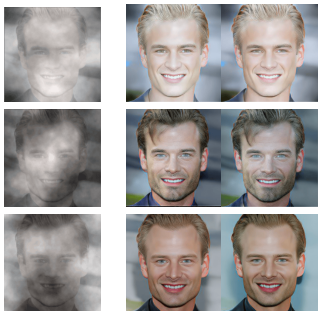


Figure 8: CODE outputs adapt faithfully to variations in the image input.

stabilizes the DDIM inversion, which might sometimes be inconsistent. This allows for fewer steps in the encoding-decoding procedure and speeds up the whole editing process.

Ablation Study We propose to study the impact of each block in CODE while keeping SDEdit for comparison. Qualitative results are visible in Figure 9a. While both our editing method alone (w/o CBC) and SDEdit excel at adding extra fine-grained details, they fail at handling unknown masks or color shits efficiently. On the contrary, CBC basically fails at adding extra details but successfully recovers certain color shifts or masked areas. As a result, combining both into CODE leads to powerful synergies. Quantitative results can be seen in Table 2.

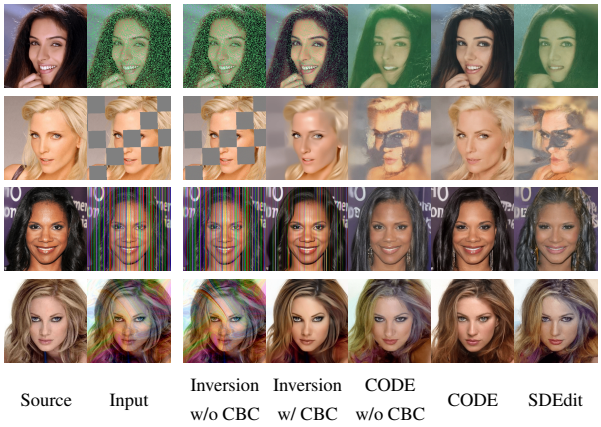
	LPIPS-Source ↓	SSIM-Input ↑	PSNR-Input ↑	FID ↓
Inputs	0.48 (0.35)	-	-	143.5 (96.3)
DDIM	0.52 (0.32)	0.89 (0.08)	30.7 (5.3)	152.2 (88.4)
DDIM w/ CBC	0.43 (0.19)	0.73 (0.17)	23.5 (5.3)	90.1 (49.5)
CODE w/o CBC	0.31 (0.11)	0.48 (0.22)	19.4 (4.5)	34.75 (14.6)
CODE w/ CBC	**0.30 (0.12)**	**0.49 (0.22)**	**19.6 (4.7)**	**30.65 (16.2)**

Table 2: Confidence-Based Clipping ablation, metrics are averaged across all corruptions.

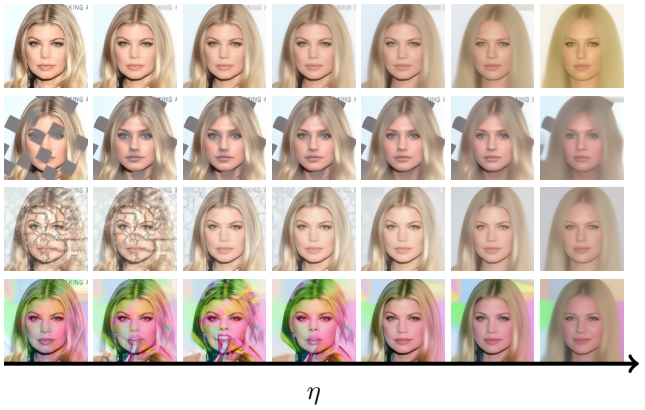
Discussion. While CODE offers enhanced versatility and control in editing, it introduces greater complexity compared to SDEdit. SDEdit's tuning is straightforward, with binary success or failure outcomes, whereas CODE's dual hyperparameter framework requires a more extensive grid search, increasing the search complexity quadratically. However, this added complexity enables CODE to achieve better results across a wider range of scenarios.

6 Conclusion

We introduce Confident Ordinary Differential Editing, a novel approach for guided image editing and synthesis that handles OoD inputs and balances realism and fidelity. Our method eliminates the need for retraining, finetuning, data augmentation, or paired data, and it integrates seamlessly with any pre-trained Diffusion Model. CODE excels in addressing



(a) Ablation Study.



(b) DDIM inversion using CBC with different values for confidence parameter η.

Figure 9: (a) Ablation Study (b) DDIM inversion using CBC with different values for the confidence parameter η.

a wide array of corruptions, outperforming existing methods that often rely on handcrafted features. As an evolution of SDEdit, our approach provides enhanced control, variety, and capabilities in editing by disentangling the original method and introducing additional hyperparameters. These new parameters not only offer deeper insights into the functioning of Diffusion Models' latent spaces but also enable more diverse editing strategies. Furthermore, we introduce a Confidence-Based Clipping method that synergizes effectively with our editing technique, allowing the disregard of unlikely pixels or areas in a completely agnostic manner. Finally, our extensive study of the different components at play offers a greater understanding of the underlying mechanics of diffusion models, enriching the field's knowledge base. Our findings reveal that CODE surpasses SDEdit in versatility and quality while maintaining its strengths across various tasks, including stroke-based editing. We hope our work inspires further innovations in this domain, akin to the transformative impact of GAN inversion. Looking ahead, we see potential in automating the editing and hyperparameter search processes and exploring synergies with text-to-image synthesis.

Acknowledgements

We thank the reviewers, for their valuable comments and insights.

This research is funded by the Swiss National Science Foun- dation (SNSF) through the project Narratives from the Long Tail: Transforming Access to Audiovisual Archives (Grant: CRSII5 198632). The project description is available on: https://www.futurecinema.live/project/

References

Abdal, R.; Qin, Y.; and Wonka, P. 2019. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, 4432–4441.

Abdal, R.; Qin, Y.; and Wonka, P. 2020. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8296–8305.

Batzolis, G.; Stanczuk, J.; Schönlieb, C.-B.; and Etmann, C. 2021. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*.

Chan, K. C.; Wang, X.; Xu, X.; Gu, J.; and Loy, C. C. 2021. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14245–14254.

Chen, C.; Li, X.; Yang, L.; Lin, X.; Zhang, L.; and Wong, K.-Y. K. 2021. Progressive semantic-aware style transformation for blind face restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11896–11905.

Chen, Y.; Tai, Y.; Liu, X.; Shen, C.; and Yang, J. 2018. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2492–2501.

Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*.

Chung, H.; Kim, J.; Mccann, M. T.; Klasky, M. L.; and Ye, J. C. 2023. Diffusion Posterior Sampling for General Noisy Inverse Problems. arXiv:2209.14687.

Chung, H.; Kim, J.; and Ye, J. C. 2023. Direct Diffusion Bridge using Data Consistency for Inverse Problems. arXiv:2305.19809.

Fei, B.; Lyu, Z.; Pan, L.; Zhang, J.; Yang, W.; Luo, T.; Zhang, B.; and Dai, B. 2023. Generative Diffusion Prior for Unified Image Restoration and Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9935–9946.

Gao, J.; Zhang, J.; Liu, X.; Darrell, T.; Shelhamer, E.; and Wang, D. 2022. Back to the source: Diffusion-driven test-time adaptation. *arXiv preprint arXiv:2207.03442*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gu, Y.; Wang, X.; Xie, L.; Dong, C.; Li, G.; Shan, Y.; and Cheng, M.-M. 2022. VQFR: Blind face restoration with vector-quantized dictionary and parallel decoder. In *European Conference on Computer Vision*, 126–143. Springer.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *Proceedings of the International Conference on Learning Representations*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. arXiv:1710.10196.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35: 23593–23606.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Langevin, P. 1908. Sur la théorie du mouvement brownien. *C. R. hebd. Séanc. Acad. Sci. Paris*.

Li, B.; Xue, K.; Liu, B.; and Lai, Y.-K. 2023. BBDM: Image-to-image Translation with Brownian Bridge Diffusion Models. arXiv:2205.07680.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Fei, B.; Dai, B.; Ouyang, W.; Qiao, Y.; and Dong, C. 2023. DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior. *arXiv preprint arXiv:2308.15070*.

Liu, G.-H.; Vahdat, A.; Huang, D.-A.; Theodorou, E. A.; Nie, W.; and Anandkumar, A. 2023. I^2SB: Image-to-Image Schrödinger Bridge. arXiv:2302.05872.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

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 *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2437–2445.

Mintun, E.; Kirillov, A.; and Xie, S. 2021. On Interaction Between Augmentations and Corruptions in Natural Corruption Robustness. *Advances in neural information processing systems*.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Murata, N.; Saito, K.; Lai, C.-H.; Takida, Y.; Uesaka, T.; Mitsufuji, Y.; and Ermon, S. 2023. GibbsDDRM: A Partially Collapsed Gibbs Sampler for Solving Blind Inverse Problems with Denoising Diffusion Restoration. arXiv:2301.12686.

Oord, A. v. d.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*.

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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7474–7489.

Poirier-Ginter, Y.; and Lalonde, J.-F. 2023. Robust Unsupervised StyleGAN Image Restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22292–22301.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Song, Y.; and Ermon, S. 2020. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33: 12438–12448.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Su, X.; Song, J.; Meng, C.; and Ermon, S. 2023. Dual Diffusion Implicit Bridges for Image-to-Image Translation. In *The Eleventh International Conference on Learning Representations*.

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

Vincent, P. 2011. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7): 1661–1674.

Wang, X.; Li, Y.; Zhang, H.; and Shan, Y. 2021a. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9168–9178.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021b. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In *International Conference on Computer Vision Workshops (ICCVW)*.

Wang, Y.; Yu, J.; and Zhang, J. 2022. Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model. arXiv:2212.00490.

Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; and Gool, L. V. 2023. DiffIR: Efficient Diffusion Model for Image Restoration. arXiv:2303.09472.

Yang, P.; Zhou, S.; Tao, Q.; and Loy, C. C. 2023. PGDiff: Guiding Diffusion Models for Versatile Face Restoration via Partial Guidance. arXiv:2309.10810.

Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2021. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 672–681.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2016. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. arXiv:1506.03365.

Yu, X.; Fernando, B.; Ghanem, B.; Porikli, F.; and Hartley, R. 2018. Face Super-resolution Guided by Facial Component Heatmaps. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhou, S.; Chan, K.; Li, C.; and Loy, C. C. 2022. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35: 30599–30611.

Zhu, J.; Shen, Y.; Zhao, D.; and Zhou, B. 2020. In-domain gan inversion for real image editing. In *European conference on computer vision*, 592–608. Springer.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhu, Y.; Zhang, K.; Liang, J.; Cao, J.; Wen, B.; Timofte, R.; and Gool, L. V. 2023. Denoising Diffusion Models for Plug-and-Play Image Restoration. arXiv:2305.08995.