# Landscape Learning for Neural Network Inversion

**Ruoshi Liu**[1], **Chengzhi Mao**[1], **Purva Tendulkar**[1], **Hao Wang**[2], and **Carl Vondrick**[1]

[1]Columbia University
[2]Rutgers University

## Abstract

Many machine learning methods operate by inverting a neural network at inference time, which has become a popular technique for solving inverse problems in computer vision, robotics, and graphics. However, these methods often involve gradient descent through a highly non-convex loss landscape, causing the optimization process to be unstable and slow. We introduce a method that learns a loss landscape where gradient descent is efficient, bringing massive improvement and acceleration to the inversion process. We demonstrate this advantage on a number of methods for both generative and discriminative tasks, including GAN inversion, adversarial defense, and 3D human pose reconstruction.

## 1 Introduction

Many inference problems in machine learning are formulated as inverting a forward model $F(x)$ by optimizing an objective over the input space $x$. This approach, which we term optimization-based inference (OBI), has traditionally been used to solve a range of inverse problems in vision, graphics, robotics, recommendation systems, and security [20, 20, 30, 17, 8, 49, 13]. Recently, neural networks have emerged as the paramterization of choice for forward models [35, 42, 1, 39, 61, 54, 11, 63], which can be pretrained on large collections of data, and inverted at testing time in order to solve inference queries.

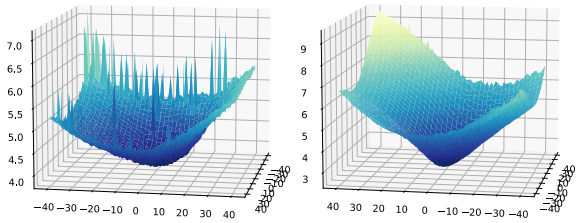



Figure 1: **Loss Landscapes Comparison.** The loss landscape of optimization-based inference (OBI) is often highly non-convex. We propose to learn a smoother loss landscape through a mapping network to accelerate the optimization procedure. Plotted from real data.

Optimization-based inference has a number of advantages over feed-forward or encoder-based inference (EBI) for neural network inversion. Since there is no encoder, OBI provides flexibility to adapt to new tasks, allowing one to define new constraints into the objective during inference [11, 2]. When observations are partially missing, OBI can adapt without additional training [39]. Moreover, OBI naturally supports generating multiple and diverse hypotheses when there is uncertainty [32, 39]. Finally, OBI has intrinsic advantages for robustness, both adapting to new data distributions as well as defending against adversarial examples [37].

However, the key bottleneck for OBI in practice is the computational efficiency and the speed of inference. Feedforward models are fast because they only require a single forward pass of a neural network, but OBI requires many (often hundreds) steps of optimization in order to obtain strong results for one example. Forward models in OBI are often trained with generative or discriminative tasks, but they are never trained for the purpose of performing gradient descent in the input space. Consequently, the loss landscape is often highly non-convex, visualized in Fig. 1 (left). This non-convexity directly causes the instability and inefficiency of the optimization.

Preprint. Under review.

In this paper, we propose a new framework which allows for faster and stabler optimization for OBI. Our key insight is to train a mapping network in the latent space that is aware of the optimization procedure at inference time. Existing optimization over the input space converges slowly, because the forward model is not designed for a test-time optimization procedure. By first collecting samples from the optimization trajectories in the latent input space, and then training the mapping network to minimize the loss on each sample of the trajectories, the mapping network will map each sample, even on the beginning of the trajectory, to an input that minimizes the loss. In doing this, the landscape becomes smoother, which allows even a few steps of optimization to obtain high quality results. We achieve this with an coordinate descent algorithm that first collects samples from the optimization trajectories in the new input space, then trains the mapping network to minimize the loss on each sample of the trajectories.

Empirical experiments and visualizations on both generative and discriminative models show that our method can significantly improve the convergence speed for optimization. We validate our approach on a diverse set of computer vision tasks, where we achieve up to 34% gain on GAN inversion [1], 18% accuracy gain on adversarial defense [37], and up to 75% performance gain on 3D human pose reconstruction [42]. In addition, our method converges an order of magnitude faster without loss in absolute performance. As our approach does not require retraining of the forward model, it can be compatible to all existing OBI methods with a differentiable forward model and objective function.

The primary contribution of this paper is an efficient optimization-based inference framework. In Sec. 2, we survey the related literature to provide an overview of forward model inversion problem. In Sec. 3, we formally define OBI, our method to learn a faster loss landscape for OBI, and a training algorithm for better generalization and robustness. In Sec. 4, we experimentally study and analyze the effectiveness of mapping network for OBI. We will release all code and models.

## 2 Related Work

The different approaches for inference with a neural network can be partitioned into either encoder-based inference, which is feedforward, or optimization-based inference, which is iterative. We briefly review these two approaches in the context of our work.

### 2.1 Encoder-based Inference

Encoder-based inference trains a neural network $F$ to directly map from the output space to the input space. Auto-encoder based approach [44] learns an encoder that map the input data to the latent space. [47, 52, 56, 43] learn an encoder from the image to the latent space in GAN. Encoder based inference requires training the encoder on the anticipated distribution in advance, which is often less effective and can fail on unexpected samples [16, 25]. In addition, encoder-based method can only produce one reconstruction for image inpainting [60], even if multiple outcomes can all be true given the partial input.

### 2.2 Optimization-based Inference

OBI methods perform inference by solving an optimization problem with gradient-based methods such as Stochastic Gradient Descent (SGD) [7] and Projected Gradient Descent (PGD) [36]. In these cases, the objective function specifies the inference task. Besides these methods which use a point estimate for the latent variable, one can estimate the posterior distribution of the latent variable through Bayesian optimization, such as SGLD [57].

Gradient based optimization methods have been used to infer the latent code of query samples in deep generative models like GANs [18] via GAN inversion [28, 23, 48, 66, 1, 3, 6, 21, 41]. Style transfer relies on gradient based optimization to change the style of the input images [24]. It can also create adversarial attacks that fool the classifier [14, 9, 38, 50]. Recently, backpropagation-based optimization has shown to be effective in defending adversarial examples [37].

Recently, constrained optimization was popularized for text-to-image synthesis by [15, 33]. They search in the latent space to produce an image that has the highest similarity with the given text as measured by a multi-modal similarity model like CLIP [45]. Test-time constrained optimization is also related to the idea of 'prompt-tuning' for large language models. [31] learn "soft prompts" to

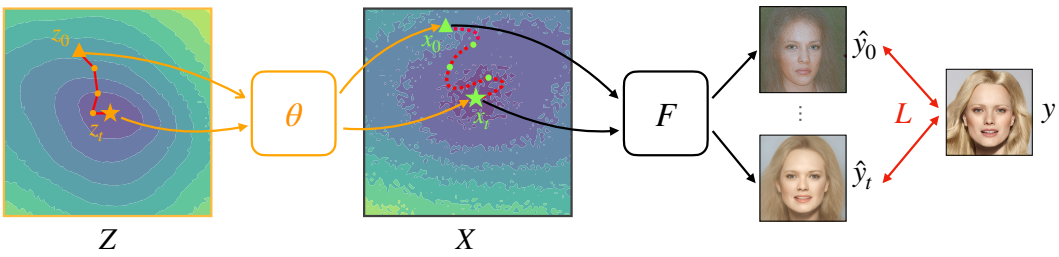



Figure 2: **Method**. The left and middle figure show the loss landscape for our latent space and the original latent space, respectively. While walking to the optimal solution in a few steps is hard in $X$ space, it can be done in our learned loss landscapes.

condition frozen language models to perform specific downstream tasks. Soft prompts are learned through backpropagation to incorporate signal from just a few labeled examples (few-shot learning).

A major challenge for optimization-based inference is how to perform efficient optimization in a highly non-convex space. To address this, input convex model [5] was proposed so that gradient descent can be performed in a convex space. [53] introduced a method to retrain the generative model such as it learns a latent manifold that is easy to optimize. When the model cannot be changed and retrained, bidirectional inference [54] and hybrid inference [66, 65] uses an encoder to provide a good initialization for the optimization-based inference in a non-convex space. Our method does not retrain the generative model, but maps the original latent space to a fast loss landscape.

## 3 Learning Landscapes for Fast Inference

We present our framework to learn a fast loss landscape for optimization-based inference (OBI) methods. In Sec. 3.1, we will define OBI. In Sec. 3.2, we will introduce our framework as well as the training objective. In Sec. 3.3, we will describe how to train our model with a coordinate descent algorithm and an experience-replay buffer.

### 3.1 Optimization-based Inference

Let $F(\mathbf{x}) = \hat{y}$ be a differentiable forward model that generates an output $\hat{y}$ given an input variable $\mathbf{x} \in X$. For example, $\hat{y}$ might be an image, and $\mathbf{x}$ might be the latent variables for a generative model. Given an observation $y$, the goal of OBI is to find the input $\hat{\mathbf{x}} \in X$ such that an objective function $L(\hat{y}, y)$ is minimized. Formally, we write this procedure as:

$$\hat{\mathbf{x}} = \underset{\mathbf{x} \in X}{\operatorname{argmin}} \, L(F(\mathbf{x}), y) \tag{1}$$

When the objective function L and the model $F$ are both differentiable, we can perform the optimization over input space $X$ with gradient descent. However, the original forward model has not been trained to perform gradient descent in the input space. Consequently, the loss landscape is highly non-convex, making the convergence slow.

### 3.2 Remapping the Input Space

Instead of operating in the original input space $X$, we will create a new space $Z$ where gradient descent is efficient and converges in a small number of iterations. To parameterize $Z$, we will use a neural network $\theta : Z \to X$ that maps from the new space $Z$ to the original space $X$. The learning problem we need solve is to estimate the parameters of $\theta$ so that there is a short gradient descent path in $Z$ from the initialization to the solution. Fig. 2 shows an overview of this setup.

We formulate an objective by rolling out the gradient updates on $\mathbf{z}$, where we write $\mathbf{z}_t \leftarrow \mathbf{z}_{t-1} + \lambda \frac{\partial L}{\partial \mathbf{z}_{t-1}}$ as the $t^{\text{th}}$ update with a step size of $\lambda$. For gradient descent in space $Z$ to be efficient, the goal of our introduced $\theta$ is to move every step $\mathbf{z}_t$ as close as possible to the global minima:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \; \mathbb{E}_{(\mathbf{z},y)} \left[ \sum_{t=0}^{T} L(F(\theta(\mathbf{z}_t)), y) \right] \quad \text{where} \quad \mathbf{z}_t = \begin{cases} 0, & t = 0 \\ \mathbf{z}_{t-1} + \lambda \frac{\partial L}{\partial \mathbf{z}_{t-1}}, & t > 0 \end{cases} \tag{2}$$

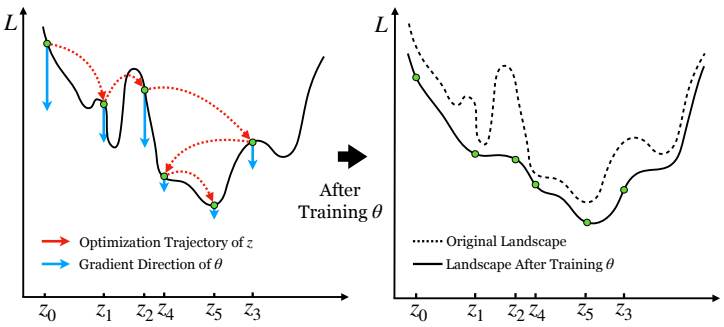



Figure 3: **Landscape Learning.** An optimization trajectory $\{\mathbf{z}_t\}_{t=0}^{5}$ collected is used to train $\theta$. $\mathbf{z}_i$ that corresponds to a higher $L_i$ will yield a higher gradient when training $\theta$. Optimization over multiple steps along the trajectory causes $\theta$ to learn *patterns* of trajectories and create a smoother loss landscape.

We visualize this process with a toy example in Fig. 3. Gradient updates on $\theta$ w.r.t multiple steps $\mathbf{z}_t$ along a trajectory will cause the loss value on each step to be lowered. By learning the parameters of $\theta$ across many examples, $\theta$ can learn the patterns of optimization trajectories in $X$. For example, $\theta$ can learn to estimate the step size in $X$ and dynamically adjust it to each example. Moreover, $\theta$ can learn to smooth non-convex regions in the landscape.

Once we obtain $\hat{\theta}$, we do inference on a new example $y$ through the standard optimization-based inference procedure, except in $Z$ space now. Given the observation $y$, we find the corresponding $\hat{\mathbf{x}}$ through the optimization:

$$\hat{\mathbf{x}} = \hat{\theta}(\hat{\mathbf{z}}) \quad \text{where} \quad \hat{\mathbf{z}} = \underset{\mathbf{z} \in Z}{\operatorname{argmin}} \; L(F(\hat{\theta}(\mathbf{z})), y) \tag{3}$$

When the inverse problem is under-constrained, one can infer multiple hypotheses for $\hat{\mathbf{x}}$ by repeating the above optimization multiple times with a different random initialization for $\mathbf{z}_0$.

### 3.3 Training

We use coordinate descent (CD) in order to train $\theta$ jointly with estimating $\mathbf{z}$ for each example in the training set. Specifically, we first fix parameters of $\theta$ and collect $N$ optimization trajectories of $\mathbf{z}$, each with length $T$. Adopting the same terminology from the robotics community for learning on continuous states [40], we term this a *experience replay buffer*. Subsequently, we randomly sample data from this buffer and train $\theta$ to optimize the loss function. We alternate between these two steps for a fixed number of iterations with gradient descent for both. Depending on the application, the training time for $\theta$ varied from one hour to one day using a four GPU server. Please see the appendix for more implementation details.

We also experimented with an online version of the training algorithm, where we update $\theta$ immediately after one update to $\mathbf{z}$. However, in our experiments, we found this resulted in a slower convergence rate. We show these comparisons in the ablation experiments.

## 4 Experiments

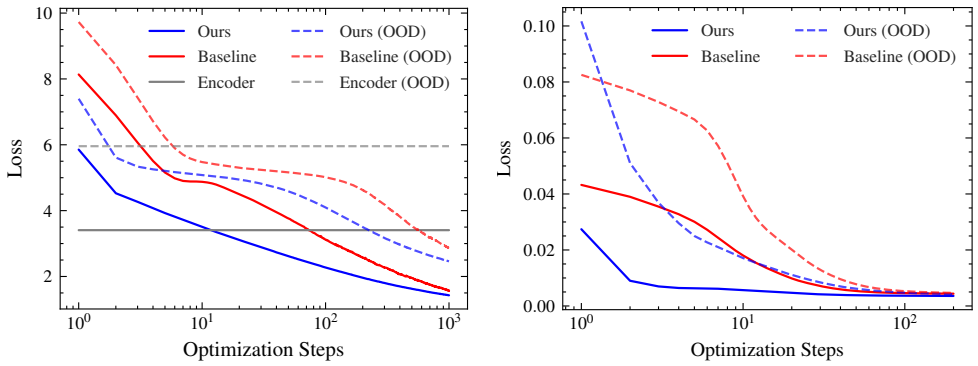
The goal of our experiments is to analyze how well our proposed method can be applied to various existing OBI methods to improve the optimization efficiency. We demonstrate application of our method to three diverse OBI methods in computer vision, including both generative models and discriminative models. For each OBI method, the inference-time optimization objective of the baseline and ours can be written as:

$$\text{Baseline: } \hat{\mathbf{x}} = \min_{\mathbf{x} \in X} L(F(\mathbf{x}), y), \qquad \text{Ours: } \hat{\mathbf{z}} = \min_{\mathbf{z} \in Z} L(F(\theta(\mathbf{z})), y) \tag{4}$$

Next, we provide the specific implementation of the loss term $L$ for each application, along with quantitative and qualitative results. We also perform experiments to understand the loss landscape in Sec. 4.4 and perform ablations on different parts of our approach in Sec. 4.5.

### 4.1 GAN Inversion

We first validate our method on StyleGAN inversion [1]. We take a pretrained generator of StyleGAN [27] denoted as $F$. Let $y$ be an observed image whose input variable we are recovering, we optimize




(a) GAN Inversion (b) 3D Human Pose Reconstruction

Figure 4: **Optimization Performance.** We visualize the trends of optimization performance compared with the baseline. In **GAN Inversion (Left)**, we evaluate all models on test splits of CelebA-HQ [26] and LSUN-cat [62] (OOD) with loss defined in Eq. 5. Since encoder-based inference doesn't involve optimization, we use a flat line to represent it. We perform 2000 steps of gradient descent for all models except encoder-based models. In **3D Human Pose Reconstruction (Right)**, we evaluate all models on test splits of GRAB [51] and PROX [19] (OOD) with loss defined in Eq. 6. We perform 200 steps of gradient descent for all models. For each step, we plot the average loss value of test splits.

the objective of Eq. 4 where the loss can be written as,

$$L(\hat{y}, y) = L_{lpips}(\hat{y}, y) + ||\hat{y} - y||_2^2 \tag{5}$$

where $L_{lpips}$ is a perceptual similarity loss introduced in [64], $\hat{y} = F(\hat{x})$ for baseline and $\hat{y} = F(\theta(\hat{z}))$ for ours. We train $\theta$ on the train split of CelebA-HQ[26] dataset and evaluate on CelebA-HQ validation split for in-distribution experiments and LSUN-Cat[62] for distribution shifting (OOD) experiments. We compare the results from our method against the state-of-the-art encoder-based GAN inversion model [47].

**Quantitative Results.** From Fig. 4a, we see that in all experiments, optimization in our space $Z$ consistently outperforms the baseline from the first optimization step to after convergence. This gap in performance is even larger when evaluated on OOD data. This suggests that the improvement in performance is not caused by memorizing the training data. Note that our image reconstruction performance after only 2 steps of optimization is able to outperform or be on-par with the baseline at 20 steps of optimization, resulting in a 10-fold improvement in efficiency. Even after convergence (after 2000 optimization steps), our reconstruction performance improves over the baseline by 15% for in-distribution evaluation and 10% for OOD evaluation. When compared with encoder-based GAN inversion [47], our method achieves better reconstruction after 11 steps of optimization for in-distribution data and 3 steps for OOD data.

**Qualitative Results.** From Fig. 8, we can see that our method already shows improvements on in-distribution data - it can almost perfectly reconstruct details like fine hair strands, the cap on the person's head, the object in the person's mouth as well-as the text on it. Interestingly, our method is able to discover and reconstruct latents for cats while the encoder-based model fails miserably as shown in Fig 8. The performance on OOD data truly highlights the benefits of our method. We also visualize how the face generations evolve over the process of optimization in Fig. 5. We can see that in just 4 steps, our method is already able to reconstruct coarse features such as face orientation, skin tone and hair color, while the baseline has hardly deviated from the initialization in regard to any of these features. Further, in Fig. 7 we visualize reconstructions from partial observations where only the center of the face (row 1) or everything other than the mouth (row 2) is visible. We can see a variety of feasible possibilities for the hidden regions (e.g., different hairstyles, lip colors, expressions, etc) showcasing the diversity of the new latent space.

### 4.2 3D Human Pose Reconstruction

In addition to image generation, our framework also works for 3D reconstruction. For this, we use VPoser [42] – a variational autoencoder [29] that has learnt a pose prior over body pose. VPoser

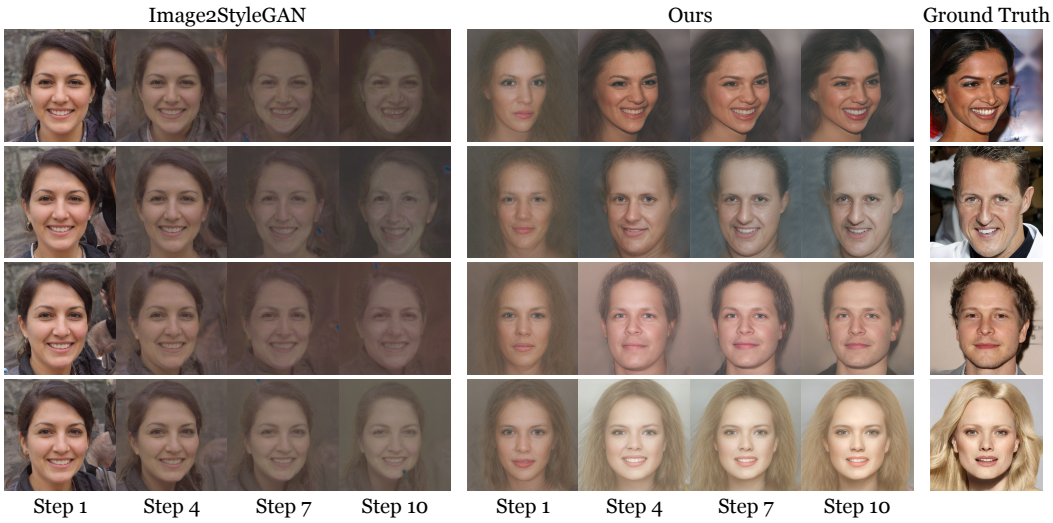



Figure 5: **Optimization Process for GAN Inversion.** Comparing optimization process of our method and the baseline in order to reconstruct the ground truth image. **Left** shows the results from the baseline where optimization is done in the original input space $X$. **Middle** shows the results from our method where optimization is done in our space $Z$. **Right** column contains the ground truth image to each example. Each row corresponds to the same example.

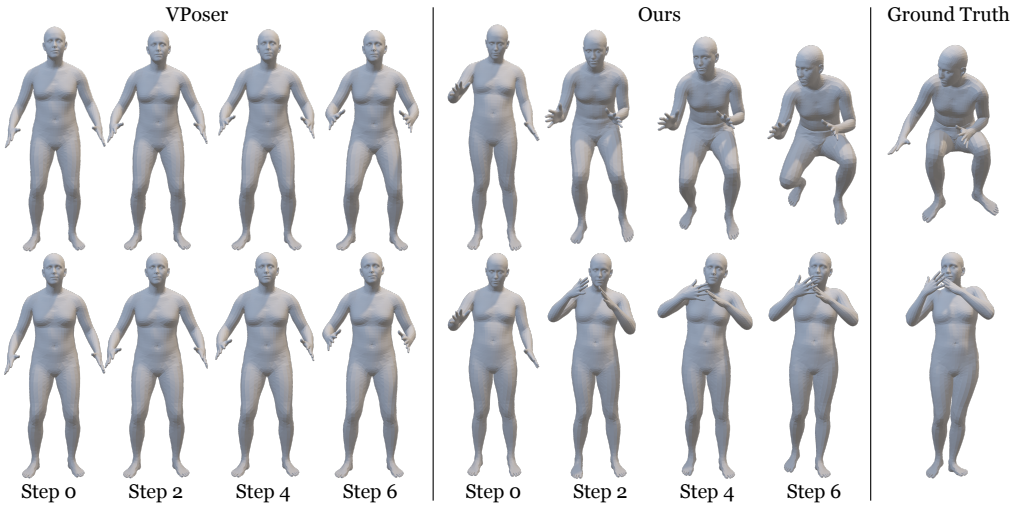



Figure 6: **Optimization Process for 3D Human Pose Reconstruction.** Results shown are for out-of-distribution PROX dataset for sitting (**Top**) and standing (**Bottom**) poses.

was trained on SMPL-X body pose parameters $y \in \mathbb{R}^{63}$ obtained by applying MoSh [34] on three publicly available human motion capture datasets: CMU [12], training set of Human3.6M [22], and the PosePrior dataset [4].

We take a pretrained VPoser decoder denoted as $F$. Our trained mapping network $\theta$ projects a vector from the new input space $\mathbf{z} \in Z$ to a vector in the original VPoser decoder's input space $\mathbf{x} \in X$. Similar to GAN Inversion, we optimize the objective of Eq. 4, where the loss function between predicted and ground truth pose parameters is,

$$L(\hat{y}, y) = ||\hat{y} - y||_2^2 \tag{6}$$

where $\hat{y} = F(\hat{x})$ for the baseline and $\hat{y} = F(\theta(\hat{z}))$ for ours. For training $\theta$, we use the GRAB dataset [51] which contains poses of humans interacting with everyday objects. We construct splits for novel video sequences – thus the test split will contain a seen human subject but a potentially unseen pose / demonstration by that subject. We evaluate on this test split for in-distribution

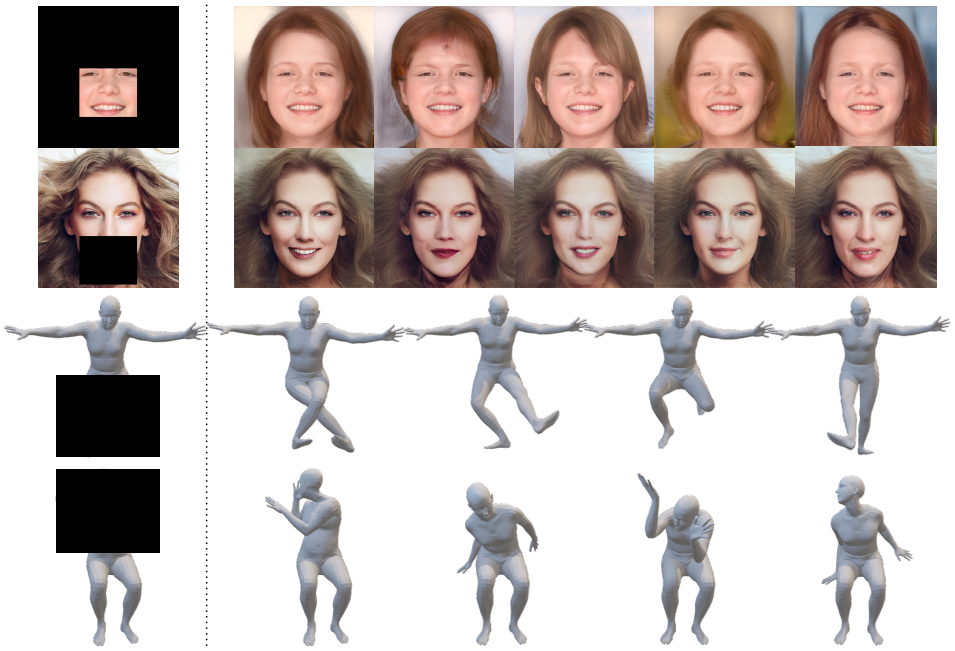



Figure 7: **Diversity of Masked Reconstructions.** We visualize reconstructions for partially observable inputs from random initializations. The masked regions are not considered for loss computation, i.e., the gradient is set to be zero. By optimizing only on the partial observation, we obtain diverse, feasible solutions for the hidden regions.

experiments and on the PROX dataset [19] for OOD experiments, which contains poses of humans interacting in 3D scenes (e.g., living room, office, etc).

**Quantitative Results.** For SMPL-X human pose reconstruction experiment, the results follow a similar trend as GAN inversion, with massive improvement in both convergence speed and final loss values after convergence (see Fig. 4b). Our method outperforms the baseline by 19% for in-distribution evaluation and 11% for OOD evaluation.

**Qualitative Results.** In Fig. 6 we visualize how the human pose reconstructions evolve over the process of optimization. Here, we observe that the reconstructions from steps 0 to 6 of the baseline are similar for both examples. On the other hand, our method caters to fast convergence for the varying examples, highlighting the general, yet efficient properties of our search space. Further, in Fig. 7 we visualize reconstructions from partial observations where the only joints visible are that of the upper body (row 3) or lower body (row 4). We obtain a wide range of feasible possibilities for the hidden joints demonstrating the diversity of the latent space.

### 4.3 Defending Adversarial Attacks

Our method is also applicable to discriminative models. A state-of-the-art defense [37] for adversarial attack optimizes the self-supervision task at inference time, such that the model can dynamically restore the corrupted structure from the input image for robust inference. Following the existing algorithm implementation, we optimize the input image via our method by minimizing the following discriminative loss function:

$$L(F(\mathbf{r}+\mathbf{a}), y) = L(F(\theta(\mathbf{z})+\mathbf{a}), y) = \mathbb{E}_{i,j}\left[-y_{ij}^{(s)} \log \frac{\exp(\cos(\mathbf{f}_i, \mathbf{f}_j))}{\sum_k \exp(\cos(\mathbf{f}_i, \mathbf{f}_k))}\right] + \lambda||\theta(\mathbf{z})||_2^2, \quad (7)$$

where $\mathbf{a}$ is the adversarial attacks that we aim to defend against, $\mathbf{r} = \theta(\mathbf{z})$ is our additive defense vector to optimize, $\mathbf{f}_i$ are the contrastive features produced by the neural network $F$ from the $i^{\text{th}}$ image instance, and $\mathbf{y}_{ij}^{(s)}$ is the indicator for the positive pairs and negative pairs.

After obtaining the mapping network $\theta$ and the input variable $\hat{\mathbf{z}}$, the prediction is $\hat{\mathbf{y}} = F'(\mathbf{a} + \theta(\mathbf{z}))$, where $F'$ is the classification model. Note that the a self-supervision loss is optimized as a proxy

| Model Type | No Optimization | Optimization Steps: 1 step Baseline | 1 step Ours | 3 steps Baseline | 3 steps Ours | 5 steps Baseline | 5 steps Ours |
|---|---|---|---|---|---|---|---|
| RO [46] | 31.99 | 34.62 | **44.65** | 36.77 | **44.23** | 38.38 | **43.43** |
| AWP [58] | 35.61 | 39.54 | **51.39** | 42.81 | **51.67** | 44.96 | **51.05** |
| MART [55] | 35.66 | 39.77 | **51.77** | 42.50 | **51.77** | 45.42 | **50.96** |
| SemiSL [10] | 29.78 | 34.53 | **52.11** | 37.27 | **51.23** | 40.93 | **49.83** |

Table 1: Experiment on improving adversarial robust accuracy. Our goal is to defend 200 steps of $L_2 = 256/255$ norm bounded attack [36], where the attack's step size is $64/255$. Our baseline is the SOTA test-time optimization-based defense [37], which minimizes the loss of self-supervision task.


Figure 8: **Comparison Against Encoder-Based Inference. Left** shows the results on the test split of CelebA-HQ; **Right** shows the results on the LSUN-cat dataset.

for increasing the robust classification accuracy. In addition, we add a $L_2$ norm decay term for the generated noise $z$ to avoid generating reversal vector that is too large.

**Quantitative Results.** We evaluate our method on four popular pretrained robust models [46, 55, 58, 10] on CIFAR-10 dataset. The results in [37] require many steps to optimize the objective to improve the adversarial robustness, which slows down the inference by hundreds of times than the original forward pass. Ideally, we desire test-time optimization that can adapt to the attacked images in just one step, causing the minimal delay at inference time. In Table 1, our method outperforms the gradient descent method in [37] by up to 18% robust accuracy using a single step, providing a real-time robust inference method. Note that our method converges after 1 step of optimization, demonstrating the effectiveness of our approach.

### 4.4 Loss Landscape

To understand the underlying cause of the significant improvement in optimization brought by our mapping network $\theta$, we visualize in Fig. 9 the loss landscape for performing optimization in the original input space $X$, and our projected space $Z$. To generate this visualization, we first perform 20 steps of optimization on the validation dataset to collect a set of recovered latents. We then perform principle component analysis (PCA) on these recovered latents to obtain two principle directions. Finally, for individual examples, we evaluate the loss for vertices on a meshgrid spanned by the two principle directions.

From the visualization, we can see that the loss landscapes of the baseline are highly non-convex and contain points whose loss are significantly higher than its neighboring regions, while our loss landscapes are significantly smoother, with the "spikes" removed. Besides, our loss landscapes also

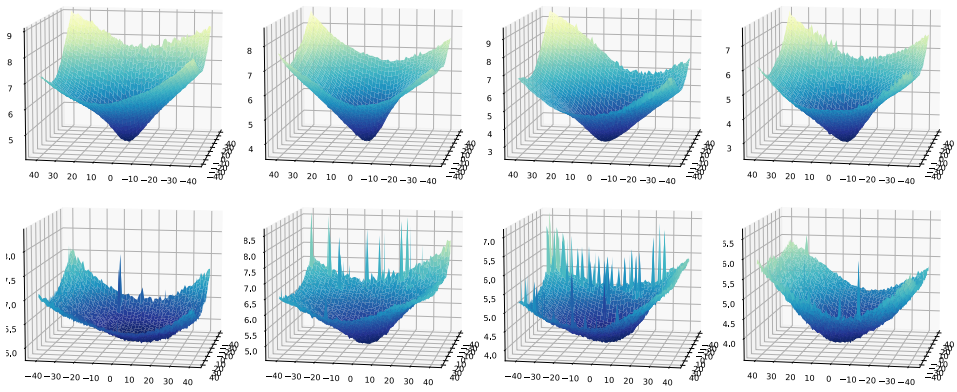


Figure 9: **Visualizing Loss Landscape (Uncurated).** Visualizing the loss landscape of StyleGAN inversion spanned by two principle directions. **Top** row shows 4 examples of the loss landscapes corresponding to our space $Z$. **Bottom** row shows the loss landscapes corresponding to the original input space $X$ for the same 4 examples.

| | Number of Steps | Full Model | Without CD and Buffer | Random $\theta$ | Baseline |
|---|---|---|---|---|---|
| In-distribution | 20 | **3.082** | 3.583 | 4.417 | 4.456 |
| OOD | 20 | **4.932** | 5.127 | 5.135 | 5.292 |
| In-distribution | 200 | **1.964** | 3.034 | 2.723 | 2.569 |
| OOD | 200 | **3.498** | 4.756 | 3.823 | 4.617 |

Table 2: **Ablation Study on Mapping Network.** Number of Steps indicates the number of optimization steps performed during inference. Evaluation metric consistent with 4.

tend to be steeper than the baseline ones. These two phenomena directly cause our method to perform gradient descent faster and stabler.

### 4.5 Ablation Study

In this section, we present an ablation study by removing the proposed coordinate descent scheme and the experience replay buffer. We also compare against a $\theta$ that is randomly initialized. From Table 2, we discovered that for in-distribution, coordinate descent and training of theta improves the optimziation performance by 14% and 30% respectively. Such gap becomes 35% and 28% for evaluation on 200 steps. For OOD data, the advantage is furthered enlarged as shown in Table 2.

One surprising result we discovered experimentally is that OBI under a randomly initialized mapping network $\theta$ consistently outperforms the baseline. We believe this is due to the fact that adding a Gaussian distribution to an underlying latent distribution of StyleGAN is beneficial in smoothing out loss landscape, making it easier to perform OBI. Similar random projection can also be found in [59].

## 5 Conclusion

This paper presents a method to accelerate optimization-based inference to invert a forward model. We propose an approach that learns a new space that is easier than the original input space to optimize with gradient descent at testing time. Our experiments and analysis on three different applications in computer vision have shown that by learning this mapping function, optimization becomes more efficient and generalizes better to out-of-distribution data. Through quantitative and qualitative analysis, we found that such improvement in optimization performance comes from a smoother loss landscape. Since optimization-based inference has many advantages over encoder-based inference, we believe methods to accelerate them will have many impacts in a variety of applications.

**Acknowledgements:** This research is based on work partially supported by the NSF NRI Award #1925157, NSF STC LEAP, the DARPA MCS program, and the DARPA CCU program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

## References


[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8296–8305, 2020.

[3] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[4] Ijaz Akhter and Michael Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. 06 2015.

[5] Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

[6] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4502–4511, 2019.

[7] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[8] Philémon Brakel, Dirk Stroobandt, and Benjamin Schrauwen. Training energy-based models for time-series imputation. *The Journal of Machine Learning Research*, 14(1):2771–2797, 2013.

[9] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy*, pages 39–57, 2017.

[10] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[11] Boyuan Chen, Robert Kwiatkowski, Carl Vondrick, and Hod Lipson. Full-body visual self-modeling of robot morphologies. *arXiv preprint arXiv:2111.06389*, 2021.

[12] CMU. Cmu mocap dataset.

[13] Jochen L Cremer, Ioannis Konstantelos, and Goran Strbac. From optimization-based machine learning to interpretable security rules for operation. *IEEE Transactions on Power Systems*, 34(5):3826–3836, 2019.

[14] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

[15] Katherine Crowson, Stella Rose Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. *ArXiv*, abs/2204.08583, 2022.

[16] Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. *arXiv preprint arXiv:2112.00719*, 2021.

[17] Justin Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pages 318–326. PMLR, 2012.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[19] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019.

[20] Carlos Hernandez, George Vogiatzis, and Roberto Cipolla. Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):548–554, 2008.

[21] Minyoung Huh, Richard Zhang, Jun-Yan Zhu, Sylvain Paris, and Aaron Hertzmann. Transforming and projecting images into class-conditional generative networks. In *European Conference on Computer Vision*, pages 17–34. Springer, 2020.

[22] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[23] Ali Jahanian, Lucy Chai, and Phillip Isola. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019.

[24] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385, 2019.

[25] Kyoungkook Kang, Seongtae Kim, and Sunghyun Cho. Gan inversion for out-of-range images with geometric transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13941–13949, 2021.

[26] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[30] Kyoung Mu Lee and C-CJ Kuo. Shape from shading with a linear triangular element surface model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(8):815–822, 1993.

[31] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *ArXiv*, abs/2104.08691, 2021.

[32] Ruoshi Liu, Sachit Menon, Chengzhi Mao, Dennis Park, Simon Stent, and Carl Vondrick. Shadows shed light on 3d objects. *arXiv*, 2022.

[33] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Haoran Su, and Qiang Liu. Fusedream: Training-free text-to-image generation with improved clip+gan space optimization. *ArXiv*, abs/2112.01573, 2021.

[34] Matthew Loper, Naureen Mahmood, and Michael J. Black. Mosh: Motion and shape capture from sparse markers. *ACM Trans. Graph.*, 33(6), nov 2014.

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[36] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[37] Chengzhi Mao, Mia Chiquier, Hao Wang, Junfeng Yang, and Carl Vondrick. Adversarial attacks are reversible with natural supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 661–671, 2021.

[38] Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[39] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. 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*, pages 2437–2445, 2020.

[40] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[41] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[42] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[43] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[44] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14104–14113, 2020.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[46] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning, 2020.

[47] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021.

[48] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020.

[49] Veselin Stoyanov, Alexander Ropson, and Jason Eisner. Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 725–733. JMLR Workshop and Conference Proceedings, 2011.

[50] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.

[51] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020.

[52] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

[53] Austin Tripp, Erik Daxberger, and José Miguel Hernández-Lobato. Sample-efficient optimization in the latent space of deep generative models via weighted retraining. *Advances in Neural Information Processing Systems*, 33:11259–11272, 2020.

[54] Hao Wang, Chengzhi Mao, Hao He, Mingmin Zhao, Tommi S Jaakkola, and Dina Katabi. Bidirectional inference networks: A class of deep bayesian networks for health profiling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 766–773, 2019.

[55] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2020.

[56] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Lu Yuan, Gang Hua, and Nenghai Yu. A simple baseline for stylegan inversion. *ArXiv*, abs/2104.07661, 2021.

[57] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

[58] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *NeurIPS*, 2020.

[59] Yan Wu, Mihaela Rosca, and Timothy Lillicrap. Deep compressed sensing. In *International Conference on Machine Learning*, pages 6850–6860. PMLR, 2019.

[60] Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5485–5493, 2017.

[61] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021.

[62] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[63] Junzhe Zhang, Xinyi Chen, Zhongang Cai, Liang Pan, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, Bo Dai, and Chen Change Loy. Unsupervised 3d shape completion through gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1768–1777, 2021.

[64] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[65] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pages 592–608. Springer, 2020.

[66] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016.

# A Algorithm

Accompanying Section 3.3, we provide a detailed coordinate ascent training algorithm with an experience replay buffer:

**Algorithm 1** Learning Mapping Network $\theta$

**Input:** Ground truth $y$, step size $\lambda_z$ and $\lambda_\theta$, number of buffers $B$, number of data samples in a buffer $N$, number of optimization steps for each sample $T$, loss function $L$, and forward model $F$.
**Output:** Learned mapping network $\theta$
Randomly initialize a mapping network $\theta$
**for** $b = 1, ..., B$ **do**
  Initialize Experience Replay Buffer $[\{\mathbf{z}_{t,i}\}_{t=1}^{T}]_{i=1}^{N}$
  **for** $i = 1, ..., N$ **do**
    $\mathbf{z}_0 \leftarrow \mathbf{0}$
    **for** $t = 1, ..., T$ **do**
      $l \leftarrow L(F(\theta(\mathbf{z})))$
      $\mathbf{z}_t \leftarrow \mathbf{z}_{t-1} + \lambda_z \frac{\partial l}{\partial \mathbf{z}_{t-1}}$
      $\mathbf{z}_{t,i} \leftarrow z_t$
    **end for**
  **end for**
  **for** $j = 1, ..., T \cdot N$ **do**
    Randomly sample $\mathbf{z}$ from previously collected buffer
    $l \leftarrow L(F(\theta(\mathbf{z})))$
    $\theta_j \leftarrow \theta_{j-1} + \lambda_\theta \frac{\partial L}{\partial \theta_{j-1}}$
  **end for**
**end for**
Return $\theta$

# B Implementation Details

We will released all code, models, and data. Here we describe our implementation details for the above algorithm.

## B.1 GAN Inversion

Mapping network $\theta$ is implemented with a 3-layer MLP. The input dimension (dimension of $Z$ space) is the same as the output dimension (dimension of $X$ space). For each intermediate output, we apply a Leaky ReLu function with a negative slope of 0.2 as an activation function. Hidden dimension of the MLP is 1024. For optimizing $z$ (collecting optimization trajectories), we use an Adam optimizer with a learning rate of 0.1. For training the mapping network $\theta$, we use an AdamW optimizer with a weight decay of 0.1 with a learning rate of 0.0001. We used the following parameter set, $T = 20$, $N = 256$, $B = 500$. For baseline, we use the implementation of [1] provided here. The pretrained weights of StyleGAN converted to PyTorch are also provided in the same link.

## B.2 3D Human Pose Reconstruction

Mapping network $\theta$ is implemented with a 3-layer MLP. The input dimension is 128 (dimension of $Z$) and the output dimension is 32 (dimension of $X$). For each intermediate output, we apply a Leaky ReLu function with a negative slope of 0.2 as an activation function. Hidden dimension of the MLP is 512. For optimizing $z$ (collecting optimization trajectories), we use an Adam optimizer with a learning rate of 0.1. For training the mapping network $\theta$, we use an AdamW optimizer with a weight decay of 0.1 and a learning rate of 0.005. We use the following parameter set, $T = 200$, $N = 40960$, $B = 500$. For a fair comparison with the baseline, we tried a range of learning rate for $\mathbf{x} \in X$ $\{0.5, 0.1, 0.05, 0.01, 0.001\}$ and select the best performing configuration for comparison.

### B.3 Defending Adversarial Attacks

Mapping network $\theta$ is implemented with a 3-layer MLP. The input dimension is 3072 (dimension of $Z$) and the output dimension is 3072 (dimension of $X$). For each intermediate output, we apply a Leaky ReLu function with a negative slope of 0.2 as an activation function. Hidden dimension of the MLP is 3072. For optimizing $z$ (collecting optimization trajectories), we use an Adam optimizer with learning rate $0.2/255$. For training the mapping network $\theta$, we use an AdamW optimizer with learning rate of 0.0001 and a weight decay of 0.1. We use the following parameter set, $T = 5$, $N = 5120$, $B = 70$. For the regularization term $\lambda$ that constrains the amplitude of the additive defense vector, we use $\lambda = 1$. We use random start instead of zero start for initializing the attack reversal vector.

## C Limitations

Optimization-based inference has intrinsic advantages to robustness, accuracy, and flexibility, which comes at the cost of additional computation time during inference. Encoder-based methods will usually perform faster because they only require a single forward pass of a neural network, while our approach requires several computational passes in both the forward and backward (gradient) direction. We believe that for many applications this trade-off will be desirable, especially in cases where accuracy is more important than speed. Our approach aims to minimize this additional computational overhead brought by optimization-based inference, and our experiments on multiple datasets show the significant computational savings compared to other optimization-based inference methods.

Unlike many other optimization-based inference algorithms, our approach also requires a training step in order to fit a suitable landscape, which requires both training time and training data. However, we believe this overhead is insignificant for most applications and we have designed our neural networks to be efficient. For example, $\theta$ is relatively lightweight, making its training time fairly marginal compared to the training of the forward model $F$. In all our experiments, we found that the training time of $\theta$ is orders of magnitudes faster than the training time for $F$.

## D Societal Impact

Optimization-based inference has a wide variety of applications broadly in computer vision, robotics, natural language processing, assistive technology, security, and healthcare. Since our proposed method provides significant acceleration to these inference techniques, we expect our work to find positive impact in these applications, where speed, accuracy, and robustness are often critical.

Our algorithm is compatible with many different types of forward models – as long as a gradient can be calculated – including neural networks. However, learned forward models are known to acquire and contain biases from the original dataset. Our approach will also inherit the same biases. While our approach to inference offers some advantages to out-of-distribution data, it does not mitigate nor correct biases. The datasets in our experiments are likely not representative of the population, and consequently also contain biases. We acknowledge these limitations, and applications of this method should be mindful of these limitations, especially in potentially sensitive scenarios. As the community continues to address biases in models and datasets, our method can be applied to these models in the future too.