# A Diffusion Model with State Estimation for Degradation-blind Inverse Imaging

**Liya Ji**[1*], **Zhefan Rao**[1*], **Sinno Jialin Pan**[2], **Chenyang Lei**[3†], **Qifeng Chen**[1†]

[1] Hong Kong University of Science and Technology
[2] The Chinese University of Hong Kong
[3] CAIR, HKISI-CAS
{lji, zraoac}@connect.ust.hk, sinnopan@cuhk.edu.hk, leichenyang7@gmail.com, cqf@ust.hk

## Abstract

Solving the task of inverse imaging problems can restore unknown clean images from input measurements that have incomplete information. Utilizing powerful generative models, such as denoising diffusion models, could better tackle the ill-posed issues of inverse problems with the distribution prior of the unknown clean images. We propose a learnable state-estimator-based diffusion model to incorporate the measurements into the reconstruction process. Our method makes efficient use of the pre-trained diffusion models with computational feasibility compared to the conditional diffusion models, which need to be trained from scratch. In addition, our pipeline does not require explicit knowledge of the image degradation operator or make the assumption of its form, unlike many other works that use the pre-trained diffusion models at the test time. The experiments on three typical inverse imaging problems (both linear and non-linear), inpainting, deblurring, and JPEG compression restoration, have comparable results with the state-of-the-art methods.

## Introduction

Computational photography strives to produce visually pleasing images that faithfully depict the original scenes they represent (Ongie et al. 2020). However, due to physical limitations such as loss of details during image transmission or out-of-focus when capturing the photograph, the measurements we obtain may only contain inaccurate or incomplete information. A wide range of problems, such as deblurring, deraining, or JPEG compression restoration, at the heart of computational photography, reduces to the crucial task of solving inverse imaging problems, which aims to reconstruct unknown images from given measurements (Diamond et al. 2017).

Measurements are usually obtained through an image degradation operator from unknown clean images in the setting of inverse imaging problems. But it is difficult that an inverse imaging problem is always ill-posed, so multiple images can fit a measurement even though they do not look like natural images at all. Utilizing the deep generative prior is beneficial to solve the ill-posed inverse problems (Pan et al. 2020) by adding the constraints to improve the quality of reconstructed images. Due to the impressive generative ability brought by denoising diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2020; Song, Meng, and Ermon 2020; Rombach et al. 2021), applying diffusion models for inverse problems may enable us to fill in the missing information with a more powerful knowledge prior about the distribution of the unknown clean images.

There are two directions in the field of diffusion-based inverse problems. Firstly, some works (Saharia et al. 2022b,a; Rombach et al. 2021) formulate inverse imaging problems as conditional diffusion models and convert the measurements into the condition embeddings of the diffusion models. However, this direction is computationally expensive since the diffusion models must be trained from scratch when a new inverse imaging task comes. Secondly, some works (Chung, Sim, and Ye 2022; Lugmayr et al. 2022; Kawar et al. 2022a; Chung et al. 2022c; Kawar et al. 2022b; Chung et al. 2022b; Wang, Yu, and Zhang 2022) adjusts the intermediate state of pre-trained diffusion models at test time by utilizing the image degradation operator. This direction fails in scenarios where the image degradation operator is unknown. In addition, their reconstruction process is not convenient for real-world applications, especially for non-deterministic inverse imaging problems, such as motion deblurring and inpainting, as their approaches require the motion kernels or inpainting masks as input. There are some works (Chung et al. 2022a; Ben Fei 2023) to tackle the degradation-blind inverse imaging problems with diffusion models. These approaches need the assumption of the form of degradation operators. On the contrary, our approach is degradation-blind, which does not need image degradation information as a prior.

Our key design is proposing a State Estimator to learn the way to incorporate the noised measurements into the intermediate state of diffusion models at every timestep. More specifically, the State Estimator automatically estimates the weighting mask for the noised measurements and the intermediate state of diffusion models. The output of the State Estimator varies from different states in order to guide the diffusion model to the desired region on the data manifold.

With the design of the State Estimator, our model has two advantages over existing baselines. Firstly, we use less data and time to train the model compared to the conditional dif-

[*]These authors contributed equally.
[†]Corresponding authors

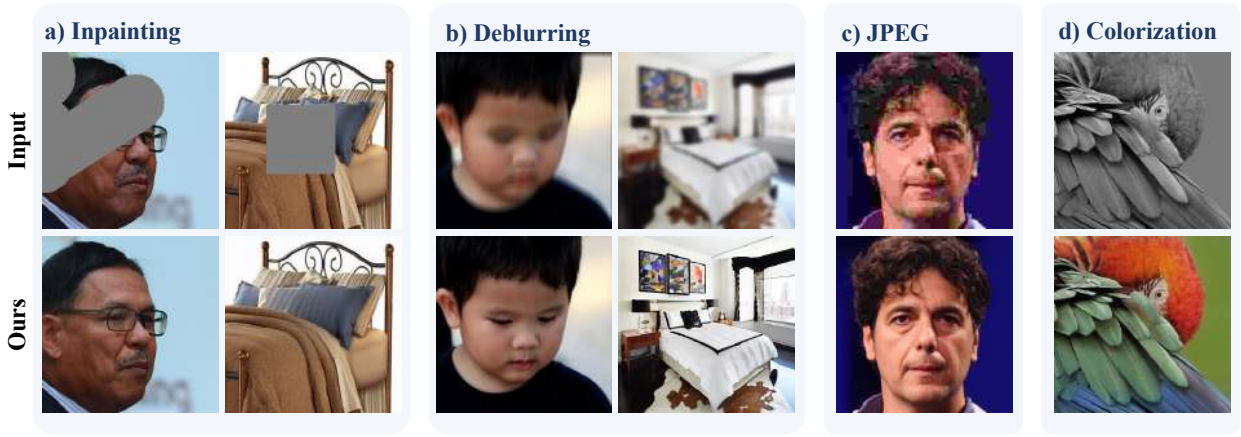


Figure 1: **The results of our state-estimator-based diffusion model for inverse imaging problems**. The image degradation kernel is not needed in our model. We show some examples of input images and reconstructed images for the typical inverse problems: (a) Inpainting, (b) Deblurring, (c) JPEG Compression Restoration, and (d) Colorization.

fusion model. We decouple the condition from the diffusion model with a learnable State Estimator so that we can directly finetune the diffusion models trained on the unknown clean data domain, which saves lots of computational resources compared to the way of training from scratch. Besides, we only need to let State Estimator learn the knowledge of the image degradation operator instead of the data distribution of clean images. Therefore, the training data used in our framework is much less than the pipeline of training conditional diffusion models. Secondly, our model can be applied to more inverse imaging problems where the image degradation operator is totally unknown or non-linear. As a comparison, most existing work needs to know the degradation form, and some works are limited to the generalization of other tasks due to the limits of imputation methods. For example, DDNM (Wang, Yu, and Zhang 2022) limits their work to linear inverse problems due to the use of SVD decomposition.

We conduct our experiments on three typical inverse imaging problems, deblurring, inpainting, and JPEG compression restoration. We evaluate our models on two standard datasets, FFHQ (Karras, Laine, and Aila 2019) and LSUN-Bedroom (Yu et al. 2015), which are comparable to current state-of-the-art models for inverse problems. Our contributions can be summarized as follows:

- We have developed a degradation-blind framework for solving inverse imaging problems using a pre-trained diffusion model. This framework can effectively solve both linear and non-linear inverse imaging problems with only a few paired samples without requiring explicit knowledge of the image degradation operator or making assumptions about its form.
- We propose a state estimation strategy for pixel-wise control and produces generative results that are semantically consistent and preserve details.
- We show the effectiveness of our model on three typical inverse imaging tasks: inpainting, deburring, JPEG compression restoration, and colorization. Without any task-specific design, our approach achieves state-of-the-art results with low latency on all tasks.

## Background

An inverse imaging problem aims to reconstruct the unknown clean image $\mathbf{x}$ from the partial measurements $\mathbf{y}$. The inverse problem can be modeled as

$$\mathbf{y} = \mathcal{A}(\mathbf{x}) + \mathbf{n}, \tag{1}$$

where $\mathcal{A}(\cdot)$ is an image degradation operator and $\mathbf{n}$ is the noise sampled from an unknown distribution (Ongie et al. 2020). The inverse problem is ill-posed. The mapping from $\mathbf{x}$ to $\mathbf{y}$ under the setting of the inverse model is many-to-one. Therefore, we need the image prior $p(\mathbf{x})$ to guarantee the quality of the results of restoration.

Traditional approaches to inverse imaging problems are mostly based on images priors, including the dark channel prior (He, Sun, and Tang 2010), the deep image prior (Ulyanov, Vedaldi, and Lempitsky 2018), Markov random fields (Zhu and Mumford 1997; Geman and Geman 1984; Roth and Black 2005), and GAN-based priors (Albright and McCloskey 2019; Creswell and Bharath 2018; Karras, Laine, and Aila 2019; Brock, Donahue, and Simonyan 2018; Zhu et al. 2016; Donahue, Krähenbühl, and Darrell 2016).

Diffusion-based inverse problems have begun to thrive with the impressive generative ability brought by the denoising diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2020; Song, Meng, and Ermon 2020; Rombach et al. 2021). One direction in diffusion-based inverse problems is conditional diffusion models (Saharia et al. 2022b,a; Rombach et al. 2021), which take the noised measurements as the condition embeddings for the diffusion models. Palette (Saharia et al. 2022a) trains the model on pairs of $\{\mathbf{x}, \mathbf{y}\}_N$ with the loss:

$$E_{\mathbf{x}_0,t,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[||\epsilon - \epsilon_\theta(\mathbf{x}_t, \sqrt{\overline{\alpha}_t}\mathbf{y} + \sqrt{(1-\overline{\alpha}_t)}\epsilon, t)||_2^2], \tag{2}$$

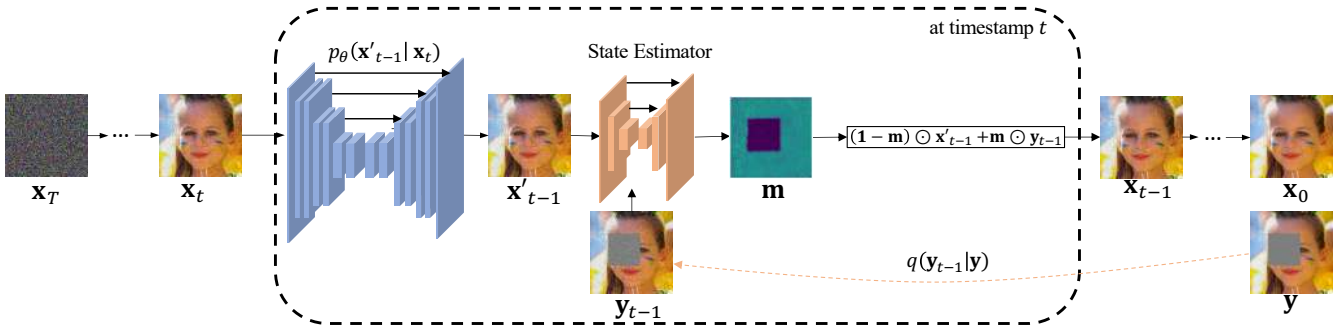


Figure 2: **Overview of our state-estimator-based diffusion model**. The output of state estimator is $\mathbf{m} \in \mathbb{R}^{h\times w\times 1}$, whose values belong to $[0, 1]$. During the training process, the parameters $\theta$ of the diffusion denoising model will be fine-tuned, and the state estimator will be learned from scratch.

where $\overline{\alpha}_t$ is the cumulative variance schedule (Ho, Jain, and Abbeel 2020). However, this direction is computationally expensive since they need to train the diffusion models from scratch conditioning on the measurements.

Another direction (Choi et al. 2021; Chung, Sim, and Ye 2022; Lugmayr et al. 2022; Kawar et al. 2022a; Chung et al. 2022c; Kawar et al. 2022b; Chung et al. 2022b; Wang, Yu, and Zhang 2022; Zhu et al. 2023; Mardani et al. 2023; Song et al. 2022) is directly adapting the pre-trained diffusion model to inverse problems with the imputation of measurements. Most of these methods need to know the image degradation operator $\mathcal{A}(\cdot)$ at the test time. Denote that $\mathbf{x}'_{t-1}$ is the output of the reverse process at step $t$. RePaint (Lugmayr et al. 2022) is a simple and effective work with pixel-wise control to address the inpainting task. They directly use the mask information $\mathbf{m}$ to incorporate the measurements by

$$\mathbf{x}_{t-1} = \mathbf{m}\odot\mathbf{y}_{t-1} + (\mathbf{1}-\mathbf{m})\odot(\mathbf{x}'_{t-1}). \tag{3}$$

However, they need to use the resampling mechanism to boost the performance so that the inference time is significantly increased. DDRM (Kawar et al. 2022a,b) and its follow-up work (Wang, Yu, and Zhang 2022) decompose the measurement operator by SVD. MCG (Chung et al. 2022c) and its follow-up work DPS (Chung et al. 2022b) utilize the gradients of $||\mathcal{A}(\hat{\mathbf{x}}_0)-\mathbf{y}||_2^2$ to adjust the intermediate state of reversion process, where $\hat{\mathbf{x}}_0$ is calculated by the Tweedie's formula (Robbins 1992).

There are some works (Song et al. 2021; Chung et al. 2022a; Ben Fei 2023) to tackle the degradation-blind inverse imaging problems with diffusion models. Chung *et al.* tries to solve the inverse problems in deblurring and imaging through turbulence. GDP (Ben Fei 2023) restores the measurements by optimizing the degradation models as well as changing the intermediate state during the denoising process. However, these methods need to know the form of the degradation operator. Recently, some controllable methods (Zhang and Agrawala 2023; Ma et al. 2023; Chefer et al. 2023) for generative methods have been proposed. ControlNet (Zhang and Agrawala 2023), is similar to our approach by finetuning the diffusion models. However, ControlNet is incapable of maintaining the details of the partial measurements.

**Latent diffusion models** Instead of working diffusion models directly on the original image RGB space, latent diffusion models (Rombach et al. 2021) first encode the image into latent space and decode the generated vector into the original space after the diffusion reverse process starting from a Gaussian noise $\mathbf{z}_T$ in the latent space. More specifically, let us assume that we have an encoder $\xi$ and a decoder $\mathcal{D}$. An RGB image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ will be encoded into the latent space $\mathbf{z} \in \mathbb{R}^{h\times w\times c}$, where $\mathbf{z} = \xi(\mathbf{x})$ and $\widetilde{\mathbf{x}} = \mathcal{D}(\mathbf{z})$. In our work, $h = \frac{H}{f}$ and $w = \frac{W}{f}$, where $f = 4$. The value of $c$ equals 3. Therefore, the loss becomes:

$$L_{LDM} = E_{\xi(\mathbf{x}_0),t,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[||\epsilon - \epsilon_\theta(\xi(\mathbf{x}_t), t)||_2^2]. \tag{4}$$

## Method

Given a measurement $\mathbf{y}$, our approach aims to learn a function $f(\cdot)$ to predict the corresponding $\mathbf{x}$ in Equation 1. Different from previous work, in our setting, the image degradation operator $\mathcal{A}(\cdot)$ is unknown both for the training and testing period. Our model learns the knowledge of $\mathcal{A}$ from the training data, consisting of pairs of $\{\mathbf{x}, \mathbf{y}\}$, where $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$.

We utilize a learned probability distribution $p_\theta(\mathbf{x})$ for the data domain $\mathcal{X}$ we aim to reconstruct. Instead of artificially designing a mechanism to incorporate the noised measurements, we propose a State Estimator to learn the way for the imputation. The State Estimator extends the flexibility and generality for different inverse problems and adapts the conditioning based on the current intermediate states of diffusion models and the noised measurements.

Figure 2 shows the whole overview of our pipeline. The structure of the State Estimator (SE) consists of a deep neural network followed by a normalization function. The output of the state-estimator-based model $\mathbf{m}$ varies from different states of the reverse process. Firstly, we will introduce our conditioning method via the state estimator to gain a comprehensive understanding of the reconstruction process. Then we describe the training framework we use and delve into the loss function that we have customized in accordance with the setting of diffusion models.

**Conditioning via State Estimator** We start the iteration from the $\mathbf{x}_T \in \mathbb{R}^{h\times w\times 3}$ in the latent space, which is sampled from Standard Normal Distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. We obtain our reconstructed image $\mathbf{x}_0$ after $T$ iterative steps. For ease of convenience, we use $\mathbf{x}$ as the variable in the latent space instead of $\mathbf{z}$. Assume that we arrive at step $t \in T, ..., 1$, we first generate the noise version $\mathbf{y}_{t-1}$ of the latent space of the measurement $\mathbf{y} \in \mathcal{Y}$:

$$\mathbf{y}_{t-1} \sim \mathcal{N}(\sqrt{\overline{\alpha}_{t-1}}\xi(\mathbf{y}), (1-\overline{\alpha}_{t-1})\mathbf{I}). \tag{5}$$

In addition, we obtain the unconditional intermediate state $\mathbf{x'}_{t-1}$ by

$$\mathbf{x'}_{t-1} \sim \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)). \tag{6}$$

A learned State Estimator $SE(\cdot)$ take $\mathbf{x'}_{t-1}$, $\mathbf{y}_{t-1}$, and $t$ as the inputs, and output a state mask $\mathbf{m_t} \in \mathbb{R}^{h\times w\times 1}$, of which each value belongs to the range $(0, 1)$:

$$\mathbf{m} = SE(\mathbf{x'}_{t-1}, \mathbf{y}_{t-1}, t). \tag{7}$$

$SE$ is a neural network with the output layer as $\frac{tanh+1}{2}$ so that its output is in the space $(0, 1)^{h\times w\times 1}$. Then an ad-hoc, called relaxing boundary, is applied to extend the space into $[0, 1]^{h\times w\times 1}$. We generate the new intermediate state $\mathbf{x}_{t-1}$ for next iteration by

$$\mathbf{x}_{t-1} = (\mathbf{1}-\mathbf{m}) \odot \mathbf{x'}_{t-1} + \mathbf{m} \odot \mathbf{y}_{t-1}. \tag{8}$$

More specifically, we use the naive version of U-net (Ronneberger, Fischer, and Brox 2015) as our $SE$ function. Our pipeline requires that the architecture of the state estimator should keep the dimensions of the input unchanged, and we want the state estimator to learn a weighting mask. Additionally, we only use 1k-2k paired images as our training data for computational efficiency. The property of U-net, capable of generating high-quality segmentation results with training on a few samples (Ronneberger, Fischer, and Brox 2015), fits our requirements of the state estimators.

In addition, we claim that the given measurements lose some information in the field of inverse imaging. For example, in the deblurring (Gaussian) task, the details of the original images have been lost. In the inpainting task, the content of the masked area is missing. The intuitive way to let the diffusion prior $\mathbf{x}_t$ fill in the missing information is the linear combination with the given measurements. In other words, for deblurring (Gaussian), the measurements provide the structure of the image, and the diffusion prior $\mathbf{x}_t$ provides the details of the image, which is verified at the visualization of the outputs of the state estimator.

**Training framework and loss function** Instead of directly using the latent diffusion loss to measure the noise in each step in Equation 4, we propose a new training framework in Algorithm 1, with the loss of calculating the distances between the restored images and the ground truth images, which yields better and more stable performance.

Due to the long sequence property, normally at least 50, of diffusion models, the restored images must be estimated on the intermediate state $\mathbf{x}_t$. As introduced in DDIM (Song, Meng, and Ermon 2020), we can predict the unknown clean image $\mathbf{x}_0$ based on the intermediate state $\mathbf{x}_t$ according to the equation:

$$\hat{\mathbf{x}}_0 = \frac{\mathbf{x}_t - \sqrt{1-\overline{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{\overline{\alpha}_t}}, \tag{9}$$

where $\epsilon_\theta(\cdot)$ stands for the diffusion function approximator, which predict noise $\epsilon$ from $\mathbf{x}_t$. Due to the inaccurate prediction of $\mathbf{x}_0$ in the early stage of the diffusion reverse process, we add a reweight function $\Upsilon_t(\cdot)$ to the loss in order to decrease the penalty when $t$ is large. For each batch $\{\mathbf{x}, \mathbf{y}\}_N$, the loss will be

$$L = E_{\mathbf{x},t}[\Upsilon_t(\mathcal{L}(\mathcal{D}(\hat{\mathbf{x}}_0), \mathbf{x}))], \tag{10}$$

where $\mathcal{D}$ is the decoder in latent diffusion models. $\mathcal{L}$ is the weighted loss of MSE and LPIPS.

Although we use the predicted $\mathbf{x}_0$ to calculate the loss, we also face gradient vanishing or exploding issues during the training process. More specifically, we use Truncated Back Propagation Through Time (TBPTT) (Williams and Zipser 1995; Sutskever 2013) to tackle these issues. Algorithm 1 illustrates the details. Denote $\eta$ to be the truncated steps we set. The parameters we optimize consist of State Estimator $SE_\tau(\cdot)$ and the diffusion model $p_\theta(\cdot)$. Within a batch, similar to the diffusion training, we randomly sample a timestep $t_0 \in$ Uniform $(\eta, ..., T)$. Then we derive $\mathbf{x}_{t_0-\eta}$ from $\mathbf{x}_T \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ using the latest parameters of models and detach the nodes before $\mathbf{x}_{t_0}$. Therefore, only the parameters falling in steps $[t_0, t_0 - \eta]$ will be optimized.

Algorithm 1: Training

**Input**: Training samples $\{\mathbf{x}, \mathbf{y}\}_N$, truncated steps $\eta$, an encoder $\xi$, a decoder $\mathcal{D}$, diffusion models $p_\theta(\cdot)$ / $\epsilon_\theta(\cdot)$ , State Estimator $SE_\tau(\cdot)$, a reweight function $\Upsilon_t(\cdot)$

**repeat**

$t_0 \sim$ Uniform$(\eta, ..., T)$

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$t_a,\ t_b = t_0,\ t_0 - \eta$

**for** $t = T, ..., t_b + 1$ **do**

$\mathbf{x'}_{t-1} \sim p_\theta(\mathbf{x'}_{t-1}|\mathbf{x}_t)$

$\mathbf{y}_{t-1} \sim q(\mathbf{y}_{t-1}|\xi(\mathbf{y}))$

$\mathbf{m} = SE_\tau(\mathbf{x'}_{t-1}, \mathbf{y}_{t-1}, t)$

$\mathbf{x}_{t-1} = (\mathbf{1}-\mathbf{m}) \odot \mathbf{x'}_{t-1} + \mathbf{m} \odot \mathbf{y}_{t-1}$

**end for**

$\hat{\mathbf{x}}_0 = \frac{\mathbf{x}_{t_b} - \sqrt{1-\overline{\alpha}_{t_b}}\epsilon_\theta(\mathbf{x}_{t_b}, t_b)}{\sqrt{\overline{\alpha}_{t_b}}}$

Only take gradient descent step between steps $[t_a, t_b]$ on $\nabla_{\theta,\tau}\ \Upsilon_t(\mathcal{L}(\mathcal{D}(\hat{\mathbf{x}}_0), \mathbf{x}))$

**until** converged

# Experiments

## Experimental settings

**Pretrained models and datasets** We utilize the pretrained diffusion model to reconstruct the original images from the measurements. Specifically, we use two unconditional models of latent diffusion models (Rombach et al. 2021; Blattmann et al. 2022) trained on FFHQ

(a) Input (b) DDRM (c) RePaint (d) DPS (e) Ours (f) Ground truth

Figure 3: **Examples for inpainting task on FFHQ 256 $\times$ 256 and LSUN-Bedroom**. Our method is able to complete the measurements with semantic consistency as well as preserve the details of the unmasked areas.

| Methods | Degradation Blindness | Modeling | FFHQ | | | LSUN-Bedroom | | | Latency (s) |
|---|---|---|---|---|---|---|---|---|---|
| | | | Votes ↑ | LPIPS↓ | FID↓ | Votes ↑ | LPIPS↓ | FID↓ | |
| MAT (Li et al. 2022) | ✔ | GAN | 36.2% | 0.0359 | 14.595 | - | - | - | - |
| PUT (Liu et al. 2022) | ✔ | Transformer | 42.6% | 0.0346 | 13.921 | - | - | - | - |
| DDRM (Kawar et al. 2022a) | ✘ | Diffusion | 17.6% | 0.0579 | 17.197 | 34.6% | 0.1175 | 29.186 | 8.37 |
| Repaint (Lugmayr et al. 2022) | ✘ | | 32.6% | 0.0445 | 12.504 | 54.2% | 0.0816 | 15.540 | 272.45 |
| DPS (Chung et al. 2022b) | ✘ | | 31.6% | 0.0978 | 29.521 | 53.0% | 0.1587 | 22.231 | 87.33 |
| Ours | ✔ | Diffusion | Reference (50%) | 0.0530 | 12.923 | Reference (50%) | 0.1074 | 16.472 | 5.32 |

Table 1: **Quantitivate results of inpainting task on FFHQ 256 $\times$ 256 and LSUN-Bedroom**. We have conducted the pair-wise comparison with our method as the reference. Votes stand for the ratio of votes compared with our method. The *p-value* of our user study is smaller than 0.05. Our method has the lowest latency among all the diffusion-based baselines.

256×256 (Karras, Laine, and Aila 2019), and LSUN-bedroom 256×256 (Yu et al. 2015). We evaluate our performance on a held-out dataset with 1000 images, which were randomly sampled from the validation datasets of FFHQ and LSUN-bedroom, respectively. We choose three representative tasks, Inpainting, Deblurring, and JPEG Compression Restoration in inverse problems, to illustrate the effectiveness of our methods.

**Implementation details** The training pairs for each task in the FFHQ dataset and LSUN-bedroom is 1000 and 2000, respectively. We train the State Estimator from scratch and fine-tune the diffusion model. The total step is around 75k for FFHQ and 150k for LSUN-Bedroom with batch size 4. We train and evaluate our model on NVIDIA A100 GPU cards. The learning rate is 0.001 for the State Estimator and $10^{-6}$ for the finetuning of the diffusion model. The hyperparameter *eta* for the diffusion model is set to 3.0 both for training and inference. The truncated step $\eta$ equals 3.

**Evaluation Metrics** We report the commonly used metrics, the peak-signal-to-noise ratio (PSNR), LPIPS (Zhang et al. 2018) (AlexNet (Krizhevsky, Sutskever, and Hinton 2017)), and FID (Heusel et al. 2017) as the quantitative metrics for all of the inversion tasks.

In addition, we also conducted a user study of 100 images for the inpainting task. We randomly chose 50 images from the held-out validation datasets of FFHQ 256 $\times$ 256 and LSUN-Bedroom, respectively. We use MTurk to conduct the user study. We have set the experiments as the pairwise comparison with our method as the reference by asking the question *'Which image is more realistic and preserves more details of the given pixels?'*. 10 workers rate every question. Therefore, there are a total of 2500 votes, with each baseline totaling 500. Table 1 shows the results of our user study. The *p-value* of our user study is smaller than 0.05, which indicates a significant convince of our results.

## Experimental results

**Inpainting** We compare our method with diffusion-based models, DPS (Chung et al. 2022b), Repaint (Lugmayr et al. 2022), DDRM (Kawar et al. 2022a) using the public pretrained weights on FFHQ and LSUN-bedroom datasets. We

| Methods | Gaussian Blur | | | | | | Motion Blur | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FFHQ | | | LSUN-Bedroom | | | FFHQ | | | LSUN-Bedroom | | |
| | PSNR↑ | LPIPS↓ | FID↓ | PSNR↑ | LPIPS↓ | FID↓ | PSNR↑ | LPIPS↓ | FID↓ | PSNR↑ | LPIPS↓ | FID↓ |
| DeblurGANV2 (Kupyn et al. 2019) | 26.30 | 0.1903 | 41.053 | 24.88 | 0.1621 | 25.177 | 15.75 | 0.4282 | 62.565 | 12.87 | 0.5954 | 61.276 |
| MPRNet (Zamir et al. 2021) | 33.42 | 0.0976 | 34.927 | 31.89 | 0.0987 | 19.494 | 20.03 | 0.3359 | 69.998 | 13.03 | 0.5481 | 36.048 |
| DPS (Chung et al. 2022b) | 25.84 | 0.1279 | 27.256 | 23.59 | 0.2117 | 22.491 | 25.27 | 0.1371 | 45.553 | 25.10 | 0.1981 | 50.916 |
| BlindDPS (Chung et al. 2022a) | 25.75 | 0.2476 | 31.805 | 21.02 | 0.3460 | 30.820 | 21.49 | 0.2244 | 33.675 | 19.69 | 0.3118 | 30.894 |
| GDP (Ben Fei 2023) | 25.69 | 0.1249 | 26.295 | 22.97 | 0.2102 | 20.847 | - | - | - | - | - | - |
| Ours | 27.98 | 0.0939 | 25.453 | 28.70 | 0.0700 | 23.047 | 25.58 | 0.1112 | 25.547 | 23.81 | 0.1461 | 20.659 |

Table 2: Quantitative results of deblurring task on FFHQ 256 $\times$ 256 and LSUN-Bedroom.

(a) Input (b) DPS (c) Ours (d) GT

Figure 4: **Examples for deblurring task on FFHQ 256 $\times$ 256 and LSUN-Bedroom**. Our method can preserve more details of the clean images based on blurred images.

(a) Input (b) DDRM (c) Ours (d) GT

Figure 5: **Examples for JPEG compression restoration on FFHQ 256 $\times$ 256 and LSUN-Bedroom**. Our methods have better quality compared to the baseline in the setting of quality factor 5.

also add the GAN-based method, MAT (Li et al. 2022), as the inpainting baselines. Except for the methods utilizing the deep generative prior, we add PUT (Liu et al. 2022), the method of training an end-to-end framework via transformers for inpainting tasks. Since all of the GAN-based and transformer methods don't provide the pre-trained weights on LSUN-bedroom datasets, we compare our methods with them only on the FFHQ dataset. The box size is set to $100 \times 100$, and the position of the box is randomly assigned both for the training and testing. The pixels inside the box are filled with Standard Gaussian noise. Unlike Palatte (Saharia et al. 2022a) only considers the loss of the pixels inside the mask, we consider the loss for the whole image since the mask is not used for the whole process.

The quantitative results of the inpainting task are shown in Table 1. We also compare the inference time with other diffusion-based methods using an NVIDIA RTX 3090 GPU in Table 1. Some representative reconstruction results are demonstrated in Figure 3. Our method can complete the measurements with semantic consistency for accessories and keep the details of the unmasked areas.

**Deblurring** We compare our methods with DeblurGANV2 (Kupyn et al. 2019), MPRNet (Zamir et al. 2021), DPS (Chung et al. 2022b), BlindDPS (Chung et al. 2022a), and GDP (Ben Fei 2023). For DPS, BlindDPS, and GDP, We use the public pre-trained weights. We have trained MPRNet and DeblurGANV2 using our same training dataset, containing pairs of blurred images and clean images. For Gaussian Blur, we use Gaussian Blur with the kernel size 16 and sigma 2.6. For Motion Blur, we follow the scripts used by DPS (Chung et al. 2022b), with the kernel size 61 and the intensity value 0. We show the quantitative results in Table 2. We also show some representative results in Figure 4. Our method preserves more details of the original images compared to the results of DPS. The results of DeblurGANV2 and MPRNet have artifacts, leading to low PSNR scores.

**JPEG Compression Restoration** JPEG compression restoration is important because of the effectiveness and popularity of the JPEG compression algorithm. The degradation process of JPEG compression is non-linear. We compare our method with the basic JPEG decoder as JPEG, QGAC (Ehrlich et al. 2020), and DDRM-JPEG (Kawar et al. 2022b) as DDRM for short. For QGAC and DDRM, we use the pre-trained weights. We train our model on the pairs with quality factor 5 and evaluate the performance for the quality factor, both 5 and 10, using the same model weights. We show the quantitative results in Table 3. We also show some

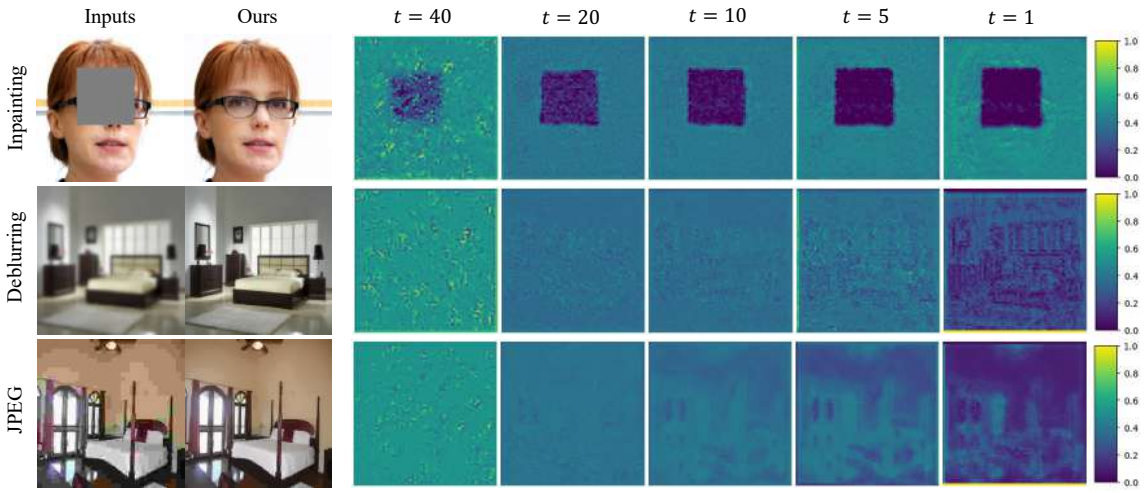


Figure 6: **Visualization of State Estimators during the reversion process**. The **yellow** indicates that the estimator prefers to pick the values from the noise version of measurements, and the **dark blue** indicates that the State Estimator prefers to pick the values from the intermediate state of the diffusion process.

| QF | Method | FFHQ | | LSUN-Bedroom | |
|---|---|---|---|---|---|
| | | PSNR↑ | LPIPS↓ | PSNR↑ | LPIPS↓ |
| 5 | JPEG | 24.35 | 0.3961 | 23.74 | 0.3303 |
| | QGAC (Ehrlich et al. 2020) | 24.99 | 0.3049 | 24.02 | 0.2761 |
| | DDRM (Kawar et al. 2022b) | 26.25 | 0.2000 | 26.02 | 0.2510 |
| | Ours | 25.43 | 0.1237 | 24.08 | 0.1682 |
| 10 | JPEG | 27.37 | 0.2230 | 26.56 | 0.2055 |
| | QGAC (Ehrlich et al. 2020) | 29.47 | 0.1810 | 28.63 | 0.1727 |
| | DDRM (Kawar et al. 2022b) | 28.91 | 0.1341 | 28.43 | 0.1495 |
| | Ours | 26.37 | 0.1068 | 25.67 | 0.1347 |

Table 3: Quantitative results of JPEG compression restoration task on FFHQ 256 $\times$ 256 and LSUN-Bedroom.

reconstruction results in Figure 5. Our methods have better quality compared to the baselines both for quality factors 5 and 10. In addition, the reason why QGAC is slightly better than JPEG for QF=5 is that the lowest value of Quality Factor QGAC can handle is 10 in the paper.

## Visualization of State Estimators

Our principle approach is letting the neural network learn how to incorporate the measurements into the sampling process. We show the visualization of the learned State Estimator through timesteps in Figure 6. From the visualization, we can see that the estimator picks the value evenly from both sides in the early stage of the reverse process. But in the last several timesteps, the measurements can only provide partial information for the reconstructed images. Interestingly, for deblurring, we could see a rough sketch of images with dark blue, proving that the diffusion model provides the details of missing information from the measurement. This estimator and the format of linear combination provide us with a new point of view to analyze the choice of State Estimator during the image restoration process.

| Task types | Conditioning methods | |
|---|---|---|
| | Concatenation | State Estimator |
| Inpainting | 0.0681 | 0.0530 |
| Gaussian Blur | 0.0848 | 0.0939 |
| Motion Blur | 0.1394 | 0.1112 |
| JPEG (QF=5) | 0.1425 | 0.1237 |
| JPEG (QF=10) | 0.1255 | 0.1068 |

Table 4: **Ablation study on the conditioning methods**. We report the LPIPS↓ score on FFHQ 256 $\times$ 256 datasets.

## Ablation Study on Conditioning Methods

We conduct the ablation study on the way of conditioning on the measurement $\mathbf{y}$. The controlled experiment is that we directly concatenate $\mathbf{y}$ with the intermediate state $\mathbf{x}_t$ along the channel. The parameters of channels concatenated are initialized with zeros and trained from scratch. We evaluate the LPIPS score for three inverse imaging tasks on FFHQ testing datasets. From Table 4, we conclude that our state estimator can achieve high-quality results while keeping the details of the measurement $\mathbf{y}$ on most of the inverse imaging tasks in this paper.

# Conclusion

We propose a unified framework for solving inverse imaging problems using a learnable State Estimator, which automatically controls the imputation of noised measurements into the reconstruction process. Our methods show effectiveness on three inverse tasks: inpainting, deblurring, and JPEG compression restoration. However, the solution space of the current method is still limited to the generative ability of diffusion models. The future work is to transfer our framework to text-to-image diffusion models or utilize domain adaptation techniques to extend the limitation.

## Acknowledgements

This project was supported by the National Key R & D Program of China under grant number 2022ZD0161501. We are also grateful for the useful discussion of Jiapeng Zhu, Qiang Wen, and Yunhao Gou for this project.

## References

Albright, M.; and McCloskey, S. 2019. Source Generator Attribution via Inversion. In *CVPR Workshops*, volume 8.

Ben Fei, L. P. J. Z. W. Y. T. L. B. Z. B. D., Zhaoyang Lyu. 2023. Generative Diffusion Prior for Unified Image Restoration and Enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Blattmann, A.; Rombach, R.; Oktay, K.; and Ommer, B. 2022. Retrieval-Augmented Diffusion Models.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–10.

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.; Kim, S.; and Ye, J. C. 2022a. Parallel Diffusion Models of Operator and Image for Blind Inverse Problems. *arXiv preprint arXiv:2211.10656*.

Chung, H.; Kim, J.; Mccann, M. T.; Klasky, M. L.; and Ye, J. C. 2022b. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*.

Chung, H.; Sim, B.; Ryu, D.; and Ye, J. C. 2022c. Improving diffusion models for inverse problems using manifold constraints. *arXiv preprint arXiv:2206.00941*.

Chung, H.; Sim, B.; and Ye, J. C. 2022. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12413–12422.

Creswell, A.; and Bharath, A. A. 2018. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 30(7): 1967–1974.

Diamond, S.; Sitzmann, V.; Heide, F.; and Wetzstein, G. 2017. Unrolled optimization with deep priors. *arXiv preprint arXiv:1705.08041*.

Donahue, J.; Krähenbühl, P.; and Darrell, T. 2016. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*.

Ehrlich, M.; Davis, L.; Lim, S.-N.; and Shrivastava, A. 2020. Quantization guided jpeg artifact correction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, 293–309. Springer.

Geman, S.; and Geman, D. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 721–741.

He, K.; Sun, J.; and Tang, X. 2010. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12): 2341–2353.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022a. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*.

Kawar, B.; Song, J.; Ermon, S.; and Elad, M. 2022b. Jpeg artifact correction using denoising diffusion restoration models. *arXiv preprint arXiv:2209.11888*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84–90.

Kupyn, O.; Martyniuk, T.; Wu, J.; and Wang, Z. 2019. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8878–8887.

Li, W.; Lin, Z.; Zhou, K.; Qi, L.; Wang, Y.; and Jia, J. 2022. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10758–10768.

Liu, Q.; Tan, Z.; Chen, D.; Chu, Q.; Dai, X.; Chen, Y.; Liu, M.; Yuan, L.; and Yu, N. 2022. Reduce Information Loss in Transformers for Pluralistic Image Inpainting.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023. Follow Your Pose: Pose-Guided Text-to-Video Generation using Pose-Free Videos. *arXiv preprint arXiv:2304.01186*.

Mardani, M.; Song, J.; Kautz, J.; and Vahdat, A. 2023. A Variational Perspective on Solving Inverse Problems with Diffusion Models. *arXiv preprint arXiv:2305.04391*.

Ongie, G.; Jalal, A.; Metzler, C. A.; Baraniuk, R. G.; Dimakis, A. G.; and Willett, R. 2020. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 39–56.

Pan, X.; Zhan, X.; Dai, B.; Lin, D.; Loy, C. C.; and Luo, P. 2020. Exploiting Deep Generative Prior for Versatile Image Restoration and Manipulation. In *European Conference on Computer Vision (ECCV)*.

Robbins, H. E. 1992. *An empirical Bayes approach to statistics*. Springer.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Roth, S.; and Black, M. J. 2005. Fields of experts: A framework for learning image priors. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, 860–867. IEEE.

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

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022b. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Song, J.; Vahdat, A.; Mardani, M.; and Kautz, J. 2022. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*.

Song, Y.; Shen, L.; Xing, L.; and Ermon, S. 2021. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*.

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*.

Sutskever, I. 2013. *Training recurrent neural networks*. University of Toronto Toronto, ON, Canada.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9446–9454.

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

Williams, R. J.; and Zipser, D. 1995. Gradient-based learning algorithms for recurrent. *Backpropagation: Theory, architectures, and applications*, 433: 17.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14821–14831.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv:2302.05543*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhu, J.-Y.; Krähenbühl, P.; Shechtman, E.; and Efros, A. A. 2016. Generative visual manipulation on the natural image manifold. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 597–613. Springer.

Zhu, S. C.; and Mumford, D. 1997. Prior learning and Gibbs reaction-diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(11): 1236–1250.

Zhu, Y.; Zhang, K.; Liang, J.; Cao, J.; Wen, B.; Timofte, R.; and Van Gool, L. 2023. Denoising Diffusion Models for Plug-and-Play Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1219–1229.