Conditional Velocity Score Estimation for Image Restoration

Ziqiang Shi, Rujie Liu
Fujitsu R&D Center
Beijing, China
shiziqiang@fujitsu.com

Abstract

This paper proposes a new image restoration method by introducing a velocity variable on top of the data position during recovery. Under the guidance of the degraded image, it can effectively and dynamically control the direction of the diffusion path in the reverse-time stochastic differential equation (SDE). So the crucial factor is how to combine the degraded signal as a guide in this second-order reverse process with velocity, especially in the moving direction as a diffusion path. To this end, we propose a conditional velocity score approximation (CVSA) method based on the Bayesian principle to approximate the true posterior conditional velocity score by the sum of a priori conditional velocity score and an observation velocity score of the degraded measurement at the current moment. Our method is versatile from two perspectives. It can be used for both non-blind restoration and blind restoration. At the same time, there is almost no requirement for the degradation operator, and both linear and nonlinear tasks are acceptable. In non-blind restoration, including deblurring, inpainting, super-resolution, phase retrieval, and blind restoration, such as deblurring experiments, CVSA is better than other methods and achieves a new state-of-the-art.

1. Introduction

In recent years, with the rapid development of generative models, especially diffusion generative models (DGM) [12, 27], many fields such as high-quality text-to-image [21], molecular modelling [32], music synthesis [17], image editing [19], and image translation [16] have flourished. Among them, the quality of image restoration based on DGMs have also been unprecedentedly improved, such as image super-resolution [9, 15], image deblurring [1, 4, 16], image inpainting [5, 29] etc. The basic principle behind these restoration methods is to use the DGM as the prior distribution of real images, and the degraded image as the observation signal to refine the prior to obtain the posterior of the ground truth image. The power of the DGM is that it can generate real images from white noise with high quality, although this diffusion path is uncontrolled and the generated images are inconsistent in content with the given degraded measurements. Therefore, most image restoration methods inject degraded images into the diffusion path as conditional information, control and deviate from the path to generate restored images [4, 15, 16, 29]. However, most of these image restoration methods are only carried out in the position space at present, lacking the modelling of the velocity of the image on the restoration path.

This paper proposes to model the image restoration in the position-velocity space. Along the image recovery path, the position and velocity of the image are diffused, coupled, and affect each other together. The rate of change of position is velocity, and the change of the speed can affect the direction in which the noise image diffuses to the posterior manifold in the Euclidean space. The key to controlling all this is the conditional velocity score (CVS) based on the observed degraded image at each moment on the path. CVS controls the evolution of the path that diffuses from the position-velocity initial variable of white noise to the ground truth image (corresponding to the degraded one), but it is intractable. By Bayes' Theorem, the CVS can be transformed into the sum of the unconditional velocity score and the observation velocity score. In the case of a pre-trained DGM, the unconditional velocity score is immediately available, and thus the only missing piece in the whole puzzle is the velocity score of the degraded signal observed at the current moment. Based on an improved Jensen inequality, we convert the degraded signal velocity score at the current moment to the velocity score at the initial moment, which can be obtained through the model of the image restoration problem. To underline the critical computations in this image restoration process, our method is termed as CVS approximation (CVSA).

In summary, this paper makes the following technical contributions,

- Via estimating CVS, we propose and realize image restoration in position-velocity space for the first time. Based on a general generative model in this space,

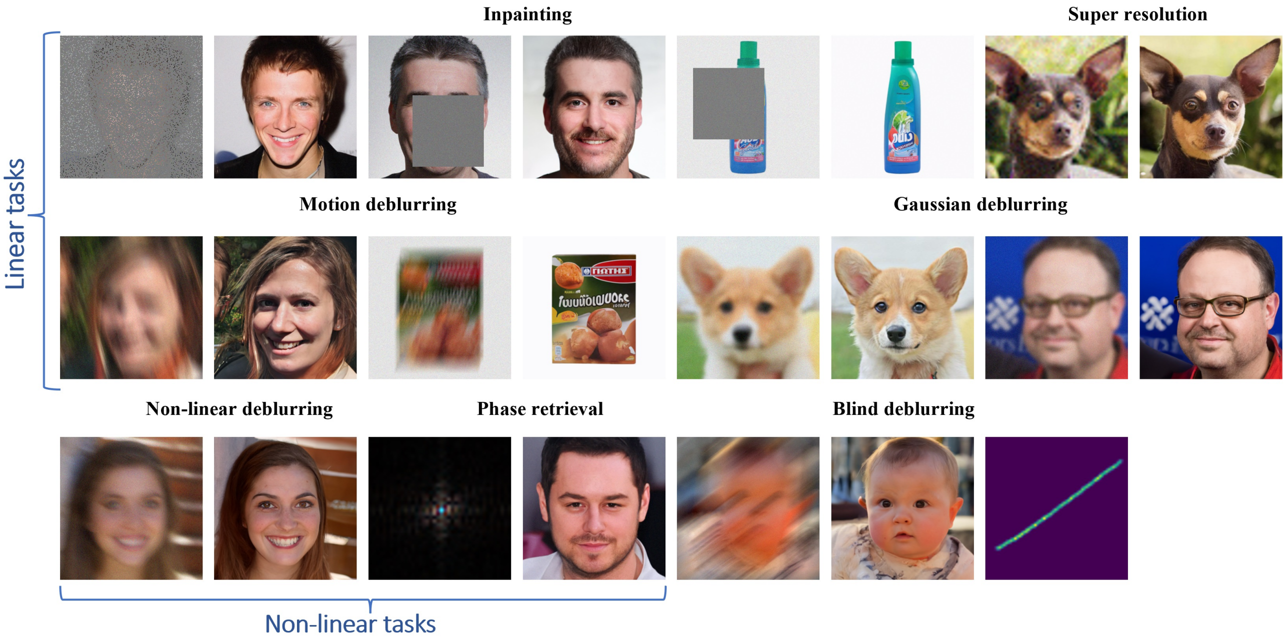

Figure 1. CVSA can be competent for a variety of different tasks, including linear, nonlinear, and even blind recovery.

CVSA does not need to be trained for specific image restoration tasks and is applicable to both linear and nonlinear tasks as shown in Figure 1.

- We propose a method to transfer the observed signal velocity score at the current moment to the calculation of the score at the initial moment, enabling the estimation of the intractable CVS.
- Experiments on multiple databases and multiple tasks, including deblurring, inpainting, super-resolution, phase retrieval, and blind deblurring, prove that CVSA is effective and achieves state-of-the-art (SOTA) results.

The following content of this paper includes: First we will introduce the related work and preliminary background; then the methodology and implementation details of CVSA will be described in Section 4; afterwards, we demonstrate the effectiveness of the method through experiments and finally summarize the full paper.

2. Related Work

Our work belongs to the field of image restoration (inverse problem solving) based on diffusion models. Research in this area is divided into two categories, one is supervised regression using paired data for each specific recovery problem; the other is a universal method for all recovery tasks based on a general DGM. There are two ways to realize the method in the former category. One is to use the degraded measurement as the conditional input and the original signal as the output to train the DGMs, such as SR3 [24], SR [31], etc. The other way is image-to-image translation, which is to directly find the continuous diffusion path between the degraded signal distribution and the ground truth image distribution through training, such as Palette [23], I^2SB [16], etc. This category of method requires a model trained separately for each restoration task, which is not universal and cannot be applied to other tasks.

Our method belongs to the second category, which is more related to CVSA. This type of method has received more and more attention due to its advantage that one model can solve all problems. There are two paradigms in this category, namely replacement-based and reconstruction-based. The philosophy behind the replacement-based method is to directly add the degraded measurement to the predictions on the diffusion path, and the resulting combined signal is used as the input for the next diffusion step. Typical methods include ΠGDM [26], Pyramid DDPM [22], MCG [6], etc. The principle of the reconstruction-based method is to correct the conditional score by approximating the classifier guidance term based on minimizing the error between the degraded signal and the constructed degraded signal predicted by the DGMs. Typical examples include ScoreSDE [27], DDRM [15], ILVR [2], CCDF [7], DPS [5], etc. Most of the above methods are carried out in the image space, lacking the simulation of the diffusion speed of the image on the restoration path. Our CVSA fills this gap.

3. Preliminaries

Diffusion models have achieved state-of-the-art performance in image restoration [5, 26, 29]. The basic principle of these methods is to add conditional information into the diffusion process from white noise to real images, such as degraded images to control the direction of the generated diffusion paths. This paper attempts to further explicitly introduce the velocity (or momentum) field to control the direction of the restoration path, so as to obtain a better and faster method.

Generative models based on first-order stochastic differential equations (SDE) can be enhanced, both in terms of the quality and speed in generation of data, by introducing a characterization of the velocity in the diffusion process. A typical representative of which is the following critically-damped Langevin diffusion (CLD) process [8]

$$\begin{cases} d\mathbf{x}_t &= M^{-1}\beta\mathbf{v}_t dt, \\ d\mathbf{v}_t &= -\beta\mathbf{x}_t dt + \Gamma\beta M^{-1}\mathrm{v}_t dt + \sqrt{2\Gamma\beta}d\mathbf{w}_t. \end{cases} \tag{1}$$

where $\mathbf{x}_t, \mathbf{v}_t \in \mathbb{R}^d$, $t \in [0,1]$ are the position (data) and momentum process respectively. $M > 0$ is mass, and it controls the degree of coupling between $\mathbf{x}_t$ and $\mathbf{v}_t$; friction coefficient $\Gamma > 0$ governs the amount of noise (from the standard Wiener process $\mathbf{w}_t$) injected into the system; scaling coefficient $\beta > 0$ guarantees that the system converges to an equilibrium state within a finite time. M and Γ must satisfy the *critical damping* condition $\Gamma^2 = 4M$, so that the system can smoothly converge to an equilibrium data distribution without oscillations.

The image restoration problem is defined as follows

$$\mathbf{y} = \mathcal{H}(\mathbf{x}) + \mathbf{n}, \tag{2}$$

where the only unknown is $\mathbf{x} \in \mathbb{R}^d$ which is the ground truth image that needs to be recovered. $\mathcal{H}$ is the measurement (or degradation) operator, it can be linear or nonlinear. For example, inpainting, super-resolution, Gaussian deblurring or motion deblurring are all linear problems, while phase retrieval and non-uniform deblurring are nonlinear problems. The method proposed in this paper can be used for both types of operators, as long as it is differentiable. $\mathbf{n} \in \mathbb{R}^e$ is the measurement noise satisfying $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}_e)$, and $\mathbf{y}$ is the observation signal. The task of image restoration is to produce $\hat{\mathbf{x}}$ given $\mathbf{y}$ under the constraints of realness ($\hat{\mathbf{x}} \sim q(\mathbf{x})$, which is the distribution of ground truth images) and consistency ($\mathbf{y} = \mathcal{H}(\hat{\mathbf{x}})$). It can be formulated as

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left[\|\mathbf{y} - \mathcal{H}(\mathbf{x}) - \mathbf{n}\|_2^2 - \lambda \log q(\mathbf{x})\right], \tag{3}$$

which is the problem to be addressed in the next section.

4. Methods

In order to obtain realness and consistency at the same time, we inject the given degraded measurement $\mathbf{y}$ as conditional information into the reverse CLD generation process

$$\begin{cases} d\mathbf{x}_t &= -M^{-1}\beta\mathbf{v}_t dt, \\ d\mathbf{v}_t &= \beta\mathbf{x}_t dt + \Gamma\beta M^{-1}\mathrm{v}_t dt \\ & \quad -2\Gamma\beta\nabla_{\mathbf{v}_t}\log p(\mathbf{x}_t, \mathbf{v}_t|\mathbf{y})dt + \sqrt{2\Gamma\beta}d\bar{\mathbf{w}}_t, \end{cases} \tag{4}$$

via the CVS $\nabla_{\mathbf{v}_t}\log p(\mathbf{x}_t, \mathbf{v}_t|\mathbf{y})$. However, since the form of conditional $\mathbf{y}$ in different image restoration tasks is different, there is no general diffusion model (or score estimation model) to calculate the CVS. And what we have is a generic unconditional velocity score estimation model [8]. Intractable CVS needs to be converted into a computable form. Through Bayes' Theorem

$$\begin{aligned} \nabla_{\mathbf{v}_t}\log p(\mathbf{x}_t, \mathbf{v}_t|\mathbf{y}) &= \nabla_{\mathbf{v}_t}\log p(\mathbf{x}_t, \mathbf{v}_t) \\ &\quad + \nabla_{\mathbf{v}_t}\log p(\mathbf{y}|\mathbf{x}_t, \mathbf{v}_t), \end{aligned} \tag{5}$$

CVS can be converted into the sum of the unconditional score $\nabla_{\mathbf{v}_t}\log p(\mathbf{x}_t, \mathbf{v}_t)$ and the observation (or measurement) velocity score (OVS) $\nabla_{\mathbf{v}_t}\log p(\mathbf{y}|\mathbf{x}_t, \mathbf{v}_t)$. Among them, the unconditional score can be obtained by a pre-trained general CLD diffusion model, while the OVS is mostly intractable, since there is no direct analytic closed relationship between diffusion state $\mathbf{x}_t, \mathbf{v}_t$ and the measurement $\mathbf{y}$. They are connected by the initial state $(\mathbf{x}_0, \mathbf{v}_0)$, where $(\mathbf{x}_t, \mathbf{v}_t)$ and $(\mathbf{x}_0, \mathbf{v}_0)$ are associated by Eq. (1), and $\mathbf{y}$ and $(\mathbf{x}_0, \mathbf{v}_0)$ are related by Eq. (2). In the next subsection, we will give an approximation of the OVS through these two connections.

4.1. Observation Velocity Score Approximation

Let $\mathbf{u}_0 = (\mathbf{x}_0, \mathbf{v}_0)^\top$, $\mathbf{u}_t = (\mathbf{x}_t, \mathbf{v}_t)^\top$, and introducing $\mathbf{u}_0$ into the OVS, and integrating over it, we get

$$p(\mathbf{y}|\mathbf{u}_t) = \int p(\mathbf{y}|\mathbf{u}_0)p(\mathbf{u}_0|\mathbf{u}_t)d\mathbf{u}_0 \simeq p(\mathbf{y}|\hat{\mathbf{u}}_0), \tag{6}$$

where $\hat{\mathbf{u}}_0 = \mathbb{E}_{\mathbf{u}_0 \sim p(\mathbf{u}_0|\mathbf{u}_t)}[\mathbf{u}_0] = \int \mathbf{u}_0 p(\mathbf{u}_0|\mathbf{u}_t)d\mathbf{u}_0$. Here we have used the convex integral version of Jensen's inequality [13]

$$\phi\left(\int \mathbf{u}f(\mathbf{u})d\mathbf{u}\right) \leq \int \phi(\mathbf{u})f(\mathbf{u})d\mathbf{u}, \tag{7}$$

where $f(\mathbf{u})$ and $\phi(\mathbf{u})$ are arbitrary probability density functions and convex functions, respectively. Due to the assumption that $\mathbf{n}$ is Gaussian in Eq. (2), thus $p(\mathbf{y}|\mathbf{u}_0)$ is a convex function in $\mathbf{u}_0$. Let $p(\mathbf{y}|\cdot) \to \phi(\cdot)$ and $p(\cdot|\mathbf{u}_t) \to f(\cdot)$ in Eq. (7), then we have

$$p\left(\mathbf{y}|\int \mathbf{u}_0 p(\mathbf{u}_0|\mathbf{u}_t)d\mathbf{u}_0\right) \leq \int p(\mathbf{y}|\mathbf{u}_0)p(\mathbf{u}_0|\mathbf{u}_t)d\mathbf{u}_0,$$

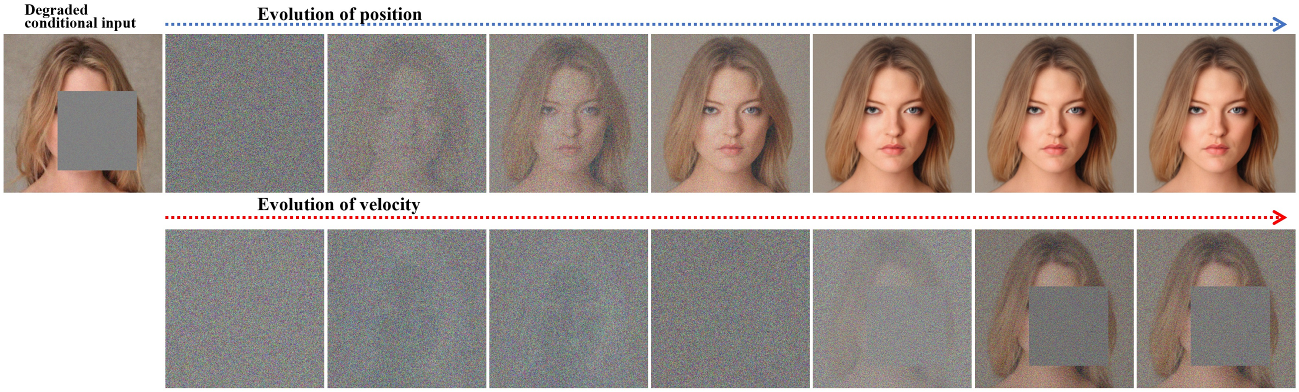

Figure 2. The position and velocity of the data on the generation path controlled by the input conditions are mutually coupled and evolve, and it can be seen that both controlled and affected by the input degraded face.

and the gap is roughly proportional to the fourth power of the distance between ground truth $\mathbf{x}_0$ and estimated $\hat{\mathbf{x}}_0$ at time t.

Proposition 4.1. *(Determination of the probability of measurement* $\mathbf{y}$ *given* $\mathbf{u}_t$*) For simplicity of notation,* $\mathbb{E}_{\mathbf{u}_0 \sim p(\mathbf{u}_0|\mathbf{u}_t)}$ *is denoted as* $\mathbb{E}$*. Under the condition of diffusion state* $\mathbf{u}_t$ *at time* t*, the probability of observing* $\mathbf{y}$ *can be derived from*

$$\begin{aligned} p(\mathbf{y}|\mathbf{u}_t) &= \mathbb{E}\left[p(\mathbf{y}|\mathbf{u}_0)\right] = p(\mathbf{y}|\hat{\mathbf{u}}_0) \\ &\quad + c\mathbb{E}\left[\left\|\mathcal{H}(\mathbf{x}_0 - \hat{\mathbf{x}}_0)\right\|^2 \left\|\mathbf{x}_0 - \hat{\mathbf{x}}_0\right\|^2\right], \end{aligned} \tag{8}$$

where

$$c = \frac{1}{\sigma^2} \int_0^1 \frac{1}{2\pi^{d/2}\sigma^{d/2}} \exp\left(-\frac{s^2}{2}\right)(1-s)\left(\frac{s^2}{\sigma^2} - 1\right) ds.$$

Proof sketch. Take $p(\mathbf{y}|s\mathbf{u}_0 + (1-s)\hat{\mathbf{u}}_0)$ as a function of s, and then perform a second-order Taylor expansion at $s = 0$. Computes the second-order derivative $d^2 p(\mathbf{y}|s\mathbf{u}_0 + (1-s)\hat{\mathbf{u}}_0)/ds^2$, bringing it into the remainder of the second-order integral. Finally extract the terms that have nothing to do with s to get Eq. (8). The detailed proof is presented in the supplementary section.

So far, Eq. (6) and Eq. (8) tell us that $\hat{\mathbf{u}}_0$ (in fact $\hat{\mathbf{x}}_0$, since the particle is assumed in stationary state at the initial moment, that is the initial $\hat{\mathbf{v}}_0$ is assumed to be $\mathbf{0}$) is the final piece of the puzzle.

4.2. Initial State Estimation

Getting $\mathbf{u}_t$ from $\mathbf{x}_0$ is easy. Eq. (1) shows that $\mathbf{u}_t$ is obtained by directly adding white noise to the velocity $\mathbf{v}_0$, that is, indirectly adding noise disturbance to the image $\mathbf{x}_0$ through several diffusion steps. They are linked by the probabilistic transition kernel of linear SDEs [25]. In terms of CLD as in Eq. (1), we have the exact relation [8]

$$p(\mathbf{u}_t|\mathbf{u}_0) = \mathcal{N}(\mathbf{u}_t; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \tag{9}$$

where

$$\begin{aligned} \boldsymbol{\mu}_t &= \begin{pmatrix} 2\beta t\Gamma^{-1}\mathbf{x}_0 + 4\beta t\Gamma^{-2}\mathbf{v}_0 + \mathbf{x}_0 \\ -\beta t\mathbf{x}_0 - 2\beta t\Gamma^{-1}\mathbf{v}_0 + \mathbf{v}_0 \end{pmatrix} e^{-2\beta t\Gamma^{-1}} \\ &= e^{-2\beta t\Gamma^{-1}} \begin{pmatrix} 2\beta t\Gamma^{-1} + 1 & 4\beta t\Gamma^{-2} \\ -\beta t & -2\beta t\Gamma^{-1} + 1 \end{pmatrix} \begin{pmatrix} \mathbf{x}_0 \\ \mathbf{v}_0 \end{pmatrix} \\ &= \boldsymbol{D}_t \begin{pmatrix} \mathbf{x}_0 \\ \mathbf{v}_0 \end{pmatrix}, \end{aligned} \tag{10}$$

and

$$\begin{aligned} \boldsymbol{\Sigma}_t &= \Sigma_t \otimes \boldsymbol{I}_d, \\ \Sigma_t &= \begin{pmatrix} \Sigma_t^{xx} & \Sigma_t^{xv} \\ \Sigma_t^{xv} & \Sigma_t^{vv} \end{pmatrix} e^{-4\beta t\Gamma^{-1}}, \\ \Sigma_t^{xx} &= \Sigma_0^{xx} + e^{4\beta t\Gamma^{-1}} - 1 + 4\beta t\Gamma^{-1}\left(\Sigma_0^{xx} - 1\right) \\ &\quad + 4\beta^2 t^2\Gamma^{-2}\left(\Sigma_0^{xx} - 2\right) + 16\beta^2 t^2\Gamma^{-4}\Sigma_0^{vv}, \\ \Sigma_t^{xv} &= -\beta^2 t\Sigma_0^{xx} + 4\beta t\Gamma^{-2}\Sigma_0^{vv} \\ &\quad - 2\beta^2 t^2\Gamma^{-1}\left(\Sigma_0^{xx} - 2\right) - 8\beta^2 t^2\Gamma^{-3}\Sigma_0^{vv}, \\ \Sigma_t^{vv} &= \frac{\Gamma^2}{4}\left(e^{4\beta t\Gamma^{-1}} - 1\right) + \beta t\Gamma + \beta^2 t^2\left(\Sigma_0^{xx} - 2\right) \\ &\quad + \Sigma_0^{vv}\left(1 + 4\beta t^2\Gamma^{-2} - 4\beta t\Gamma^{-1}\right). \end{aligned} \tag{11}$$

But it is not easy to get $\mathbf{x}_0$ from $\mathbf{u}_t$, since tens or hundreds of Gaussian diffusion steps have occurred between them. The distribution of $\mathbf{x}_0$ given $\mathbf{u}_t$ is no longer a simple Gaussian distribution, but its expectation still has a closed-form solution.

Proposition 4.2. *(Estimate the mean of the initial image from the current moment) Under the condition of* $\mathbf{u}_t$ *at time* t*, the posterior mean value of the image at time 0 can be derived from*

$$\begin{aligned} \left[\boldsymbol{\Sigma}_t^{-1}\boldsymbol{D}_t\mathbb{E}[\mathbf{u}_0|\mathbf{u}_t]\right]_\Delta &= \nabla_{\mathbf{v}_t}\log p(\mathbf{u}_t) \\ &\quad + \left[\boldsymbol{\Sigma}_t^{-1}\mathbf{u}_t\right]_\Delta, \end{aligned} \tag{12}$$

where $\Delta \doteq d+1, d+2, \cdots 2d$ indicates to take the second half of this vector.

Proof sketch. First, $\mathbf{x}$ and $\mathbf{v}$ are bundled together, and the relationship between $\mathbf{u}_t$ and $\mathbf{u}_0$ is simplified by using the fact that the probability density transition matrix (or disturbance matrix) of the linear SDEs is a Gaussian distribution. This relationship is then separated into the product of a distribution containing only $\mathbf{u}_t$ and a simple power distribution containing $\mathbf{u}_t$ and $\mathbf{u}_0$. The last step is to separate $\mathbf{x}$ and $\mathbf{v}$, that is, to take the $\log$ and find the score respective to $\mathbf{v}$, and the result can be obtained by doing an inverse matrix multiplication. For detailed derivation please refer to the supplementary section.

Remark 4.3. The distribution of $\mathbf{v}_0$ is simple and clear, and the initial value is taken as $\mathbf{0}$, although it is a Gaussian distribution with variance $\lambda M \boldsymbol{I}_d$. Therefore, the only unknown quantity in Eq. (12) is $\hat{\mathbf{x}}_0$, so it can be obtained.

Now from Eq. (2), Eq. (6), and $\mathbf{v}_0 = \mathbf{0}$ we know that

$$\begin{aligned} \nabla_{\mathbf{v}_t} \log p(\mathbf{y}|\mathbf{u}_t) &\simeq \nabla_{\mathbf{v}_t} \log p(\mathbf{y}|\hat{\mathbf{x}}_0, \mathbf{0}) \\ &= \nabla_{\mathbf{v}_t} \left[-\frac{1}{2\sigma^2} \left\| \mathbf{y} - \mathcal{H}(\hat{\mathbf{x}}_0) \right\|_2^2 \right], \end{aligned} \tag{13}$$

which can be obtained through the automatic differentiation function `torch.autograd` [20].

Bringing this OVS approximation in Eq. (13) into the decomposition of CVS in Eq. (5), and then into the reverse generative CLD in Eq. (4) with conditional input, we can get the overall image restoration algorithm CVSA, which is summarized in the next section.

4.3. Algorithm

For the clarity of the algorithm description, all elements of the $\boldsymbol{D}_t$ in Eq. (10) and $\boldsymbol{\Sigma}_t^{-1}$ in Eq. (11) matrices are listed as follows

$$\boldsymbol{D}_t = \begin{pmatrix} D_t^{xx} & D_t^{xv} \\ D_t^{vx} & D_t^{vv} \end{pmatrix}, \quad \boldsymbol{\Sigma}_t^{-1} = \begin{pmatrix} \sigma_t^{xx} & \sigma_t^{xv} \\ \sigma_t^{xv} & \sigma_t^{vv} \end{pmatrix}, \tag{14}$$

and the whole restoration process of CVSA is shown in Algorithm 1. Figure 2 shows how CVSA controls the evolution of position and velocity under the influence of the initial degraded image conditional input in the case of mutual coupling.

4.4. CVSA for Blind Restoration

As can be seen from Algorithm 1, we need to know the exact form of $\mathcal{H}$ to use CVSA. But in practical applications, such as blind deblurring, the exact hyperparameters in the blurring kernel of $\mathcal{H}$ are not known, and the problem in Eq. (3) becomes

$$\begin{aligned} \hat{\mathbf{x}}, \hat{\mathbf{k}} = \arg\min_{\mathbf{x},\mathbf{k}} [&\|\mathbf{y} - \mathbf{k} * \mathbf{x} - \mathbf{n}\|_2^2 \\ &- \lambda_1 \log q_1(\mathbf{x}) - \lambda_2 \log q_2(\mathbf{k})]. \end{aligned} \tag{15}$$

Algorithm 1 CVSA for image restoration

```
Input and initialization: Observed measurement $\mathbf{y}$, the degradation operator $\mathcal{H}$ in Eq. (2), $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, $\mathbf{v}_T \sim \mathcal{N}(\mathbf{0}, M\boldsymbol{I}_d)$, $\mathbf{u}_T = (\mathbf{x}_T, \mathbf{v}_T)^\top$, $dt = 1/T$, $\mathbf{v}_0 = \mathbf{0}$, and the pre-trained general velocity score ($\nabla_{\mathbf{v}_t} \log p(\mathbf{u}_t)$) prediction model is $\mathfrak{S}_\theta(\mathbf{u}_t, t)$.
for $t = T, T-1, \cdots, 1$
    #Initial state estimation from current moment
    $d_t \leftarrow \sigma_t^{xv} * D_t^{xx} + \sigma_t^{vv} * D_t^{vx}$,
    $\hat{\mathbf{x}}_0 \leftarrow \left[\sigma_t^{xv} * \mathbf{x}_t + \sigma_t^{vv} * \mathbf{v}_t + \mathfrak{S}_\theta(\mathbf{u}_t, t)\right] / d_t$.
    #Approximate the CVS
    $\mathbf{s}_t \leftarrow \mathfrak{S}_\theta(\mathbf{u}_t, t) + \nabla_{\mathbf{v}_t} \left[ -\frac{1}{2\sigma^2} \left\| \mathbf{y} - \mathcal{H}(\hat{\mathbf{x}}_0) \right\|_2^2 \right]$.
    #Use the Eq. (4) to update $\mathbf{x}$ and $\mathbf{v}$, that is
    $\mathbf{x}_{t-1} \leftarrow \mathbf{x}_t - M^{-1}\beta\mathbf{v}_t dt$, and
    $\mathbf{v}_{t-1} \leftarrow \mathbf{v}_t + \beta\mathbf{x}_t dt + \Gamma\beta M^{-1}\mathrm{v}_t dt - 2\Gamma\beta\mathbf{s}_t dt$
    #Do not add Wiener noise in the last step
    If $t > 0$, then $\mathbf{v}_{t-1} \leftarrow \mathbf{v}_{t-1} + \sqrt{2\Gamma\beta}\sqrt{dt}\mathbf{z}$,
    where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$.
Output: $\mathbf{x}_0$.
```

It can be seen from the above problem that $\mathbf{k}$ and $\mathbf{x}$ are in a symmetrical position, and neither of them has particularity. Therefore, CVSA can also be used to restore $\mathbf{k}$ just like restoring $\mathbf{x}$, that is to say, $\mathbf{k}$ and $\mathbf{x}$ can be predicted at the same time only when a blurred image $\mathbf{y}$ is observed. The model and process of predicting $\mathbf{k}$ are similar to those of $\mathbf{x}$ and its complete calculation process is summarized in Algorithm 2.

5. Experiments

5.1. Datasets and Setup

We tested CVSA on four different datasets all at 256×256 resolution, including CelebA-HQ [18], AFHQ [3], FFHQ [14], and Products-6k [10].

For the image degradation operator $\mathcal{H}$ in Eq. (2) and the corresponding restoration tasks, we tested and compared the performance of CVSA under 8 configurations, namely (1) Gaussian blur, using a kernel with a size of 61×61 and a standard deviation of 3.0; (2) Motion blur, using the third-party generated code[1] with a kernel of size 61×61 and intensity of 0.5; (3) For box-type inpainting, randomly mask out the signals on all RGB channels with a size of 128×129; (4) For random-type inpainting, 92% of the pixels on all channels are randomly masked; (5) Non-linear blurring, using an algorithm called Blur Kernel Space encoding [28] to generate; (6) Phase recovery, where the degradation operator is a Fourier transform, and only its amplitude information is taken; (7) Super-resolution, using bicubic downsampling; (8) Blind deblurring, which has both motion blur and Gaussian blur, the size of both kernels are 61×61 with a Motion Blur strength of 0.5 and a Gaussian Blur strength of

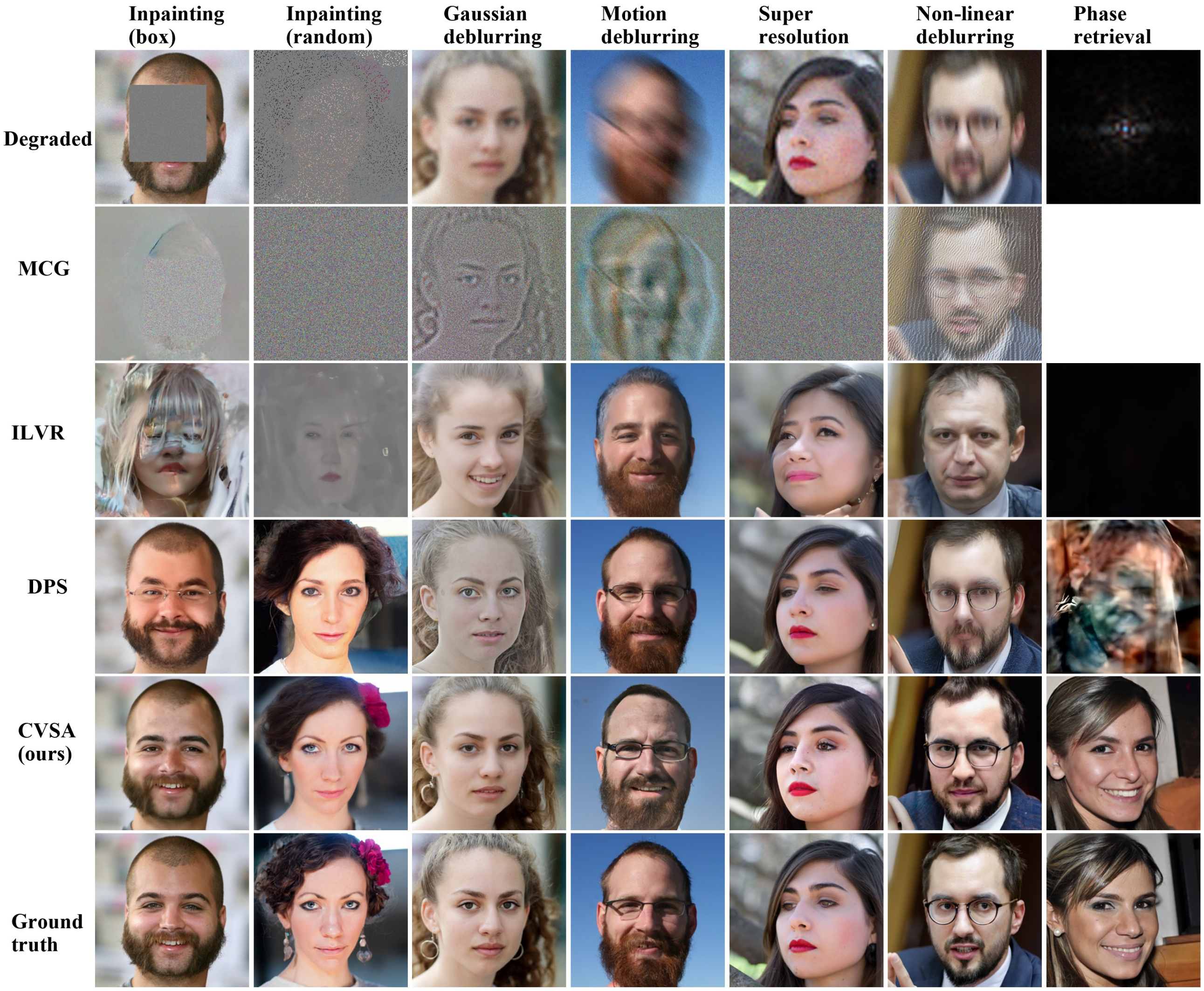

Figure 3. Qualitative comparison of different algorithms on FFHQ.

3.0. $\mathbf{n}$ was the same in all tasks, and σ was 0.05.

CVSA was compared with several of the best methods at present, including denoising diffusion restoration models (DDRM) [15], diffusion posterior sampling (DPS) [5], manifold constrained gradient (MCG) [6], iterative latent variable refinement (ILVR), and parallel DPS (BlindDPS) [4]. For all these methods, we use their publicly available pre-trained checkpoints. However not all methods have available models on all data. For example, DDRM has only CelebA-HQ checkpoints; DPS, MCG, and BlindDPS only have pre-trained models on FFHQ, while ILVR has available models on all datasets except Products-6k.

5.2. Results and Discussion

We have done quantitative evaluations based on peak signal-to-noise-ratio (PSNR) [30], structural similarity index (SSIM) [30], and Fréchet Inception Distance (FID) [11] distances for image restoration qualities. Mean square error (MSE) is used to evaluate the recovery accuracy of the kernel in blind deblurring.

The results of five **linear restoration** tasks are shown in Table 1. It can be seen that CVSA is better than almost all other comparison methods on FFHQ, CelebA-HQ, and AFHQ-dog in PSNR, SSIM, and FID. Compared with the previous SOTA method DPS, CVSA has a relative improvement of 5% in SSIM and a relative decrease of 10% in FID. Compared with ILVR on the three datasets, CVSA has an average performance improvement of 50% on both

[1] https://github.com/LeviBorodenko/motionblur

Table 1. PSNR, SSIM, and FID in a comparative study of different SOTA image restoration methods with our CVSA on FFHQ, CelebA-HQ, and AFHQ-dog under 5 different linear degradation operators. **Bold**: best on each dataset.

Data	Method	4 × SR	Inpainting (box)	Inpainting (random)	Deblurring (Gauss)	Deblurring (motion)
		PSNR↑/SSIM↑/FID↓	PSNR↑/SSIM↑/FID↓	PSNR↑/SSIM↑/FID↓	PSNR↑/SSIM↑/FID↓	PSNR↑/SSIM↑/FID↓
FFHQ	MCG	18.05 / 0.245 / 172.1	10.98 / 0.206 / 337.7	10.50 / 0.047 / 477.2	11.33 / 0.147 / 299.6	11.27 / 0.093 / 328.6
	ILVR	17.30 / 0.413 / 40.84	15.09 / 0.352 / 83.23	11.39 / 0.353 / 171.4	17.19 / 0.409 / 41.07	17.12 / 0.410 / 42.22
	DPS	23.64 / 0.666 / 35.63	23.22 / 0.800 / 30.06	23.64 / 0.686 / 34.38	**24.85** / 0.702 / 29.15	23.31 / 0.652 / 31.64
	CVSA	**26.65 / 0.690 / 30.86**	**24.22 / 0.857 / 27.41**	**24.22 / 0.724 / 30.87**	23.10 / **0.738 / 28.83**	**23.60 / 0.674 / 29.77**
CelebA	DDRM	29.06 / 0.828 / 38.87	-	15.49 / 0.421 / 149.6	31.16 / 0.870 / 28.04	-
	ILVR	18.03 / 0.457 / 41.44	15.31 / 0.395 / 71.40	10.96 / 0.366 / 194.6	17.90 / 0.455 / 41.63	17.92 / 0.456 / 42.18
	CVSA	**33.14 / 0.831 / 33.47**	**30.53 / 0.574 / 38.06**	**13.44 / 0.456 / 14.35**	**38.76 / 0.892 / 26.76**	**40.40 / 0.554 / 35.00**
AFHQ	ILVR	17.60 / 0.378 / 33.70	14.89 / 0.319 / 60.88	11.73 / 0.313 / 210.27	17.41 / 0.373 / 32.91	17.31 / 0.371 / 33.54
	CVSA	**26.70 / 0.743 / 22.15**	**25.20 / 0.696 / 28.54**	**23.59 / 0.686 / 27.33**	**24.13 / 0.701 / 24.19**	**23.53 / 0.635 / 23.37**

Algorithm 2 CVSA for blind deblurring

```
Input and initialization: Observed signal $\mathbf{y}$, $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, $\mathbf{v}_T \sim \mathcal{N}(\mathbf{0}, M\boldsymbol{I}_d)$, $\mathbf{u}_T = (\mathbf{x}_T, \mathbf{v}_T)^\top$, $dt = 1/T$, $\mathbf{v}_0 = \mathbf{0}$, and the pre-trained general score $(\nabla_{\mathbf{v}_t} \log p(\mathbf{u}_t))$ prediction model is $\mathfrak{S}_\theta(\mathbf{u}_t, t)$; $\mathbf{k}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$, $\mathbf{v}'_T \sim \mathcal{N}(\mathbf{0}, M\boldsymbol{I}_k)$, $\mathbf{u}'_T = (\mathbf{k}_T, \mathrm{v}'_T)^\top$, $\mathbf{v}'_0 = \mathbf{0}$, and the pre-trained unconditional velocity score $(\nabla_{\mathbf{v}'_t} \log p(\mathbf{u}'_t))$ prediction model is $\mathfrak{S}_{\theta'}(\mathbf{u}'_t, t)$.
for $t = T, T-1, \cdots, 1$
    #Initial image state prediction from current moment
    $d_t = \sigma_t^{xv} * D_t^{xx} + \sigma_t^{vv} * D_t^{vx}$,
    $\hat{\mathbf{x}}_0 \leftarrow \left[\sigma_t^{xv} * \mathbf{x}_t + \sigma_t^{vv} * \mathbf{v}_t + \mathfrak{S}_\theta(\mathbf{u}_t, t)\right] / d_t$.
    #Initial kernel state prediction from current moment
    $d_t = \sigma_t^{kv'} * D_t^{kk} + \sigma_t^{v'v'} * D_t^{v'k}$,
    $\hat{\mathbf{k}}_0 \leftarrow \left[\sigma_t^{kv'} * \mathbf{k}_t + \sigma_t^{v'v'} * \mathbf{v}'_t + \mathfrak{S}_{\theta'}(\mathbf{u}'_t, t)\right] / d_t$.
    #Approximate the CVS for image
    $\mathbf{s}_t = \mathfrak{S}_\theta(\mathbf{u}_t, t) + \nabla_{\mathbf{v}_t}\left[-\frac{1}{2\sigma^2}\left\|\mathbf{y} - \hat{\mathbf{k}}_0 * \hat{\mathbf{x}}_0\right\|_2^2\right]$.
    #Approximate the CVS for kernel
    $\mathbf{s}'_t = \mathfrak{S}_{\theta'}(\mathbf{u}'_t, t) + \nabla_{\mathbf{v}'_t}\left[-\frac{1}{2\sigma^2}\left\|\mathbf{y} - \hat{\mathbf{k}}_0 * \hat{\mathbf{x}}_0\right\|_2^2\right]$.
    # Use Eq. (4) to update
    $\mathbf{x}_{t-1} \leftarrow \mathbf{x}_t - M^{-1}\beta\mathbf{v}_t dt$,
    $\mathbf{v}_{t-1} \leftarrow \mathbf{v}_t + \beta\mathbf{x}_t dt + \Gamma\beta M^{-1}\mathrm{v}_t dt - 2\Gamma\beta\mathbf{s}_t dt$.
    Then use the Eq. (4) to update,
    $\mathbf{k}_{t-1} \leftarrow \mathbf{k}_t - M^{-1}\beta\mathbf{v}'_t dt$,
    $\mathbf{v}'_{t-1} \leftarrow \mathbf{v}'_t + \beta\mathbf{k}_t dt + \Gamma\beta M^{-1}\mathrm{v}'_t dt - 2\Gamma\beta\mathbf{s}'_t dt$.
    #Do not add Wiener noise in the last step
    If $t > 0$, then $\mathbf{v}_{t-1} \leftarrow \mathbf{v}_{t-1} + \sqrt{2\Gamma\beta}\sqrt{dt}\mathbf{z}$,
    $\mathbf{v}'_{t-1} \leftarrow \mathbf{v}'_{t-1} + \sqrt{2\Gamma\beta}\sqrt{dt}\mathbf{z}'$,
    where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ and $\mathbf{z}' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_k)$.
Output: $\mathbf{x}_0$ and $\mathbf{k}_0$.
```

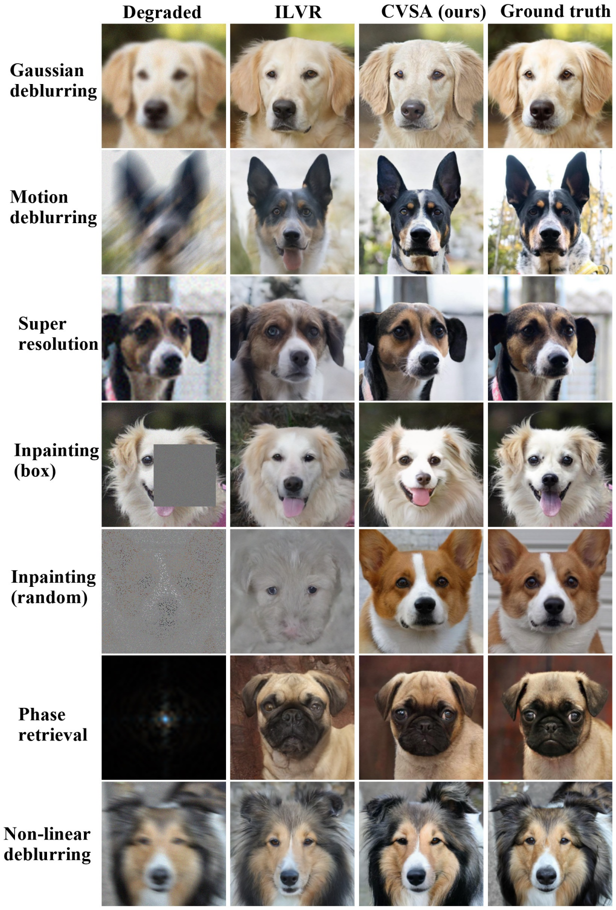

Figure 4. Qualitative comparison of different algorithms on AFHQ-dog.

PSNR and SSIM, and an average 30% decrease on FID. Figure 3, 4, and 5 show the qualitative results on the three datasets, and it can be seen that CVSA is better than other methods in terms of authenticity and consistency. There are also sample results on two **nonlinear restoration** tasks in Figure 3, 4, and 5, which also show that CVSA is superior to other comparison methods. In particular, other methods have failed on the phase retrieval task, and CVSA can still restore the ground truth image. Table 2 proves this advantage from the quantitative perspective. On FFHQ, CVSA outperforms both DPS and ILVR on three image quality metrics. On CelebA-HQ and AFHQ, CVSA improves PSNR by at least 40%, SSIM by at least 50%, and

Table 2. PSNR, SSIM, and FID in a comparative study of different SOTA image restoration methods with our CVSA in phase retrieval and non-linear deblurring on FFHQ, CelebA-HQ, and AFHQ-dog. **Bold**: best on each dataset.

Data	Method	Phase retrieval	Non-linear deblurring
		PSNR↑/SSIM↑/FID↓	PSNR↑/SSIM↑/FID↓
FFHQ	MCG	-	14.86 / 0.245 / 200.0
	ILVR	6.128 / 0.069 / 369.7	17.19 / 0.410 / 41.44
	DPS	11.81 / 0.318 / 169.2	22.87 / 0.636 / 37.68
	CVSA	**12.45 / 0.326 / 50.87**	**25.25 / 0.653 / 34.80**
CelebA	ILVR	6.248 / 0.081 / 311.4	17.81 / 0.451 / 41.27
	CVSA	**15.20 / 0.421 / 97.78**	**24.54 / 0.662 / 25.04**
AFHQ	ILVR	6.035 / 0.061 / 218.5	17.33 / 0.371 / 33.95
	CVSA	**11.10 / 0.245 / 44.33**	**24.19 / 0.593 / 27.40**

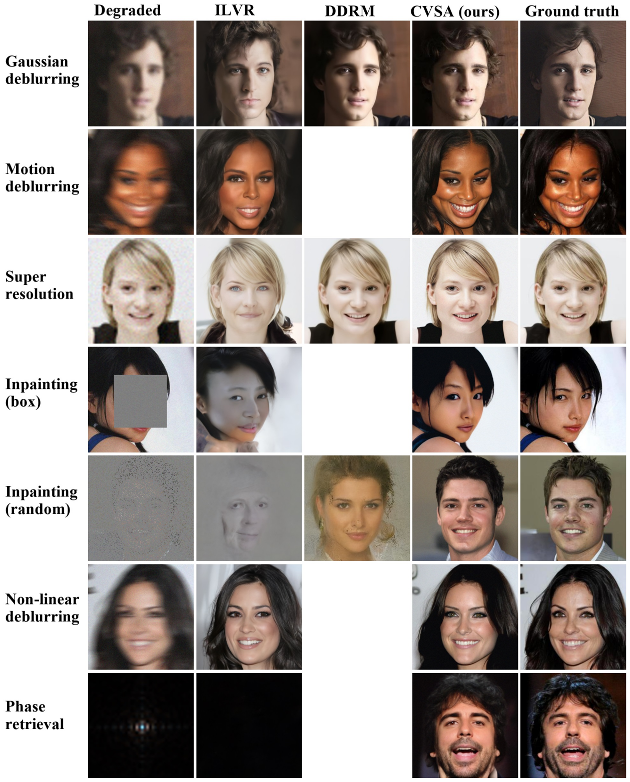

Figure 5. Qualitative comparison of different algorithms on CelebA-HQ.

Table 3. PSNR, SSIM, FID, and MSE (for kernel estimation) in a comparative study of different SOTA blind deblurring methods with our CVSA on FFHQ. **Bold**: best.

Method	Motion deblurring	Gaussian deblurring
	PSNR↑/SSIM↑/FID↓/MSE↓	PSNR↑/SSIM↑/FID↓/MSE↓
BlindDPS	23.66 / 0.684 / 29.62 / 0.130	26.18 / 0.757 / 26.18 / 0.101
CVSA	**25.83 / 0.700 / 26.92 / 0.116**	**28.90 / 0.781 / 24.23 / 0.071**

FID by more than 50% on average compared with ILVR.

Table 3 shows the performance on **blind deblurring** task. Although the performance of CVSA on the four metrics is better than the previous SOTA method BlindDPS, the advantage is not obvious. Therefore, how to further improve the performance of CVSA in blind recovery is an interesting future direction. At the same time, its application can also be extended to other blind restoration tasks, such as imaging through turbulence.

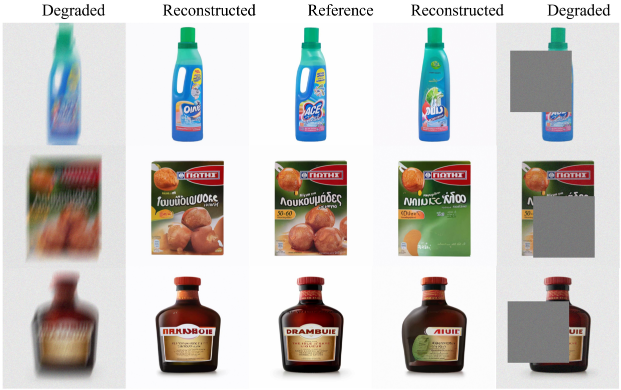

Figure 6. CVSA can be applied to the preprocessing of complex and fast automatic checkout for de-occlusion and de-blurring, improving checkout efficiency and commodity recognition performance.

CVSA is a model that can be applied to all kinds of datasets, not just images of human faces or dogs. Figure 6 shows that the image of products in the supermarket can also be restored very well. This preprocessing method can overcome hand occlusion and quick motion blurring problems in the automatic checkout, thereby improving performance and efficiency. This is a good **application** candidate for the algorithm in this paper.

6. Conclusion

By introducing a position-velocity space and a second-order Langevin inverse SDE conditioned on degraded images in this space, we propose an image restoration method called CVSA. The key point of this method is to realize the approximation of the difficult CVS by transferring the measurement probability at current moment to the initial moment. CVSA is suitable for linear and nonlinear tasks, non-blind and blind restoration, and does not require separate training for specific tasks. Experimental results show that CVSA is better than previous SOTA methods.

In our method, the velocity score estimates $p(\mathbf{y}|\hat{\mathbf{u}}_0)$ of the degraded image at the current moment is the lower bound of the true value $p(\mathbf{y}|\mathbf{u}_t)$. A more accurate estimate can be obtained if the remainder of the Jenssen inequality estimate is added, which we expect to be achieved by Monte Carlo sampling of the estimate. This is the direction we want to try in future work.

References

[1] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. *Proceedings of International Conference on Computer Vision*, 2023. 1

[2] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14367–14376, 2021. 2

[3] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 5

[4] Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye. Parallel diffusion models of operator and image for blind inverse problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6059–6069, 2023. 1, 6

[5] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 1, 2, 3, 6

[6] Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Fast diffusion sampler for inverse problems by geometric decomposition. *arXiv preprint arXiv:2303.05754*, 2023. 2, 6

[7] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. 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*, pages 12413–12422, 2022. 2

[8] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. *arXiv preprint arXiv:2112.07068*, 2021. 3, 4

[9] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10021–10030, 2023. 1

[10] Kostas Georgiadis, Giorgos Kordopatis-Zilos, Fotis Kalaganis, Panagiotis Migkotzidis, Elisavet Chatzilari, Valasia Panakidou, Kyriakos Pantouvakis, Savvas Tortopidis, Symeon Papadopoulos, Spiros Nikolopoulos, et al. Products-6k: a large-scale groceries product recognition dataset. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*, pages 1–7, 2021. 5

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020. 1

[13] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906. 3

[14] 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. 5

[15] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 1, 2, 6

[16] Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I2sb: Image-to-image schrödinger bridge. *arXiv preprint arXiv:2302.05872*, 2023. 1, 2

[17] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11020–11028, 2022. 1

[18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 5

[19] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 1

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[22] Dohoon Ryu and Jong Chul Ye. Pyramidal denoising diffusion probabilistic models. *arXiv preprint arXiv:2208.01864*, 2022. 2

[23] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[24] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 2

[25] Simo Särkkä and Arno Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019. 4

[26] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023. 2, 3

[27] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 2

[28] Phong Tran, Anh Tuan Tran, Quynh Phung, and Minh Hoai. Explore image deblurring via encoded blur kernel space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11956–11965, 2021. 5

[29] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022. 1, 3

[30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[31] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16293–16303, 2022. 2

[32] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022. 1