Uncertainty Visualization via Low-Dimensional Posterior Projections

Omer Yair, Elias Nehme & Tomer Michaeli
Technion - Israel Institute of Technology
Haifa, Israel
omeryair@gmail.com, seliasne@gmail.com, tomer.m@ee.technion.ac.il

Abstract

In ill-posed inverse problems, it is commonly desirable to obtain insight into the full spectrum of plausible solutions, rather than extracting only a single reconstruction. Information about the plausible solutions and their likelihoods is encoded in the posterior distribution. However, for high-dimensional data, this distribution is challenging to visualize. In this work, we introduce a new approach for estimating and visualizing posteriors by employing energy-based models (EBMs) over low-dimensional subspaces. Specifically, we train a conditional EBM that receives an input measurement and a set of directions that span some low-dimensional subspace of solutions, and outputs the probability density function of the posterior within that space. We demonstrate the effectiveness of our method across a diverse range of datasets and image restoration problems, showcasing its strength in uncertainty quantification and visualization. As we show, our method outperforms a baseline that projects samples from a diffusion-based posterior sampler, while being orders of magnitude faster. Furthermore, it is more accurate than a baseline that assumes a Gaussian posterior. Code is available at https://github.com/yairomer/PPDE

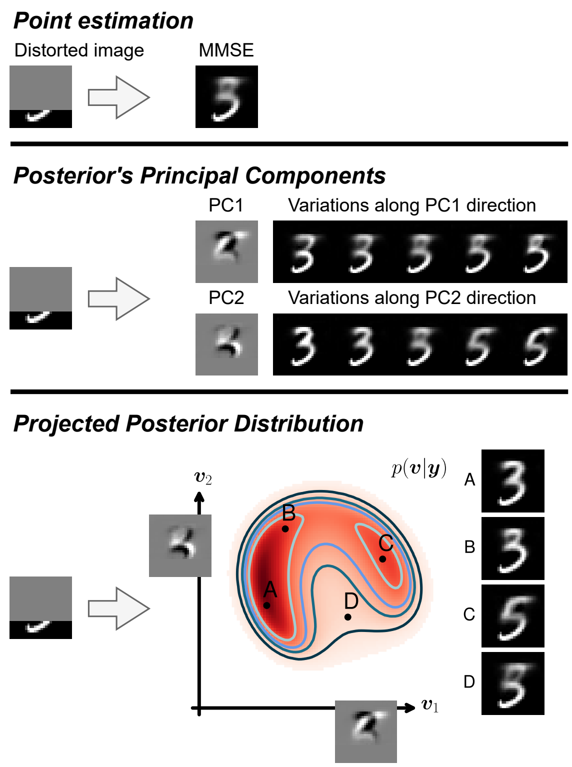


Figure 1. Informed uncertainty visualization. Point estimation methods receive a distorted image and output only a single solution, *e.g.,* the MMSE estimator (top). NPPC [41] complements MMSE estimators with input-adaptive uncertainty directions (principal components of the posterior) without modeling output likelihood (middle). Our method (bottom) learns the input-adaptive projected posterior distribution, facilitating a likelihood-informed uncertainty visualization.

1. Introduction

Interpreting and communicating prediction uncertainty plays a key role in advancing trustworthy models that could assist decision-makers. In the context of imaging, many practical problems are ill-posed, so that a range of plausible explanations exist for any given input. In such cases, it is beneficial to provide the user with tools to efficiently explore and visualize the set of all admissible solutions. This is especially crucial in safety-critical domains such as scientific and medical image analysis [9, 15, 44, 47], where mistaken predictions could influence human life.

The information about the plausible solutions and their likelihoods is fully encoded in the posterior distribution. However, high-dimensional posteriors are hard to estimate and practically impossible to visualize. One way to visualize uncertainty is to settle with generating samples from the posterior [5, 10, 26, 29, 32, 43, 49, 50, 59, 64]. However, navigating many samples per input is highly inefficient and often an impractical way of communicating uncertainty [12]. Indeed, in complex domains and high levels of uncertainty, users may need to examine hundreds of posterior samples to confidently confirm or refute their suspicion about the unobserved ground-truth image. Several works proposed to manipulate the sampling process to produce a small set of meaningful samples highlighting posterior diversity [12, 36, 54, 66]. Nonetheless, methods that are based on state-of-the-art (diffusion based) samplers are unacceptably slow due to their iterative sampling process.

Another way to summarize uncertainty in image restoration problems is via per-pixel heatmaps [1, 2, 21, 27, 63]. However, such maps ignore the correlations between pixels in the recovered image and thus typically result in non-semantic and unnecessarily inflated uncertainty estimates that are of limited practical use.

Recently, several works proposed to visualize posterior uncertainty by traversing along the principal components (PCs) of the posterior [4, 41]. These visualizations shed light on the main modes of variation along which the solution could vary given the input, and have thus been demonstrated to be natural for signals with strong correlations, like images. These methods are also mathematically appealing since one-dimensional statistics of the projected posterior along the PCs (like variances or quantiles) provide a tighter and more accurate description of the underlying uncertainty [4]. Nonetheless, methods in this category do not provide estimates for the likelihoods of the projected solutions within the low-dimensional subspace. This hinders the user's ability to interpret and visualize the effective support of the projected posterior.

In this work, we propose to model the projected posterior using a conditional energy-based model (EBM) that receives a degraded measurement and an affine subspace (parameterized by an origin point and a set of directions) and can output the projection of the posterior probability for any queried point within that subspace. This allows us to visualize the projected posterior within any low-dimensional space (*e.g.,* the one spanned by the dominant posterior PCs), facilitating informed navigation of posterior uncertainty in both one and two dimensions (Fig. 1). Our method, which we coin *projected posterior distribution estimation* (PPDE), is a general technique for uncertainty visualization that is seamlessly transferable across tasks and datasets. We demonstrate it with the affine subspace that passes through the point prediction provided by an MSE-optimized model and spanned by the posterior PCs provided by the NPPC method [41] and . However, our method is not constrained to this specific choice and can wrap around any point estimate and an accompanying set of directions.

As we illustrate, our proposed visualization approach cannot be achieved with methods that explicitly learn the high-dimensional posterior distribution, like conditional invertible models (*e.g.,* SRFlow [31]) or score-based models that enable the calculation of the exact likelihood using the probability flow ODE [57, 58]. In particular, the naive approach of evaluating and plotting the posterior along a 2D or 1D slice through the high-dimensional distribution gives meaningless results, as it will almost surely miss the high-probability regions in space (see Fig. 3 and the explanation in Sec. 3.1). By contrast, our approach corresponds to *projecting* the distribution onto the 2D or 1D slice (*i.e.,* integrating over all other dimensions), an effect that is practically impossible to achieve with models that output the density in the high-dimensional space.

We demonstrate the practical benefit of our method on multiple inverse problems in imaging. Our learned posteriors are quantitatively compared to a Gaussian approximation obtained from [41], as well as to a baseline that projects samples generated by a diffusion-based posterior sampler and applies kernel-density estimation (KDE) in the low-dimensional space. In both cases, we show a significant improvement in sample log-likelihood across tasks and datasets. This illustrates the benefit offered by our method compared to works that approximate the posterior with a Gaussian [13, 37, 39, 42] or attempt to rely on posterior samplers [4], setting forth a new approach to proper uncertainty visualization.

2. Related Work

In the deep learning literature (*e.g.,* [16, 27]), predictive uncertainty is often decomposed into two main sources: (i) epistemic/model uncertainty [7, 24, 28, 30, 34, 40, 45, 46, 52], which stems from imperfect knowledge of model parameters and can be reduced by acquiring additional training data, and (ii) aleatoric/data uncertainty [2, 26, 27, 43, 63, 64], which is inherent to the task (*e.g.,* due to the degradation and measurement noise), and cannot be reduced even with infinite data. We focus on the latter.

Until recently, the majority of methods for uncertainty quantification in imaging problems focused on per-pixel estimates *e.g.,* in the form of variance heatmaps [27] or confidence intervals [2]. However, these techniques ignore inter-pixel correlations, which are very important in visual data. Distribution-free methods such as risk-controlling prediction sets (RCPS) [3] treat all pixels jointly by aggregating and controlling their risk (*e.g.,* in image segmentation). However, RCPS only provides upper and lower bounds on the solution set without the ability to navigate within-set possibilities, and it also requires an extra data split which may not be readily available in data-scarce settings. A follow-up work [53], employed a similar idea in the latent space of StyleGAN, demonstrating some informative visualizations of uncertainty. However, this work is limited in its ability to treat real images and it requires previously identified disentangled latent directions.

Another alternative for exploring data uncertainty is posterior sampling using conditional generative models such as

[5, 10, 11, 25, 26, 29, 31, 32, 43, 49, 50, 59, 64], with score-based/diffusion models [20, 23, 55, 56] pulling ahead in the last two years. Posterior sampling can in principle offer a large solution set for any given input, which can, in turn, be summarized to useful uncertainty estimates *e.g.,* using PCA [4]. However, this strategy is extremely slow with unbearable run times for modern state-of-the-art models, despite promising recent efforts to speed them up [33, 38, 51, 60].

Some works proposed to approximate posteriors with more simple distributions that account for pixel correlations (*e.g.,* correlated Gaussians [13, 37, 39, 42]), which are fast to infer. Most recently, two methods were proposed to output the top PCs of the posterior distribution directly without imposing a distributional assumption [35, 41]. Specifically, NPPC [41] was shown to provide useful uncertainty estimates while being extremely fast. Here we use it as a building block to demonstrate our method.

A visualization of projected posterior distributions, as we propose here, can theoretically be achieved by generating many posterior samples with some stochastic inverse-problem solver (*e.g.,* [64]), projecting them onto the desired subspace, and estimating the distribution of the low-dimensional projections using *e.g.,* kernel density estimation (KDE). However, this approach is impractical for real-world applications as the computational burden of generating a sufficient number of samples is very high with state-of-the-art posterior samplers. Furthermore, as we show, our method outperforms this approach in terms of average negative log-likelihood (NLL) even when using a large sample.

Another way to obtain an approximation of the projected posterior was recently proposed in the concurrent work of [35]. Specifically, this work presents a training-free method for estimating the one-dimensional projected posterior distribution along any desired direction in the task of Gaussian denoising. Our approach offers two key advantages over this work: (i) It is applicable to arbitrary inverse problems and not only to Gaussian denoising; (ii) It is applicable to arbitrary subspace dimensions (*e.g.,* 2D) and not constrained to only 1D projections.

3. Method

3.1. Projecting the posterior distribution

Our goal is to visualize the uncertainty when predicting a signal $\boldsymbol{x}$ based on measurements $\boldsymbol{y}$. In the context of imaging, $\boldsymbol{y}$ often represents a degraded version of $\boldsymbol{x}$ (*e.g.,* noisy, blurry). We assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ are realizations of random vectors $\mathbf{x}$ and $\mathbf{y}$, respectively. The desired uncertainty we aim to explore is therefore encapsulated in the posterior distribution $p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y})$. As mentioned earlier, this distribution is typically defined over a high-dimensional space, and therefore visualizing it requires projection onto a lower-dimensional manifold (*e.g.,* 1D or 2D).

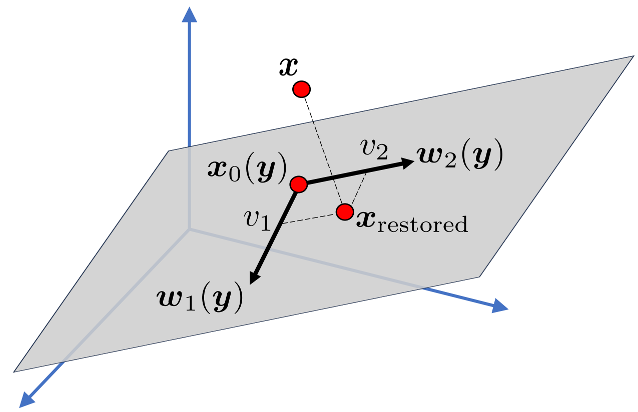


Figure 2. Notations illustration. An image $\boldsymbol{x}$ is projected onto a measurement-adaptive affine subspace $\mathcal{A}(\boldsymbol{y})$, and represented by its projection coefficients $\boldsymbol{v} = [v_1, v_2]^\top$. The restored image within the subspace is denoted by $\boldsymbol{x}_{\text{restored}}$.

It is important to note that the manifold that best suits uncertainty visualization may generally depend on the observation $\boldsymbol{y}$. Here we choose the manifold to be an input-dependent affine subspace of the form

$$\left\{\boldsymbol{x} : \boldsymbol{x} = \boldsymbol{x}_0(\boldsymbol{y}) + \boldsymbol{W}(\boldsymbol{y})\boldsymbol{v},\ \boldsymbol{v} \in \mathbb{R}^K\right\}, \tag{1}$$

where $\boldsymbol{x}_0(\boldsymbol{y})$ is an origin point and $\boldsymbol{W}(\boldsymbol{y})$ is a matrix with K orthonormal columns,

$$\boldsymbol{W}(\boldsymbol{y}) = \begin{bmatrix} | & | & & | \\ \boldsymbol{w}_1(\boldsymbol{y}) & \boldsymbol{w}_2(\boldsymbol{y}) & \dots & \boldsymbol{w}_K(\boldsymbol{y}) \\ | & | & & | \end{bmatrix}. \tag{2}$$

For conciceness, we denote the pair $\{\boldsymbol{x}_0(\boldsymbol{y}), \boldsymbol{W}(\boldsymbol{y})\}$ by

$$\mathcal{A}(\boldsymbol{y}) = \{\boldsymbol{x}_0(\boldsymbol{y}), W(\boldsymbol{y})\} \tag{3}$$

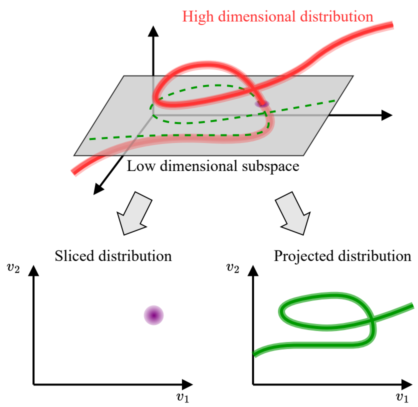
and with slight abuse of terminology, refer to $\mathcal{A}(\boldsymbol{y})$ as our affine subspace (or just subspace). Our method can be used along with any method of producing a subspace $\mathcal{A}(\boldsymbol{y})$ for a given input $\boldsymbol{y}$. Here we choose $\boldsymbol{x}_0(\boldsymbol{y})$ to be a minimum MSE (MMSE) predictor, and use the NPPC method [41] for producing $\boldsymbol{W}(\boldsymbol{y})$ (see Sec. 3.2).

For a given input $\boldsymbol{y}$, our goal is to visualize the posterior distribution of the coordinates of $\mathbf{x}$ projected onto the subspace $\mathcal{A}(\boldsymbol{y})$, which we denote by

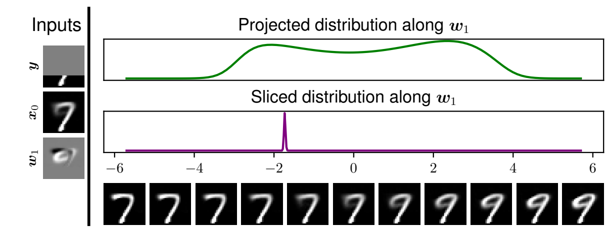
$$\mathbf{v} = \boldsymbol{W}(\boldsymbol{y})^\top(\mathbf{x} - \boldsymbol{x}_0(\boldsymbol{y})). \tag{4}$$

Namely, we are interested in the *projected posterior distribution* (PPD) $p_{\mathbf{v}|\mathbf{y}}(\boldsymbol{v}|\boldsymbol{y})$. When exploring this distribution, a user can visually inspect any particular $\boldsymbol{v}$ by projecting it back into pixel space as

$$\boldsymbol{x}_{\text{restored}} = \boldsymbol{x}_0(\boldsymbol{y}) + \boldsymbol{W}(\boldsymbol{y})\boldsymbol{v}. \tag{5}$$



(a) Schematic illustration.



(b) Digit inpainting.

Figure 3. Posterior slicing vs. projection. (a) Schematic illustrating the difference between the slice and the projection of a high-dimensional manifold onto a low-dimensional subspace. (b) Sliced and projected posterior comparison for a 1D affine subspace defined by the posterior mean and the first PC in the task of digit inpainting. The sliced posterior has been computed using an EBM.

Figure 2 summarizes our notations.

It is instructive to note the difference between the PPD and a sliced posterior distribution (*i.e.,* the high-dimensional posterior evaluated on $\mathcal{A}(\boldsymbol{y})$). Any $\boldsymbol{x}$ can be expressed in terms of its projection onto $\mathcal{A}(\boldsymbol{y})$ and its projection onto the orthogonal complement of $\mathcal{A}(\boldsymbol{y})$ as $\boldsymbol{x} = \boldsymbol{x}_0(\boldsymbol{y}) + \boldsymbol{W}(\boldsymbol{y})\boldsymbol{v} + \boldsymbol{W}^{\perp}(\boldsymbol{y})\boldsymbol{u}$. Here, $\boldsymbol{W}^{\perp}(\boldsymbol{y})$ is a matrix with orthonormal columns that span $\text{Range}^{\perp}\{\boldsymbol{W}(\boldsymbol{y})\}$, and $\boldsymbol{u}$ corresponds to the coordinates within that subspace. Now, the sliced posterior distribution corresponds to evaluating the posterior at $\boldsymbol{u} = 0$ in this decomposition. Namely, it is given by[1]

$$f^{\text{sliced}}(\boldsymbol{v}|\boldsymbol{y}) = p_{\mathbf{x}|\mathbf{y}}\left(\boldsymbol{x}_0(\boldsymbol{y}) + \boldsymbol{W}(\boldsymbol{y})\boldsymbol{v}|\boldsymbol{y}\right). \tag{6}$$

In contrast, the PPD at a point $\boldsymbol{v}$ corresponds to the integral of the posterior over all points whose projection onto $\mathcal{A}(\boldsymbol{y})$ has coordinates $\boldsymbol{v}$. In other words, the PPD is the marginal distribution of $\boldsymbol{v}$ produced by integrating out $\boldsymbol{u}$,

$$p_{\mathbf{v}|\mathbf{y}}(\boldsymbol{v}|\boldsymbol{y}) = \int p_{\mathbf{x}|\mathbf{y}}\left(\boldsymbol{x}_0(\boldsymbol{y}) + \boldsymbol{W}(\boldsymbol{y})\boldsymbol{v} + \boldsymbol{W}^{\perp}(\boldsymbol{y})\boldsymbol{u}\middle|\boldsymbol{y}\right) d\boldsymbol{u}. \tag{7}$$

[1]The function f^{sliced} is not a density function as it is not normalized.

There exist settings in which f^{sliced} equals $p_{\mathbf{v}|\mathbf{y}}$ (up to a normalization constant). This is the case, for example, when $\mathbf{u}$ and $\mathbf{v}$ are statistically independent given $\mathbf{y}$. However, in imaging inverse problems, this seems to rarely be the case. Indeed, the distribution of natural images is known to be concentrated near a low-dimensional manifold within the high-dimensional ambient space. Therefore, any low-dimensional subspace almost always misses the high-density regions of the distribution. On the other hand, the PPD provides meaningful information regardless of whether the slice passes through the high-density regions of the posterior or not. This is illustrated in Fig. 3.

3.2. Choosing the subspace

Selecting an appropriate subspace is crucial for producing a meaningful PPD visualization. The ideal subspace should be either 1D or 2D to facilitate visualization. Moreover, the origin $\boldsymbol{x}_0(\boldsymbol{y})$ and directions $\boldsymbol{W}(\boldsymbol{y})$ should disclose as much posterior variance as possible, hopefully in a semantically meaningful manner. A natural choice is to take the origin to be the minimum mean square error (MMSE) prediction, $\boldsymbol{x}_0(\boldsymbol{y}) \approx \mathbb{E}[\mathbf{x}|\mathbf{y} = \boldsymbol{y}]$, and the directions $\boldsymbol{w}_1(\boldsymbol{y}), \dots, \boldsymbol{w}_K(\boldsymbol{y})$ to be the top K PCs of the posterior. NPPC [41] outputs both $\boldsymbol{x}_0(\boldsymbol{y})$ and $\boldsymbol{w}_1(\boldsymbol{y}), \dots, \boldsymbol{w}_K(\boldsymbol{y})$, rendering it a suitable candidate for fast training and inference. In Fig. 3, we exemplify this choice for the task of image inpainting where moving about the origin $\boldsymbol{x}_0(\boldsymbol{y})$ along the selected direction $\boldsymbol{w}_1(\boldsymbol{y})$ changes digit identity.

3.3. Architecture

We propose to learn the conditional distribution $p_{\mathbf{v}|\mathbf{y}}(\boldsymbol{v}|\boldsymbol{y})$ using a conditional EBM. Unlike normalizing flows, EBMs do not require specialized architectures, and are thus more flexible in their design. EBMs are usually trained with variants of contrastive divergence (CD) [6, 8, 19, 61] or noise contrastive estimation (NCE) [18], with various recent efforts to stabilize and improve training [14, 17, 67].

As depicted in Fig. 4, we propose to design the architecture using two parts: A (heavy) feature extractor with a classifier-style architecture and a lightweight small MLP with a few linear layers. The feature extractor receives the measurement $\boldsymbol{y}$ and the set of vectors defining the subspace $\mathcal{A}(\boldsymbol{y})$, and outputs a 1D feature vector $\boldsymbol{h}(\boldsymbol{y})$. This feature vector is then fed along with a queried point $\boldsymbol{v}$ into the second part that outputs $p_{\mathbf{v}|\mathbf{y}}(\boldsymbol{v}|\boldsymbol{y})$. Using this architecture we expect the intermediate feature vector to encapsulate all the information regarding the 2D distribution and have the MLP just translate this data into a distribution function.

This segregation of the architecture serves the purpose of lowering computational complexity. The rationale behind it is the following: For a given measurement $\boldsymbol{y}$, we would like to query $p_{\mathbf{v}|\mathbf{y}}(\boldsymbol{v}|\boldsymbol{y})$ for many values of $\boldsymbol{v}$. This is true both during training when running MCMC chains to

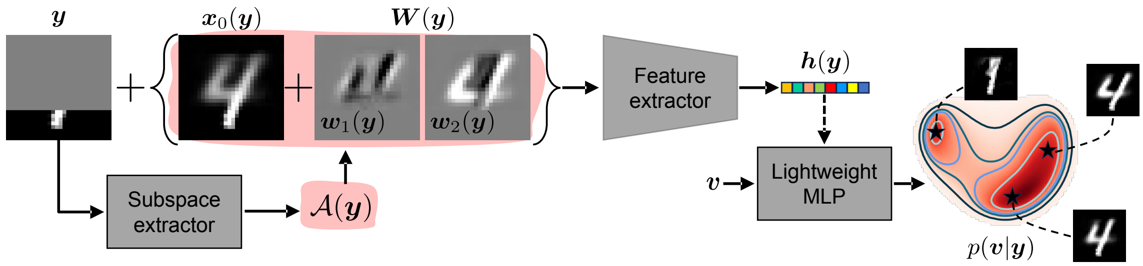
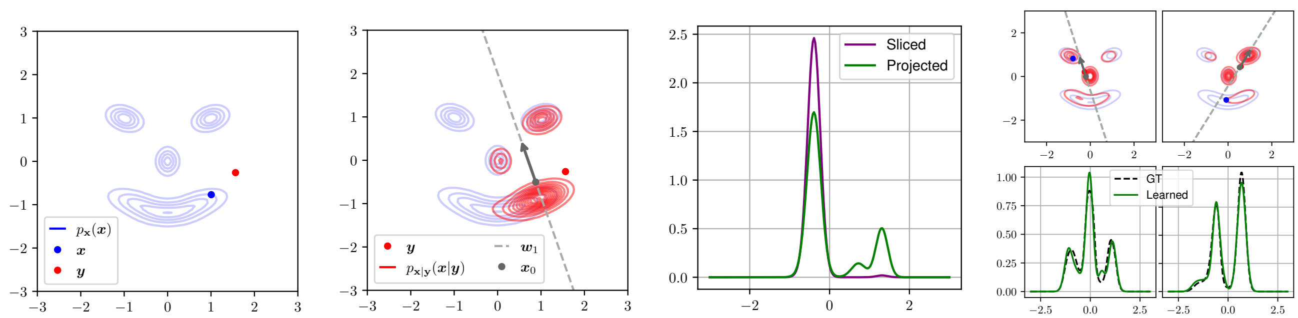


Figure 4. Architecture overview. The degraded image $\boldsymbol{y}$ is first fed to a pre-selected subspace extractor that outputs an input-adaptive subspace $\mathcal{A}(\boldsymbol{y}) = \{\boldsymbol{W}(\boldsymbol{y}), \boldsymbol{x}_0(\boldsymbol{y})\}$. Afterward, the degraded image $\boldsymbol{y}$ and the extracted subspace are fed to a feature extractor that outputs a feature vector $h(\boldsymbol{y})$. The resulting $h(\boldsymbol{y})$ modulates a lightweight MLP that outputs $p(\boldsymbol{v}|\boldsymbol{y})$ for any query $\boldsymbol{v}$. The resulting projected distribution can then be navigated to visualize posterior uncertainty.



(a) $p_{\mathbf{x}}(\boldsymbol{x})$ and a sample $(\boldsymbol{x}_i, \boldsymbol{y}_i)$. (b) $p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y})$ and $\mathcal{A}(\boldsymbol{y})$. (c) $f^{\text{sliced}}(\boldsymbol{v}|\boldsymbol{y})$ vs. $p_{\mathbf{v}|\mathbf{y}}(\boldsymbol{v}|\boldsymbol{y})$. (d) Learned vs. GT $p_{\mathbf{v}|\mathbf{y}}(\boldsymbol{v}|\boldsymbol{y})$.

Figure 5. Gaussian mixture denoising. (a) Prior distribution $p_{\mathbf{x}}(\boldsymbol{x})$, and a sample from the joint distribution $p_{\mathbf{x},\mathbf{y}}(\boldsymbol{x}, \boldsymbol{y})$. (b) Full posterior distribution $p_{\mathbf{x}|\mathbf{y}}(\boldsymbol{x}|\boldsymbol{y})$, and the selected 1D subspace $\mathcal{A}(\boldsymbol{y}) = \{\boldsymbol{x}_0(\boldsymbol{y}), \boldsymbol{W}(\boldsymbol{y})\}$. (c) Sliced $f^{\text{sliced}}(\boldsymbol{v}|\boldsymbol{y})$ (purple) vs. projected $p_{\mathbf{v}|\mathbf{y}}(\boldsymbol{v}|\boldsymbol{y})$ (green) posterior distribution along the 1D subspace (dashed gray line in (b)). (d) Comparison of the GT and the learned projected posterior.

produce contrastive samples, and during testing when evaluating the PPD on a grid of $\boldsymbol{v}$'s for plotting. Therefore, for a given measurement $\boldsymbol{y}$ and subspace $\mathcal{A}(\boldsymbol{y})$, this inner separation of the architecture allows us to evaluate the feature extractor only once while the lightweight MLP can be queried dozens of times incurring only a minor computational burden. The conditioning mechanism we adopt here is similar to the one employed for timestep encoding in diffusion models [20], with adaptive shifting and scaling of features reminiscent of AdaIN [22].

To train our conditional EBM, we use a variant of CD [19]. In its basic form, CD fits a given parametric function $E(\boldsymbol{v}; \boldsymbol{\theta})$ to the (unnormalized) negative log distribution function $-\log \tilde{p}_{\mathbf{v}}(\boldsymbol{v})$. In each training step, an MCMC process (*e.g.,* Langevin dynamics) is applied to a batch of data samples $\boldsymbol{v}$ to produce contrastive samples $\boldsymbol{v}_{\text{neg}}$. The model's parameters $\boldsymbol{\theta}$ are then updated as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \nabla_{\boldsymbol{\theta}} \left[\mathbb{E}_{\boldsymbol{v}}[E(\boldsymbol{v}; \boldsymbol{\theta})] - \mathbb{E}_{\boldsymbol{v}_{\text{neg}}}[E(\boldsymbol{v}_{\text{neg}}; \boldsymbol{\theta})] \right] \Big|_{\boldsymbol{\theta}_t}. \tag{8}$$

This process gradually increases the likelihood that the model assigns to the samples in the training set and decreases the likelihood that the model assigns to their contrastive counterparts, until convergence. At test time, the resulting model can output the probability density $p_{\mathbf{v}}(\boldsymbol{v})$ (up to an unknown normalization constant) for any queried $\boldsymbol{v}$.

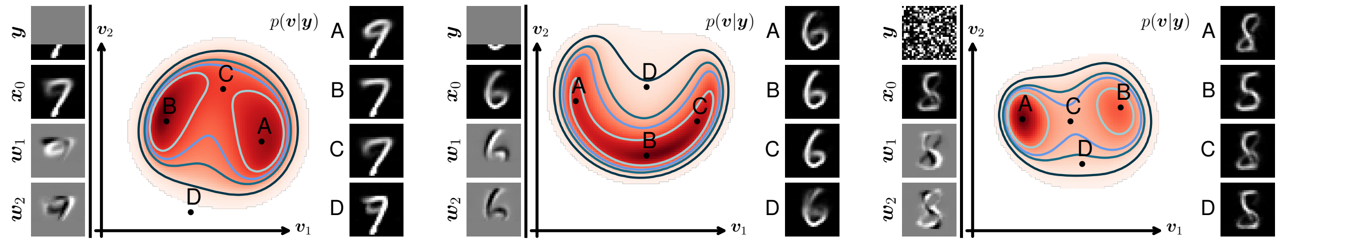
The only adjustment needed for using CD to learn a *conditional* distribution, is to add the feature vector $\boldsymbol{h}(\boldsymbol{y})$ as an additional input to the model, so that $E(\boldsymbol{v}, \boldsymbol{h}(\boldsymbol{y}); \boldsymbol{\theta}) = -\log \tilde{p}_{\mathbf{v}|\mathbf{y}}(\boldsymbol{v}|\boldsymbol{y})$. Here, we use the CD variant from [65], which learns a series of distributions by employing multiple levels of additive noise, similarly to diffusion models.

4. Experiments

We now illustrate our method on several tasks and datasets. In all our experiments with visual data, the subspace $\mathcal{A}(\boldsymbol{y})$ is spanned by the top two PCs of the posterior, obtained from NPPC [41]. Full details about the architectures, the per-task settings and additional results are in the appendix.

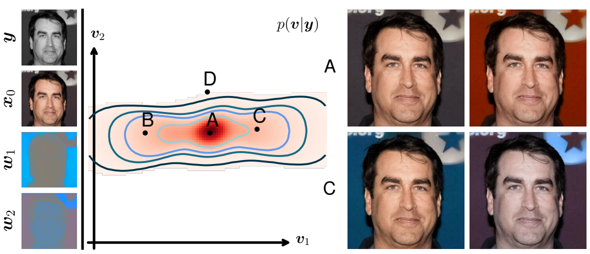
4.1. Toy example in 2D

Figure 5 illustrates the application of PPDE to a 2D toy example with 1D projections. Here, $\mathbf{x}$ is sampled from a mixture of six Gaussians (arranged as a face), and $\mathbf{y}$ is a noisy version of $\mathbf{x}$. The prior distribution $p_{\mathbf{x}}(\boldsymbol{x})$ and an exem-

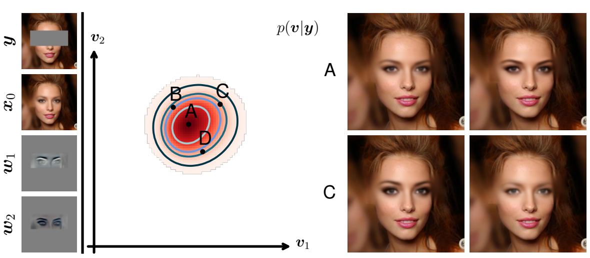


(a) Inpainting (b) Inpainting (c) Denoising

Figure 6. MNIST inpainting and denoising. (a),(b) Application of PPDE to image inpainting from only the 8 bottom pixel rows of the image. The areas within the contours correspond to $50\%, 80\%, 90\%, 98\%$ of the total probability mass. Note how the PPD reveals posterior multi-modality (*e.g.,* it shows the digit is either a “7” or a “9”). Similarly, it also reveals intra-digit variations such as the circular part of the digit “6”. (c) Application of PPDE to image denoising with an extreme noise level of $\sigma_\varepsilon = 1$. PPDE reveals a similar posterior multi-modality (*e.g.,* the digit is either an “8” or a “5”).



(a) Colorization



(b) Inpainting

Figure 7. Restoration of face images from CelebA-HQ. The PPD reveals a variety of plausible background and skin colors in colorization (a), and a range of plausible eye-shapes in inpainting (b). In addition, the PPD shows the likelihoods of the different restorations, distinguishing between plausible and less plausible solutions (such as points A and D, respectively, in (a)).

plar sample from the joint distribution $(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\mathbf{x},\mathbf{y}}$ are presented in Fig. 5a. For this simple case, the posterior distribution can be calculated analytically (see appendix for the derivation) and is also a mixture of Gaussians (Fig. 5b). Note that for this illustration the posterior itself is two-dimensional, and therefore can be plotted and visualized in full. However, as explained earlier, this is not the case for high-dimensional distributions and is precisely the problem our method aims to solve by projection. Here, this example serves as a sanity check enabling us to benchmark our results against a known ground truth.

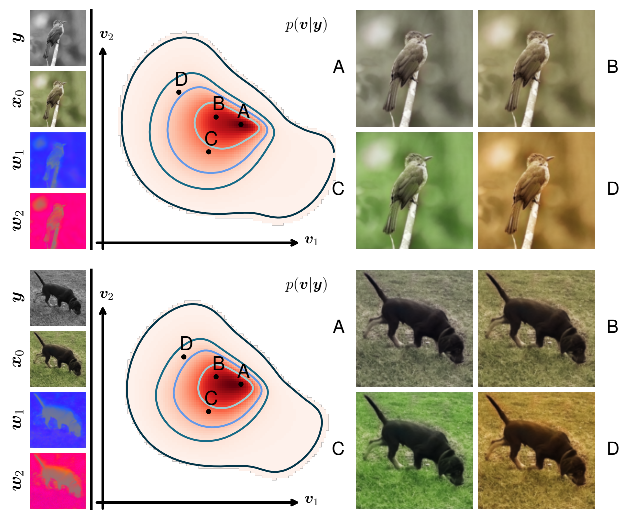


Figure 8. Application of PPDE to natural image colorization on ImageNet, revealing the distribution of different solutions.

To demonstrate our method, we selected an arbitrary 1D subspace on which we project and plot the posterior. The origin point and a direction defining this subspace are shown in Fig. 5b. As can be seen in Fig. 5c, the PPD on the selected 1D subspace (green line) can provide us with insights regarding the behavior of the full posterior, such as the fact that it is composed of three modes with different weights. In addition, Fig. 5c also shows the probability density of the high dimensional (2D) posterior along the 1D line, which we refer to as the sliced posterior (purple line). Comparing both plots, it is easy to see why the sliced posterior is less informative as not all posterior modes appear in the slice projection.

4.2. Common restoration problems

Handwritten Digits. Figure 6 demonstrate PPDE on inpainting and denoising of handwritten digits from the

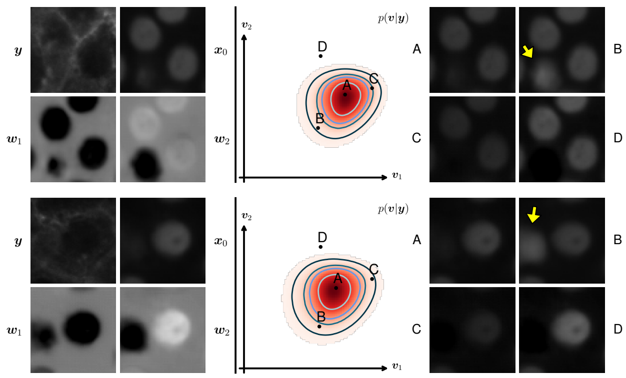


Figure 9. Biological image-to-image transfer. The PPD outlines the probability of possible solutions for the task of reconstructing fluorescent microscopy images using one type of dye from another. As seen in these examples, the PPD can assist in visualizing the fact that a given input might produce plausible solutions with varying amounts of cells (highlighted by yellow arrows).

MNIST dataset. In the inpainting task, we used a mask that covers the top 70% of the image, and in denoising we added noise of standard deviation $\sigma_\varepsilon = 1$. As can be seen, for both tasks, PPDE reveals the inherent bimodality of the posterior with two different possible digits being likely given the measurement $\boldsymbol{y}$. In addition, note that when the digit identity is easier to infer from $\boldsymbol{y}$, PPDE reveals uncertainty corresponding to intricate intra-digit variations.

Faces. Figure 7 presents results for colorization and inpainting of face images from the CelebA-HQ 256×256 dataset. In the appendix, we further provide results for $8\times$ super-resolution with a bicubic downsampling filter and a crop area taken from SRFlow [31]. As can be seen, PPDE reveals the unique structure of the projected posterior distribution, which outlines the region of plausible solutions in the projected space.

Natural images. Figure 8 shows colorization results for natural images from the ImageNet 1K dataset [48]. The distributions in this case are often not multi-modal; however, they are also far from being axis-aligned Gaussians. This demonstrate the practical benefit of PPD visualizations. More examples are available in the Appendix.

Biological image-to-image translation. Finally, we test our method on a nonlinear image-to-image translation task. Here, we evaluate PPDE on the dataset from [62], where a microscopic biological specimen imaged with one fluorescent dye is transformed to appear as if it was imaged with another. This "virtual staining" task is of immense importance in biological imaging, as it enables correlative imaging using fewer fluorescent dyes. Figure 9 presents the results of PPDE on this challenging task. The resulting distribution discloses a range of plausible solutions in the projected space corresponding to different numbers of cells. This example demonstrates how PPDE can be used as a quantitative tool for assessing the variance in cell counting applications.

Table 1. Negative log likelihood comparison. NPPC and our method were computed on the entire test set. The KDE approach was computed for 100 test images due to computational constraints.

Task	NPPC (full test)	KDE (100 samples)	Ours (full test)
MNIST			
Inpainting	4.19 ± 0.02	3.9 ± 0.09	**3.72 ± 0.01**
Denoising	3.41 ± 0.02	3.26 ± 0.07	**3.04 ± 0.01**
CelebA			
Inpainting Eyes	6.38 ± 0.08	8.278 ± 1.1	**5.37 ± 0.03**
Colorization	10.3 ± 0.2	7.7 ± 0.3	**7.39 ± 0.03**
Super-resolution	5.12 ± 0.3	-	**2.84 ± 0.1**
ImageNet			
Colorization	10.15 ± 0.06	-	**6.86 ± 0.01**

4.3. Quantitative comparisons

Comparison to NPPC. NPPC outputs the posterior mean $\boldsymbol{x}_0(\boldsymbol{y})$, alongside the top K posterior PCs $\boldsymbol{W}(\boldsymbol{y})$ and variances $\{\sigma_k^2(\boldsymbol{y})\}_{k=1}^K$ for a given input image $\boldsymbol{y}$. While it was originally proposed as a distribution-free method, the predicted variances can be used to construct a 2D Gaussian approximation of the projected posterior covariance. Let $\boldsymbol{\Sigma}(\boldsymbol{y})$ denote the diagonal matrix that has $[\sigma_1^2(\boldsymbol{y}), \sigma_2^2(\boldsymbol{y})]^\top$ along its main diagonal. Then, the projected posterior distribution can be approximated as

$$p_{\mathbf{v}|\mathbf{y}}(\boldsymbol{v}|\boldsymbol{y}) \approx \mathcal{N}(\boldsymbol{v}; 0, \boldsymbol{\Sigma}(\boldsymbol{y})). \quad (9)$$

We compare our results to this baseline in Table 1, showing a significant improvement in sample negative log likelihood (NLL) in favor of PPDE.

Comparison to posterior sampling + KDE. The aim of our density estimator is to estimate the PPD in one or two dimensions. Given a method that can sample from the posterior (*e.g.,* DDNM [64]), a possible baseline approach to achieve this goal is as follows: (i) Generate N samples from the posterior $\{\boldsymbol{x}_i\}$; (ii) Compute PCA using $\{\boldsymbol{x}_i\}$ and project the samples onto the space spanned by the top two PCs; (iii) Estimate the PPD in 2D from the projected sam-

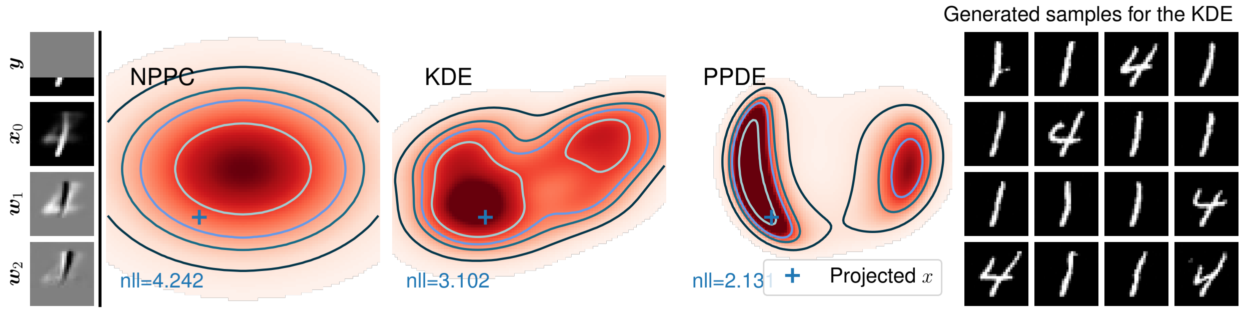


Figure 10. PPD comparison on MNIST. On the left we plot the input measurement $\boldsymbol{y}$, and the input-adaptive subspace $\mathcal{A}(\boldsymbol{y}) = \{\boldsymbol{x}_0(\boldsymbol{y}), \boldsymbol{w}_1(\boldsymbol{y}), \boldsymbol{w}_2(\boldsymbol{y})\}$. In the middle, we plot the estimated PPDs obtained from a Gaussian approximation using NPPC, from the KDE approach, and from our PPDE. The projected ground truth posterior sample $\boldsymbol{x}$ is marked by a blue cross, and its likelihood under the estimated PPD is shown on the bottom left of the corresponding plot. On the right, we show a few posterior samples out of the 100 used in the KDE approach. These were obtained using an EBM trained on MNIST.

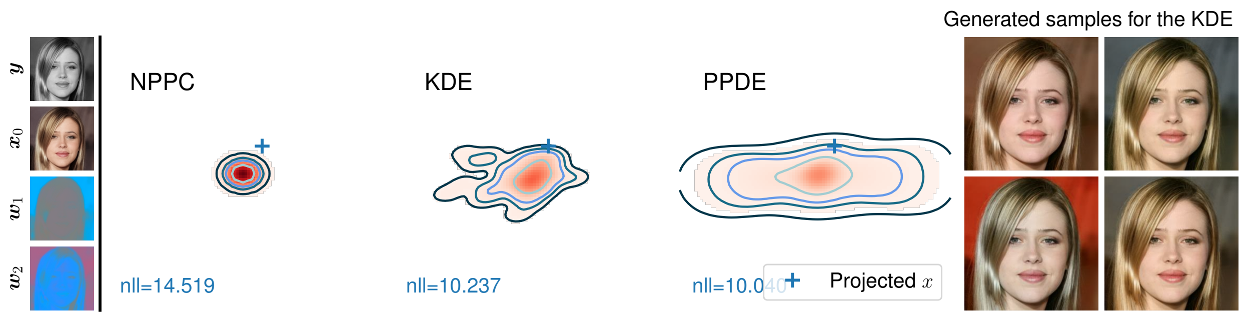


Figure 11. PPD comparisons on CelebA-HQ. The layout is the same as Fig. 10. Posterior samples on the right were obtained using DDNM.

ples $\{\boldsymbol{v}_i\}$ using standard techniques such as *kernel density estimation* (KDE). Comparisons against this baseline are reported in Table 1. For producing the KDE samples from the posterior we have used: a conditional EBM for MNIST, DPS [10][2] for face inpainting[3], and DDNM for face colorization, selecting the best sampler according to their performance and available trained models. For each task/dataset we randomly chose 100 test images and sampled 100 posterior samples per image. Compared to this (impractical) baseline, our technique is orders of magnitude faster as this approach requires costly repeated posterior sampling to estimate the projected density with enough accuracy. Figures 10 and 11 present visual examples of the approximated PPD with NPPC, KDE, and PPDE. PPDE exhibits superior NLL on average while being extremely faster than KDE, and only slightly slower than NPPC.

5. Discussion and conclusion

We presented a new method for *informed* uncertainty visualization and demonstrated its practical benefit across various tasks and datasets. Compared to the proposed baselines, our method brought a significant advantage both qualitatively and quantitatively in terms of improved sample NLL. The main limitation of our method lies in the fact that visualization beyond two or three dimensions becomes very difficult. This limits our supported analysis to 2 or 3 PCs, which may be insufficient to faithfully represent rich posteriors. Our method is particularly suited for tasks where posteriors have a spectrally-concentrated covariance, such as in image colorization. However, for severely ill-posed inverse problems where a large number of PCs is required to project the posterior in an informative manner, our strategy becomes impractical, warranting further research of condensed alternative representations of uncertainty.

[2]For DPS we have used the public model trained on FFHQ (and not on CelebA). Therefore, the model is not expected to produce accurate samples. The comparison to the PPD it produces is presented here only as an additional rough baseline.

[3]We also experimented with MAT [29], which gave inferior results.

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020. 2

[2] Anastasios N Angelopoulos, Amit P Kohli, Stephen Bates, Michael I Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, 2022. 2

[3] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68 (6):1–34, 2021. 2

[4] Omer Belhasin, Yaniv Romano, Daniel Freedman, Ehud Rivlin, and Michael Elad. Principal uncertainty quantification with spatial correlation for image restoration problems. *arXiv preprint arXiv:2305.10124*, 2023. 2, 3

[5] Matthew Bendel, Rizwan Ahmad, and Philip Schniter. A regularized conditional gan for posterior sampling in inverse problems. *arXiv preprint arXiv:2210.13389*, 2022. 1, 3

[6] Yoshua Bengio and Olivier Delalleau. Justifying and generalizing contrastive divergence. *Neural computation*, 21(6): 1601–1621, 2009. 4

[7] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015. 2

[8] Miguel A Carreira-Perpinan and Geoffrey Hinton. On contrastive divergence learning. In *International workshop on artificial intelligence and statistics*, pages 33–40. PMLR, 2005. 4

[9] Eric M Christiansen, Samuel J Yang, D Michael Ando, Ashkan Javaherian, Gaia Skibinski, Scott Lipnick, Elliot Mount, Alison O'neil, Kevan Shah, Alicia K Lee, et al. In silico labeling: predicting fluorescent labels in unlabeled images. *Cell*, 173(3):792–803, 2018. 1

[10] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 3, 8

[11] Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Fast diffusion sampler for inverse problems by geometric decomposition. *arXiv preprint arXiv:2303.05754*, 2023. 3

[12] Noa Cohen, Hila Manor, Yuval Bahat, and Tomer Michaeli. From posterior sampling to meaningful diversity in image restoration. *arXiv preprint arXiv:2310.16047*, 2023. 1, 2

[13] Garoe Dorta, Sara Vicente, Lourdes Agapito, Neill DF Campbell, and Ivor Simpson. Structured uncertainty prediction networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5477–5485, 2018. 2, 3

[14] Yilun Du, Shuang Li, Joshua B Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy based models. In *Energy Based Models Workshop-ICLR 2021*, 2021. 4

[15] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019. 1

[16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2

[17] Ruiqi Gao, Yang Song, Ben Poole, Ying Nian Wu, and Diederik P Kingma. Learning energy-based models by diffusion recovery likelihood. In *International Conference on Learning Representations*, 2020. 4

[18] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 4

[19] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8): 1771–1800, 2002. 4, 5

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 5

[21] Eliahu Horwitz and Yedid Hoshen. Conffusion: Confidence intervals for diffusion models. *arXiv preprint arXiv:2211.09795*, 2022. 2

[22] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 5

[23] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. 3

[24] Pavel Izmailov, Wesley J Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Subspace inference for bayesian deep learning. In *Uncertainty in Artificial Intelligence*, pages 1169–1179. PMLR, 2020. 2

[25] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems*, 34:21757–21769, 2021. 3

[26] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3

[27] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 2

[28] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2

[29] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022. 1, 3, 8

[30] Christos Louizos and Max Welling. Multiplicative normalizing flows for variational bayesian neural networks. In *International Conference on Machine Learning*, pages 2218–2227. PMLR, 2017. 2

[31] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *ECCV*, 2020. 2, 3, 7

[32] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 1, 3

[33] Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021. 3

[34] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018. 2

[35] Hila Manor and Tomer Michaeli. On the posterior distribution in denoising: Application to uncertainty quantification. *arXiv preprint arXiv:2309.13598*, 2023. 3

[36] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1429–1437, 2019. 2

[37] Chenlin Meng, Yang Song, Wenzhe Li, and Stefano Ermon. Estimating high order gradients of the data distribution by denoising. *Advances in Neural Information Processing Systems*, 34:25359–25369, 2021. 2, 3

[38] Chenlin Meng, Ruiqi Gao, Diederik P Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022. 3

[39] Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in neural information processing systems*, 33:12756–12767, 2020. 2, 3

[40] Radford M Neal. *Bayesian learning for neural networks*. Springer Science & Business Media, 2012. 2

[41] Elias Nehme, Omer Yair, and Tomer Michaeli. Uncertainty quantification via neural posterior principal components. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 3, 4, 5

[42] Frank Nussbaum, Jakob Gawlikowski, and Julia Niebling. Structuring uncertainty for fine-grained sampling in stochastic segmentation networks. *Advances in Neural Information Processing Systems*, 35:27678–27691, 2022. 2, 3

[43] Guy Ohayon, Theo Adrai, Gregory Vaksman, Michael Elad, and Peyman Milanfar. High perceptual quality image denoising with a posterior sampling cgan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1805–1813, 2021. 1, 2, 3

[44] Chawin Ounkomol, Sharmishtaa Seshamani, Mary M Maleckar, Forrest Collman, and Gregory R Johnson. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nature methods*, 15(11): 917–920, 2018. 1

[45] Tim Pearce, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International conference on machine learning*, pages 4075–4084. PMLR, 2018. 2

[46] Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*. International Conference on Representation Learning, 2018. 2

[47] Yair Rivenson, Hongda Wang, Zhensong Wei, Kevin de Haan, Yibo Zhang, Yichen Wu, Harun Günaydın, Jonathan E Zuckerman, Thomas Chong, Anthony E Sisk, et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nature biomedical engineering*, 3(6):466–477, 2019. 1

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 7

[49] 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. 1, 3

[50] 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*, 2022. 1, 3

[51] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 3

[52] Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *International conference on machine learning*, pages 1218–1226. PMLR, 2015. 2

[53] Swami Sankaranarayanan, Anastasios Nikolas Angelopoulos, Stephen Bates, Yaniv Romano, and Phillip Isola. Semantic uncertainty intervals for disentangled latent spaces. In *Advances in Neural Information Processing Systems*, 2022. 2

[54] Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11492–11501, 2022. 2

[55] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[56] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 3

[57] 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. 2

[58] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34: 1415–1428, 2021. 2

[59] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2021. 1, 3

[60] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 3

[61] Ilya Sutskever and Tijmen Tieleman. On the convergence properties of contrastive divergence. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 789–795. JMLR Workshop and Conference Proceedings, 2010. 4

[62] Lucas von Chamier, Romain F Laine, Johanna Jukkala, Christoph Spahn, Daniel Krentzel, Elias Nehme, Martina Lerche, Sara Hernández-Pérez, Pieta K Mattila, Eleni Karinou, et al. Democratising deep learning for microscopy with zerocostdl4mic. *Nature communications*, 12(1):2276, 2021. 7

[63] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019. 2

[64] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 3, 7

[65] Omer Yair and Tomer Michaeli. Thinking fourth dimensionally: Treating time as a random variable in ebms. *https://openreview.net/forum?id=m0fEJ2bvwpw*, 2022. 5

[66] Ning Yu, Ke Li, Peng Zhou, Jitendra Malik, Larry Davis, and Mario Fritz. Inclusive gan: Improving data and minority coverage in generative models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 377–393. Springer, 2020. 2

[67] Yang Zhao, Jianwen Xie, and Ping Li. Learning energy-based generative models via coarse-to-fine expanding and sampling. In *International Conference on Learning Representations*, 2020. 4