

# Image Restoration by Deep Projected GSURE

Shady Abu-Hussein  
Tel-Aviv University  
shadya@mail.tau.ac.il

Tom Tirer  
Tel-Aviv University  
tomtirer@mail.tau.ac.il

Se Young Chun  
Ulsan National Institute of  
Science and Technology  
sychun@unist.ac.kr

Yonina C. Eldar  
Weizmann Institute of Science  
yonina.eldar@weizmann.ac.il

Raja Giryes  
Tel-Aviv University  
raja@tauex.tau.ac.il

## Abstract

*Ill-posed inverse problems appear in many image processing applications, such as deblurring and super-resolution. In recent years, solutions that are based on deep Convolutional Neural Networks (CNNs) have shown great promise. Yet, most of these techniques, which train CNNs using external data, are restricted to the observation models that have been used in the training phase. A recent alternative that does not have this drawback relies on learning the target image using internal learning. One such prominent example is the Deep Image Prior (DIP) technique that trains a network directly on the input image with a least-squares loss. In this paper, we propose a new image restoration framework that is based on minimizing a loss function that includes a "projected-version" of the Generalized Stein Unbiased Risk Estimator (GSURE) and parameterization of the latent image by a CNN. We demonstrate two ways to use our framework. In the first one, where no explicit prior is used, we show that the proposed approach outperforms other internal learning methods, such as DIP. In the second one, we show that our GSURE-based loss leads to improved performance when used within a plug-and-play priors scheme.*

## 1. Introduction

Inverse problems appear in many image processing applications, where the reconstruction of an unknown latent image  $\mathbf{x} \in \mathbb{R}^n$  from its given corrupted version  $\mathbf{y} \in \mathbb{R}^m$  is required. In many image-restoration tasks the observed image  $\mathbf{y}$  can be expressed by the following linear model

$$\mathbf{y} = \mathbf{Hx} + \mathbf{e}, \quad (1)$$

where  $\mathbf{H} \in \mathbb{R}^{m \times n}$  is the measurement operator with  $m \leq n$ , and  $\mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$  is an additive white Gaussian noise.

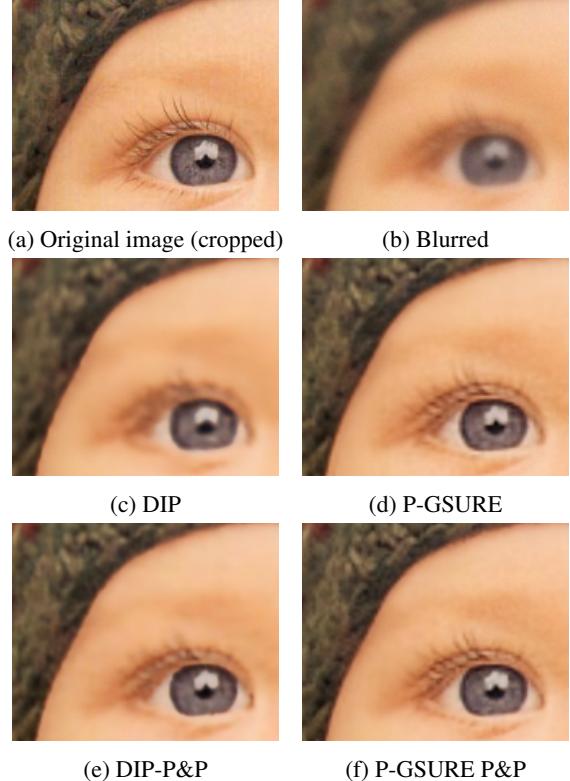


Figure 1: Deblurring of the baby image from Set5, blurred using the scenario 2 model from Table 2. (a) The original image (cropped) (b) The observed blurry and noisy image (c) Reconstruction using DIP (d) Reconstruction using P-GSURE (proposed method) (e) Reconstruction using DIP based P&P-ADMM (with BM3D prior) (f) Reconstruction using P-GSURE based P&P-ADMM (proposed method).

For example, when  $\mathbf{H}$  is a blur operator we refer to restoring

$\mathbf{x}$  from  $\mathbf{y}$  as a deblurring problem, and when  $\mathbf{H}$  is an anti-aliasing filtering followed by sub-sampling we refer to it as super-resolution (SR).

Image restoration problems represented by (1) are usually ill-posed, in the sense that fitting the measurements  $\mathbf{y}$  alone is not enough for a successful reconstruction of  $\mathbf{x}$ . Therefore, the use of some prior assumptions on  $\mathbf{x}$  is inevitable. Accordingly, many reconstruction approaches are based on a minimization of a cost function formulated as

$$\ell(\tilde{\mathbf{x}}; \mathbf{y}) + \beta s(\tilde{\mathbf{x}}), \quad (2)$$

where  $\tilde{\mathbf{x}}$  is the optimization variable,  $\ell(\cdot; \mathbf{y})$  is the fidelity term,  $s(\cdot)$  is the prior / regularization term, and  $\beta > 0$  controls the strength of the regularization. The fidelity term assures compliance to the measurements  $\mathbf{y}$ , while the prior term encourages the reconstruction to satisfy some prior assumptions on  $\mathbf{x}$ , such as sparsity or non-local repeating patterns.

Over the past decade, Convolutional Neural Networks (CNNs) have shown great potential when it comes to solving imaging inverse problems as they circumvent the hardness of designing priors for natural images [20, 9, 38, 28]. However, in order to perform well, the CNNs require that the observation model at test time is the exact model used in the training phase. For instance, in the SR task, it is common to train the CNNs to reconstruct  $\mathbf{x}$  from  $\mathbf{y}$  that is observed using the bicubic down-sampling kernel [9, 20]. But when examined on other down-sampling kernels these networks perform very poorly [33, 41, 1].

An alternative to the offline training of CNNs with a pre-determined observation model (which may not match the model at test time) is “internal learning”, *i.e.* training that is based only on the given observed image [33, 44]. One popular internal learning method is Deep Image Prior (DIP) [44], where a CNN with a set of parameters  $\theta$  (and a fixed random input) is employed to parameterize the latent image  $\tilde{\mathbf{x}} = \mathbf{f}(\theta)$ . It is then used to minimize the Least-Squares (LS) loss (with no use of an explicit prior  $s(\cdot)$ ),

$$\ell_{\text{LS}}(\mathbf{f}(\theta)) = \|\mathbf{y} - \mathbf{H}\mathbf{f}(\theta)\|_2^2.$$

Using such a scheme implicitly benefits from the fact that the CNN architecture promotes non-local repeating patterns, which are common in natural images. However, because DIP uses an over-parameterized CNN it may overfit the noise. Therefore, it requires an accurate early stopping, and its performance degrades dramatically when faced with heavy noise. This inspired follow-up works to incorporate DIP with additional explicit regularizations [6, 21, 22, 50].

In this paper, we propose a new image restoration framework that is based on minimizing a loss function, where instead of a plain fidelity term  $\ell(\tilde{\mathbf{x}})$ , we use a “projected-version” of the Generalized Stein Unbiased Risk Estimator

(GSURE) [37, 10] together with a parameterization of the latent image by a CNN. We demonstrate two ways to use our framework. In the first one, where no explicit prior is used, we show that the proposed approach outperforms other internal learning methods, such as DIP [44] and ZSSR [33]. In the second one, we show that our GSURE-based loss leads to improved performance when used within a plug-and-play priors scheme [45, 30, 5, 40, 22], where an additional prior  $s(\cdot)$  is imposed through the application of existing denoisers.

## 2. Related work

The self-similarity of natural images (*i.e.* the phenomenon of recurrence of patterns within and across scales) has been shown to be a powerful prior for image reconstruction [3, 27, 7, 13, 15]. Recent empirical works, such as DIP [44] and zero-shot super-resolution (ZSSR) [33] showed that the self-similarity prior imposed by CNN architectures is sufficient for recovering an image simply by training a deep neural network from scratch at test time using only the observed degraded image. While ZSSR focused only on SR, the approach in DIP considered general inverse problems and led to many follow-up papers [11, 48, 31, 32, 39, 46, 14].

In the presence of noise, DIP requires an accurate early stopping to avoid overfitting the noise. To mitigate this issue, and further improve the performance, several papers incorporate it with additional explicit priors [6, 21, 22, 50].

Another line of work, which is related to our paper, focuses on offline unsupervised learning: training CNNs on noisy training sets (without ground truth clean images). One such work is Noise2Noise [19], where the training set includes two independent noise realizations per image, and the strategy is examined for various noise models and some applications, including compressive sensing MR recovery [19]. Another method, proposed in [34], is based on Stein’s unbiased risk estimator (SURE) [37] (which is an unbiased estimate for the estimation MSE when the ground truth is unknown). This method trains CNNs with a training set of a single noise realization per image and SURE as the loss function. Yet, similarly to the original SURE [37], it is limited to Gaussian denoising. Another paper that uses SURE to train CNNs for Gaussian denoising is [24], where the performance is also demonstrated when training using only a single noisy image. These works have been extended to train networks (without ground truth data) for image recovery from undersampled compressive sensing measurements but the training is performed in this case using multiple images.

Over the years, SURE has been extended to more general noise models such as mixed Poisson-Gaussian model [18] or a general exponential family [10]. The Generalized SURE (GSURE) in [10] has been used for offline unsu-

pervised training (without ground truth data) of networks for image recovery from undersampled compressive sensing measurements [24]. Another work, combined the regular SURE with the LDAMP technique [26] for solving the same problem using also offline training [49]. In contrast to these two approaches that perform offline training on a large set of images with pre-determined observation models, in this paper we will use GSURE to solve ill-posed problems by training a CNN using only the observed image.

Using SURE to train a CNN requires evaluating the divergence of the network’s output with respect to its input, as well as the gradient of the divergence with respect to the network’s parameters. Most previous works [34, 49, 24] have used Monte-Carlo (MC) approximation of the divergence, as proposed in [29], due to fast and easy computation, which we employ also in our paper. Alternative approximation of the divergence, based on random projections of the network’s output, has been shown in [35].

The image restoration framework, presented in this paper, is demonstrated with no explicit prior (similarly to DIP), and also when an additional prior is used. In the latter case, we are inspired by the Plug-and-Play Priors (P&P) approach [45, 25, 30, 5, 16, 40], where the prior used for solving different inverse problems is imposed through the application of denoising operations. Specifically, in this paper, we impose the additional prior using the BM3D denoiser [7]. Note that several P&P methods have also shown excellent results using offline trained CNN denoisers [23, 40, 47, 40]. We note that there exists a method [22] that combines DIP and P&P. The major difference between our approach and [22] is in the fact that [22] uses the plain LS fidelity term while we use a GSURE-based loss function. GSURE has the advantage over DIP that it leads to a more stable training that is less sensitive than DIP to the parameters of the network and the early stopping in the training.

### 3. Method

In this paper, we employ the generalized SURE (GSURE), which has been developed in [10]. Let  $\hat{\mathbf{x}}(\mathbf{y})$  be a (weakly) differentiable estimator of the deterministic unknown  $\mathbf{x}$ . The original SURE formula [37] provides an unbiased estimate of the MSE,  $\mathbb{E}\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2$ , for the case where  $\mathbf{e}$  is white Gaussian noise and  $\mathbf{H}$  equals the identity matrix  $\mathbf{I}_n$ . The work in [10] has proposed the project GSURE, which generalizes the MSE estimate to more cases, where one of them is the ill-posed linear model that is considered in Equation (1).

While previous works used this “projected” GSURE mainly for tuning only a few hyper-parameters of a pre-determined estimator  $\hat{\mathbf{x}}$  (*i.e.* by setting them to values that minimize Equation (4)) [12], in this paper, we proposed to extend the role that GSURE plays in estimating  $\mathbf{x}$ . Specifi-

cally, we propose to estimate  $\mathbf{x}$  (from scratch) as the minimizer of the cost function

$$\ell_{\text{GSURE}}(\tilde{\mathbf{x}}) + \beta s(\tilde{\mathbf{x}}), \quad (3)$$

where  $\tilde{\mathbf{x}}$  is the optimization variable,  $s(\tilde{\mathbf{x}})$  is a prior term that can improve the estimation and  $\beta$  is a positive hyper-parameter that balances the two terms.

Before introducing our method, we first discuss the use of the projected GSURE and its formulation. Then we show how it can be used for network training from a single example. This is followed by a derivation of a plug and play based approach allows incorporating prior knowledge through a denoiser to the single image training. Finally, we draw a connection between our formulation and other recently used loss functions.

#### 3.1. The projected GSURE

In the case where  $m \leq n$  and  $\mathbf{H}$  has a full row rank, its pseudoinverse is given by  $\mathbf{H}^\dagger := \mathbf{H}^T(\mathbf{H}\mathbf{H}^T)^{-1}$  and the projection onto  $\mathcal{R}(\mathbf{H}^T)$  (*i.e.* the subspace spanned by the rows of  $\mathbf{H}$ ) is given by  $\mathbf{P}_H := \mathbf{H}^\dagger\mathbf{H}$ . As the observation model is ill-posed, the GSURE provides an estimate of the MSE only in  $\mathcal{R}(\mathbf{H}^T)$ , *i.e.*, of the projected MSE  $\mathbb{E}\|\mathbf{P}_H\hat{\mathbf{x}} - \mathbf{P}_H\mathbf{x}\|_2^2$ . The GSURE estimate is given by

$$\begin{aligned} \ell_{\text{GSURE}}(\hat{\mathbf{x}}(\mathbf{u})) = & c + \|\mathbf{P}_H\hat{\mathbf{x}}(\mathbf{u})\|_2^2 - 2\hat{\mathbf{x}}^T(\mathbf{u})\hat{\mathbf{x}}_{\text{ML}} \\ & + 2\text{div}_{\mathbf{u}}(\mathbf{P}_H\hat{\mathbf{x}}(\mathbf{u})), \end{aligned} \quad (4)$$

where  $c$  is a term that does not depend on  $\hat{\mathbf{x}}$ ,  $\mathbf{u} \in \mathbb{R}^n$  is a sufficient statistic (for the estimation of  $\mathbf{x}$  from  $\mathbf{y}$ ) that  $\hat{\mathbf{x}}$  gets as input,  $\hat{\mathbf{x}}_{\text{ML}}$  is the maximum likelihood estimator and the last term denotes the divergence

$$\text{div}_{\mathbf{u}}(\mathbf{P}_H\hat{\mathbf{x}}(\mathbf{u})) = \sum_{i=1}^n \frac{\partial[\mathbf{P}_H\hat{\mathbf{x}}(\mathbf{u})]_i}{\partial u_i}. \quad (5)$$

For the considered model, typical choices are  $\mathbf{u} = \frac{1}{\sigma^2}\mathbf{H}^T\mathbf{y}$ <sup>1</sup> and  $\hat{\mathbf{x}}_{\text{ML}} = \mathbf{H}^\dagger\mathbf{y}$ .

In practice, the assumption that  $\mathbf{H}$  has a full row-rank does not hold always. Moreover, some of the singular values of  $\mathbf{H}$  may be very close to zero such that even if  $\mathbf{H}$  has a full row-rank, numerically the inversion is unstable and therefore leads to poor MSE estimations. This is particularly important in the deblurring problem where many of the blur operators have a full rank in theory but in practice, many of their singular values are close to zero and therefore should be treated as such. In this case, to calculate the pseudo-inverse of  $\mathbf{H}$ , we take the SVD decomposition  $\mathbf{H} = \mathbf{U}\Sigma\mathbf{V}^T$ , where  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_m, 0, \dots, 0)$ , and set the pseudo inverse to be  $\mathbf{H}^\dagger = \mathbf{V}\Sigma^\dagger\mathbf{U}^T$ . The pseudo inverse  $\Sigma^\dagger$ , which is a diagonal matrix, is defined such that

<sup>1</sup>Notice that although the GSURE relies on knowing the noise level  $\sigma$ ,  $\sigma$  can be estimated from the  $\mathbf{y}$  using techniques such as [4].

$(\Sigma^\dagger)_{i,i} = 1/\sigma_i$  if  $i \leq m$  and  $\sigma_i > \xi$  (for a given small threshold  $\xi$ ) and  $(\Sigma^\dagger)_{i,i} = 0$  otherwise.

This formulation for the Pseudo-inverse can be easily computed in the deblurring case (with circular padding) since the blur kernels are shift-invariant and therefore they are diagonalized by the Fourier transform. Thus, to apply the Pseudo-inverse of  $\mathbf{H}$  we can simply apply the Fourier transform on the input, then multiply it with the inverse eigenvalues of the kernel (that are larger than the threshold  $\xi$ ) and then apply the inverse Fourier transform.

Notice that the formula of the projected GSURE remains the same also when we use the thresholding of the small singular values. The only difference is in the calculation of  $\mathbf{H}^\dagger$ , which then sets also  $\mathbf{P}_H$  in Equation (4).

Another element in Equation (4) that deserves a special care is the divergence term. It is essentially the only term in Equation (4) that requires connecting  $\tilde{\mathbf{x}}$  with an input  $\mathbf{u}$ . In this paper we resolve this issue by parameterizing  $\tilde{\mathbf{x}}$  by a CNN with weights  $\theta$  and input  $\mathbf{u} = \frac{1}{\sigma^2} \mathbf{H}^T \mathbf{y}$ , i.e.  $\tilde{\mathbf{x}} = \mathbf{f}(\mathbf{u}; \theta)$ . While this differentiable connection between  $\tilde{\mathbf{x}}$  and  $\mathbf{u}$  allows to compute the derivatives required in Equation (5), to ease the computation we use the MC approximation of [29]:

$$\text{div}_{\mathbf{u}}(\mathbf{P}_H \tilde{\mathbf{x}}(\mathbf{u})) \approx \mathbf{g}^T \mathbf{P}_H \frac{\tilde{\mathbf{x}}(\mathbf{u} + \epsilon \mathbf{g}) - \tilde{\mathbf{x}}(\mathbf{u})}{\epsilon}, \quad (6)$$

where  $\mathbf{g}$  is a Gaussian vector drawn from  $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$  and  $\epsilon$  is a fixed small positive value.

To summarize, our framework for estimating  $\mathbf{x}$  is based on optimizing the weights of the CNN  $\mathbf{f}(\mathbf{u}; \theta)$  by minimizing the loss function

$$L(\theta) = \tilde{\ell}_{\text{GSURE}}(\mathbf{f}(\mathbf{u}; \theta)) + \beta s(\mathbf{f}(\mathbf{u}; \theta)), \quad (7)$$

where

$$\begin{aligned} \tilde{\ell}_{\text{GSURE}}(\mathbf{f}(\mathbf{u}; \theta)) = & \|\mathbf{P}_H \mathbf{f}(\mathbf{u}; \theta)\|_2^2 - 2\mathbf{f}^T(\mathbf{u}; \theta) \mathbf{H}^\dagger \mathbf{y} \\ & + 2\mathbf{g}^T \mathbf{P}_H \frac{\mathbf{f}(\mathbf{u} + \epsilon \mathbf{g}; \theta) - \mathbf{f}(\mathbf{u}; \theta)}{\epsilon}. \end{aligned} \quad (8)$$

As mentioned above, in tasks like deblurring and super-resolution, the operators  $\mathbf{H}$ ,  $\mathbf{H}^\dagger$  and  $\mathbf{P}_H$  have a fast implementation using the Fast Fourier Transform (FFT). In addition, fast automatic differentiation of  $\tilde{\ell}_{\text{GSURE}}(\mathbf{f}(\mathbf{u}; \theta))$  with respect to  $\theta$  can be obtained using popular software packages, such as TensorFlow and PyTorch. Note that the random variable  $\mathbf{g}$  is drawn in each iterations of the optimization (this improves the expectation of the MC approximation). After obtaining the optimized weights,  $\hat{\theta}$ , the latent image is estimated by  $\hat{\mathbf{x}} = \mathbf{f}(\mathbf{u}; \hat{\theta})$ .

We now turn to describe two instantiations of Equation (7): one without the prior function  $s(\cdot)$  and the second when the additional prior is implicitly imposed by a “plug-and-play” denoiser.

### 3.2. GSURE-based network training

Standard neural network training techniques for inverse problems in image processing require a large amount of training data as they are usually trained based on pairs of the input deteriorated image  $\mathbf{y}$  and the true clean image  $\mathbf{x}$ . Under the assumption that the image statistics in the training resemble the one of the input image to the network, then the fact that the network was trained to minimize the MSE on the training images is expected to lead to a small MSE also on the input test image. The GSURE gives us an opportunity to treat the MSE of the input image directly as it provides an unbiased estimate for it. This is particularly important in cases where the image statistics or the measurement model are different than the training setup.

Having the estimation of the GSURE for the MSE (or projected MSE), we use it to train a network just a single image by directly minimizing its (estimated projected) MSE. Notice that although we do not use any explicit image prior in this setting, the fact that we use a neural network for the reconstruction implicitly imposes a prior as shown in the DIP approach [44]. Moreover, the implicit regularization of optimization methods like SGD and Adam is another source of regularization [36]. Therefore, the first reconstruction method that we examine in this paper is solely based on minimizing  $L(\theta) = \tilde{\ell}_{\text{GSURE}}(\mathbf{f}(\mathbf{u}; \theta))$  using Adam [17].

While the GSURE based approach bears similarity to the original DIP work, it differs from it in several ways. First, it is theoretically motivated by minimizing an estimate of the (projected) MSE. Second, it uses a different loss function —  $\tilde{\ell}_{\text{GSURE}}(\mathbf{f}(\mathbf{u}; \theta))$  instead of  $\ell_{\text{LS}}(\mathbf{f}(\mathbf{u}; \theta))$ . Third, it uses a different input for the CNN — the sufficient statistic  $\mathbf{u}$  rather than a large tensor of random noise. As will be shown in the experiments section, GSURE based training not only outperforms the original DIP in terms of PSNR but also does not require accurate early stopping to avoid noise overfitting. Another related work is a recent paper [50] that solves deblurring tasks using DIP with the BP loss (instead of the plain LS loss as in regular DIP), which is equivalent to omitting the divergence term in GSURE (this also allows their CNN to get random noise as input). Yet, the divergence term that we have in the GSURE stabilizes the effect of the noise on the BP loss, which as shown in [50] is sensitive to noise and requires accurate early stopping. We discuss more the relationship to this loss in Section 3.4.

### 3.3. GSURE-based P&P-ADMM

In an ideal case, where GSURE would provide a perfect estimation for the MSE, it would have been sufficient to train a neural network using GSURE to get “perfect” results. Yet, it is only an estimate of the MSE and with the projection it estimates only of the projected MSE. Thus, using it alone for minimization is not sufficient and adding another image prior  $s(\cdot)$  can lead to improved reconstruc-

tion.

To achieve this goal, we follow the P&P denoisers concept [45]. We modify Equation (7) using variable splitting and obtain the following constrained optimization problem

$$\begin{aligned} \min_{\theta, \mathbf{z}} \quad & \tilde{\ell}_{\text{GSURE}}(\mathbf{f}(\mathbf{u}; \theta)) + \beta s(\mathbf{z}) \\ \text{s.t. } \quad & \mathbf{z} = \mathbf{f}(\mathbf{u}; \theta). \end{aligned} \quad (9)$$

This problem can be solved using ADMM (a review on ADMM can be found in [2]). Specifically, we construct the augmented Lagrangian

$$\begin{aligned} L_\rho = & \tilde{\ell}_{\text{GSURE}}(\mathbf{f}(\mathbf{u}; \theta)) + \beta s(\mathbf{z}) + \tilde{\mathbf{v}}^T(\mathbf{f}(\mathbf{u}; \theta) - \mathbf{z}) \\ & + \frac{\rho}{2} \|\mathbf{f}(\mathbf{u}; \theta) - \mathbf{z}\|_2^2 \\ = & \tilde{\ell}_{\text{GSURE}}(\mathbf{f}(\mathbf{u}; \theta)) + \beta s(\mathbf{z}) + \frac{\rho}{2} \|\mathbf{f}(\mathbf{u}; \theta) - \mathbf{z} + \mathbf{v}\|_2^2 \\ & - \frac{\rho}{2} \|\mathbf{v}\|_2^2, \end{aligned} \quad (10)$$

where  $\tilde{\mathbf{v}}$  is the dual variable,  $\mathbf{v} = \frac{1}{\rho} \tilde{\mathbf{v}}$  is the scaled dual variable, and  $\rho$  is the ADMM penalty parameter. Then, we initialize  $\mathbf{v}_0 = \mathbf{0}$  and  $\mathbf{z}_0 = \hat{\mathbf{x}}_{\text{ML}}$ , and repeat the following three steps for  $k = 1, \dots, N_{\text{iter}}$

$$\begin{aligned} \theta_k = & \operatorname{argmin}_{\theta} \tilde{\ell}_{\text{GSURE}}(\mathbf{f}(\mathbf{u}; \theta)) \\ & + \frac{\rho}{2} \|\mathbf{f}(\mathbf{u}; \theta) - \mathbf{z}_{k-1} + \mathbf{v}_{k-1}\|_2^2, \\ \mathbf{z}_k = & \operatorname{argmin}_{\mathbf{z}} \frac{1}{2\beta/\rho} \|\mathbf{f}(\mathbf{u}; \theta_k) + \mathbf{v}_{k-1} - \mathbf{z}\|_2^2 + s(\mathbf{z}), \\ \mathbf{v}_k = & \mathbf{v}_{k-1} + \mathbf{f}(\mathbf{u}; \theta_k) - \mathbf{z}_k. \end{aligned} \quad (11)$$

The first step can be optimized via Adam with warm-start initialization (given by the previous  $\theta_{k-1}$ ). The second step describes obtaining  $\mathbf{z}_k$  using a denoiser for white Gaussian noise of variance  $\beta/\rho$  applied on  $\mathbf{f}(\mathbf{u}; \theta_k) + \mathbf{v}_{k-1}$ . Following the P&P denoisers concept, in our experiments we implement this step by an existing denoiser  $\mathbf{z}_k = \mathcal{D}(\mathbf{f}(\mathbf{u}; \theta_k) + \mathbf{v}_{k-1}; \sqrt{\beta/\rho})$  (e.g., colored BM3D [7]) that implicitly imposes the prior  $s(\cdot)$  on the reconstruction.

Note that in typical P&P schemes, which use plain fidelity terms, the first step in Equation (11) inverts the observation model with very minor regularization (the simple  $\ell_2$ -norm term that is weighted by  $\rho/2$ ). In contrast, in our approach, this step enjoys both the advantage of GSURE over LS (*e.g.* better handling the noise via the divergence term) and the implicit prior of the CNN architecture.

### 3.4. Relation to other loss functions

Let us present a motivation for optimizing Equation (3) (or its CNN parameterized form Equation (7)) by comparing it with alternative cost functions. Perhaps the most common approach to estimate  $\mathbf{x}$  from observations modeled

by Equation (1) is to minimize cost function of the form  $\ell_{\text{LS}}(\tilde{\mathbf{x}}) + \beta s(\tilde{\mathbf{x}})$ , where  $\ell_{\text{LS}}(\tilde{\mathbf{x}}) = \|\mathbf{y} - \mathbf{H}\tilde{\mathbf{x}}\|_2^2$  is the least squares (LS) fidelity term [22, 45, 47].

Recently, an alternative fidelity term  $\ell_{\text{BP}}(\tilde{\mathbf{x}}) = \|\mathbf{H}^\dagger(\mathbf{y} - \mathbf{H}\tilde{\mathbf{x}})\|_2^2$ , dubbed “back-projection” (BP) term, has been shown to have advantages over LS for different ill-posed problems, such as deblurring and super-resolution, where the condition number of  $\mathbf{H}$ , *i.e.* the ratio between the largest and smallest singular values of  $\mathbf{H}$ , is large [40, 42, 43]. However, the tradeoff is that the BP term is more sensitive than LS to noise amplification when  $\mathbf{H}$  has small singular values.

Now, observe that if we ignore the divergence term in Equation (4) then minimizing  $\ell_{\text{GSURE}}(\tilde{\mathbf{x}})$  (where  $\hat{\mathbf{x}}_{\text{ML}} = \mathbf{H}^\dagger \mathbf{y}$ ) is equivalent to minimizing  $\ell_{\text{BP}}(\tilde{\mathbf{x}})$ . This follows from (recall that  $\mathbf{P}_H = \mathbf{P}_H^T$  and  $\mathbf{P}_H \mathbf{H}^\dagger = \mathbf{H}^\dagger$ ) the fact that

$$\begin{aligned} & \operatorname{argmin}_{\tilde{\mathbf{x}}} \|\mathbf{H}^\dagger(\mathbf{y} - \mathbf{H}\tilde{\mathbf{x}})\|_2^2 \\ = & \operatorname{argmin}_{\tilde{\mathbf{x}}} \|\mathbf{P}_H \tilde{\mathbf{x}}\|_2^2 - 2\tilde{\mathbf{x}}^T \mathbf{P}_H \mathbf{H}^\dagger \mathbf{y} + \|\mathbf{H}^\dagger \mathbf{y}\|_2^2 \\ = & \operatorname{argmin}_{\tilde{\mathbf{x}}} \|\mathbf{P}_H \tilde{\mathbf{x}}\|_2^2 - 2\tilde{\mathbf{x}}^T \mathbf{H}^\dagger \mathbf{y}. \end{aligned} \quad (12)$$

Therefore,  $\ell_{\text{GSURE}}(\tilde{\mathbf{x}})$  is expected to have advantages over  $\ell_{\text{LS}}(\tilde{\mathbf{x}})$  that are similar to those of  $\ell_{\text{BP}}(\tilde{\mathbf{x}})$ . Moreover, note that the term  $\operatorname{div}_{\mathbf{u}}(\mathbf{P}_H \tilde{\mathbf{x}}(\mathbf{u}))$  in  $\ell_{\text{GSURE}}(\tilde{\mathbf{x}})$  constrains the changes in  $\tilde{\mathbf{x}}$  due to perturbations in the observations, which provides a regularization against noise amplification. Therefore,  $\ell_{\text{GSURE}}(\tilde{\mathbf{x}})$  is also expected to be more suitable than  $\ell_{\text{BP}}(\tilde{\mathbf{x}})$  to handle high noise levels.

## 4. Experiments

We demonstrate our method on two common image processing tasks on the Set5 and Set14 datasets. We first present the experiments for deblurring, in which a restoration of the original sharp image from the blurry and noisy observation is desired. This is followed by evaluating our approach on the SR task, where we seek to recover the original High-Resolution (HR) image from its given Low-Resolution (LR) and noisy version.

In each task, we examine the proposed GSURE-based method with and without an additional BM3D prior. In all the experiments, we compare our results with those obtained by using DIP instead of GSURE (recall that they mainly differ in their loss functions and in their inputs). For both GSURE and DIP, we use the same ‘skip’ network used in the original DIP code published by [44]. In the “internal learning” SR experiments (where only the self-similarity of the observed image is used as prior), we also compare to the results of ZSSR [33]. In the experiments that exploit the BM3D prior, we also present the results of the plain P&P method (that does not include the CNN parameterization) [45].

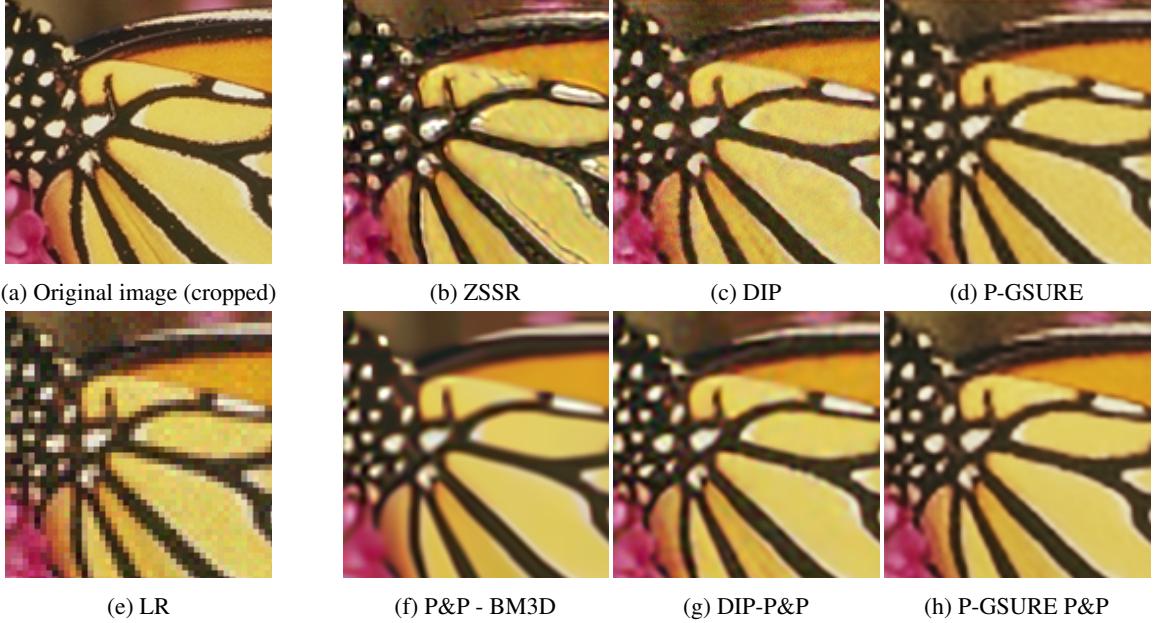


Figure 2: Super-resolution using the scenario 6 model from Table 4, tested on the butterfly image from Set5, (a) The original image cropped (b) The observed low-resolution and noisy image (c) Reconstruction using ZSSR (d) Reconstruction using DIP (e) Reconstruction using P-GSURE (proposed method) (f) P&P-ADMM method (using BM3D prior) (g) Reconstruction using DIP based P&P-ADMM (with BM3D prior) (h) Reconstruction using P-GSURE based P&P-ADMM (proposed method).

We use  $\epsilon = 10^{-6}$  for approximating the divergence in (6) for all GSURE scenarios and optimize the network weights using ADAM [17] with learning rate of  $10^{-2}$  for both DIP and GSURE. For achieving the best performance in DIP, we run the algorithm in all scenarios on all the images in Set5, then we average the iterations number in which the maximum PSNR is achieved, and use that as the early stopping criterion in each scenario (6 values, one for each scenario). In GSURE, however, we take the reconstruction at the minimum value of  $\ell_{GSURE}(\hat{\mathbf{x}})$ , since it is on par with the MSE value. Notice that unlike in DIP, with GSURE we do not need the ground truth information to set the stopping criterion, as demonstrated in Figures 3 and 4. In P&P based GSURE and DIP we take the reconstruction from the last ADMM iteration.

#### 4.1. Deblurring

In the image deblurring problem,  $\mathbf{H}$  is a blur operator defined as a circular convolution with a blur filter  $h$ . Thus, it has fast implementation using the Fast Fourier Transform (FFT). Accordingly,  $\mathbf{H}^T$  is a blur operator with a flipped version of the blur kernel  $h$ . As a result, calculating  $\mathbf{H}\mathbf{H}^T$  can be achieved by simply multiplying the FFTs of the blur kernel and its flipped version.

We adopt the test scenarios used in [8], whose blur kernels  $h$  and noise levels  $\sigma$  are presented in Table 2. Note

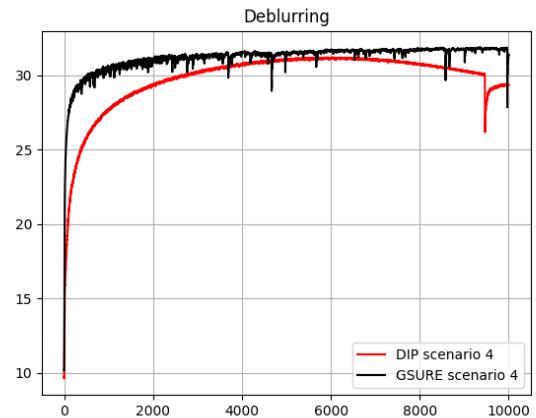


Figure 3: Deblurring PSNR results of scenario 4 averaged on Set5.

that in these scenarios the condition number of  $\mathbf{H}$  is extremely high, hence, when computing  $(\mathbf{H}\mathbf{H}^T)^\dagger$ , we zero (rather than invert) all the places where the eigenvalues are smaller than  $\xi$  (the specific values are given in Table 2).

Under this setup, we run GSURE for 4000 iterations and take the reconstruction that minimizes  $\ell_{GSURE}(\hat{\mathbf{x}})$ . Practically, GSURE does not need any early stopping criterion,

Table 1: Deblurring results (averaged PSNR) for the different scenarios

Scenario	Set5		Set14		Set5			Set14		
	DIP	P-GSURE	DIP	P-GSURE	P&P	P&P-DIP	P&P-GSURE	P&P	P&P-DIP	P&P-GSURE
1	30.36	<b>31.95</b>	27.10	<b>30.14</b>	<b>33.36</b>	32.18	33.30	30.72	27.49	<b>30.97</b>
2	30.24	<b>31.39</b>	27.01	<b>28.28</b>	31.25	31.5	<b>31.82</b>	28.27	27.46	<b>28.92</b>
3	28.42	<b>30.28</b>	25.17	<b>27.75</b>	<b>32.60</b>	29.68	30.38	<b>30.31</b>	25.85	27.58
4	31.16	<b>31.99</b>	26.77	<b>29.49</b>	31.85	31.23	<b>32.53</b>	29.29	28.85	<b>29.7</b>
5	30.41	<b>30.93</b>	26.91	<b>28.06</b>	31.80	31.64	<b>32.05</b>	28.66	27.55	<b>28.71</b>
6	33.44	<b>35.55</b>	31.32	<b>34.21</b>	34.55	33.32	<b>35.67</b>	33.69	29.74	<b>34.45</b>

Table 2: Deblurring test scenarios

Scenario	$h(x_1, x_2)$	$\sigma^2$	$\xi$
1	$1/(1 + x_1^2 + x_2^2)$	2	$5 \times 10^{-2}$
2	$1/(1 + x_1^2 + x_2^2)$	8	$1 \times 10^{-1}$
3	9 × 9 uniform	0.3	$5 \times 10^{-3}$
4	$[1, 4, 6, 4, 1]^T [1, 4, 6, 4, 1]/256$	49	$1 \times 10^{-1}$
5	Gaussian with $std = 1.6$	4	$5 \times 10^{-2}$
6	Gaussian with $std = 0.4$	64	0

since the divergence term in Equation (4) prevents the network from fitting the noise even when  $\sigma$  is high. Therefore, the objective in Equation (7) can be conveniently minimized until convergence, in contrast to DIP, where careful early stopping criteria need to be examined in order to achieve optimal results and prevent it from fitting the noise (as can be seen in Figure 3).

The average PSNR results of our method compared to DIP are presented in the left columns of Table 1. The superiority of our method over DIP is clear.

One may appreciate the visual quality of the GSURE based reconstruction by looking at Figure 1 and Figure 5. The first is from scenario 2 and the second from scenario 5 in Table 2.

We also examine the different methods within the ADMM-based P&P scheme (presented with GSURE loss in Equation (11)). In P&P-GSURE, we set  $\beta_i/\rho = 0.01$ , where  $\beta_i = \{0.75, 0.75, 4, 1, 2, 1.5\}$  and  $i \in \{1, 2, 3, 4, 5, 6\}$  indicates the scenario number. We initialize  $\mathbf{v} = 0$  and  $\mathbf{z} = \hat{\mathbf{x}}_{ML}$ , and run 250 ADMM iterations, while performing the first step in (11) with 20 iterations of Adam with learning rate of  $10^{-3}$ . For P&P-DIP we set  $\beta/\rho = 1$ , where  $\beta = 0.1$ , initialize  $\mathbf{v} = 0$  and  $\mathbf{z} = \hat{\mathbf{x}}_{ML}$ , and run 250 ADMM iterations, where in each iteration the first step in Equation (11) is implemented using 20 iterations of Adam with learning rate set to  $10^{-2}$ . For the plain P&P (without GSURE or DIP) we use  $\beta_i = \{0.85, 0.85, 0.9, 0.8, 0.85, 0.85\}$  and  $\rho_i = \{2, 1, 3, 1, 2, 2\}/255$ . In all cases we implement the second step in Equation (11) using the python version of the BM3D denoiser [7] with noise level  $\sqrt{\beta/\rho}$ .

Because the projected GSURE measure is an estimate of

the true projected MSE (up to a constant value), one could find the optimal hyper-parameters  $\beta$  and  $\rho$  that gives the lowest projected GSURE value. However, in the case of DIP, the true ground-truth image is required for tuning these parameters, which can be complicated.

The average PSNR results of combining the P&P approach with GSURE and DIP are presented in the right columns of Table 1. It can be seen that incorporating GSURE and DIP with BM3D prior via the P&P approach usually yields better results than their “internal learning” application. Interestingly, P&P-DIP is often no better than the plain P&P scheme (which uses only the BM3D prior). Yet, the proposed P&P-GSURE obtains the best results in most cases.

Finally, several visual examples of all the methods are displayed in Figure 5. It can be seen that our GSURE-based restoration yields sharper images with better textures.

## 4.2. Super-Resolution

In super-resolution  $\mathbf{H}$  is defined as applying an anti-aliasing filter followed by sub-sampling by factor  $\alpha > 1$ .  $\mathbf{H}$  is a circular convolution with filter  $h$ , therefore it can be applied efficiently using Fast Fourier Transform (FFT), then by sub-sampling the inverse FFT of the result we obtain the LR image. Accordingly,  $\mathbf{H}^T$  in this case is defined as zero-padding of the input in between its samples, followed by applying a flipped version of  $h$ , which also can be applied using the FFT. Using the poly-phase decomposition (see, e.g. [5])  $\mathbf{H}\mathbf{H}^T$  can be achieved by multiplying the FFT of  $h$  and its flipped version followed by sub-sampling the inverse FFT.

We test our approach on the test scenarios detailed in Table 4. In these scenarios  $\mathbf{H}$  is ill-posed, therefore when calculating  $(\mathbf{H}\mathbf{H}^T)^\dagger$  we also zero all the eigenvalues smaller than  $\xi$  (as in the deblurring case).

The average PSNR results of our method compared to DIP and ZSSR are presented in the right columns of Table 3, and demonstrate the superiority of our method.

Similar to the deblurring task, we examine the performance of DIP and GSURE within the ADMM-based P&P scheme. For P&P-GSURE, we set  $\beta = 100$  and  $\beta/\rho = 10$ , initialize  $\mathbf{v} = 0$  and  $\mathbf{z} = \hat{\mathbf{x}}_{ML}$ , and run 50 ADMM it-

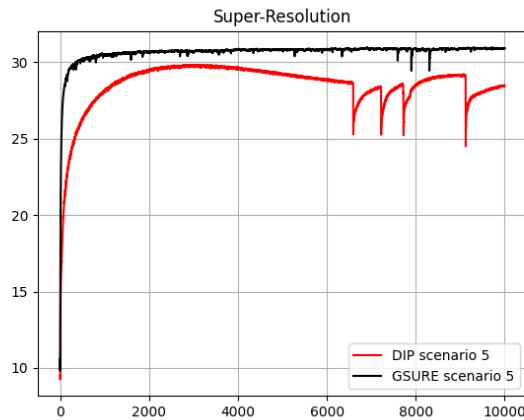


Figure 4: Super-Resolution results of scenario 5 averaged on Set5.

erations, while performing the first step in (11) with 100 iterations of Adam with learning rate of  $10^{-3}$ . In P&P-DIP, we set  $\beta/\rho = 1$ , where  $\beta = 0.1$  for scenarios  $\{1,3,4\}$  and  $\beta = 1.5$  for scenarios  $\{2,5,6\}$ . We run 20 ADMM iterations, where in each iteration the first step in Equation (11) is implemented using 250 iterations of Adam with learning rate set to  $10^{-3}$ . For the plain P&P we use  $\beta_i = \{0.85, 0.9, 0.85, 0.85, 0.9, 0.9\} \times 2$  and  $\rho_i = \{2, 3, 2, 2, 3, 3\}/255$ . Similar to deblurring, we implement the second step in Equation (11) using the python version of the BM3D denoiser [7] with noise level  $\sqrt{\beta/\rho}$ .

Similar to deblurring, note that the hyper-parameters  $\beta$  and  $\rho$  can be set by finding the minimum value of the projected GSURE, in contrast to the least-squares objective where one needs to tune them carefully in order to achieve optimal performance.

The average PSNR results of combining the P&P approach with GSURE and DIP are presented in the left columns of Table 3. Again, incorporating GSURE and DIP with BM3D prior via the P&P approach usually yields better results than their “internal learning” applications. Notice how again the proposed P&P-GSURE obtains the best results in most scenarios.

Finally, several visual examples of all the methods are displayed in Figure 2. It can be seen that our GSURE-based restoration yields sharper images with better textures.

## 5. Conclusion

In this work, we have presented a general framework for solving inverse problems using a neural network on a single image without the need for pre-training on a large dataset. Our approach relies on an unbiased estimator of the MSE, namely the generalized SURE that approximates the true

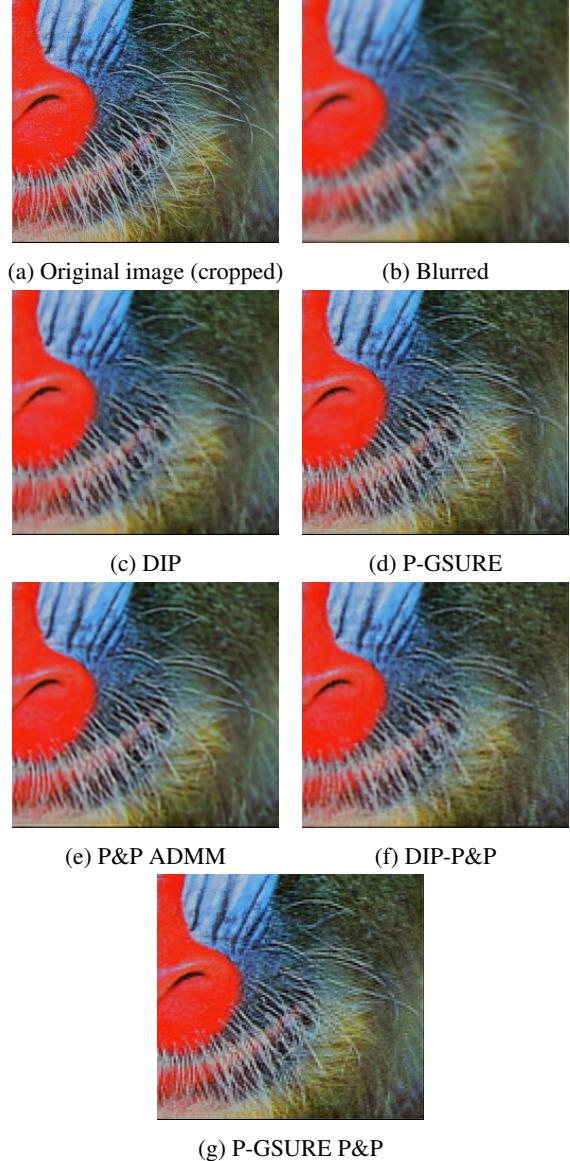


Figure 5: Deblurring of the baboon image from Set14, blurred using the model of scenario 5 from Table 2. (a) The original image cropped (b) The observed blurry and noisy image (c) Reconstruction using DIP (d) Reconstruction using P-GSURE (proposed method) (e) Reconstruction using P&P-ADMM method (using BM3D prior) (f) Reconstruction using DIP based P&P-ADMM (g) Reconstruction using P-GSURE based P&P-ADMM (proposed method).

error. This allows training a network directly on the input image without external data.

Our approach improves over the DIP strategy, which train networks based on the least-squares loss. Moreover, it has the advantage that it does not need to rely on early stopping and less sensitive to the used architecture. In the

Table 3: SR results (averaged PSNR) of the methods for the different scenarios

Scenario	Set5			Set14			Set5			Set14		
	DIP	ZSSR	GSURE	DIP	ZSSR	GSURE	P&P	P&P-DIP	P&P-GSURE	P&P	P&P-DIP	P&P-GSURE
1	28.78	26.77	<b>29.41</b>	26.10	25.03	<b>26.34</b>	28.71	29.26	<b>29.59</b>	26.12	26.67	<b>29.77</b>
2	26.61	25.34	<b>28.09</b>	24.79	24.13	<b>25.68</b>	27.28	27.70	<b>28.41</b>	25.2	25.59	<b>25.93</b>
3	31.49	<b>32.14</b>	32.04	28.38	26.07	<b>29.05</b>	<b>32.83</b>	31.92	32.79	29.57	29.14	<b>29.58</b>
4	29.49	27.33	<b>29.48</b>	26.14	25.23	<b>26.65</b>	29.82	29.95	<b>29.92</b>	26.90	<b>27.19</b>	27.00
5	29.67	29.71	<b>31.09</b>	27.08	25.21	<b>28.32</b>	30.43	31.15	<b>31.46</b>	27.90	28.47	<b>28.77</b>
6	27.60	25.97	<b>28.79</b>	25.41	24.45	<b>26.20</b>	27.92	28.51	<b>28.94</b>	25.63	26.30	<b>26.46</b>

Table 4: SR test scenarios

Scenario	$h(x_1, x_2)$	$\alpha$	$\sigma^2$	$\xi$
1	Gaussian with $std = 1.6$	3	$10$	$1 \times 10^{-2}$
2	Gaussian with $std = 1.6$	3	49	$1 \times 10^{-2}$
3	Bicubic	2	10	$1 \times 10^{-2}$
4	Bicubic	3	10	$1 \times 10^{-2}$
5	Bicubic	2	49	$1 \times 10^{-2}$
6	Bicubic	3	49	$1 \times 10^{-2}$

work, for the sake of comparison, we have used the same architecture proposed in the DIP work. Yet, when we changed the network model, our approach still achieved good results while the performance of DIP decreased more significantly.

While in this work we focus on using the GSURE formulation for white Gaussian noise, it might be possible in follow-up work to support a larger variety of noise types using the formulation in [10] that supports also colored noise and other types of noise. Moreover, while our work focused on deblurring and super-resolution, our proposed framework can be easily used also for other inverse problems in image processing such as joint demosaicing and denoising.

## Acknowledgments

The work is supported by the ERC-StG (No. 757497) grants.

## References

- [1] Shady Abu Hussein, Tom Tirer, and Raja Giryes. Correction filter for single image super-resolution: Robustifying off-the-shelf deep super-resolvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. [2](#)
- [2] Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011. [5](#)
- [3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005. [2](#)
- [4] Ce Liu, W. T. Freeman, R. Szeliski, and Sing Bing Kang. Noise estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 901–908, 2006. [3](#)
- [5] Stanley H. Chan, Xiran Wang, and Omar A. Elgendy. Plug-and-play ADMM for image restoration: Fixed point convergence and applications. *CoRR*, abs/1605.01710, 2016. [2](#), [3](#), [7](#)
- [6] Zezhou Cheng, Matheus Gadelha, Subhransu Maji, and Daniel Sheldon. A bayesian perspective on the deep image prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. [2](#)
- [7] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007. [2](#), [3](#), [5](#), [7](#), [8](#)
- [8] A. Danielyan, V. Katkovnik, and K. Egiazarian. Bm3d frames and variational image deblurring. *IEEE Transactions on Image Processing*, 21(4):1715–1728, 2012. [6](#)
- [9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. [2](#)
- [10] Y C Eldar. Generalized SURE for Exponential Families: Applications to Regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481, Jan. 2009. [2](#), [3](#), [9](#)
- [11] Yossi Gandelsman, Assaf Shocher, and Michal Irani. “double-dip”: Unsupervised image decomposition via coupled deep-image-priors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2019. [2](#)
- [12] R. Giryes, M. Elad, and Y.C. Eldar. The projected gsure for automatic parameter tuning in iterative shrinkage methods. *Applied and Computational Harmonic Analysis*, 30(3):407 – 422, 2011. [3](#)
- [13] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 349–356. IEEE, 2009. [2](#)
- [14] Rana Hanocka, Gal Metzger, Raja Giryes, and Daniel Cohen-Or. Point2mesh: A self-prior for deformable meshes. *ACM Trans. Graph.*, 39(4), July 2020. [2](#)
- [15] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015. [2](#)

- [16] Ulugbek S Kamilov, Hassan Mansour, and Brendt Wohlberg. A plug-and-play priors approach for solving nonlinear imaging inverse problems. *IEEE Signal Processing Letters*, 24(12):1872–1876, 2017. 3
- [17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4, 6
- [18] Yoann Le Montagner, Elsa D Angelini, and Jean-Christophe Olivo-Marin. An Unbiased Risk Estimator for Image Denoising in the Presence of Mixed Poisson–Gaussian Noise. *IEEE Transactions on Image Processing*, 23(3):1255–1268, Aug. 2014. 2
- [19] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In *International Conference on Machine Learning 35 (ICML)*, pages 2965–2974, 2018. 2
- [20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2
- [21] Jiaming Liu, Yu Sun, Xiaojian Xu, and Ulugbek S Kamilov. Image restoration using total variation regularized deep image prior. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7715–7719. IEEE, 2019. 2
- [22] Gary Mataev, Peyman Milanfar, and Michael Elad. Deepred: Deep image prior powered by red. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 3, 5
- [23] Tim Meinhardt, Michael Moeller, Caner Hazirbas, and Daniel Cremers. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In *ICCV*, 2017. 3
- [24] C. Metzler, A. Mousavi, R. Heckel, and R. Baraniuk. Unsupervised Learning with Stein’s Unbiased Risk Estimator. In *International Biomedical and Astronomical Signal Processing (BASP) Frontiers workshop*, 2019. 2, 3
- [25] Christopher A Metzler, Arian Maleki, and Richard G Baraniuk. From denoising to compressed sensing. *IEEE Transactions on Information Theory*, 62(9):5117–5144, 2016. 3
- [26] Christopher A. Metzler, Ali Mousavi, and Richard G. Baraniuk. Learned d-amp: Principled neural network based compressive image recovery. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 1770–1781, 2017. 3
- [27] Yves Meyer. *Oscillating patterns in image processing and nonlinear evolution equations: the fifteenth Dean Jacqueline B. Lewis memorial lectures*, volume 22. American Mathematical Soc., 2001. 2
- [28] Seungjun Nah, Tae Hyung Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 257–265, 2016. 2
- [29] S Ramani, T Blu, and M Unser. Monte-Carlo Sure: A Black-Box Optimization of Regularization Parameters for General Denoising Algorithms. *IEEE Transactions on Image Processing*, 17(9):1540–1554, Aug. 2008. 3, 4
- [30] Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017. 2, 3
- [31] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. *arXiv preprint arXiv:1905.01164*, 2019. 2
- [32] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. Ingan: Capturing and retargeting the “dna” of a natural image. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2
- [33] Assaf Shocher, Nadav Cohen, and Michal Irani. “zero-shot” super-resolution using deep internal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3118–3126, 2018. 2, 5
- [34] Shakarim Soltanayev and Se Young Chun. Training deep learning based denoisers without ground truth data. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 3261–3271, 2018. 2, 3
- [35] S. Soltanayev, R. Giryes, S. Y. Chun, and Y. C. Eldar. On divergence approximations for unsupervised training of deep denoisers based on stein’s unbiased risk estimator. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3592–3596, 2020. 3
- [36] M. Soltanolkotabi, A. Javanmard, and J. D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2019. 4
- [37] Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, Nov. 1981. 2, 3
- [38] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 769–777, 2015. 2
- [39] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A Efros, and Moritz Hardt. Test-time training for out-of-distribution generalization. *arXiv preprint arXiv:1909.13231*, 2019. 2
- [40] Tom Tirer and Raja Giryes. Image restoration by iterative denoising and backward projections. *IEEE Transactions on Image Processing*, 28(3):1220–1234, 2018. 2, 3, 5
- [41] Tom Tirer and Raja Giryes. Super-resolution via image-adapted denoising CNNs: Incorporating external and internal learning. *IEEE Signal Processing Letters*, 26(7):1080–1084, 2019. 2
- [42] Tom Tirer and Raja Giryes. Back-projection based fidelity term for ill-posed linear inverse problems. *IEEE Transactions on Image Processing*, 29:6164–6179, 2020. 5
- [43] Tom Tirer and Raja Giryes. On the convergence rate of projected gradient descent for a back-projection based objective. *arXiv preprint arXiv:2005.00959*, 2020. 5
- [44] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018. 2, 4, 5

- [45] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948. IEEE, 2013. [2](#), [3](#), [5](#)
- [46] Francis Williams, Teseo Schneider, Claudio Silva, Denis Zorin, Joan Bruna, and Daniele Panozzo. Deep geometric prior for surface reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10130–10139, 2019. [2](#)
- [47] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3929–3938, 2017. [3](#), [5](#)
- [48] Yang Zhou, Zhen Zhu, Xiang Bai, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Non-stationary texture synthesis by adversarial expansion. *ACM Trans. Graph.*, 37(4):49:1–49:13, July 2018. [2](#)
- [49] Magauiya Zhussip, Shakarim Soltanayev, and Se Young Chun. Training deep learning based image denoisers from undersampled measurements without ground truth and without image prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. [3](#)
- [50] Jenny Zukerman, Tom Tirer, and Raja Giryes. BP-DIP: A backprojection based deep image prior. *arXiv preprint arXiv:2003.05417*, 2020. [2](#), [4](#)