Discriminative Transfer Learning for General Image Restoration

Lei Xiao, Felix Heide, Wolfgang Heidrich, Bernhard Schölkopf, Michael Hirsch

***Abstract*—Recently, several discriminative learning approaches have been proposed for effective image restoration, achieving convincing trade-off between image quality and computational efficiency. However, these methods require separate training for each restoration task (e.g., denoising, deblurring, demosaicing) and problem condition (e.g., noise level of input images). This makes it time-consuming and difficult to encompass all tasks and conditions during training. In this paper, we propose a discriminative transfer learning method that incorporates formal proximal optimization and discriminative learning for general image restoration. The method requires a single-pass discriminative training and allows for reuse across various problems and conditions while achieving an efficiency comparable to previous discriminative approaches. Furthermore, after being trained, our model can be easily transferred to new likelihood terms to solve untrained tasks, or be combined with existing priors to further improve image restoration quality.**

***Index Terms*—Image restoration, discriminative learning, proximal optimization**

I. Introduction

Low-level vision problems, such as denoising, deconvolution and demosaicing, have to be addressed as part of most imaging and vision systems. Although a large body of work covers these classical problems, low-level vision is still a very active area. The reason is that, from a Bayesian perspective, solving them as statistical estimation problems does not only rely on models for the likelihood (i.e. the reconstruction task), but also on natural image priors as a key component.

A variety of models for natural image statistics have been explored in the past. Traditionally, models for gradient statistics [1], [2], including total variation (TV), have been a popular choice. Another line of works explores patch-based image statistics, either as per-patch sparse model [3], [4] or modeling non-local similarity between patches [5], [6], [7]. These prior models are general in the sense that they can be applied for various likelihoods, with the image formation and noise setting as parameters. However, the resulting optimization problems are prohibitively expensive in many cases, rendering them impractical for many real-time tasks especially on mobile platforms.

Recently, a number of works [8], [9] have addressed this issue by truncating the iterative optimization and learning discriminative image priors, tailored to a specific reconstruction task (likelihood) and optimization approach. While these methods allow to trade-off quality with the computational budget for a given application, the learned models are highly specialized for the image formation model and noise parameters, in contrast to optimization-based approaches. Since each individual problem instantiation requires costly learning and storing of the model coefficients, current proposals for learned models are impractical for vision applications with dynamically changing (often continuous) parameters. This is a common scenario in most real-world image processing settings, as well as applications in engineering and scientific imaging that rely on the ability to rapidly prototype methods.

In this paper, we combine discriminative learning techniques with formal proximal optimization methods to learn generic models that can be truly transferred across problem domains while achieving comparable efficiency as previous discriminative approaches. Using proximal optimization methods [10], [11], [12] allows us to decouple the likelihood and prior, which is key to learning such shared models. It also means that we can rely on well-researched physically-motivated models for the likelihood, while learning priors from example data. We verify our technique using the same model for a variety of diverse low-level image reconstruction tasks and problem conditions, demonstrating the effectiveness and versatility of our approach. After training, our approach benefits from the proximal splitting techniques, and can be naturally transferred to new likelihood terms for untrained restoration tasks, or it can be combined with existing state-of-the-art priors to further improve the reconstruction quality. This is impossible with previous discriminative methods. In particular, we make the following contributions:

- We propose a discriminative transfer learning technique for general image restoration. It requires a single-pass discriminative training and transfers across different restoration tasks and problem conditions.
- We show that our approach is general by demonstrating its robustness for diverse low-level problems, such as denoising, deconvolution, inpainting, and for varying noise

Manuscript received June 28, 2017; revised March 10, 2018; accepted April 23, 2018. This work was in part supported by King Abdullah University of Science and Technology under individual baseline funding. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Weisi Lin.

L. Xiao is currently with Oculus Research, a division of Oculus VR, LLC.. The work represented by this paper was done when L. Xiao was with the Department of Computer Science at University of British Columbia. F. Heide is with the Department of Electrical Engineering at Stanford University. W. Heidrich is with the Visual Computing Center at King Abdullah University of Science and Technology, and the Department of Computer Science at University of British Columbia. B. Schölkopf and M. Hirsch are with Max Planck Institute for Intelligent Systems.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the authors. The material includes additional results and comparisons. The source code is available at https://github.com/lxgh/DTL. Contact lei.xiao@oculus.com for further questions about this work.

settings.

- We show that, while being general, our method achieves comparable computational efficiency as previous discriminative approaches, making it suitable for processing high-resolution images on mobile imaging systems.
- We show that our method can naturally be combined with existing likelihood terms and priors after being trained. This allows our method to process untrained restoration tasks and take advantage of previous successful work on image priors (e.g., color and non-local similarity priors).

II. Related work

Image restoration aims at computationally enhancing the quality of images by undoing the adverse effects of image degradation such as noise and blur. As a key area of image and signal processing it is an extremely well studied problem and a plethora of methods exists, see for example [13] for a recent survey. Through the successful application of machine learning and data-driven approaches, image restoration has seen revived interest and much progress in recent years. Broadly speaking, recently proposed methods can be grouped into three classes: *classical* approaches that make no explicit use of machine learning, *generative* approaches that aim at probabilistic models of undegraded natural images and *discriminative* approaches that try to learn a direct mapping from degraded to clean images. Unlike classical methods, methods belonging to the latter two classes depend on the availability of training data.

Classical models: This class of methods focus on local image statistics and aim at maintaining edges. Examples include total variation [1], bilateral filtering [14], anisotropic diffusion models [15] and kernel regression (KR) [16]. More recent methods exploit the non-local statistics of images with the fundamental observation that similar patches often can be found within an image. Representative work include the non-local mean (NLM) method [17], block-matching and 3D filtering (BM3D) [5] and non-local variants of sparse and low-rank representation methods [18], [6], [7], [19], [20], [21]. Specifically, BM3D extends the non-local similarity idea first introduced in NLM, however combines them through collaborative patch-filtering steps instead of simple pixel averaging. The non-local sparse representation methods, e.g. learned simultaneous sparse coding (LSSC), explore the patch similarity idea while enforcing similar patches to have similar coefficients in transform domains. The weighted nuclear norm minimization (WNNM) method [7] filters similar patches together by applying low-rank constraints on singular value decomposition of patch stacks. NLR-CS [20] applies the non-local low rank constraint to compressive sensing for image recovery. The group-based sparse representation (GSR) [21] method models natural images in the domain of group sparsity and exploits the intrinsic local sparsity and non-local similarity simultaneously. While effective search for similar patches/pixels is important for these non-local methods, the extensively used mean-square-error (MSE) as a similarity metric appears ineffective for images with high noise and distortion [22]. More recent methods use perceptually motivated similarity metrics (e.g. structural similarity (SSIM) and gradient magnitude similarity deviation (GMSD)) for improved restoration quality [23], [24].

TABLE I: Analysis of state-of-the-art methods. In the table, "Transferable" means the model can be used for different restoration tasks and problem conditions; "Modular" means the method can be combined with other existing priors at test time. EPLL, plug-and-play (P&P), BM3D are representative generative methods, and CSF and TRD are representative discriminative methods. Our approach DTL is able to combine the strengths of both generative and discriminative models.

	EPLL	P&P	BM3D	CSF	TRD	DTL
Efficiency				✓	✓	✓
Parallelism				✓	✓	✓
Transferable	✓	✓	✓			✓
Modular	✓	✓	✓			✓

Generative learning models: This class of methods seek to learn probabilistic models of undegraded natural images. A simple yet powerful subclass include models that approximate the sparse gradient distribution of natural images, e.g. the ℓ_p-norm $(0 < p < 1)$ constraint on image derivatives [25], [2], [26]. More expressive generative models include k-singular value decomposition (KSVD) [3], convolutional sparse coding (CSC) [27], [28], [29], fields of experts (FoE) [30] and expected patch log likelihood (EPLL) [4]. While KSVD and CSC assume patches in an image can be approximated by a linear combination of a few atoms from an overcomplete dictionary that is learned from training data, FoE learns a set of filters whose responses on an image (i.e. the convolution of the image and the filter) are assumed to be sparse. EPLL models image patches through Gaussian Mixture Models (GMM) and applies this patch prior to the whole image through half-quadratic splitting (HQS) approach [10].

Another line of research, which is closely related to our approach, is the *plug-and-play* technique [31], [32], [33], [34]. In these methods, Gaussian denoisers are utilized as image regularizers for solving general inverse problems, through splitting optimization methods such as ADMM [12]. The fundamental difference between these methods and our approach is that they utilize an existing *generative* Gaussian denoiser, while our approach learns all parameters by discriminative learning thus achieving a better trade-off between high quality and time efficiency.

Generative models have in common that they are agnostic to the image restoration task, i.e. they are transferable to any image degradation and can be combined in a modular fashion with any likelihood and additional priors at test time. The downside is that they are typically expensive to solve, hampering their applications in real-time tasks especially on mobile platforms.

Discriminative learning models: This class of methods have recently become increasingly popular for image restoration due to their attractive tradeoff between high image restoration quality and efficiency at test time. Some representative examples of such methods include trainable random field models such as separable Markov random field (MRFSepa) [35] regression tree fields (RTF) [36], cascaded shrinkage fields

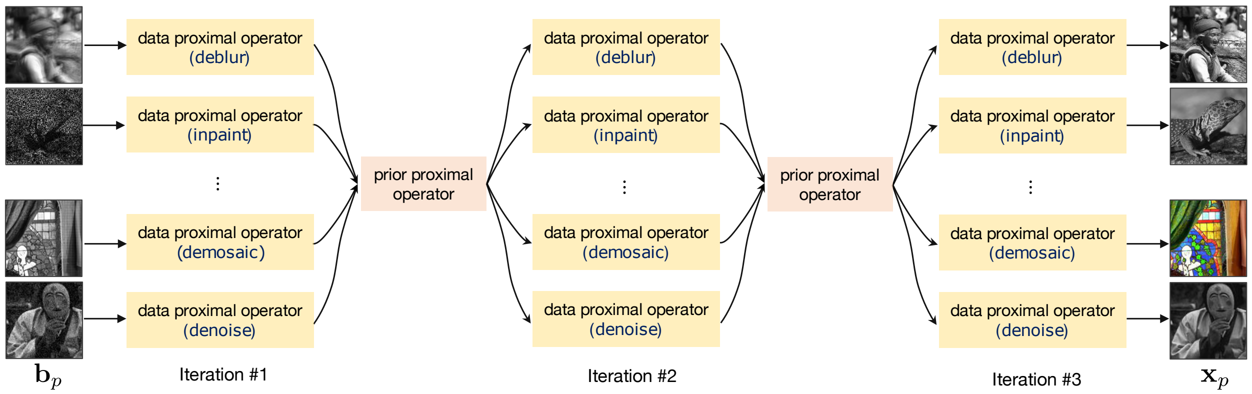

data proximal operator (deblur)
data proximal operator (inpaint)
⋮
data proximal operator (demosaic)
data proximal operator (denoise)
prior proximal operator
$\mathbf{b}_p$
Iteration #1
Iteration #2
Iteration #3
$\mathbf{x}_p$

Fig. 1: The architecture of our method. Input images are drawn from various restoration tasks and problem conditions. Each iteration uses the same model parameters, forming a recurrent network.

(CSF) [8], trainable nonlinear reaction diffusion (TRD) models [9] and their extensions [37], [38], [39]. The state-of-the-art CSF and TRD methods can be derived from the FoE model [30] by unrolling corresponding optimization iterations to be feed-forward networks, where the parameters of each network are trained by minimizing the error between its output images and ground truth for each specific task. Another line of research apply neural networks for image restoration, such as multi-layer perceptrons [40], deep convolutional networks [41], [42], [43] and deep recurrent neural networks [44].

Discriminative approaches owe their computational efficiency at run-time to a particular feed-forward structure whose trainable parameters are optimized for a particular task during training. Those learned parameters are then kept fixed at test-time resulting in a fixed computational cost. On the downside, discriminative models do not generalize across tasks and typically necessitate separate feed-forward architectures and separate training for each restoration task (denoising, demosaicing, deblurring, etc.) as well as every possible image degradation (noise level, Bayer pattern, blur kernel, etc.).

In this work, we propose the *discriminative transfer learning* (DTL) technique that is able to combine the strengths of both generative and discriminative models: it maintains the flexibility of generative models, but at the same time enjoys the computational efficiency of discriminative models. While in spirit our approach is akin to the recently proposed method of Rosenbaum and Weiss [45], who equipped the successful EPLL model with a discriminative prediction step, the key idea in our approach is to use proximal optimization techniques [10], [11], [12] that allow the decoupling of likelihood and prior and therewith share the full advantages of a Bayesian generative modeling approach.

Table I summarizes the properties of the most prominent state-of-the-art methods and puts our own proposed approach into perspective.

III. Proposed method

A. Diversity of data likelihood

The seminal work of fields-of-experts (FoE) [30] generalizes the form of filter response based regularizers in the objective function given in Eq. 1. The vectors $\mathbf{b}$ and $\mathbf{x}$ represent the observed and latent (desired) image respectively, the matrix $\mathbf{A}$ is the sensing operator, $\mathbf{F}_i$ represents 2D convolution with filter $\mathbf{f}_i$, and ϕ_i represents the penalty function on corresponding filter responses $\mathbf{F}_i\mathbf{x}$. The positive scalar λ controls the relative weight between the data fidelity (likelihood) and the regularization term.

$$\frac{\lambda}{2}||\mathbf{b}-\mathbf{A}\mathbf{x}||_2^2+\sum_{i=1}^{N}\phi_i(\mathbf{F}_i\mathbf{x}) \tag{1}$$

The well-known anisotropic total-variation regularizer can be viewed as a special case of the FoE model where $\mathbf{f}_i$ is the derivative operator ∇, and ϕ_i the ℓ_1 norm.

While there are various types of restoration tasks (e.g., denoising, deblurring, demosaicing) and problem parameters (e.g., noise level of input images), each problem has its own sensing matrix $\mathbf{A}$ and optimal fidelity weight λ. For example, $\mathbf{A}$ is an identity matrix for denoising, a convolution operator for deblurring, a binary diagonal matrix for demosaicing, and a random matrix for compressive sensing [46]. λ depends on both the task and its parameters in order to produce the best quality results.

The state-of-the-art discriminative learning methods (CSF [8], TRD [9]) derive an end-to-end feed-forward model from Eq. 1 for each specific restoration task, and train this model to map the degraded input images directly to the output. These methods have demonstrated a great trade-off between high-quality and time-efficiency, however, as an inherent problem of the discriminative learning procedure, they require separate training for each restoration task and problem condition. Given the diversity of data likelihoods in image restoration, this fundamental drawback of discriminative models makes it time-consuming and difficult to encompass all tasks and conditions during training.

B. Decoupling likelihood and prior

It is difficult to directly minimize Eq. 1 when the penalty function ϕ_i is non-linear and/or non-smooth (e.g. ℓ_p norm, $0 \leq p \leq 1$). Proximal algorithms [12], [10], [47] instead relax Eq. 1 and split the original problem into several easier subproblems that are solved alternately until convergence.

Algorithm 1 Proposed algorithm

```
Input: degraded image $\mathbf{b}$
Output: recovered image $\mathbf{x}$
$\mathbf{x}^0 = \mathbf{b}, \rho^1 = 1$ (initialization)
for $t = 1$ to $T$ do
    (Update $\mathbf{z}^t$ by Eq. 6 below)
    $\mathbf{z}_0^t = \mathbf{x}^{t-1}$
    for $k = 1$ to $K$ do
        $\mathbf{z}_k^t = \mathbf{z}_{k-1}^t - \sum_{i=1}^N {\mathbf{F}_i^k}^\mathsf{T} \psi_i^k(\mathbf{F}_i^k \mathbf{z}_{k-1}^t)$
    end for
    $\mathbf{z}^t = \mathbf{z}_K^t$
    (Update $\mathbf{x}^t$ by Eq. 4 below)
    $\mathbf{x}^t = \mathrm{argmin}_\mathbf{x} \lambda||\mathbf{b} - \mathbf{A}\mathbf{x}||_2^2 + \rho^t||\mathbf{z}^t - \mathbf{x}||_2^2$
    $\rho^{t+1} = 2\rho^t$
end for
```

In this paper we employ the half-quadratic-splitting (HQS) algorithm [10] to relax Eq. 1, as it typically requires much fewer iterations to converge compared to other proximal methods such as ADMM [12] and PD [47]. The relaxed objective function is given in Eq. 2:

$$\frac{\lambda}{2}||\mathbf{b} - \mathbf{A}\mathbf{x}||_2^2 + \frac{\rho}{2}||\mathbf{z} - \mathbf{x}||_2^2 + \sum_{i=1}^N \phi_i(\mathbf{F}_i\mathbf{z}), \tag{2}$$

where a slack variable $\mathbf{z}$ is introduced to approximate $\mathbf{x}$, and ρ is a positive scalar.

With the HQS algorithm, Eq. 2 is iteratively minimized by solving for the slack variable $\mathbf{z}$ and the latent image $\mathbf{x}$ alternately as in Eq. 3 and 4 ($t = 1, 2, ..., T$).

Prior proximal operator:

$$\mathbf{z}^t = \underset{\mathbf{z}}{\mathrm{argmin}} \left(\frac{\rho^t}{2}||\mathbf{z} - \mathbf{x}^{t-1}||_2^2 + \sum_{i=1}^N \phi_i(\mathbf{F}_i\mathbf{z}) \right), \tag{3}$$

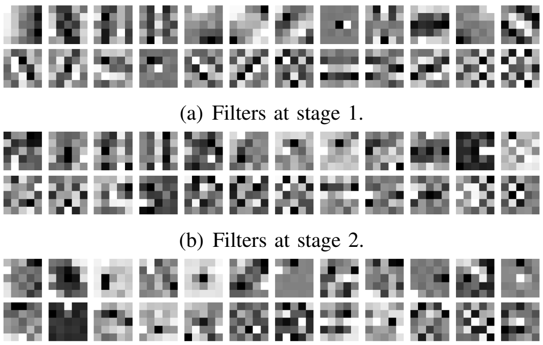
Data proximal operator:

$$\mathbf{x}^t = \underset{\mathbf{x}}{\mathrm{argmin}} \left(\lambda||\mathbf{b} - \mathbf{A}\mathbf{x}||_2^2 + \rho^t||\mathbf{z}^t - \mathbf{x}||_2^2 \right), \tag{4}$$

where ρ^t increases as the iteration continues. This forces $\mathbf{z}$ to become an increasingly good approximation of $\mathbf{x}$, thus making Eq. 2 an increasingly good proxy for Eq. 1.

Note that, while most related approaches including CSF relax Eq. 1 by splitting on $\mathbf{F}_i\mathbf{x}$, we split on $\mathbf{x}$ instead. This is critical for deriving our approach. With this new splitting strategy, the prior term and the data likelihood term in the original objective Eq. 1 are now separated into two subproblems that we call the "prior proximal operator" (Eq. 3) and the "data proximal operator" (Eq. 4), respectively.

We notice that recent plug-and-play work [31], [32], [33] adopt similar proximal splitting strategy as our method though with the ADMM framework. However, while plug-and-play methods adopt existing generic Gaussian denoiser for the prior proximal operator, our method trains the prior proximal operator and other parameters used in the optimization algorithm with discriminative learning technique. This makes our method share the advantage of discriminative restoration methods, that is, achieving great trade-off between high quality and time efficiency.



(a) Filters at stage 1.

(b) Filters at stage 2.

(c) Filters at stage 3.

Fig. 2: Trained filters at each stage (k in Eq. 6) of the proximal operator $\mathbf{prox}_\Theta$ in our model (3 stages each with 24 5×5 filters).

C. Discriminative transfer learning

While the data proximal operator in Eq. 4 is task-dependent because both the sensing matrix $\mathbf{A}$ and fidelity weight λ are problem-specific as explained in Sec. III-A, the prior proximal operator (i.e. $\mathbf{z}^t$-update step in Eq. 3) is independent of the original restoration tasks and problem conditions.

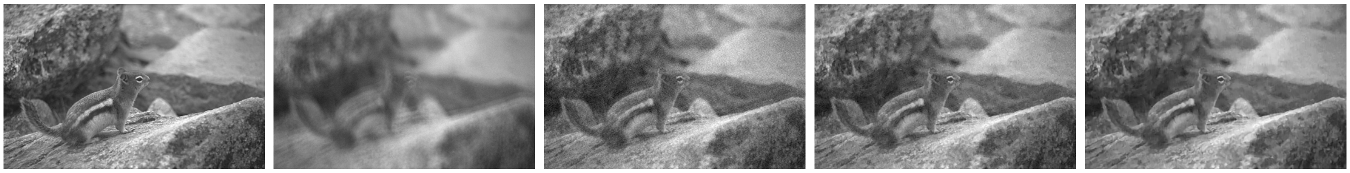
This leads to our main insight: *Discriminative learning models can be made transferable by using them in place of the prior proximal operator, embedded in a proximal optimization algorithm.* This allows us to generalize a single discriminative learning model to a very large class of problems, i.e. any linear inverse imaging problem, while simultaneously overcoming the need for problem-specific retraining. Moreover, it enables learning the task-dependent parameter λ in the data proximal operator for each problem in a single training pass, eliminating tedious hand-tuning at test time. As will be further explained later, we train various restoration tasks and problem conditions simultaneously.

Benefiting from our new splitting strategy, the prior proximal operator in Eq. 3 can be interpreted as a Gaussian denoiser on the intermediate image $\mathbf{x}^{t-1}$, since the least-squares consensus term is equivalent to a Gaussian denoising term. This inspires us to utilize existing discriminative models that have been successfully used for denoising (e.g. CSF, TRD).

For convenience, we denote the prior proximal operator as $\mathbf{prox}_\Theta$, i.e.

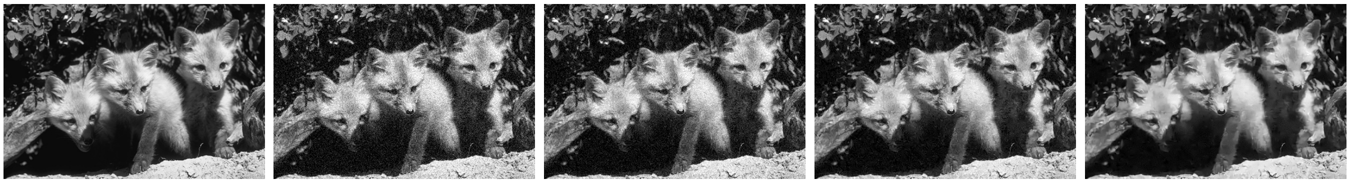
$$\mathbf{z}^t := \mathbf{prox}_\Theta(\mathbf{x}^{t-1}, \rho^t), \tag{5}$$

where the model parameter Θ includes a number of filters $\mathbf{f}_i$ and corresponding penalty functions ϕ_i. Inspired by the state-of-the-art discriminative methods [8], [9], we propose to learn the model $\mathbf{prox}_\Theta$, and the fidelity weight scalar λ, from training data. Recall that with our new splitting strategy introduced in Sec. III-B, the image prior and data-fidelity term in the original objective (Eq. 1) are contained in two separate subproblems (Eq. 3 and 4). This makes it possible to train together an ensemble of diverse tasks (e.g., denoising, deblurring, or with different noise levels) each of which has its own data proximal operator, while learning a single prior proximal operator $\mathbf{prox}_\Theta$ that is shared across tasks. This is



(a) Ground truth (b) Blurry input (23.37dB) (c) Iter 1 (27.32dB) (d) Iter 2 (28.48dB) (e) Iter 3 (29.36dB)

Fig. 3: Results at each HQS iteration of our method on non-blind deconvolution with a 25×25 PSF and noise level $\sigma = 3$.



(a) Ground truth (b) Noisy input (20.18dB) (c) Iter 1 (22.85dB) (d) Iter 2 (25.93dB) (e) Iter 3 (28.14dB)

Fig. 4: Results at each HQS iteration of our method on image denoising with noise level $\sigma = 25$.

in contrast to state-of-the-art discriminative methods such as CSF and TRD which train separate models for each task.

For clarity, in Fig. 1 we visualize the architecture of our method. The input images may represent various restoration tasks and problem conditions. At each HQS iteration, each image $\mathbf{x}_p^t$ from problem p is updated by *its own* data proximal operator in Eq. 4 which contains separate trainable fidelity weight λ_p and pre-defined sensing matrix $\mathbf{A}_p$; then each slack image $\mathbf{z}_p^t$ is updated by the same, *shared* prior proximal operator implemented by a learned, discriminative model.

Recurrent network: Note that in Fig. 1 each HQS iteration uses exactly the same model parameters, forming a recurrent network akin to [44]. This is in contrast to previous discriminative learning methods including CSF and TRD, which form feed-forward networks. Our recurrent network architecture maintains the convergence property of the proximal optimization algorithm (HQS), and is critical for our method to transfer between various tasks and problem conditions.

Shared prior proximal operator: While *any* discriminative Gaussian denoising model could be used as $\mathbf{prox}_\Theta$ in our framework, we specifically propose to use the multi-stage non-linear diffusion process that is modified from the TRD model, for its efficiency. The model is given in Eq. 6.

$$\begin{aligned} \mathbf{z}_k^t &= \mathbf{z}_{k-1}^t - \sum_{i=1}^{N} \mathbf{F}_i^{k\top} \psi_i^k(\mathbf{F}_i^k \mathbf{z}_{k-1}^t), \\ s.t. \quad & \mathbf{z}_0^t = \mathbf{x}^{t-1}, \quad k = 1, 2, ..., K, \end{aligned} \tag{6}$$

where k is the stage index, filters $\mathbf{F}_i^k$, function ψ_i^k are trainable model parameters at each stage, and $\mathbf{z}_0^t$ is the initial value of $\mathbf{z}_k^t$. Note that, different from TRD, our model does not contain the reaction term which would be $-\rho^t \alpha_k (\mathbf{z}_{k-1}^t - \mathbf{x}^{t-1})$ with step size α_k. The main reasons for this modification are:

- The data constraint is contained in $\mathbf{x}^t$ update in Eq. 4;
- More importantly, by dropping the reaction term our model gets rid of the weight ρ^t which changes at each HQS iteration. Therefore, our proximal operator $\mathbf{prox}_\Theta(\mathbf{x}^{t-1}, \rho^t)$ is simplified to be:

$$\mathbf{z}^t := \mathbf{prox}_\Theta(\mathbf{x}^{t-1}) \tag{7}$$

The set of trainable parameters Ω in our method includes λ's for each problem class p (restoration task and problem condition), and $\Theta = \{\mathbf{F}_i^k, \psi_i^k\}$ in the prior proximal operator shared across different classes, i.e. $\Omega = \{\lambda_p, \Theta\}$. Even though the scalar parameters λ_p are trained, our method allows users to override them at test time to handle non-trained problem classes or specific inputs as we will show in Sec. IV. This contrasts to previous discriminative approaches whose model parameters are all fixed at test time. The subscript p indicating the problem class in λ_p is omitted below for convenience. The values of ρ^t are pre-selected: $\rho^1 = 1$ and $\rho^t = 2\rho^{t-1}$ for $t > 1$.

Note that a multi-stage model as in Eq. 6 is not possible if we split on $\mathbf{F}_i\mathbf{x}$ instead of $\mathbf{x}$ in Eq. 1 and 2. For clarity, an overview of the proposed algorithm is given in Alg. 1.

D. Training

We consider denoising and deconvolution tasks at training, where the sensing operator $\mathbf{A}$ is an identity matrix, or a block circulant matrix with circulant blocks that represents 2D convolution with randomly drawn blur kernels respectively. In denoising tasks, the $\mathbf{x}^t$ update in Eq. 4 has a closed-form solution:

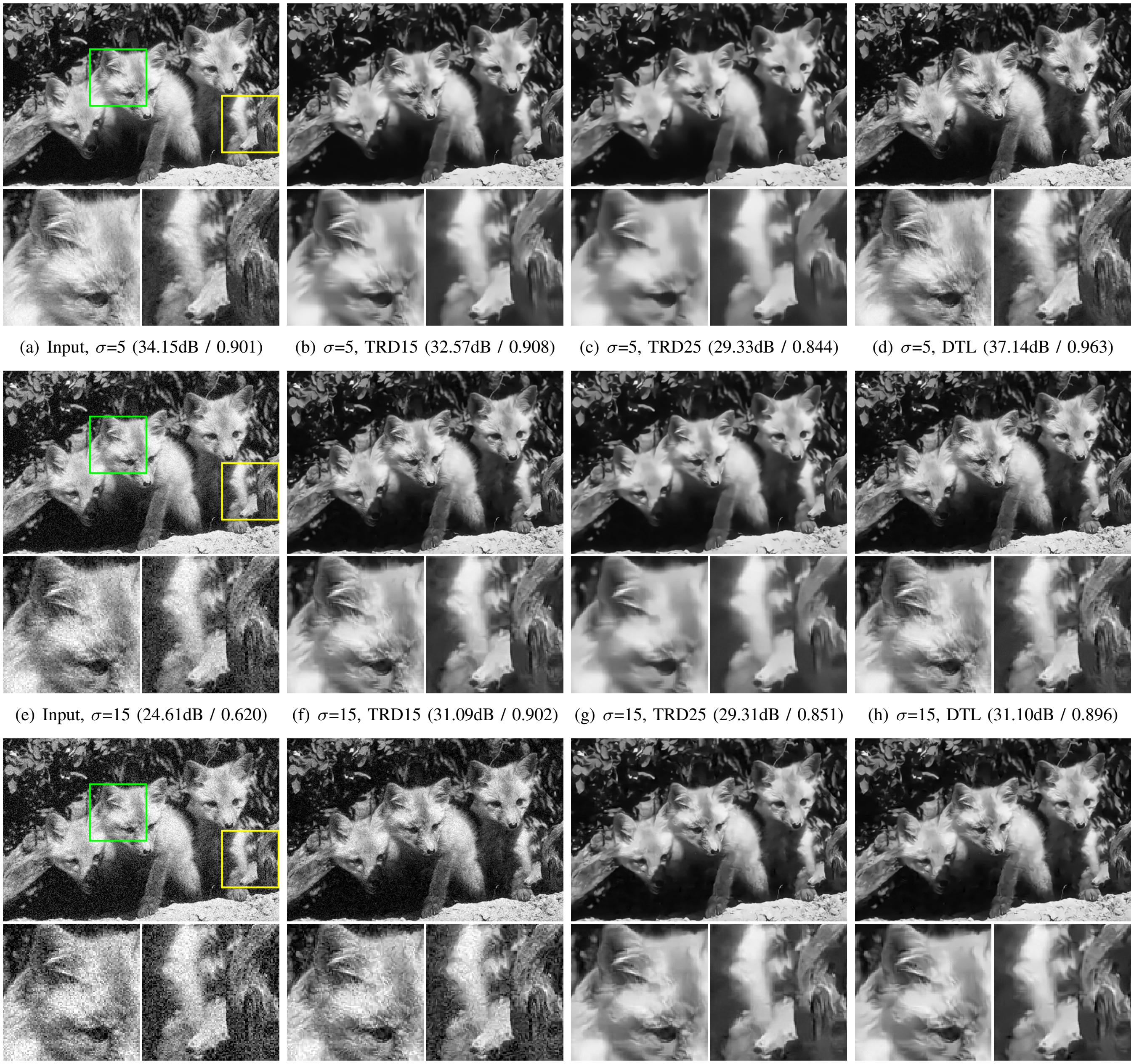
$$\mathbf{x}^t = (\lambda \mathbf{b} + \rho^t \mathbf{z}^t)/(\lambda + \rho^t) \tag{8}$$

In deconvolution tasks, the $\mathbf{x}^t$ update in Eq. 4 has a closed-form solution in the Fourier domain:

$$\mathbf{x}^t = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\lambda \mathbf{A}^\top \mathbf{b} + \rho^t \mathbf{z}^t)}{\mathcal{F}(\lambda \mathbf{A}^\top \mathbf{A} + \rho^t)}\right), \tag{9}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ represent Fourier and inverse Fourier transform respectively. Note that, compared to CSF, our method does not require FFT computations for denoising tasks. We use the L-BFGS solver [48] with analytic gradient computation for training. The training loss function ℓ is defined as the negative average Peak Signal-to-Noise Ratio (PSNR) of reconstructed images. The gradient of ℓ w.r.t. the model parameters $\Omega = \{\lambda_p, \Theta\}$ is computed by accumulating gradients at all HQS iterations, i.e.

$$\frac{\partial \ell}{\partial \Omega} = \sum_{t=1}^{T} \left(\frac{\partial \mathbf{x}^t}{\partial \lambda} \frac{\partial \ell}{\partial \mathbf{x}^t} + \frac{\partial \mathbf{z}^t}{\partial \Theta} \frac{\partial \mathbf{x}^t}{\partial \mathbf{z}^t} \frac{\partial \ell}{\partial \mathbf{x}^t} \right). \tag{10}$$



(a) Input, σ=5 (34.15dB / 0.901) (b) σ=5, TRD15 (32.57dB / 0.908) (c) σ=5, TRD25 (29.33dB / 0.844) (d) σ=5, DTL (37.14dB / 0.963)

(e) Input, σ=15 (24.61dB / 0.620) (f) σ=15, TRD15 (31.09dB / 0.902) (g) σ=15, TRD25 (29.31dB / 0.851) (h) σ=15, DTL (31.10dB / 0.896)

(i) Input, σ=25 (20.17dB / 0.441) (j) σ=25, TRD15 (23.74dB / 0.589) (k) σ=25, TRD25 (28.44dB / 0.845) (l) σ=25, DTL (28.45dB / 0.837)

Fig. 5: Generality analysis. (a)-(d), (e)-(h) and (i)-(l) show the input noisy images and denoised results by each method, at noise level $\sigma = 5, 15, 25$ respectively. Inside the bracket of each sub-caption PSNR (dB) and SSIM values are reported. While the performance of TRD deteriorates quickly as the noise level at test differs from the level at training (it either fails to remove noise or over-smoothes textures), the proposed method DTL is more generic and works robustly for various noise levels. More quantitative comparisons can be found in Fig. 6.

The 1D functions ψ_i^k in Eq. 6 are parameterized as a linear combination of equidistant-positioned Gaussian kernels whose weights are trainable.

Progressive training: A progressive scheme is proposed to make the training more effective. First, we set the number of HQS iterations to be 1, and train $\lambda's$ and the model Θ of each stage in $\mathbf{prox}_\Theta$ in a greedy fashion. Then, we gradually increase the number of HQS iterations from 1 to T where at each step the model $\Omega = \{\lambda, \Theta\}$ is refined from the result of the previous step. The L-BFGS iterations are set to be 200 for the greedy training steps, and 100 for the refining steps. Fig. 2 shows examples of learned filters in $\mathbf{prox}_\Theta$. In our final implementation, the filter size is chosen to be 5 by 5 pixels, which yields a good trade-off between result quality and time efficiency at test time. This observation is straightforward and consistent with previous work CSF and TRD where detailed experimental analysis can be found.

E. Connection and difference with related methods

In this section we emphasize the fundamental algorithmic differences between our method with several closely related prior work that are partly discussed in Section II. As given in Table I, our method is able to combine the strengths of both generic and discriminative methods.

Plug-and-play priors ([31], [32], [33], [43]): Like the plug-and-play work, our method uses formal optimization with a proximal operator framework. However, while plug-and-play methods adopt an existing generic Gaussian denoiser for the prior proximal operator, our method trains the prior proximal operator with discriminative learning technique. This makes our method share the advantage of discriminative restoration methods, that is, achieving great trade-off between high quality and time efficiency.

Discriminative learning methods ([36], [8], [9], [41], [40]): Previous discriminative learning methods require separate training for each restoration task (denoise, deblur, demosaic) and problem condition (noise levels, blur kernels). This makes it time-consuming and difficult to encompass all tasks and conditions during training. In contrast, by incorporating discriminative learning with formal proximal optimization, our method only requires a single-pass training and allows for reuse across various problems and conditions while achieving an efficiency comparable to previous discriminative approaches.

Cascaded shrinkage fields (CSF) [8]: Our method reuses the same model parameters in each iteration, while the splitting weight ρ is increased after each iteration. Doing so retains the convergence properties of the proximal optimization method. In contrast, CSF trains a different model for each iteration, and ρ is not kept as a separate parameter. Hence, CSF loses the convergence property of HQS. In addition, the splitting strategy CSF adopted does not separate data and regularizer terms (see Eq. 4-6 in [8]), which makes it impossible to share models across different tasks. Moreover, the splitting approach employed in CSF prohibits closed-form solutions for masked imaging problems, e.g. demosaicking, inpainting, joint inpainting and denoising.

IV. Results

A. Denoising and generality analysis

We compare the proposed discriminative transfer learning (DTL) method with state-of-the-art image denoising techniques, including KSVD [3], FoE [30], BM3D [5], LSSC [18], WNNM [7], EPLL [4], opt-MRF [49], ARF [50], CSF [8] and TRD [9]. The subscript in CSF_5 and TRD_5 indicates the number of cascaded stages (each stage has different model parameters). The subscript and superscript in our method DTL_3^5 indicate the number of diffusion stages ($K = 3$ in Alg. 1) in the prior proximal operator $\mathbf{prox}_\Theta$, and the number of HQS iterations ($T = 5$ in Alg. 1), respectively. Note that the complexity (size) of our model is linear in K, but independent of T. CSF, TRD and DTL use 24 filters of size 5×5 pixels at all stages in this section.

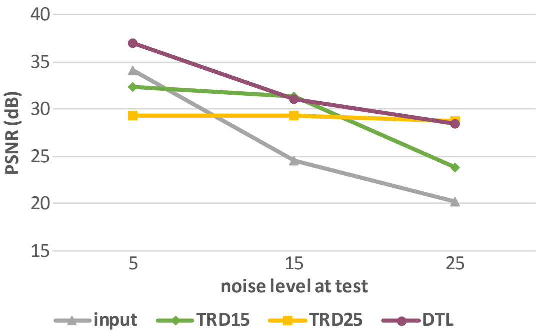


Fig. 6: Analysis of model generality on image denoising. “TRD15” denotes the TRD model trained at noise $\sigma = 15$, and “TRD25” trained at noise $\sigma = 25$. While our model DTL is trained with mixed noise levels in a single pass and used at various noise levels, TRD has been specialized to a single noise level matching the test images. Although it outperforms our method at exactly matching noise levels, quality drops down quickly when the test noise levels differs slightly from the trained ones. In contrast, our DTL model is robust across a wide range of noise levels. Example visual results are given in Fig. 5 and supplementary material.

TABLE II: Average PSNR (dB) on 68 images from [30] for denoising.

KSVD	FoE	BM3D	LSSC	WNNM	EPLL
30.87	30.99	31.08	31.27	31.37	31.19
opt-MRF	ARF	CSF_5	TRD_5	DTL_3^3	DTL_3^5
31.18	30.70	31.14	31.30	30.92	31.02

The compared discriminative methods, CSF_5 and TRD_5 both are trained at single noise level $\sigma = 15$ that is the same as the test images. In contrast, our model is trained on 400 images (100×100 pixels) cropped from [30] with random and discrete noise levels (standard deviation σ) varying between 5 and 25. The images with the same noise level share the same data fidelity weight λ at training.

TABLE III: Average PSNR (dB) on 32 images from [51] for non-blind deconvolution.

Input	Levin [25]	Schmidt [52]	CSF_3	DTL_3^3
22.86	32.73	33.97	33.48	34.34

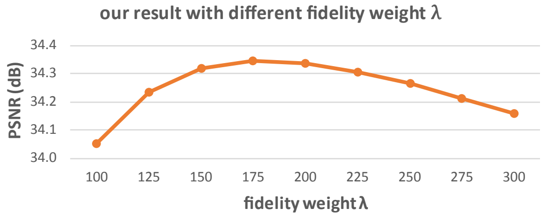


Fig. 7: Our results with different fidelity weight λ for the non-blind deconvolution experiment reported in Table III.

Generality analysis: To verify the generality of our method on varying noise levels, we test our model DTL_3^3 (trained with varying noise levels in a single pass) and two TRD models (trained at specific noise levels 15 and 25) on 3 sets of 68 images with noise $\sigma = 5, 15, 25$ respectively.

TABLE IV: Test with different HQS iterations (T) and model stages (K) for image denoising. Average PSNR (dB) results on 68 images from [51] with noise $\sigma = 15$ and 25 are reported (before and after "/" in each cell respectively).

		# HQS iterations		
		1	3	5
# stages	1	29.84 / 26.81	30.91 / 28.12	30.98 / 28.28
	3	30.56 / 27.82	30.92 / 28.19	31.02 / 28.42
	5	30.56 / 27.83	30.94 / 28.18	-

The average PSNR values are shown in Fig. 6. Although performing slightly below the TRD model trained for the exact noise level used at test time, our method is more generic and works robustly for various noise levels. The performance of the discriminative TRD method drops down quickly as the problem condition (i.e. noise level) at test differs from its training data (i.e., it either fails to remove noise or over-smoothes textures). In sharp contrast to discriminative methods (CSF, TRD, etc), which are inherently specialized for a given problem setting, i.e. noise level, the proposed approach transfers across different problem settings. In Fig. 5 we show example images from this analysis for visual comparison. In our model, the learned parameter λ for each noise level is $\lambda = 20.706$ for $\sigma = 5$, $\lambda = 2.475$ for $\sigma = 15$, and $\lambda = 0.033$ for $\sigma = 25$.

All compared methods are evaluated on the 68 test images from [30] and the averaged PSNR values are reported in Table II. The compared discriminative methods (CSF, TRD, etc) were trained for exactly the same noise level as the test images (i.e. the best case for them), while our model was trained with mixed noise levels and works robustly for arbitrary noise levels. Our results are comparable to generic methods such as KSVD, FoE and BM3D, and very close to discriminative methods such as CSF$_5$, while at the same time being much more time-efficient.

B. Analysis of convergence and model complexity

To better understand the convergence properties of our method, we show the intermediate results of each HQS iteration of our method DTL$_3^5$ on the denoising task in Fig. 4. The result image quality is progressively and significantly improved with each HQS iteration.

In above Sec. IV-A we demonstrate the results of our method trained with 3 and 5 HQS iterations, and 3 diffusion stages (i.e. DTL$_3^3$ and DTL$_3^5$). These hyper-parameters are chosen to balance the result quality and run-time efficiency (discussed in following Sec. IV-C) of the trained models. To further understand the tradeoff between model complexity and the number of HQS iteration, we report test results in Table IV for models trained with a varying number of HQS iterations (T in Alg. 1) and stages in $\mathbf{prox}_\Theta$ (K in Alg. 1).

C. Run-time comparison

TABLE V: Runtime (seconds) comparison.

Image size	256^2	512^2	1024^2	2048^2	4096^2
WNNM	157.73	657.75	2759.79	-	-
EPLL	29.21	111.52	463.71	-	-
BM3D	0.78	3.45	15.24	62.81	275.39
CSF$_5$	1.23	2.22	7.35	27.08	93.66
TRD$_5$	0.39	0.71	2.01	7.57	29.09
DTL$_3^3$	0.60	1.19	3.45	12.97	56.19
DTL$_3^3$ (Halide)	0.11	0.26	1.60	5.61	20.85

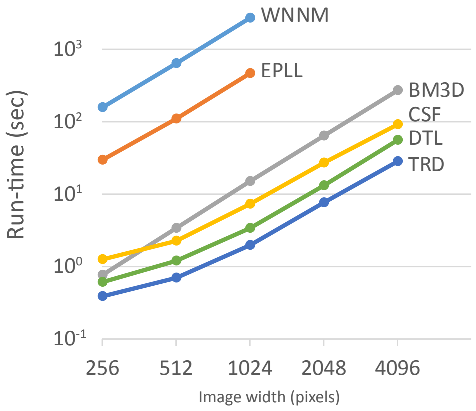


Fig. 8: Visualization of the runtime comparison that is reported in Table V.

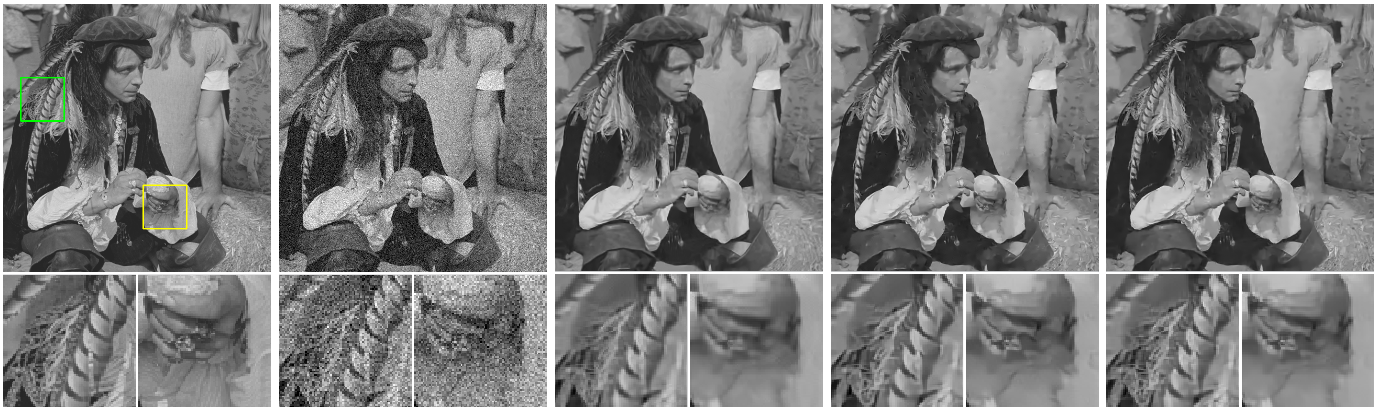
In Table V and Fig. 8 we compare the run-time of our method and state-of-the-art methods. The experiments were performed on a laptop computer with Intel i7-4720HQ CPU and 16GB RAM. WNNM and EPLL ran out-of-memory for images over 4 megapixels in our experiments. CSF$_5$, TRD$_5$ and DTL$_3^3$ all use "parfor" setting in Matlab. DTL$_3^3$ is significantly faster than all compared generic methods (WNNM, EPLL, BM3D) and even the discriminative method CSF$_5$. Run-time of DTL$_3^3$ is about 1.5 times that of TRD$_5$, which is expected as they use 5 versus 9 diffusion steps in total. In addition, we implement our method in Halide language [53], which has become popular recently for high-performance image processing applications, and report the run-time on the same CPU as mentioned above.

D. Deconvolution

In this experiment, we train a model DTL$_3^3$ with an ensemble of denoising and deconvolution tasks on 400 images (100×100 pixels) cropped from [30], in which 250 images are generated for denoising tasks with random noise levels σ varying between 5 and 25, and the other 150 images are generated by blurring the images with random 25×25 kernels (PSFs) and then adding Gaussian noise with σ ranging between 1 and 5. All images are quantized to 8 bits.

We compare our method with state-of-the-art non-blind deconvolution methods including Levin et al. [25], Schmidt et al. [52] and CSF [8]. Note that TRD [9] does not support non-blind deconvolution. We test the methods on the benchmark dataset from [51] which contains 32 images and report the average PSNR values in Table III.

As said in Sec. III-C, while the scalar weight λ is trained, our method allows users to override it at test time for untrained problem classes or specific inputs. Fig. 7 shows our results with different λ on the experiments compared in Table III.



(a) Ground truth (b) Input (20.17dB / 0.333) (c) BM3D (29.62dB / 0.804) (d) DTL$^5_3$ (29.48dB / 0.793) (e) DTL$^5_3$+BM3D (29.74dB / 0.809)

Fig. 9: Experiment on combining a non-local patch similarity prior (BM3D) with our model after being trained. The input noise level $\sigma = 25$. Inside the bracket of each sub-caption PSNR (dB) and SSIM values are shown. (c) BM3D performs well in removing noise especially in smooth regions but usually over-smoothes edges and textures. (d) DTL$^5_3$ well preserves sharp edges however sometimes introduces artifacts in smooth regions when the input noise level is high. (e) the hybrid method improves the result both visually and quantitatively. Please zoom in for better view.

Within a fairly wide range of λ, our method outperforms the previous methods.

To better understand the convergence properties of our method, we show the intermediate deconvolution results of our method at each HQS iteration in Fig. 3.

We further test the above model DTL$^3_3$ trained with ensemble tasks on the denoising experiment in Table II. The resulting average PSNR is 30.98dB, which is comparable to the result (30.92dB) with the model trained only on the denoising task.

E. Modularity with existing priors

As shown above, even though the fidelity weight λ is trainable, our method allows users to override its value at test time. This property also makes it possible to combine our model (after being trained) with existing state-of-the-art priors at test time, in which case λ typically needs to be adjusted. This allows our method to take advantage of previous successful work on image priors. Again, this is not possible with previous discriminative methods (CSF, TRD).

In Fig. 9 we show an example to incorporate a non-local patch similarity prior (BM3D [5]) with our method to further improve the denoising quality. BM3D performs well in removing noise especially in smooth regions but usually over-smoothes edges and textures. Our original model (DTL$^5_3$) well preserves sharp edges however sometimes introduces artifacts in smooth regions when the input noise level is high. By combining those two methods, which is easy with our HQS framework, the result is improved both visually and quantitatively.

We give the derivation of the proposed hybrid method below. Let $\mathcal{S}(\mathbf{x})$ represents the non-local patch similarity prior. The objective function is:

$$\frac{\lambda}{2}||\mathbf{b}-\mathbf{A}\mathbf{x}||_2^2+\sum_{i=1}^{N}\phi_i(\mathbf{F}_i\mathbf{x})+\tau\mathcal{S}(\mathbf{x}) \tag{11}$$

Applying the HQS technique described in Sec. III, we relax the objective to be:

$$\begin{aligned}\frac{\lambda}{2}||\mathbf{b}-\mathbf{A}\mathbf{x}||_2^2+\frac{\rho}{2}||\mathbf{z}-\mathbf{x}||_2^2+\sum_{i=1}^{N}\phi_i(\mathbf{F}_i\mathbf{z})\\ +\frac{\rho_s}{2}||\mathbf{v}-\mathbf{x}||_2^2+\tau\mathcal{S}(\mathbf{v})\end{aligned} \tag{12}$$

Then we minimize Eq. 12 by alternately solving the following 3 subproblems:

$$\begin{aligned}\mathbf{z}^t &= \mathbf{prox}_{\Theta}(\mathbf{x}^{t-1})\\ \mathbf{v}^t &= \operatorname*{argmin}_{\mathbf{v}}\frac{\rho_s^t}{2}||\mathbf{v}-\mathbf{x}^{t-1}||_2^2+\tau\mathcal{S}(\mathbf{v})\approx \text{BM3D}(\mathbf{x}^{t-1},\frac{\tau}{\rho_s^t})\\ \mathbf{x}^t &= \operatorname*{argmin}_{\mathbf{x}}\lambda||\mathbf{b}-\mathbf{A}\mathbf{x}||_2^2+\rho^t||\mathbf{z}^t-\mathbf{x}||_2^2+\rho_s^t||\mathbf{v}^t-\mathbf{x}||_2^2,\end{aligned} \tag{13}$$

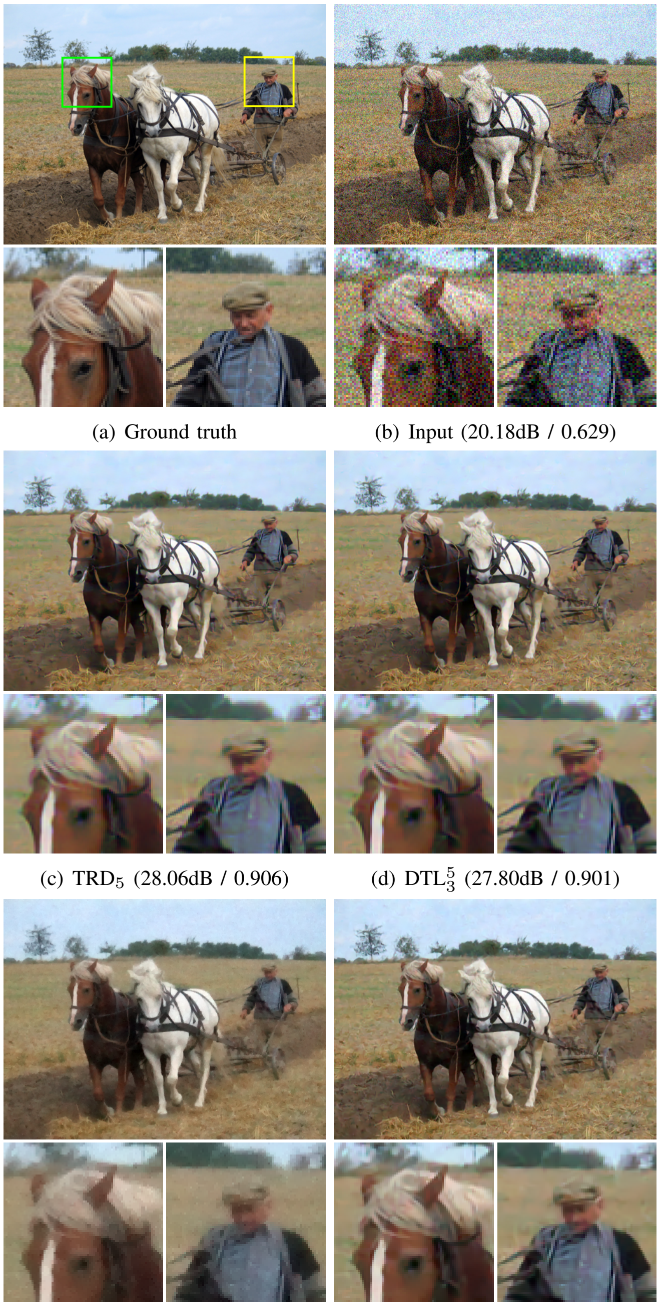
where $\mathbf{prox}_{\Theta}$ is from our previous training, and the $\mathbf{v}^t$ subproblem is approximated by running BM3D software on $\mathbf{x}^{t-1}$ with noise parameter τ/ρ_s^t following [54].

Similarly, our method can incorporate color image priors (e.g., cross-channel edge-concurrence prior [54]) to improve test results on color images, despite our model being trained on gray-scale images. Specifically, let the color image prior be

$$\mathcal{C}(\mathbf{x})=\sum_{i,j\in\{R,G,B\},i\neq j}||\nabla\mathbf{x}_i-\nabla\mathbf{x}_j||_1, \tag{14}$$

where the subscripts i, j represent color channel (RGB) and $\mathbf{x}_i$ represents the i-th channel image component. The objective function of the hybrid method is

$$\frac{\lambda}{2}||\mathbf{b}-\mathbf{A}\mathbf{x}||_2^2+\sum_{i=1}^{N}\phi_i(\mathbf{F}_i\mathbf{x})+\tau\mathcal{C}(\mathbf{x}), \tag{15}$$



(a) Ground truth

(b) Input (20.18dB / 0.629)

(c) TRD_5 (28.06dB / 0.906)

(d) DTL_3^5 (27.80dB / 0.901)

(e) TV + color prior (26.89dB / 0.881)

(f) DTL_3^5 + color prior (28.69dB / 0.917)

Fig. 10: Experiment on incorporating a color prior [54] with our model after being trained. The input noise level $\sigma = 25$. (e,f) show the results of combining total variation (TV) denoising with a cross-channel prior, and our method with a cross-channel prior, respectively. PSNR (dB) and SSIM values are reported inside the bracket of each sub-caption. Please zoom in for better view.

with the following optimization subproblems

$$\begin{aligned}
\mathbf{z}^t &= \mathbf{prox}_\Theta(\mathbf{x}^{t-1}) \\
\mathbf{v}^t &= \underset{\mathbf{v}}{\operatorname{argmin}} \frac{\rho_s^t}{2}||\mathbf{v} - \mathbf{x}^{t-1}||_2^2 + \tau\mathcal{C}(\mathbf{v}) \\
\mathbf{x}^t &= \underset{\mathbf{x}}{\operatorname{argmin}} \lambda||\mathbf{b} - \mathbf{A}\mathbf{x}||_2^2 + \rho^t||\mathbf{z}^t - \mathbf{x}||_2^2 + \rho_s^t||\mathbf{v}^t - \mathbf{x}||_2^2
\end{aligned} \tag{16}$$

An example is shown in Fig. 10. The hybrid method shares the advantages of our original model that effectively preserves edges and textures and the cross-channel prior that reduces color artifacts.

Compared with Alg. 1, the hybrid method described in this section requires extra computation of the incorporated prior (i.e. the slack variable $\mathbf{v}^t$ in Eq. 13 and 16) at each HQS iteration and therefore results in additional computation cost and run-time at test. Assuming c be the time cost of computing $\mathbf{v}^t$, the total run-time increase is approximately cT, where T is the number of HQS iterations.

The ability of our method to easily combine a discriminative model with other priors at test time is a key feature of our approach: it allows our method to take advantage of existing and future work on image priors, especially when those priors are difficult to learn from data by training, such as the non-local similarity and low rank priors [5], [7]. This is not possible with previous discriminative methods.

F. Transferability to unseen tasks

Our method allows for new data-fidelity terms that are not contained in training, with no need for re-training. We demonstrate this flexibility with an experiment on the *joint denoising and inpainting* task shown in Fig. 11. To clarify, in this paper we refer "inpainting" to be the problem of pixel interpolation rather than hole-filling. In this experiment, 60% pixels of the input image are missing, and the measured 40% pixels are corrupted with Gaussian noise with $\sigma = 15$. This is a challenging problem to solve, due to the co-existence of missing pixels and strong noise on the measured pixels.

Let vector $\mathbf{a}$ be the binary mask for measured pixels. The sensing matrix $\mathbf{A}$ in Eq. 1, assumed to be known, is a binary diagonal matrix (hence $\mathbf{A} = \mathbf{A}^\mathsf{T} = \mathbf{A}^\mathsf{T}\mathbf{A}$) with diagonal elements $\mathbf{a}$. To reuse our model trained on denoising/deconvolution tasks, we only need to specify $\mathbf{A}$ and λ. The subproblems of our HQS framework are given in Eq. 17.

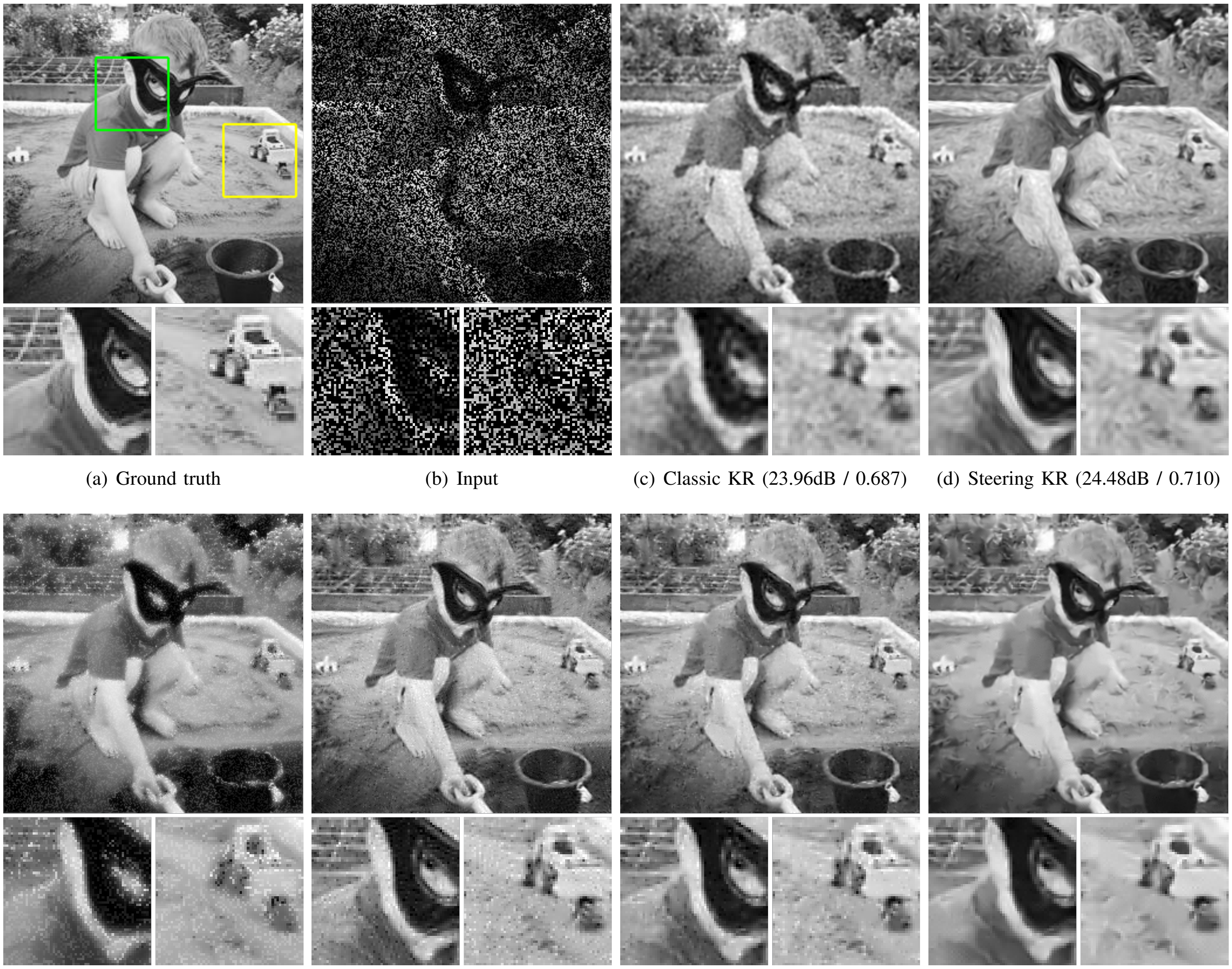
$$\begin{aligned}
\mathbf{z}^t &= \mathbf{prox}_\Theta(\mathbf{x}^{t-1}), \\
\mathbf{x}^t &= (\lambda\mathbf{A}^\mathsf{T}\mathbf{b} + \rho^t\mathbf{z}^t)/(\lambda\mathbf{a} + \rho^t)
\end{aligned} \tag{17}$$

We compare our method with the state-of-the-art methods including kernel regression (KR) [16], NLR-CS [20], GSR [21] and EPLL [4]. We observe that KR, NLR-CS and GSR have limited performance on this joint denoising and inpainting problem, which is more common in practice than the pure (noiseless) inpainting problem. As shown in Fig. 11, DTL outperforms the compared methods in both PSNR and SSIM. Meanwhile, DTL is significantly faster (over two orders of magnitude) than most of the compared methods. The detailed numbers are included in Fig. 11.

As demonstrated above, our method is able to transfer the learned discriminative model to unseen tasks without re-training, which has not been possible with prior discriminative learning methods.

V. Conclusion

In this paper, we proposed a discriminative transfer learning framework for general image restoration. By combining



(a) Ground truth (b) Input (c) Classic KR (23.96dB / 0.687) (d) Steering KR (24.48dB / 0.710)

(e) NLR-CS (20.85dB / 0.626) (f) GSR (24.24dB / 0.667) (g) EPLL (24.89dB / 0.725) (h) DTL$_3^5$ (25.10dB / 0.749)

Fig. 11: Experiment on *joint denoising and inpainting* task. The input image misses 60% pixels, and is corrupted with noise $\sigma = 15$. Inside brackets of certain subcaptions are PSNR (dB) and SSIM values of corresponding result images. DTL outperforms kernel regression (KR) method with either classic or steering kernels, NLR-CS, GSR and EPLL in both PSNR and SSIM. Meanwhile, DTL is significantly faster (over two orders of magnitude) than most of the compared methods. The method run-times are 1.0 second for DTL, 2.0 seconds for classic KR, 2.9 seconds for steering KR, 119.5 seconds for NLR-CS, 551.9 seconds for GSR, and 296.8 seconds for EPLL.

advanced proximal optimization algorithms and discriminative learning techniques, a single training process leads to a transferable model useful for a variety of image restoration tasks and problem conditions. Furthermore, our method is flexible and can be combined with existing priors and likelihood terms after being trained, allowing us to improve image quality on a task at hand. In spite of this generality, our method achieves comparable run-time efficiency as previous discriminative approaches, making it suitable for high-resolution image restoration and mobile vision applications.

In this work we adopt a variant of the TRD model for the prior proximal operator as an example. We believe replacing it with future more expressive discriminative models will further improve the results of our framework. We also believe that our framework that incorporates advanced optimization with discriminative learning techniques can be extended to deep learning, for training more compact and shareable models, and might prove useful for high-level vision problems.

Appendix A
Derivation of analytic gradients

In this section we give the derivation for the computation of several analytic gradients that are required for training.

The HQS iterations ($t = 1, 2, ..., T$) in our method read as follows:

$$\begin{aligned} \mathbf{z}^t &= \mathbf{prox}_{\Theta}(\mathbf{x}^{t-1}) \\ \mathbf{x}^t &= \underbrace{\left(\lambda \mathbf{A}^\top \mathbf{A} + \rho^t\right)^{-1}}_{\mathbf{\Pi}_t} \underbrace{\left(\lambda \mathbf{A}^\top \mathbf{b} + \rho^t \mathbf{z}^t\right)}_{\mathbf{\Lambda}_t} \end{aligned} \tag{18}$$

For both denoising and deconvolution tasks, the $\mathbf{x}^t$-update in Eq. 18 has a closed-form solution, which is given in Eq. 8 and 9. Here, we present a derivation with the more general formula $\mathbf{x}^t = \mathbf{\Pi}_t^{-1}\mathbf{\Lambda}_t$.

In our method, the trainable parameters $\Omega = \{\lambda_p, \Theta\}$ include the fidelity weight λ_p for each problem class p, and the model parameters Θ of the prior proximal operator that is shared across all problem classes. The training loss function ℓ is defined as the negative of the average Peak Signal-to-Noise Ratio (PSNR) between the reconstructed and ground truth images. The gradient of the loss ℓ w.r.t. Θ is computed by averaging the gradients of all images, while the gradient of the loss ℓ w.r.t. λ_p is computed by averaging the gradients of only those images that belong to class p. For convenience, we give the derivations for one image, and omit the class label p below.

$$\ell = -20\log_{10}\left(\frac{255\sqrt{M}}{||\mathbf{x}^T - \mathbf{x}_{\text{true}}||_2}\right), \tag{19}$$

where M is the number of pixels in each image, $\mathbf{x}_{\text{true}}$ is the ground truth image, and $\mathbf{x}^T$ is the reconstructed image.

$$\frac{\partial \ell}{\partial \Omega} == \sum_{t=1}^{T}\left(\frac{\partial \mathbf{x}^t}{\partial \lambda} + \frac{\partial \mathbf{z}^t}{\partial \Theta}\frac{\partial \mathbf{x}^t}{\partial \mathbf{z}^t}\right)\frac{\partial \ell}{\partial \mathbf{x}^t} \tag{20}$$

Next, we provide the derivation for the partial derivative terms in Eq. 20:

$$\begin{aligned}\frac{\partial \mathbf{x}^t}{\partial \lambda} &= \frac{\partial \mathbf{\Lambda}_t}{\partial \lambda}\mathbf{\Pi}_t^{-1} - \left(\mathbf{\Pi}_t^{-1}\mathbf{\Lambda}_t\right)^{\mathsf{T}}\frac{\partial \mathbf{\Pi}_t}{\partial \lambda}\mathbf{\Pi}_t^{-1} \\ &= \left(\mathbf{A}^{\mathsf{T}}\mathbf{b} - \mathbf{A}^{\mathsf{T}}\mathbf{A}\mathbf{x}^t\right)^{\mathsf{T}}\mathbf{\Pi}_t^{-1}\end{aligned} \tag{21}$$

$$\frac{\partial \mathbf{x}^t}{\partial \mathbf{z}^t} = \frac{\partial \mathbf{\Lambda}_t}{\partial \mathbf{z}^t}\mathbf{\Pi}_t^{-1} - \left(\mathbf{\Pi}_t^{-1}\mathbf{\Lambda}_t\right)^{\mathsf{T}}\frac{\partial \mathbf{\Pi}_t}{\partial \mathbf{z}^t}\mathbf{\Pi}_t^{-1} = \rho^t\mathbf{\Pi}_t^{-1} \tag{22}$$

$$\frac{\partial \ell}{\partial \mathbf{x}^{t-1}} = \frac{\partial \mathbf{z}^t}{\partial \mathbf{x}^{t-1}}\frac{\partial \mathbf{x}^t}{\partial \mathbf{z}^t}\frac{\partial \ell}{\partial \mathbf{x}^t} = \rho^t\frac{\partial \mathbf{z}^t}{\partial \mathbf{x}^{t-1}}\mathbf{\Pi}_t^{-1}\frac{\partial \ell}{\partial \mathbf{x}^t} \tag{23}$$

To compute the gradient of $\mathbf{z}^t$ w.r.t. $\mathbf{x}^{t-1}$ and Θ, i.e. $\partial\mathbf{z}^t/\partial\mathbf{x}^{t-1}$ and $\partial\mathbf{z}^t/\partial\Theta$, we notice that in Eq. 6, $\mathbf{x}^{t-1}$ only appears at the first stage $k = 1$:

$$\begin{aligned}\mathbf{z}_1^t &= \mathbf{z}_0^t - \sum_{i=1}^{N}\mathbf{F}_i^{1\,\mathsf{T}}\psi_i^1(\mathbf{F}_i^1\mathbf{z}_0^t) \\ &= \mathbf{x}^{t-1} - \sum_{i=1}^{N}\mathbf{F}_i^{1\,\mathsf{T}}\psi_i^1(\mathbf{F}_i^1\mathbf{x}^{t-1})\end{aligned} \tag{24}$$

Therefore,

$$\begin{aligned}\frac{\partial \mathbf{z}^t}{\partial \mathbf{x}^{t-1}} &= \frac{\partial \mathbf{z}_1^t}{\partial \mathbf{x}^{t-1}}\frac{\partial \mathbf{z}_K^t}{\partial \mathbf{z}_1^t} \\ &= \left(\mathbf{I} - \sum_{i=1}^{N}\mathbf{F}_i^{1\,\mathsf{T}}{\psi'}_i^1(\mathbf{F}_i^1\mathbf{x}^{t-1})\mathbf{F}_i^1\right)\frac{\partial \mathbf{z}_K^t}{\partial \mathbf{z}_1^t},\end{aligned} \tag{25}$$

where $\mathbf{I}$ is an identity matrix, and $\partial\mathbf{z}_K^t/\partial\mathbf{z}_1^t$ can be computed by following the rule:

$$\frac{\partial \mathbf{z}_k^t}{\partial \mathbf{z}_{k-1}^t} = \mathbf{I} - \sum_{i=1}^{N}\mathbf{F}_i^{k\,\mathsf{T}}{\psi'}_i^k(\mathbf{F}_i^k\mathbf{z}_{k-1}^t)\mathbf{F}_i^k \tag{26}$$

$\partial\mathbf{z}^t/\partial\Theta$ in Eq. 20 is composed of $\partial\mathbf{z}^t/\partial\mathbf{f}_i^k$ and $\partial\mathbf{z}^t/\partial\psi_i^k$, which are computed as follows:

$$\frac{\partial \mathbf{z}^t}{\partial \mathbf{f}_i^k} = \frac{\partial \mathbf{z}_k^t}{\partial \mathbf{f}_i^k}\frac{\partial \mathbf{z}_K^t}{\partial \mathbf{z}_k^t} = -\frac{\partial \mathbf{F}_i^{k\,\mathsf{T}}\psi_i^k(\mathbf{F}_i^k\mathbf{z}_{k-1}^t)}{\partial \mathbf{f}_i^k}\frac{\partial \mathbf{z}_K^t}{\partial \mathbf{z}_k^t} \tag{27}$$

$$\frac{\partial \mathbf{z}^t}{\partial \psi_i^k} = \frac{\partial \mathbf{z}_k^t}{\partial \psi_i^k}\frac{\partial \mathbf{z}_K^t}{\partial \mathbf{z}_k^t} = -\frac{\partial \mathbf{F}_i^{k\,\mathsf{T}}\psi_i^k(\mathbf{F}_i^k\mathbf{z}_{k-1}^t)}{\partial \psi_i^k}\frac{\partial \mathbf{z}_K^t}{\partial \mathbf{z}_k^t} \tag{28}$$

References

[1] L. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[2] D. Krishnan and R. Fergus, “Fast image deconvolution using hyper-laplacian priors,” in *Conference on Neural Information Processing Systems (NIPS)*, 2009, pp. 1033–1041.

[3] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[4] D. Zoran and Y. Weiss, “From learning models of natural image patches to whole image restoration,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 479–486.

[5] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, “Image denoising by sparse 3-D transform-domain collaborative filtering,” *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[6] W. Dong, L. Zhang, G. Shi, and X. Li, “Nonlocally centralized sparse representation for image restoration,” *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1620–1630, 2013.

[7] S. Gu, L. Zhang, W. Zuo, and X. Feng, “Weighted nuclear norm minimization with application to image denoising,” in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 2862–2869.

[8] U. Schmidt and S. Roth, “Shrinkage fields for effective image restoration,” in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 2774–2781.

[9] Y. Chen, W. Yu, and T. Pock, “On learning optimized reaction diffusion processes for effective image restoration,” in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2015.

[10] D. Geman and C. Yang, “Nonlinear image recovery with half-quadratic regularization,” *IEEE Transactions on Image Processing*, vol. 4, no. 7, pp. 932–946, 1995.

[11] N. Parikh and S. Boyd, “Proximal algorithms,” *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2013.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[13] P. Milanfar, “A tour of modern image filtering: New insights and methods, both practical and theoretical,” *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 106–128, 2013.

[14] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” in *Computer Vision (ICCV), 1998 IEEE International Conference on*, 1998, pp. 839–846.

[15] J. Weickert, *Anisotropic diffusion in image processing*. ECMI Series, Teubner-Verlag, Stuttgart, Germany, 1998.

[16] H. Takeda, S. Farsiu, and P. Milanfar, “Kernel regression for image processing and reconstruction,” *IEEE Transactions on image processing*, vol. 16, no. 2, pp. 349–366, 2007.

[17] B. C. A. Buades and J. M. Morel, “A review of image denoising algorithms, with a new one,” *Multiscale Modeling and Simulation*, vol. 4, no. 2, pp. 490–530, 2005.

[18] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, “Non-local sparse models for image restoration,” in *Computer Vision (ICCV), 2009 IEEE International Conference on*. IEEE, 2009, pp. 2272–2279.

[19] H. Talebi and P. Milanfar, “Global image denoising,” *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 755–768, 2014.

[20] W. Dong, G. Shi, X. Li, Y. Ma, and F. Huang, “Compressive sensing via nonlocal low-rank regularization,” *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3618–3632, 2014.

[21] J. Zhang, D. Zhao, and W. Gao, “Group-based sparse representation for image restoration,” *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3336–3351, 2014.

[22] Z. Wang and A. C. Bovik, “Mean squared error: Love it or leave it? a new look at signal fidelity measures,” *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[23] A. Rehman and Z. Wang, “Ssim-based non-local means image denoising,” in *Image Processing (ICIP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 217–220.

[24] V. Jakhetiya, W. Lin, S. P. Jaiswal, S. C. Guntuku, and O. C. Au, “Maximum a posterior and perceptually motivated reconstruction algorithm: a generic framework,” *IEEE Transactions on Multimedia*, vol. 19, no. 1, pp. 93–106, 2017.

[25] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, “Image and depth from a conventional camera with a coded aperture,” *ACM transactions on graphics (TOG)*, vol. 26, no. 3, p. 70, 2007.

[26] D. Krishnan, T. Tay, and R. Fergus, “Blind deconvolution using a normalized sparsity measure,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 233–240.

[27] H. Bristow, A. Eriksson, and S. Lucey, “Fast convolutional sparse coding,” in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 391–398.

[28] B. Wohlberg, “Efficient convolutional sparse coding,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7173–7177.

[29] F. Heide, W. Heidrich, and G. Wetzstein, “Fast and flexible convolutional sparse coding,” in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2015, pp. 5135–5143.

[30] S. Roth and M. Black, “Fields of experts,” *International Journal of Computer Vision*, vol. 82, no. 2, pp. 205–229, 2009.

[31] S. Sreehari, S. Venkatakrishnan, B. Wohlberg, G. T. Buzzard, L. F. Drummy, J. P. Simmons, and C. A. Bouman, “Plug-and-play priors for bright field electron tomography and sparse interpolation,” *IEEE Transactions on Computational Imaging*, vol. 2, no. 4, pp. 408–423, 2016.

[32] A. Brifman, Y. Romano, and M. Elad, “Turning a denoiser into a super-resolver using plug and play priors,” in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1404–1408.

[33] Y. Romano, M. Elad, and P. Milanfar, “The little engine that could: Regularization by denoising (red),” *arXiv preprint arXiv:1611.02862*, 2016.

[34] S. A. Bigdeli, M. Zwicker, P. Favaro, and M. Jin, “Deep mean-shift priors for image restoration,” in *Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 763–772.

[35] J. Sun and M. F. Tappen, “Separable markov random field model and its applications in low level vision,” *IEEE transactions on image processing*, vol. 22, no. 1, pp. 402–407, 2013.

[36] J. Jancsary, S. Nowozin, T. Sharp, and C. Rother, “Regression tree fields - an efficient, non-parametric approach to image labeling problems,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012.

[37] K. Schelten, S. Nowozin, J. Jancsary, C. Rother, and S. Roth, “Interleaved regression tree field cascades for blind image deconvolution,” in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*. IEEE, 2015, pp. 494–501.

[38] L. Xiao, J. Wang, W. Heidrich, and M. Hirsch, “Learning high-order filters for efficient blind deconvolution of document photographs,” in *Computer Vision (ECCV), 2016 European Conference on*. Springer, 2016, pp. 734–749.

[39] T. Klatzer, K. Hammernik, P. Knobelreiter, and T. Pock, “Learning joint demosaicing and denoising based on sequential energy minimization,” in *Computational Photography (ICCP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–11.

[40] H. C. Burger, C. J. Schuler, and S. Harmeling, “Image denoising: Can plain neural networks compete with BM3D?” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2392–2399.

[41] V. Jain and H. Seung, “Natural image denoising with convolutional networks,” in *Conference on Neural Information Processing Systems (NIPS)*, vol. 21, 2009, pp. 769–776.

[42] L. Xu, J. S. Ren, C. Liu, and J. Jia, “Deep convolutional neural network for image deconvolution,” in *Conference on Neural Information Processing Systems (NIPS)*, 2014, pp. 1790–1798.

[43] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand, “Deep joint demosaicking and denoising,” *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, p. 191, 2016.

[44] S. Wang, S. Fidler, and R. Urtasun, “Proximal deep structured models,” in *Conference on Neural Information Processing Systems (NIPS)*, 2016, pp. 865–873.

[45] D. Rosenbaum and Y. Weiss, “The return of the gating network: Combining generative models and discriminative training in natural image priors,” in *Conference on Neural Information Processing Systems (NIPS)*, 2015, pp. 2683–2691.

[46] E. J. Candès and M. B. Wakin, “An introduction to compressive sampling,” *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[47] A. Chambolle and T. Pock, “A first-order primal-dual algorithm for convex problems with applications to imaging,” *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[48] M. Schmidt, “minfunc: unconstrained differentiable multivariate optimization in matlab,” http://www.cs.ubc.ca/s̃chmidtm/Software/minFunc.html.

[49] Y. Chen, T. Pock, R. Ranftl, and H. Bischof, “Revisiting loss-specific training of filter-based mrfs for image restoration,” in *German Conference on Pattern Recognition 2013*. Springer, 2013, pp. 271–281.

[50] A. Barbu, “Training an active random field for real-time image denoising,” *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2451–2462, 2009.

[51] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, “Efficient marginal likelihood optimization in blind deconvolution,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 2657–2664.

[52] U. Schmidt, C. Rother, S. Nowozin, J. Jancsary, and S. Roth, “Discriminative non-blind deblurring,” in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 604–611.

[53] J. Ragan-Kelley, C. Barnes, A. Adams, S. Paris, F. Durand, and S. Amarasinghe, “Halide: a language and compiler for optimizing parallelism, locality, and recomputation in image processing pipelines,” *ACM SIGPLAN Notices*, vol. 48, no. 6, pp. 519–530, 2013.

[54] F. Heide, M. Steinberger, Y.-T. Tsai, M. Rouf, D. Pajak, D. Reddy, O. Gallo, J. Liu, W. Heidrich, K. Egiazarian *et al.*, “Flexisp: a flexible camera image processing framework,” *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, p. 231, 2014.

Lei Xiao is a Research Scientist at Oculus Research, a division of Oculus VR, LLC.. He received the Ph.D. degree in Computer Science from University of British Columbia in 2017, under the supervision of Dr. Wolfgang Heidrich. Before that, he received the B.S. degree in Biomedical Engineering from Huazhong University of Science and Technology in 2009, and the M.S. degree in Computer Engineering from University of New Mexico in 2012. His research focuses on computational imaging, display and machine learning.

Felix Heide is a Postdoctoral Researcher at Stanford University and the co-founder and CTO of Algolux. He is interested in the theory and application of computational imaging and vision systems. Researching imaging systems end-to-end, his work lies at the intersection of optics, machine learning, optimization, computer graphics and computer vision. He has co-authored over 25 publications and filed 6 patents. He received his Ph.D. in 2016 at the University of British Columbia under the advisement of Dr. Wolfgang Heidrich. His doctoral dissertation won the Alain Fournier Ph.D. Dissertation Award and the SIGGRAPH outstanding doctoral dissertation award.

Wolfgang Heidrich is a Professor of Computer Science and the Director of the Visual Computing Center at King Abdullah University of Science and Technology. He received his PhD in Computer Science from the University of Erlangen in 1999, and then worked as a Research Associate in the Computer Graphics Group of the Max Planck Institute for Computer Science in Saarbrucken, Germany, before joining the faculty of the University of British Columbia in 2000, initially as a Assistant, then Associate and Full Professor, and finally Dolby Research Chair. In 2014, he joined King Abdullah University of Science and Technology while continuing to affiliated with University of British Columbia until 2018. His research interests lie at the intersection of computer graphics, computer vision, imaging, and optics. In particular, he has worked on computational imaging and displays, high dynamic range imaging and display, image-based modeling, measuring, and rendering, geometry acquisition, GPU-based rendering, and global illumination. He has written well over 200 refereed publications on these subjects and has served on numerous program committees. His work on High Dynamic Range Displays served as the basis for the technology behind Brightside Technologies, which was acquired by Dolby in 2007. In 2016, he was the papers chair for both SIGGRAPH ASIA and ICCP. He is the recipient of a 2014 Humboldt Research Award.

Bernhard Schölkopf is a Director at the Max Planck Institute for Intelligent Systems in Tübingen, Germany, where he heads the Department of Empirical Inference. His scientific interests are in machine learning and causal inference. He has applied his methods to a number of different fields, ranging from biomedical problems to computational photography and astronomy. He worked at AT&T Bell Labs, at GMD FIRST, Berlin, and at Microsoft Research Cambridge, UK, before becoming a Max Planck director in 2001. He is a member of the German Academy of Sciences (Leopoldina), has won the Royal Society Milner Award and the Leibniz Prize, and is an Amazon Distinguished Scholar. He co-founded the series of Machine Learning Summer Schools, and serves as co-editor-in-chief for the Journal of Machine Learning Research, an early development in open access and today the field's flagship journal.

Michael Hirsch is an Affiliated Researcher at the Max Planck Institute for Intelligent Systems and is leading a research group on computational imaging in the department of Empirical Inference. His research interests cover a wide range of signal and image processing problems in scientific imaging as well as computational photography. He studied physics and mathematics at the University of Erlangen and at Imperial College London. He received a Diploma in theoretical physics in 2007, before joining the Max Planck Institute for Biological Cybernetics. After his doctoral studies he worked as a post-doctoral researcher at University College London from 2011 to 2014.