




Variations on the Convolutional Sparse Coding Model

Ives Rey-Otero , Jeremias Sulam , *Member, IEEE*, and Michael Elad , *Fellow, IEEE*

Abstract—Over the past decade, the celebrated sparse representation model has achieved impressive results in various signal and image processing tasks. A convolutional version of this model, termed convolutional sparse coding (CSC), has been recently reintroduced and extensively studied. CSC brings a natural remedy to the limitation of typical sparse enforcing approaches of handling global and high-dimensional signals by local, patch-based, processing. While the classic field of sparse representations has been able to cater for the diverse challenges of different signal processing tasks by considering a wide range of problem formulations, almost all available algorithms that deploy the CSC model consider the same $\ell_1 - \ell_2$ problem form. As we argue in this paper, this CSC pursuit formulation is also too restrictive as it fails to explicitly exploit some local characteristics of the signal. This work expands the range of formulations for the CSC model by proposing two convex alternatives that merge global norms with local penalties and constraints. The main contribution of this work is the derivation of efficient and provably converging algorithms to solve these new sparse coding formulations.

Index Terms—Sparse representation, convolutional sparse coding, parallel proximal algorithm, convex optimization.

I. INTRODUCTION

THE sparse representation model [1] is a central tool for a wide range of inverse problems in image processing, such as denoising [2], [3], super-resolution [4], [5], image deblurring [6], [7] and more. This model assumes that natural signals can be represented as a sparse linear combination of a few columns, called atoms, taken from a matrix called dictionary. The problem of recovering the sparse decomposition of a given signal over a (typically overcomplete) dictionary is called *sparse coding* or pursuit. Such an inverse problem is usually formulated as an optimization objective seeking to minimize the ℓ_0 pseudo-norm, or its convex relaxation, the ℓ_1 -norm, while allowing for a good¹ signal reconstruction. An effective deployment of the sparse representation model calls for the identification of a dictionary that suites the data treated. This is known as the

dictionary learning problem, of finding the best sparsifying dictionary that fits a large set of signal examples [8], [9].

Alas, when it comes to the need to process global high-dimensional signals (e.g., complete images), the sparse representation model hits strong barriers. Dictionary learning is completely intractable in such cases due to its too high memory and computational requirements. In addition, the global pursuit fails to grasp local varying behaviors in the signal, thus leading to inferior treatment of the overall data. Because of these reasons, it has become a common practice to split the global signal into small overlapping blocks, or patches, identify the dictionary that best models these patches, and then sparse code and reconstruct each of these blocks independently before averaging them back into a global signal [2]. Although practical and effective [10], this patch-based strategy is inherently limited since it does not account for the natural dependencies that exist between adjacent or overlapping patches, and therefore it cannot ensure a coherent reconstruction of the global signal [11], [12].

This limitation of the patch-based strategy has been tackled in two ways. One way maintains the patch-based strategy while extending it by modifying the objective so as to bridge the gap between local prior and global reconstruction. This is achieved either by taking into account the self-similarities of natural images [3], [7], by exploiting their multi-scale nature [12]–[14], or by explicitly requiring the reconstructed global signal to be consistent with the local prior [11], [15]. The second way consists in dropping the heuristic patch-based strategy altogether in favor of global, yet computationally tractable and locally-aware, models. Such is the case of the CSC [16]–[18], allowing the pursuit to be performed directly on the global signal by imposing a specific banded convolutional structure on the global dictionary. This implies, naturally, that the signal of interest is a superposition of a few local atoms shifted to different positions. And so, while the CSC is a global model, it has patch-based flavor to it and in addition, learning its dictionary is within reach [19].

Recent years have seen a renewed interest in the CSC model, including a thorough theoretical analysis along with new pursuit and dictionary learning algorithms for it, and its deployment to problems such as image inpainting, super-resolution, dynamic range imaging, and pattern classification [19]–[26]. Nevertheless, the research activity on the CSC model is still in its infancy. In particular, while the classic sparse representation model has assembled an extensive toolbox of problem formulations, diverse sparsity promoting penalty functions along with countless pursuit algorithms (with greedy, relaxation and Bayesian alternatives), most pursuit approaches to recover the CSC representation Γ from a global signal X and a convolutional dictionary D rely on minimizing the same $\ell_2 - \ell_1$ objective,

Manuscript received September 24, 2018; revised August 22, 2019 and November 19, 2019; accepted December 11, 2019. Date of publication January 6, 2020; date of current version January 21, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Qingjiang Shi. The research leading to these results has received funding in part by the European Research Council under EUs 7th Framework Program, ERC under Grant 320649, and in part by Israel Science Foundation (ISF) under Grant 1770/14. (*Corresponding author: Ives Rey-Otero.*)

The authors are with the Computer Science Department, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: ives.rey.otero@gmail.com; jsulam@cs.technion.ac.il; elad@cs.technion.ac.il).

Digital Object Identifier 10.1109/TSP.2020.2964239

¹The desired representation accuracy, or fitting, is problem dependent and it varies for different applications.

namely

$$\underset{\Gamma}{\text{minimize}} \quad \frac{1}{2} \|X - D\Gamma\|_2^2 + \lambda \|\Gamma\|_1, \quad (1)$$

where λ is a Lagrangian parameter. This problem formulation is too restrictive and dull. Indeed, both terms in this formulation, the ℓ_2 reconstruction term and the ℓ_1 sparsity promoting penalty, are global quantities - as is the scalar Lagrangian parameter λ that controls the trade-off between them. This contrasts with state-of-the-art patch-based methods where sparsity is controlled locally, typically through a per-patch constraint on the maximum number of non-zeros or on the *maximal allowed patch error* [2]. This calls for alternative problem formulations where local sparsity and local representation errors are explicitly taken into account in the global model.

An additional motivation for an alternative formulation of the CSC pursuit stems from the findings of [27], which is the first work to derive a theoretical analysis framework for the CSC model. In order to leverage the convolutional structure in this pursuit problem, the authors in [27] advocate for a new notion of local sparsity. In particular, they provide recovery and stability guarantees conditioned on the sparsity of each representation portion responsible for encoding individual patches, as opposed to the traditional global ℓ_0 norm. The CSC pursuit formulations proposed in the present work aim at explicitly controlling the sparsity level in these portions of the representation vectors, called *stripes*. The first formulation employs the $\ell_{1,\infty}$ norm as the sparsity promoting function, providing a convex relaxation of the $\ell_{0,\infty}$ pseudo-norm that was introduced in [27] and explored further in [28], [29]. The second formulation controls the sparsity of the stripes by considering the maximum reconstruction error on each patch simultaneously, via an $\ell_{2,\infty}$ norm. Such an approach is motivated by patch averaging techniques that have been successfully deployed for denoising and other inverse problems [2], [10]. We derive, for each of these two formulations, simple, efficient, and provably converging algorithms.

The remainder of the paper is organized as follows. Section II introduces notations and definitions for the CSC model that we use throughout the paper. The two proposed alternate formulations, the $\ell_2 - \ell_{1,\infty}$ and $\ell_{2,\infty} - \ell_1$, are discussed in Section III and Section IV respectively, along with derivations of algorithms to solve them. Section V illustrates their behavior and performance in a series of experiments. Section VI contains a final discussion.

II. CONVOLUTIONAL SPARSE CODING

Throughout the paper, an image of size $H \times W$ is represented in its vectorized form as a vector X of length $N = HW$. Similarly, image patches of size $n \times n$ are represented in vectorized form as vectors of length n^2 . We denote R_i , the patch extraction operator that extracts from the vectorized image, the image patch at the i -th position.² Naturally, R_i^T denotes the operator that positions, within the vectorized image, a n^2 -long vectorized

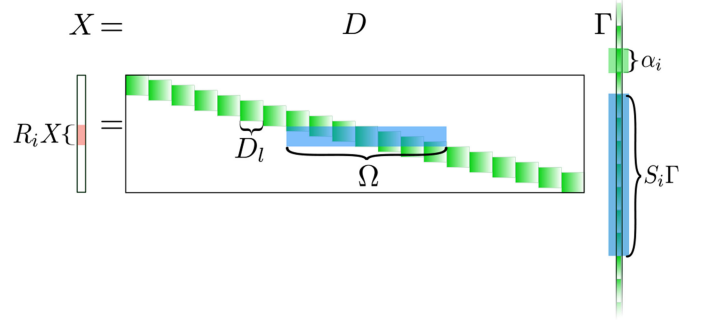


Fig. 1. Illustration of the CSC model for the 1D case. At the global scale, the image X can be decomposed into the product of the global convolutional dictionary D and a global sparse representation Γ . At the patch scale, the patch $R_i X$ can be decomposed into the product of the stripe dictionary Ω and the stripe representation vector $S_i \Gamma$.

patch in the i -th position and pads the rest of the entries with zeroes.

The CSC model assumes that X can be decomposed as $X = D\Gamma$, with D denoting the global convolutional dictionary of size $N \times Nm$, and Γ denoting the corresponding global sparse representation vector of length Nm . The global convolutional dictionary D is built as the concatenation of m (block-) circulant matrices of size $N \times N$, each representing one convolution. These convolutions employ small support filters of size $n \times n$, thus causing the above-mentioned circulant matrices to be narrowly banded. Another way to describe D is by combining all the shifted versions of a local dictionary $D_l \in \mathbb{R}^{n^2 \times m}$ composed of the m vectorized 2D filters. Such construction is best illustrated by expressing the global signal in terms of the local dictionary, $X = \sum_{i=1}^N R_i^T D_l \alpha_i$. In this expression, the quantity $D_l \alpha_i$ is called a slice, with α_i being the portion of the sparse representation vector Γ , called *needle*, that encodes the slice [27]. It is important to stress that slices are not patches but rather simpler components that are combined to form patches.

To better understand which parts of the dictionary D and of the sparse vector Γ represent an isolated patch, it is convenient to consider the patch extraction operator R_i and apply it to the system of equations $X = D\Gamma$. This yields the system $R_i X = R_i D\Gamma$ consisting of the n^2 rows relating to the patch pixels. Due to the banded structure of D , the extracted rows $R_i D$ contain only a subset of $(2n-1)^2 m$ columns that are not trivially zeros. Denoting by S_i^T the operator that extracts such columns and rewriting our system of equations as $R_i X = R_i D S_i^T S_i \Gamma$ make two interesting entities come to light. The first is the vector $S_i \Gamma$, a subset of $(2n-1)^2 m$ coefficients of Γ called the *stripe* that entirely encodes the patch $R_i X$. The second entity is the sub-matrix $\Omega = R_i D S_i^T \in \mathbb{R}^{n^2 \times (2n-1)^2 m}$, called the *stripe dictionary*, which multiplies the stripe vector $S_i \Gamma$ to reconstruct the patch. These two entities were first defined and discussed in [27]. The notations and definitions employed in the remainder of the paper are illustrated in Figure 1 and summarized in Table I.

For the CSC model in its most common formulation, the $\ell_2 - \ell_1$, a variety of algorithms have been proposed [20], [22], [30]–[34]. All of them use the ADMM framework [35] as their workhorse to solve Problem (1) but differ in the subproblems

²By assuming that the image is extended beyond its borders via periodization, the number of $n \times n$ patches that can be extracted from the image equals N , its total number of pixels.

TABLE I
SUMMARY OF NOTATIONS

X, Y	: image in vectorized form of length $N = HW$.
D	: global convolutional dictionary of size $N \times mN$.
Γ	: global sparse vector of length mN .
D_l	: local dictionary of size $n^2 \times m$.
Γ_i	: stripe for position i .
Ω	: stripe dictionary of size $n^2 \times (2n-1)^2 m$.
R_i	: patch extraction operator of size $n^2 \times N$.
S_i	: slice extraction operator of size $(2n-1)^2 m \times mN$.

in which they decompose it into. See [36] for a comparative review.

III. THE $\ell_2 - \ell_{1,\infty}$ CSC FORMULATION

The first alternate formulation that we explore drops the global ℓ_1 as a sparsity promoting penalty and uses instead a mixed norm function, adding an explicit and local control of sparsity. This is motivated by the work in [27], whose analysis centers around a new notion of local sparsity, the $\ell_{0,\infty}$. This measure, instead of quantifying the total number of non-zeros in a vector, reports the ℓ_0 norm of the *densest* stripe:

$$\|\Gamma\|_{0,\infty} = \max_i \|S_i \Gamma\|_0. \quad (2)$$

Such a localized norm is a somewhat more appropriate measure of sparsity in the convolutional setting, since with it one is able to significantly improve on the theoretical guarantees for the CSC model [27]. Although that work established that the $\ell_2 - \ell_1$ formulation approximates the solution to an $\ell_{0,\infty}$ problem, it also conjectured that further improvement could be achieved by considering a new $\ell_{1,\infty}$ -norm. This norm, defined as $\|\Gamma\|_{1,\infty} = \max_i \|S_i \Gamma\|_1$, will be the center of our current discussion: the $\ell_2 - \ell_{1,\infty}$ formulation,

$$\min_{\Gamma} \frac{1}{2} \|X - D\Gamma\|_2^2 + \lambda \|\Gamma\|_{1,\infty}. \quad (3)$$

The $\ell_{1,\infty}$ is nothing but a mixed norm on the global representation Γ . Mixed-norms have been commonly used in signal processing to promote various types of structure in the sparsity pattern [37]. In the context of the CSC model, using this mixed norm is expected to promote a distribution of non-zero coefficients that makes use of more diverse local atoms and is less affected by the global attributes of the image.

This formulation, in fact, first appeared in [29], which proposed two algorithms to solve Problem (3). The first is a nested ADMM algorithm, in which one of the updates involves a multi-block ADMM solver. Using a multi-block ADMM poses a practical challenge, as it does not enjoy the same convergence guarantees of the standard ADMM and requires delicate parameter tuning [38]. To alleviate this problem, the second algorithm proposed in [29] maps Problem (3) to a non-negative problem. This second algorithm relies on standard ADMM formulation combined with the standard DFT-domain Sherman-Morrison approach [32] and is faster and easier to setup than the first one. We will revisit this alternative in our experimental comparison.

A. The Proposed Algorithm

Recalling the $\ell_2 - \ell_{1,\infty}$ formulation in Equation (3), consider N splitting variables $\{\gamma_i\}_{i=1}^N$, so as to rewrite the problem equivalently as

$$\begin{aligned} & \underset{\Gamma, \{\gamma_i\}}{\text{minimize}} && \frac{1}{2} \|Y - D\Gamma\|_2^2 + \lambda \max_i \|\gamma_i\|_1 \\ & \text{subject to} && \forall i, \gamma_i = S_i \Gamma. \end{aligned} \quad (4)$$

This constrained minimization problem is handled by considering its augmented Lagrangian:

$$\frac{1}{2} \|Y - D\Gamma\|_2^2 + \lambda \max_i \|\gamma_i\|_1 + \frac{\rho}{2} \sum_i \|\gamma_i - S_i \Gamma + u_i\|_2^2, \quad (5)$$

where $\{u_i\}_{i=1}^N$ denote the scaled dual-variables associated with each equality constraint $\gamma_i = S_i \Gamma$. The ADMM algorithm [35] minimizes this augmented Lagrangian by alternatively updating the variable Γ and the set of splitting variables $\{\gamma_i\}_{i=1}^N$. Formally, an iteration of the ADMM algorithm consists of the following steps:

$$\begin{aligned} \Gamma^{(k)} &:= \arg \min_{\Gamma} \frac{1}{2} \|Y - D\Gamma\|_2^2 \\ &+ \frac{\rho}{2} \sum_i \left\| \gamma_i^{(k-1)} - S_i \Gamma + u_i^{(k-1)} \right\|_2^2. \end{aligned} \quad (6)$$

$$\begin{aligned} \{\gamma_i^{(k)}\} &:= \arg \min_{\{\gamma_i\}} \lambda \max_i \|\gamma_i\|_1 \\ &+ \frac{\rho}{2} \sum_i \left\| \gamma_i - S_i \Gamma^{(k)} + u_i^{(k-1)} \right\|_2^2. \end{aligned} \quad (7)$$

$$u_i^{(k)} := u_i^{(k-1)} + \gamma_i^{(k)} - S_i \Gamma^{(k)}. \quad (8)$$

The update of Γ in Equation (6) is straightforward, as it is a least-square minimization that boils down to solving the linear system of equations

$$\begin{aligned} \left(D^T D + \rho \sum_i S_i^T S_i \right) \Gamma &= D^T Y \\ &+ \rho \sum_i S_i^T (\gamma_i + u_i). \end{aligned} \quad (9)$$

Bearing in mind that fast implementations are widely available for the convolution D^T and the transpose convolution D , and using the fact that $\sum_i S_i^T S_i = (2n-1)^2 I$, this *regularized* least-square minimization can be carried out efficiently and reliably via a few iterations of the conjugate gradient method [39].

The updates of the variables $\{\gamma_i\}_{i=1}^N$ in Equation (7) are seemingly more complicated, due to the max operation between the different stripes and the fact that they overlap. To make it more manageable, we cast the Problem (7) in epigraph form as

$$\begin{aligned} & \underset{\{\gamma_i\}, t}{\text{minimize}} && \lambda t + \frac{\rho}{2} \sum_i \left\| \gamma_i - S_i \Gamma^{(k+1)} + u_i^{(k)} \right\|_2^2, \\ & \text{subject to} && \forall i, \|\gamma_i\|_1 \leq t. \end{aligned} \quad (10)$$

Here, the initial problem with variables $\{\gamma_i\}_{i=1}^N$ has just been replaced with an equivalent minimization over variables $\{\gamma_i\}_{i=1}^N$ and t . Note that, for a fixed value of variable t , this new objective in Equation (10) is now separable in the variables $\{\gamma_i\}_{i=1}^N$. More precisely, it can be broken down into N separate minimization problems

$$\begin{aligned} \bar{\gamma}_i(t) &:= \arg \min_{\gamma_i} \left\| \gamma_i - S_i \Gamma^{(k)} + u_i^{(k-1)} \right\|_2^2, \\ \text{subject to } & \|\gamma_i\|_1 \leq t. \end{aligned} \quad (11)$$

Each of these is simply a projection onto the ℓ_1 -ball [40] that can be performed via the shrinkage operator:³

$$\bar{\gamma}_i(t) = S_{\lambda^*} \left(S_i \Gamma^{(k)} - u_i^{(k-1)} \right), \quad (12)$$

where the shrinkage parameter λ^* can be efficiently estimated by sorting the vector's coefficients and computing over them a cumulative sum (see [40] for details).

In this way, solving the initial problem (7) boils down to finding the optimal t leading to the minimum of the objective, namely $\{\gamma_i^{(k)}\}_{i=1}^N = \{\gamma_i(t^*)\}_{i=1}^N$ with

$$t^* := \arg \min_t \left(\lambda t + \sum_i \left\| \bar{\gamma}_i(t) - S_i \Gamma^{(k)} + u_i^{(k-1)} \right\|_2^2 \right). \quad (13)$$

As a sum of an affine function and squared distances to the ℓ_1 ball of radius t , the previous objective is a convex function of t . Indeed, the distance to the ℓ_1 ball is a convex function of the radius t (see Proposition 1 in Appendix A). Leveraging the unimodality of the objective, we can iteratively estimate the location of its minimum via a simple ternary-search, which only requires the evaluation of function values.

This simple algorithm, by not involving an over-sensitive Lagrange multiplier setting, and by enjoying the convergence properties of the standard ADMM is simpler in practice than the first algorithm proposed in [29], namely the nested ADMM method. In practice, it will also be slightly faster than the efficient alternative proposed in [29].

IV. THE $\ell_{2,\infty} - \ell_1$ CSC FORMULATION

We move on to consider our second formulation, of explicitly incorporating a local control on the CSC model. This is inspired by the patch-based strategy for image denoising and other inverse problems. Recall that patch-based sparse denoising methods [2], [10] control the sparsity level on each patch by upper-bounding the patch reconstruction error. We will borrow such an idea, and translate it into the convolutional setting.

For a noisy image Y , patch methods rely on a global objective of the form

$$\begin{aligned} \text{minimize}_{\{\beta_i\}, X} \quad & \frac{\lambda}{2} \|X - Y\|_2^2 + \sum_i \|\beta_i\|_0 \\ \text{subject to} \quad & \forall i, \|D_i \beta_i - R_i X\|_2^2 \leq T, \end{aligned} \quad (14)$$

³ $S_{\lambda}(\mathbf{x})$ denotes the shrinkage operator, formally $S_{\lambda}(\mathbf{x}) = \text{sign}(\mathbf{x}) \odot \max(|\mathbf{x}| - \lambda, 0)$, with \odot denoting the element-wise product.

where β_i is the sparse vector for the patch $R_i X$ and the upper-bound T over the patch reconstruction error is typically set to $C n^2 \sigma_{\text{noise}}^2$, the assumed patch noise level (up to a multiplicative constant). This is typically solved via a block-coordinate descent algorithm, which means first initializing $X = Y$ and seeking the sparsest α_i for each patch via the set of local problems

$$\begin{aligned} \text{minimize}_{\beta_i} \quad & \|\beta_i\|_0 \\ \text{subject to} \quad & \|D_i \beta_i - R_i Y\|_2^2 \leq T, \end{aligned} \quad (15)$$

which yields a reconstruction for each overlapping patch and, in turn, an intermediary global reconstruction $\frac{1}{n^2} \sum_i R_i^T D_i \beta_i$. While state-of-the-art methods typically consider approximate solutions through greedy pursuit algorithms, it is also possible to consider an ℓ_1 relaxation of the same sparse coding problem. We will employ the latter option in order to benefit from the resulting convexity of the problem.

The second stage of the block-coordinate descent algorithm consists in updating the estimate of X , the restored image, by solving the least-square problem in closed form [2] according to:

$$X = \left(\lambda I + \sum_i R_i^T R_i \right)^{-1} \left(\lambda Y + \sum_i R_i^T D_i \beta_i \right), \quad (16)$$

essentially averaging the input signal Y with the patch-averaging estimate $\frac{1}{n^2} \sum_i R_i^T D_i \beta_i$.

In order to bring this classic approach into a convolutional setting, note that the CSC global representation Γ can be decomposed into its constituent *needles*, and so $\sum_i \|\alpha_i\|_1 = \|\Gamma\|_1$. Recalling the definitions and notations in Section II, a patch from the reconstructed image $R_i X$ in the CSC model can be equivalently written as $R_i X = R_i D \Gamma = \Omega S_i \Gamma$. With these elements, the problem in (14) can be naturally transformed into

$$\begin{aligned} \text{minimize}_{\Gamma, X} \quad & \frac{\lambda}{2} \|X - Y\|_2^2 + \|\Gamma\|_1 \\ \text{subject to} \quad & \forall i, \|\Omega S_i \Gamma - R_i X\|_2^2 \leq T. \end{aligned} \quad (17)$$

One might indeed adopt a similar block-coordinate descent strategy for this problem as well. After an initialization of $X = Y$, the first step considers the resulting $\ell_{2,\infty} - \ell_1$ formulation:

$$\begin{aligned} \text{minimize}_{\Gamma} \quad & \|\Gamma\|_1 \\ \text{subject to} \quad & \forall i, \|\Omega S_i \Gamma - R_i Y\|_2^2 \leq T, \end{aligned} \quad (18)$$

where the constraint on patch reconstruction considers the stripe dictionary. Again, the second stage consists in updating the estimate of X by solving the least-square problem

$$X = \left(\lambda I + \sum_i R_i^T R_i \right)^{-1} \left(\lambda Y + \sum_i R_i^T \Omega S_i \Gamma \right). \quad (19)$$

whose solution, since $\sum_i R_i^T \Omega S_i \Gamma = n^2 D \Gamma$ and since $\sum_i R_i^T R_i = n^2 I$, boils down to an average between the input image and the intermediary global reconstruction $D \Gamma$. In this manner, and similarly to the patch-averaging strategy, the trade-off between sparsity and reconstruction is controlled locally via

an upper-bound on the reconstruction error of each individual patch. However, while in the original method each vector β_i encodes one patch in disregard with other patches, now each needle α_i becomes part of various stripes $S_i\Gamma$ and therefore contributes in various patches. In other words, the classic patch-averaging approach performs these pursuit independently, whereas this convolutional counterpart will need to update all needles jointly.

In what follows, we show that this seemingly complex problem can in fact be addressed by using traditional ℓ_1 solvers such as the Fast Iterative Shrinkage-Tresholding Algorithm (FISTA) [41] in conjunction with the Parallel Proximal Algorithm (PPXA).

A. Proposed Algorithm

PPXA is a generic convex optimization algorithm introduced by Combettes and Pesquet [42], [43] that extends the Douglas-Rachford algorithm and aims to minimize an objective of the form

$$\underset{x}{\text{minimize}} \sum_i^N f_i(x), \quad (20)$$

where each f_i is a convex function that admits an easy-to-compute proximal operator [44], [45]. Recall that the proximity operator $\text{prox}_{f_i}(y) : \mathbb{R}^N \rightarrow \mathbb{R}^N$ of f_i is defined by

$$\text{prox}_{f_i}(y) := \arg \min_x f_i(x) + \frac{1}{2} \|x - y\|_2^2. \quad (21)$$

In our context, PPXA offers a way to manage the explicit use of overlapping stripes. Indeed, by encapsulating each inequality constraint into its corresponding indicator function, the objective in Equation (18) can be recast as a sum, namely

$$\underset{\Gamma}{\text{minimize}} \sum_{i=1}^N \left(\frac{1}{N} \|\Gamma\|_1 + \mathcal{I}_{\{\|\Omega S_i \Gamma - R_i Y\|_2^2 \leq T\}} \right), \quad (22)$$

where $\mathcal{I}_{\{\|\Omega S_i \Gamma - R_i Y\|_2^2 \leq T\}}$ denotes the indicator function⁴ on the constraint feasibility set. The successful deployment of the PPXA algorithm for this problem depends on our ability to compute, for each patch, the proximal operator

$$\begin{aligned} \text{prox}_{f_i}(\Gamma) := \arg \min_{\hat{\Gamma}} \quad & \|\hat{\Gamma}\|_1 + \frac{1}{2 N \mu} \|\Gamma - \hat{\Gamma}\|_2^2 \\ & + \mathcal{I}_{\{\|\Omega S_i \hat{\Gamma} - R_i Y\|_2^2 \leq T\}}, \end{aligned} \quad (23)$$

with parameter μ scaling the least-square term. The solution to the above problem is also the solution to a Lagrangian

$$\arg \min_{\hat{\Gamma}} \|\hat{\Gamma}\|_1 + \frac{1}{2 N \mu} \|\Gamma - \hat{\Gamma}\|_2^2 + \lambda_i^* \|R_i(D\hat{\Gamma} - Y)\|_2^2, \quad (24)$$

in which the Lagrange multiplier is set to an optimal value λ_i^* : the *smallest* Lagrange multiplier such that the inequality constraint is satisfied. Observe that, while transitioning from Equation (23) to Equation (24), we moved from Ω to D , in order to pose the algorithm w.r.t. the global dictionary. Fortunately, for a given Lagrangian multiplier λ_i , such objective can be efficiently

minimized by a proximal gradient method such as (ISTA) [46] or its fast version FISTA [41]. Indeed, denoting $g_i(\hat{\Gamma}, \lambda_i) := \frac{1}{2 N \mu} \|\Gamma - \hat{\Gamma}\|_2^2 + \lambda_i \|R_i(D\hat{\Gamma} - Y)\|_2^2$, ISTA and FISTA revolve around the update step

$$\hat{\Gamma}^{(k+1)} = \mathcal{S}_{t_k} \left(\hat{\Gamma}^{(k)} + t_k \frac{\partial g_i}{\partial \hat{\Gamma}}(\hat{\Gamma}^{(k)}, \lambda_i) \right), \quad (25)$$

where t_k denotes the step-size.⁵ The dominant effort here is the evaluation of the gradient of g_i with respect to $\hat{\Gamma}$. This boils down to the computation of convolutions. Running FISTA successively with warm-start initialization allows to estimate the minimizer for different values of λ_i with only few extra iterations. This allows to use a binary-search scheme to estimate the optimal Lagrange multiplier λ_i^* which in turn provides the solution to the proximal operator in Equation (23).

Armed with this procedure to compute the proximal operators, an iteration of the PPXA algorithm boils down to the following steps:

- 1) Compute the proximal operators for each patch

$$\forall i = 1 \dots N, \quad \hat{\Gamma}_i^{(l)} = \text{prox}_{f_i}(\Gamma_i^{(l)}), \quad (26)$$

following the procedure described above. The evaluations can be carried out in parallel.

- 2) Aggregate the solutions

$$\hat{\Gamma}^{(l)} = \frac{1}{N} \sum_i^N \hat{\Gamma}_i^{(l)}. \quad (27)$$

- 3) Update the estimate of Γ along with the auxiliary variables Γ_i

$$\begin{aligned} \forall i, \quad \Gamma_i^{(l+1)} &= \Gamma_i^{(n)} + \rho_l (2\hat{\Gamma}^{(l)} - \Gamma^{(l)} - \hat{\Gamma}_i^{(l)}), \\ \Gamma^{(l+1)} &= \Gamma^{(l)} + \rho_l (\hat{\Gamma}^{(l)} - \Gamma^{(l)}), \end{aligned} \quad (28)$$

where ρ_l denotes the relaxation parameter⁶ on this iteration. The sequence of sparse vector estimates $\Gamma^{(l)}$ is proven to converge to the solution of the $\ell_{2,\infty} - \ell_1$ CSC problem (18) [42]. Note that using FISTA in conjunction with PPXA makes it possible to take full advantage of GPU hardware and high-level libraries for fast convolutions, in contrast with most sparse coding algorithm that operate in the Fourier domain [20], [22].

B. Extension Via Weighted Stripe Dictionary

The method described above for the $\ell_{2,\infty} - \ell_1$ formulation brings an additional level of flexibility by offering a generic way to enforce a wider range of structured sparsity. Indeed, because the proposed method splits the global pursuit into parallel pursuits on each stripe, a specific local structure can be imposed on individual stripes. This can be achieved naturally by simply weighting the columns of the stripe dictionary, so as to

⁵For convergence, the step-size t_k must satisfy $t_k \leq \frac{1}{\lambda_{\max}}$, where λ_{\max} denotes the maximum eigenvalue of ∇g_i which can be approximated efficiently via the power method.

⁶To guaranty convergence, the relaxation parameters (ρ_l) must satisfy $\sum_{l \in \mathbb{N}} \rho_l (2 - \rho_l) = +\infty$.

⁴The indicator function \mathcal{I}_S equals 0 inside the set S and ∞ elsewhere.

relatively promote or penalize the use of certain atoms. Formally this corresponds to

$$\begin{aligned} & \underset{\Gamma}{\text{minimize}} && \|\Gamma\|_1 \\ & \text{subject to} && \forall i, \|\Omega W_i S_i \Gamma - R_i Y\|_2^2 \leq T, \end{aligned} \quad (29)$$

where W_i denotes the weighting diagonal matrix relative to the i -th patch.⁷ In the context of the proposed algorithm, this boils down to an extra weighting within each FISTA iterations.

One particularly interesting application of such strategy consists in combining the CSC and patch-averaging models. Such a combination allows for the benefits of both the global and local models, which respective performances on various tasks are increasingly well understood. From an analysis stand point, being able to examine the entire spectrum separating the CSC model and the patch-averaging approach is highly valuable, as the understand of their precise inter-relation has been of interest to the image processing community [47]. With the proposed method, such combination can be achieved via a mere re-weighting of the columns that amounts to replacing the stripe dictionary with the convex combination

$$\Omega_\theta = (1 - \theta)\Omega + \theta n^2 \bar{D}_l, \quad (30)$$

with $0 \leq \theta \leq 1$ and with \bar{D}_l denoting the local dictionary padded with zero columns. The parameter θ allows to regulate the level of patch aggregation that has been proven to be critical in denoising problems [47]. Setting $\theta = 0$ corresponds to the $\ell_{2,\infty} - \ell_1$ CSC formulation above. By increasing θ , filters which locations are shifted with respect to the patch are increasingly penalized. Setting $\theta = 1$ is synonymous with the patch averaging strategy in which the reconstruction relies exclusively on D_l and none of its shifted atoms. As an illustration, let us local-normalize test image *barbara* and sparse-code it with the resulting problem,

$$\begin{aligned} & \underset{\Gamma}{\text{minimize}} && \|\Gamma\|_1 \\ & \text{subject to} && \forall i, \|\Omega_\theta S_i \Gamma - R_i Y\|_2^2 \leq T, \end{aligned} \quad (31)$$

where parameter θ ranges from 0 ($\ell_{2,\infty} - \ell_1$ CSC) to 1 (patch averaging). Figure 2(a) shows the average representation error $\|\Omega_\theta S_i \Gamma - R_i Y\|_2$ (in blue) and the average Euclidean distance between individual slices and patches $\|n^2 \bar{D}_l S_i \Gamma - R_i Y\|_2$ (in red) as a functions of the parameter θ . Threshold T in (31) is plotted as a green dotted line. In accordance to the inequality constraints in Problem (18), the patch reconstruction error stays below the threshold T irrespective of parameter θ . On the other hand, and as expected, the Euclidean distance between slices and patches is above the threshold T , as it is the combination of overlapping slices, rather than an isolated slice, that approximates the patch. However, as θ increases, the term $\Omega_\theta S_i \Gamma$ in the representation error in Problem (31) is increasingly similar to a slice $n^2 \bar{D}_l \alpha$. This in turn constrains the individual slices to better approximate the corresponding patch on their own.

⁷Note that to be consistent with the global CSC model, the set of matrices $\{W_i\}$ must satisfy the relation $D = \frac{1}{n^2} \sum R_i^T \Omega W_i S_i$.

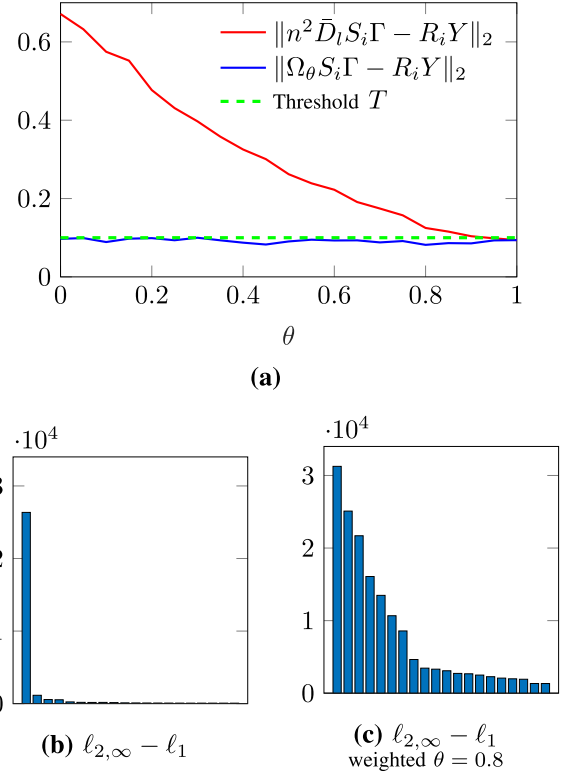


Fig. 2. Effect of replacing the stripe dictionary Ω with the convex combination $\Omega_\theta = (1 - \theta)\Omega + \theta n^2 \bar{D}_l$, with $0 \leq \theta \leq 1$. Test image *barbara* is sparse-coded using formulation (31) for various values of parameter θ . (a) The average reconstruction error $\|\Omega_\theta S_i \Gamma - R_i Y\|_2$ (in blue) and the average Euclidean distance between patches and slices $\|n^2 \bar{D}_l S_i \Gamma - R_i Y\|_2$ (in red) are plotted as functions of θ . Threshold T in (31) is plotted as a green dotted line. In accordance to (31), the reconstruction error remains below T for any θ . As θ increases, individual slices $n^2 \bar{D}_l S_i \Gamma$ become increasingly similar to patches on their own. Weighted stripe dictionary mitigates imbalances in the distribution of used atoms. (b) Number of non-zero coefficients for each of the 20 most commonly used atoms for the non-weighted $\ell_{2,\infty} - \ell_2$ formulation. (c) In contrast, the weighted $\ell_{2,\infty} - \ell_2$ formulation with $\theta = 0.8$ leads to more diverse local atoms being used.

An additional benefit of the weighted extension is that it helps mitigate imbalance in the atom usage distribution, a typical problem affecting the CSC model. Indeed, consider the sparse-coding of test image *barbara* using the non-weighted $\ell_{2,\infty} - \ell_1$ formulation. In Figure 2(b), which depicts how often the first 20 atoms in the local dictionary are used in the solution Γ , shows that one atom is predominantly used. In fact, most of the needles in Γ contain at most just one active atom, and many of them (about 70%) remain completely empty. This behavior is characteristic of the CSC model because, while patch-based approaches rely solely on the local dictionary atoms to encode a patch, the CSC pursuit can rely on the atoms as well as their shifts. In practice, the CSC pursuit tends to use less diverse atoms and favors instead a juxtaposition of the simplest atom shifted at different locations to reconstruct the image. For a CSC based dictionary learning method, this tendency is problematic since an unbalanced selection of atoms during sparse-coding results in one atom being predominantly updated at the expense of all others. The weighted formulation offers a remedy to this problem. Indeed, Figure 2(c) shows the number of non-zero

coefficients for the weighted $\ell_{2,\infty} - \ell_1$ formulation with $\theta = 0.8$. Even though this formulation for $\theta = 0.8$ is consistent with the global CSC model, it leads to more diverse local atoms being used.

V. EXPERIMENTS

To illustrate the behavior and performance of the proposed formulations, we now move to consider two image processing applications: the texture-cartoon separation problem and inpainting.

A. $\ell_2 - \ell_{1,\infty}$ for Texture-Cartoon Separation

We illustrate the $\ell_2 - \ell_{1,\infty}$ formulation on the texture-cartoon separation task. This problem consists in decomposing an input image X into a piecewise smooth component (cartoon) X_c and a texture component X_t such that $X = X_c + X_t$. The typical prior for the cartoon component X_c is based on the total variation norm, denoted $\|X_c\|_{TV}$, which penalizes oscillations. In addition, we propose to assume that the texture component X_t admits a decomposition $X_t = D_t \Gamma$ where D_t is a convolutional texture dictionary and Γ is the solution of the $\ell_2 - \ell_{1,\infty}$ CSC formulation. Under these assumptions, the task of texture and cartoon separation boils down to a minimization problem over three variables: the cartoon component X_c , the CSC representation Γ and a convolutional texture dictionary D_t , namely

$$\underset{\Gamma, D_t, X_c}{\text{minimize}} \quad \frac{1}{2} \|X - D_t \Gamma - X_c\|_2^2 + \lambda \|\Gamma\|_{1,\infty} + \zeta \|X_c\|_{TV}, \quad (32)$$

with parameter ζ controlling the level of TV regularization penalizing oscillations in X_c . Such minimization is carried out iteratively in a block-coordinated manner until convergence. Each iteration consists of the three following steps:

$$X_c^{(k+1)} := \arg \min_{X_c} \frac{1}{2} \|X - D_t^{(k)} \Gamma^{(k)} - X_c\|_2^2 + \zeta \|X_c\|_{TV} \quad (33)$$

$$\Gamma^{(k+1)} := \arg \min_{\Gamma} \frac{1}{2} \|X - D_t^{(k)} \Gamma - X_c^{(k+1)}\|_2^2 + \lambda \|\Gamma\|_{1,\infty} \quad (34)$$

$$D_t^{(k+1)} := \arg \min_{D_t} \frac{1}{2} \|X - D_t \Gamma^{(k+1)} - X_c^{(k+1)}\|_2^2. \quad (35)$$

A TV denoiser.⁸ is used to solve Problem (33) while Problem (34) relies on our $\ell_2 - \ell_{1,\infty}$ solver. For the dictionary update, one option is to use a standard patch-based dictionary learning such as K-SVD using overlapping patches as training sets and the needles of the current Γ estimate. However this would not be consistent with the CSC model. Indeed, the patch would then be assumed to stem from the local dictionary alone, disregarding all the contributions of shifted atoms to its reconstruction. We adopt instead a more coherent alternative that was recently proposed in [28] in which standard dictionary update

⁸The TV denoiser used here is the publicly available implementation of [48].

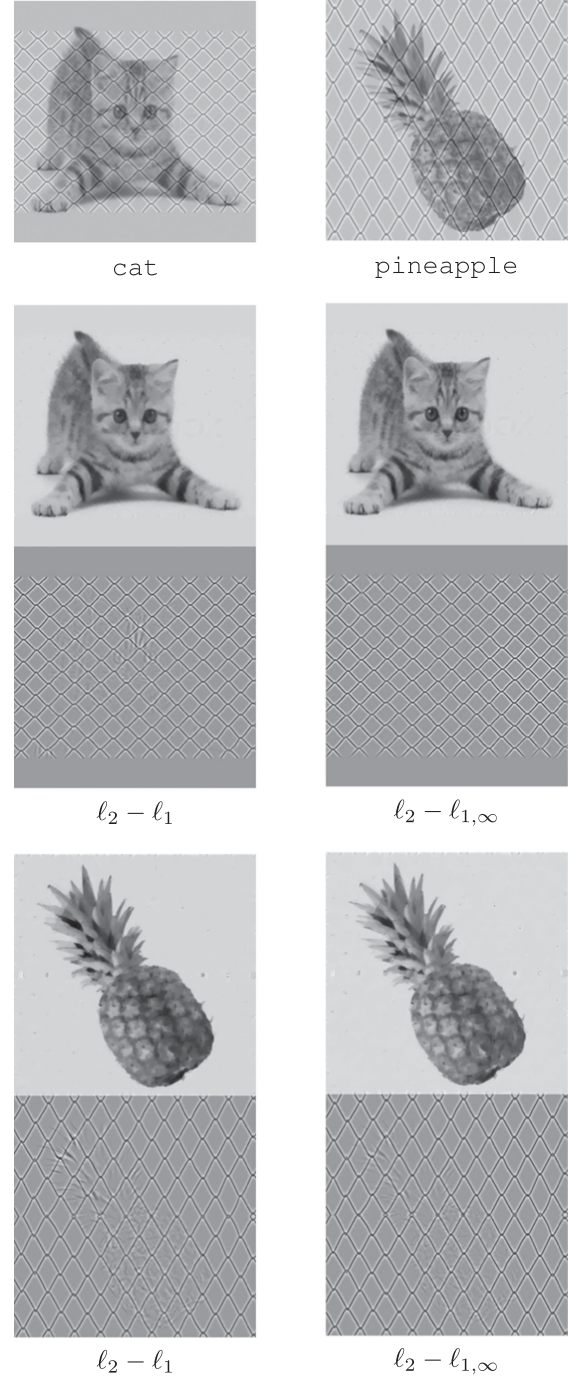


Fig. 3. Noiseless texture-cartoon separation. Comparing the $\ell_2 - \ell_{1,\infty}$ and $\ell_2 - \ell_1$ formulations. The input images consist of the test image cat and pineapple.

procedures are adapted to a convolutional setting and carried out via conjugate gradient descent [39] in conjunction with fast convolution computations. The proposed method is applied to the test images cat and pineapple, the results of our method are shown in Figure 3 along with the results from the $\ell_1 - \ell_2$ based method in [30]. The algorithm relies on GPU/CUDA based implementations for faster convolutions. The computation time for a the sparse coding of a 256×256 is 156 seconds. While it compares favorably to the fastest algorithm proposed in [29]

TABLE II

IMAGE INPAINTING. THE $\ell_2 - \ell_1$ BASED METHOD OF [30] AND [20] ARE COMPARED TO THE PROPOSED METHODS: THE $\ell_{2,\infty} - \ell_1$ FORMULATION AND ITS VARIANT WITH A WEIGHTED STRIPE DICTIONARY, AND THE $\ell_2 - \ell_{1,\infty}$. IN THE FIRST AND SECOND BLOCKS, THE LOCAL DICTIONARY IS PRETRAINED FROM THE `fruit` DATASET USING THE METHOD FROM [30]. METHODS IN THE FIRST BLOCK ARE BASED ON THE $\ell_2 - \ell_1$ CLASSIC FORMULATION WHILE THE SECOND BLOCK CONSIDERS THE ALTERNATIVE FORMULATIONS. THE $\ell_{2,\infty}$ PRIOR IMPROVES OVER THE BEST $\ell_2 - \ell_1$ BASED METHOD FORMULATION. THE WEIGHTED STRIPE DICTIONARY Ω_θ WITH $\theta = 0.8$ BRINGS AN ADDITIONAL IMPROVEMENT IN PSNR OVER THE STANDARD $\ell_{2,\infty}$ BY PROMOTING PATCH AVERAGING. THE $\ell_2 - \ell_{1,\infty}$ VARIANT ON THE OTHER HAND IS OUTPERFORMED BY THE OTHER FORMULATION IN MOST CASES. IN THE RESULT REPORTED IN THE THIRD BLOCK, THE LOCAL DICTIONARY USED IS LEARNED FROM THE CORRUPTED IMAGE. IN THIS SCENARIO, THE WEIGHTED $\ell_{2,\infty} - \ell_1$ FORMULATION WITH $\theta = 0.8$ GENERALLY OUTPERFORMS [30]

	barbara	lena	boat	hill	house	couple	man
Heide <i>et al.</i> [20]	11.00	11.77	10.29	10.37	10.18	11.99	11.60
Papayan <i>et al.</i> [30]	11.67	11.92	10.33	10.66	10.56	12.25	11.84
Wohlberg [32]	11.65	11.56	10.13	10.10	10.67	10.58	11.32
Wang <i>et al.</i> [34]	11.75	11.74	10.29	10.63	10.25	12.03	11.73
$\ell_{2,\infty} - \ell_1$	11.65	11.99	10.39	10.55	10.60	12.34	11.91
Weighted $\ell_{2,\infty} - \ell_1$,	11.78	12.13	10.58	10.65	10.62	12.46	11.98
$\ell_2 - \ell_{1,\infty}$ Ours	10.92	11.23	10.26	10.11	9.91	11.65	11.31
$\ell_2 - \ell_{1,\infty}$ Wohlberg [29]	10.73	10.16	10.39	9.87	9.99	11.34	11.16
Papayan <i>et al.</i> [30], image specific D_l	15.20	12.35	11.60	10.90	11.70	12.41	11.71
weighted $\ell_{2,\infty} - \ell_1$, image specific D_l	16.11	12.29	11.93	11.22	12.13	13.16	12.05

(533 s), it is nevertheless slower than methods for the $\ell_1 - \ell_2$ formulation (7.6 s for [30]).

B. Inpainting

We illustrate the behavior of the proposed variants on the classic problem of image inpainting. Let us consider an image X and a diagonal binary matrix M , which masks the entries in X in which $M_{i,i} = 0$. Image inpainting is the process of filling in missing areas in an image in a realistic manner. That is, given the corrupted image $Y = MX$, the task consists in estimating the original signal X .

Estimating the original signal via the $\ell_{2,\infty} - \ell_1$ CSC requires solving the problem

$$\begin{aligned} & \underset{\Gamma}{\text{minimize}} \quad \|\Gamma\|_1 \\ & \text{subject to} \quad \forall i, \|R_i(MD\Gamma - Y)\|_2^2 \leq T_i, \end{aligned} \quad (36)$$

where the constraint on the representation accuracy incorporates the binary matrix M , and where the threshold T_i is set on a patch-by-patch basis to reflect the varying numbers of active pixels in each patch. Minimizing this objective requires only a slight modification of the algorithm described above, namely incorporating the mask into the function g_i and its gradient. The PPXA relaxation parameter is set to $\lambda_l = 1.6$ and the scaling factor in the proximal operator is set to $\mu = 100$. The minimization was performed with the weighted formulation introduced in Section IV with 10 values of the blending parameter θ ranging from 0 to 1. Similarly, estimating the original signal via the $\ell_2 - \ell_{1,\infty}$ formulation requires solving the problem

$$\underset{\Gamma}{\text{minimize}} \quad \frac{1}{2} \|M(Y - D\Gamma)\|_2^2 + \lambda \|\Gamma\|_{1,\infty}, \quad (37)$$

which in practice only requires adapting the least-square minimization stage for the update of Γ in (6).

We follow the experimental setting in [20]. In particular, input images are mean-subtracted and contrast normalized, the mask M is set to discard 50% of the pixel values. The formulations proposed in this work are compared to four

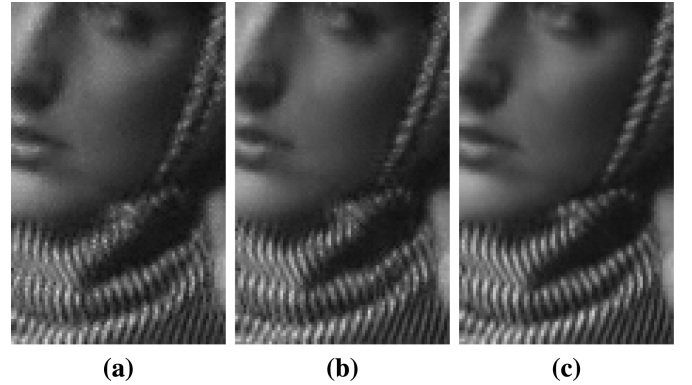


Fig. 4. Visual comparison on a cropped region extracted from inpainting estimations for test image barbara. The input image is mean-subtracted, contrast normalized, and 50% of its pixels are discarded. (a) $\ell_2 - \ell_{1,\infty}$, PSNR = 10.92. (b) $\ell_{2,\infty} - \ell_1$, PSNR = 11.65. (c) weighted $\ell_{2,\infty} - \ell_1$ with $\theta = 0.8$, PSNR = 11.78.

existing convex relaxation-based algorithms: three methods operating in the DFT-domain [20], [32], [34] and the slice-based approach of [30].

Table II contains the peak signal-to-noise ratio (PSNR) on a set of publicly available standard test images. In the first two blocks of experiments, the local dictionary is pretrained from the `fruit` dataset, using the method from [30]. The method based on the $\ell_{2,\infty} - \ell_1$ formulation outperforms the method proposed in [20] and slightly improves over the slice-based approach of [30] and the scalable online convolutional sparse coding of [34]. The best performance are obtained in general with the weighted $\ell_{2,\infty} - \ell_1$ with $\theta = 0.8$, which formulation tends to promote an averaging of similar local estimates. The $\ell_2 - \ell_{1,\infty}$ formulation does not in general lead to improved results for inpainting, not any more that the algorithm proposed in [29] for the same formulation. Figure 4 shows crops of inpainted results for test image barbara for the proposed formulation.

Significant additional improvements are achieved when learning the local dictionary D_l from the corrupted image. The third block in Table II contains the inpainting PSNR obtained in this scenario for the slice-based method [30] and for the weighted

$\ell_{2,\infty} - \ell_1$ used along the dictionary update proposed in [28]. In this context, the weighting of the stripe dictionary is particularly beneficial as it encourages more atoms to be used and therefore updated. The alternative formulations come however at a cost in terms of speed, with the execution times averaging 103 seconds and 124 seconds for the $\ell_2 - \ell_{1,\infty}$ and $\ell_{2,\infty} - \ell_1$ formulations respectively, compared to 12 seconds on average for the slice-based algorithm [30].

VI. CONCLUSION

While enjoying a renewed interest in recent years, the CSC model has been almost exclusively considered in its $\ell_2 - \ell_1$ formulation. In the present work, we expanded the formulations for the CSC with two alternative formulations, namely the $\ell_2 - \ell_{1,\infty}$ and $\ell_{2,\infty} - \ell_1$ formulations in which mixed-norms, alter how the spatial distributions of non-zero coefficients are controlled. For both formulations, we derived algorithms that rely on the ADMM and PPXA algorithms. The algorithms are simple and easy to implement. Their convergence naturally follows from the convergence properties of the two standard convex optimization framework they build on. We examined the performance and behavior of the proposed formulation on two image processing tasks: inpainting and cartoon texture separation. Furthermore, we showed that the $\ell_{2,\infty} - \ell_1$ formulation in particular opens the door to a wide variety of structured sparsity, that could bring additional practical benefits while still being consistent with the CSC model. An interesting example of such structured sparsity was offered in the combination of the CSC and patch-averaging models, showing that such a mixture provides improved performance. Finally, we envision that similar combinations of global and local sparse priors, within the proposed unifying framework, will allow to further benefits in several other restoration problems.

APPENDIX

Proposition 1: For a point y and the ℓ_1 -ball of radius r , $\mathcal{B}_r := \{x, \text{s.t. } \|x\|_1 \leq r\}$, the distance between y and the ball

$$d(y, \mathcal{B}_r) := \inf \{\|x - y\|_2, \mid x \in \mathcal{B}_r\},$$

is a convex function of the ball radius r .

Proof: From the ℓ_1 -norm triangle inequality, it comes that for any convex combination of two radii $\theta r_1 + (1 - \theta)r_2$, with $0 \leq \theta \leq 1$, we have the inclusion

$$\theta \mathcal{B}_{r_1} + (1 - \theta) \mathcal{B}_{r_2} \subset \mathcal{B}_{\theta r_1 + (1 - \theta)r_2},$$

where $\theta \mathcal{B}_{r_1}$ denotes the set of points $\{\theta x_1 \mid x_1 \in \mathcal{B}_{r_1}\}$. In particular, for the nearest points to y in \mathcal{B}_{r_1} and \mathcal{B}_{r_2} respectively, i.e., for $x_1 \in \mathcal{B}_{r_1}$ such that $\|y - x_1\|_2 = d(y, \mathcal{B}_{r_1})$ and $x_2 \in \mathcal{B}_{r_2}$ such that $\|y - x_2\|_2 = d(y, \mathcal{B}_{r_2})$, we have

$$\theta x_1 + (1 - \theta)x_2 \in \mathcal{B}_{\theta r_1 + (1 - \theta)r_2},$$

and therefore

$$\|y - (\theta x_1 + (1 - \theta)x_2)\|_2 \geq d(y, \mathcal{B}_{\theta r_1 + (1 - \theta)r_2}).$$

Finally, from the Euclidean norm triangle inequality, it comes that

$$\theta d(y, \mathcal{B}_{r_1}) + (1 - \theta)d(y, \mathcal{B}_{r_2}) \geq d(y, \mathcal{B}_{\theta r_1 + (1 - \theta)r_2})$$

which proves that $r \mapsto d(y, \mathcal{B}_r)$ is convex. ■

REFERENCES

- [1] M. Elad, *Sparse and Redundant Representations - From Theory to Applications in Signal and Image Processing*. Berlin, Germany: Springer, 2010.
- [2] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.
- [3] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proc. IEEE 12th Int. Conf. Comput. Vision*, 2009, pp. 2272–2279.
- [4] Y. Romano, M. Protter, and M. Elad, "Single image interpolation via adaptive non-local sparsity-based modeling," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 3085–3098, Jul. 2014.
- [5] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.
- [6] G. Yu, G. Sapiro, and S. Mallat, "Solving inverse problems with piecewise linear estimators: From Gaussian mixture models to structured sparsity," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2481–2499, May 2012.
- [7] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1620–1630, Apr. 2013.
- [8] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
- [9] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1999, vol. 5, pp. 2443–2446.
- [10] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, Jan. 2008.
- [11] J. Sulam and M. Elad, "Expected patch log likelihood with a sparse prior," in *Proc. Int. Workshop Energy Minimization Methods Comput. Vision Pattern Recognit.*, 2015, pp. 99–111.
- [12] V. Pappas and M. Elad, "Multi-scale patch-based image restoration," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 249–261, Jan. 2016.
- [13] J. Mairal, G. Sapiro, and M. Elad, "Learning multiscale sparse representations for image and video restoration," *Multiscale Model. Simul.*, vol. 7, no. 1, pp. 214–241, 2008.
- [14] J. Sulam, B. Ophir, and M. Elad, "Image denoising through multi-scale learnt dictionaries," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 808–812.
- [15] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *Proc. IEEE Int. Conf. Comput. Vision*, 2011, pp. 479–486.
- [16] R. Grosse, R. Raina, H. Kwong, and A. Y. Ng, "Shift-invariance sparse coding for audio classification," in *Proc. 23rd Conf. Uncertainty Artif. Int.*, Vancouver, BC, Canada, 2007, pp. 149–158.
- [17] J. Thiagarajan, K. Ramamurthy, and A. Spanias, "Shift-invariant sparse representation of images using learned dictionaries," in *Proc. IEEE Workshop Mach. Learn. Signal Process.*, 2008, pp. 145–150.
- [18] C. Rusu, B. Dumitrescu, and S. A. Tsafaris, "Explicit shift-invariant dictionary learning," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 6–9, Jan. 2014.
- [19] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 391–398.
- [20] F. Heide, W. Heidrich, and G. Wetzstein, "Fast and flexible convolutional sparse coding," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 5135–5143.
- [21] B. Kong and C. C. Fowlkes, "Fast convolutional sparse coding (FCSC)," Dept. Comput. Sci., Univ. California, Irvine, California, Tech. Rep., vol. 3, 2014.
- [22] B. Wohlberg, "Efficient convolutional sparse coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 7173–7177.
- [23] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang, "Convolutional sparse coding for image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 1823–1831.

- [24] F. Yellin, B. D. Haeffele, and R. Vidal, "Blood cell detection and counting in holographic lens-free imaging by convolutional sparse dictionary learning and coding," in *Proc. IEEE 14th Int. Symp. Biomed. Imag.*, 2017, pp. 650–653.
- [25] A. Serrano, F. Heide, D. Gutierrez, G. Wetzstein, and B. Masia, "Convolutional sparse coding for high dynamic range imaging," in *Computer Graphics Forum*. Lisbon, Portugal: Wiley Online Library, 2016, vol. 35, pp. 153–163.
- [26] E. Skau and C. Garcia-Cardona, "Tomographic reconstruction via 3D convolutional dictionary learning," in *Proc. IEEE 13th Image, Video, Multidimensional Signal Process. Workshop*, 2018, pp. 1–5.
- [27] V. Pappayan, J. Sulam, and M. Elad, "Working locally thinking globally: Theoretical guarantees for convolutional sparse coding," *IEEE Trans. Signal Process.*, vol. 65, no. 21, pp. 5687–5701, Nov. 2017.
- [28] E. Plaut and R. Giryes, "Matching pursuit based convolutional sparse coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6847–6851.
- [29] B. Wohlberg, "Convolutional sparse coding with overlapping group norms," Aug. 2017, *arXiv:1708.09038*.
- [30] V. Pappayan, Y. Romano, M. Elad, and J. Sulam, "Convolutional dictionary learning via local processing," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5306–5314.
- [31] E. Zisselman, J. Sulam, and M. Elad, "A local block coordinate descent algorithm for the CSC model," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2019, pp. 8208–8217.
- [32] B. Wohlberg, "Efficient algorithms for convolutional sparse representations," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 301–315, Jan. 2016.
- [33] E. Skau and B. Wohlberg, "A fast parallel algorithm for convolutional sparse coding," in *Proc. IEEE 13th Image, Video, Multidimensional Signal Process. Workshop*, 2018, pp. 1–5.
- [34] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Scalable online convolutional sparse coding," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4850–4859, Oct. 2018.
- [35] S. Boyd *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.
- [36] C. Garcia-Cardona and B. Wohlberg, "Convolutional dictionary learning: A comparative review and new algorithms," *IEEE Trans. Comput. Imag.*, vol. 4, no. 3, pp. 366–381, Sep. 2018.
- [37] M. Kowalski, "Sparse regression using mixed norms," *Appl. Comput. Harmonic Anal.*, vol. 27, no. 3, pp. 303–324, 2009.
- [38] M. Tao and X. Yuan, "Convergence analysis of the direct extension of ADMM for multiple-block separable convex minimization," *Advances Comput. Math.*, vol. 44, no. 3, pp. 773–813, 2018.
- [39] C. T. Kelley, *Iterative Methods for Optimization*, vol. 18. Philadelphia, PA, USA: SIAM, 1999.
- [40] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l_1 -ball for learning in high dimensions," in *Proc. 25th Int. Conf. Mach. Learn.* ACM, 2008, pp. 272–279.
- [41] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
- [42] P. L. Combettes and J.-C. Pesquet, "A proximal decomposition method for solving convex variational inverse problems," *Inverse Problems*, vol. 24, no. 6, 2008, Art. no. 065014.
- [43] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point Algorithms for Inverse Problems in Science and Engineering*, Berlin, Germany: Springer, 2011, pp. 185–212.
- [44] N. Parikh *et al.*, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2014.
- [45] H. H. Bauschke *et al.*, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, vol. 408. Berlin, Germany: Springer, 2011.
- [46] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.: A J. Issued by Courant Inst. Math. Sci.*, vol. 57, no. 11, pp. 1413–1457, 2004.
- [47] D. Carrera, G. Boracchi, A. Foi, and B. Wohlberg, "Sparse overcomplete denoising: Aggregation versus global optimization," *IEEE Signal Process. Lett.*, vol. 24, no. 10, pp. 1468–1472, Oct. 2017.
- [48] S. H. Chan, R. Khoshabeh, K. B. Gibson, P. E. Gill, and T. Q. Nguyen, "An augmented Lagrangian method for total variation video restoration," *IEEE Trans. Image Process.*, vol. 20, no. 11, pp. 3097–3111, Nov. 2011.