SUB-DIP: OPTIMIZATION ON A SUBSPACE WITH DEEP IMAGE PRIOR
REGULARIZATION AND APPLICATION TO SUPERRESOLUTION

Alexander Sagel
fortiss GmbH, Munich, Germany
sagel@fortiss.org

Aline Roumy, Christine Guillemot
Inria, Rennes, France
FIRSTNAME.LASTNAME@inria.fr

ABSTRACT
The Deep Image Prior has been recently introduced to solve
inverse problems in image processing with no need for train-
ing data other than the image itself. This property is of great
interest, in particular when there is a lack of data to perform
the training. However, the original training algorithm of the
Deep Image Prior constrains the reconstructed image to be on
a manifold described by a convolutional neural network. For
some problems, this neglects prior knowledge and can render
certain regularizers ineffective. This work proposes an alter-
native approach that relaxes this constraint and fully exploits
all prior knowledge. We evaluate our algorithm on the prob-
lem of reconstructing a high-resolution image from a down-
sampled version and observe a significant improvement over
the original Deep Image Prior algorithm.

Index Terms— Image reconstruction, Image restoration,
Inverse problems, Neural networks

1. INTRODUCTION
Inverse problems appear in a variety of image processing ap-
lications, e.g., image denoising, inpainting or superresolu-
tion. A common strategy for dealing with ill-posed inverse
problems consists in introducing some prior knowledge on
the kind of typical images we try to restore, which helps re-
stricting the class of admissible solutions.

In their simplest form, priors are handcrafted regulariz-
ers that promote certain properties assumed about the data
at hand. A prominent example is Total Variation (TV) reg-
ularization that favors piecewise smooth solutions. More
sophisticated priors require some form of learning procedure
applied to the input image in order to infer additional struc-
tural information. For instance, it is common to assume that
localized patches of a natural image have a sparse representa-

ever, unlike the traditional “shallow” algorithms mentioned
above, they entirely rely on very large data sets.

The Deep Image Prior (DIP) has recently demon-
strated that a deep convolutional neural network (CNN) can
still provide superior performance on different kinds of in-
verse problems, even though it is trained exclusively on the
input image itself, thus leveraging the advantages of tradi-
tional and deep learning algorithms.

In the original formulation, DIP searches for a CNN-
generated image that fulfills best the constraint posed by the
inverse problem. Thus, it models the set of possible solutions
as the manifold described by the CNN and the constraints
posed by the inverse problem as a real-valued energy func-
tion to be minimized. Our work proposes the opposite point
of view. The feasible set is defined by the inverse problem,
while DIP is merely a regularizing energy function to be
minimized. We show that this inversion of perspective can
improve DIP performance on superresolution problems.

2. DIP: BACKGROUND AND PRIOR WORK
Many inverse problems consist of reconstructing an image
\( x \in \mathbb{R}^n \) from a measurement \( y \in \mathbb{R}^d \), \( d < n \), i.e.
\[
y = Ax + \eta,
\]
where \( A \) represents a linear degradation operator. The term \( \eta \) corresponds to an optional Gaussian noise.

DIP approximates the solution of the underdetermined
linear Eq. (1) by employing a CNN. Specifically, given a
CNN
\[
T_\theta : \mathbb{R}^n \rightarrow \mathbb{R}^n
\]
parameterized by its trainable weights \( \theta \in \mathbb{R}^p \), and a fixed
input \( z \in \mathbb{R}^n \), DIP estimates the minimizer \( \theta^* \) of the objective
\[
f_{\text{DIP}}(\theta) = \| AT_\theta(z) - y \|^2,
\]
and returns
\[
x_{\text{DIP}} = T_{\theta^*}(z), \text{ s.t. } \theta^* = \arg\min_{\theta \in \mathbb{R}^p} f_{\text{DIP}}(\theta)
\]
as the approximated solution of Eq. (1).

Even though the DIP model was introduced only recently,
some important research focusing on the formulation of its

Alexander Sagel carried out the research at Inria in Rennes, France. This
work was supported by the EU H2020 Research and Innovation Programme
under grant agreement No 694122 (ERC advanced grant CLIM).
3. A SUBSPACE INDUCED DIP OBJECTIVE

We can interpret the DIP objective in Eq. (1) in terms of the involved sets. If we neglect the noise term in Eq. (1), then the solution set is an affine space. Let us denote this space by \( V \)

\[
V = \{ x | Ax = y \}
\]

and let

\[
T = \{ T_\theta(z) | \theta \in \mathbb{R}^p \}
\]

denote the manifold described by the CNN. We define the \( A \)-weighted distance \( d_A \) of a point \( x \) to a set \( S \) as

\[
d_A(x,S) = \min_{s \in S} \| Ax - s \|^2.
\]

Since, \( \forall s \in V, \| Ax - y \| = \| Ax - \hat{s} \| = \min_{s \in V} \| Ax - s \| \), DIP essentially returns the point \( x^* \in T \) that minimizes the \( A \)-weighted distance to \( V \), i.e.

\[
x^*_{\text{DIP}} = \arg \min_{x \in T} \| Ax - y \|^2 = \arg \min_{x \in T} d_A(x,V).
\]

However, constraining the solution to \( T \) has disadvantages. To begin with, \( V \) and \( T \) are usually disjoint, meaning that the result will never exactly fulfill the constraints posed by the inverse problem. For instance, a superresolved image \( x^* \) obtained from a downscaled version \( y \) via Eq. (1), will not result in \( y \) when downsampled again. While this may be desirable in problems involving noise, in cases like noiseless superresolution or inpainting, it is not. But even in cases, where noise is involved, it is unlikely that \( T \) is the ideal model of a natural image manifold. Another problem with Eq. (7) can arise when used in combination with regularizers to promote certain features of \( x \) that are not captured by \( T \). To illustrate this problem, consider the extreme case, where the employed regularizer \( R(x) \) is such that \( T \) is a level set, i.e.

\[
R(x) = \text{const. } \forall x \in T.
\]

Then, adding \( R(x) \) to Eq. (7) has no effect on the optimization. A naive remedy to these problems is to search for an \( x_{\text{btwn}} \) somewhere between \( T \) and \( V \) by means of the objective

\[
\min_{x_{\text{btwn}} \in \mathbb{R}^n, \theta \in \mathbb{R}^p} \| T_\theta(z) - x_{\text{btwn}} \|^2 + \lambda \| Ax_{\text{btwn}} - y \|^2,
\]

where \( \lambda > 0 \) is a tuning parameter. Unfortunately, we observed that this approach rarely has success due to local minima. Specifically, \( x_{\text{btwn}} \) tends to quickly converge to early estimations of \( T_\theta(z) \) and not evolve any further.

As an alternative, we propose an approach based on an inverted perspective of Eq. (7). Specifically, we aim to search for a point in \( V \) that minimizes some distance to \( T \), i.e., for an appropriate choice of some full rank matrix \( B \in \mathbb{R}^{m \times n}, m \leq n \), we formulate the objective

\[
x^* = \arg \min_{x \in V} d_B(x,T).
\]

The matrix \( B \) can be seen as an additional degree of freedom in the objective that allows us to measure the distance in a way such that the algorithm does not get stuck at early estimations of \( x \). Fig. 1 visualizes the optimization procedures of DIP and \( \text{SUB-DIP} \) in the following. DIP is performing its optimization (blue curve) by searching for the point on \( T \) that minimizes the distance (red dashed line) to \( V \). In our approach, the optimization (red curve) is performed by searching for the point on \( V \) that minimizes the distance to \( T \). In principle, Eq. (9) could be solved by combining the classical DIP algorithm with appropriate projection operations. However, such an approach would not allow for an additional regularization of \( x \), which is why we prefer to optimize Eq. (9) directly. Eq. (9) does not exhibit the regularizer problem of DIP, since a regularizer with \( V \) as a level set contradicts the very definition of Eq. (1).

By construction, the solution in Eq. (9) exactly fulfills the constraint in Eq. (1) for \( \eta = 0 \). This noiseless case is the main focus of our work. However, when only noisy observations are available, the solution \( x \) must depart from the subspace \( V \). To do so, we can modify Eq. (9) by incorporating a relaxation term \( \xi \in V_\perp \), where \( V_\perp \) is the row space of \( A \), i.e.

\[
\min_{x_{\text{btwn}} \in \mathbb{R}^n, \theta \in \mathbb{R}^p} \| T_\theta(z) - x_{\text{btwn}} \|^2 + \lambda \| Ax_{\text{btwn}} - y \|^2 + \| \xi \|^2,
\]

where \( \xi \in V_\perp \). This modified objective can be solved for any initial point \( x_0 \) by means of the optimization

\[
\min_{x_{\text{btwn}} \in \mathbb{R}^n, \theta \in \mathbb{R}^p} \| T_\theta(z) - x_{\text{btwn}} \|^2 + \lambda \| Ax_{\text{btwn}} - y \|^2 + \| \xi \|^2,
\]

where \( \lambda > 0 \) is a tuning parameter. Unfortunately, we observed that this approach rarely has success due to local minima. Specifically, \( x_{\text{btwn}} \) tends to quickly converge to early estimations of \( T_\theta(z) \) and not evolve any further.

As an alternative, we propose an approach based on an inverted perspective of Eq. (7). Specifically, we aim to search for a point in \( V \) that minimizes some distance to \( T \), i.e., for an appropriate choice of some full rank matrix \( B \in \mathbb{R}^{m \times n}, m \leq n \), we formulate the objective

\[
x^* = \arg \min_{x \in V} d_B(x,T).
\]

The matrix \( B \) can be seen as an additional degree of freedom in the objective that allows us to measure the distance in a way such that the algorithm does not get stuck at early estimations of \( x \). Fig. 1 visualizes the optimization procedures of DIP and \( \text{SUB-DIP} \) in the following. DIP is performing its optimization (blue curve) by searching for the point on \( T \) that minimizes the distance (red dashed line) to \( V \). In our approach, the optimization (red curve) is performed by searching for the point on \( V \) that minimizes the distance to \( T \). In principle, Eq. (9) could be solved by combining the classical DIP algorithm with appropriate projection operations. However, such an approach would not allow for an additional regularization of \( x \), which is why we prefer to optimize Eq. (9) directly. Eq. (9) does not exhibit the regularizer problem of DIP, since a regularizer with \( V \) as a level set contradicts the very definition of Eq. (1).

By construction, the solution in Eq. (9) exactly fulfills the constraint in Eq. (1) for \( \eta = 0 \). This noiseless case is the main focus of our work. However, when only noisy observations are available, the solution \( x \) must depart from the subspace \( V \). To do so, we can modify Eq. (9) by incorporating a relaxation term \( \xi \in V_\perp \), where \( V_\perp \) is the row space of \( A \), i.e.

\[
\min_{x_{\text{btwn}} \in \mathbb{R}^n, \theta \in \mathbb{R}^p} \| T_\theta(z) - x_{\text{btwn}} \|^2 + \lambda \| Ax_{\text{btwn}} - y \|^2 + \| \xi \|^2.
\]
orthogonal complement to the vector space that is parallel to \( \mathbb{V} \). A relaxed version of Eq. (9) can be thus formulated as

\[
x^*, \xi^* = \arg \min_{x \in \mathbb{V}, \xi \in \mathbb{V}_\perp} d_B(x + \xi, T) + \lambda \|\xi\|^2,
\]

where \( \lambda > 0 \) is a tuning parameter that should be chosen inversely proportional to the noise variance.

The formulations in Eq. (7) and Eq. (10) allow us to better exploit the constraints in Eq. (1), either by limiting the solution set entirely to \( \mathbb{V} \) or by penalizing the distance to \( \mathbb{V} \).

Additionally to not having the disadvantages of the original DIP formulation as discussed in the beginning of this section, we observed that the formulations in Eq. (7) and Eq. (10) are less prone to numerical failures, when used in combination with intricate regularizers. This is due to the fact that unlike Eq. (7), \( x \) is not a function of \( \theta \) and thus the gradient does not need to be backpropagated through several layers of the CNN. Finally, by choosing the weighting matrix \( B \) appropriately, we can avoid getting stuck in local minima as it is the case for Eq. (8).

4. APPLICATION TO SUPERRESOLUTION

In the following, we describe an algorithm that applies SUB-DIP to the problem of superresolution. For the sake of clarity, we derive the procedure on 1D signals but emphasize that the generalization to 2D signals is straightforward. Indeed, experiments are carried out on 2D images.

4.1. Superresolution Algorithm

Superresolution refers to the problem of reconstructing a high-resolution signal \( x \in \mathbb{R}^{\tau d} \) from a low-resolution version \( y \in \mathbb{R}^{d} \) with \( \tau \in \mathbb{N} \) being the magnification factor. It is generally assumed that \( y \) results from subsequent filtering of \( x \in \mathbb{R}^{\tau d} \) with a filter \( h \in \mathbb{R}^{L} \), and downsampling by the factor \( \tau \) i.e.

\[
y = DS_\tau[h * x],
\]

where DS_\tau is a subsampling operator and * denotes convolution\(^1\). Eq. (11) can be written in the form of Eq. (1). Without loss of generality, we assume that \( L = k\tau, k \in \mathbb{N} \). The vector \( h \) can thus be split into \( k \) subvectors \( h_i \in \mathbb{R}^{\tau} \) as follows.

\[
h^\top = [h_1^\top \ h_2^\top \ \cdots \ h_{k-1}^\top \ h_k^\top].
\]

Let us define the matrix \( H \in \mathbb{R}^{d \times \tau d} \) as

\[
H = \begin{bmatrix}
h_1^\top \\
h_2^\top \\
\vdots \\
h_k^\top \\
h_1^\top \\
h_2^\top \\
\vdots \\
h_{k-1}^\top \\
h_1^\top \\
h_2^\top
\end{bmatrix}.
\]

Then, Eq. (11) is equivalent to \( y = Hx \). With \( x_0 = H^\top y \), the solution set can be written as

\[
\mathbb{V} = \{x_0 + G^\top s | s \in \mathbb{R}^{(\tau-1)d}\},
\]

where \( G \) is a matrix with columns that span the kernel \( \ker(H) \) of \( H \). We can now rewrite Eq. (9) to fit our superresolution problem, as follows.

\[
\min_{x \in \mathbb{V}} d_B(x, T) = \min_{s \in \mathbb{R}^{(\tau-1)d}} ||B(x_0 + G^\top s - T_\theta(z))||^2. \tag{15}
\]

Similarly to Eq. (10), we can include noise into Eq. (15) via\(^2\)

\[
\min_{s \in \mathbb{R}^{(\tau-1)d}, \theta \in \mathbb{R}^p} ||B(x_0 + G^\top s + H^\top \xi - T_\theta(z))||^2 + \lambda_\xi \|\xi\|^2. \tag{16}
\]

Equations (15) and (16) can be optimized by applying any common variation of gradient descent to \( \theta \) and \( s \). The optimization procedure of SUB-DIP is almost identical to the one of DIP, but includes a gradient step for \( s \) in each iteration.

The matrices \( B \) and \( G \) are extremely large for real-world signals and the backpropagation of the gradient is computationally only feasible, if they exhibit a convolutional or transpose-convolutional structure. In other words, we need to find a set of filters that can replace the respective matrix multiplications by convolutions.

For \( G \), we need to make sure, that the chosen filters \( g^1, \ldots, g^{r-1} \) span \( \ker(H) \). Candidates for \( g^1, \ldots, g^{r-1} \) can be generated in the following way. We consider the submatrix \( H_{i, \tau(d-k)+1:\tau d} \) of \( H \) that contains only its last \( L \) columns. Note that only the last \( 2k-1 \) rows of the matrix are non-zero.

We can then generate an orthogonal basis \( \gamma_1, \ldots, \gamma_{L-1-2k} \) of \( \ker(H_{i, \tau(d-k)+1:\tau d}) \). Then, for any matrix \( \Gamma_i \in \mathbb{R}^{d \times \tau d} \) that is constructed from a basis element \( \gamma_i \) in the same manner as \( H \) is constructed from \( h_i \), the rows lie in \( \ker(H) \). The matrix \( B \) is chosen to fulfill two properties. It should be easily realizable by convolutional operations and allows us to decompose Eq. (15) and Eq. (16) into its components on \( \mathbb{V} \) and \( \mathbb{V}_\perp \). We thus fix

\[
B = [H^\top \ \sqrt{\lambda_G} G^\top]^\top,
\]

where \( \lambda_G > 0 \) is a parameter. Since \( HG^\top = 0 \), we get

\[
||B(x_0 + G^\top s - T_\theta(z))||^2 = ||H(x_0 - T_\theta(z))||^2 + \lambda_G ||G^\top s - T_\theta(z)||^2,
\]

for the Eq. (15) and a similar decomposition for Eq. (16). The two terms correspond to the known, and interpolated parts of \( x \). Choosing an appropriate weighting factor \( \lambda_G \) permits us to avoid running into early local minima, as mentioned before. By putting less emphasis on the second term, we can ensure that \( T_\theta(z) \) does not get optimized to approach meaningless early estimations of the unknown parts of \( x \).

\(^1\) Mathematically, the operation is actually a cross-correlation

\(^2\) \( H^\top \xi \) is in \( \mathbb{V}_\perp \), because \( \mathbb{V}_\perp \) is the row space of \( H \).
4.2. TV Regularization

Our new formulation, allows to easily introduce new regularizers. We illustrate this by considering TV. It was shown [7] that TV regularization can further improve the performance of DIP. For a 2D image $I$, a TV regularizer is formulated as

$$r_{TV}(I) = \sum_{i,j} \sqrt{(I_{i,j} - I_{i-1,j})^2 + (I_{i,j} - I_{i,j-1})^2}. \quad (19)$$

With TV regularization Eq. (9) becomes

$$x^* = \arg\min_{x \in V} d_B(x, T) + \lambda_{TV} r_{TV}(x). \quad (20)$$

5. EXPERIMENTS

We evaluate our algorithm on the 4x superresolution of the Set5 [11] dataset. To this end, we calculate the PSNR values for the superresolved images reconstructed by our algorithm. The emphasis of this section is to perform an ablation study that compares the original DIP to SUB-DIP. We reuse the official DIP implementation that was made publicly available [12]. We directly optimize the objectives in Eq. (15) and Eq. (16) by means of the Adam optimizer and fix $\lambda_G = 0.1$. The same parameters for DIP as for SUB-DIP are used. In particular, a Lanczos filter is used as $h$, even though the actual filter that was used to perform the downsampling is unknown to us. Since many superresolution algorithms operate on gray-scale images only, it is common to calculate the PSNR exclusively on the luminance channel of an image. Here, we are more interested in how SUB-DIP compares to other RGB algorithms, in particular the original DIP formulation, which is why we compute the PSNR on all three RGB channels. This explains why the values reported in [5], are slightly higher. DIP performs some cropping in order to create images with dimensions divisible by 32. For the sake of comparability, we thus crop the ground truth images accordingly, as well as the images generated by all other baseline algorithms, obtained from [12].

5.1. Noiseless Superresolution

Often, superresolution does not involve any noise. This is for instance the case when the downsampling was done digitally. In that case, we can search for the solution directly on $V$ by means of Eq. (15). We evaluate DIP and SUB-DIP both without any regularization as well as with an additional TV term as described in Section 4.2 with $\lambda_{TV} = 10^{-3}/n$. Table 1 shows the reconstruction results.

SUB-DIP consistently outperforms DIP without TV regularization. With an added TV term, DIP is also outperformed by SUB-DIP, except for the “Woman” image. The advantage of our formulation of the DIP objective can be observed in Fig. 2. The manifold $T$ tends not to capture certain directional structures at very high frequencies. Restricting the solution to $T$ thus causes a blurring out of such structures.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Baby</th>
<th>Bird</th>
<th>Butterfly</th>
<th>Head</th>
<th>Woman</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicubic</td>
<td>30.43</td>
<td>28.09</td>
<td>20.90</td>
<td>28.72</td>
<td>25.39</td>
<td>26.71</td>
</tr>
<tr>
<td>LapSRN</td>
<td>32.02</td>
<td>30.42</td>
<td>25.52</td>
<td>29.62</td>
<td>29.24</td>
<td>29.36</td>
</tr>
<tr>
<td>DIP</td>
<td>29.85</td>
<td>29.62</td>
<td>24.83</td>
<td>28.49</td>
<td>27.00</td>
<td>27.96</td>
</tr>
<tr>
<td>DIP+TV</td>
<td>29.87</td>
<td>29.52</td>
<td>24.60</td>
<td>28.62</td>
<td>27.20</td>
<td>27.96</td>
</tr>
<tr>
<td>Ours</td>
<td>31.38</td>
<td>29.99</td>
<td>24.84</td>
<td>28.74</td>
<td>27.01</td>
<td>28.39</td>
</tr>
</tbody>
</table>

Table 1: PSNR values for Set5. Best and second-best results are written in red, and blue respectively.

This is why, for instance, the DIP struggles to reproduce fine details such as eyelashes, while SUB-DIP does not have this problem to that extent. State-of-the-art deep learning methods [13, 14, 15] such as LapSRN [3] achieve generally better performance than DIP or SUB-DIP, but require training on large datasets, which is not always feasible.

5.2. Noisy Superresolution

We test the approach in Eq. (16) by adding Gaussian noise with $\sigma_{noise} = 0.02 \cdot I_{max}$ to the downsampled image $y$, where $I_{max}$ corresponds to the maximal pixel value. The noise regulization weight is set to $\lambda_G = 3 \cdot 10^{-4} / (d \cdot \sigma_{noise}^2)$. Even though our method can no longer exploit the subspace assumption, it still achieves a slight improvement over classical DIP, as can be seen in Table 2.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Baby</th>
<th>Bird</th>
<th>Butterfly</th>
<th>Head</th>
<th>Woman</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIP</td>
<td>29.24</td>
<td>27.70</td>
<td>23.81</td>
<td>26.95</td>
<td>26.17</td>
<td>26.78</td>
</tr>
<tr>
<td>Ours</td>
<td>29.74</td>
<td>28.22</td>
<td>23.89</td>
<td>27.47</td>
<td>26.40</td>
<td>27.15</td>
</tr>
</tbody>
</table>

Table 2: PSNR values for Set 5 with noise

6. CONCLUSION

In this work, we presented a novel approach to leverage the Deep Image Prior by formulating an optimization procedure on the solution set of an inverse problem. We have described how to apply this approach on the problem of superresolution and have demonstrated a significant improvement of the reconstruction results.
7. REFERENCES


