Beyond Variational Models and Self-Similarity in Super-Resolution: Unfolding Models and Multi-Head Attention

Ivan Pereira-Sánchez[1,2][a], Eloi Sans[1][b], Julia Navarro[1,2][c] and Joan Duran[1,2][d]
[1] *Departament de Ciències Matemàtiques i Informàtica, Universitat de les Illes Balears (UIB), Spain*
[2] *Institute of Applied Computing and Community Code (IAC3), Spain*

Keywords: Unfolding, Multi-Head Attention, Image Super-Resolution, Variational Methods, Nonlocal Regularization, Self-Similarity.

Abstract: Classical variational methods for solving image processing problems are more interpretable and flexible than pure deep learning approaches, but their performance is limited by the use of rigid priors. Deep unfolding networks combine the strengths of both by unfolding the steps of the optimization algorithm used to estimate the minimizer of an energy functional into a deep learning framework. In this paper, we propose an unfolding approach to extend a variational model exploiting self-similarity of natural images in the data fidelity term for single-image super-resolution. The proximal, downsampling and upsampling operators are written in terms of a neural network specifically designed for each purpose. Moreover, we include a new multi-head attention module to replace the nonlocal term in the original formulation. A comprehensive evaluation covering a wide range of sampling factors and noise realizations proves the benefits of the proposed unfolding techniques. The model shows to better preserve image geometry while being robust to noise.

1 INTRODUCTION

The goal of image super resolution is to recover a high-resolution image from a low-resolution observation of it. This field has significant importance due to its numerous practical applications, including satellite imaging, biometric information identification, medical imaging, remote sensing, microscopy image processing, surveillance, multimedia industry, video enhancement or astrological studies, among others (Lepcha et al., 2023).

In the literature, a wide variety of methods have been proposed to address image super-resolution. One may roughly divide them into interpolation, model-based, and learning-based methods.

Interpolation methods are commonly used for quickly generating zoomed images, providing a good compromise between efficiency and quality. These methods typically estimate pixel intensity using fixed kernels, such as local variance coefficients (Keys, 1981), or adaptive structural kernels (Li and Orchard, 2001; Zhang and Wu, 2006). While these methods are efficient, they can introduce artifacts. For applications demanding greater precision, more sophisticated techniques are required.

Model-based methods assume that the low-resolution observed image is derived from the sought high-resolution one after applying a sequence of operators, usually blur filtering and downsampling. Reversing such a process is an ill-posed inverse problem, thus prior knowledge on the structure of natural images must be assumed to regularize it. The most popular model-based methods are variational models, which define an energy functional that induces a high energy when the priors are not fulfilled. The total variation (TV) semi-norm (Rudin et al., 1992) and nonlocal regularization terms, which exploit image self-similarities (Duran et al., 2014; Wang et al., 2019; Gilboa and Osher, 2009), have been mainly used as regularization terms in the variational framework. Several variational methods have been proposed in the literature to deal with image super-resolution (Yue et al., 2016). In this setting, the total variation (TV) semi-norm (Babacan et al., 2008) and nonlocal regularization exploiting image self-similarities (Dong et al., 2011; Dong et al., 2013; Zhang et al., 2012) have been mainly used as priors. Recently, (Pereira-Sánchez et al., 2022) proposed to

[a] https://orcid.org/0000-0002-9400-5720
[b] https://orcid.org/0009-0000-7702-2425
[c] https://orcid.org/0000-0003-3667-7008
[d] https://orcid.org/0000-0003-0043-1663

Pereira-Sánchez, I., Sans, E., Navarro, J. and Duran, J.
Beyond Variational Models and Self-Similarity in Super-Resolution: Unfolding Models and Multi-Head Attention.
DOI: 10.5220/0012395400003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 3: VISAPP*, pages 190-198
ISBN: 978-989-758-679-8; ISSN: 2184-4321

leverage the self-similarity in the fidelity term rather than in the regularization term.

In the last decade, a growing number of deep learning-based super-resolution methods have been suggested in the literature and showed promising results. These approaches can be categorized depending on their architectures. In this setting, we can find residual connections (Tai et al., 2017; Lim et al., 2017), generative adversarial networks (Bell-Kligler et al., 2019), or attention modules (Dai et al., 2019). In fact, attention modules are based on the assumption that images are self-similar and are the key of vision transformers, which have become remarkably popular (Wang et al., 2022; Lu et al., 2022).

The use of observation models makes variational methods robust to distortions, but their performance is limited by rigid hand-crafted priors. On the contrary, data-driven learning approaches can easily learn natural priors, but are less flexible and interpretable. Deep unfolding networks (Monga et al., 2021) combine the strengths of both. The general idea consists in unfolding the steps of the optimization algorithm used to estimate the minimizer of an energy functional into a deep learning framework. Since these networks do not model the complete problem but particular operations, architectures are shallower than pure learning approaches. Besides, the unfolding process avoids the need of defining an explicit prior in the model-based formulation. This results in efficient and highly interpretable methods whose network architectures can be trained on acceptable-sized datasets.

In this paper, we propose to extend the variational model in (Pereira-Sánchez et al., 2022) with algorithm unfolding techniques. This model leverages the self-similarity of natural images in the fidelity term. We use the proximal gradient algorithm to compute the minimizer of the proposed energy functional and replace all proximal, downsampling and upsampling operators with a neural network specifically tailored to model each function. In addition, we introduce a new multi-head attention module to take the place of the nonlocal term in the original formulation. Extensive experiments under different sampling factors and noise realizations of the proposed approach demonstrate the effectiveness of the presented method.

The rest of the paper is organized as follows. Section 2 presents the proposed model. Section 3 covers implementation details. Experimental results in comparison with other methods are included in Section 4, while Section 5 analyses the different components of our model. Finally, Section 6 provides the conclusions and outlines the future work. The source code of the proposed method is publicly available at https://github.com/TAMI-UIB/UNLDSR.

2 PROPOSED MODEL

The most common observation model in image super-resolution relates the observed data f with the underlying image u via

$$f = DBu + \eta, \tag{1}$$

where D is a decimation operator and B is a low-pass filter. Then, DB is a linear operator modelling the degradation of u and η is the realization of i.i.d. zero-mean noise. Since solving (1) is an ill-posed inverse problem, the choice of a good prior is required.

2.1 Variational Formulation

Let us denote the low resolution image as $f \in \mathbb{R}^{C\times N}$, where N is the number of pixels and C is the number of spectral bands, and the restored image as $u \in \mathbb{R}^{C\times M} := X$, with $M > N$.

Classical variational models for image restoration include the fidelity term

$$\|DBu - f\|^2 = \sum_{k=1}^{C}\sum_{i=1}^{N}\left((DBu)_{k,i} - f_{k,i}\right)^2. \tag{2}$$

In order to take advantage of image self-similarity, the work in (Pereira-Sánchez et al., 2022) extends the previous energy to a nonlocal framework, replacing the previous term by a nonlocal data term. We propose to enhance the performance of this nonlocal fidelity term by incorporating learning-based modules in the variational framework. In particular, we consider the super-resolution variational model

$$\min_{u\in X} R(u) + \frac{\lambda}{2}\|(DBu - f)_\omega\|^2, \tag{3}$$

where $R : X \to [-\infty, +\infty]$ is a strictly convex, coercive and lower-semicontinuous functional that plays the role of the regularitzation term, $\lambda > 0$ is a trade-off parameter, and $(DBu - f)_\omega \in \mathbb{R}^{C\times N\times N}$ is defined as

$$\|(DBu - f)_\omega\|^2 := \sum_{k=1}^{C}\sum_{i,j=1}^{N}\omega_{i,j}\left((DBu)_{k,i} - f_{k,j}\right)^2.$$

The weights $\{\omega_{i,j}\}$ are originally computed on the observed data f, taking into account both the spatial closeness and the spectral similarity in f.

Since the proposed energy (3) is also strictly convex, coercive and lower-semicontinuous, the existence and uniqueness of minimizer is guaranteed. In contrast to (Pereira-Sánchez et al., 2022), the prior R does not have a specific form in our formulation, since it will be folded into the learning-based scheme. This will permit to learn the regularization on u from training data.

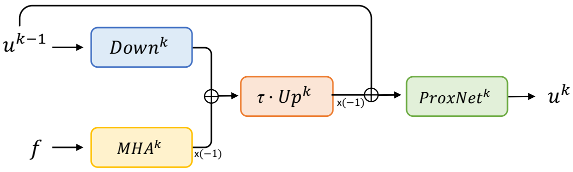


Figure 1: Illustration of one stage of the proposed model, which unfolds Equation (5). *Down*^{k} represents the downsampling DB, and *Up*^{k} is its transposed counterpart $(DB)^T$. *MHA*^{k} denotes the multi-attention module that substitutes the nonlocal operator in the data term, and *ProxNet*^{k} is the architecture that replaces the proximity operator.

2.2 Proximal Gradient

We use the proximal gradient algorithm (Combettes and Wajs, 2005; Chambolle and Pock, 2016) to minimize (3). Given a Hilbert space Z, when considering a proper, lower semi-continuous convex function $R : Z \to [-\infty, +\infty]$, the associated proximal operator is defined as

$$\text{prox}_{\tau R}(x) = \text{argmin}_{y \in Z} R(y) + \frac{1}{2\tau} \|y - x\|^2, \qquad (4)$$

which can be seen as a generalization of the projection operators.

The iterative scheme given by the proximal gradient algorithm becomes

$$\begin{aligned} u^k &= \text{prox}_{\tau R}\left(u^k - \tau \nabla F_\omega(u^k)\right) \\ &= \text{prox}_{\tau R}\left(u^k - \tau (DB)^T (DBu - f_\omega)\right), \end{aligned} \qquad (5)$$

where $(DB)^T$ is the transposed operator of DB and $f_\omega = \sum_{j \in X} \omega_{i,j} f_j$.

2.3 Algorithm Unfolding

We propose to unfold (5) and replace the operators involved the iterative scheme with learning-based networks. In particular, we replace $\text{prox}_{\tau R}$ by *ProxNet*^{k}, DB with *Down*^{k}, $(DB)^T$ with *Up*^{k} and the nonlocal operator by a multi-head attention module *MHA*^{k}. With these replacements we allow the operators to learn the intrinsic properties and geometry of natural images. In addition, by rewriting the proximity operator with a neural network, we avoid the need of specifying the prior on u, which will be implicitly modelled by the network.

In the following we accurately describe these networks. Also, in the rest of the paper we refer to each step k in the iterative scheme as a *stage*. Figure 1 illustrates one stage of the proposed approach. In the first stage, u^0 is initialized with bicubic interpolation.

2.3.1 From Proximal Operator to Residual Network

The proximal operator defined in (4) can be reduced to

$$\begin{aligned} \hat{y} = \text{prox}_{\tau R}(x) &\iff x \in \hat{y} + \tau \partial R(\hat{y}) \\ &\iff \hat{y} \in (Id + \tau \partial R)^{-1}(x), \end{aligned}$$

where ∂R is the sub-differential operator. Therefore, we can view the proximal operator as the inverse of a perturbation of the identity. Given this perspective, residual networks are good candidates for replacing the proximal operator in the unfolding context. They consist of a convolutional neural network followed by a skip connection, also known as a residual connection, making it an approach to identity.

Then, we replace the $\text{prox}_{\tau R}$ function by a residual network *ProxNet*^{k}. We maintain the same architecture in all stages but without sharing the weights. This architecture is depicted in Figure 2.

2.3.2 Upsampling and Downsampling Operators

The classic downsampling operator DB is modelled with a convolution by a Gaussian kernel of σ_{blur} standard deviation, followed by a decimation operator. We propose to learn this operator by modelling it with two 2-dimensional convolutions. The first convolution substitutes the Gaussian convolution and their parameters are determined by the standard deviation σ_{blur}. While the decimation operator is replaced by another convolution which stride equals the sampling factor.

On the other hand, the upsampling operator $(DB)^T$ is classically modelled by a zero-padding operator followed by a Gaussian convolution of the same standard deviation. We replace this upsampling by a concatenation of transposed 2-dimensional convolutions, which serve the purpose of zero padding, followed by a 2-dimensional convolution that replaces the Gaussian convolution. To determine the concatenation of transposed convolutions, we consider the prime decomposition of sampling factors, sorted in ascending order as $s = p_1^{r_1} \cdot \ldots \cdot p_m^{r_m}$. This decomposition determines the steps of the concatenation. For a prime number p, we denote $\mathcal{TC}_p$ as the transposed convolution with a stride equal to p. Using this notation, we can describe the concatenation as the following composition:

$$\overbrace{\underbrace{\mathcal{TC}_{p_m} \circ \cdots \circ \mathcal{TC}_{p_m}}_{r_m} \circ \cdots \circ \underbrace{\mathcal{TC}_{p_1} \circ \cdots \circ \mathcal{TC}_{p_1}}_{r_1}}^{m}.$$

This formula is designed to simplify the upsampling operation by decomposing it into the smallest feasible steps.

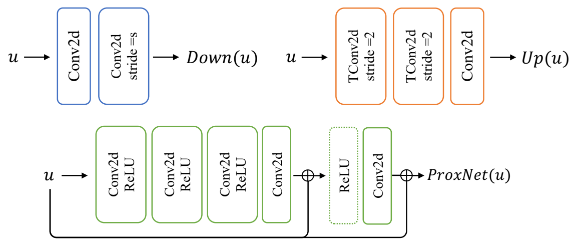


Figure 2: Diagrams of the downsampling and upsampling architectures for a sampling factor of 4, and *ProxNet* network. We keep the same architectures at each stage, but weights are not shared.

These replacements lead to the new downsampling and upsampling operators, $Down^k$ and Up^k, respectively, which appear in Figure 1. Importantly, the involved parameters in both upsampling and downsampling are not shared per stage. An example illustrating the proposed downsampling and upsampling processes for a sampling factor of 4 is shown in Figure 2.

2.3.3 From Self-Similarity to Multi-Head Attention

In the variational model, the weights $\{\omega_{i,j}\}$ are computed on the observed data f, taking into account the similarities in f. This similarity is computed by considering a whole patch around each pixel and using the Euclidean distance across channels. For computational purposes, nonlocal interactions are limited to pixels at a certain distance. In practice, the weights are defined as

$$\omega_{i,j} = \frac{1}{\Gamma_i} \exp\left(- \frac{\left\| f(P_i) - f(P_j) \right\|^2}{h^2_{\text{sim}}} \right),$$

if $\|i-j\|_\infty \leq \nu$ and zero otherwise. In this setting, $\nu \in \mathbb{Z}^+$ determines the size of the window to search for similar pixels, P_i denotes a patch centered at pixel i, and Γ_i is a normalitzation factor. The filtering parameters $h_{\text{spt}} > 0$ measure how fast the weights decay with increasing dissmilarity between patches.

In (Wang et al., 2018), this expression is simplified by considering $\omega_{i,j} = \frac{1}{\Gamma_i} \exp(f(i)^T f(j))$, and by computing the similarity using the dot product while employing the spectral representation of the image x_i at pixel i. The primary motivation for this simplification is that the dot product is computationally more efficient, and the difference between the distance and the dot product does not yield significant differences in practice. Furthermore, they propose a simple extension of these weights by incorporating a convolution operation, such that,

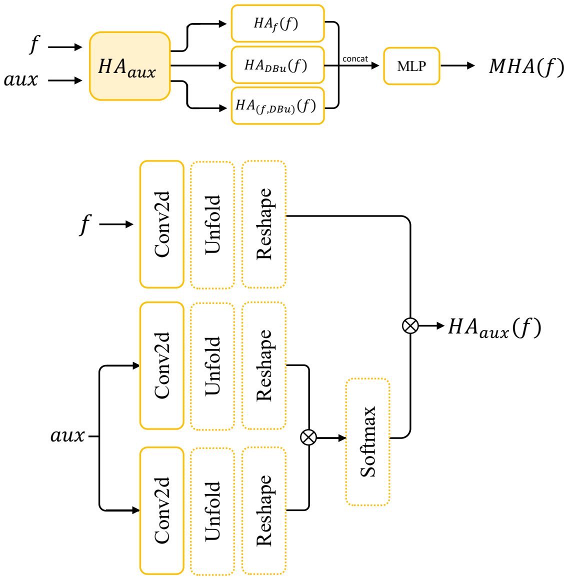
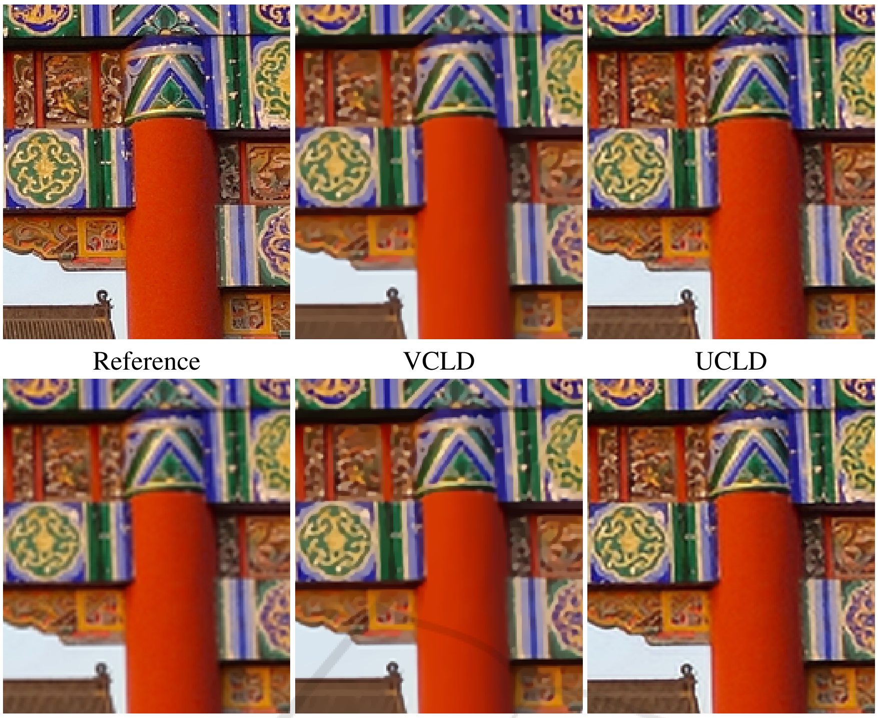


Figure 3: Illustration of the multi-head attention and self-attention networks. The same architecture is used at each stage without sharing weights.

$$\begin{aligned} \omega_{i,j} =& \frac{1}{\Gamma_i} \exp((\theta * f)(i)^T (\phi * f)(j))) \\ =& \text{softmax}\left((\theta * f)(i)^T (\phi * f)(j) \right). \end{aligned}$$

Finally, they compute the nonlocal filter, denoted as $f_\omega = \sum_j w_{i,j} f_j$, by reshaping the data and performing matrix multiplication between the last two dimensions, which can be efficiently executed on a GPU. This module, followed by a residual connection, is referred to as self-attention. This architecture has been widely used and forms the basis of the well-known transformer block, which concatenates the previous architecture and follows it with a multi-layer perceptron (MLP).

In our approach, we propose to write the variational nonlocal module as a multi-head attention one. Instead of working with the spectral representation of individual pixels, we propose to compute the distance between pixels using the spectral representation of patches, to obtain more robust weights. Moreover, we concatenate three self-attention layers, employing three different auxiliary images for weight computation: f, DBu, and the concatenation of both. The architecture is detailed in Figure 3.



Reference VCLD UCLD

BIC VNLD UNLD (ours)

Figure 4: Close-ups of the results for single-image super-resolution on *temple*, with $s = 2$, $\sigma_{blur} = 0.75$ and $\sigma_{noise} = 0$. While the variational methods give oversmooth results, both unfolded models better preserve the geometry and textures of the image.

3 IMPLEMENTATION DETAILS

3.1 Convolution Parameters

The convolution that replaces the Gaussian operators in the downsampling and upsampling operators is determined by the blur standard deviation, denoted as σ_{blur}. The kernel size is calculated as $[4\sigma_{blur}] + 1$, and the padding is set to the integer division of the kernel size by 2. The padding consists of replicating the last pixel at the boundary. This convolution operates independently on channel information, and the weights are initialized with a Gaussian kernel.

For the transposed convolutions of the upsampling operator, when given a prime factor p, the stride is set equal to p, the kernel size is set to the smallest odd number strictly greater than p, and consequently, the padding is established as the integer division of the kernel size by 2. In the case of the convolution that replaces the decimation operator, the stride is set equal to p, the kernel size is chosen as the smallest odd number greater than the sampling factor, and the padding is determined in the same way than the previous convolutions.

For all the convolutions in *ProxNet* we employ a kernel size of 3 and a padding of 1. The number of features added in the first convolution before the residual block is 16. These convolutions do not have bias terms, except for the last convolution, which includes a bias.

Finally, the involved convolutions in the attention module *MHA* have a kernel size equal to one and therefore no padding is needed, any of these convolution have bias. The window size defined for the neighbors in the multi-head attention is set to 5 for all self-attentions. However, the patch size is equal to 3 for the self-attentions that use f and DBu, and it is set to 1 for the one that uses the concatenation of both.

3.2 Training

We trained our complete system in an end-to-end manner using the mean square error as loss function for 400 epochs. The learning rate was initially set to 0.001 and was updated by a factor of 0.75 whenever the loss for the training data did not decrease by more than 0.0001 during the last 10 epochs. Regarding the number of stages, a detailed study of its influence is conducted in Section 5. For most experiments we

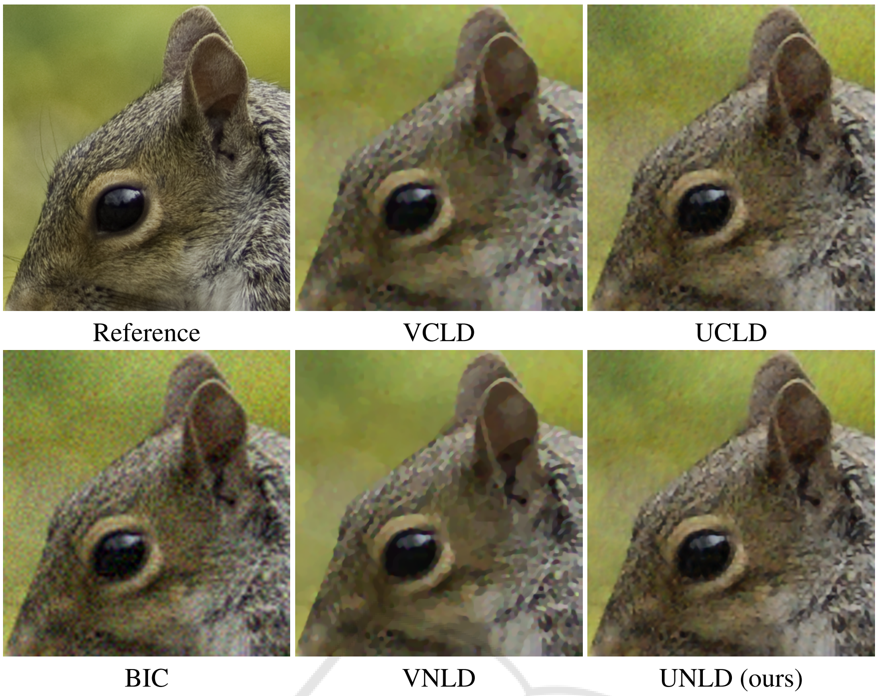


Figure 5: Close-ups of the results for single-image super-resolution on *squirrel*, with $s = 4$, $\sigma_{blur} = 1.4$ and $\sigma_{noise} = 10$. While the variational methods face difficulties in reconstructing the textures of the images, the unfolded results yield better results. Additionally, UNLD proves to be more effective in removing noise when compared to UCLD.

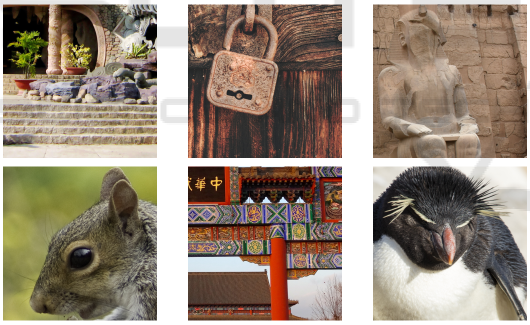


Figure 6: Set of images used for validation. From left to right and from top to botom: *stairs, lock, statue, squirrel, temple, penguin.*

have chosen to set it at 5 because it strikes a good balance between performance and computational cost.

3.3 Dataset

We have selected 19 images from the DIV2K dataset and cropped each of these images to a size of 512x512. These crops have been divided into 13 for training, with the remaining 6 for validation, considering these crops as the reference images. Figure 6 displays the validation data. The low-resolution images are generated by applying a Gaussian kernel, followed by bicubic interpolation for downsampling, and then adding Gaussian noise with a standard deviation in the range of $[0, 255]$.

4 RESULTS

In this section, we assess the performance of the proposed unfolded nonlocal model (UNLD) and compare it with other methods. In particular, we compare with the unfolding of the variational model that uses the classic data fidelity term from Equation (2) (UCLD). This comparison will prove the influence of the multi-head attention module. Also, we include two variational models, the conventional classic data (VCLD), and the one with the nonlocal data term (VNLD) from (Pereira-Sánchez et al., 2022). Both of these variational methods use total variation as regularization. Additionally, we include the bicubic (BIC) interpolation method in our evaluation.

The comparison is conducted by using sampling factors of 2 and 4, with standard blur deviations of 0.75 and 1.4, respectively. In both cases, the models have been evaluated for different realizations of noise, with standard deviations of 0, 5, 10, and 25. The parameters of the variational methods have been optimized using the image *penguin* from the validation set.

Table 1: Results for a sampling factor of 2, a blur level of 0.75, and various noise levels: 0, 5, 10, and 25. The comparison is based on PSNR and SSIM metrics, with the averages computed across the entire validation data. Bold numbers indicate the best metrics for the mean, and the italic numbers represent the second-best metrics. Notably, the proposed UNLD achieves the best metrics across all noise levels.

	BIC		**VCLD**		**VNLD**		**UCLD**		**UNLD**	
Noise	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
0	26,33	0,784	28,29	0,811	*29,40*	*0,871*	29,31	0,870	**29,57**	**0,876**
5	25,69	0,716	27,61	0,781	27,95	0,805	*28,39*	*0,822*	**28,49**	**0,826**
10	24,33	0,595	26,58	0,738	26,63	0,739	*27,01*	*0,758*	**27,30**	**0,775**
25	20,18	0,342	24,02	0,597	24,48	0,624	*24,79*	*0,648*	**24,86**	**0,654**

Table 2: Results for a sampling factor of 4, a blur level of 1.4, and various noise levels: 0, 5, 10, and 25. The comparison is based on PSNR and SSIM metrics, with the averages computed across the entire validation data. Bold numbers indicate the best metrics for the mean, and italic numbers represent the second-best metrics. Notably, the unfolded nonlocal data achieves the best metrics across all noise levels.

	BIC		**VCLD**		**VNLD**		**UCLD**		**UNLD**	
Noise	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
0	24,37	0,640	24,22	0,647	24,24	0,658	*24,77*	*0,670*	**24,81**	**0,672**
5	23,96	0,598	23,88	0,609	23,89	0,616	*24,34*	*0,630*	**24,37**	**0,633**
10	22,98	0,517	23,33	0,569	23,37	0,575	*23,77*	*0,590*	**23,88**	**0,601**
25	19,53	0,318	22,04	0,498	22,11	0,503	*22,35*	*0,515*	**22,38**	**0,518**

The quantitative results can be found in Tables 1 and 2 for sampling factors of 2 and 4, respectively. Our method achieves the best metrics for both sampling factors and for all noise realizations. Moreover, the second-best metrics are achieved by the other unfolding method. This corroborates the fact that unfolding methods are capable of learning the intrinsic and natural geometry of the images better than variational methods. Furthermore, the superiority of UNLD over UCLD demonstrates the importance of the multi-head attention module.

Moreover, the qualitative results are showcased in Figures 4 and 5. The visual comparison aligns with the quantitative results. In both examples, variational methods VNLD and VCLD provide oversmooth solutions while their unfolded versions UNLD and UCLD better preserve the image geometry and textures. However, in Figure 5 the proposed UNLD shows to be superior to UCLD in removing noise. When assessing the performance of both models, it becomes evident that the nonlocal model consistently outperforms its counterpart when subjected to varying degrees of noise.

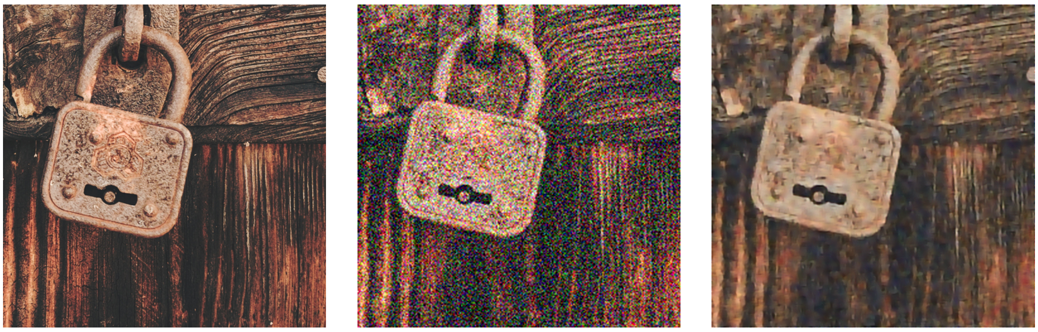
5 ABLATION STUDY

In this section, we evaluate the influence of different components of our system. First, we explore the efficacy and robustness of various residual networks architectures within our context. Four proposals of residual networks were considered to replace the proximal network *ProxNet*. These proposals can be divided into two different approaches. The first approach incorporates the argument in the proximal network provided by Equation (5), i.e., $u - \tau\nabla F_{\omega}$. Within this approach, we tested two similar architectures: one with batch normalization, denoted as Net_1, and the other without batch normalization, denoted as Net_2. The second approach introduces a concatenation of u and ∇F_{ω} instead. Within this approach, we tested two similar architectures: one without adding features to u and ∇F_{ω}, denoted as Net_3, and the other adding features, denoted as Net_4. Table 3 demonstrates that the second option delivers the best performance, indeed, this architecture is the one used in our model, as shown in Figure 2.

Next, we study the behaviour of the model under different samplings of the data. We evaluated the model with samplings 2, 3, 4 and 8 and blur 0.75, 1, 1.4, 1.8, respectively, all of them with absence of noise. The results of both unfolding models is shown in Table 4. As expected, the greater the sampling the lower metrics, but in every scenario the nonlocal model UNLD outperforms the classic model UCLD. This behaviour confirms the adaptability of the model to different data types.

Also, we have tested the performance of UNLD in a scenario with high noise level. Figure 7 illustrates that the model can reconstruct an image even with a noise level of 50 in the case of sampling 2 and blur 0.75.

Finally, we checked the influence of the selected number of stages in the iterative process. We trained UCLD and UNLD models with 3, 5, and 7 stages.



Reference BIC UNLD

Figure 7: Example showcasing the robustness to high noise level of UNLD. The images were generated from an input image with a sampling factor of 2, a blur level of 0.75, and a noise level of 50.

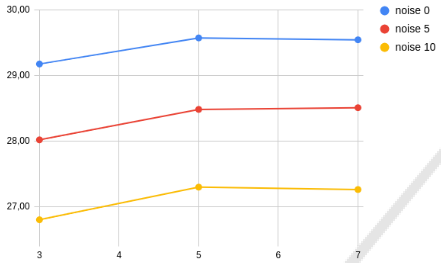


Figure 8: Evolution of PSNR as a function of the number of stages and noise values. Each PSNR value corresponds to the average value over the validation set.

Table 3: Results for sampling factor 2, blur level 0.75 and 3 stages, and different architectures for the *ProxNet*. The comparison is conducted in terms of average PSNR. The bold numbers indicate the best metrics. Notably, the Net_2 outperforms all other architectures.

ProxNet	***Net1***	***Net2***	***Net3***	***Net4***
PSNR	25,676	**29,160**	29,041	28,827

Results are reported in Figure 8 in terms of average PSNR. While from 3 to 5 there is a significant increase in performance, from 5 to 7 the performance is just slightly better. This behaviour is consistent across all noise levels. Therefore, we selected 5 stages as a good compromise between performance and computational cost.

Table 4: Average PSNR values of UCLD and UNLD with sampling factors of 2, 3, 4, and 8, and with blur levels of 0.75, 1, 1.4, and 1.8, respectively.

Sampling	**2**	**3**	**4**	**8**
UCLD	29,31	25,97	24,77	22,06
UNLD	29,57	26,08	24,81	22,12

6 CONCLUSIONS

In this work, we proposed an unfolding algorithm to extend a variational model that includes a nonlocal data fidelity term. After the minimization of the energy through the proximal gradient algorithm, we replaced all proximal, downsampling and upsampling operators with neural networks. While we wrote the nonlocal module in terms of a learning-based multi-head attention.

Experimental results demonstrated that learning-based methods can enhance the performance of variational models by leveraging prior knowledge from a dataset, rather than solely relying on traditional regularization terms. Additionally, we have shown that the benefits of self-similarity persist in the unfolded network when using multi-head attention, making the model more robust to noise.

As future work we plan to continue exploring such models with novel architectures that further improve performance. Furthermore, we aim to experiment with different energy terms from the wide array of models available in the literature.

ACKNOWLEDGEMENTS

This work is part of the MaLiSat project TED2021-132644B-I00, funded by MCIN/AEI/10.13039/501100011033/ and by the European Union NextGenerationEU/PRTR, and also of the MoLaLIP project PID2021-125711OB-I00, financed by MCIN/AEI/10.13039/501100011033/FEDER, EU. Also, we are grateful for the funding provided by the Conselleria de Fons Europeus, Universitat i Cultura (GOIB). The authors gratefully acknowledge the computer resources at Artemisa, funded by the EU

ERDF and Comunitat Valenciana and the technical support provided by IFIC (CSIC-UV).

REFERENCES

Babacan, S. D., Molina, R., and Katsaggelos, A. K. (2008). Total variation super resolution using a variational approach. In *2008 15th IEEE International Conference on Image Processing*, pages 641–644. IEEE.

Bell-Kligler, S., Shocher, A., and Irani, M. (2019). Blind super-resolution kernel estimation using an internal-gan. *Advances in Neural Information Processing Systems*, 32.

Chambolle, A. and Pock, T. (2016). An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319.

Combettes, P. L. and Wajs, V. R. (2005). Signal recovery by proximal forward-backward splitting. *Multiscale modeling & simulation*, 4(4):1168–1200.

Dai, T., Cai, J., Zhang, Y., Xia, S.-T., and Zhang, L. (2019). Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074.

Dong, W., Zhang, L., Shi, G., and Li, X. (2013). Nonlocally centralized sparse representation for image restoration. *IEEE Transactions on Image Processing*, 22(4):1620–1630.

Dong, W., Zhang, L., Shi, G., and Wu, X. (2011). Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857.

Duran, J., Buades, A., Coll, B., and Sbert, C. (2014). A nonlocal variational model for pansharpening image fusion. *SIAM Journal on Imaging Sciences*, 7(2):761–796.

Gilboa, G. and Osher, S. (2009). Nonlocal operators with applications to image processing. *Multiscale Modeling & Simulation*, 7(3):1005–1028.

Keys, R. (1981). Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1153–1160.

Lepcha, D. C., Goyal, B., Dogra, A., and Goyal, V. (2023). Image super-resolution: A comprehensive review, recent trends, challenges and applications. *Information Fusion*, 91:230–260.

Li, X. and Orchard, M. (2001). New edge-directed interpolation. *IEEE Transactions on Image Processing*, 10(10):1521–1527.

Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K. (2017). Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144.

Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., and Zeng, T. (2022). Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 457–466.

Monga, V., Li, Y., and Eldar, Y. C. (2021). Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44.

Pereira-Sánchez, I., Navarro, J., and Duran, J. (2022). What if image self-similarity can be better exploited in data fidelity terms? In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3697–3701. IEEE.

Rudin, L., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268.

Tai, Y., Yang, J., and Liu, X. (2017). Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155.

Wang, W., Li, F., and Ng, M. (2019). Structural similarity-based nonlocal variational models for image restoration. *IEEE Transactions on Image Processing*, 28(9):4260–4272.

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Z., Chen, Y., Shao, W., Li, H., and Zhang, L. (2022). Swinfuse: A residual swin transformer fusion network for infrared and visible images. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12.

Yue, L., Shen, H., Li, J., Yuan, Q., Zhang, H., and Zhang, L. (2016). Image super-resolution: The techniques, applications, and future. *Signal processing*, 128:389–408.

Zhang, K., Gao, X., Tao, D., and Li, X. (2012). Single image super-resolution with non-local means and steering kernel regression. *IEEE Transactions on Image Processing*, 21(11):4544–4556.

Zhang, L. and Wu, X. (2006). An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE Transactions on Image Processing*, 15(8):2226–2238.