IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 December 2023, accepted 13 January 2024, date of publication 17 January 2024, date of current version 25 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3355195

RESEARCH ARTICLE

Video Super-Resolution Using Plug-and-Play Priors

MATINA CH. ZERVA AND LISIMACHOS P. KONDI, (Senior Member, IEEE)
Department of Computer Science and Engineering, University of Ioannina, 45110 Ioannina, Greece

Corresponding author: Lisimachos P. Kondi (lkon@uoi.gr)

This work was supported in part by the Project "Dioni: Computing Infrastructure for Big-Data Processing and Analysis" co-funded by European Union through European Regional Development Fund (ERDF) under Grant 5047222; and in part by Greece through Operational Program "Competitiveness, Entrepreneurship and Innovation" under Grant NSRF 2014-2020.

ABSTRACT Video super-resolution is a fundamental task in computer vision, aiming to enhance the resolution and visual quality of low-resolution videos. Plug-and-Play Priors is one of the most widely used frameworks for solving computational imaging problems by integrating physical and learned models. Traditional approaches often rely on handcrafted priors, which are computationally expensive and may not generalize well to diverse video content. In this paper, we propose a novel approach for video super-resolution using Plug-and-Play Priors with motion estimation. By leveraging the power of deep learning and the flexibility of the Plug-and-Play framework, our method achieves promising results while maintaining computational efficiency. Experimental results on benchmark datasets demonstrate the superiority of our approach in terms of both quantitative metrics and visual quality.

INDEX TERMS Video, super-resolution, plug-and-play, motion estimation.

I. INTRODUCTION

Super-resolution (SR) involves the generation of high-resolution images or videos from their low-resolution counterparts, presenting a complex challenge within the field of computer vision. Its applications span diverse domains, including medical imaging, surveillance, remote sensing, and multimedia. The enhancement of resolution and visual quality in low-resolution videos is a particularly demanding task within super-resolution, as it aims to address issues like motion, subsampling, additive noise, and point spread function (PSF) blurring between frames in a low-resolution (LR) sequence [1].

Researchers have, over time, introduced various techniques and algorithms to tackle the intricate problem of super-resolution. Within a low-resolution sequence, each frame captures only a fraction of the original high-resolution (HR) image's information due to inherent degradations. However, frames with subpixel motion offer unique partial information of the original HR image. Consequently, with sufficient LR frames containing distinct information, the HR image can be reconstructed through digital image or video processing [2].

In the realm of video super-resolution, the accurate estimation of motion assumes a pivotal role. This process, essential for enhancing the resolution of low-resolution videos, involves aligning and consolidating information from multiple frames to generate a high-resolution output [3].

The deterioration of images in video super-resolution commonly involves the representation of a linear blur, motion, subsampling, and Gaussian noise. This is typically conceptualized through an observation model, assuming the acquisition of multiple low-resolution (LR) images through a specific process [4]. According to this model, the LR input images are obtained from the high-resolution (HR) original scene through operations such as warping, blurring, and downsampling. It is assumed that the HR image remains constant during the acquisition of several LR images [5].

Numerous algorithms and techniques have been proposed over the years to address the enhancement of resolution in both images and videos. The initial attempt was made by Tsai and Huang [6], utilizing the shifting property of the Fourier Transform and the aliasing relationship between the continuous Fourier transform (CFT) and the discrete Fourier

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

transform (DFT). Tekalp et al. [7] extended this method, incorporating a least squares approach to solve a system of equations and introducing a linear shift-invariant (LSI) blur point spread function (PSF). Kim et al. [8] further improved this technique by introducing a weighted least squares algorithm to handle noisy data. However, these methods are limited to scenarios where global motion is known in advance. Other spatial domain methods include the projection onto convex sets (POCS) approach introduced by Stark and Oskoui [9]. This method intersects convex constraint sets representing desirable image characteristics, such as positivity, bounded energy, fidelity to the data, and smoothness, with the HR image space. POCS has been extended to handle time-varying motion blur [10] and [11], using block matching or phase correlation for registration parameter estimation [10].

Stochastic methods form another category of resolution enhancement algorithms, with maximum likelihood (ML) and maximum a posteriori (MAP) approaches falling under this group [12]. The MAP estimation, employing an edge-preserving Huber-Markov random field image prior, is examined in [13], [14], and [15]. Resolution enhancement with simultaneous registration parameter estimation is proposed in [16], [17], [18], and [19]. This method uses a Gibbs-Markov random fields (GMRF) image prior with a local clique. The regularization parameter is crucial to the HR image reconstruction, and the L-curve method is employed for its estimation in [20], selecting the desired "L-corner" or point with maximum curvature on the L-curve.

A thorough comprehension of the point spread function (PSF) and precise registration of subpixel motion are crucial elements for reconstructing high-resolution (HR) images. However, in practical applications, ensuring accurate knowledge of these parameters is often challenging. Lee and Kang [21] presented a regularized adaptive HR reconstruction method that accommodates inaccurate subpixel registration. Assuming Gaussian noise for the registration error, with a standard deviation (STD) proportional to the registration error's magnitude, two approaches were developed to estimate the regularization parameter for each low-resolution (LR) frame (channel). Experimental results demonstrated the convergence of these methods to a unique global solution, although the synergy of these approaches was not extensively demonstrated. In [22], a hierarchical Bayesian framework was employed to address image restoration in the presence of partially known blurs, using stationary zero-mean white noise to model the unknown component of the PSF. Evidence analysis (EA) was utilized to propose two iterative algorithms resembling the regularized constrained total least squares filter and the linear minimum mean square-error filter [23], [24], [25].

Robust super-resolution techniques have been introduced in [23], [24], and [25], specifically designed to handle anomalies (data that deviate from the model). In [23], the iterative HR image acquisition method incorporates a median filter, showcasing robustness when errors from outliers are symmetrically distributed. However, determining whether bias arises from aliasing or outlier information requires a threshold, and the method's mathematical justification is not thoroughly examined. In [24] and [25], a robust super-resolution approach was proposed, incorporating the norm in both the regularization term and the measurement term of the penalty function. A robust regularization based on bilateral priors was introduced to accommodate various data and noise models, providing mathematical support for a "shift and add" approach related to norm minimization when relative motion is purely translational, and the PSF and decimation factor are common to all LR images.

Subsequently, the methodology introduced in [16], [17], [18], and [19] was extended to handle scenarios where low-resolution (LR) frames suffer from additive white Gaussian noise (AWGN) with varying variances in each frame [18]. The fundamental idea involves adjusting the residual term of the cost function by the inverse of the variance for each frame (channel) when AWGN with distinct variances is the sole additional noise source in the LR images. Moreover, to mitigate errors introduced by other types of noise during the resolution enhancement reconstruction phase, weighting should be applied to each channel. Additionally, He and Kondi proposed an image super-resolution algorithm in [4] that takes into account imprecise estimates of registration parameters and the point spread function. These inaccurate estimates, coupled with additive Gaussian noise in the LR image sequence, result in varying noise levels for each frame. In the proposed algorithm, LR frames are adaptively weighted based on their reliability, and the regularization parameter is simultaneously estimated, assuming a translational motion model.

Image super-resolution using deep learning has gained significant attention due to its ability to generate high-resolution images from low-resolution inputs. Various deep learning architectures and methods have been proposed for image super-resolution. Among them there is SRCNN [26], FSRCNN [27], ESPCN [28], VDSR [29], SRGAN [30], EDSR [31], RCAN [32], IDPT [33] and DBTC [34]. The emergence of deep learning has showcased the substantial potential of convolutional neural networks (CNNs) in video super-resolution. Tao et al. [35] introduced a CNN-based framework for video super-resolution that effectively harnessed both spatial and temporal information. Their network learned spatio-temporal dependencies in videos, leading to improved resolution and visual quality.

To further reinforce the performance of CNN-based video super-resolution, researchers explored the incorporation of recurrent neural networks (RNNs) to model long-term temporal dependencies. Caballero et al. [36] proposed a recurrent video super-resolution network (RVSR) that integrated a recurrent structure to capture temporal information across frames. The recurrent connections facilitated a better understanding of temporal dynamics, resulting in superior super-resolution outcomes.

In addition to approaches based on deep learning, there have been endeavors to exploit alternative priors and constraints in video super-resolution. For example, Huang et al. [37] proposed a method that incorporates non-local self-similarity to harness redundancy within video frames. By enforcing self-similarity constraints, their approach achieved enhanced reconstruction quality and reduced artifacts.

Another direction in video super-resolution involves the utilization of generative adversarial networks (GANs). Ledig et al. [30] introduced an SRGAN-based framework for single-image super-resolution, later extended to address video super-resolution. With a generator-discriminator architecture, SRGAN effectively captured high-frequency details, resulting in visually pleasing super-resolved videos.

Moreover, researchers have explored the fusion of multiple frames to enhance the resolution of video sequences. Huang et al. [38] proposed a multi-frame video super-resolution method that combines a temporal fusion module with a spatial attention mechanism. By selectively fusing information from multiple frames, their approach achieved improved super-resolution results.

It is crucial to emphasize that the assessment of video super-resolution methods relies significantly on evaluation metrics and datasets. The use of benchmark datasets, such as Vimeo-90K [39] and REDS [40], has facilitated fair comparisons and benchmarking of various algorithms.

The Plug-and-Play Priors (PPP) framework is recognized as one of the extensively used methodologies for addressing computational imaging challenges through the integration of physical and learned models. PPP takes advantage of high-fidelity physical sensor models and robust machine learning techniques for data pre-modeling, incorporating cutting-edge reconstruction algorithms. PPP algorithms follow a cycle of minimizing data fidelity terms to uphold data consistency and enforcing learned regularization through image denoising [41]. Recent achievements of PPP algorithms span applications in biomicroscopy, computed tomography, magnetic resonance imaging, and joint ptychotomography [42].

This article proposes a video super-resolution method, based on the Plug-and-Play (PnP) framework. To our knowledge this is the first attempt to use PnP framework in video super-resolution, using motion estimation.

II. PLUG-AND-PLAY PRIORS

Plug-and-Play Priors (PPP) stands out as a widely adopted framework that integrates physical and learned models to address computational imaging challenges. It is a robust framework that merges conventional optimization techniques with modern denoising methods and priors to efficiently tackle inverse problems [43]. Initially introduced by Venkatakrishnan et al. [42], PPP has garnered significant attention across various domains of computer vision and image processing. This literature review delves into key contributions that have shaped the development and application of PPP.

The original PPP framework proposed by Venkatakrishnan et al. [42] showcased its efficacy in solving inverse problems, such as image denoising and deblurring. Their work demonstrated that by alternately applying denoising and data fidelity steps, PPP achieves state-of-the-art results. The denoising step employs robust algorithms like Non-Local Means (NLM) or Block-matching and 3D filtering (BM3D) [44] to eliminate noise and enhance image quality. The data fidelity step ensures consistency between the denoised image and the observed measurements. Despite the original formulation relying on ADMM [45], PPP proves equally effective when combined with other proximal algorithms like primal-dual splitting (PDS) [46] and fast iterative shrinkage/thresholding algorithm (FISTA) [47].

To further enhance denoising capabilities within PPP, Zhang et al. [48] introduced a deep denoising network named DnCNN. Integrating DnCNN into the PPP framework demonstrated its effectiveness in tasks such as image super-resolution and inpainting. The utilization of deep neural networks within PPP provides a more flexible and potent denoising tool, surpassing traditional handcrafted denoisers in performance.

Ghassab and Bouguila [49] explored the utilization of a Student-t mixture model as a promising tool for the reconstruction of video super-resolution. The Student-t mixture model, renowned for its heavy tail, was deemed robust and well-suited for the prior of video frame patches, offering a mixture model with a rich log-likelihood for information retrieval. Edge-preserving filtering was implemented to address potential data uncertainties and preserve areas with abrupt lighting changes in video frames. The Plug-and-Play Priors (PPP) structure was subsequently employed to integrate the Student-t mixture prior model and edge-preserving filtering into the super-resolution algorithm. Empirical evaluations conducted on various video frame sets, demonstrated the effectiveness of the proposed algorithm. Comparisons with eight other state-of-the-art super-resolution methods affirmed that the proposed framework generally outperforms others across different super-resolution scales, even in the absence of leveraging motion estimation to exploit frame correlations.

PnP-ADMM is widely recognized for its efficiency and fast empirical convergence within the realm of frequently employed operators in computational imaging. However, it demands the computation of the proximal map, in contrast to PnP-FISTA, which solely requires the computation of the gradient ∇g. While the gradient is theoretically less complex than the proximal map, numerous applications enable the efficient computation or approximation of the proximal map. General techniques such as conjugate gradient or specialized methods, particularly when the forward model incorporates a spatial blurring operator computed through fast Fourier transform (FFT), can be employed for this purpose [50].

The incorporation of an extra state variable, employed as an initiation for the proximal minimization problem, streamlines this procedure. An iterative solver, commencing from this initialization, performs a series of steps to estimate the minimization effectively. This state variable also converges with the outer loop, resulting in decreased computational requirements through partial updates while maintaining the accuracy of the final solution [51].

In the research work reported in [52], scientists introduce a straightforward and robust super-resolution framework applicable to individual images and easily adaptable to videos. The foundation of the framework is rooted in the observation that the denoising of both images and videos can be effectively accomplished through various methods. By leveraging the Plug-and-Play-Prior framework and adopting the Regularization-by-Denoising (RED) approach, the researchers illustrate how denoisers can be harnessed to tackle both Single-Image Super-Resolution (SISR) and Video Super-Resolution (VSR) challenges using a unified formulation. Instead of incorporating motion estimation between frames, the VBM3D video denoiser was employed in this approach.

Our paper attempts to introduce a PnP method for video super-resolution, using motion estimation, which has not been done yet.

III. OUR METHOD

The acquisition model we are assuming is:

$$\mathbf{y} = A\mathbf{x} + \varepsilon, \tag{1}$$

where:

- $\mathbf{y}$ is the full set of LR frames, described as $\mathbf{y} = [\,\mathbf{y}_1{}^T, \mathbf{y}_2{}^T, \ldots, \mathbf{y}_p{}^T\,]^T$, where $\mathbf{y}_k, k = 1, 2, \ldots, p$ are the p LR images. Each observed LR image is of size $N_1 \times N_2$. Let the kth LR image be denoted in lexicographic notation as $\mathbf{y}_k = [\,y_{k,1}, y_{k,2}, \ldots, y_{k,M}\,]^T$, for $k = 1, 2, \ldots, p$ and $M = N_1N_2$.
- $\mathbf{x}$ is the desired HR image, of size $L_1N_1 \times L_2N_2$, written in lexicographical notation as the vector $\mathbf{x} = [\,x_1, x_2, \ldots, x_N\,]^T$, where $N = L_1N_1L_2N_2$ and L_1 and L_2 represent the up-sampling factors in the horizontal and vertical directions, respectively. $\mathbf{x}$ is the ideal un-degraded image that is sampled at or above the Nyquist rate from a continuous scene which is assumed to be band-limited.
- $\varepsilon = [\,\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p\,]^T$, where ε_k is the noise vector for frame k and contains independent zero-mean Gaussian random variables.
- $A = [\,A_1, A_2, \ldots, A_p\,]^T$ is the degradation matrix which performs the operations of blur, motion and subsampling.

Assuming that each LR image is corrupted by additive noise, we can then represent the observation model as [5]:

$$\mathbf{y}_k = A_k\mathbf{x} + \varepsilon_k \text{ for } 1 \le k \le p \tag{2}$$

where

$$A_k = SB_kM_k. \tag{3}$$

M_k is a warp matrix of size $L_1N_1L_2N_2 \times L_1N_1L_2N_2$, B_k represents a $L_1N_1L_2N_2 \times L_1N_1L_2N_2$ blur matrix, and S is a $N_1N_2 \times L_1N_1L_2N_2$ subsampling matrix. In our case $B_k = I$, since we assumed no added blur on video frames.

The goal is to find the estimate $\hat{\mathbf{x}}$ of the HR image $\mathbf{x}$ from the p LR images y_k by minimizing the cost function

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x}) \text{ with } f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}), \tag{4}$$

where $g(\mathbf{x}) = \sum_{k=1}^{p} \frac{1}{2}\|A_k\mathbf{x} - \mathbf{y}_k\|_2^2$ is the "fidelity to the data" term, and $h(\mathbf{x})$ is the regularization term, which offers some prior knowledge about $\mathbf{x}$. In this work, we utilize the Plug-and-Play Prior methodology, where $h(\mathbf{x})$ is not explicitly defined. Instead, the ADMM algorithm is modified so that the proximal operator that depends on $h(\mathbf{x})$ is replaced by a denoising neural network [53].

We next outline the steps of the proposed algorithm.

1) The first step of our algorithm is to evaluate the term M_k from Eq. (3), by using optical flow motion estimation. The motion estimation method used is a popular optical flow method, called the Farneback algorithm, named after its creator, Gunnar Farneback. The Farneback algorithm generates an image pyramid, where each level has a lower resolution compared to the previous level. The Farneback method employs a dense approach, meaning it estimates the motion vector for every pixel in the image. The algorithm consists of the following steps [54]:
   a) *Preprocessing*: The input frames are preprocessed to enhance their quality. Preprocessing steps include noise reduction, image denoising, and color space conversion.
   b) *Image pyramids*: The Farneback algorithm constructs a Gaussian pyramid for each frame. This involves creating a series of downsampled versions of the original image, forming a hierarchy of images with decreasing resolution. The pyramids enable capturing motion at multiple scales, improving the accuracy of the optical flow estimation.
   c) *Optical flow estimation*: For each level of the pyramid, the Farneback algorithm computes the optical flow using a combination of polynomial expansion and spatial filtering. It estimates the local flow vectors by calculating the phase difference between the polynomials corresponding to neighboring image patches.
   d) *Upsampling and refinement*: Once the optical flow is computed at the coarsest level of the pyramid, it is successively refined by upsampling the flow field and incorporating the local information from higher-resolution levels. This refinement process improves the accuracy of the

flow estimation, particularly for small and fast-moving objects.

The result of the Farneback method is a dense optical flow field, where each pixel has an associated motion vector. These vectors represent the direction and magnitude of the motion of objects in the scene between consecutive frames.

We assume that one of the LR images, $\mathbf{y}_{mid}$ (typically the middle one), is produced from the HR image $\mathbf{x}$, by applying only downsampling, without motion shift. Thus, $M_{mid} = I$. Optical flow is calculated between $\mathbf{y}_{mid}$ and the rest of the LR images. Following that, we get M_k for the remaining $p-1$ images.

2) The second step of our algorithm is based on the PnP-ADMM method. Specifically, we run PnP-ADMM, following the steps described in Algorithm 1 until convergence, where $\mathbf{x}^0$ is the initial value of the HR image, obtained from $\mathbf{y}_{mid}$ multiplied by the pseudo-inverse of $\mathbf{A}_{mid}$, followed by denoising using DnCNN, while D is the image denoising operator (neural network) and g is defined as $g(\mathbf{x}) = \sum_{k=1}^{p} \frac{1}{2}\|A_k\mathbf{x} - \mathbf{y}_k\|_2^2$.

Algorithm 1 : PnP-ADMM [42]

1: $\mathbf{u}^0 = \mathbf{0}, \mathbf{x}^0,$ and $\gamma > 0$
2: **for** $k = 1, 2, \ldots, t$ **do**
3: $\quad \mathbf{z}^k \leftarrow prox_{\gamma g}(\mathbf{x}^{k-1} - \mathbf{u}^{k-1})$
4: $\quad \mathbf{x}^k \leftarrow D(\mathbf{z}^k + \mathbf{u}^{k-1})$
5: $\quad \mathbf{u}^k \leftarrow \mathbf{u}^{k-1} + (\mathbf{z}^k - \mathbf{x}^k)$
6: **end for**
7: **return** $\mathbf{x}^t$

One important property of ADMM is that it does not explicitly require knowledge of $g(\mathbf{x})$ or their gradients, relying instead on the proximal operator, which is defined as:

$$prox_{\gamma g}(\mathbf{x}) := \arg\min_{\mathbf{x}\in\mathbb{R}^N} \{\frac{1}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 + \gamma g(\mathbf{x})\}. \quad (5)$$

IV. PROOF OF CONVERGENCE

The crucial conceptual observation lies in the fact that PnP algorithms incorporating black-box denoisers often struggle to address optimization problems. In other words, while the original ADMM algorithm effectively solves the optimization problem, the introduction of a black-box denoiser, denoted as D, disrupts this process by eliminating a corresponding function h for minimization. Specifically, the numerical assessment of widely employed denoisers, such as BM3D and DnCNN, demonstrates that their Jacobians lack symmetry, suggesting that these denoisers do not function as either gradient descent steps or proximal maps [55].

Nevertheless, it remains feasible to establish a criterion for the converged solution in PnP by employing a consensus equilibrium formulation, as proposed by [56].

$$\mathbf{x} = G(\mathbf{x} - \mathbf{u}) \text{ and } \mathbf{x} = D(\mathbf{x} + \mathbf{u}), \quad (6)$$

where $G := prox_g$ and $\mathbf{x}$, $\mathbf{u}$ are the converged values of PnP-ADMM.

FIGURE 1. Original "Calendar" Image.

FIGURE 2. Original "City" Image.

Notably, within the consensus equilibrium expression in (6), $\mathbf{x}$ represents the final reconstruction and $\mathbf{u}$ can be construed as noise, eliminated by the denoiser in $\mathbf{x} = D(\mathbf{x}+\mathbf{u})$ on one hand and counterbalanced by the fidelity to the data effect in $\mathbf{x} = G(\mathbf{x} - \mathbf{u}\mathbf{x})$ on the other. To derive (6), it is important to recognize that the fixed points $\mathbf{z}$, $\mathbf{x}$, and $\mathbf{u}$ of the PnP-ADMM iteration satisfy

$$\mathbf{z} = G(\mathbf{x} - \mathbf{u}), \mathbf{x} = D(\mathbf{z} + \mathbf{u}), \mathbf{u} = \mathbf{u} + \mathbf{z} - \mathbf{x}. \quad (7)$$

From the last equation we conclude that $\mathbf{x} = \mathbf{z}$, which leads directly to (6). Also, the first-order optimality condition for the minimization problem $\mathbf{x} = G(\mathbf{x} - \mathbf{u}) = prox_{\gamma g}(\mathbf{x} - \mathbf{u})$ is $0 = \mathbf{x} - (\mathbf{x} - \mathbf{u}) + \gamma\nabla g(\mathbf{x})$, so $\mathbf{u} = -\gamma\nabla g(\mathbf{x})$.

The application of monotone operator theory, as outlined in [57], allows for the illustration of the convergence of PnP algorithms. In this approach, the initial phase involves identifying a fixed point for a high-dimensional operator that can be iteratively used to discover a solution, provided the appropriate assumptions are met. In the proof of PnP-ADMM convergence presented in [56] and [58], the initial step is

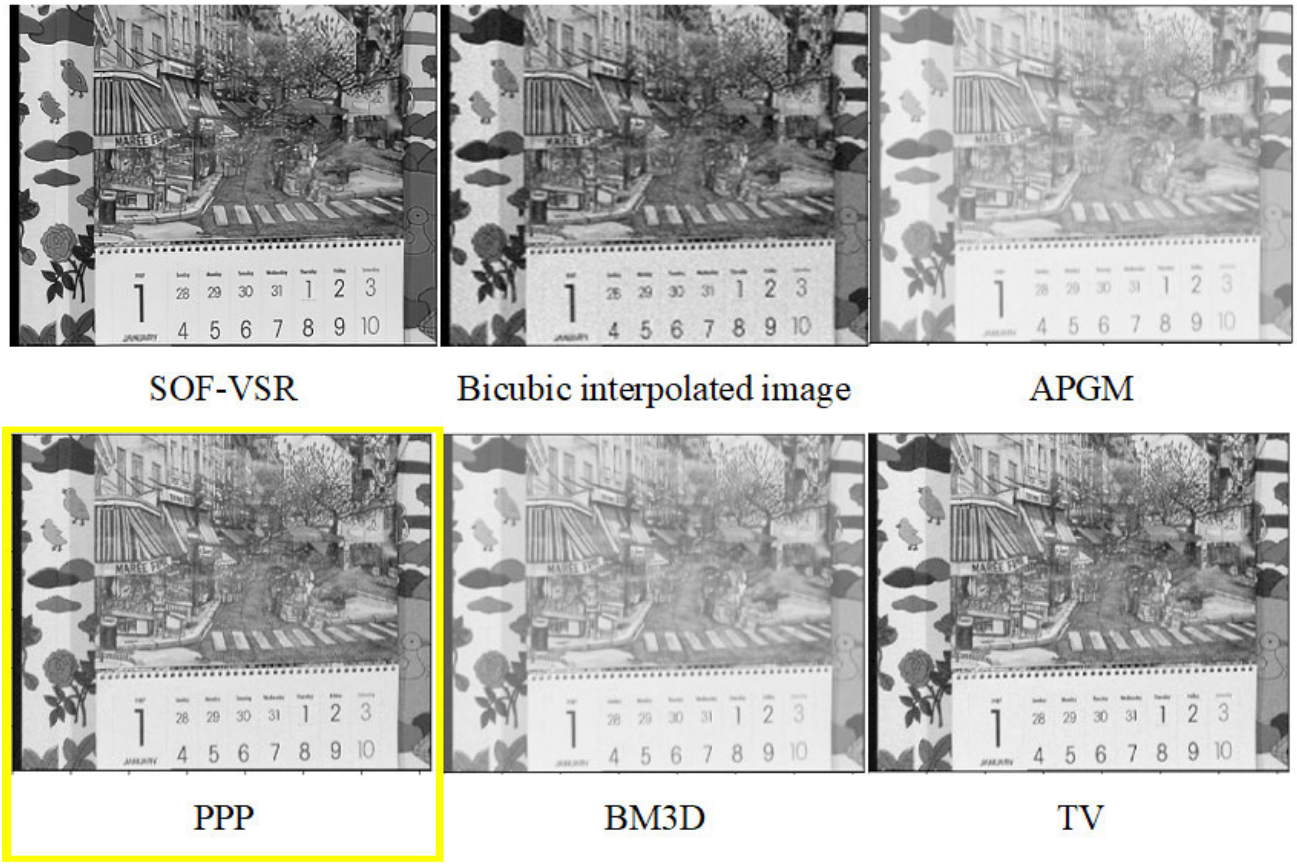


FIGURE 3. Result of "Calendar" Image.

to establish a one-to-one correspondence between the fixed points of PnP-ADMM and those of the operator:

$$T := (2G - I)(2D - I). \tag{8}$$

After a linear coordinate transformation, Algorithm 1 is essentially identical to the Mann iterations of T, expressed as $\mathbf{v}^k \leftarrow \frac{1}{2}\mathbf{v}^{k-1} + \frac{1}{2}T(\mathbf{v}^{k-1})$ [56]. This results in linear convergence towards a unique fixed point when T functions as a contraction, a condition satisfied when g is strongly convex and $R := I - D$ serves as a suitably strong contraction [58]. Weaker conditions lead to sublinear convergence, reaching a potentially non-unique fixed point [59]. Additional notable theoretical findings on PnP-ADMM encompass its convergence for implicit proximal operators [43], applicability with bounded denoisers [60], and suitability for linearized Gaussian mixture model (GMM) denoisers [50]. Even CNN-based denoisers can be trained to meet these contractive, non-expansive, or Lipschitz conditions through the implementation of spectral normalization techniques [58], [61]. Conversely, when g exhibits only mild convexity and the denoiser D is strongly non-expansive, the iteration converges sublinearly towards its fixed point [62].

V. RESULTS

We implemented our PnP method in SCICO [63], which is an open source library for computational imaging that includes implementations of PnP algorithms.

We conducted extensive experiments on benchmark subsets "calendar" and "city", from Vid4 dataset to evaluate the performance of our proposed method. Specifically, we used $p = 3$ frames, with the second in order being the zero-motion image, and we added Gaussian noise with a standard deviation of 0.02. The up-sampling factors in the horizontal and vertical directions were $L_1 = L_2 = 4$. For the denoising operator D, the DnCNN neural network [48] was used, as it was pre-trained by SCICO. Finally, we compared our results against other successful video super-resolution techniques in terms of both quantitative metrics, such as PSNR (Peak signal-to-noise ratio), and visual quality.

The results that were compared to ours were acquired by APGM (accelerated proximal gradient method) [45], BM3D (Block-matching and 3D filtering) [64], Total Variation [44], bicubic, SOF-VSR (Super-resolving Optical Flow for Video Super-Resolution) [65] and EDVR (Video Restoration with Enhanced Deformable Convolutional Networks) [66].

Table 1 show PSNR results for the two datasets for all the methods tested. It can be seen that average PSNR for our method is 22.86 dB for "Calendar" dataset and 25.74 dB for "City" dataset, while all the other methods have lower values. The highest PSNR values for Frame 17 of "Calendar" (Fig. 1) and Frame 14 of "City" (Fig. 2).

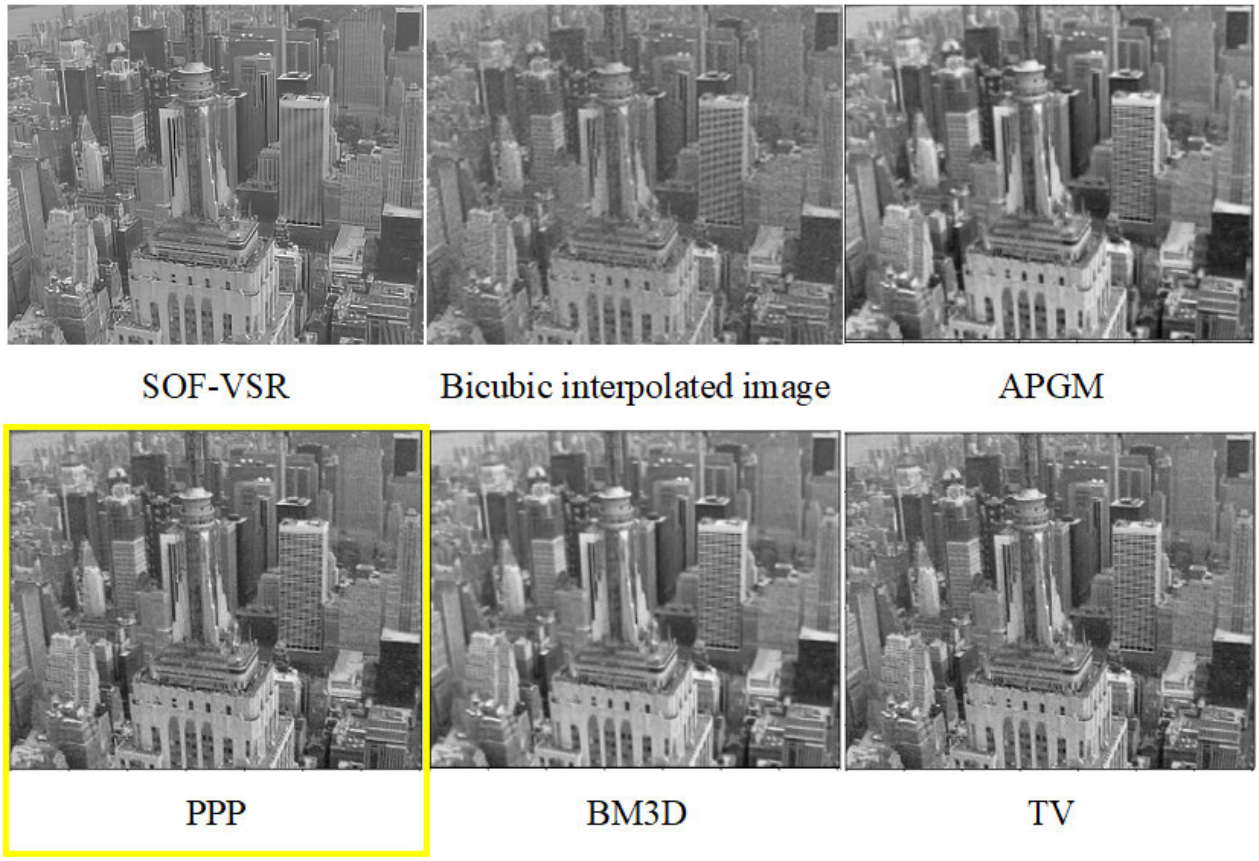


FIGURE 4. Result of "City" Image.

TABLE 1. Average PSNR values for the two datasets for all the methods.

	Ours	APGM	BD3M	TV	Bicubic	SOF-VSR	EDVR
Calendar	**22.86**	20.58	21.09	21.66	19.36	21.69	21.69
City	**25.74**	23.91	24.37	25.13	22.61	25.62	25.51

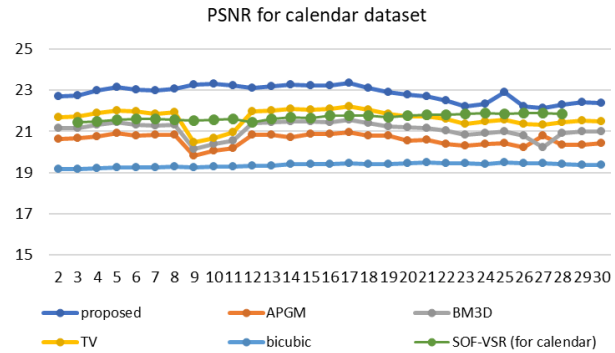


FIGURE 5. PSNR values of the 29 images of "Calendar" dataset for all the methods tested.

Apart from the numerical results, the visual proofs are also in favor of our method, since the super-resolved pictures are clearer than the pictures produced with the other methods. Examples of the results can be seen in Fig. 3 and Fig. 4, which are the results for the original Fig. 1 and Fig. 2. It should be noted that there is no image result for EDVR, since results were taken from [67].

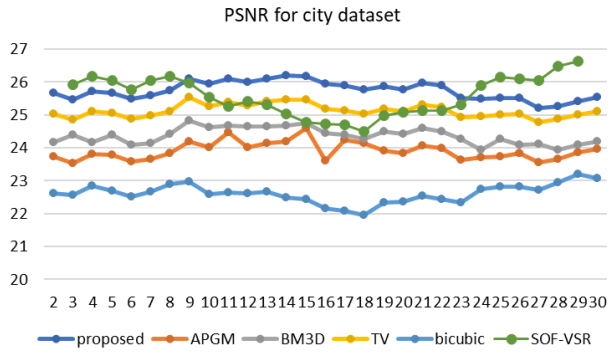


FIGURE 6. PSNR values of the 29 images of "City" dataset for all the methods tested.

Fig. 5 and Fig. 6 show the results in terms of PSNR for the images of "Calendar" and "City" datasets accordingly, for all the methods tested.

Frames 9, 10 and 11 from "Calendar" dataset show a much lower PSNR for APGM, BM3D, and TV, because these images have greater difference from the others and these methods are more motion-sensitive than ours.

The results demonstrate the superior performance of our approach in terms of reconstruction accuracy and preservation of fine details and textures. It is worth mentioning that our method needs no training, since DnCNN is pre-trained. Finally, the runtime of our method per frame is 12 seconds, ran in Google Colab with T4 GPU.

VI. CONCLUSION

PnP techniques have established themselves as a standard tool for computational imaging since their introduction in 2013. They have been utilized in a remarkable variety of applications that provide cutting-edge performance. They were arguably the first practical approach to integrating learned models with imaging physics to solve inverse imaging issues when they were first introduced. The ease with which they can be implemented was a major factor in their rapid popularity. Since then, alternative strategies have emerged that, in some cases, result in improved reconstruction performance; however, this is achieved at the expense of a potentially time-consuming and data-dependent application-specific training procedure. In this paper, we proposed a PnP method for video super-resolution (resolution enhancement) with motion estimation. The convergence property of the proposed algorithm is analyzed in detail. More importantly, experimental results show the validity of our algorithm and its superiority compared to other state-of-the-art methods.

ACKNOWLEDGMENT

The publication of the article in OA mode was financially supported by HEAL-Link.

REFERENCES

[1] S. Borman and R. Stevenson, "Spatial resolution enhancement of low-resolution image sequences—A comprehensive review with directions for future research," Lab. Image Sugnal Anal., Univ. Notre Dame, Notre Dame, IN, USA, Tech. Rep., 1998.
[2] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal Process.*, vol. 128, pp. 389–408, Nov. 2016.
[3] P. Torr and A. Zisserman, "Feature based methods for structure and motion estimation," in *Proc. Vis. Algorithms, Theory Pract.*, Mar. 2000, pp. 278–294.
[4] H. He and L. P. Kondi, "An image super-resolution algorithm for different error levels per frame," *IEEE Trans. Image Process.*, vol. 15, no. 3, pp. 592–603, Mar. 2006.
[5] S. C. Park, M. K. Park, and M. Gi Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, May 2003.
[6] R. Y. Tsai and T. S. Huang, "Multiframe image restoration and registration," in *Proc. Adv. Comput. Vis. Image Process.*, 1984, pp. 317–339.
[7] A. M. Tekalp, M. K. Ozkan, and M. I. Sezan, "High-resolution image reconstruction from lower-resolution image sequences and space-varying image restoration," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 3, Apr. 1992, pp. 169–172.
[8] S. P. Kim, N. K. Bose, and H. M. Valenzuela, "Recursive reconstruction of high resolution image from noisy undersampled multiframes," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 6, pp. 1013–1027, Jun. 1990.
[9] H. Stark and P. Oskoui, "High-resolution image recovery from image-plane arrays, using convex projections," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 6, no. 11, pp. 1715–1726, 1989.
[10] A. J. Patti, M. I. Sezan, and A. M. Tekalp, "Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time," *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1064–1076, Aug. 1997.
[11] P. E. Eren, M. I. Sezan, and A. M. Tekalp, "Robust, object-based high-resolution image reconstruction from low-resolution video," *IEEE Trans. Image Process.*, vol. 6, no. 10, pp. 1446–1451, Oct. 1997.
[12] B. C. Tom, N. P. Galatsanos, and A. K. Katsaggelos, "Reconstruction of a high resolution image from multiple low resolution images," in *Super-Resolution Imaging*. Boston, MA, USA: Springer, 2000, pp. 73–105.
[13] R. R. Schultz and R. L. Stevenson, "A Bayesian approach to image expansion for improved definition," *IEEE Trans. Image Process.*, vol. 3, no. 3, pp. 233–242, May 1994.
[14] R. R. Schultz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Trans. Image Process.*, vol. 5, no. 6, pp. 996–1011, Jun. 1996.
[15] R. R. Schultz and R. L. Stevenson, "Improved definition video frame enhancement," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, May 1995, pp. 2169–2172.
[16] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint MAP registration and high-resolution image estimation using a sequence of undersampled images," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1621–1633, Dec. 1997.
[17] L. Kondi, D. Scribner, and J. Schuler, "A comparison of digital image resolution enhancement techniques," *Proc. SPIE*, vol. 4719, pp. 220–229, Jul. 2002.
[18] H. He and L. P. Kondi, "Resolution enhancement of video sequences with simultaneous estimation of the regularization parameters," *Proc. SPIE*, vol. 5022, pp. 1123–1133, May 2003.
[19] H. He and L. P. Kondi, "MAP based resolution enhancement of video sequences using a Huber–Markov random field image prior model," in *Proc. Int. Conf. Image Process.*, vol. 2, Sep. 2003, pp. 933–936.
[20] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1998.
[21] E. S. Lee and M. Gi Kang, "Regularized adaptive high-resolution image reconstruction considering inaccurate subpixel registration," *IEEE Trans. Image Process.*, vol. 12, no. 7, pp. 826–837, Jul. 2003.
[22] N. P. Galatsanos, V. Z. Mesarovic, R. Molina, and A. K. Katsaggelos, "Hierarchical Bayesian image restoration from partially known blurs," *IEEE Trans. Image Process.*, vol. 9, no. 10, pp. 1784–1797, Oct. 2000.
[23] A. Zomet, A. Rav-Acha, and S. Peleg, "Robust super-resolution," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Dec. 2001, pp. 645–650.
[24] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Robust shift and add approach to superresolution," *Proc. SPIE*, vol. 5203, pp. 121–130, Oct. 2003.
[25] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.
[26] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016, doi: 10.1109/TPAMI.2015.2439281.
[27] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 391–407.
[28] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2016, pp. 1874–1883, doi: 10.1109/CVPR.2016.207.
[29] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.
[30] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[31] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.
[32] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.
[33] J. Shi, Y. Wang, S. Dong, X. Hong, Z. Yu, F. Wang, C. Wang, and Y. Gong, "IDPT: Interconnected dual pyramid transformer for face super-resolution," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, L. D. Raedt, Ed., Jul. 2022, pp. 1306–1312, doi: 10.24963/ijcai.2022/182.
[34] J. Shi, Y. Wang, Z. Yu, G. Li, X. Hong, F. Wang, and Y. Gong, "Exploiting multi-scale parallel self-attention and local variation via dual-branch transformer-CNN structure for face super-resolution," *IEEE Trans. Multimedia*, 2023.

[35] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4482–4490.

[36] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5572–5581.

[37] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.

[38] J. B. Huang, A. Singh, N. Ahuja, and E. Learned-Miller, "Multi-scale convolutional neural networks for high-resolution image inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2016, pp. 1110–1118.

[39] S. W. Oh, J. Y. Lee, K. Sunkavalli, S. J. Kim, and I. S. Kweon, "Video super-resolution via bidirectional recurrent convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 678–694.

[40] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3883–3891.

[41] A. Brifman, Y. Romano, and M. Elad, "Turning a denoiser into a super-resolver using plug and play priors," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1404–1408.

[42] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Dec. 2013, pp. 945–948.

[43] S. Sreehari, S. V. Venkatakrishnan, B. Wohlberg, G. T. Buzzard, L. F. Drummy, J. P. Simmons, and C. A. Bouman, "Plug-and-play priors for bright field electron tomography and sparse interpolation," *IEEE Trans. Comput. Imag.*, vol. 2, no. 4, pp. 408–423, Dec. 2016.

[44] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image restoration by sparse 3D transform-domain collaborative filtering," *Proc. SPIE*, vol. 6812, Feb. 2008, Art. no. 681207.

[45] U. S. Kamilov, H. Mansour, and B. Wohlberg, "A plug-and-play priors approach for solving nonlinear imaging inverse problems," *IEEE Signal Process. Lett.*, vol. 24, no. 12, pp. 1872–1876, Dec. 2017.

[46] S. Ono, "Primal-dual plug-and-play image restoration," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1108–1112, Aug. 2017.

[47] 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, Jan. 2009.

[48] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Feb. 2017.

[49] V. K. Ghassab and N. Bouguila, "Plug-and-play video super-resolution using edge-preserving filtering," *Comput. Vis. Image Understand.*, vol. 216, Feb. 2022, Art. no. 103359.

[50] A. M. Teodoro, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "A convergent image fusion algorithm using scene-adapted Gaussian-mixture-based denoising," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 451–463, Jan. 2019.

[51] V. Sridhar, X. Wang, G. T. Buzzard, and C. A. Bouman, "Distributed iterative CT reconstruction using multi-agent consensus equilibrium," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1153–1166, 2020.

[52] A. Brifman, Y. Romano, and M. Elad, "Unified single-image and video super-resolution via denoising algorithms," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 6063–6076, Dec. 2019.

[53] U. S. Kamilov, C. A. Bouman, G. T. Buzzard, and B. Wohlberg, "Plug-and-play methods for integrating physical and learned models in computational imaging: Theory, algorithms, and applications," *IEEE Signal Process. Mag.*, vol. 40, no. 1, pp. 85–97, Jan. 2023.

[54] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. 13th Scand. Conf. Image Anal. (SCIA)*, vol. 2749, Jun. 2003, pp. 363–370.

[55] E. T. Reehorst and P. Schniter, "Regularization by denoising: Clarifications and new interpretations," *IEEE Trans. Comput. Imag.*, vol. 5, no. 1, pp. 52–67, Mar. 2019.

[56] G. T. Buzzard, S. H. Chan, S. Sreehari, and C. A. Bouman, "Plug-and-play unplugged: Optimization-free reconstruction using consensus equilibrium," *SIAM J. Imag. Sci.*, vol. 11, no. 3, pp. 2001–2020, Jan. 2018.

[57] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Cham, Switzerland: Springer, 2017.

[58] E. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin, "Plug-and-play methods provably converge with properly trained denoisers," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, Jun. 2019, pp. 5546–5557.

[59] Y. Sun, Z. Wu, X. Xu, B. Wohlberg, and U. S. Kamilov, "Scalable plug-and-play ADMM with convergence guarantees," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 849–863, 2021.

[60] Z. Chen, S. Gu, G. Lu, and D. Xu, "Exploiting intra-slice and inter-slice redundancy for learning-based lossless volumetric image compression," *IEEE Trans. Image Process.*, vol. 31, pp. 1697–1707, 2022.

[61] Y. Sun, J. Liu, X. Xu, B. Wohlberg, and U. S. Kamilov, "Block coordinate regularization by denoising," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2019, p. 382–392.

[62] Y. Sun, B. Wohlberg, and U. S. Kamilov, "An online plug-and-play algorithm for regularized image reconstruction," *IEEE Trans. Comput. Imag.*, vol. 5, no. 3, pp. 395–408, Sep. 2019.

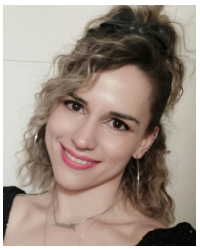
[63] T. Balke, F. Davis, C. Garcia-Cardona, S. Majee, M. McCann, L. Pfister, and B. Wohlberg, "Scientific computational imaging code (SCICO)," *J. Open Source Softw.*, vol. 7, no. 78, p. 4722, Oct. 2022.

[64] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D, Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, Nov. 1992.

[65] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, and W. An, "Deep video super-resolution using HR optical flow estimation," *IEEE Trans. Image Process.*, vol. 29, pp. 4323–4336, 2020.

[66] X. Wang, K. Yu, K. C. Chan, C. Dong, and C. C. Loy. (2020). *BasicSR*. [Online]. Available: https://github.com/xinntao/BasicSR

[67] Y. Li, P. Jin, F. Yang, C. Liu, M.-H. Yang, and P. Milanfar, "COMISR: Compression-informed video super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2523–2532.



MATINA CH. ZERVA was born in Greece, in 1989. She received the B.Sc. degree in computer science from the Department of Computer Science and Engineering, University of Ioannina, Greece, and the M.Sc. degree in applied mathematics and computer science from the Department of Mathematics. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, University of Ioannina. She is also a Researcher with the Department of Computer Science and Engineering, University of Ioannina. She is also a Computer Teacher in a private computer school she owns. Her research interests include medical images compression and enhancement.

LISIMACHOS P. KONDI (Senior Member, IEEE) received the Diploma degree in electrical engineering from the Aristotle University of Thessaloniki, Greece, in 1994, and the M.S. and Ph.D. degrees in electrical and computer engineering from Northwestern University, Evanston, IL, USA, in 1996 and 1999, respectively. He is currently a Professor with the Department of Computer Science and Engineering, University of Ioannina, Greece. Previously, he was a Faculty Member of the University at Buffalo, The State University of New York, USA, and has held summer appointments with the Naval Research Laboratory, Washington, DC, USA, and the Air Force Research Laboratory, Rome, NY, USA. He is the coauthor of a book titled *4G Wireless Video Communications* (Wiley, 2009). His research interests include signal and image processing and communications, including image and video compression and transmission over wireless channels and the internet, perceptual image and video quality estimation, sparse representations and compressive sensing, super-resolution of video sequences, and shape coding. He has been an Associate Editor of the *EURASIP Journal on Advances in Signal Processing* since 2005, IEEE TRANSACTIONS ON IMAGE PROCESSING, since 2018, the IEEE SIGNAL PROCESSING LETTERS (2008–2012), and the *International Journal of Distributed Sensor Networks* (2013–2015). He was the Technical Program Committee (TPC) Chair of the International Conference on Digital Signal Processing (DSP), Santorini, Greece, in 2013, and the Technical Program Committee Chair of the IEEE International Conference on Image Processing (ICIP), Athens, Greece, in 2018.