International Journal of Computer Vision (2021) 129:2827–2845
https://doi.org/10.1007/s11263-021-01510-7

Deep Maximum a Posterior Estimator for Video Denoising

Lu Sun[1] · Weisheng Dong[1] · Xin Li[2] · Jinjian Wu[1] · Leida Li[1] · Guangming Shi[1]

Received: 14 December 2020 / Accepted: 9 July 2021 / Published online: 4 August 2021

Abstract
Unlike the maturity of image denoising research, video denoising has remained a challenging problem. A fundamental issue at the core of the video denoising (VD) problem is how to efficiently remove noise by exploiting temporal redundancy in video frames in a principled manner. Based on the maximum a posterior (MAP) estimation framework and recent advances in deep learning, we present a novel deep MAP-based video denoising method named *MAP-VDNet* with adaptive temporal fusion and deep image prior. The proposed MAP-based VD algorithm allows computationally efficient untangling of motion estimation (frame alignment) and image restoration (denoising). To address the misalignment issue, we also present a robust multi-frame fusion strategy for predicting spatially varying fusion weights by a neural network. To facilitate end-to-end optimization, we unfold the proposed iterative MAP-based VD algorithm into a deep convolutional network named *MAP-VDNet*. Extensive experimental results on three popular video datasets have shown that the proposed *MAP-VDNet* significantly outperforms current state-of-the-art VD techniques such as ViDeNN and FastDVDnet. The code is available at https://see.xidian.edu.cn/faculty/wsdong/Projects/MAP-VDNet.htm.

Keywords Video denoising · Model-guided construction · Deep convolutional neural network · Multiframe fusion

1 Introduction

The field of image denoising has advanced rapidly in the past 20 years. Existing image denoising methods can be roughly classified into two categories—i.e., model-based (Portilla et al. 2003; Dabov et al. 2007; Dong et al. 2013; Gu et al. 2014) and deep learning-based (Zhang et al. 2017a; Tai et al. 2017; Zhang et al. 2018a). Model-based methods formulate image denoising as a Bayesian estimation problem in a probabilistic setting or a variational optimization problem in a deterministic setting. Through mathematical construction of an image prior or a regularization function, we can develop principled solutions to the image denoising problem (Dabov et al. 2007; Dong et al. 2013). In the literature of image denoising, sparsity-based prior models or regularization methods are among the most popular, which lead to state-of-the-art denoising performance in model-based approaches (Dabov et al. 2007; Dong et al. 2013).

Deep learning-based approaches to image denoising have been extensively studied in recent years. Instead of mathematical construction, learning-based denoising methods directly learn a nonlinear mapping function from the space of noisy images to that of clean images (Zhang et al. 2017a; Tai et al. 2017; Zhang et al. 2018a). In view of the notorious vanishing gradient problem (Bengio et al. 1994), novel architectures such as residual and densely connected networks (Huang et al. 2017; Zhang et al. 2018b) have been developed. Other methods have exploited the domain knowledge during the construction of network architectures—e.g., unfolding iterative denoising algorithms into deep convolutional neural networks (DCNN) (Wisdom et al. 2017; Hershey et al. 2014; Dong et al. 2018b; Bertocchi et al. 2020). DCNN-based

Communicated by Dong Xu.

✉ Weisheng Dong
wsdong@mail.xidian.edu.cn

Lu Sun
sunlu@stu.xidian.edu.cn

Xin Li
xin.li@ieee.org

Jinjian Wu
jinjian.wu@mail.xidian.edu.cn

Leida Li
ldli@xidian.edu.cn

Guangming Shi
gmshi@xidian.edu.cn

[1] Xidian University, Xi'an, China

[2] West Virginia University, Morgantown, USA

image denoising techniques (Tai et al. 2017; Zhang et al. 2017a; Dong et al. 2018b) have achieved current state-of-the-art performance due to their powerful learning capabilities.

By contrast, video denoising (VD) has been a relatively underresearched problem in the literature (Ji et al. 2010; Liu and Freeman 2010). Model-based approaches toward video denoising are often built upon well-known image denoising algorithms (Varghese and Wang 2010; Mahmoudi and Sapiro 2005; Buades et al. 2016; Maggioni et al. 2012). e.g., Gaussian scale mixture (GSM) denoising (Portilla et al. 2003) was extended in Varghese and Wang (2010), nonlocal means denoising (Buades et al. 2005) became Mahmoudi and Sapiro (2005), Buades et al. (2016) and BM3D denoising (Dabov et al. 2007) evolved into BM4D for video denoising (Maggioni et al. 2012). Besides, some model-based algorithms (Dong et al. 2018a; Pablo and Jean 2018) have obtained superior denoising performance. Most recently, a flurry of deep learning-based VD methods (Mildenhall et al. 2018; Godard et al. 2018; Xue et al. 2019; Claus and van Gemert 2019; Ehret et al. 2019; Tassano et al. 2019, 2020; Davy et al. 2019) have been developed. Most of these methods first align noisy frames and then perform denoising on the aligned noisy frames with a standard DCNN. Frame alignment can be achieved by optical flow estimation network (Ranjan and Black 2017), kernel prediction network (Mildenhall et al. 2018), and MEMC-Net (Bao et al. 2019).

While recent deep learning-based VD methods address the VD problem by directly learning a mapping function from the noisy frames to the desired clean frames, these methods are heuristic and ignore the observation model that characterizes the formation of noisy frames. Inspired by the analogy between burst and video denoising (Godard et al. 2018), we propose a maximal a posterior (MAP) estimation framework for VD with DCNN denoising prior and robust multiframe alignment/fusion. By explicitly taking the frame alignment errors (due to misregistration) into the observation model, we have derived a principled MAP estimation solution for better VD performance. As both steps of single-frame denoising and multiframe fusion can be implemented by convolutional neural networks, our MAP-based VD algorithm can be unfolded into a network implementation, allowing further optimization of network parameters through end-to-end training. In addition to the excellent performance, the resulting network (denoted as *MAP-VDNet*) admits a Bayesian interpretation from the MAP perspective. The key technical contributions of this work are summarized as follows.

- MAP-inspired network. We first propose an iterative MAP-based VD algorithm with DCNN denoising prior, where each iteration can be efficiently computed. Then, a MAP-inspired VD network is constructed by unfolding the iterative video denoising algorithm into a multistage implementation.
- Robust multiframe fusion. To address the inevitable alignment errors (due to misregistration), we propose to fuse the aligned noisy frames with spatially variant weights learned by a kernel prediction network (KPN, Mildenhall et al. (2018)). Such an analogy between burst and video denoising has not been addressed in the open literature to the best of our knowledge.
- Excellent denoising performance. The proposed MAP-VDNet has dramatically advanced the state-of-the-art by over 1 dB while using a comparable number of model parameters as shown in Fig. 1. The fast implementation of MAP-VDNet strikes an improved trade-off between the cost and the performance over other competing methods.
- Good generalization property. In addition to VD, this work can be readily generalized to other video restoration tasks such as superresolution and compression artifact reduction. Promising experimental results for video superresolution and compression artifact reduction have verified the good generalization property of the proposed MAP-VDNet.

2 Related Works

2.1 Deep Image Denoising

Image denoising has been extensively studied, not only due to its wide applications. As the simplest inverse problem, numerous image priors or modeling techniques have been used to evaluate their performance for image denoising (Portilla et al. 2003; Dabov et al. 2007; Zoran and Weiss 2011; Dong et al. 2013; Gu et al. 2014). Before the popularity of deep learning, the most popular image prior is the sparse representation over learned dictionaries (Elad and Aharon 2006; Dong et al. 2013), assuming that image patches can be well approximated by a few representative atoms learned from training images. Such methods were highly related to the maximum a posterior (MAP) techniques, where the prior distributions of the sparse coefficients play a key. In addition to the sparsity, the nonlocal self-similarity of natural images was also considered in developing an improved probability model (Dong et al. 2013), leading to significant improvements. Inspired by the great successes of DCNNs for image classification, DCNNs have also been adopted for image denoising (Zhang et al. 2017a; Tai et al. 2017; Zhang et al. 2018a). In Zhang et al. (2017b), the DCNNs were proposed to learn the residual images that are the differences between clean and noisy images. More sophisticated network architectures considering long-term dependencies have also been proposed (Tai et al. 2017). Instead of designing the network as a black box, the domain knowledge has also been incorporated into the DCNNs- i.e., unfolding the iterative processes into deep networks (Dong et al. 2018b), showing

highly competitive performance. Most recently, neural architecture search (NAS) has found successful application into image denoising too (e.g., Zhang et al. (2020)).

2.2 Deep Video Denoising

Deep learning for VD has recently received increasing attention. In Xue et al. (2019), a neural network with a trainable motion estimation component and a video processing component was developed for video denoising and superresolution. Another VD network based on explicit alignment is RViDeNet (Yue et al. 2020) which takes a supervised learning approach. One of the common limitations to existing learning-based VD is the heuristic concatenation of alignment and denoising subnetworks. We conjecture that there is still much room for performance improvement by pursuing end-to-end optimization. Deep learning for video denoising without explicit motion estimation has also been studied in the literature. A 3D deformable kernel was developed for video denoising in Xu et al. (2019a) aiming at more efficiently sampling the pixels across the spatio-temporal space. A nonlocal extension of the convolutional neural network (CNN) was constructed in Davy et al. (2019) for VD and demonstrated highly competent performance to the best model-based VD VNLB (Pablo and Jean 2018). A low-complexity alternative called Kalman filtering for frame-recursive video denoising has also been considered in Arias and Morel (2019). Most recently, a VNLnet (Davy et al. 2019), a FastDVDnet (Tassano et al. 2020), and a spatio-temporal pixel aggregation network (ST-PAN, Xu et al. (2020)) has achieved the current state-of-the-art performance in VD.

2.3 Deep Burst Denoising

Deep burst denoising (Godard et al. 2018) refers to the problem of image restoration in burst imaging (Hasinoff et al. 2016), which is a special case in the situation of low light imaging (e.g., nighttime environment). Burst denoising can be viewed as the middle ground between image and video denoising. It dealt with the restoration of multiple noisy frames, but the amount of camera and object motion is usually assumed to be small. Therefore, the alignment of multiple frames is faster for burst denoising—e.g., commonly adopted global registration techniques such as Homography-based (Tekalp 2015) and local feature-based such as Lucas-Kanade tracking (Lucas and Kanade 1981) are deemed sufficient. Existing works on deep burst denoising mostly fall in the framework of kernel prediction network (KPN, Mildenhall et al. (2018)). The basic idea behind the construction of KPN is to predict spatially varying kernels that can both align and restore noisy frames. However, previous works including KPN and other methods (Godard et al. 2018; Xu et al. 2019a, 2020) have not considered the general cases of large unknown misalignments which are common in VD (as illustrated in Fig. 2). How to model structure and signal-dependent noise arising from misalignment sets up the stage for this work.

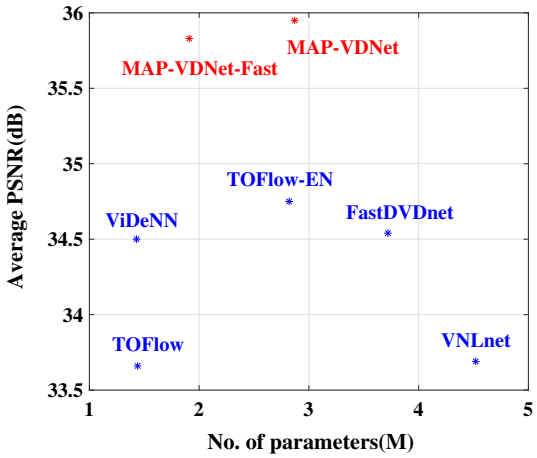

Fig. 1 Comparisons of performance and no. of parameters. Results are evaluated on DTMC-HD test set for $\sigma_n = 25$

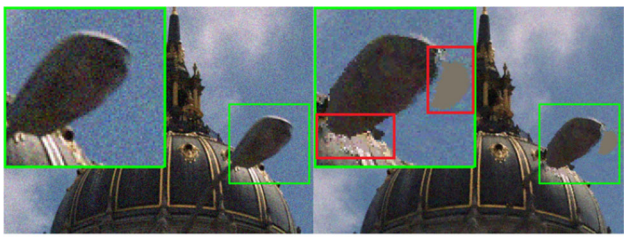

Fig. 2 The illustration of structural noise introduced by the alignment module. **a** the reference frame, **b** the adjacent aligned frame (note undesirable artifacts as highlighted by red boxes). Note that these artifacts are signal-dependent and behave differently from additive white Gaussian noise (Color figure online)

3 Proposed MAP Estimator for Video Denoising

3.1 MAP Estimator with i.i.d Gaussian Likelihood

For VD, we aim at recovering each clean frame $\boldsymbol{x} \in \mathbb{R}^N$ (the frame index is omitted for simplicity) from a set of adjacent noisy frames $\boldsymbol{y}_j$, $j = -J, \ldots, J$. Unlike image denoising, a noisy frame can have multiple observations because it is highly correlated with adjacent frames due to temporal redundancy. A commonly used video frame observation model can be expressed as

$$\boldsymbol{y}_j = \tilde{\mathbf{W}}_j \boldsymbol{x} + \boldsymbol{n}_j, j = -J, \ldots, J, \quad (1)$$

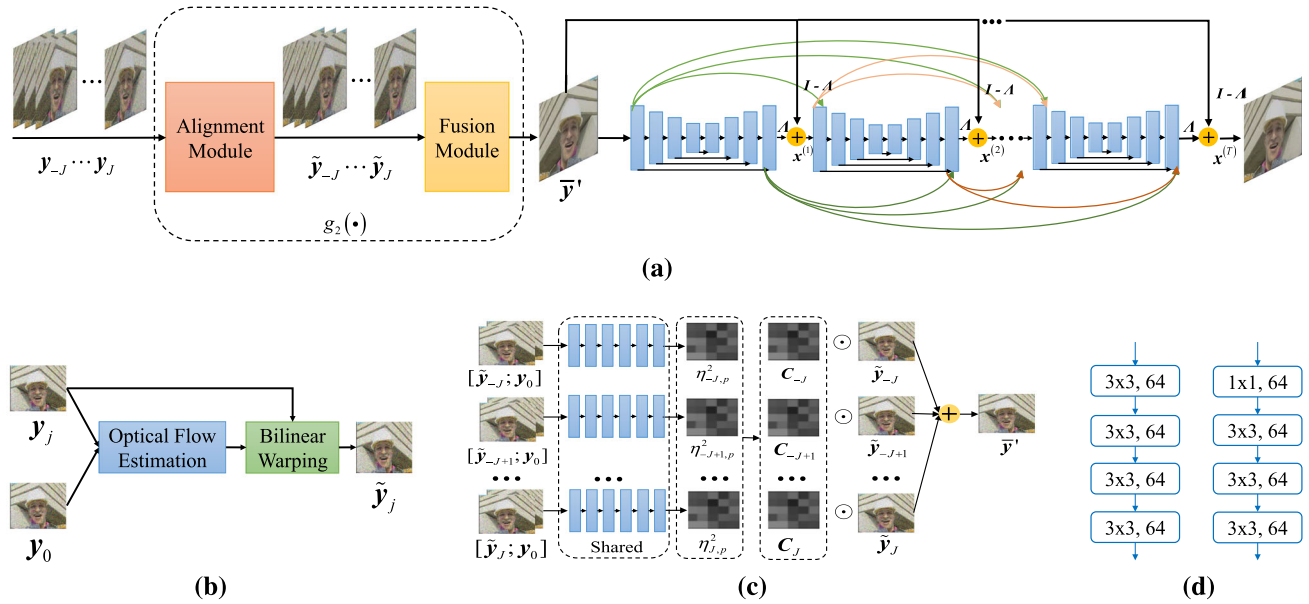

Fig. 3 The architecture of the proposed model-guided DCNN for video denoising. **a** The overall architecture of the proposed network, **b** the architecture of the alignment module, **c** the architecture of the robust multi-frame fusion, and **d** the architecture of the encoding and decoding block of the U-net denoiser

Table 1 The average PSNR (dB) and SSIM results of the variants of the proposed method for noise level $\sigma_n = 25$

Variants	Vimeo	ASU	DTMC-HD	No. of Para.(M)
MAP-VDNet	39.12	34.75	35.95	2.87
	0.9646	0.9478	0.9246	
MAP-VDNet-Fast	38.73	34.37	35.83	1.90
	0.9598	0.9444	0.9225	
MAP-VDNet-NAF	38.82	34.42	35.63	2.77
	0.9619	0.9432	0.9199	
MAP-VDNet-DnCNN	38.91	34.62	35.88	2.00
	0.9624	0.9459	0.9222	
VD-Unet	38.65	34.26	35.56	2.50
	0.9611	0.9433	0.9194	
VD-Unet-Fast	38.44	33.71	35.48	1.54
	0.9594	0.9406	0.9193	

where $\tilde{\mathbf{W}}_j \in \mathbb{R}^{N\times N}$ denotes the warping matrix (or an operator) that maps from the frame of interest $\boldsymbol{x}$ to its adjacent frame $\boldsymbol{x}_j$ (e.g., optical flow (Ranjan and Black 2017)), and $\boldsymbol{n}_j \in \mathbb{R}^N$ is the additive Gaussian noise. In the above formulation, $\tilde{\mathbf{W}}_0$ is simply an identity matrix for the central frame at $j = 0$. The above observation model has been adopted in Kokkinos and Lefkimmiatis (2019) for burst denoising. However, the tangle of the warping matrices $\tilde{\mathbf{W}}_j$ with $\boldsymbol{x}$ making the optimization of $\boldsymbol{x}$ very difficult, i.e., one has to solve a sequence of large matrix inverse problems. In this paper, we propose the following noisy frame observation model, as

$$\boldsymbol{x} = \mathbf{W}_j \boldsymbol{x}_j = \mathbf{W}_j \boldsymbol{y}_j + \tilde{\boldsymbol{n}}_j,\ \ j = -J, \ldots, J, \tag{2}$$

where we have used $\boldsymbol{y}_j = \boldsymbol{x}_j + \boldsymbol{n}_j$, $\tilde{\boldsymbol{n}}_j = -\mathbf{W}_j \boldsymbol{n}_j$, and $\mathbf{W}_j$ denotes the warping matrix from $\boldsymbol{x}_j$ to the central frame $\boldsymbol{x}$. For simplicity, we assume $\tilde{\boldsymbol{n}}_j$ is still Gaussian with variance σ_n^2. Based on such multi-frame observation model and the assumption that $\boldsymbol{x}$ and $\mathbf{W}_j$ are independent, we propose the following Maximum a Posterior (MAP) estimation

$$(\boldsymbol{x}, \mathbf{W}_j) = \underset{\boldsymbol{x},\mathbf{W}_j}{\operatorname{argmax}} \sum_{j=-J}^{J} \log P(\boldsymbol{y}_j|\boldsymbol{x}, \mathbf{W}_j) + \log P(\boldsymbol{x}) + \log P(\mathbf{W}_j), \tag{3}$$

where $P(\boldsymbol{y}_j|\boldsymbol{x})$ is the Gaussian likelihood, which can be expressed as

$$P(\boldsymbol{y}_j|\boldsymbol{x},\mathbf{W}_j) \propto \exp\left(-\frac{1}{\sigma_n^2}||\boldsymbol{x}-\mathbf{W}_j\boldsymbol{y}_j||_2^2\right). \tag{4}$$

For the prior terms $P(\boldsymbol{x})$ and $P(\mathbf{W}_j)$, we use a general form to describe

$$P(\boldsymbol{x}) \propto \exp(-\gamma\Omega(\boldsymbol{x})), \tag{5a}$$
$$P(\mathbf{W}_j) \propto \exp(-\delta\Phi_1(\mathbf{W}_j)), \tag{5b}$$

where $\Omega(\cdot)$ and $\Phi_1(\cdot)$ denote energy functions related to $\boldsymbol{x}$ and $\mathbf{W}_j$ respectively, and γ and δ are the corresponding weights. By substituting the Gaussian likelihood into Eq. (4) and prior terms of Eq. (5) in the MAP estimation of Eq. (3), we can obtain the following objective function for multiframe denoising

$$(\boldsymbol{x},\mathbf{W}_j) = \underset{\boldsymbol{x},\mathbf{W}_j}{\operatorname{argmin}} \sum_{j=-J}^{J} \eta||\boldsymbol{x}-\mathbf{W}_j\boldsymbol{y}_j||_2^2 + \gamma\Omega(\boldsymbol{x}) + \delta\Phi_1(\mathbf{W}_j), \tag{6}$$

where $\eta = 1/\sigma_n^2$. Regarding the prior term $P(\boldsymbol{x})$, instead of mathematical construction of $\Omega(\boldsymbol{x})$, we propose to adopt the DCNN-based image prior (Dong et al. 2018b) by introducing an auxiliary variablea $\boldsymbol{v}$, the objective function of Eq. (6) could be translated into

$$(\boldsymbol{x},\boldsymbol{v},\mathbf{W}_j) = \underset{\boldsymbol{x},\boldsymbol{v},\mathbf{W}_j}{\operatorname{argmin}} \sum_{j=-J}^{J} \eta||\boldsymbol{x}-\mathbf{W}_j\boldsymbol{y}_j||_2^2 + \lambda||\boldsymbol{x}-\boldsymbol{v}||_2^2 + \gamma\Omega(\boldsymbol{v}) + \delta\Phi_1(\mathbf{W}_j). \tag{7}$$

It could be observed from Eq. (7) that the multiframe denoising problem could be solved by iteratively optimizing the following three subproblems

$$\mathbf{W}_j = \underset{\mathbf{W}_j}{\operatorname{argmin}} \sum_{j=-J}^{J} \eta||\boldsymbol{x}-\mathbf{W}_j\boldsymbol{y}_j||_2^2 + \delta\Phi_1(\mathbf{W}_j), \tag{8a}$$
$$\boldsymbol{v} = \underset{\boldsymbol{v}}{\operatorname{argmin}}\ \lambda||\boldsymbol{x}-\boldsymbol{v}||_2^2 + \gamma\Omega(\boldsymbol{v}), \tag{8b}$$
$$\boldsymbol{x} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \sum_{j=-J}^{J} \eta||\boldsymbol{x}-\mathbf{W}_j\boldsymbol{y}_j||_2^2 + \lambda||\boldsymbol{x}-\boldsymbol{v}||_2^2. \tag{8c}$$

Regarding the $\mathbf{W}_j$-subproblem, instead of estimating $\mathbf{W}_j$ with a specific prior, we propose to estimate $\mathbf{W}_j^{(t+1)}$ from $\boldsymbol{x}$ and $\boldsymbol{y}_j$ using a DCNN function. Since $\boldsymbol{x}$ is not available, we use the current estimation $\boldsymbol{x}^{(t)}$ to compute $\mathbf{W}_j$ as follows

$$\mathbf{W}_j{}^{(t+1)} = g_1(\boldsymbol{x}^{(t)}, \boldsymbol{y}_j), \tag{9}$$

where $g_1(\cdot)$ represents a DCNN function. Besides, as demonstrated in Eq. (8b), the $\boldsymbol{v}$-subproblem could be considered as a single-frame denoising problem, and $\boldsymbol{v}$ can be updated as

$$\boldsymbol{v}^{(t+1)} = f(\boldsymbol{x}^{(t)}), \tag{10}$$

where $f(\cdot)$ denotes the DCNN denoising function. Concerning the $\boldsymbol{x}$-subproblem, since both the terms are l_2-norm in Eq. (8c), the above objective function of $\boldsymbol{x}$-subproblem admits a closed-form solution as

$$\boldsymbol{x}^{(t+1)} = (1-a)\overline{\boldsymbol{y}}^{(t+1)} + a\boldsymbol{v}^{(t+1)}, \tag{11}$$

where $a = \frac{\lambda}{(2J+1)\eta+\lambda}$ and $\overline{\boldsymbol{y}}^{(t+1)} = \frac{1}{(2J+1)}\sum_{j=-J}^{J}\mathbf{W}_j{}^{(t+1)}\boldsymbol{y}_j$. In principle, through the iterative computations of $\mathbf{W}_j$, $\boldsymbol{v}$ and $\boldsymbol{x}$, the multiframe denoising problem can be efficiently solved.

Although the warping matrix $\mathbf{W}_j$ can be estimated by any existing motion compensation method, its computational complexity is often prohibitive. For example, popular deep motion compensation methods (Ranjan and Black 2017; Wang et al. 2019) tend to use a multi-scale architecture to deal with large displacements between frames. Even with a fast optical flow estimation method such as Kroeger et al. (2016), iterative computation of the alignment matrix would involve frequent data type conversion between GPU and CPU, which renders an additional burden to limited computational resources. To tackle this problem, we opt to update $\mathbf{W}_j$ *just once* in our iterations, and then iteratively optimize $\boldsymbol{x}$ and $\boldsymbol{v}$ with a fixed $\mathbf{W}_j$. In practice, we use $\boldsymbol{y}_0$ to initialize $\boldsymbol{x}^{(t)}$ in Eq. (9) and $\mathbf{W}_j$ can be initially estimated by $\mathbf{W}_j = g_1(\boldsymbol{y}_0, \boldsymbol{y}_j)$. Then the MAP estimator leads to a simple multiframe denoising scheme—i.e., iteratively computing

$$\boldsymbol{x}^{(t+1)} = (1-a)\overline{\boldsymbol{y}} + af(\boldsymbol{x}^{(t)}), \tag{12}$$

where $\overline{\boldsymbol{y}} = \frac{1}{(2J+1)}\sum_{j=-J}^{J}\mathbf{W}_j\boldsymbol{y}_j$. Regarding the starting point $\boldsymbol{x}^{(0)}$ in Eq. (12), a natural selection is the weighted average of the aligned frames—i.e., $\boldsymbol{x}^{(0)} = \overline{\boldsymbol{y}}$. The reason is that the averaging of the m noisy observations of a frame will reduce the noise variance from σ_n^2 to $\frac{\sigma_n^2}{m}$ (Chang et al. 2000). Thus, if the temporal alignment is perfect, the averaging of the aligned noisy frames will significantly reduce the noise variance by a factor equaling to the total number of aligned noisy frames. However, due to the inevitable misalignment errors from one-time alignment iteration, further refinement

Table 2 The PSNR(dB) and SSIM performance comparison for different values of J at the noise level of $\sigma_n = 25$

Dataset	$J=1$		$J=2$		$J=3$	
Vimeo	38.27	0.9581	38.88	0.9626	39.12	0.9646
DTMC-HD	35.28	0.9172	35.80	0.9222	35.95	0.9246

of MAP estimation has to be performed based on the baseline method. We will refine the noise observation model and derive a new MAP estimation for the updated observation model next.

3.2 MAP Estimator with Non-i.i.d Gaussian Likelihood

In Eq. (4), we assume that the warped noise $\mathbf{W}_j \boldsymbol{n}_j$ is still Gaussian with variance σ_n^2. However, when the motion estimation is inaccurate, some pixels in the adjacent frame after alignment might become misaligned with the reference frame causing undesirable artifacts as shown in Fig. 2. Those artifacts are signal-dependent and behave differently from additive white Gaussian noise. Therefore, the noise $\tilde{\boldsymbol{n}}_j$ in the aligned noisy frame is more complex than identically distributed (i.i.d) Gaussian. To address this issue, we propose a non-i.i.d Gaussian likelihood for multiframe denoising inspired by the analogy between burst denoising and video denoising (Mildenhall et al. 2018). To the best of our knowledge, such an adversary effect of motion estimation on video denoising has not been explored in the open literature.

To properly take such systematic errors into account, we introduce a new alignment error term $\boldsymbol{e}_j$ to represent the potential deviation of an aligned frame from the reference frame. More specifically, we propose to rewrite the observation model in Eq. (2) as

$$\boldsymbol{x} = \mathbf{W}_j \boldsymbol{x}_j + \boldsymbol{e}_j = \tilde{\boldsymbol{y}}_j + \tilde{\boldsymbol{n}}_j + \boldsymbol{e}_j, \;\; j = -J, \ldots, J, \tag{13}$$

where $\tilde{\boldsymbol{y}}_j = \mathbf{W}_j \boldsymbol{y}_j$, and $\boldsymbol{e}_j$ denotes the alignment error. Generally, $\boldsymbol{x} - \tilde{\boldsymbol{y}}_j$ generally no longer satisfies the i.i.d Gaussian assumption due to the interference from alignment error term $\boldsymbol{e}_j$. Instead, it is more appropriate to pursue an adaptive estimation about the standard deviation of spatially varying noise on a pixel-by-pixel basis. It follows that we can extend the i.i.d Gaussian likelihood term in Eq. (4) into its non-i.i.d version, as

$$\begin{aligned} P(\boldsymbol{y}_j|\boldsymbol{x}, \mathbf{W}_j, \boldsymbol{\sigma}_j) &= \prod_p P(\boldsymbol{y}_{j,p}|\boldsymbol{x}_p, \tilde{\boldsymbol{y}}_{j,p}, \sigma_{j,p}) \\ &= \prod_p \frac{1}{\sigma_{j,p}\sqrt{2\pi}} \exp[-\sum_p \frac{1}{\sigma_{j,p}^2}(\boldsymbol{x}_p - \tilde{\boldsymbol{y}}_{j,p})^2], \end{aligned} \tag{14}$$

where $\tilde{\boldsymbol{y}}_{j,p}$ and $\boldsymbol{x}_p$ denote the p-th elements of $\tilde{\boldsymbol{y}}_j$ and $\boldsymbol{x}$, respectively, and $\sigma_{j,p}$ is the per-pixel noise standard deviation after alignment. Obviously, $\sigma_{j,p}$ is relevant to the warping matrix $\mathbf{W}_j$, thus we propose to jointly estimate $\mathbf{W}_j$ and $\boldsymbol{\sigma}_j$ under the MAP estimation framework as follows

$$\begin{aligned} (\boldsymbol{x}, \mathbf{W}_j, \eta_j) = \underset{\boldsymbol{x}, \mathbf{W}_j, \eta_j}{\operatorname{argmax}} & \sum_{j=-J}^{J} \log P(\boldsymbol{y}_j|\boldsymbol{x}, \mathbf{W}_j, \eta_j) \\ & + \log P(\boldsymbol{x}) + \log P(\mathbf{W}_j, \eta_j), \end{aligned} \tag{15}$$

where $\boldsymbol{\eta}_j = 1/\boldsymbol{\sigma}_j$. Similarly, by substituting the non-i.i.d likelihood term into the MAP estimator of Eq. (15) and introducing a denoising prior term, we can extend the objective function of Eq. (7) into the following spatially adaptive formulation

$$\begin{aligned} (\boldsymbol{x}, \boldsymbol{v}, \mathbf{W}_j, \boldsymbol{\eta}_j) = \underset{\boldsymbol{x}, \boldsymbol{v}, \mathbf{W}_j, \boldsymbol{\eta}_j}{\operatorname{argmin}} & \sum_{j=-J}^{J} ||\boldsymbol{\Sigma}_j(\boldsymbol{x} - \mathbf{W}_j \boldsymbol{y}_j)||_2^2 \\ & + \gamma \Omega(\boldsymbol{v}) + \lambda ||\boldsymbol{x} - \boldsymbol{v}||_2^2 + \delta \Phi_2(\mathbf{W}_j, \boldsymbol{\eta}_j) \\ & + \log(\boldsymbol{\eta}_j \sqrt{2\pi}), \end{aligned} \tag{16}$$

where $\Phi_2(\cdot)$ represents the energy function related to joint distribution of $\mathbf{W}_j$ and $\boldsymbol{\eta}_j$, and the weighting matrix $\boldsymbol{\Sigma}_j = \operatorname{diag}(\eta_{j,p}) \in \mathbb{R}^{N \times N}$ is a diagonal matrix whose diagonal entries ($\eta_{j,p} = 1/\sigma_{j,p}$) represent the reciprocal of the standard deviation for each individual noisy pixel. Similar to Eq. (7), we can solve the above optimization problem by the following formulas

$$\begin{aligned} (\mathbf{W}_j, \boldsymbol{\eta}_j) = \underset{\mathbf{W}_j, \boldsymbol{\eta}_j}{\operatorname{argmin}} & \sum_{j=-J}^{J} ||\boldsymbol{\Sigma}_j(\boldsymbol{x} - \mathbf{W}_j \boldsymbol{y}_j)||_2^2 + \delta \Phi_2(\mathbf{W}_j, \boldsymbol{\eta}_j) \\ & + \log(\boldsymbol{\eta}_j \sqrt{2\pi}), \end{aligned} \tag{17a}$$

$$\boldsymbol{v} = \underset{\boldsymbol{v}}{\operatorname{argmin}} \, \lambda ||\boldsymbol{x} - \boldsymbol{v}||_2^2 + \gamma \Omega(\boldsymbol{v}), \tag{17b}$$

$$\boldsymbol{x} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \sum_{j=-J}^{J} ||\boldsymbol{\Sigma}_j(\boldsymbol{x} - \mathbf{W}_j \boldsymbol{y}_j)||_2^2 + \lambda ||\boldsymbol{x} - \boldsymbol{v}||_2^2. \tag{17c}$$

Instead of constructing a specific prior of $\Phi_2(\mathbf{W}_j, \boldsymbol{\eta}_j)$, we propose to jointly estimate $\mathbf{W}_j$ and $\boldsymbol{\eta}_j$ from $\boldsymbol{x}$ and $\boldsymbol{y}_j$ with a DCNN function. Considering the unavailability of $\boldsymbol{x}$, we exploit the intermediate estimation $\boldsymbol{x}^{(t)}$. Similarly, we initialize $\boldsymbol{x}^{(t)}$ with $\boldsymbol{y}_0$, and then conduct one iteration to save the computational overhead. Therefore, $\mathbf{W}_j$ and $\boldsymbol{\eta}_j$ can be estimated as

$$(\mathbf{W}_j, \boldsymbol{\eta}_j) = g_2(\boldsymbol{y}_0, \boldsymbol{y}_j), \tag{18}$$

where $g_2(\cdot)$ is a DCNN function to estimate $\mathbf{W}_j$ and $\boldsymbol{\eta}_j$. Then $\boldsymbol{x}$ could be updated as the following

$$\begin{aligned} \boldsymbol{x}^{(t+1)} &= (\sum_{j=-J}^{J} \boldsymbol{\Sigma}_j^\top \boldsymbol{\Sigma}_j + \lambda \mathbf{I})^{-1} (\sum_{j=-J}^{J} \boldsymbol{\Sigma}_j^\top \boldsymbol{\Sigma}_j \tilde{\boldsymbol{y}}_j + \lambda f(\boldsymbol{x}^{(t)})), \\ &= (\mathbf{I} - \boldsymbol{\Lambda}) \overline{\boldsymbol{y}}' + \boldsymbol{\Lambda} f(\boldsymbol{x}^{(t)}), \end{aligned} \tag{19}$$

Algorithm 1 Proposed MAP estimation for video denoising

- **Initialization**:
  (1) Set λ, $\mathbf{\Lambda}$ and $\mathbf{C}_j$, $j = -J, \cdots, J$, $t = 0$;
  (2) Align $\boldsymbol{y}_j$ to $\boldsymbol{y}$ as $\tilde{\boldsymbol{y}}_j = \mathbf{W}_j \boldsymbol{y}_j$ and perform the fusion as $\overline{\boldsymbol{y}}' = \sum_{j=-J}^{J} \mathbf{C}_j \tilde{\boldsymbol{y}}_j$;
  (3) Initialize $\boldsymbol{x}$ as $\boldsymbol{x}^{(0)} = \overline{\boldsymbol{y}}'$;
- **While** does not converge **do**
  (4) Compute $\boldsymbol{x}^{(t+1)} = (\mathbf{I} - \mathbf{\Lambda})\overline{\boldsymbol{y}}' + \mathbf{\Lambda} f(\boldsymbol{x}^{(t)})$;
  (5) $t = t + 1$.

End while

Output: $\boldsymbol{x}^{(t)}$

where $\overline{\boldsymbol{y}}' = \sum_{j=-J}^{J} \mathbf{C}_j \tilde{\boldsymbol{y}}_j$ denotes the spatially adaptive fusion of the aligned noisy frames, $\mathbf{C}_j = \text{diag}(\frac{\eta_{j,p}^2}{\sum_{j=-J}^{J} \eta_{j,p}^2}) \in \mathbb{R}^{N\times N}$ and $\mathbf{\Lambda} = \text{diag}(\frac{\lambda}{\sum_{j=-J}^{J} \eta_{j,p}^2 + \lambda}) \in \mathbb{R}^{N\times N}$ are diagonal matrices. The fusion matrix $\mathbf{C}_j$ guides the spatially adaptive averaging of the aligned noisy frames, and thus Eq. (19) effectively implements the idea of robust multi-frame fusion, which is conceptually similar to per-pixel kernel prediction in burst denoising (Mildenhall et al. 2018). From Eq. (19), we can see that the MAP estimator with non-i.i.d Gaussian likelihood also leads to a simple multiframe denoising scheme. For the initial estimate $\boldsymbol{x}^{(0)}$ in Eq. (19), we use the adaptively averaged frame as the starting point, i.e., $\boldsymbol{x}^{(0)} = \overline{\boldsymbol{y}}'$. Comparing with Eq. (12), we can see that the updating scheme of Eq. (12) is a special case of Eq. (19) by setting $\mathbf{C}_j$ and $\mathbf{\Lambda}$ as fixed scalar variables. Regarding the per-pixel fusion filters $\boldsymbol{\eta}_j$, we will use the deep network to adaptively optimize the fusion weights as described in the next section. Through the estimation of filters $\eta_{j,p}^2 \in \mathbb{R}^N (j = -J, \ldots, J)$, the diagonal matrices $\mathbf{C}_j$ and $\mathbf{\Lambda}$ can be efficiently computed. The proposed MAP-based video denoising algorithm can be summarized in **Algorithm 1**.

In the conventional wisdom of model-based denoising, an implementation of **Algorithm 1** requires many iterations to converge. Moreover, it will be difficult to jointly optimize the denoiser $f(\cdot)$, the alignment operators $\mathbf{W}_j$, the adaptive fusion weights $\mathbf{C}_j$, and λ (note that most parameters in model-based denoising are often hand-crafted). Such observation motivates us to implement the proposed MAP estimator with a deep neural network, and so all components along with their parameters can be jointly optimized through end-to-end training. In particular, regarding the per-pixel fusion weights $\mathbf{C}_j$ that plays an important role in robust multiframe fusion, we will use the deep network to adaptively optimize the fusion weights as described in the next section.

4 Deep Neural Network Implementation

Although the frame alignment, multiframe fusion, and image denoising can all be implemented by deep neural networks (DNN), an exact DNN-based implementation of the proposed MAP-based VD algorithm is not straightforward (Hershey et al. 2014). Unfolding MAP-based VD algorithm takes more effort than its image denoising counterpart (Dong et al. 2018b). The overall network architecture of the unfolded MAP-based Video Denoising Network, dubbed *MAP-VDNet*, is presented in Fig. 3a. Note that in our current implementation, the $g_2(\cdot)$ function in Eq. (18) is realized by a two-step procedure, which illustrated as the alignment module to compute $\mathbf{W}_j \boldsymbol{y}_j$ and fusion module to estimate $\boldsymbol{\eta}_j$, respectively. As shown in Fig. 3a, the input noisy frames $\boldsymbol{y}_j$ are first aligned by an alignment module, which can be any existing alignment method in principle. The aligned noisy frames $\tilde{\boldsymbol{y}}_j$ are then fed into the fusion module, which performs a spatially adaptive fusion of $\tilde{\boldsymbol{y}}_j$. The fused noisy frame $\overline{\boldsymbol{y}}'$ as the initial estimate of $\boldsymbol{x}$ is then passed to the image denoiser $f(\cdot)$. The output of the denoiser $f(\cdot)$ is finally combined with the fused frame $\overline{\boldsymbol{y}}'$ for producing an improved estimate of $\boldsymbol{x}$. Such a process will be iterated for several stages corresponding to the unfolding of the main loop in **Algorithm 1**. Thus, the proposed MAP-VDNet as shown in Fig. 3a provides an exact implementation of **Algorithm 1**. Next, we will provide the details of each module.

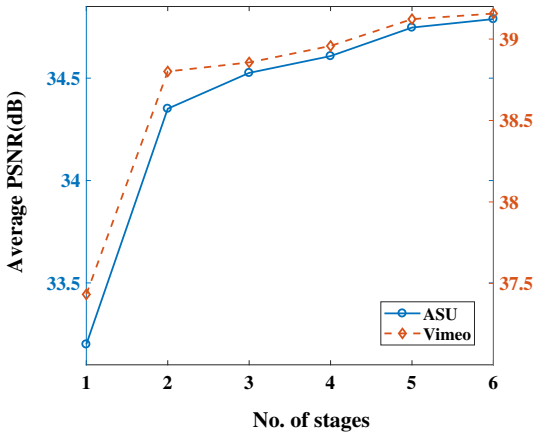

Fig. 4 The average PSNR curves as a function of the no. of stages T of the proposed *MAP-VDNet* for $\sigma_n = 25$

Table 3 The PSNR(dB) and SSIM performance comparison for the effect of dense connections between the denoisers when $\sigma_n = 25$

Dataset	MAP-VDNet-wo		MAP-VDNet	
Vimeo	39.00	0.9637	39.12	0.9646
DTMC-HD	35.81	0.9231	35.95	0.9246

4.1 Alignment Module

This network implements step (2) in **Algorithm 1**. The adjacent $2J$ noisy frames need to be aligned with the frame of interest $\boldsymbol{y}$ first. As shown in Fig. 3b, each frame $\boldsymbol{y}_j$ and the

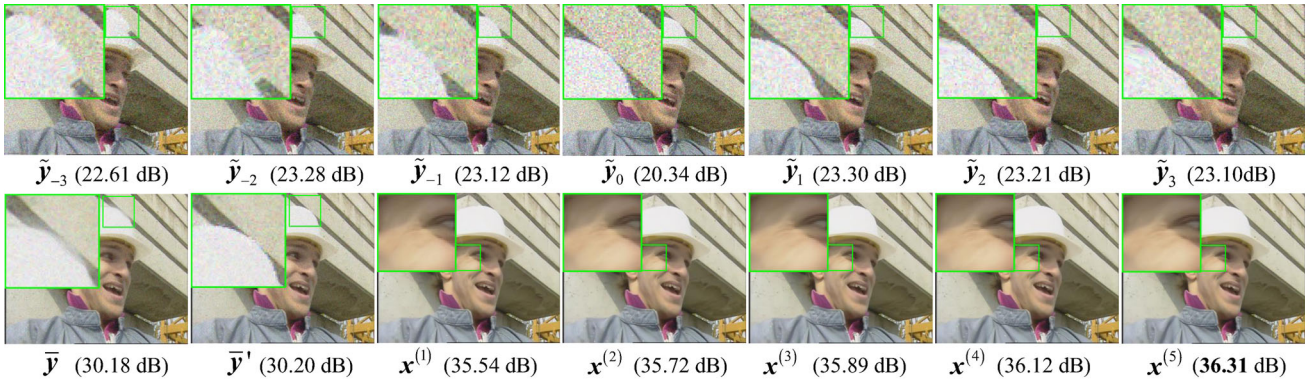

Fig. 5 Denoising intermediate visual and PSNR results of aligned frames $\tilde{\boldsymbol{y}}_j (j = -3, \ldots, 3)$, the average fused frame $\overline{\boldsymbol{y}}$, the spatially adaptive fused frame $\overline{\boldsymbol{y}}'$, and outputs from all stages $\boldsymbol{x}^{(t)} (t = 1, \ldots, 5)$ for a noisy frame of *Foreman* video of ASU test set with noise level $\sigma_n = 25$. Note that the average fused frame $\overline{\boldsymbol{y}}$ would generate undesirable artifacts caused by alignment errors, while the proposed spatially adaptive fused frame $\overline{\boldsymbol{y}}'$ could restore shaper edges (Please zoom in to see better details)

reference frame $\boldsymbol{y}$ is first used to compute the optical flow, then the estimated optical flow fields are then used to warp the noisy frame $\boldsymbol{y}_j$ to the reference frame $\boldsymbol{y}$. Either deep learning-based or conventional optical flow estimation methods can be used to estimate the optical flows. In Sect. 5.2, we have used two optical flow estimation methods—i.e., the Spynet (Ranjan and Black 2017) and the fast inverse search-based method (Kroeger et al. 2016) (corresponding to the name MAP-VDNet-Fast in Fig. 1) to compare the effect of different alignment modules. The estimated optical flow fields are then used to warp the noisy frame $\boldsymbol{y}_j$ to the reference frame $\boldsymbol{y}$ by the spatial transformer network of (Jaderberg et al. 2015) before fusion.

4.2 Fusion Module

In addition to concatenation-based fusion, we have proposed two fusion strategies, as discussed in Sect. 3- i.e., the uniform average fusion $\overline{\boldsymbol{y}} = \frac{1}{(2J+1)} \sum_{j=-J}^{J} \tilde{\boldsymbol{y}}_j$ and the spatially adaptive fusion $\overline{\boldsymbol{y}}' = \sum_{j=-J}^{J} \mathbf{C}_j \tilde{\boldsymbol{y}}_j$. The former fusion method is simply the uniform average of all aligned noisy frames; the later fusion method performs per-pixel adaptive fusion according to the noise variance. For per-pixel adaptive fusion, we propose to estimate the filters $\eta_{j,p}^2$ by a deep neural network, which is analogous to the kernel prediction network (KPN) used in burst denoising (Mildenhall et al. 2018). Similar to KPN, the objective of fusion module is to derive spatially-adaptive fusion weights $\mathbf{C}_j$. Unlike KPN (Mildenhall et al. 2018), we only need to estimate the per-pixel 1×1 kernels as the frames have been aligned by the optical flow-based alignment method.

As shown in Fig. 3c, each aligned frame $\tilde{\boldsymbol{y}}_j$ and the central frame $\boldsymbol{y}$ are first concatenated and fed into a DCNN to predict 1×1 fusion kernels. As the input frames have been aligned, we employed a compact DCNN as the backbone to predict filters $\eta_{j,p}^2$, which contains five residual blocks, each residual block consists of two convolutional layers with 3×3 kernels and ReLU nonlinearity to generate 32-channel feature maps. Note that all $2J$ DCNN modules share the same parameters. Then spatially-adaptive fusion weights $\mathbf{C}_j = \text{diag}(\frac{\eta_{j,p}^2}{\sum_{j=-J}^{J} \eta_{j,p}^2})$ could be calculated using the filters $\eta_{j,p}^2$. Finally, the aligned frames $\tilde{\boldsymbol{y}}_j$ are averaged using fusion weights $\mathbf{C}_j$ to generate the fused frame $\overline{\boldsymbol{y}}'$. Through the estimation of a spatially varying kernel, our multiframe fusion becomes more robust in the presence of misregistration errors and other nonuniformly distributed noise (e.g., signal-dependent compression artifacts).

4.3 Denoising Module

This module implements the denoising function $f(\cdot)$ in **Algorithm 1** and in principle, any existing image denoising network can be used as the denoising module. Here, we opt to use the U-net of Ronneberger et al. (2015) as the backbone network for the denoiser, which consists of an encoder and a decoder. The encoder and decoder contain five encoding blocks (EB) and four decoding blocks (DB), respectively. As shown in Fig. 3d, except the last EB, each EB consists of four convolutional layers with 3×3 kernels to generate 64-channel feature maps and followed by a downsampling layer to reduce the spatial resolution by a scaling factor of two, while the last EB only contains four convolutional layers to produce 64-channel feature maps and does not conduct downsampling. Each DB contains five convolutional layers. The first layer uses 1×1 kernels to reduce the number of feature channels from 128 to 64, and other layers produce the 64-channel feature maps with 3×3 kernels. Each DB is followed by a deconvolution layer to increase the spatial

Table 4 The average PSNR/SSIM video denoising results on Vimeo, ASU, and DTMC-HD dataset at different noise levels (boldface highlights the best)

Datasets	Vimeo			ASU			DTMC-HD		
σ_n	15	25	50	15	25	50	15	25	50
V-BM4D (Maggioni et al. 2012)	37.00	33.55	27.97	34.89	32.13	27.67	35.74	33.45	29.38
	0.9112	0.8663	0.7816	0.9416	0.9030	0.8194	0.9147	0.8740	0.7946
RTA-LSM (Dong et al. 2018a)	37.80	34.12	28.50	34.79	31.99	27.36	36.50	34.22	29.90
	0.9246	0.8879	0.8179	0.9474	0.9155	0.8345	0.9290	0.8996	0.8285
VNLB (Pablo and Jean 2018)	38.30	34.76	28.59	36.57	33.77	28.75	**37.63**	35.52	30.41
	0.9250	0.8917	0.8151	0.9587	0.9346	0.8643	0.9395	0.9149	0.8340
TOFlow (Xue et al. 2019)	37.70	35.58	32.83	34.00	31.90	29.04	35.49	33.66	31.13
	0.9645	0.9210	0.9184	0.9402	0.9125	0.8538	0.9164	0.8861	0.8287
TOFlow-EN	39.38	37.26	34.45	35.36	33.17	30.45	36.48	34.75	32.28
	0.9645	0.9477	0.9184	0.9529	0.9307	0.8888	0.9325	0.9078	0.8610
ViDeNN (Claus and van Gemert 2019)	39.85	37.22	34.36	35.52	33.28	30.14	36.38	34.50	31.87
	0.9676	0.9511	0.9148	0.9536	0.9304	0.8795	0.9320	0.9066	0.8548
VNLnet (Davy et al. 2019)	38.74	36.09	32.45	34.56	31.88	28.32	35.86	33.69	30.58
	0.9592	0.9354	0.8823	0.9440	0.9117	0.8405	0.9253	0.8938	0.8262
FastDVDnet (Tassano et al. 2020)	39.35	37.55	34.92	35.32	33.22	30.56	36.23	34.54	32.33
	0.9639	0.9519	0.9250	0.9510	0.9314	0.8917	0.9303	0.9084	0.8677
MAP-VDNet-Fast	40.97	38.73	35.57	36.54	34.57	30.98	37.57	35.83	**33.17**
	0.9727	0.9598	0.9332	0.9626	0.9444	0.9014	**0.9420**	0.9225	**0.8809**
MAP-VDNet	**41.14**	**39.12**	**35.62**	**36.83**	**34.75**	**31.02**	37.57	**35.95**	33.15
	0.9750	**0.9646**	**0.9343**	**0.9632**	**0.9478**	**0.9034**	**0.9420**	**0.9246**	0.8802

resolution of the feature maps with a scaling factor of two. To compensate for the loss of spatial information, the upsampled feature maps are concatenated with the feature maps of the same spatial dimension from the encoder. To save the total number of parameters, all denoising networks in the consecutive T stages are enforced to share the same network parameters.

Following **Algorithm 1**, the T denoising networks process the intermediate image $\boldsymbol{x}^{(t)}$ independently and thus cannot exploit the features from the previous denoising networks. Inspired by the success of the densenet (Huang et al. 2017), we propose to connect those feature maps from the previous denoisers to the following denoisers. As shown in Fig. 3a, the feature maps of the first encoding and last decoding blocks are connected to those of the subsequent denoisers. These long skip connections also help alleviate the notorious vanishing gradient problem (Bengio et al. 1994).

4.4 Network Training and Extensions

When using the deep optical flow estimation method for frame alignment, the alignment module is first pretrained on the training images of the Vimeo dataset (Xue et al. 2019). Then we jointly train the overall network by minimizing the following loss function

$$\boldsymbol{\Theta} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \sum_{i=1}^{M} \sum_{j=-J}^{J} ||\mathcal{F}(\boldsymbol{y}_{i,j}; \boldsymbol{\Theta}) - \boldsymbol{x}_i||_1 \tag{20}$$

where $\boldsymbol{x}_i$ denotes the ith central clean frame to be recovered from its adjacent noisy frames $\boldsymbol{y}_{i,j}$, $j = -J, \ldots, J$ and $\mathcal{F}(\cdot; \boldsymbol{\Theta})$ denotes the function of the overall VD network with parameters $\boldsymbol{\Theta}$. The ADAM optimizer (Kingma and Ba 2015) was employed to train the network with parameters setting $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and the learning rate 10^{-4}. The proposed network is implemented under PyTorch platform and trained using 4 Nvidia Titan XP GPUs.

In addition to synthetic noise, we have conducted real-world raw video denoising experiments on the CRVD dataset (Yue et al. 2020). Specifically, we use raw data of 11 indoor scenes from CRVD dataset as our dataset, we randomly choose video frames of scene 8 and scene 9 as the test dataset, and video frames of other dynamic indoor scenes as the training dataset. In our current implementation, we first pack the Bayer images into 4 channels (RGBG) and then clip these packed frames into image patches with the size of 128×128. Finally, we derive 72,900 sequences for training and other 16,200 sequences for testing, each sequence contains 7 temporal-sequential real noisy image patches and a

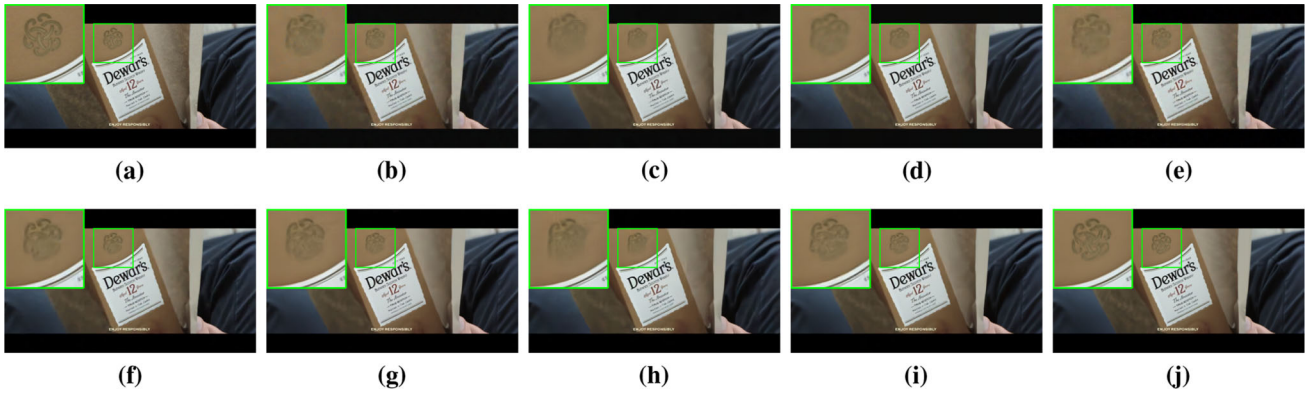

Fig. 6 Denoising results for a noisy frame from Vimeo test set with noise level $\sigma_n = 25$. **a** Original frame; denoised frame by **b** V-BM4D (Maggioni et al. 2012) (31.66 dB, 0.7169), **c** RTA-LSM (Dong et al. 2018a) (32.27 dB, 0.7429), **d** VNLB (Pablo and Jean 2018) (32.27 dB, 0.7427), **e** TOFlow (Xue et al. 2019) (35.13 dB, 0.8925), **f** TOFlow-EN (36.75 dB, 0.9484), **g** ViDeNN (Claus and van Gemert 2019) (37.65 dB, 0.9533), **h** VNLnet (Davy et al. 2019) (36.63 dB, 0.9462), **i** FastDVDnet (Tassano et al. 2020) (37.25 dB, 0.9546), and **j** MAP-VDNet (39.30 dB, 0.9627)

clean reference image patch, and there is no overlap between the training dataset and the test dataset.

Furthermore, we have extended the proposed model to other video processing tasks such as video superresolution and compressed video artifact reduction. For the task of video superresolution, we use the widely used training dataset—e.g., Vimeo dataset (Xue et al. 2019). Following the common setting (Jo et al. 2018; Yi et al. 2019; Isobe et al. 2020), the low-resolution frame was generated by first Gaussian filtering with a standard deviation of 1.6 and then $\times 4$ downsampling. Our network was trained on RGB channels and tested on the Y channel and RGB channels for video superresolution. For the task of compressed video artifact reduction, Vid70 (Yang et al. 2018) is a commonly used dataset in previous studies. To make a fair comparison, we apply the same training set and test set as Yang et al. (2018): 60 sequences in Vid70 for training and the remaining 10 sequences for testing. Besides, all compressed sequences were generated by HEVC standard, using HM 16.20 LDP mode with $QP = 37$ and 42. Similar to other competing algorithms, our network was trained and tested on the Y channel for compressed video artifact reduction.

5 Experimental Results

5.1 Experimental Setup

The proposed *MAP-VDNet* was trained using the Vimeo dataset (Xue et al. 2019), which consists of 91,701 sequences collected from 38,990 video clips. All collected sequences are divided into the training and test parts. Here, we consider two types of noise—i.e., additive white Gaussian and signal-dependent noise. In addition to the Vimeo dataset, we have also tested the proposed *MAP-VDNet* method on two sets of videos with different resolutions that are randomly selected from Derf's Test Media Collection,[1] denoted as ASU and DTMC-HD datasets, respectively. The sequences in the ASU dataset have a spatial resolution of 352×288 with a maximum of 300 frames, and the sequences in the DTMC-HD dataset have been downsampled to a resolution of 960×540 with 75 frames. Additive Gaussian noises with different variances were added to the clean frames to simulate noisy frames. Our network has been trained and tested on the RGB channels for VD. In our implementation, we adopted two optical flow estimation methods, i.e., the deep learning-based Spynet method (Ranjan and Black 2017) and the fast optical flow estimation method based on dense inverse search (Kroeger et al. 2016) which has been integrated into the OpenCV library. More comparative results are available at: https://see.xidian.edu.cn/faculty/wsdong/Projects/MAP-VDNet.htm.

5.2 Ablation Study

The effect of the core components To verify the effect of the core building blocks (i.e., frame alignment, adaptive fusion, and the denoiser module) of the proposed method on the denoising performance, we have implemented several variants of the proposed method as summarized below.

- *MAP-VDNet*: it denotes the proposed method that used the Spynet (Ranjan and Black 2017) for frame alignment, spatially adaptive fusion, and U-net denoiser.
- *MAP-VDNet-Fast*: it denotes the proposed method that used the fast optical flow estimation method (Kroeger et al. 2016), spatially adaptive fusion, and U-net denoiser.

[1] https://media.xiph.org/video/derf/.

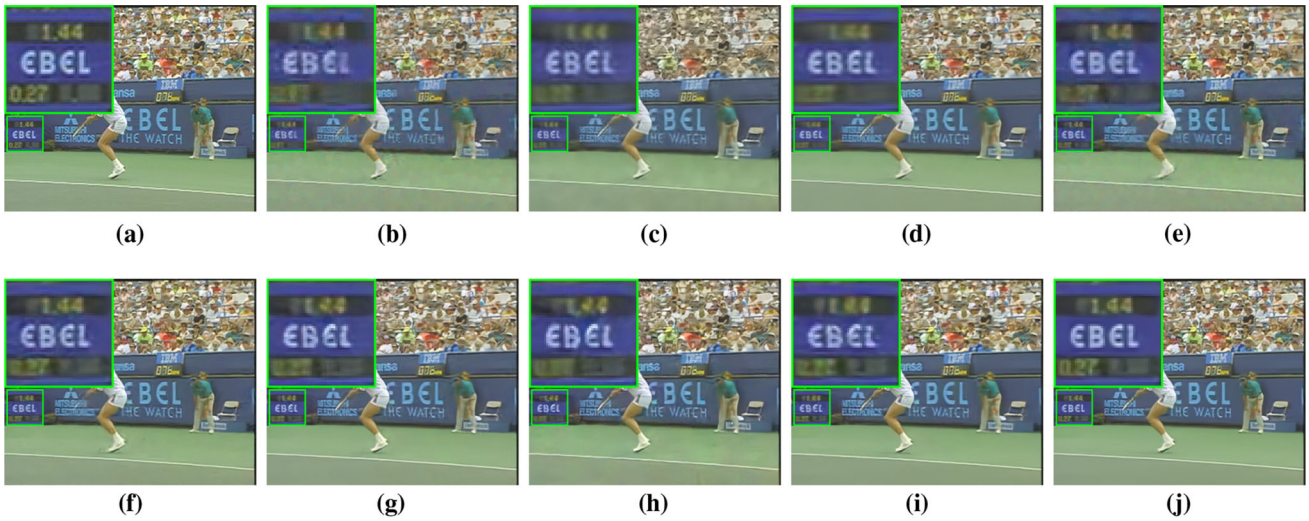

Fig. 7 Denoising results for a noisy frame of *Stefan* video of ASU test set with noise level $\sigma_n = 50$. **a** Original frame; denoised frame by **b** V-BM4D (Maggioni et al. 2012) (25.16 dB, 0.7965), **c** RTA-LSM (Dong et al. 2018a) (25.98 dB, 0.8731), **d** VNLB (Pablo and Jean 2018) (26.89 dB, 0.8885), **e** TOFlow (Xue et al. 2019) (26.70 dB, 0.8808), **f** TOFlow-EN (28.10 dB, 0.9155), **g** ViDeNN (Claus and van Gemert 2019) (26.74 dB, 0.8905), **h** VNLnet (Davy et al. 2019) (26.10 dB, 0.8751), **i** FastDVDnet (Tassano et al. 2020) (26.87 dB, 0.8951), and **j** MAP-VDNet (29.12 dB, 0.9312)

– *MAP-VDNet-NAF*: it denotes the proposed method that used the Spynet (Ranjan and Black 2017) for frame alignment, non-adaptive fusion, and U-net denoiser.
– *MAP-VDNet-DnCNN*: it denotes the proposed method that used the Spynet (Ranjan and Black 2017) for frame alignment, spatially adaptive fusion, and DnCNN denoiser (Zhang et al. 2017a).

The above variants of the proposed method contain five denoising stages, i.e., $T = 5$. To verify the effectiveness of the proposed MAP-VDNet, we also implemented the following video denoising method,

– VD-Unet: it denotes the video denoising method that first uses the Spynet (Ranjan and Black 2017) to align the noisy frames and then sends the aligned $2J + 1$ frames to the U-net denoiser to output the denoised central frame.
– VD-Unet-Fast: it denotes the video denoising method that first uses the fast optical flow estimation method (Kroeger et al. 2016) to align the noisy frames and then sends the aligned $2J + 1$ frames to the U-net denoiser to output the denoised central frame.

In the above methods, we set $J = 3$. The average PSNR and SSIM results of these variants of the proposed method for noise level 25 are shown in Table 1. From Table 1 one can see that all variants of the proposed method outperform the VD-Unet and VD-Unet-Fast methods. The improvement over the VD-Unet-Fast method is much larger, verifying the effectiveness of the proposed MAP-VDNet method. When using fast optical flow estimation, the MAP-VDNet-Fast is slightly worse than MAP-VDNet. Without using the spatially adaptive fusion, the performance of MAP-VDNet-NAF is also worse than MAP-VDNet, and the performance gap is up to 0.32 dB, which demonstrates the effectiveness of the proposed spatially-adaptive fusion. When comparing MAP-VDNet-DnCNN with MAP-VDNet, we can see that performance gaps between them are small, which are in the range of (0.07–0.21) dB, verifying that the proposed method is non-sensitive to the choice of the deep denoisers.

The effect of the number of stages To show how the number of stages T (i.e., the number of iterations in **Algorithm 1**) affects the denoising performance, we have compared the proposed *MAP-VDNet* method with different stages. Fig. 4 shows the average PSNR performance as a function of $T \in [1, 6]$ for $\sigma_n = 25$ on two test datasets. It can be observed from Fig. 4 that more stages do lead to improved denoising performance. However, the performance improvements quickly and become saturated when $T \geq 5$. For the balance of performance and computational complexity, we have manually set $T = 5$ in our current implementation.

The effect of the number of frames To verify the impact of the number of frames on VD performance, we have conducted comparative studies of the proposed *MAP-VDNet* with different numbers of input noisy frames: $J = 1$, $J = 2$, and $J = 3$. The average PSNR/SSIM results on Vimeo and DTMC-HD datasets are shown in Table 2. It can be observed that the performance monotonically increases as the number of input frames increases. Therefore, we have set $J = 3$ in our current implementation.

The effect of dense connections between the denoisers Finally, to demonstrate the effects of dense connections between the denoisers, we have conducted experiments comparing the networks with and without dense connections (denoted as *MAP-VDNet-wo*). The average results on Vimeo and DTMC-HD datasets are shown in Table 3. One can see that dense connections between denoisers are added, the average improvements on Vimeo and DTMC-HD datasets are 0.12 dB and 0.14 dB, respectively. These results verify the effectiveness of dense connections between denoisers in *MAP-VDNet*.

The inference process of the proposed iterative method In Fig. 5, we show visual and PSNR comparisons of some intermediate variables (e.g., aligned frames $\tilde{\boldsymbol{y}}_j (j = -J, \ldots, J)$, the average fused frame $\overline{\boldsymbol{y}}$, the spatially adaptive fused frame $\overline{\boldsymbol{y}}'$ and outputs from all stages $\boldsymbol{x}^{(t)} (t = 1 \ldots, 5)$) to provide a more detailed description for the inference process of the proposed method. As shown in Fig. 3b, we first feed the reference frame $\boldsymbol{y}_0$ and each noisy frame $\boldsymbol{y}_j (j = -J, \ldots, J, j \neq 0)$ into the alignment module and derive aligned frames $\tilde{\boldsymbol{y}}_j$. From the visualization in the top row of Fig. 5, we can see that there still exists some misaligned pixels in the aligned noisy frames $\tilde{\boldsymbol{y}}_j (j = -J, \ldots, J, j \neq 0)$. Inaccurate alignment could reduce denoising performance. To avoid the adverse influence of alignment bias, we introduce the spatially adaptive fusion derived from the MAP estimation with non-i.i.d Gaussian likelihood. Through visual comparisons of the average fusion $\overline{\boldsymbol{y}} = \frac{1}{(2J+1)} \sum_{j=-J}^{J} \tilde{\boldsymbol{y}}_j$ and the spatially adaptive fusion $\overline{\boldsymbol{y}}' = \sum_{j=-J}^{J} \mathbf{C}_j \tilde{\boldsymbol{y}}_j$ in Fig. 5, it can be observed that average fusion would yield undesirable artifacts caused by the alignment bias, while the output from spatially adaptive fusion has sharper edges. Therefore, spatially adaptive fusion is more effective in handling large-motion videos. Finally, we implement iterative optimizations of Eq. (19), and the visual results from all intermediate stages are demonstrated in the bottom row of Fig. 5. According to intermediate comparative results from different stages (e.g., $\boldsymbol{x}^{(t)}$), one can see that more image details could be restored and the denoising performance (PSNR) has been improved with the increasing number of stages, the final output $\boldsymbol{x}^{(5)}$ achieved the best denoising performance.

5.3 Comparison with Other State-of-the-Art Methods

We have compared the performance of the proposed *MAP-VDNet-Fast* and *MAP-VDNet* methods with several other state-of-the-art methods including both model-based (i.e., V-BM4D (Maggioni et al. 2012), RTA-LSM (Dong et al. 2018a) and VNLB (Pablo and Jean 2018)) and recently developed deep learning-based methods (i.e., TOFlow (Xue et al. 2019), ViDeNN (Claus and van Gemert 2019), VNLnet (Davy et al. 2019) and FastDVDnet (Tassano et al. 2020)). For a fair comparison, we have improved the TOFlow method (Xue et al. 2019) by adding more convolutional residual blocks, such that the total number of network parameters of TOFlow is comparable with that of the proposed *MAP-VDNet*. The enhanced TOFlow method is denoted as TOFlow-EN. All deep-learning based competing methods (i.e., TOFlow (Xue et al. 2019), TOFlow-EN, ViDeNN (Claus and van Gemert 2019), VNLnet (Davy et al. 2019) and FastDVDnet (Tassano et al. 2020)) were retrained on the same Vimeo training dataset, and we set the same number of input frames for these competing methods (i.e., 7) for fairness. Specifically, for VNLnet, we have retrained the color denoising model and the search space of the nonlocal patches is reduced to seven frames, and for ViDeNN and FastDVDnet, we increased the number of input frames to 7 and leading to a three-step cascaded version of FastDVDnet.

Table 4 shows the average results of the test methods on the Vimeo, ASU, and DTMC-HD test datasets. From Table 4, we can see that the VNLB method (Pablo and Jean 2018) is the most competitive algorithm among model-based VD methods. By adding more convolutional layers, the TOFlow-EN method significantly outperforms the original TOFlow method and becomes competitive with other deep learning-based state-of-the-art video denoising methods such as ViDeNN and FastDVDnet methods. With a similar number of parameters, the performance gains of MAP-VDNet over TOFlow-EN demonstrate the effectiveness of our maximum a posterior estimation framework. Both the proposed MAP-VDNet-Fast and MAP-VDNet methods significantly outperform the other competing methods. With a more accurate estimation of optical flow, the MAP-VDNet method performs slightly better than its counterpart that uses a fast optical flow estimation method. The visual comparisons of the denoised frames by the test methods are shown in Figs. 6, 7 and 8. From these figures, one can see that the proposed *MAP-VDNet* method can reproduce sharper edges and more details than the other competing methods.

5.4 Video Denoising with Signal-Dependent Noise

Although the additive Gaussian noise is widely used in existing denoising studies, the distribution of the real noise of imaging sensors is much more complicated. To improve the denoising performance on real-world noisy images, more realistic signal-dependent noise models have been proposed (Foi et al. 2008). Here, we have adopted the Poisson-Gaussian noise model (Foi et al. 2008) to simulate more realistic noisy frames as follows

$$y_{j,p} \sim \mathcal{N}(x_{j,p}, \sigma_r^2 + \sigma_s x_{j,p}), \tag{21}$$

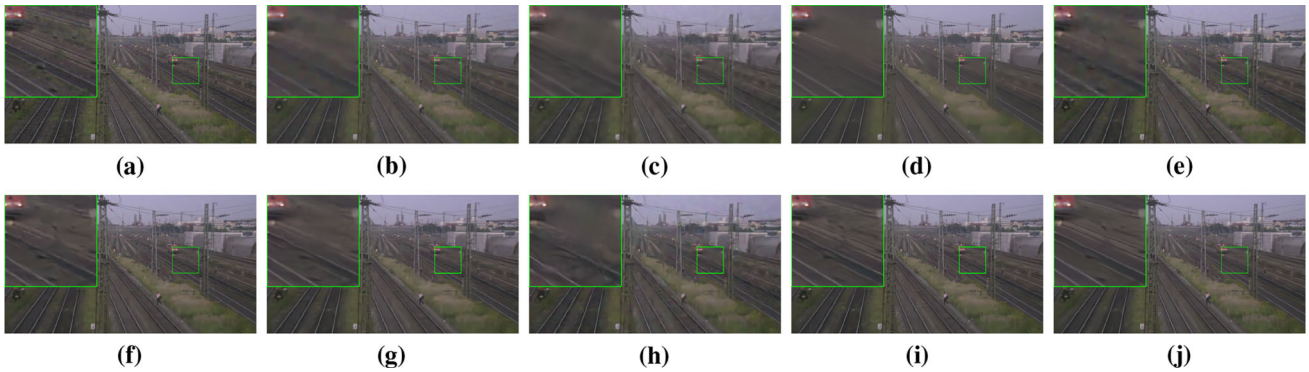

Fig. 8 Denoising results for a noisy frame of *Station* video of DTMC-HD test set with noise level $\sigma_n = 50$. **a** Original frame; denoised frame by **b** V-BM4D (Maggioni et al. 2012) (30.74 dB, 0.7445), **c** RTA-LSM (Dong et al. 2018a) (30.63 dB, 0.7682), **d** VNLB (Pablo and Jean 2018) (31.78 dB, 0.8020), **e** TOFlow (Xue et al. 2019) (31.99 dB, 0.8029), **f** TOFlow-EN (32.99 dB, 0.8391), **g** ViDeNN (Claus and van Gemert 2019) (32.72 dB, 0.8274), **h** VNLnet (Davy et al. 2019) (30.79 dB, 0.7759), **i** FastDVDnet (Tassano et al. 2020) (33.25 dB, 0.8497), and **j** MAP-VDNet (33.82 dB, 0.8608)

Table 5 The average PSNR/SSIM denoising results on different test sets with signal-dependent noise (boldface highlights the best)

Dataset	Vimeo		ASU		DTMC-HD	
Noisy	22.10	0.3882	23.36	0.5826	21.60	0.4234
CBDNet (Guo et al. 2019)	33.49	0.8636	30.67	0.8687	31.69	0.8309
TOFlow (Xue et al. 2019)	33.04	0.8602	30.55	0.8770	33.33	0.8817
TOFlow-EN	36.30	0.9296	32.34	0.9088	34.46	0.9021
MAP-VDNet	**37.08**	**0.9329**	**33.17**	**0.9307**	**35.14**	**0.9189**

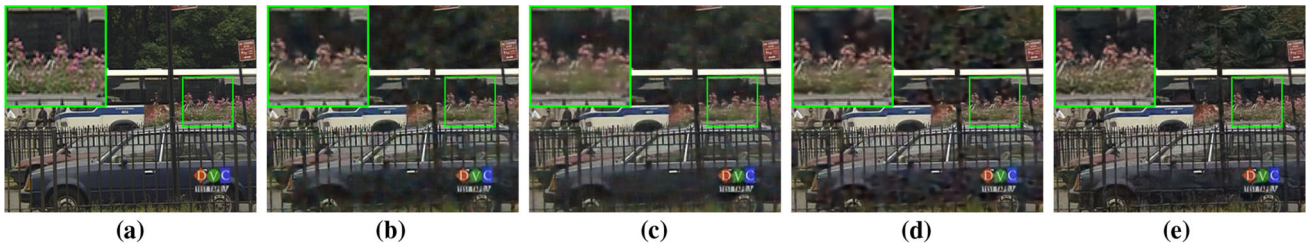

Fig. 9 Denoising results of a noisy frame from *Bus* sequence of ASU test set for signal-dependent noise. **a** The original frame; the frames denoised by **b** CBDNet (Guo et al. 2019) (26.42 dB, 0.7729), **c** TOFlow (Xue et al. 2019) (26.36 dB, 0.7839), **d** TOFlow-EN (26.65 dB, 0.7842), and **e** MAP-VDNet (28.09 dB, 0.8612)

where $x_{j,p}$ is the clean pixel at position p, σ_r and σ_s are related to the sensor gain (ISO) of cameras. Similar to Guo et al. (2019), we simulate the noisy frames by randomly sampling σ_s and σ_r from the ranges $[0, 0.16]$ and $[0, 0.06]$, respectively. Then we retrained the proposed MAP-VDNet, the TOFlow (Xue et al. 2019) and its enhanced version TOFlow-EN on the new noisy training dataset. We have also compared with the recently developed blind image denoising method—i.e., CBDNet (Guo et al. 2019) that was also trained with the same realistic noise model. Table 5 shows the denoising results by the test methods on the Vimeo, ASU and DTMC-HD datasets. From these tables, we can see that *MAP-VDNet* performs much better than the competing methods. The average PSNR gains over the TOFlow-EN method (the second best) are 0.78 dB, 0.83 dB and 0.68 '1 on Vimeo, ASU and DTMC-HD datasets, respectively. Parts of the denoised frames by the test methods are shown in Fig. 9. It can be seen that the proposed *MAP-VDNet* can faithfully recover more details around the texture regions than other competing methods.

5.5 Raw Video Denoising with Real-World Noise

Recently, a raw video dataset with realistic noise is proposed, i.e., the CRVD dataset (Yue et al. 2020). Since the characteristics of real-world noise are quite different from that of simulated noise, we conduct raw video denoising comparisons on the CRVD dataset (Yue et al. 2020) to verify

the real-world noise removal ability of the proposed MAP-VDNet.

The competing methods include the V-BM4D method (Maggioni et al. 2012), the TOFlow-EN method and the RViDeNet method (Yue et al. 2020). For fairness, we retrained TOFlow-EN and RViDeNet on the same CRVD training dataset and increased the input frames to 7 for RViDeNet (Yue et al. 2020). Due to the difference between sRGB video denoising and raw video denoising, some adaptions need to be conducted for these sRGB video denoising methods. For example, for 4-channel raw data, the V-BM4D method need to denoise the data of each channel separately. Regarding TOFlow-EN and MAP-VDNet methods, to exploit the information among channels, we average the 4-channel raw data to estimate optical flow in the alignment module, other processes are similar to sRGB video denoising. When training RViDeNet, since we don't consider the image signal processing (ISP) pipeline, we no longer apply the sRGB domain loss in Yue et al. (2020), and we follow the training setting in Yue et al. (2020) that first pretrain a denoiser using an additional MOT Challenge dataset (Milan et al. 2016) and then keep the predenoising network fixed when training the RViDeNet on the same CRVD training dataset.

The real raw video denoising results have been shown in Table 6, one can see that all deep-learning-based video denoising methods outperform the model-based V-BM4D method. Among the supervised learning methods, the PSNR gain of the proposed MAP-VDNet over TOFlow-EN with a comparable number of parameters is 3.26 dB for the CRVD test dataset. In addition, even though the number of parameters of RViDeNet is approximately triple of ours, the proposed method still outperforms RViDeNet by 1.46 dB. The experimental results in Table 6 have proved that the proposed MAP-VDNet could achieve excellent performance for the task of real-world raw video denoising.

5.6 Video Superresolution

Recently, video superresolution has been widely studied in Caballero et al. (2017); Tao et al. (2017); Sajjadi et al. (2018); Jo et al. (2018); Xue et al. (2019); Haris et al. (2019); Wang et al. (2019); Yi et al. (2019); Tian et al. (2020); Isobe et al. (2020). In this paper, to verify the generalization ability of our proposed framework, we make the proposed MAP-VDNet applied to the video superresolution by adding an upscale module before the alignment module in Fig. 3a. For the architecture of the $\times 4$ up-scale module, we employ a simple up-scale network without any special design and a sub-pixel convolutional layer (Shi et al. 2016) to conduct the upsampling operation.

The competing methods consist of the FRVSR method (Sajjadi et al. 2018), the DUF method (Jo et al. 2018), the TOFlow method (Xue et al. 2019), the RBPN method (Haris et al. 2019), the EDVR method (Wang et al. 2019), the PFNL method (Yi et al. 2019) and the TGA method (Isobe et al. 2020). Note that other comparative results come from their own publication or recent work (Isobe et al. 2020). Since Vid4 (Liu and Sun 2014) is a benchmark test set for video superresolution, we demonstrate the comparisons of the performance and number of parameters with other competing methods on Vid4 test set for $\times 4$ video superresolution in Table 7. From the results, one can see that the proposed MAP-VDNet outperforms most competing methods except the TGA method (Isobe et al. 2020). Although the performance reflected by PSNR of our method is slightly lower than the TGA method (Isobe et al. 2020), another metric (SSIM) demonstrates the performance of our method is better and the number of parameters of the proposed MAP-VDNet is far less than the TGA method (Isobe et al. 2020). The visual comparisons are illustrated in Fig. 10, the proposed MAP-VDNet could restore more clear details than other competing methods.

5.7 Compressed Video Artifact Reduction

In addition to video denoising and video superresolution, compressed video quality enhancement (Yang et al. 2017, 2018; Lu et al. 2018; Yang et al. 2019a; Xu et al. 2019b; Yang et al. 2019b; Deng et al. 2020; Guan et al. 2021 is an important and challenging video processing task. The quality fluctuation among compressed frames increases the difficulty of recovery. We verify the effectiveness of our MAP-VDNet model on the compressed video artifact reduction task. We compared with the DSCNN method (Yang et al. 2017), the MFQE method (Yang et al. 2018), the QG-ConvLSTM method (Yang et al. 2019a), the NL-ConvLSTM method (Xu et al. 2019b) and the MFQE 2.0 method (Guan et al. 2021). And the results of other competing algorithms are obtained from their publications. Among the competing methods, the MFQE 2.0 method (Guan et al. 2021) was trained on a more comprehensive training dataset which consists of 108 videos. Due to the sensitivity of MFQE 2.0 to the training dataset, we reported the results in their published paper. Note that these competing methods usually design their algorithms specifically for compressed video frames encoded by a fixed quality factor (QP value) -e.g., the NL-ConvLSTM method (Xu et al. 2019b) introduces a non-local strategy into the ConvLSTM module to learn spatiotemporal correlation across frames, and the MFQE methods (Yang et al. 2018; Guan et al. 2021) exploit peak quality frames that located by the detector to enhance the quality of adjacent compressed frames. Their generalization property often remains questionable; by contrast, our algorithm is not optimized for the task of compressed video quality enhancement but still achieves excellent performance. Table 8 illustrates the average PSNR gain (dB) on the test sequences of the

Table 6 The average PSNR/SSIM denoising results on CRVD test dataset for real raw video denoising (boldface highlights the best)

Metrics	PSNR (dB)	SSIM	Para. (M)
Noisy	32.34	0.7145	–
V-BM4D (Maggioni et al. 2012)	42.66	0.9712	–
TOFlow-EN	44.23	0.9825	2.87
RViDeNet (Yue et al. 2020)	46.03	0.9881	8.58
MAP-VDNet	47.49	0.9912	2.85

Table 7 The average PSNR/SSIM super-resolution results on Vid4 test set for $\times 4$ video super-resolution (boldface highlights the best)

Dataset	Y channel		RGB channels		Para. (M)
Bicubic	21.81	0.5458	20.36	0.5140	–
TOFlow* (Xue et al. 2019)	25.84	0.7659	24.39	0.7438	1.4
FRVSR† (Sajjadi et al. 2018)	26.69	0.8220	–		5.1
RBPN* (Haris et al. 2019)	27.17	0.8205	25.65	0.7997	12.1
EDVR-L† (Wang et al. 2019)	27.35	0.8264	25.83	0.8077	20.6
DUF-52L* (Jo et al. 2018)	27.38	0.8329	25.91	0.8166	5.8
PFNL† (Yi et al. 2019)	27.40	0.8384	–		3.0
TGA† (Isobe et al. 2020)	**27.59**	0.8419	**26.10**	0.8254	5.8
MAP-VDNet	27.49	**0.8438**	25.95	**0.8257**	3.0

† denotes the results from their publications, and * denotes the results from Isobe et al. (2020)

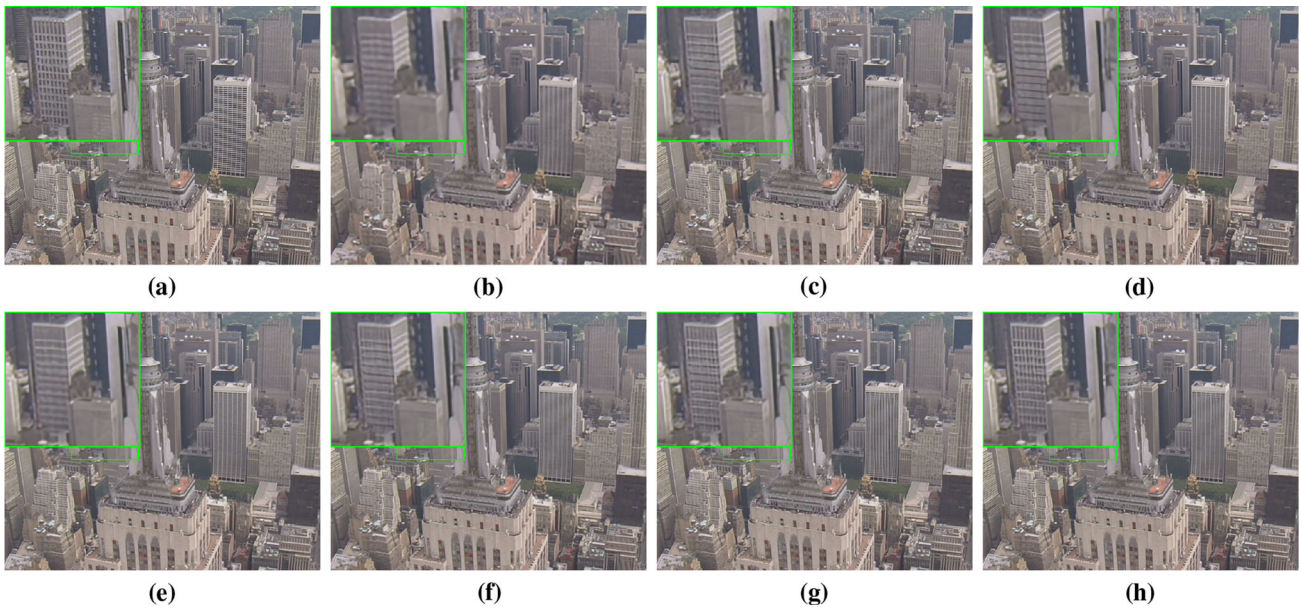

Fig. 10 Video superresolution results ($\times 4$ up-sampling) for a low-resolution frame of *City* video of Vid4 test set. **a** Original frame; superresolved frame by **b** TOFlow (Xue et al. 2019) (25.40 dB, 0.7081), **c** FRVSR (Sajjadi et al. 2018) (26.78 dB, 0.8136), **d** RBPN (Haris et al. 2019) (26.51 dB, 0.7929), **e** EDVR-L (Wang et al. 2019) (26.86 dB, 0.7995), **f** DUF-52L (Jo et al. 2018) (26.89 dB, 0.8147), **g** PFNL (Yi et al. 2019) (26.94 dB, 0.8298), **h** MAP-VDNet (27.28 dB, 0.8369)

Table 8 The average ΔPSNR (dB) on test sequences of the Vid70 dataset (boldface highlights the best)

QP	Seq.	DS- CNN	MFQE 1.0	QG-Conv LSTM	NL-Conv LSTM	MFQE 2.0	MAP- VDNet
37	1	0.492	0.402	–	0.501	**0.723**	0.479
	2	0.458	0.484	–	0.563	**0.728**	0.575
	3	0.271	0.394	–	0.439	0.594	**0.607**
	4	0.393	0.550	–	0.598	**0.735**	0.664
	5	0.356	0.598	–	0.658	**0.719**	0.640
	6	0.435	0.390	–	0.394	**0.476**	0.456
	7	0.277	0.472	–	0.483	0.550	**0.554**
	8	0.230	0.604	–	0.971	0.775	**1.050**
	9	0.271	0.438	–	0.576	**0.579**	0.524
	10	0.274	0.772	–	0.827	**0.920**	0.889
	Ave.	0.344	0.510	0.587	0.601	**0.680**	0.644
42	Ave.	0.364	0.461	0.601	**0.614**	–	0.582

1: *MaD* 2: *BasketballPass* 3: *RaceHorses* 4: *Vidyo1* 5: *Vidyo3* 6: *Vidyo4* 7: *Kimono* 8: *TunnelFlag* 9: *BarScene* 10: *PeopleOnStreet*

Table 9 Running time on 90 frames of size 352 × 288

Methods	TOFlow	TOFlow-EN	ViDeNN	VNLnet
Time (s)	17.97	23.02	22.03	28.30
Methods	FastDVDnet	MAP-VDNet-Fast		MAP-VDNet
Time (s)	11.52	16.79		27.40

$Vid70$ dataset. The results in Table 8 have shown that the proposed MAP-VDNet model is also effective for the task of compressed video artifact reduction, which justifies its excellent generalization property.

5.8 Complexity Analysis and Discussions

We have compared the proposed *MAP-VDNet-Fast* and *MAP-VDNet* method with other deep-learning-based competing methods in Fig. 1. The comparison shows the trade-off between the complexity (as measured by the number of parameters) and the denoising performance (average PSNR values). It can be observed that with a similar number of parameters, our *MAP-VDNet* outperforms TOFlow-EN over 1dB on DTMC-HD test set for $\sigma_n = 25$. With fewer number of parameters, our *MAP-VDNet-Fast* could also achieve excellent denoising performance compared to other competing video denoising algorithms. Additionally, Table 9 shows the comparison of the actual running time of different denoising methods on a Nvidia Titan XP GPU. By using a fast optical flow estimation method that is implemented with OpenCV library and only runs on CPU, the running time of our *MAP-VDNet-Fast* is comparable to that of TOFlow (note that ours outperforms TOFlow by over 2dB as shown in Fig. 1).

Differences with other deep unfolding networks To the best of our knowledge, this is the first work that uses optimization formation to guide the design of deep neural networks for video denoising. When compared with the image denoising problem (Zhang et al. 2017b; Dong et al. 2018b), the video denoising problem is more difficult to handle because the temporal redundancy among adjacent frames needs to be modeled and exploited. In this paper, we formulate the multiframe fusion module in the MAP estimation architecture, and each iteration can be efficiently computed by the proposed MAP-VDNet. Moreover, this work explicitly takes misalignment errors into account and demonstrates a principled solution based on robust multiframe fusion.

6 Conclusions

In this paper, we propose a novel MAP-based video denoising algorithm *MAP-VDNet*. Unlike existing model-based Bayesian video denoising, we strive to optimize the parameters of a network-based denoising algorithm without involving hand-crafted procedures. Different from previous deep learning-based methods with heuristically designed network architectures, the construction of *MAP-VDNet* is based on an explainable optimization-inspired solution to video denoising derived from the classical MAP estimation framework. Specifically, we first propose an iterative MAP-based video denoising algorithm based on a realistic observation model of noisy frames, which allows us to solve the denoising problem

in a principled manner. Combining a DCNN-based image denoiser module with an optical flow-based image alignment module, we then demonstrate how to unfold the iterative video denoising algorithm into a multistage implementation in which both image denoising and image fusion modules can be jointly trained. Moreover, we propose a robust multiframe fusion scheme by predicting the adaptive fusion coefficient on a pixel-by-pixel basis, which further improves the performance of video denoising in the presence of misalignment. Extensive experimental results on three video datasets show that the proposed method significantly outperforms the existing video denoising methods in terms of both objective and subjective quality of restored video frames. Additionally, we have achieved promising experimental results on other video restoration tasks such as real-world video denoising, video superresolution and compressed video artifact reduction, which demonstrate the good generalization property of the proposed MAP-VDNet.

Acknowledgements This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0101400, and the Natural Science Foundation of China under Grants 61991451, 61632019, 61621005, and 61836008. Xin Li's work is partially supported by the NSF under Grants IIS-1951504 and OAC-1940855, the DoJ/NIJ under Grant NIJ 2018-75-CX-0032, and the WV Higher Education Policy Commission Grant (HEPC.dsr.18.5).

References

Arias, P., & Morel, J. (2019). Kalman filtering of patches for frame-recursive video denoising. In *IEEE conference on computer vision and pattern recognition workshops, CVPR workshops 2019*, Long Beach, CA, USA, June 16–20, 2019 (pp. 1917–1926).

Bao, W., Lai, W. S., Zhang, X., Gao, Z., & Yang, M. H. (2019). Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, *5*(2), 157–166.

Bertocchi, C., Chouzenoux, E., Corbineau, M. C., Pesquet, J. C., & Prato, M. (2020). Deep unfolding of a proximal interior point method for image restoration. *Inverse Problems*, *36*(3), 034005.

Buades, A., Coll, B., & Morel, J. M. (2005). A non-local algorithm for image denoising. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, *2*, 60–65.

Buades, A., Lisani, J. L., & Miladinović, M. (2016). Patch-based video denoising with optical flow estimation. *IEEE Transactions on Image Processing*, *25*(6), 2573–2586.

Caballero, J., Ledig, C., Aitken, A. P., Acosta, A., Totz, J., Wang, Z., & Shi, W. (2017). Real-time video super-resolution with spatio-temporal networks and motion compensation. In *2017 IEEE conference on computer vision and pattern recognition, CVPR 2017*, Honolulu, HI, USA, July 21–26, 2017 (pp 2848–2857).

Chang, S. G., Yu, B., & Vetterli, M. (2000). Wavelet thresholding for multiple noisy image copies. *IEEE Transactions on Image Processing*, *9*(9), 1631–1635.

Claus, M., & van Gemert, J. (2019). Videnn: Deep blind video denoising. In *IEEE conference on computer vision and pattern recognition workshops, CVPR workshops 2019*, Long Beach, CA, USA, June 16–20, 2019 (pp. 1843–1852).

Dabov, K., Foi, A., Katkovnik, V., & Egiazarian, K. (2007). Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, *16*(8), 2080–2095.

Davy, A., Ehret, T., Morel, J. M., Arias, P., & Facciolo, G. (2019). A non-local CNN for video denoising. In *2019 IEEE international conference on image processing (ICIP)* (pp. 2409–2413).

Deng, J., Wang, L., Pu, S., & Zhuo, C. (2020). Spatio-temporal deformable convolution for compressed video quality enhancement. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 10696–10703).

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

Dong, W., Huang, T., Shi, G., Ma, Y., & Li, X. (2018a). Robust tensor approximation with Laplacian scale mixture modeling for multiframe image and video denoising. *IEEE Journal of Selected Topics in Signal Processing*, *12*(6), 1435–1448.

Dong, W., Wang, P., Yin, W., Shi, G., Wu, F., & Lu, X. (2018b). Denoising prior driven deep neural network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(10), 2305–2318.

Ehret, T., Davy, A., Morel, J. M., Facciolo, G., & Arias, P. (2019). Model-blind video denoising via frame-to-frame training. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 11369–11378).

Elad, M., & Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, *15*(12), 3736–3745.

Foi, A., Trimeche, M., Katkovnik, V., & Egiazarian, K. (2008). Practical Poissonian–Gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, *17*(10), 1737–1754.

Godard, C., Matzen, K., & Uyttendaele, M. (2018). Deep burst denoising. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 538–554).

Gu, S., Zhang, L., Zuo, W., & Feng, X. (2014). Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2862–2869).

Guan, Z., Xing, Q., Xu, M., Yang, R., Liu, T., & Wang, Z. (2021). MFQE 2.0: A new approach for multi-frame quality enhancement on compressed video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(3), 949–963.

Guo, S., Yan, Z., Zhang, K., Zuo, W., & Zhang, L. (2019). Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1712–1722).

Haris, M., Shakhnarovich, G., & Ukita, N. (2019). Recurrent back-projection network for video super-resolution. In *IEEE conference on computer vision and pattern recognition, CVPR 2019*, Long Beach, CA, USA, June 16–20, 2019 (pp. 3897–3906).

Hasinoff, S. W., Sharlet, D., Geiss, R., Adams, A., Barron, J. T., Kainz, F., et al. (2016). Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (TOG)*, *35*(6), 1–12.

Hershey, J. R., Roux, J. L., & Weninger, F. (2014). Deep unfolding: Model-based inspiration of novel deep architectures. arXiv preprint arXiv:1409.2574

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Isobe, T., Li, S., Jia, X., Yuan, S., Slabaugh, G. G., Xu, C., Li, Y., Wang, S., & Tian, Q. (2020). Video super-resolution with temporal group

attention. In *2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020*, Seattle, WA, USA, June 13–19, 2020 (pp. 8005–8014).

Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial transformer networks. In *Advances in neural information processing systems* (pp. 2017–2025).

Ji, H., Liu, C., Shen, Z., & Xu, Y. (2010). Robust video denoising using low rank matrix completion. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE (pp. 1791–1798).

Jo, Y., Oh, S. W., Kang, J., & Kim, S. J. (2018) Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *2018 IEEE conference on computer vision and pattern recognition, CVPR 2018*, Salt Lake City, UT, USA, June 18–22, 2018 (pp. 3224–3232).

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd international conference on learning representations, ICLR 2015*, San Diego, CA, USA, May 7–9, 2015, *Conference track proceedings*.

Kokkinos, F., & Lefkimmiatis, S. (2019). Iterative residual CNNs for burst photography applications. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5929–5938).

Kroeger, T., Timofte, R., Dai, D., & Gool, L. V. (2016). Fast optical flow using dense inverse search. In *European conference on computer vision* (pp. 471–488).

Liu, C., & Freeman, W. T. (2010). A high-quality video denoising algorithm based on reliable motion estimation. In *European conference on computer vision* (pp. 706–719). Springer.

Liu, C., & Sun, D. (2014). On Bayesian adaptive video super resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(2), 346–360.

Lu, G., Ouyang, W., Xu, D., Zhang, X., Gao, Z., & Sun, M. (2018). Deep Kalman filtering network for video compression artifact reduction. In *Computer Vision—ECCV 2018—15th European conference*, Munich, Germany, September 8–14, 2018, Proceedings, Part XIV (pp. 591–608).

Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on artificial intelligence, IJCAI '81*, Vancouver, BC, Canada, August 24–28, 1981 (pp. 674–679).

Maggioni, M., Boracchi, G., Foi, A., & Egiazarian, K. (2012). Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms. *IEEE Transactions on Image Processing*, *21*(9), 3952–3966.

Mahmoudi, M., & Sapiro, G. (2005). Fast image and video denoising via nonlocal means of similar neighborhoods. *IEEE Signal Processing Letters*, *12*(12), 839–842.

Milan, A., Leal-Taixé, L., Reid, I.D., Roth, S., & Schindler, K. (2016). MOT16: A benchmark for multi-object tracking. CoRR arXiv:1603.00831.

Mildenhall, B., Barron, J. T., Chen, J., Sharlet, D., Ng, R.,& Carroll, R. (2018). Burst denoising with kernel prediction networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2502–2510).

Pablo, A., & Jean, M. M. (2018). Video denoising via empirical Bayesian estimation of space-time patches. *Journal of Mathematical Imaging and Vision*, *60*(1), 70–93.

Portilla, J., Strela, V., Wainwright, M. J., & Simoncelli, E. P. (2003). Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Transactions on Image processing*, *12*(11), 1338–1351.

Ranjan, A., Black, M. J. (2017). Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4161–4170).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Sajjadi, M. S. M., Vemulapalli, R., & Brown, M. (2018). Frame-recurrent video super-resolution. In *2018 IEEE conference on computer vision and pattern recognition, CVPR 2018*, Salt Lake City, UT, USA, June 18–22, 2018 (pp. 6626–6634).

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE conference on computer vision and pattern recognition, CVPR 2016*, Las Vegas, NV, USA, June 27–30, 2016 (pp. 1874–1883).

Tai, Y., Yang, J., Liu, X., & Xu, C. (2017). Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision* (pp. 4539–4547).

Tao, X., Gao, H., Liao, R., Wang, J., & Jia, J. (2017). Detail-revealing deep video super-resolution. In *IEEE International conference on computer vision, ICCV 2017* Venice, Italy, October 22-29, 2017 (pp. 4482–4490).

Tassano, M., Delon, J., & Veit, T. (2019). DVDNET: A fast network for deep video denoising. In *IEEE international conference on image processing, ICIP* (pp. 1805–1809).

Tassano, M., Delon, J., & Veit, T. (2020). Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020*, Seattle, WA, USA, June 13–19, 2020 (pp. 1351–1360).

Tekalp, A. M. (2015). *Digital video processing*. Englewood Cliffs: Prentice Hall.

Tian, Y., Zhang, Y., Fu, Y., & Xu, C. (2020). TDAN: Temporally-deformable alignment network for video super-resolution. In *2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020*, Seattle, WA, USA, June 13–19, 2020 (pp. 3357–3366).

Varghese, G., & Wang, Z. (2010). Video denoising based on a spatiotemporal Gaussian scale mixture model. *IEEE Transactions on Circuits and Systems for Video Technology*, *20*(7), 1032–1040.

Wang, X., Chan, K. C. K., Yu, K., Dong, C., & Loy, C. C. (2019). EDVR: video restoration with enhanced deformable convolutional networks. In *IEEE conference on computer vision and pattern recognition workshops, CVPR workshops 2019*, Long Beach, CA, USA, June 16-20, 2019 (pp. 1954–1963).-

Wisdom, S., Powers, T., Pitton, J., & Atlas, L. (2017). Building recurrent networks by unfolding iterative thresholding for sequential sparse recovery. *2017 IEEE international conference on acoustics* (pp. 4346–4350). IEEE: Speech and Signal Processing (ICASSP).

Xu, X., Li, M., Sun, W. (2019a). Learning deformable kernels for image and video denoising. arXiv preprint arXiv:1904.06903

Xu, X., Li, M., Sun, W., & Yang, M. H. (2020). Learning spatial and spatio-temporal pixel aggregations for image and video denoising. *IEEE Transactions on Image Processing*, *29*, 7153–7165.

Xu, Y., Gao, L., Tian, K., Zhou, S., & Sun, H. (2019b). Non-local ConvLSTM for video compression artifact reduction. In *IEEE/CVF international conference on computer vision, ICCV* (pp. 7042–7051).

Xue, T., Chen, B., Wu, J., Wei, D., & Freeman, W. T. (2019). Video enhancement with task-oriented flow. *International Journal of Computer Vision*, *127*(8), 1106–1125.

Yang, R., Xu, M., & Wang, Z. (2017). Decoder-side HEVC quality enhancement with scalable convolutional neural network. In *IEEE international conference on multimedia and expo, ICME* (pp. 817–822).

Yang, R., Xu, M., Wang, Z., & Li, T. (2018). Multi-frame quality enhancement for compressed video. In *IEEE conference on computer vision and pattern recognition* (pp. 6664–6673).

Yang, R., Sun, X., Xu, M., & Zeng, W. (2019a). Quality-gated convolutional lstm for enhancing compressed video. In *IEEE international*

conference on multimedia and expo, ICME 2019, Shanghai, China, July 8–12, 2019 (pp. 532–537).

Yang, R., Xu, M., Liu, T., Wang, Z., & Guan, Z. (2019b). Enhancing quality for HEVC compressed videos. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(7), 2039–2054.

Yi, P., Wang, Z., Jiang, K., Jiang, J., & Ma, J. (2019). Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *2019 IEEE/CVF international conference on computer vision, ICCV 2019*, Seoul, Korea (South), October 27–November 2, 2019 (pp. 3106–3115).

Yue, H., Cao, C., Liao, L., Chu, R., & Yang, J. (2020). Supervised raw video denoising with a benchmark dataset on dynamic scenes. In *2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020*, Seattle, WA, USA, June 13–19, 2020 (pp. 2298–2307).

Zhang, H., Li, Y., Chen, H., & Shen, C. (2020). Memory-efficient hierarchical neural architecture search for image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3657–3666).

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017a). Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, *26*(7), 3142–3155.

Zhang, K., Zuo, W., Gu, S., & Zhang, L. (2017b). Learning deep CNN denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3929–3938).

Zhang, K., Zuo, W., & Zhang, L. (2018a). Ffdnet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, *27*(9), 4608–4622.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y. (2018b). Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2472–2481).

Zoran, D., & Weiss, Y. (2011). From learning models of natural image patches to whole image restoration. In *IEEE international conference on computer vision, ICCV 2011*, Barcelona, Spain, November 6–13, 2011 (pp. 479–486).