IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Concurrent Video Denoising and Deblurring for Dynamic Scenes

EFKLIDIS KATSAROS[1] (Student Member, IEEE), PIOTR K. OSTROWSKI[2] (Student Member, IEEE), DANIEL WĘSIERSKI[3,4] (Member, IEEE), and ANNA JEZIERSKA[1,4], (Member, IEEE).

[1]Department of Biomedical Engineering, Faculty of Electronics, Telecommunications, and Informatics, Gdańsk University of Technology, Poland
[2]Department of Robotics and Decision Systems, Faculty of Electronics, Telecommunications, and Informatics, Gdańsk University of Technology, Poland
[3]Multimedia Systems Department, Faculty of Electronics, Telecommunications, and Informatics, Gdańsk University of Technology, Poland
[4]Department of Modelling and Optimization of Dynamical Systems, Systems Research Institute, Warsaw, Poland

Corresponding author: Efklidis Katsaros (e-mail: efkleidis.katsaros@pg.edu.pl).

This work was supported in part by The National Centre for Research and Development, Poland, under grant agreement POIR.01.01.01-00-0076/19.

ABSTRACT Dynamic scene video deblurring is a challenging task due to the spatially variant blur inflicted by independently moving objects and camera shakes. Recent deep learning works bypass the ill-posedness of explicitly deriving the blur kernel by learning pixel-to-pixel mappings, which is commonly enhanced by larger region awareness. This is a difficult yet simplified scenario because noise is neglected when it is omnipresent in a wide spectrum of video processing applications. Despite its relevance, the problem of concurrent noise and dynamic blur has not yet been addressed in the deep learning literature. To this end, we analyze existing state-of-the-art deblurring methods and encounter their limitations in handling non-uniform blur under strong noise conditions. Thereafter, we propose a first-to-date work that addresses blur- and noise-free frame recovery by casting the restoration problem into a multi-task learning framework. Our contribution is threefold: **a)** We propose R2-D4, a multi-scale encoder architecture attached to two cascaded decoders performing the restoration task in two steps. **b)** We design multi-scale residual dense modules, bolstered by our modulated efficient channel attention, to enhance the encoder representations via augmenting deformable convolutions to capture longer-range and object-specific context that assists blur kernel estimation under strong noise. **c)** We perform extensive experiments and evaluate state-of-the-art approaches on a publicly available dataset under different noise levels. The proposed method performs favorably under all noise levels while retaining a reasonably low computational and memory footprint.

INDEX TERMS deblurring, denoising, multi-task learning, video enhancement

I. INTRODUCTION

VIDEOS aim at faithfully reflecting the motion in dynamic scenes but concurrent motion blur and noise can severely obscure scene perception. Vision sensors are reaching new and complex environments ranging from medicine, marine and robotics to night vision. However, hardware typically bears application-specific limitations and poses challenges for video enhancement. Improving visual outputs finds applications in visualization environments where the user can assess the scene more accurately and react. Moreover, enhanced video processing facilitates downstream computer vision tasks and improves performance in general video understanding. Although algorithms should address hardware limitations and account for adversarial physical phenomena by enhancing the video output, satisfying the objectives of a real-world application is a demanding task in practice.

Proper calibration of the sensor requires adjustment of the exposure time. While a longer time of exposure increases the number of photons and thus allows the sensor to capture scenes with less noise, it increases the risk of motion blur when the camera shakes and objects move. However, a small exposure time causes noise. Numerous methods have been proposed to address the deblurring task, ranging from spatially invariant [1]–[6] to spatially variant blur [7]–[12]. Meanwhile, many approaches have been proposed for denoising with remarkable results [13]–[16]. However, deeply learnt, dynamic scene video denoising and deblurring have

been addressed only as independent tasks. Severe noise has been recently addressed in video enhancement, but only for static scenes, that is without motion blur, for example in the context of low-light imaging [17]. The problem of spatially variant motion blur, related to independently moving objects in the presence of noise, has not yet been addressed in the deep learning literature. Not only is it an intrinsically challenging problem, but relevant research is also limited by the difficulty in constructing such labeled datasets [17], [18]. For instance, [12] used a beam splitter to construct real blurry-sharp frames, whereas [18] emulated motion by manually moving objects and used a fixed tripod to capture multiple frames of the same scene before generating real noisy-clean pairs of frames by averaging the noisy instantiations of the same scene. In this study, we rely on the realistic blurry dataset of [12] and the realistic Poisson-Gaussian noise model [18], [19].

The problem at hand raises questions. *Should different models be tailored to individually address denoising and deblurring tasks? How robust are deep video deblurring methods with increasing noise levels?* To answer our questions, we first developed a deep learning system for video deblurring under strong noise. We demonstrate that the sequential utilization of off-the-shelf state-of-the-art video denoising and deblurring algorithms is ineffective. The former oversmooths the output since it is not constrained to retain the blur kernel. Moreover, such a configuration would be suboptimal because individual methods require individual feature extraction modules, while motion estimation and local frame features between the two tasks are essentially shareable.

Contributions: To address the aforementioned limitations, we propose R2-D4, the first-to-date deeply learned network that leverages the feature-sharing potential of multi-task learning (MTL) to increase model efficiency and jointly address dynamic video denoising and deblurring. Our main contributions are summarized as follows: **R2-D4:** We propose R2-D4, a novel, MTL-inspired, cascaded convolutional architecture utilizing two decoders to denoise and deblur input frames in stages. R2-D4 employs a tailored feature alignment module that leverages deformable convolutions at the feature level. **MS-RDM:** We propose multiscale residual dense modules to learn coarse-to-fine, dense representations, enhanced by MECA, a novel extension of the efficient channel attention module [20] to further modulate deformable convolutions and increase restoration performance while retaining the number of FLOPs. **Experiments:** We extensively benchmark existing deblurring approaches under different levels of noise on a real, publicly available dataset and show that state-of-the-art deblurring networks bear noise-removing capacity, yet R2-D4 performs consistently better.

II. RELATED WORK

A. DEBLURRING IN THE PRESENCE OF NOISE

Image deblurring in the presence of noise is a fundamental, widely studied subject. Traditionally, a convolution model has been employed for spatially invariant blur and additive Gaussian noise [21] or more realistic Poisson-Gaussian noise [19]. To solve the corresponding ill-posed inverse problem, some studies have resorted to variational methods that incorporate prior information on the unknown clean image, such as promoting sparsity [22], or some prior learned from data [1]–[4]. The other strategy combines the advantages of deep neural networks and variational approaches by linking each layer of a deep network to one iteration of the baseline iterative algorithm, and learning the algorithm hyperparameters from data by using deep unfolding methods [5], [6], [23]. Although the aforementioned methods produce very good results, they are typically limited to simplistic scenarios, which are spatially invariant blur kernels. Moreover they hardly scale to videos because of their relatively high computational complexity. More realistic scenarios with both noise and unknown spatially variant blur have not been extensively addressed in the literature. Existing optimization-based methods require knowledge of the noise level and rely on relatively simple priors assuming a piece-wise constant change of the blur kernel in space [24], [25] or deal with a simplified blur model [26], [27]. The more general space-variant blur model is well studied in the context of deep learning.

B. DEEP VIDEO DEBLURRING

Deep video deblurring methods rely on the expressiveness of stacked convolution filters to deal with dynamic scenes, where blur is due to both camera shakes and independently moving objects. Su et al. [28] proposed an encoder-decoder architecture to align the input frames via its intrinsic multiscale property and showed that warping the input frames with optical flow introduces negligible performance gains but significant warping artifacts. Similarly, Zhou et al. [10] performed implicit frame alignment on the feature level by learning alignment kernels to overcome inaccurate optical flow estimation via a recurrent design. Wang et al. [29] proposed EDVR, a general-purpose reconstruction network for video restoration tasks, including deblurring, denoising, and super-resolution. The authors performed feature-level alignment with a multi-scale cascaded module, comprising deformable convolutions, before spatio-temporal attentive fusion followed a reconstruction module to restore the corrupted input. Zhong et al. [12] introduced a recurrent network that extracts the features frame-wise and pre-processes them with a spatio-temporal attention module that emphasizes the important ones to be passed to the reconstruction decoder that generates the output image. Pan et al. [30] proposed a cascaded algorithm that relies on optical flow estimation to restore the latent frame. However, despite the number of successful studies on dynamic video deblurring, no study has yet addressed this task in the presence of noise.

C. CHANNEL ATTENTION

Channel attention mechanisms have become a prevalent building block in vision since they enable enhanced channel-wise feature learning by highlighting informative features

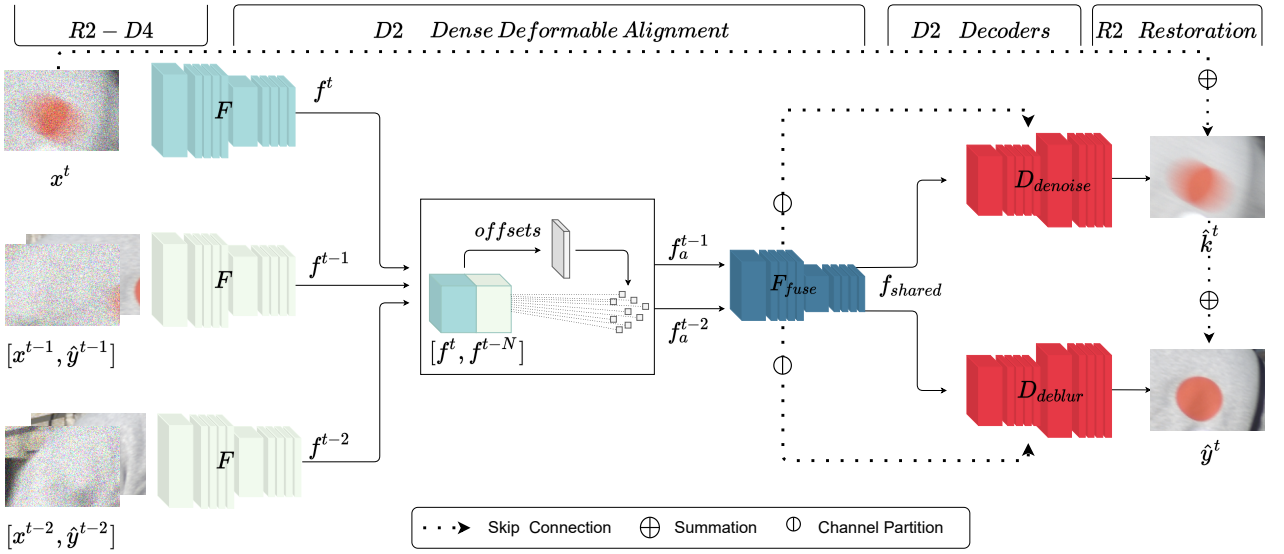


FIGURE 1. The proposed R2-D4 architecture restores the reference frame (R2) via cascaded denoising and deblurring (D2) after aligning its features with the neighboring ones via the dense deformable (D2) alignment module.

and suppressing irrelevant ones at low cost. Hu et al. [31] performed channel-wise global average pooling and employed linear projections with fully connected layers to first reduce and then redeem the channel dimensionality. The features were rescaled using learned weights. Similarly, Woo et al. [32] augmented linear projections with both the global average and max pooling. More recently, Wang et al. [20] argued that dimensionality reduction impairs channel relationships and performs projection directly on the input channel dimension with efficient 1D convolutions. Motivated by the efficiency of channel attention modules, we incorporate them into our work by extending the ECA to further enrich the attentive representations while retaining its efficacy.

D. MULTI-TASK LEARNING

Multi-task Learning constitutes the paradigm where different tasks are learnt simultaneously [33], typically through hard-parameter sharing. MTL is an especially desirable setup under a synergic task configuration. With minor gradient conflicts, it increases the model efficiency by restraining the computational budget and leverages the underlying data structure more effectively by utilizing joint signals from different labels [34]. With regard to the problem at hand, both denoising and deblurring benefit from end-to-end, accurate alignment. Inspired by the success of cascaded restoration in stages [29], [30], [35], we cast the restoration problem on an MTL framework that shares features between the cascaded decoders.

E. DATASETS

Many video deblurring datasets [8], [28], [36] have been introduced to facilitate research in the field. Earlier works [8], [28] utilized high-fps cameras to approximate spatially variant blur via frame averaging over a temporal window. Despite their wide adoption, their use comes at the expense of consistent artifact generation incurred by excessive frame averaging to increase blur. As a result, the frames from [8] exhibit ghosting artifacts. In contrast, [28] used a smaller temporal window at the expense of adequate blur generation. Most recently, the beam splitter dataset [12] (BSD) has been constructed using cameras with different exposure times that record the same scene through a beam splitter. The authors introduced three different exposure configurations, yielding datasets of three different blur levels. However, the datasets captured the outdoor scenes. Obtaining sharp and blurry pairs of frames for low-light scenes remains unaddressed. For instance, the recently published ARID [37] is a low-light dataset that motivates the proposed problem, exhibiting both noise and blur, but lacks the respective paired clear frames.

III. PROBLEM FORMULATION

Let $x \in \mathbb{R}^Q$ be a vector of observations related to an original signal $y \in [0, +\infty)^N$ through the model

$$x = \alpha z(y/\alpha) + w \tag{1}$$

where $\alpha \in (0, +\infty)$ is a scaling parameter, $z(y) = \big(z_i(y)\big)_{1\le i\le Q}$ and $w = (w_i)_{1\le i\le Q}$ are the realizations of mutually independent random vectors $Z(y) = \big(Z_i(y)\big)_{1\le i\le Q}$ and $W = (W_i)_{1\le i\le Q}$ with independent components. It is further assumed that, for every $i \in \{1, \dots, Q\}$, $Z_i(y) \sim \mathcal{P}([Hy]_i)$ and $W_i \sim \mathcal{N}(0, \sigma^2)$, where $\mathcal{P}, \mathcal{N}$ denote the Poisson and Gaussian distributions respectively, $\sigma \in (0, +\infty)$ is the standard deviation of the Gaussian noise component, and $H \in [0, +\infty)^{Q\times N}$ is a matrix modeling the degradation process, i.e. a heterogeneous motion blur kernel map with different blur kernels for each pixel in y. Let h^i represent the kernel from H that operates on a region of the

image centered at location i such that

$$k_i = [Hy]_i \tag{2}$$

Thus, for each i, we have $\mathcal{P}(k_i) = \mathcal{P}\left(\sum_j h_j^i y_{i+j}\right)$. In the context of deep learning, the original video signal can be recovered by some network $\mathcal{F}$ with parameters Θ. Hence, given T consecutive, corrupted frames $\left(x^t\right)_{1\leq t\leq T}$, the optimal set of Θ is derived by minimizing the criterion:

$$\mathcal{L}(\Theta) = L\left(\mathcal{F}_\Theta(x_i^{t-N}, \ldots, x_i^t), y_i^t, k_i^t\right) \tag{3}$$

where L denotes some quality measure function e.g. ℓ_2 squared norm or ℓ_1 norm. More recently, perceptually motivated strategies [8], [10], [38] have been considered to restore realistic image structures by augmenting the optimization criterion via either GAN-based [39] adversarial training [8], [38] or perceptual loss terms [40]. Here, $\mathcal{F}_\Theta(x_i^{t-N}, \ldots, x_i^t)$ yields two outputs of $\hat{k}^t$ and $\hat{y}^t$.

IV. PROPOSED METHOD

Given N consecutive corrupted frames $x^{[t-N:t]}$ and $N-1$ previously restored frames $\hat{y}^{[t-N:t-1]}$, our method obtains $\hat{y}^t$ via a cascaded, two-stage restoration. The proposed R2-D4 network consists of a shared, dense, deformable (D2) feature alignment module, followed by a convolutional feature fusion and two decoders performing denoising and deblurring sequentially (D2) to restore the frames via a two-stage (R2) cascaded process, as illustrated in Fig. 1. The shared D2 module processes the current x^t and previous $\{x^{t-N}, \hat{y}^{t-N}\}$ frames to extract features at each time step. Subsequently, the asymmetric offsets are estimated to align the neighboring frame features with the reference frame features. Thereafter, the aligned features are fused before the two decoders leverage the shared features to denoise and deblur the current frame sequentially in a cascaded manner.

The D2 alignment module, described in Sec. IV-B1, employs modulated deformable convolutions [41] to align frames at the feature level and does not estimate the optical flow that is harder under strong noise. Common issues arising from optical flow include computational inefficiency and generation of motion artifacts. Feature alignment is further improved via our multiscale residual dense modules (MS-RDMs), described in Sec. IV-A2, which leverage dilated convolutions to capture a longer-range context. MS-RDBs essentially serve as a pre-processing step before deformable convolutions, aggregating features with increased effective receptive fields, thus facilitating the known deformable offset estimation issue [29], [42]. MS-RDBs are further enhanced with our modulated efficient channel attention blocks, as explained in Sec. IV-A1.

The R2 two-stage cascaded restoration process utilizes decoders to denoise and deblur corrupted frames sequentially under an efficient MTL framework. Accurate feature alignment benefits both denoising and deblurring. Moreover, the features are expanded channel-wise upon fusion, increasing the model capacity at the lowest resolution to accommodate both tasks sufficiently. Finally, the two-stage cascaded process has been shown to yield increased performance on many restoration tasks and is therefore integrated into R2-D4 through the two decoders under the proposed feature-sharing scheme. As illustrated in Fig. 1, additional residual connections from x_t to the first-stage output $\hat{k}_t$, and from the latter to the second-stage $\hat{y}_t$ are used to facilitate the training.

A. PROPOSED BLOCKS

1) Modulated Efficient Channel Attention

Self-supervised channel attention blocks have become ubiquitous since they highlight informative and suppress non-relevant features. Wang et al. [20] proposed an efficient 1D convolution (ECA) on globally averaged input channels to determine the attentive weights, as illustrated in the top half of Fig. 3. Formally, ECA is denoted as follows:

$$ECA_k = \sigma \circ \mathcal{C}_{c,1\times k} \circ GAP \tag{4}$$

where GAP is the global average pooling operation, σ is the sigmoid function, and $\mathcal{C}_{c,1\times k}$ is a 1D convolutional operation with c output channels and a kernel of size k, where the latter is typically determined adaptively as a function of the input channels. Formally, assuming a feature cube $f_c \in \mathbb{R}^{H\times W\times C}$, the channel-wise attention weights are then derived as follows:

$$\tilde{f}_c = ECA_k(f_c) \tag{5}$$

Then, $\tilde{f}_c$ is multiplied by the input features f to obtain the attended $\tilde{f}$.

Despite the success of channel attention modules, they are often difficult to optimize and converge to uniform distributions of the channel weights. To alleviate such issues and facilitate the gradient flow during the backward pass, we propose to complement globally averaged features with max-pooled features as in CBAM [32], under the efficient 1D convolution configuration of [20]. The modulated efficient channel attention module, termed MECA, is illustrated in the bottom half of Fig. 3. In contrast to ECA, we perform both global average and max pooling (MP) on the features f channel-wise to obtain f'_c, and we denote the concatenation of GAP and MP channels as MGAP. By adopting the notation in Eq. 4, MECA is defined as:

$$MECA_k = \sigma \circ \mathcal{C}_{c,2\times k} \circ MGAP \tag{6}$$

The attended weights are derived similarly to Eq. 5 and multiplied by the input features to obtain the attented features. Notably, MECA retains the efficiency of 1D convolution in capturing the local cross-channel interactions but learns an essentially more effective projection, utilizing two channels of informative cues instead of solely the globally averaged ones. MECA is an easy-to-plug module that can be integrated into all standard architectures for any vision task.

2) Multiscale Residual Dense Module

Residual blocks [43] have been a popular choice [10], [16], [38], [44] in image and video restoration. More recently,

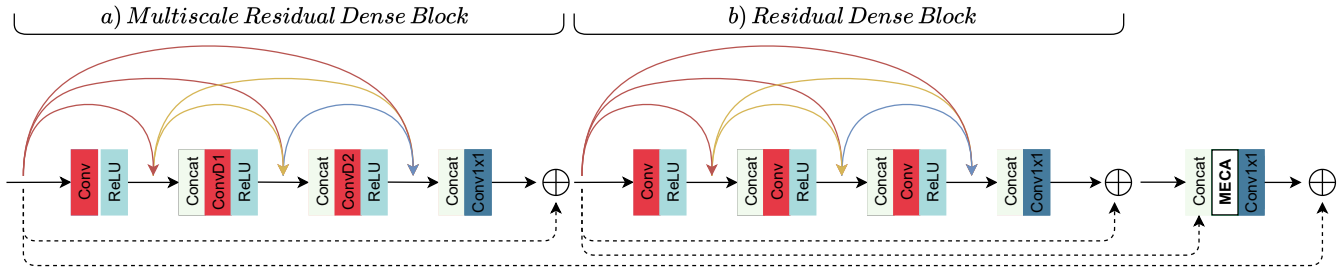


FIGURE 2. The proposed Multiscale Residual Dense Module learns enhanced hierarchical representations via its coarse-to-fine design. The MS-RDB (a) block mines coarser features with increasing dilation rates whereas the second RDB (b) block learns finer details.

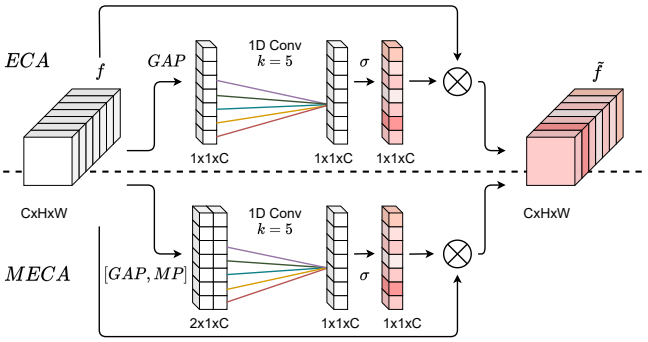


FIGURE 3. Proposed modulated efficient channel attention.

residual dense blocks (RDBs) [45] exploited dense connections between layers to extract richer hierarchical features while instantiating a contiguous memory (CM) mechanism to further enhance the learned representations.

Residual dense blocks (RDBs) typically consist of l convolutional kernels and a 'growth factor' hyperparameter g. As shown in Fig. 2b, each layer receives the feature maps from the previous stage, convolves them with a 3×3 kernel that yields g additional channels and concatenates them with the previous ones before passing them to the next layer. Each block is then followed by a 1×1 convolution to aggregate the signal and stabilize the training before the residual summation. Formally, a single RDB with 3 layers can be denoted as follows:

$$\begin{aligned}\mathcal{RDB} = \mathcal{C}_{c,1\times1} \circ \mathcal{CAT}_{c+3g} \circ \mathcal{C}_{g,3\times3} \\ \circ \mathcal{CAT}_{c+2g} \circ \mathcal{C}_{g,3\times3} \circ \mathcal{CAT}_{c+g} \circ \mathcal{C}_{g,3\times3}\end{aligned} \tag{7}$$

where $\mathcal{C}_{c,k\times k}$, R and $\mathcal{CAT}_c$ are the $k \times k$ convolution operation, the activation function and the concatenation function respectively. The subscript c denotes the number of output channels after each convolution and concatenation. Stacking b such residual dense blocks gives rise to RDB cells [12], where the output of each block is sequentially processed by the next block. For clarity, we term them residual dense modules (RDMs). In RDMs, all subsequent RDB outputs are concatenated and fed into another 1×1 convolution before the residual summation at the module level.

In this work, we introduce multiscale residual dense modules (MS-RDMs) to efficiently increase the effective receptive field by spatially augmenting the hierarchical features in a coarse-to-fine manner. As illustrated in Fig. 2, MS-RDMs are designed via an MS-RDB that captures a hierarchically coarser context via kernel dilation followed by a simple non-dilated RDB to complement hierarchical features with fine details. Regarding the MS-RDB, layers are progressively enhanced with larger dilation rates to hierarchically capture a longer-range context. As depicted in Fig. 2a, the MS-RDB block is defined as:

$$\begin{aligned}\mathcal{RDB}_{\mathcal{MS}} = \mathcal{C}_{c,1\times1,0} \circ \mathcal{CAT}_{c+3g} \circ \mathcal{C}_{g,3\times3,2} \\ \circ \mathcal{CAT}_{c+2g} \circ \mathcal{C}_{g,3\times3,1} \circ \mathcal{CAT}_{c+g} \circ \mathcal{C}_{g,3\times3,0}\end{aligned} \tag{8}$$

where $\mathcal{C}_{c,k\times k,d}$ denotes, again, the convolution, but dilated with a rate of d. Upon concatenation of the coarse and fine block features and before the 1×1 convolutional aggregation, we perform channel-wise attention via the proposed $MECA_k$. Similarly, the resultant RDM_{MS} is defined as:

$$\begin{aligned}\mathcal{RDM}_{\mathcal{MS}} = \mathcal{C}_{c,1\times1} \circ MECA_7 \\ \circ \mathcal{CAT}_{2c} \circ \mathcal{RDB} \circ \mathcal{RDB}_{\mathcal{MS}}\end{aligned} \tag{9}$$

The proposed MS-RDM reformulation enlarges the effective receptive field, which in turn renders the CM mechanism spatially more aware. The coarse-to-fine hierarchical features mine spatially aware representations and serve as a preprocessing step for deformable offset estimation.

B. RESTORATION EN CASCADE

1) Dense Deformable Alignment

At each time step t, the network receives the current frame x_t and previous corrupted and restored $\{x^{t-N}, \hat{y}^{t-N}\}$ ones. Leveraging previously restored frames encourages temporal coherence by reducing flickering and has been shown to yield improved performance [10]. At each time step, the respective features are computed using the following block:

$$F = \mathcal{RDM}_{\mathcal{MS}32} \circ \mathcal{C}_{32,3\times3,2} \circ \mathcal{RDM}_{\mathcal{MS}16} \circ \mathcal{C}_{16,3\times3,1} \tag{10}$$

where $\mathcal{C}_{c,k,s}$ denotes a $k \times k$ convolution with a stride of s, and c output channels and $\mathcal{RDM}_{\mathcal{MS}g}$ is the multiscale residual dense block with a growth factor g. As illustrated

in Fig. 1, R2-D4 contains two sets of weights: one for the current x^t and one for each past $\{x^{t-N}, \hat{y}^{t-N}\}$. Likewise,

$$f^t = F(x^t), \qquad f^{t-N} = F(x^{t-N}, \hat{y}^{t-N}) \tag{11}$$

where N is set to 2. Weight sharing for past frame features increases the training efficiency and accelerates inference by reusing f^{t-2} at each time step.

The current and previous frames are then aligned using deformable convolutional layers. A deformable module enables the modeling of geometric transformations through asymmetric kernels so that output features can capture object-specific contexts that assist blur kernel estimation. The leveraging of deformable convolutions under the proposed scheme has three advantages. First, it discards the necessity for erroneous and computationally expensive optical flow estimations. Second, it performs alignment on the deeper feature levels instead of the image level. This has been shown to improve performance [10], [29] because the layers prior to the deformable modules encode features that are tailored to the alignment. Third, estimating deformation offsets on the coarse-to-fine features extracted from MS-RDMs assists in modulated offset estimation and improves performance.

Each modulated deformable layer consists of two convolutions. The first layer learns the offset displacements and the modulating scalars that determine the amplitude of the output features. The second layer employs the modulated offsets and learns the filter weights, as in ordinary convolution. The deformable convolution is denoted as

$$\mathcal{DC} = \mathcal{C}^D_{128,3\times3,1} \circ \mathcal{C}_{27,3\times3,1} \tag{12}$$

where $\mathcal{C}_{3k^2,k\times k,s}$ is the $k \times k$ convolutional kernel with a stride of s, estimating the $2k^2$ offsets and respective k^2 modulation scalars from the concatenated c frame features and $\mathcal{C}^D_{c,k\times k,s}$ denotes the actual deformable convolution with c output channels. Correspondingly, the aligned features are defined as follows:

$$f_a^{t-N} = \mathcal{DC}(f^t, f^{t-N}) \tag{13}$$

The fusion of the aligned features is then performed via:

$$\begin{aligned} F_{fuse} = \mathcal{RDB}_{32} \circ \mathcal{RDB}_{32} \circ \mathcal{C}_{128,3\times3,2} \\ \circ \mathcal{RDB}_{32} \circ \mathcal{RDB}_{32} \circ \mathcal{C}_{128,3\times3,1} \end{aligned} \tag{14}$$

Note that simple RDBs without dilation rates are employed for fusion because a spatially wider context does not strengthen the feature representations at smaller scales. Because the number of past frames is $N = 2$, the output and shared features are defined as

$$f_{shared} = F_{fuse}(f_a^{t-1}, f_a^{t-2}) \tag{15}$$

2) Cascaded Decoders

The decoders share identical architectures. They are optimized to upsample the shared features and yield denoised and deblurred outputs (D2) sequentially. As shown in [46], transposed convolutions often generate checkerboard artifacts. To overcome these problems during feature upsampling, many studies have resorted to bilinear upsampling followed by convolution [35], [47]. Although we confirm that bilinear upsampling eliminates artifacts, it leads to a loss of spatial information. Therefore, we resort to convolutional channel-wise expansion followed by pixel shuffling [48] to reduce gridding artifacts and preserve spatial details. Denoting the upsampling layers as PS, each decoder can be expressed as follows:

$$\begin{aligned} \mathcal{D} = \mathcal{C}_{3,3\times3,1} \circ \mathcal{RDB}_{16} \circ \mathcal{PS}_{32} \circ \mathcal{C}_{128,3\times3,1} \\ \circ \mathcal{RDB}_{32} \circ \mathcal{PS}_{64} \circ \mathcal{C}_{256,3\times3,1} \end{aligned} \tag{16}$$

Assuming two such instantiations for denoising and deblurring as $\mathcal{D}_{den}$ and $\mathcal{D}_{deb}$, the intermediate denoised and restored output frames are defined as:

$$\hat{k}^t = \mathcal{D}_{den}(f_{shared}) \tag{17}$$

$$\hat{y}^t = \mathcal{D}_{deb}(f_{shared}) + \mathcal{D}_{den}(f_{shared}) \tag{18}$$

We utilize skip connections from the encoder to the decoders to preserve spatial information and facilitate training, as is common in UNet-based [49] methods. Instead of concatenating the encoding channels with both decoders, we restructure the gradient flow by dissecting the former, say $f \in \mathbb{R}^{H\times W\times C}$, in two groups f_{den}, $f_{deb} \in \mathbb{R}^{H\times W\times C/2}$, each specialized for the decoder's task, as illustrated in Fig. 1. Likewise, f_{den} and f_{deb} receive task-specific gradients in addition to the shared gradients. As a result, f_{den} focuses on the global noise distribution, whereas f_{deb} is specialized in recovering the blur-free frame.

(a) Bright (b) Dark (c) Low-light

FIGURE 4. Examples from the BSD dataset under different illumination configurations. The bright, dark, and low-light data correspond to α values of 0.5, 1.9 and 7.1 and are used to generate low, moderate and severe noise, respectively.

V. EXPERIMENTS

In this section, we present the experiments that (i) compare the performance of R2-D4 with state-of-the art video deblurring methods and investigate their robustness at different

Method	N	Par (M)	GFLOPs	Low	Moderate	Severe
STFAN	2	5.4	188.9*	29.23 0.875	29.06 0.868	28.57 0.858
ESTRNN	3	2.3	142.9	30.52 0.905	29.90 0.892	29.07 0.872
CDVD (1)	5	16.2	-	30.40 0.906	29.92 0.894	29.06 0.875
CDVD (2)	5	16.2	-	30.53 **0.911**	30.12 **0.900**	29.17 **0.880**
R2-D3	3	4.4	216.9	30.82 0.905	30.19 0.886	29.17 0.870
R2-D4$^-$	3	5.1	270.7	30.93 0.907	30.22 0.890	29.18 0.870
R2-D4	3	5.1	270.7	**31.10** 0.910	**30.32** 0.894	**29.33** / 0.876

TABLE 1. PSNR (top) and SSIM (bottom) results at three noise levels. GFLOPs* for STFAN did not include their FAC layers. The bold and underlined results indicate the first and second rank, respectively.

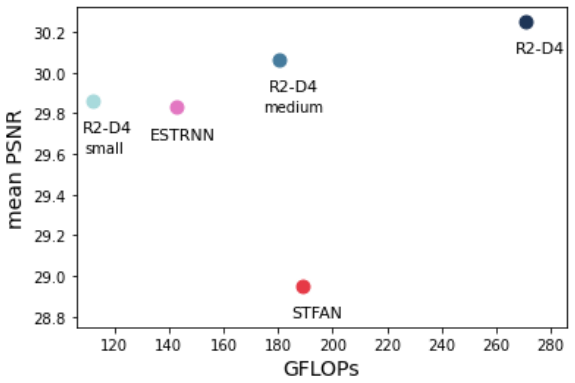


FIGURE 5. Mean PSNR *vs.* GFLOPs for three R2-D4 variants compared to ESTRNN and STFAN.

noise levels, (ii) assess the impact of the proposed blocks on R2-D4 architecture and (iii) assess the impact of the MTL configuration. All experiments use the "3ms24ms" version of BSD that has the strongest level of blur. The evaluation protocol contains 60 training (30K pairs), 20 validation (10K pairs) and 20 test (15K pairs) sequences with a resolution of 640×480. Poisson-Gaussian noise was generated using Eq. (1) on the blurry frames, with the noise parameters $\{\alpha, \sigma\}$ equal to $\{0.5, 0.9\}$, $\{1.9, 1.7\}$ and $\{7.1, 3.3\}$ for low, moderate, and severe noise, respectively. Note that the choice of α parameters in Eq. (1) simulates different illumination conditions ranging from brighter to low-light images to increase the values of α (see Fig. 4). In the model considered in Eq. (1), the corrupted data $z(y/\alpha)$ are further normalized back to the common range by multiplying with α. The generated shot noise distribution is typical for bright, dark, and low-light images for α equal to 0.5, 1.9, and 7.1, respectively.

A. LOSS FUNCTIONS

The R2-D4 parameters are derived by optimizing Eq.(3), where L is a weighted sum of ℓ_2 squared norms, i.e.

$$L = L_{blur} + \lambda_1 L_{noise} + \lambda_2 L_{perceptual}, \tag{19}$$

where

$$L_{blur} = \frac{1}{CHW}||y^t - \hat{y}^t||^2, \tag{20}$$

$$L_{noise} = \frac{1}{CHW}||k^t - \hat{k}^t||^2, \tag{21}$$

and

$$L_{perceptual} = \frac{1}{C_\phi H_\phi W_\phi}||\phi_{VGG}(y^t) - \phi_{VGG}(\hat{y}^t)||^2. \tag{22}$$

The definition of $L_{perceptual}$ is adopted from [40], where ϕ_{VGG} denotes the VGG-19 features [50] extracted from the 3th layer and C_ϕ, H_ϕ, W_ϕ denote the corresponding feature dimensions. The scalar values C, H, and W refer to the image channel, height, and width, respectively, and the weights are set to $\lambda_1 = 0.6$ and $\lambda_2 = 0.01$.

B. METHODS

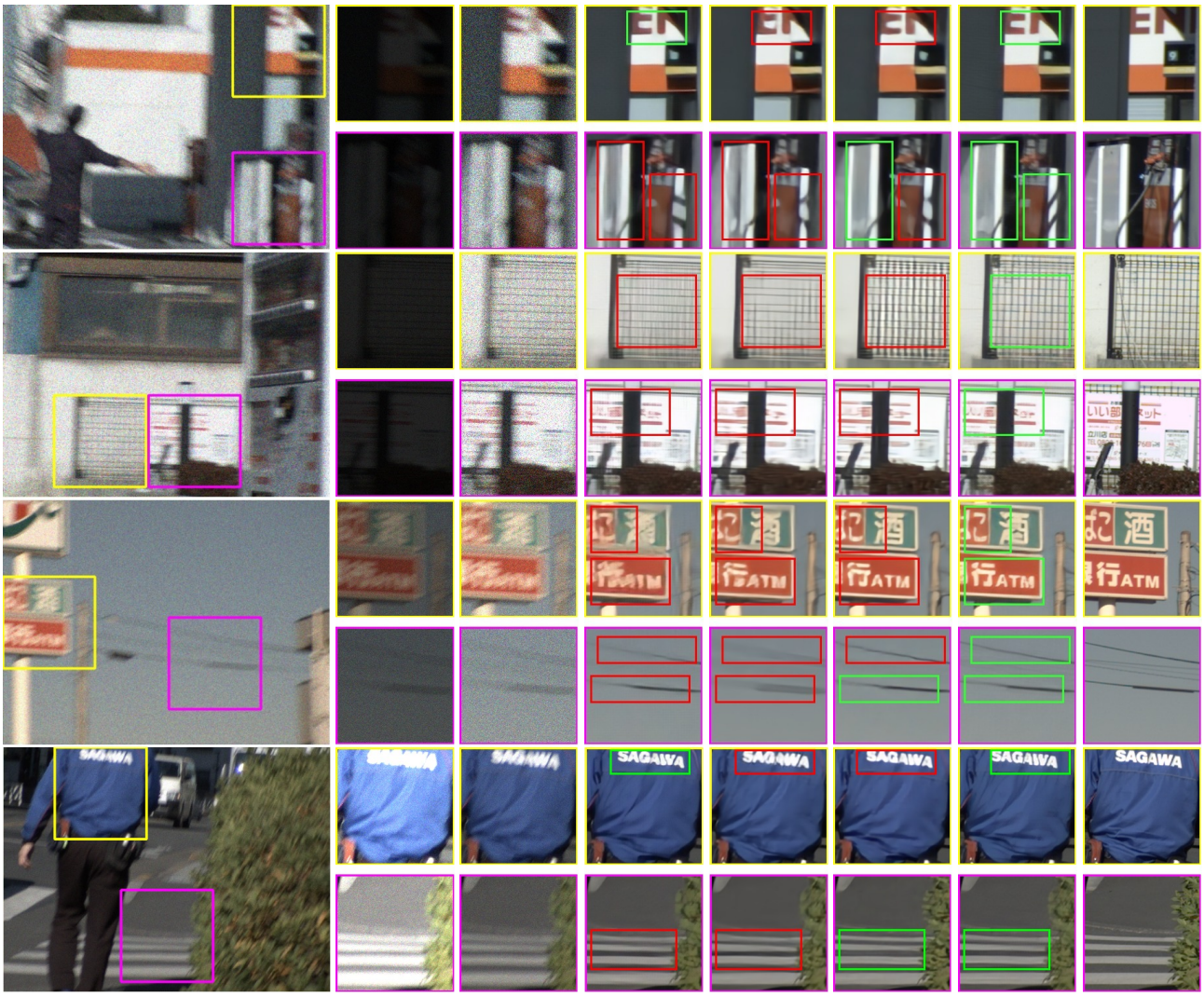
First, we examine the performance of a naive system in which two methods operate sequentially. We trained Fast-DVDNet [13] for denoising, followed by STFAN [10] for deblurring. Second, we compare R2-D4 with state-of-the-art models: STFAN [10], ESTRNN [12] with 15 blocks and only past frames, and CDVD-TSP [30]. In our ablation study, we investigated the effectiveness of our MTL setup by comparing it to R2-D3, which uses only a single decoder. Subsequently, we verify the impact of the proposed blocks on our feature alignment module by R2-D4$^-$, defined as: (i) MECA substituted with ECA; (ii) MS-RDB modules substituted with simple RDB modules, thus retaining GFLOPs. Next, we reduce the number of channels in the decoders and the fusion module, thereby obtaining the reduced but more computationally efficient "small" and "medium" R2-D4 variants.

C. SETUP

Our experiments are performed with PyTorch on an Nvidia Tesla V100 for 250 epochs. Adam [51] was used as the optimizer with a learning rate of 1.5×10^{-4} decayed to 10^{-6} via the cosine annealing strategy [52]. The networks are trained with sequences of 30 frames and a batch size of 1. The frames are randomly augmented with horizontal and vertical flips. Experiments for state-of-the-art methods follow the official, publicly available implementations.

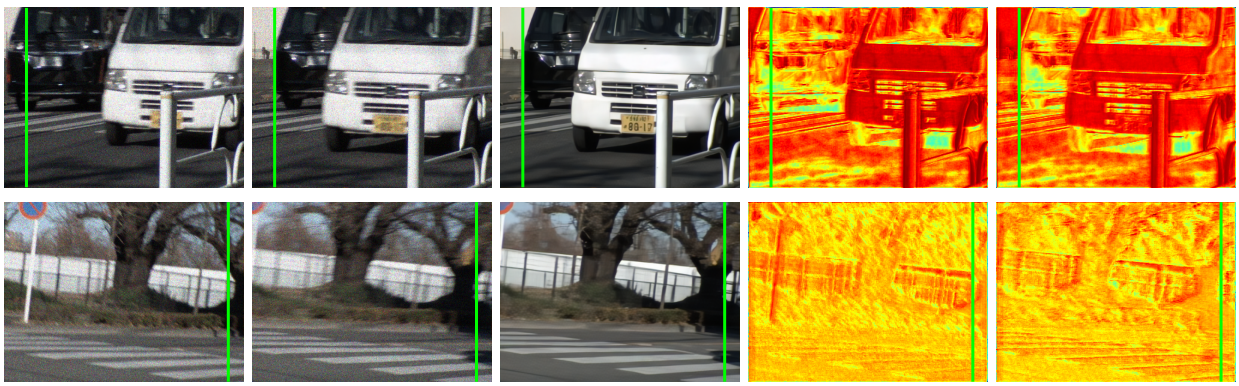
D. RESULTS

The naive approach is not trained end-to-end and thus over-smooths the input frames achieving a PSNR of 28.40 and an SSIM of 0.850 for the severe noise setting. The results of the end-to-end methods are listed in Table 1. Interestingly, our experiments show that deblurring methods bear some noise-removal capacity, although R2-D4 performs better than



(a) Corrupted frame (b) Lighting (c) Patch (d) STFAN (e) ESTRNN (f) CDVD (g) R2-D4 (h) GT

FIGURE 6. Qualitative Results. The frames were normalized to the same range. In zoomed areas, red and green rectangles highlight artifacts and more accurate reconstructions, respectively. The first, second, third and fourth rows were generated with severe, severe, moderate and low noise respectively. Column (b) demonstrates varying illumination conditions under which noise was generated whereas column (c) shows patches that are normalized to the common image scale.



(a) Corrupted frame x^{t-2} (b) Corrupted frame x^t (c) Restored frame $\hat{y}^t$ (d) Offsets at $t-2$ (e) Offsets at t

FIGURE 7. Visualization of deformable offsets. R2-D4 adapts the offsets for independently moving or uniform motion scenarios.

STFAN and ESTRNN in both PSNR and SSIM. Moreover, it performs higher in PSNR and on par in SSIM with the computationally expensive, cascaded version of CDVD-TSP (2), which performs two passes over the corrupted frames and uses five input frames. As shown in Table 1, the performance increased over the compared methods across all levels of noise. The second decoder and the proposed blocks clearly contribute to performance gains, increasing mean PSNR by 0.19 dB and 0.15 dB compared to R2-D3 and R2-D4$^-$, respectively. Last, Fig. 5 shows that while the small R2-D4 variant has 30% fewer GFLOPs in comparison to ESTRNN, it performs better than both STFAN and ESTRNN.

R2-D4 benefits from accurate feature alignment under strong noise and recovers fine-grained frame details (see Fig. 6). One can observe that STFAN often fails to align features producing hallucinations, as seen in the gas tube (top row) and in the fence (middle row). For the same examples, the ESTRNN tends to oversmooth the output. CDVD-TSP performs better but tends to yield piecewise constant artifacts despite its larger complexity, which is visible in the fence example. R2-D4 performs implicit feature alignment and dynamically adapts offsets over time, as illustrated in Fig. 7. The top row illustrates the scenario of independently moving objects, whereas the bottom row depicts the uniform motion caused by camera movement. The offset variance is higher for the former; R2-D4 mines the spatio-temporal boundaries and aggregates the object-specific context. The spatial responses for the second case show a smaller variance as the learned offsets exhibit similar directions. R2-D4 dynamically adapts offsets in the case at hand.

VI. CONCLUSION

In this paper we study dynamic scene video deblurring under strong noise. Although such acquisition settings arise frequently in practice, the problem is challenging and new in the deep learning literature. We demonstrate that state-of-the-art deblurring methods have some denoising capacity, but the proposed R2-D4 method outperforms them owing to an MTL-inspired, cascaded yet efficient architecture, enhanced with MS-RDM modules. Future research aims to bridge the gap between synthetically generated and real datasets with raw video sequences of dynamic scenes with natural noise.

REFERENCES

[1] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3929–3938.

[2] J. Zhang, J. Pan, W.-S. Lai, R. W. Lau, and M.-H. Yang, "Learning fully convolutional networks for iterative non-blind deconvolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3817–3825.

[3] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (RED)," *SIAM Journal on Imaging Sciences*, vol. 10, no. 4, pp. 1804–1844, 2017.

[4] E. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin, "Plug-and-play methods provably converge with properly trained denoisers," in *International Conference on Machine Learning*, 2019, pp. 5546–5557.

[5] C. Bertocchi, E. Chouzenoux, M.-C. Corbineau, J.-C. Pesquet, and M. Prato, "Deep unfolding of a proximal interior point method for image restoration," *Inverse Problems*, vol. 36, no. 3, p. 034005, 2020.

[6] C. Agarwal, S. Khobahi, A. Bose, M. Soltanalian, and D. Schonfeld, "DEEP-URL: A model-aware approach to blind deconvolution based on deep unfolded Richardson-Lucy network," in *IEEE International Conference on Image Processing*, 2020, pp. 3299–3303.

[7] S. Dai and Y. Wu, "Removing partial blur in a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2544–2551.

[8] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3883–3891.

[9] D. Gong, J. Yang, L. Liu, Y. Zhang, I. Reid, C. Shen, A. Van Den Hengel, and Q. Shi, "From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2319–2328.

[10] S. Zhou, J. Zhang, J. Pan, H. Xie, W. Zuo, and J. Ren, "Spatio-temporal filter adaptive network for video deblurring," in *IEEE International Conference on Computer Vision*, 2019, pp. 2482–2491.

[11] J. Wu, X. Yu, D. Liu, M. Chandraker, and Z. Wang, "DAVID: Dual-attentional video deblurring," in *IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2376–2385.

[12] Z. Zhong, Y. Gao, Y. Zheng, and B. Zheng, "Efficient spatio-temporal recurrent neural network for video deblurring," in *European Conference on Computer Vision*, 2020, pp. 191–207.

[13] M. Tassano, J. Delon, and T. Veit, "Fastdvdnet: Towards real-time deep video denoising without flow estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1354–1363.

[14] ——, "Dvdnet: A fast network for deep video denoising," in *IEEE International Conference on Image Processing*, 2019, pp. 1805–1809.

[15] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, 2019.

[16] S. Anwar and N. Barnes, "Real image denoising with feature attention," in *IEEE International Conference on Computer Vision*, 2019, pp. 3155–3164.

[17] C. Chen, Q. Chen, M. N. Do, and V. Koltun, "Seeing motion in the dark," in *IEEE International Conference on Computer Vision*, 2019, pp. 3185–3194.

[18] H. Yue, C. Cao, L. Liao, R. Chu, and J. Yang, "Supervised raw video denoising with a benchmark dataset on dynamic scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2301–2310.

[19] E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot, "A convex approach for image restoration with exact Poisson–Gaussian likelihood," *SIAM Journal on Imaging Sciences*, vol. 8, no. 4, pp. 2662–2682, 2015.

[20] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 531–11 539.

[21] M. Maggioni, G. Boracchi, A. Foi, and K. Egiazarian, "Video denoising, deblocking, and enhancement through separable 4-D nonlocal spatiotemporal transforms," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 3952–3966, 2012.

[22] T. Hyun Kim and K. Mu Lee, "Generalized video deblurring for dynamic scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5426–5434.

[23] D. Gong, Z. Zhang, Q. Shi, A. van den Hengel, C. Shen, and Y. Zhang, "Learning deep gradient descent optimization for image deconvolution," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5468–5482, 2020.

[24] S. Ben Hadj, L. Blanc-Féraud, and G. Aubert, "Space variant blind image restoration," *SIAM Journal on Imaging Sciences*, vol. 7, no. 4, pp. 2196–2225, 2014.

[25] L. Bar, N. Sochen, and N. Kiryati, "Blind space-variant single-image restoration of defocus blur," in *Scale Space and Variational Methods in Computer Vision*, F. Lauze, Y. Dong, and A. B. Dahl, Eds., vol. 10302. Cham: Springer International Publishing, 2017, pp. 109–120.

[26] L. Xu, S. Zheng, and J. Jia, "Unnatural L0 sparse representation for natural image deblurring," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1107–1114.

[27] Y. Zhang and K. Hirakawa, "Blind deblurring and denoising of images corrupted by unidirectional object motion blur and sensor noise," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4129–4144, 2016.

[28] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1279–1288.

[29] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.
[30] J. Pan, H. Bai, and J. Tang, "Cascaded deep video deblurring using temporal sharpness prior," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3043–3051.
[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
[32] S. Woo, J. Park, J.-Y. Lee, and I. Kweon, "CBAM: Convolutional block attention module," in *European Conference on Computer Vision*, 2018, pp. 3–19.
[33] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.
[34] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, and S. Savarese, "Which tasks should be learned together in multi-task learning?" in *International Conference on Machine Learning*, 2020, pp. 9120–9132.
[35] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 821–14 831.
[36] Y. Yuan, W. Su, and D. Ma, "Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3555–3564.
[37] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, "ARID: A new dataset for recognizing action in the dark," in *International Workshop on Deep Learning for Human Activity Recognition*, 2021, pp. 70–84.
[38] K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu, and H. Li, "Adversarial spatio-temporal learning for video deblurring," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 291–301, 2018.
[39] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.
[40] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016, pp. 694–711.
[41] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9308–9316.
[42] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Understanding deformable alignment in video super-resolution," *arXiv preprint arXiv:2009.07265*, vol. 2, no. 3, 2020.
[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[44] Y. Qiu, R. Wang, D. Tao, and J. Cheng, "Embedded block residual network: A recursive restoration model for single-image super-resolution," in *IEEE International Conference on Computer Vision*, 2019, pp. 4180–4189.
[45] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2472–2481.
[46] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016.
[47] K. Purohit and A. Rajagopalan, "Region-adaptive dense network for efficient motion deblurring," in *AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 882–11 889.
[48] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.
[49] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.
[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[52] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

EFKLIDIS KATSAROS received the B.Sc. degree in Mathematics, from the Aristotle University of Thessaloniki, Greece in 2016 and the M.Sc. degree in Data Science: Statistical Science, cum laude, from Leiden University, the Netherlands, in 2019. His main research interests lie in computer vision and machine learning. Since 2020, he has been a researcher at the Gdańsk University of Technology, Poland, where he simultaneously pursues a Ph.D. degree in multi-task learning for computer vision in medical applications at the Department of Biomedical Engineering.

PIOTR K. OSTROWSKI received the B.Eng. degree in control engineering and robotics in 2017, and the M.Sc. degree in Computer Science in 2018 from the Gdańsk University of Technology. He is a Ph.D. candidate at the Gdańsk University of Technology. His research interests include machine learning and computer vision with an emphasis on efficient video processing.

DANIEL WĘSIERSKI is an assistant professor at the Gdańsk University of Technology and at the Systems Research Institute. He received the M.Sc. degree from the Gdańsk University of Technology, Poland, in 2007 for his thesis on stereovision robot gripping systems that he developed at ThyssenKrupp in Bremen, Germany. He then developed vision and robot systems for testing car control units at Volkswagen R&D in Wolfsburg, Germany. He received his Ph.D. degree from Télécom SudParis, France, in 2013 for developing visual tracking algorithms of flexible and articulated objects. He is a recipient of the best paper awards at MICCAI CARE workshop in 2015 and 2017 for vision-based tracking algorithms of surgical instruments. He has led application-oriented national R&D projects and has participated in two European projects. He has extensive experience in computer vision, image processing, and teaching machines on uncertain data with a focus on medical applications, including minimally invasive surgery, dentistry, and electroencephalography. His basic research interests include machine learning under uncertainty, denoising, optimization, and model-based vision.

ANNA JEZIERSKA received the Ph.D. degree from the University Paris l'Est, Paris, France, in 2013. She received the Ph.D. award Prix solennels de la Chancellerie des universites de Paris in 2014. Since 2014, she has been with the Systems Research Institute of the Polish Academy of Sciences at the Department of Mathematical Modelling and Optimization of Dynamical Systems, where she is currently an adjunct Professor. She received an award from the Minister of Science Scholarships for outstanding young scientists in Poland in 2015. In 2015, she joined the Gdańsk University of Technology, where she has been working with the Department of Biomedical Engineering at the faculty of Electronics, Telecommunications and Informatics. She has led national and international research projects on image enhancement, large scale optimization problems and numerical methods for biology and medicine. She has coauthored more than 20 Web of Science-indexed publications in the field of image processing.