# Burst Denoising via Temporally Shifted Wavelet Transforms

Xuejian Rong[†,§], Denis Demandolx[†], Kevin Matzen[†],
Priyam Chatterjee[†], and Yingli Tian[§]

[†]Facebook [§]CUNY
{xrong,denisd,matzen,priyamc}@fb.com
ytian@ccny.cuny.edu

**Abstract.** Mobile photography has made great strides in recent years. However, low light imaging remains a challenge. Long exposures can improve signal-to-noise ratio (SNR) but undesirable motion blur can occur when capturing dynamic scenes. Consequently, imaging pipelines often rely on computational photography to improve SNR by fusing multiple short exposures. Recent deep network-based methods have been shown to generate visually pleasing results by fusing these exposures in a sophisticated manner, but often at a higher computational cost.

We propose an end-to-end trainable burst denoising pipeline which jointly captures high-resolution and high-frequency deep features derived from wavelet transforms. In our model, precious local details are preserved in high-frequency sub-band features to enhance the final perceptual quality, while the low-frequency sub-band features carry structural information for faithful reconstruction and final objective quality. The model is designed to accommodate variable-length burst captures via temporal feature shifting while incurring only marginal computational overhead, and further trained with a realistic noise model for the generalization to real environments. Using these techniques, our method attains state-of-the-art performance on perceptual quality, while being an order of magnitude faster.

**Keywords:** Burst Denoising, Wavelet Transform, Deep Learning

## 1 Introduction

Image and video denoising are fundamental low-level vision tasks that have been studied for decades. Noise reduction is even more critical with the explosive growth of mobile cameras with small apertures and sensors that have limited light capture capabilities, making it difficult to acquire images in low light conditions. An effective denoising model could improve the visual quality of captures under such constraints, and immediately underpins many downstream computer vision and image processing applications.

The traditional approach to dealing with noise under low light situations is to increase the shutter time, allowing the sensor to collect more light to reduce the dominant photon noise. However, this approach requires that the camera be stable during the exposure and cannot handle object motion and corresponding motion blur well. An alternate approach is to capture multiple short-exposure frames (or bursts) to approximately capture an equivalent number of photons as long exposure, while also avoiding motion blur. Although producing a final high SNR image requires image registration which adds computational overhead, with advances in camera technology and dedicated compute units, burst capturing mode is now quite popular (e.g. Deep Fusion [1] and Night Sight [2].)

In the last decade, researchers have proposed various multi-frame noise reduction techniques for burst captures or videos [3,4,5,6]. Many recently proposed burst denoising techniques employ deep learning to improve the state-of-the-art [7,8,9,10,11,12]. However, they are typically not efficient enough for potential deployment (especially on edge devices). Subsequently, they are often trained and evaluated on unrealistic noise models which do not generally account for the signal-dependent and spatially correlated nature of actual noise [13].

In this paper, we propose a more practical end-to-end trainable burst denoising pipeline, which produces visually pleasing results while being significantly faster than the state-of-the-art. Specifically, we start from a 2D model which is capable of encoding high-frequency and high-resolution features from each burst frame, and then extend it to a pseudo-3D model to handle an arbitrary length of burst frame sequence, as presented in Sec. 3 in detail. Our proposed 2D model utilizes both high-resolution and high-frequency deep features, and is trained with realistic noise modeling. The motivation is that the high-resolution features have been validated in recent literature [14,15,16] to be drastically beneficial for preserving fine and detailed information during representation encoding, which is critical for image restoration tasks, especially for denoising. Moreover, an explicit feature decomposition is expected to guarantee the preserving of local details since the implicit decomposition of multi-scale high-resolution architectures might not be enough. After constructing the 2D model, it is further extended to handle burst frames of varying lengths using channel-wise temporal feature shifts, which utilizes temporal cues through all burst frames with limited overhead.

For any burst denoising solution to be practical, it needs to address a few major challenges. Firstly, it needs to be efficient, especially when considering resource constrained devices. Second, it needs to be flexible and scalable, being able to handle an arbitrary length of burst frames. Third, it needs to not only pursue objective quality, but also enhance perceptual quality and balance the trade-off between them as indicated in [17]. Our proposed approach is designed under the guideline to address all these challenges.

Our main contributions are summarized as follows:

**Performance:** Multi-frame denoising tasks are notably efficiency-demanding, and usually require real-time efficiency on terminals and edge devices. We present a novel, end-to-end trainable, deep convolutional burst denoising framework capable of achieving state-of-the-art performance both qualitatively and quantitatively. Additionally, reduced computational requirements demonstrate the efficiency of our model.

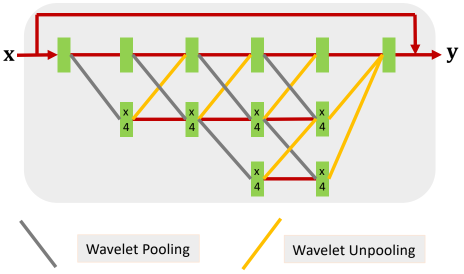


**Fig. 1:** Our proposed architecture for processing one corrupted frame, $x$, to recover one reconstructed frame, $y$. Inter-frame communication is facilitated through mechanisms illustrated in Figure 2.

**Features:** To our knowledge, this is the first work that brings up the joint high-resolution and high-frequency feature extraction and fusion in denoising tasks. Systematic ablation studies and experiments are conducted to validate the efficacy of extracted features on burst denoising.

**Flexibility:** Our model can be used in both single-image (photo) and multi-image input (burst captures or videos) scenarios. For the multi-image input scenario, it could work in either a bidirectional offline manner (burst captures) or a unidirectional online manner (real-time video streams).

## 2 Background and Related Work

**Single-image denoising.** Image denoising has been extensively studied, with many traditional methods being proposed to take advantage of the specific statistics of natural images for reducing noise, including anisotropic diffusion [18], bilateral filters [19], total variation methods [20], domain transform methods such as wavelet transform [21,22], non-local patch-based methods [23], sparsity-based methods [24], and notably, the BM3D method [25]. Due to the popularity of deep neural networks, image denoising algorithms [26,27,28,29,30,31,32,33] have achieved a significant boost in performance. Notable denoising neural networks, DnCNN [26], and IrCNN [28] predict the residual noise present in the image instead of the denoised image, and recently, CBDNet [13] was proposed as a blind denoising model for real photographs. However, it is typically not satisfyingly effective and efficient to apply single-image denoising methods on multi-image data to generate consistent non-flickering results, without utilizing temporal cues.

**Multi-image denoising.** Beyond the longstanding single-image denoising problem, many algorithms have also been proposed for multi-image (burst or video) denoising. Many ideas are shared between the two tasks, though burst denoising methods usually tend to generate or select one denoised frame as output, while video denoising methods aim to generate frame-by-frame output. When burst images or videos are available, noise can be reduced using spatial and temporal correlations. For example, BM3D has been extended to videos by filtering 3D blocks formed by grouping similar 2D patches (VBM3D) [4] or 4D blocks formed by grouping similar spatio-temporal volumes (VBM4D) [5]. [34] proposed a video denoising approach via the empirical Bayesian estimation of space-time patches. Liu et al. [6] in the case of processed images and Hasinoff et al. [10] in the case of raw images showed how to achieve good denoising performance in terms of PSNR and much higher speeds by exploiting temporal redundancy and averaging time-consistent pixels. Godard et al. [7] proposed a recurrent network for multi-frame denoising, where the burst sequence needs to be pre-warped to the reference frame. Mildenhall et al. [8] designed a convolutional network architecture that predicts spatially varying kernels for each of the input frames. These adaptive kernels can be applied to the input burst to correct for small misalignment and generate a clean output. Kokkinos et al. [11] proposed an iterative approach for both burst denoising and demosaicking tasks. Xu et al. [35] extended this approach to 3D deformable kernels for video denoising to sample pixels across the spatial-temporal space. We focus on the burst denoising task in this paper.

**Wavelets in Convolutional Networks.** Wavelets were originally proposed to separate data into different space-frequency components for component-wise analysis at multiple scales, and has been widely used in various image processing tasks [36]. Recently, different works have been proposed to incorporate wavelets with CNNs on various tasks, including deep feature dimension reduction [37], style transfer [38], super-resolution [39], and image denoising [27]. Different with most previous wavelet-based denoising networks which interleave wavelet transform with convolutional layers, we focus on utilizing wavelet transforms only by replacing the feature rescaling step similar to [37] and [38] to explicitly decompose the convolutional features to high-frequency and low-frequency sub-band features, and incorporate with the multi-scale network design.

The rest of this paper is organized as follows: Section 3 presents our proposed model for burst denoising, including the 2D model for feature extraction, and extended pseudo-3D model for temporal feature fusion. The data preparation and noise modeling are described in detail in Section 4. Section 5 demonstrates and analyzes the experimental results. Limitations and failure cases are presented in Section 5.6. We conclude this paper in Section 6.

# 3 Methodology

## 3.1 Overview

Our focus in this work is to generate a single high-quality clean image from a burst of $2N$ noisy frames ($\{X_0, ..., X_{2N-1}\}$) captured by a handheld camera. We consider $X_N$ or $X_{N+1}$, namely the center frame, as the reference frame. We focus on the 8-bit sRGB camera output images as input instead of RAW images since most phones do not retain raw photos (even when supported). To be more practical, we developed our model to process images that are pre-processed by an imaging pipeline. This is more challenging since the noise model is often

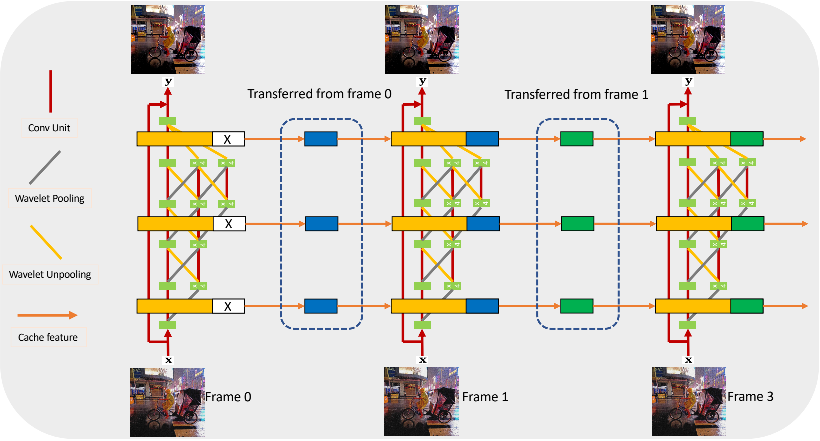


**Fig. 2: The main schematic of our proposed model in unidirectional processing mode.** For each burst frame, the 2D model will individually extract high-resolution and high-frequency features for later fusion. Then interleaved Temporal Shift Modules facilitate inter-frame communication and extend the 2D model to a full pseudo-3D model. As the forward pass of each per-frame branch is executed, a subset of activations are stored in a separate buffer and the corresponding activations from the previous frame are copied to take their place.

considerably altered by different operations such as demosaicking, tone-mapping, and denoising.

In accordance with the goals for our burst denoising task, we designed a new deep learning-based pipeline to process noisy bursts. In the following subsections, we introduce our network architecture, high-resolution and high-frequency feature extraction, temporal feature fusion mechanism, adaptive and conditional versions of the model, and training objective. The architecture of our proposed model is demonstrated in Fig. 2 using the unidirectional frame-by-frame processing manner as an example. The whole pipeline starts from the 2D model to extract high-frequency and high-resolution features from each burst frame, and is extended to the pseudo-3D version through temporal feature shifting. We explored different options for the 2D feature extraction and 3D feature aggregation, and will present in detail in the following subsections. Specifically, Sec. 3.2 demonstrates the design for maintaining high-resolution features and explicitly separating high-frequency features, and Sec. 3.3 further demonstrate the different feature fusion regimes along the temporal dimension.

### 3.2 Features matter in burst denoising

We first introduce how we built the 2D model for extracting the high-resolution and high-frequency features from each frame. This proposed 2D model can be directly used for the single image denoising task if needed.

**High-resolution features.** Recently high-to-low convolution and on-the-fly fusion techniques have been proposed for various computer vision tasks [14,15,16]. This helps maintain high-resolution representations through the whole convolutional feature extraction process. Our 2D model builds upon HRNet [15], one of the recently proposed multi-resolution convolution and fusion architectures, for high-resolution feature encoding. Our motivation is that the high-resolution features are expected to enhance the objective quality of the generated best denoised frame, while maintaining local detail. Our 2D network structure is illustrated below containing 3 parallel streams.

$$\begin{tikzcd}
\mathcal{N}_{11} \arrow[r] \arrow[dr] & \mathcal{N}_{21} \arrow[r] & \mathcal{N}_{31} \\
 & \mathcal{N}_{22} \arrow[r] \arrow[dr] & \mathcal{N}_{32} \\
 & & \mathcal{N}_{33}
\end{tikzcd} \tag{1}$$

where $\mathcal{N}_{s,r}$ is a sub-stream in the $s$th stage and $r$ is the resolution index. The resolution index of the first stream is $r = 1$. The resolution of index $r$ is $\frac{1}{2^{r-1}}$ of the resolution of the first stream. The highest resolution features are preserved along the top stream $\mathcal{N}_{s,1}$, and fused along with other streams at last.

**High-frequency features.** Besides maintaining high-resolution features, to achieve high visual quality, a denoising model should faithfully recover the structural information of a given noisy frame while removing noise. Though the aforementioned multi-scale architectures [14,15,16] are usually designed to implicitly decompose features into different frequencies, we find that an explicit decomposition is also beneficial to the multi-frame denoising task. (see Sec. 5).

Inspired by recent work on wavelet pooling and unpooling for convolutional networks [37,38], we propose to integrate wavelet decomposition along with the branch of high-resolution features in the multi-scale learning. That being that, the maintained high-resolution features are explicitly decomposed to different frequency bands, and processed on-the-fly before fused by wavelet unpooling. With several popular wavelets designs being proposed before, we finally choose Haar wavelet to efficiently split the original features into channels that capture different frequency bands. It results in better denoising and corresponding signal reconstruction. Specifically in our model, Haar wavelet pooling has four kernels, $\{\mathrm{LL}^\top\ \mathrm{LH}^\top\ \mathrm{HL}^\top\ \mathrm{HH}^\top\}$, where the low and high pass filters are

$$L^\top = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \end{bmatrix}, \quad H^\top = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 & 1 \end{bmatrix}. \tag{2}$$

Thus, unlike common pooling operations, the output of the Haar wavelet pooling has four channels. Here, the low-pass filter captures smooth surface and texture while the high-pass filters extract vertical, horizontal, and diagonal edge-like information. For simplicity, we denote the output of each kernel as LL, LH, HL, and HH, respectively.

One favorable property of wavelet pooling is that the original signal can be reconstructed by mirroring its operation; i.e., wavelet unpooling, as illustrated in detail in the supplementary doc. More precisely, wavelet unpooling can fully recover the original signal by performing a component-wise transposed-convolution.

With this property, our proposed model can reconstruct an image with minimal information loss and noise amplification. In contrast, there is no exact inverse for max-pooling so that it is difficult for the encoder-decoder alike networks in previous work to fully recover the signal. To sum, both the high-resolution and high-frequency design are utilized in our final 2D model capable of frame-by-frame denoising, which will be further extended to a (pseudo-) 3D model for multi-frame burst denoising as instructed in Sec. 3.3.

### 3.3 Temporal fusion of deep features

In this subsection, we introduce the different temporal feature fusion mechanisms we utilized for aggregating the information from all burst frames, including the conventional 3D convolution, and two pseudo-3D regimes, temporal max-pooling and temporal feature shifting. The final model adopts the temporal feature shifting as the feature fusion mechanism.

**3D Convolution.** For multi-frame feature extraction, there usually exist two flavors of 3D convolution methods, i.e. either feeding all frames into the network at once (offline mode), or feeding a certain number of frames (e.g. 3 or 5) in a unidirectional sliding window manner (online mode). For burst denoising, the 3D network aggregates all frames and jointly learn spatio-temporal features. However, 3D CNNs are computationally expensive [40] and more prone to over-fitting. As a result, we use 3D convolution as one version of our feature fusion regime, and further investigate 2 pseudo-3D learning mechanisms for the ablation study.

**Temporal Max-pooling.** Compared to relatively computationally intensive real 3D convolutions, pseudo-3D convolution strategies such as temporal max-pooling have been recently proposed as an alternative for handling feature fusion. Zaheer et al. [41] and Qi et al. [42] show that any function that maps an unordered set into a regular vector (or an image) can be approximated by a neural network. Aittala et al. [9] successfully applied this idea to the burst deblurring task. In this version of our model, the individual input frame of the set are first processed separately by identical neural networks with tied weights, yielding a vector (or an image) of features for each. The features are then pooled by a symmetric operation, by evaluating either the mean or maximum value of each feature across the members. This scheme gives the individual frames in the burst a principled mechanism to contribute their local feature capturing the likely content of the sharp image.

**Temporal Feature Shifting.** Another alternative to 3D convolution method, the Temporal Shift Module (TSM) [43] has been successfully used for video understanding, and is ported as a feature fusion mechanism for our burst denoising model. Specifically, given a burst image sequence $B$, we take all $2N$ frames $\{X_0, ..., X_{2N-1}\}$ in the sequence. The aforementioned 2D CNN baseline model would process each of the frames individually with no temporal modeling, and the output results are averaged to give the final burst denoising prediction. In contrast, the TSM module has the same parameters and time cost of computation as 2D model. During inference, the frames are processed independently similar to 2D CNNs. The TSM is then inserted in each residual block (similar to temporal max-pooling) which enables temporal information fusion at no computational overhead. In our final model, TSM shifts a small proportion (typically 1/8) of channels along the temporal dimension, enabling the temporal multiply-accumulate to be computed inside the 2D convolution of channels instead of using an explicit time dimension. The temporal shift can be either uni- or bi-directional. In contrast to temporal max-pooling, TSM's can preserve information ordering. This allows the model to handle scene motion and object motion appearing in consecutive burst frames. For each inserted temporal shift module, the temporal receptive field will be enlarged by 2, as if running a convolution with the kernel size of 3 along the temporal dimension. Hence, the final integrated model has a very large temporal receptive field to conduct highly complicated temporal modeling.

Using the unidirectional temporal feature shifting for burst image denoising has some unique advantages. First, for each frame, we only need to replace and cache 1/8 of the features, without any extra computations. Hence, the latency of per-frame prediction is almost the same as the 2D CNN baseline. Both 3D Convolution and Temporal max-pooling methods need all the frames to be fed in the network at once for the inference, which leads to increased latency. Additionally, the temporal feature shifting enables temporal fusion on the fly at all levels, improving the model's robustness to scene motions. In contrast, most online methods only allow late temporal fusion after feature extraction.

### 3.4 Loss function

With ground truth reference image $Y$ and the denoised frame $\hat{Y}$ in linear space, we directly use a $L_1$ loss on both the pixel intensities and gradient intensities to train the proposed denoising network:

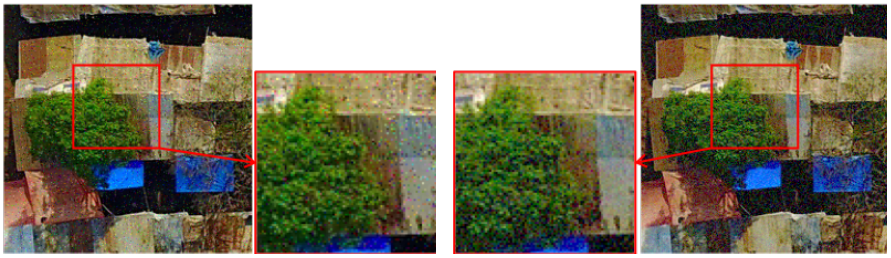
$$\ell(\hat{Y}, Y) = \lambda_1 \left\| \hat{Y} - Y \right\|_1 + \lambda_2 \left\| \nabla \hat{Y} - \nabla Y \right\|_1 . \tag{3}$$

which tries to make the average of all denoising estimations close to the ground truth $Y$, with $\lambda_1 + \lambda_2 = 1$ (both set to a constant of 0.5 in our experiments). For fair comparisons with previous approaches, we do not utilize any adversarial training mechanism or perceptual loss in the proposed model to favor the perceptual metric. [8] proposed to add annealing loss for multi-frame training to avoid the convergence at an undesirable local minima in training, which we finally did not use as we have not noticed significant differences in our experiments.

## 4 Data preparation

### 4.1 Camera Pipeline.

In this work, we follow the camera simulation pipeline in [44] to model realistic noise, and generate noisy burst sequences for all the synthetic training samples.



*Gaussian noise modeling* *Realistic noise modeling*

**Fig. 3:** Noisy image comparison between Gaussian noise modeling and our realistic noise modeling with inverse ISP.

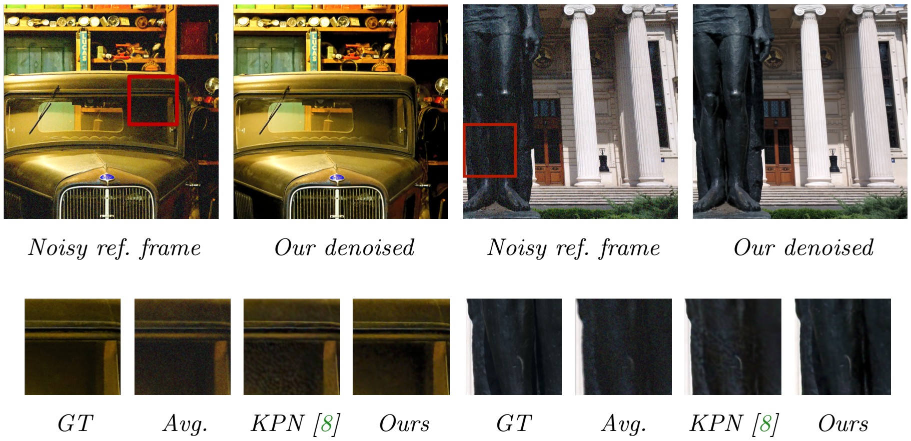
We encourage readers to check our brief review in the *supplementary document* for more details of the camera pipeline we choose and corresponding comparisons. Specifically for the noise modeling, following [44], the noise parameters and factors (Gaussian and Poisson noise as additive and multiplicative operations) are randomly determined for all training samples, as 0-0.1 for Gaussian standard deviation, 0-0.02 for Poisson multiplication factor, with post-processing. For static burst synthesis, slight camera motion is simulated with random cropping and jittering (2-8 pixels) of a single image from the DIV2K dataset. Unfortunately, scene/object motion cannot be easily added without extra processing such as inpainting and depth/boundary prediction. For dynamic burst synthesis, we only add noises since camera and object motion already exist in the original data of the Vimeo90K dataset.

### 4.2 Datasets and synthetic burst generation

**DIV2K dataset.** DIV2K (DIVerse 2K resolution high-quality images) [45] is a dataset composed of 800 high (2K) resolution images for training with 100 images each for validation and testing respectively (the testing set is not publicly available). We used the available high-resolution images with the *static* noisy bursts synthesis for training and testing respectively.

**Vimeo90K dataset.** Vimeo90K [46] is built upon $5,846$ selected videos from *vimeo.com*, which covers large variety of scenes and actions. The subset we used for training contains $91,701$ 7-frame sequences, extracted from 39K selected video clips. We used all available $91,701$ 7-frame sequences (original clean images) with our realistic noise modeling for the *dynamic* noisy bursts synthesis and training (original data containing various camera and scene/object motion as demonstrated on the dataset page). Note that though our proposed model supports by design bursts of arbitrary size, this is not the case for the network proposed in [8]. For a fair comparison, we duplicate the last frame in each 7-frame sequence to generate a 8-frame burst for training.

**Real noisy burst captures.** We also captured a collection of realistic burst sequences with a handheld device. Most of the bursts are captured under low-light conditions where burst denoising is typically most useful. There are no ground-truth clean frames available for this set. We used these real noisy data to qualitatively evaluate the generalization capability of our model trained on synthetic bursts with realistic noise modeling.



*Noisy ref. frame* *Our denoised* *Noisy ref. frame* *Our denoised*

*GT* *Avg.* *KPN [8]* *Ours* *GT* *Avg.* *KPN [8]* *Ours*

**Fig. 4:** The qualitative comparison of our burst denoising model with state-of-the-arts on synthetic burst captures. Our method well preserves the local details while avoiding the over-smooth issue. It also introduces less distortion than KPN in terms of the perceptual quality.

# 5 Experimental Results

## 5.1 Overview

After setting up the burst denoising benchmarks, we conducted various experiments for the ablation study on different modules of our model, burst image denoising, and efficiency analysis. We would like to compare our method to two most recent deep learning based approaches [11,47]. However, there is no model or testing data released from [47] as a preprint work. The proposed method in [11] strictly depends on the Enhanced Correlation Coefficient (ECC) estimation of the warping matrix that aligns every observation to the reference frame while all the other methods, including our proposed model, do not have this requirement. While there is no easy way to conduct a fair comparison with similar training regimes, we qualitatively evaluate our method on the public released testing set from [11] (see supplementary), and also tested its computational efficiency in Table 3 for a better experimental completeness.

**Implementation and training details.** We implement our method using PyTorch [48]. The full proposed model is trained on two NVIDIA GeForce RTX 2080 Ti GPU with 11 GB of memory each. We use the Adam optimizer [49] and the batch size is 16. The initial learning rate is $10^{-3}$ and is reduced to $3 \times 10^{-5}$ after 80 epochs, which takes around 26 hours in total. For the burst image denoising experiments, most DIV2K and Vimeo90K data are utilized for training, and the remaining DIV2K and Vimeo90K data are used for testing on static burst denoising and dynamic burst denoising respectively. For fair comparisons, all learning-based approaches such as KPN are trained from scratch on the DIV2K/Vimeo90K based data,

**Table 1:** *Ablation study.* Investigation of the contribution of different modules in our proposed model. The best results in PSNR (dB) on values on DIV2K [45] are reported. The + symbol indicates the optional modules, and the × symbols indicates the exclusive modules and only one of them would work.

| **Model** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| + high-res feature | | | | ✓ | | ✓ | ✓ | ✓ |
| + high-freq feature | | | | | ✓ | ✓ | ✓ | ✓ |
| × 3D convolution | ✓ | | | | | ✓ | | |
| × Temporal max-pooling | | ✓ | | | | | ✓ | |
| × Temporal feature shifting | | | ✓ | ✓ | ✓ | | | ✓ |
| PSNR (in dB) | 32.35 | 32.18 | 32.34 | 32.79 | 32.75 | 32.82 | 32.73 | **32.88** |

**Evaluation metrics.** We evaluate our proposed model and state-of-the-art methods based on the standard Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) image quality metrics. Furthermore, there usually exists a trade-off between objective quality and perceptual quality in image restoration tasks, as indicated in [17] and [39]. Compared with other burst denoising approaches, our model better preserves the local details in reconstruction due to utilizing the high-resolution and high-frequency features, thus is prone to better perceptual quality and alleviates the commonly occurred over-smoothing issues such as in VBM4D [5]. Therefore, besides the two objective metrics, we also introduce one recently popular deep learning-based perceptual metric, namely Learned Perceptual Image Patch Similarity (LPIPS) [50], for a more comprehensive comparison and analysis. LPIPS is more sensitive to the distortions from deep learning-based representations, and prone to the human perceptual judgments. For fair comparisons, we do not utilize any adversarial training mechanism or perceptual loss in the proposed model to favor this metric.

### 5.2 Ablation study

To comprehensively analyze the contributions and significance of different modules in our proposed method, we first quantitatively assess our full model with a set

**Table 2:** *The mean PSNR, SSIM, and LPIPS burst denoising results of our proposed model compared with state-of-the-art algorithms evaluated on the synthetic static burst denoising data generated from DIV2K (left) and dynamic data generated from Vimeo90K (right).*

| **Method** | **PSNR↑** | **SSIM↑** | **LPIPS↓** | **PSNR↑** | **SSIM↑** | **LPIPS↓** |
|---|---|---|---|---|---|---|
| Noisy reference frame | 23.38 | 0.6841 | 0.3361 | 23.33 | 0.5337 | 0.4255 |
| Burst average | 24.73 | 0.6923 | 0.3159 | 21.39 | 0.5164 | 0.4513 |
| VBM4D [5] | 30.81 | 0.8637 | 0.2166 | 27.05 | 0.6922 | 0.2869 |
| VNLB [34] | 31.52 | 0.8841 | 0.1824 | 27.44 | 0.7043 | 0.2720 |
| KPN [8] | 32.64 | **0.8920** | 0.1772 | **28.67** | **0.7319** | **0.2357** |
| Our Proposed | **32.88** | 0.8855 | **0.1742** | 28.61 | 0.7246 | 0.2366 |

of ablations on a synthetic test set, followed by an analysis of its interpretability. We choose the basic U-Net like residual denoising model as the baseline backbone 2D model for ablation study in feature encoding. The main designs we aim to analyze in our model include: 1) High-resolution feature encoding: w/ or w/o the three-stage multi-scale high-resolution convolutional network; 2) High-frequency feature encoding: w/ or w/o the interleaved wavelet pooling and unpooling operations to explicitly decompose the feature maps. For temporal feature fusion, only one mechanism will be integrated with 2D backbones to produce the final denoised output.

As illustrated in Table 1, when we directly utilize the baseline backbone model without preserving high-resolution and decomposing high-frequency features, the denoising network performs fairly well on the testing data. Both the high-resolution and high-frequency features would consistently boost the denoising performance, though the high-resolution features tend to make slightly more contributions.

As to the temporal feature fusion mechanisms, temporal feature shifting performs comparably well as the 3D convolution with the baseline model, and both outperforms the temporal max-pooling mechanism. However, temporal feature shifting demonstrates the capability to be better integrated with high-resolution and high-frequency features, most possibly due to the ability to learn shifting the most salient features temporally, and better aggregate cues from all available burst frames. Combining all the best practices, our full model achieves the top performance on the testing benchmark.

### 5.3 Burst denoising qualitative evaluation.

We evaluate different methods first on *static* noisy bursts derived from DIV2k, and then on *dynamic* noisy bursts derived from Vimeo90K. Here, *static* means the noisy bursts are generated by altering a single image with random jitter, disturbance, and shifts, then, finally, realistic noise is added. Therefore, these bursts contain no scene motion, only camera motion. In contrast, the *dynamic* noisy bursts contain scene motion that cannot be fully modeled by camera stabilization alone.

Fig. 4 shows examples of the denoising results from the proposed model and state-of-the-arts. As claimed in recent kernel sampling based methods such as [8] and [47], the patch-based methods such as VBM4D, and direct prediction based deep learning methods which directly synthesize the denoising results (either producing the clean or residual image) is prone to generate over-smoothed results, and lose the local detail. However, the qualitative results validate the efficacy of our method, which effectively adopts the high-quality features, and faithfully synthesizes and reconstructs the details.

Also, the slight simulated camera motion are equally well tackled by our proposed model, though the kernel-sampling based methods such as KPN are naturally more robust to slight disturbances, and still outperforms the proposed method by a little margin on Vimeo90K dataset.

**Table 3:** *Comparison of inference latency (running time per frame) during testing.* We report the time of each method to denoise one color frame of resolution 720 × 720. For burst denoising, the overall latency can be computed by multiplying the per-frame latency with the total number of burst frames (e.g., 8 in our experiments). All deep learning based methods are evaluated on one single NVIDIA GeForce RTX 2080 Ti GPU. *Note: Values displayed for [34] and [11] do not include the time required for pre-processing (e.g., motion or warping matrix estimation), which can be time costing.

| **Method** | **Latency (s)** ↓ | **Megapixels/s** ↑ |
|---|---|---|
| VNLB* [34] | 387.24 | 0.0052 |
| VBM4D [5] | 131.49 | 0.0153 |
| KPN (GPU) [8] | 0.597 | 0.8685 |
| Iterative (GPU)* [11] | 0.140 | 3.7029 |
| Our Proposed (GPU) | **0.064** | **8.1403** |

### 5.4 Burst denoising quantitative evaluation.

Furthermore, we quantitatively evaluate our model along with state-of-the-arts on both the *static* and *dynamic* testing set, and the results are demonstrated in Table 2. Generally, our proposed model performs comparably decently with KPN on different evaluation metrics. In terms of the objective quality (PSNR and SSIM), KPN leads a slight margin, especially on the *dynamic* data which contain scene and object motion. However, our proposed model consistently outperforms all state-of-the-arts in terms of perceptual quality, evaluated by the deep learning based metric *LPIPS*.

### 5.5 Algorithm efficiency

While achieving comparable performances on burst denoising, we further evaluate all models on efficiency. As illustrated in Table 3, benefiting from the efficient pseudo-3D feature fusion design, our model is significantly faster than state-of-the-arts. Empirically, the speed-up is mainly due to the concise network design and the way temporality is handled. The Haar wavelet transform is efficient, though bringing an increase of space complexity (more intermediate feature maps to store) of the model.

### 5.6 Limitations.

The main limitation of our proposed model is that it is still trained in a non-blind denoising fashion without taking a noise estimation map as input from an individual noise estimator such as the ones proposed in [13,51], thus lacking the capability of adaptive noise-aware burst denoising which is recently popular in single-image denoising methods. Fig. 2 in the *supplementary document* demonstrates a failure of our model on a severely corrupted burst, which is considerably noisier than examples seen during training. The model struggles to recover the corrupted detail, but instead produces unsatisfactory artifacts. Integrating the proposed model with a noise level estimation mechanism is a promising future research direction to mitigate this problem.

### 5.7 Generalization to real burst captures.

Finally, we evaluate the generalization capability of our model on real burst noisy frames. The results qualitatively validates that our proposed model performs reasonably well on real noisy burst captures, while only being trained on synthetic data with realistic noise modeling. We aim to collect more real noisy bursts along with corresponding long-exposure shots as the ground truth, though the problem is that this regime only works for static scenes.

## 6 Conclusion and future work

We propose the first unified end-to-end trainable deep burst denoising framework which effectively utilizes high-resolution and high-frequency features. The proposed model excels both quantitatively and qualitatively, while facilitating joint spatial-temporal modeling through burst frames at no extra cost. Compared to the prior state-of-the-art, our the proposed framework is more scalable, lighter, and faster, thus enabling low-latency burst denoising on edge devices.

Inspired by recent work on extreme low-light image [52] and video [53] enhancement, our future work is to support burst shots captured in the extreme dark environments where noise is significantly amplified. Also, we aim to incorporate a noise estimator for burst captures to support blind (noise-aware) denoising.

## References

1. : A Deep Look into the iPhone's new Deep Fusion Feature. https://tinyurl.com/deepfusion Accessed: 2019-11-04. 2
2. : Night Sight: Seeing in the Dark on Pixel Phones. https://tinyurl.com/googlenightsight Accessed: 2019-11-04. 2
3. Buades, T., Lou, Y., Morel, J.M., Tang, Z.: A note on multi-image denoising. In: 2009 International Workshop on Local and Non-Local Approximation in Image Processing, IEEE (2009) 1–15 2
4. Liu, C., Freeman, W.T.: A high-quality video denoising algorithm based on reliable motion estimation. In: ECCV, Springer (2010) 706–719 2, 4
5. Maggioni, M., Boracchi, G., Foi, A., Egiazarian, K.: Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms. IEEE Transactions on image processing (TIP) **21**(9) (2012) 3952–3966 2, 4, 11, 12, 13
6. Liu, Z., Yuan, L., Tang, X., Uyttendaele, M., Sun, J.: Fast burst images denoising. ACM Transactions on Graphics (TOG) **33**(6) (2014) 232 2, 4
7. Godard, C., Matzen, K., Uyttendaele, M.: Deep Burst Denoising. In: ECCV. (2018) 538–554 2, 4
8. Mildenhall, B., Barron, J.T., Chen, J., Sharlet, D., Ng, R., Carroll, R.: Burst denoising with kernel prediction networks. In: CVPR. (2018) 2502–2510 2, 4, 8, 9, 10, 12, 13
9. Aittala, M., Durand, F.: Burst image deblurring using permutation invariant convolutional neural networks. In: ECCV. (2018) 731–747 2, 7
10. Hasinoff, S.W., Sharlet, D., Geiss, R., Adams, A., Barron, J.T., Kainz, F., Chen, J., Levoy, M.: Burst photography for high dynamic range and low-light imaging on mobile cameras. ACM Transactions on Graphics (TOG) **35**(6) (2016) 192 2, 4
11. Kokkinos, F., Lefkimmiatis, S.: Iterative residual cnns for burst photography applications. In: CVPR. (2019) 5929–5938 2, 4, 10, 13
12. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: ICCV. (2017) 764–773 2
13. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. In: CVPR. (2019) 1712–1722 2, 3, 14
14. Ke, T.W., Maire, M., Yu, S.X.: Multigrid neural architectures. In: CVPR. (2017) 6665–6673 2, 6
15. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. arXiv preprint arXiv:1908.07919 (2019) 2, 6
16. Chen, Y., Fang, H., Xu, B., Yan, Z., Kalantidis, Y., Rohrbach, M., Yan, S., Feng, J.: Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. arXiv preprint arXiv:1904.05049 (2019) 2, 6
17. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6228–6237 2, 11
18. Weickert, J.: Anisotropic diffusion in image processing. Teubner, Stuttgart (1998) 3
19. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: ICCV. (1998) 839–846 3
20. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena **60**(1-4) (1992) 259–268 3

21. Antonini, M., Barlaud, M., Mathieu, P., Daubechies, I.: Image coding using wavelet transform. IEEE Transactions on image processing (TIP) **1**(2) (1992) 205–220 3
22. Portilla, J., Strela, V., Wainwright, M.J., Simoncelli, E.P.: Image denoising using scale mixtures of gaussians in the wavelet domain. Trans. Img. Proc. **12**(11) (November 2003) 1338–1351 3
23. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: CVPR. Volume 2., IEEE (2005) 60–65 3
24. Elad, M., Aharon, M.: Image denoising via learned dictionaries and sparse representation. In: CVPR. Volume 1., IEEE (2006) 895–900 3
25. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Transactions on Image Processing (TIP) **16** (2007) 2080–2095 3
26. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Transactions on Image Processing (TIP) **26**(7) (2017) 3142–3155 3
27. Liu, P., Zhang, H., Zhang, K., Lin, L., Zuo, W.: Multi-level wavelet-cnn for image restoration. In: CVPR Workshop. (2018) 773–782 3, 4
28. Zhang, K., Zuo, W., Gu, S., Zhang, L.: Learning deep cnn denoiser prior for image restoration. In: CVPR. (2017) 3929–3938 3
29. Laine, S., Lehtinen, J., Aila, T.: High-quality self-supervised deep image denoising. arXiv preprint arXiv:1901.10277 (2019) 3
30. Batson, J., Royer, L.: Noise2self: Blind denoising by self-supervision. In: ICML. (2019) 524–533 3
31. Anwar, S., Barnes, N.: Real image denoising with feature attention. In: ICCV. (2019) 3
32. Cha, S., Moon, T.: Fully convolutional pixel adaptive image denoiser. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 4160–4169 3
33. Gu, S.e.a.: Self-guided network for fast image denoising. In: ICCV. (2019) 3
34. Arias, P., Morel, J.M.: Video denoising via empirical bayesian estimation of space-time patches. Journal of Mathematical Imaging and Vision **60**(1) (2018) 70–93 4, 12, 13
35. Xu, J., Huang, Y., Liu, L., Zhu, F., Hou, X., Shao, L.: Noisy-as-clean: Learning unsupervised denoising from the corrupted image. arXiv preprint arXiv:1906.06878 (2019) 4
36. Wang, J.Z.: Wavelets and imaging informatics: A review of the literature. Journal of Biomedical Informatics **34**(2) (2001) 129–141 4
37. Williams, T., Li, R.: Wavelet pooling for convolutional neural networks. In: ICLR. (2018) 4, 6
38. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic style transfer via wavelet transforms. In: ICCV. (2019) 4, 6
39. Deng, X., Yang, R., Xu, M., Dragotti, P.L.: Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution (2019) 4, 11
40. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P.: End-to-end learning of geometry and context for deep stereo regression. In: ICCV. (2017) 7
41. Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J.: Deep sets. In: NeurIPS. (2017) 3391–3401 7
42. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: NeurIPS. (2017) 5099–5108 7
43. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: ICCV. (2019) 7

44. Jaroensri, R., Biscarrat, C., Aittala, M., Durand, F.: Generating training data for denoising real rgb images via camera pipeline simulation. arXiv preprint arXiv:1904.08825 (2019) 8, 9
45. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: CVPR Workshop. (July 2017) 9, 11
46. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. International Journal of Computer Vision (IJCV) **127**(8) (2019) 1106–1125 9
47. Xu, X., Li, M., Sun, W.: Learning deformable kernels for image and video denoising. arXiv preprint arXiv:1904.06903 (2019) 10, 13
48. Steiner, B., DeVito, Z., Chintala, S., Gross, S., Paszke, A., Massa, F., Lerer, A., Chanan, G., Lin, Z., Yang, E., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS **32** (2019) 11
49. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 11
50. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. (June 2018) 11
51. Zhou, Y., Jiao, J., Huang, H., Wang, Y., Wang, J., Shi, H., Huang, T.: When awgn-based denoiser meets real noises. arXiv preprint arXiv:1904.03485 (2019) 14
52. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: CVPR. (2018) 3291–3300 14
53. Chen, C., Chen, Q., Do, M., Koltun, V.: Seeing motion in the dark. In: ICCV. (2019) 14