This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Rethinking Compressive Sensing: A Compression Framework for Video Super-Resolution

Ruthy Katz* Adi Teitel* Moran Mordechay Adi Falik Eli Bery Maya Mayberg
Corephotonics, Israel
{rkatz, ateitel, mmordechay, afalik, ebery, mmayberg}@corephotonics.com

Abstract

Sensors in mobile phones have significantly increased in dimensions, reaching an impressive 200M pixels. However, the output resolution of images and videos stored on a device is still constrained. In most modern high-resolution imaging sensors, pixel binning is performed to maintain a good signal-to-noise ratio (SNR) and accommodate memory and bandwidth resource limitations. This presents a challenge, as extensive spatial compression leads to irreversible data loss. Although Video Super-Resolution (VSR) methods can be applied to reconstruct high-resolution videos, they may still lead to unsatisfactory reconstruction results. To address this challenge, we rethink the role of Compressive Sensing (CS) in the video domain. We propose a novel on-sensor sampling scheme and examine different dynamic binning mask layouts with temporal variations. For reconstruction, we use conventional VSR models and demonstrate that combining them with our compression strategy enhances image quality. Extensive experiments on benchmark datasets confirm that our method improves reconstruction quality both quantitatively and qualitatively.

1. Introduction

Each year, 1.6 trillion photos are captured worldwide using smartphones, establishing mobile phones as the dominant force in the photography market. This is connected to the ever-improving quality of smartphone cameras and the expansion of image sensor dimensions. Despite a substantial increase in sensor size - reaching up to 200M pixels [10, 46] - the default image file save settings remain limited to a modest 12.5M pixels for still images and Full HD (2M pixels) for video sequences.

In high-resolution sensors, physical constraints impose limitations as smaller pixels are densely packed onto the sensor surface. While this theoretically may help capture more details, it can reduce the per-pixel photon reception, resulting in lower SNR. Such sensors can acquire extremely high-resolution images under optimal lighting conditions, but perform poorly in low-light environments. Moreover, higher pixel counts demand increased processing bandwidth and storage capacity, as they generate substantial volumes of data for processing. To achieve higher signal-to-noise ratios while maintaining high frame rates, pixel binning is commonly integrated at the hardware level [19, 26].

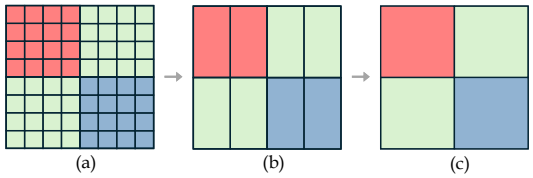


Figure 1. Illustration of sensor layout with a 4×4 pixel structure cell (a) and the conventional 16Sum binning process with a 16:1 compression rate. Where (b) is an intermediate stage: 8-summation and (c) is the final binned low-resolution output.

For 200M pixel sensors [10, 46], implementing a 16:1 binning ratio, denoted as 16Sum, reduces the output to a 12.5M pixel image, resulting in significantly larger effective pixels, that can capture more light. The binning process aggregates a 4×4 pixel structure by first summing the values of 4×2 pixels [46], then averaging the two resulting sums to produce a single pixel output, as illustrated in Fig. 1. However, pixel binning presents a fundamental trade-off: while it reduces image noise, it simultaneously decreases spatial resolution and fine detail preservation. When upscaling the images back to a higher resolution during post-capture, the output quality is limited to either utilizing mobile digital zoom, or reconstruction using Single-Image Super-Resolution (SISR), or VSR methods.

VSR aims to restore a series of high-resolution (HR) frames by exploiting the temporal information within low-resolution (LR) frames [32]. It produces Super-Resolution (SR) results with higher accuracy than SISR, by leveraging

*Equal contribution

the complementary sub-pixel information from multiple video frames. Nevertheless, there are several limitations for the existing VSR methods, which lead to unsatisfactory quality of the results. One limitation is the underlying assumption regarding motion in a scene. In regions characterized by minimal to no motion, temporal correlations across frames may not offer additional information to enhance the super-resolved output.

In this work, we propose a novel compression framework, that addresses this challenge and allows for a more robust SR reconstruction. Inspired by Image Compressive Sensing (ICS) methods [6, 9, 44], we suggest encoding a series of video frames using different binning-masks layouts with temporal variations. While traditional CS involves taking multiple measurements of a single image, this approach is impractical for maintaining high frame rates. To address this, we leverage the assumption that subtle changes between consecutive frames can be addressed as noise, and shift the problem to the video domain. We introduce a compression block to obtain on-sensor compressed measurements from HR images, acquiring a single measurement per frame. However, implementing such non-regular sampling, where some pixels are excluded from the sampling process, lowers the fill-factor and leads to low SNR [23, 24, 39, 40]. Since such sensitivity loss is unacceptable for most practical image sensors, we propose exploiting the existing 2-stage binning process (Fig. 1) to take two measurements per frame using complementary masks, forming a complete tiling of the sensor area. This approach enables binning of eight HR pixels, achieving both high SNR and frame rates comparable to conventional rectangular binning. Our proposal includes multiple mask designs (Fig 6), with image reconstruction performed using standard VSR techniques.

To the best of our knowledge, we are the first to apply CS formulation for spatial compression and reconstruction of video frames. This is a departure from classic CS theory [3, 12, 15], which assumes multiple measurements of the same signal. Moreover, we will show that our method preserves user experience and can be seamlessly integrated into the image acquisition pipeline of smartphones. It ensures that the compressed measurements stored on the device can be directly accessed and visualized as images, unlike conventional CS methods, which generate a compressed matrix. Extensive experiments are carried out, showing that using our compression approach combined with VSR models outperforms the reconstruction quality compared to conventional square binning. We select three representative VSR models [4, 30, 49] to show the generality of our proposed compression and perform the majority of our experiments using BasicVSR [4]. Our approach is not limited to on-sensor image acquisition process and can also further benefit VSR tasks that have control over the acquisition of LR images.

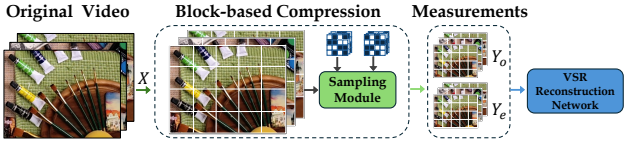


Figure 2. Pipeline illustration of the compression and reconstruction flow. Images are divided into blocks and sampled using the sampling module. Block size is bigger for better visualization.

2. Related work

2.1. Compressive sensing

CS [3, 12, 15] is a theoretical framework that requires significantly fewer measurements than those dictated by Nyquist sampling, assuming the signal is sparse in some domain. Let $x \in R^N$ be the original signal, and $\Phi \in R^{M \times N}$ be a controlled measurements matrix, where $M < N$. Then the measurements $y \in R^M$ are given by:

$$y = \Phi x = A\theta, \quad A = \Phi\Psi, \tag{1}$$

where $\Psi \in R^{N \times N}$ is the sparsifying basis, $A \in R^{M \times N}$ is the effective measurement matrix and $\theta \in R^N$ is the sparse representation of x. Recovering x from the measurements y is an ill-posed under-determined inverse problem. However, the sparse property of the signal x enables its recovery.

Over the past years, there have been a large number of classical algorithms proposed in the literature for solving this optimization problem [7, 11, 18, 25, 34, 45, 48]. Some works have developed optimization algorithms for Φ [1, 8, 16, 17, 35], often by minimizing the mutual coherence of A [13–15]. The authors of [38] proposed using random binary matrices to measure light intensities corrupted by Poisson noise. Each element in the measurement matrix is 0 with probability p and $\frac{1}{M}$ with probability $1-p$, where $p \in (0, 1)$ is a given parameter. The authors formulate recovery guarantees using such matrices, and show that the optimal value of p is $p = 0.5$.

Natural images are known for sparsity in various transform domains [3], thus CS theory can be applied to image acquisition. Due to the high resolution of the image, Block-based Compressive Sensing (BCS) [22] was introduced to improve the efficiency of image sampling by employing a block-by-block sampling. In recent years, with the rise of deep neural networks (DNNs), many DNN-based models have gained much success for image CS reconstruction and achieved state-of-the-art (SOTA) results [21, 43, 44, 51]. A major limitation of these methods is that the compressed signal cannot be visualized as an image due to significant

compression and an incompatible number of channels, typically only one (grayscale) or three (RGB) are supported. This limits their applicability to scenarios that do not require visualization or direct interaction, thereby making them unsuitable for smartphones.

2.2. Compressive image sensors

Several works propose implementations for compressive sensing in hardware. In [28, 36] the authors implement a block-based compressive image sensor with random binary measurements matrix using multiplexers and $\Sigma\Delta$ analog to digital converters (ADCs). While in [28, 36] each pixel can participate in multiple measurements, in [29, 37] each pixel is measured exactly once. In [29] the authors implement a column based compressive sensor using a deterministic measurements matrix. Pixels are summed along the column in groups of 4 by a $\Sigma\Delta$ ADC in order to acquire digital measurements. This is done for all the columns in parallel. In [37] the sensor is block-based, but the compression within the block is only along the rows. The pixels of the row are inputs to a switched capacitor (SC) matrix multiplier, which implements multiplication by random ± 1 values and summation in groups of 4. Then, the compressed row is quantized by ADCs. This is done sequentially for all the rows in the block. To the best of our knowledge, no compressive sensing system implements frame-varying dynamic binning masks. However, this capability could be achieved by adapting existing designs with appropriate modifications to support dynamic spatial binning.

2.3. Non-regular sampling

In [23], the authors propose an end-to-end neural network based solution that is capable of super resolving single images that were downscaled using different non-regular sensor layouts, such as quarter sampling and three-quarter sampling. Their proposal is to replace the bicubic upsampling used as initial reconstruction for SISR algorithms with a data-driven approach, using a locally fully connected reconstruction (LFCR) network, which results in an improved reconstruction quality.

In patent application [20], the inventors introduce a Tetromino sensor, in which the pixels contributing to a certain measurement form the shape of the letters T, L or Z. As in [29, 37], each pixel is measured exactly once. Recent work [24] investigates T-Tetromino pixel layouts, where each T-shaped element aggregates four higher-resolution pixels (Fig. 5a.). The authors perform single image reconstruction with various algorithms, motivated by the premise that non-regular T-shaped sampling can outperform conventional square binning in terms of reconstruction quality. This is due to intrinsic limitations of regular sampling in reconstruction algorithms, introducing aliasing in the presence of high-frequency scene components.

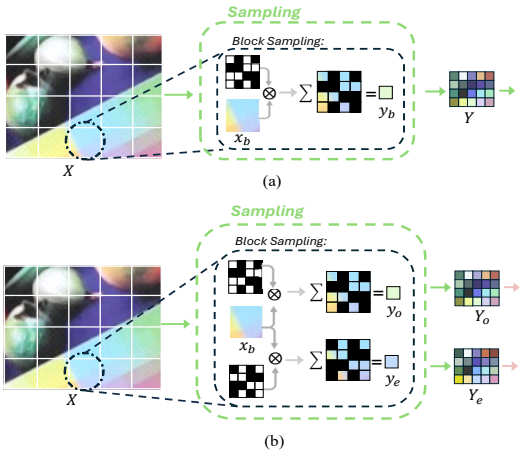


Figure 3. Sampling flow for single sample per frame (a) and two samples per frame (b). For simplicity, we show a single frame as an input and not a video sequence. Block size is also bigger for better visualization.

2.4. Video super resolution

VSR enhances video quality by reconstructing HR frames from their LR counterparts, addressing image degradation challenges. The distinction from image super-resolution lies in the utilization of inter-frame information. Most VSR networks learn residual HR feature maps which are then added to an initial reconstruction that is usually based on the bi-linear interpolation. Since adjacent frames are often correlated but misaligned, an ongoing research in VSR focuses on temporal alignment. Early methods performed the alignment in the image space [2, 27]. However, BasicVSR [4] claimed that feature alignment can achieve superior results and adopted bidirectional propagation to effectively exploit information from the entire input video. basicVSR++ [5] redesigned the architecture in [4] by proposing twice bidirectional propagation and second-order flow guided deformable convolution alignment. RVRT [30] proposed to process local neighboring frames in parallel within a globally recurrent framework, combined with guided deformable attention for clip-to-clip alignment. To enhance robustness against inaccurate optical flow, PSRT [42] proposed to align patches by averaging the motions within predefined grids. Recently, IART [49] achieved state-of-the-art results by proposing an implicit resampling based alignment, which uses a coordinate network and a window-based cross-attention.

In this work, we select representative networks to demonstrate the advantages of temporally varying sampling masks in enhancing VSR reconstruction quality.

3. Method

The pipeline of our compression and reconstruction framework is demonstrated in Fig 2. Each HR frame is sampled block-wise with non-overlapping sampling masks. The LR frames are used as input to VSR network.

3.1. Motivation

Since reconstruction from LR to HR images is common to both VSR methods and ICS frameworks, we reconsider the CS problem and examine its various components with respect to VSR. Images downscaled by a factor of 4 (compression ratio (CR) of $1/16$), commonly used as LR inputs in VSR benchmarks, can be represented as the outcome of BCS with a single sampling mask applied to a 4×4 block. The sensor pixel layout as presented in Fig. 1, defines the downscale ratio, thus we set the block size to $B = 4$, which contrasts with the larger block size of $B = 32$ typically used in CS frameworks [41].

In CS theory, the same signal is sampled multiple times. However, conventional CS approaches are impractical for image sensors due to hardware constraints, where each pixel can be accessed only once, and real-time capture requires high frame rates. To address this, we propose to sample a video sequence. In static regions, subtle frame to frame changes can be modeled as noise, which allows for CS theory to remain applicable. We can harness multi-frame processing infrastructure, already used in mobile devices for noise reduction, for HR frame reconstruction. Additionally, most CS measurements cannot be interpreted as images, restricting it from applications needing visualization or direct use. In contrast, our approach preserves the image-like nature of the measurements, allowing for direct visualization.

We propose several sampling mask layouts, drawing from three sources: existing sensor architectures, CS theory, and the T-Tetromino-based patterns introduced by [24].

3.2. Sampling

In BCS strategy [22], the input image is divided into non-overlapping blocks of size $B \times B$. Each block is sampled separately and M measurements are taken, where $M < B^2$. For an image X of size $H \times W \times 1$, this sampling operation results in measurements Y of size $\frac{H}{B} \times \frac{W}{B} \times M$. We can represent the block-sampling process with matrix multiplication: $y_b = \Phi x_b$, where Φ is the measurements matrix, $y_b \in R^M$ is the sampled block measurements, $x_b \in R^{F \cdot B^2}$ is the input video block reshaped to a vector, and F is the number of frames. The CR is therefore $\frac{M}{F \cdot B^2}$. We introduce two sampling strategies as can be seen in Fig. 3.

Single sample per frame. The basic proposed scheme is to sample each frame once, i.e. $M = F$, resulting in CR of $1/B^2$. For $B = 4$, this yields $CR = 1/16$. This setup enables direct comparison with VSR models trained using bicubic downsampling or traditional binning, where only one sample per frame is taken, maintaining the same CR of $1/16$. The compressed frame remains an image, similar to 16Sum or bicubic downsampling, but only a subset of the pixels are utilized. The measurements matrix $\Phi \in R^{M \times F \cdot B^2}$ can be written as:

$$\Phi = \begin{pmatrix} v_1 & \bar{0} & \cdots & \bar{0} \\ \bar{0} & v_2 & \cdots & \bar{0} \\ \vdots & \vdots & \cdots & \vdots \\ \bar{0} & \bar{0} & \cdots & v_F \end{pmatrix}, \quad \bar{0} \in R^{1 \times B^2}, \tag{2}$$

where $\bar{0}$ is a zero row vector, and $v_f \in R^{1 \times B^2}, 1 \leq f \leq F$ is a row vector containing the elements of a $B \times B$ mask applied on the f-th frame. Due to hardware limitations, our sampling masks must be binary. Following [38], we set each mask to be sparse by exactly 50%. For $B = 4$, precisely 8 pixels equal to 1 and the complementary 8 pixels equal to 0. This pixel sampling strategy, which utilizes only a subset of available pixels, results in a decreased fill-factor and consequently a lower SNR.

Two samples per frame. To address this limitation, inspired by the common two stage 16Sum binning with an intermediate 8Sum stage (Fig. 1), we propose the use of two orthogonal masks per frame, i.e. $M = 2F$. Each frame is sampled twice, with a mask and its complementary mask, thus each pixel is summed only once. By enabling the use of all available image data, we achieve 100% fill-factor. Since each measurement is an image, summing the two measurements, when needed for visualization, contributes to a high SNR and is equivalent to 16Sum sampling. This preserves the image structure and maintains the integrity of the imaging pipeline. The measurements matrix $\Phi \in R^{M \times F \cdot B^2}$ is of the form:

$$\Phi = \begin{pmatrix} v_1 & \bar{0} & \cdots & \bar{0} \\ 1 - v_1 & \bar{0} & \cdots & \bar{0} \\ \vdots & \vdots & \cdots & \vdots \\ \bar{0} & \bar{0} & \cdots & v_F \\ \bar{0} & \bar{0} & \cdots & 1 - v_F \end{pmatrix}. \tag{3}$$

We divide the measurements vector y_b in the following manner:

$$y_o = \Phi_o x_b, \quad y_e = \Phi_e x_b, \tag{4}$$

where $\Phi_o, \Phi_e \in R^{\frac{M}{2} \times F \cdot B^2}$ are obtained by taking only the odd and even rows of Φ respectively. The CR is now $2/B^2$, and becomes $1/8$ for $B = 4$.

3.3. Mask design

Since adjacent frames are very similar, they can be modeled as the same frame with different noise values. We therefore consider the matrix $\tilde{\Phi} \in R^{M \times B^2}$:

$$\tilde{\Phi}_1 = \begin{pmatrix} v_1 \\ \vdots \\ v_F \end{pmatrix}, \quad \tilde{\Phi}_2 = \begin{pmatrix} v_1 \\ 1 - v_1 \\ \vdots \\ v_F \\ 1 - v_F \end{pmatrix}, \tag{5}$$

where $\tilde{\Phi}_1$, $\tilde{\Phi}_2$ are used in single-sample and two-samples cases respectively. Several choices for the masks elements $v_f, 1 \leq f \leq F$ are examined below. For all the configurations we use $F = 7$ to match the number of frames in the commonly used dataset Vimeo-90K [50].

Single sample per frame (illustrated in Fig. 4):

a) *16Sum* - Standard square sensor binning, with $B = 4$. For each frame 16 pixels are summed in order to generate the measurements. Using mask representation, each mask element $v_f, 1 \leq f \leq F$ is a vector of all ones.

b) *Random* - Random matrices are widely employed in CS [3, 12, 38]. Following this principle, we consider randomly selecting 8 non-zero entries in each mask element $v_f, 1 \leq f \leq F$. We will show that this CS approach can achieve superior performance compared to the standard 16-sum binning that is currently implemented on sensors. This expectation can be further reinforced by the premise that non-regular sampling patterns can outperform regular binning [24].

Two samples per frame (illustrated in Fig. 6):

a) *8Sum* - The on-sensor 16Sum binning process involves an intermediate stage of 4×2 charge accumulation, followed by analog summation and averaging using a single ADC. The 8Sum binning is columns-based, as in Fig.1. We analyze the potential reconstruction benefits of bypassing the summation stage replacing it with two ADCs. In this case $v_f = u, 1 \leq f \leq F$, where u equals one for pixels corresponding to notation A of the block and zero otherwise. This mask remains static over time.

b) *8Sum alternate* - We further examine the reconstruction benefit obtained by alternating between columns-based and rows-based binning, harnessing another existing on-sensor binning structure of a 2×2 pixels cell that forms 50M pixels images. We utilize this existing infrastructure and suggest alternating summation of pairs of fours. The binning is columns-based for odd frames and rows-based for even frames.

c) *Tetromino basic* - Recent work [24] considers T-Tetromino pixel layouts for SISR task, where each T-shaped element aggregates four higher-resolution pixels (Fig. 5a.), based on the assumption that this non-regular sampling can surpass conventional square binning. Four measurements are taken from each 4×4 block in the HR image, and the measurements matrix $\Phi \in R^{4\times 16}$ is given by:

$$\Phi^T = \begin{pmatrix} t_1 & t_2 & t_3 & t_4 \end{pmatrix}, \tag{6}$$

where $t_i \in R^{16\times 1}$ equals one for pixels belong to the i-th T-Tetromino element and zero otherwise. We examine whether this sampling pattern can outperform 8Sum rectangular binning. To achieve a desirable downscaling factor of 4, our binning mask layout differs from [24] and requires two T-Tetromino elements. We employ combinations of elements pairs to achieve one mask. The masks elements $v_f, 1 \leq f \leq F$ are given by repetitions of the sequence $t_1 + t_2, t_1 + t_3, t_1 + t_4$, and the complementary masks are given by repetitions of $t_3 + t_4, t_2 + t_4, t_2 + t_3$.

d) *Tetromino extended* - We aim to expand the set of available masks pairs options, which is currently limited to three. To accomplish this, we horizontally flip the four T-Tetromino elements, generating four additional elements, as illustrated in Fig. 5b. The masks elements $v_f, 1 \leq f \leq F$ are given by repetitions of the sequence $t_1 + t_2, t_1 + t_3, t_1 + t_4, t_5 + t_6, t_5 + t_7, t_5 + t_8$.

e) *Random* - As in the single-sample-per-frame case, we randomly pick 8 non-zero entries in each mask element $v_f, 1 \leq f \leq F$ based on CS theory.

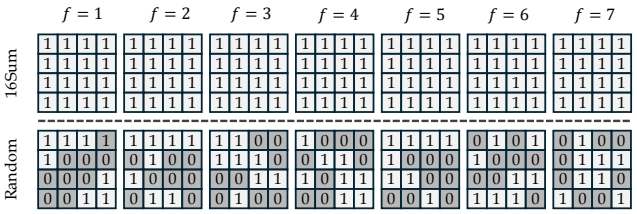


Figure 4. Single-sample-per-frame measurement masks: 16Sum (top) and Random (bottom).

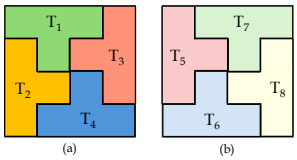


Figure 5. 4×4 T-Tetromino cell as suggested in [24] (a). To enrich the mask options, an extended version is presented in (b)

Hardware implementation. The proposed masks do not require new technologies. Instead, sensors can be designed using existing tool chains. Nevertheless, the proposed designs require hardware capable of implementing dynamic mask patterns, where the mask values change from frame to frame. Although existing compressive sensors, to the best of our knowledge, do not meet this requirement, the implementation suggested in [37] could be adapted for this purpose. The described matrix generator employs multiplexers to cyclically update mask values with each new row scan. This implementation could be adapted to update mask values every time a new frame is scanned. Random binary matrices can be implemented as shown in [28, 36]. Since our design ensures that each pixel is measured only once, a more efficient implementation, such as described in [37], can be employed. In this case, the following modifications would be required:

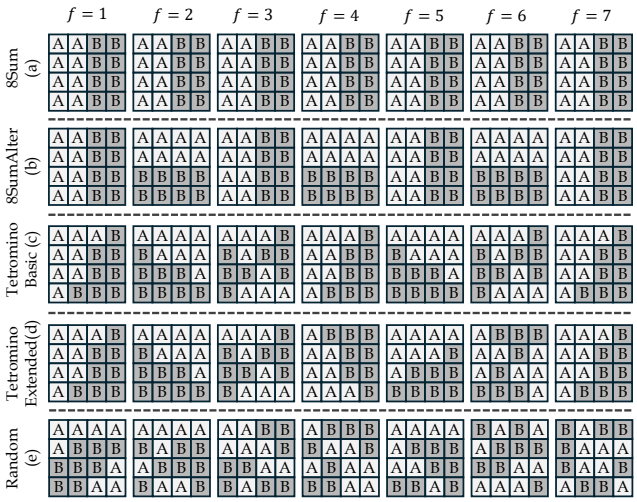


Figure 6. Two-samples-per-frame masks illustration. Starting from static masks to time-varying masks with different sampling strategies. Pixels with the value of one in the mask element $v_f, 1 \leq f \leq 7$, are marked as A, and pixels with the value of zero in v_f (and one in the complementary mask element $1 - v_f$) are marked with B.

1. Replacing the ADCs in [37] with $\Sigma\Delta$ ADCs, and summing the compressed rows along the columns in groups of 4 as in [29].
2. Updating the SC matrix multiplier to implement standard binary values instead of ± 1 values.
3. For the two-samples-per-frame mode, twice as much SC matrix multipliers and ADCs are needed.

3.4. Reconstruction

We employ VSR models for reconstruction as shown in Fig. 2. Our first hypothesis is that incorporating time modality into the down-sampling acquisition process can contribute to the VSR task. To validate this, we utilize the single-sample-per-frame configuration with 16Sum and random measurements masks (Fig. 4). These two sampling schemes can be directly compared to original VSR models trained with bicubic downsampling, requiring no modifications to the networks.

Next, we explore which sampling masks are most beneficial for VSR. Given that a 50% fill-factor is not acceptable in most imaging applications, we utilize two-samples-per-frame configurations (Fig. 6), and test their impact on reconstruction quality. Since this approach provides two LR inputs per frame, a modification to the network is required.

Network adaptation for two-samples-per-frame.
Two LR measurements are captured, and these two measurements are concatenated in the channels dimension. Instead of a 3-channels RGB image, we use a 6-channels input and change the first layer of the VSR model respectively. This design implicitly guides the model that the two measurements are drawn from the same frame. The output is also a 6-channels image and the last layer is changed accordingly. We add a pointwise convolution layer to reduce the output back to a 3-channel RGB image. The video sequence duration remains unchanged, unlike explicitly using measurement pairs, which doubles it, thereby reducing inference time. Given the benefits of channel stacking in both inference efficiency and feature utilization, we adopt this configuration as our solution. We conduct ablation studies on alternative network adaptations and present these approaches, reconstruction results, and inference times in the supplementary material. The results demonstrate that channel stacking outperforms other solutions in both reconstruction quality and time efficiency.

Initial reconstruction modification.
Most VSR architectures rely on bilinear interpolation for the initial reconstruction, with the network learning only the residual HR features to refine this estimate [4, 30, 49]. Following [23, 24], we investigate whether using an end-to-end neural network solution with our non-regular sampling offers any advantage. Specifically, we replace bilinear interpolation with the LFCR model for initial reconstruction. Modifications were required in the LFCR network to align with our compression scheme. The details of the LFCR implementation and the quantitative results are provided in the supplementary material. Our results indicate that VSR models perform effectively without this modification and that bilinear interpolation is sufficient, simplifying the overall approach. By omitting LFCR, we reduce both training time and inference complexity, making our approach more efficient.

The final reconstruction flow is illustrated in Fig. 7 using basicVSR [4] architecture.

4. Experiments

Datasets and settings. We consider the widely-used dataset for training: Vimeo-90K [50]. This benchmark has 64,612 and 7,824 video sequences for training and testing, respectively. Each sequence contains 7 frames. We evaluate our models and report results on Vimeo-90K-T and Vid4 [31] which has 4 video sequences. During training the images are randomly cropped to 256×256 image patches, which corresponds to 64×64 LR patches. Data augmentation using random horizontal and vertical flipping was applied. As in [4], the networks trained with Adam optimizer, using Charbonnier penalty loss. Initial learning rate was set to 1×10^{-4} and decreased with the Cosine Annealing scheme[33]. Models were trained for $300,000$ iterations with mini-batch size of 2. Detailed experimental settings are provided in the supplementary material.

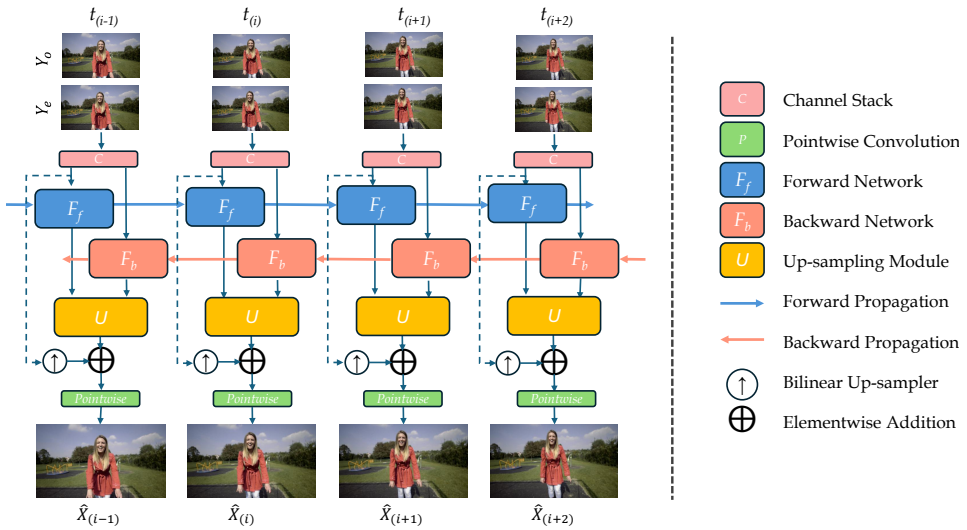


Figure 7. Proposed two-samples per frame reconstruction flow using BasicVSR architecture. Two LR sequences are used as inputs.

4.1. Quantitative and qualitative comparisons

Quantitative comparison. We performed experiments to evaluate the reconstruction quality using our proposed sampling mask designs combined with representative VSR models and report peak-signal-to-noise-ratio (PSNR), structural similarity index measure (SSIM) [47] and learned-perceptual-image-patch-similarity(LPIPS) [52] metrics.

Table 1 presents quantitative comparison of three sampling configurations in the single-sample-per-frame case combined with three representative VSR models (BasicVSR [4], IART [49], RVRT [30]). To ensure fair comparison, we trained and tested all models with the same settings and inputs. BasicVSR is a first-order VSR model, making it well-suited for applications on edge devices. Therefore, for the two-samples-per-frame case, we evaluate the reconstruction quality using our proposed mask designs in combination with BasicVSR. The results are presented in Table 2, with mask designs ordered by their temporal dynamics. As shown in Table 1, all models exhibit a consistent trend: bicubic down-sampling yields the poorest results, likely due to frequency attenuation and information loss. Comparing 16Sum with random mask sampling, our findings align with CS theory, which suggests that applying non-regular and more dynamic sampling masks to the sequence enhances performance. Superior reconstruction results were achieved using random masks, despite their 50% fill factor, across all models and metrics for both Vid4 and Vimeo-90K-T datasets. This pattern is further reinforced by the results in Table 2. The non-dynamic 8Sum masks perform worst, while highly dynamic random masks improve results but fall short of Tetromino masks. This performance gap may be due to T-Tetromino elements containing adjacent pixels, thus enhancing edge preservation. Such distinctions are expected, as VSR task differs from conventional CS problems. Furthermore, the theoretical results regarding random matrices in CS theory [12], were obtained under the assumption that $M, N \to \infty$, and therefore are only applicable for high matrix dimensions (typically $N > 1000$), while in our case $N = 16$.

Model	Params(M)	Sampling	Vimeo90K-T	Vid4
BasicVSR [4]	6.3	Bicubic	37.30/ 0.9461/ 0.1006	27.16/ 0.8195/ 0.2524
BasicVSR [4]	6.3	16Sum	37.82/ 0.9538/ 0.0799	27.88/ 0.8553/ 0.1920
BasicVSR [4]	6.3	Random	39.00/ 0.9670/ 0.0544	28.91/ 0.8905/ 0.1271
RVRT [30]	10.8	Bicubic	38.10/ 0.9524/ 0.0882	27.62/ 0.8345/ 0.2180
RVRT [30]	10.8	16Sum	39.07/ 0.9628/ 0.0591	29.30/ 0.8806/ 0.1523
RVRT [30]	10.8	Random	40.65/ 0.9754/ 0.0383	30.27/ 0.9145/ 0.0936
IART [49]	13.4	Bicubic	38.09/ 0.9524/ 0.0897	27.63/ 0.8340/ 0.2373
IART [49]	13.4	16Sum	39.37/ 0.9644/ 0.0564	29.05/ 0.8816/ 0.1570
IART [49]	13.4	Random	40.99/ 0.9768/ 0.0348	30.71/ 0.9197/ 0.0827

Table 1. Quantitative comparison (PSNR/SSIM/LPIPS) on Vimeo-90K-T [50] and Vid4 [31] for ×4 VSR in single sample-per-frame case (CR=$\frac{1}{16}$), with different sampling schemes. Best results of each model are marked in red.

To further validate our claim, we trained BasicVSR using random and Tetromino masks, which remain constant over time, for the two-samples-per-frame case. The results, presented in Table 2 indicate that such non-regular, non-dynamic sampling outperforms the rectangular, non-dynamic 8Sum sampling but falls short compared to all dynamic sampling approaches. This confirms our premise that introducing temporal dynamics to non-regular sampling enhances reconstruction performance in VSR (see supplementary material for further details).



Figure 8. Visual comparison on an example from Vimeo-90K-T using BasicVSR [4] with various two samples-per-frame sampling schemes mask layouts. The frame number is shown in the figure. Zoom in for better visualization.

Sampling	Vimeo90K-T	Vid4
8Sum	39.2199/ 0.9652/ 0.0555	28.9752/ 0.8879/ 0.1450
Tetromino constant	40.0938/ 0.9711/ 0.0398	29.9249/ 0.9162/ 0.0859
Random constant	40.2931/ 0.9724/ 0.0401	30.8335/ 0.9210/ 0.0841
8Sum alternate	40.7928/ 0.9762/ 0.038	31.2395/ 0.9290/ 0.0945
Tetromino basic	41.4386/ 0.9791/ 0.0320	31.8586/ 0.9371/ 0.0720
Tetromino extended	41.5675/ 0.9799/ 0.0312	31.8636/ 0.9375/ 0.0707
Random	41.0379/ 0.9778/ 0.0335	31.2786/ 0.9292/ 0.0741

Table 2. Quantitative comparison (PSNR/SSIM/LPIPS) on Vimeo-90K-T [50] and Vid4 [31] for ×4 VSR in two samples-per-frame case (CR=$\frac{1}{8}$), with different static (upper) and dynamic (lower) compression schemes combined with BasicVSR [4] model. Best results are marked in red.

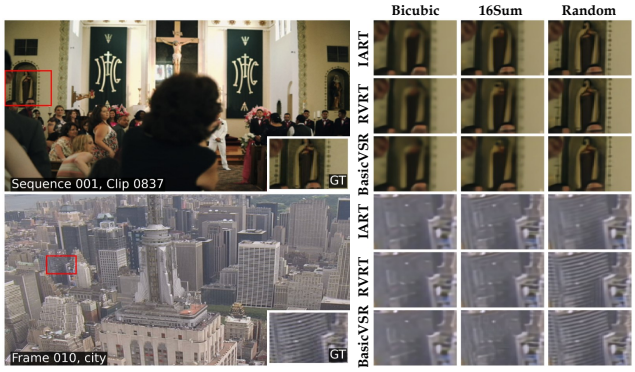


Figure 9. Visual comparison for examples from Vimeo-90K-T and Vid4 datasets on different VSR models (IART, [49] RVRT [30], BasicVSR [4]) with single sample-per-frame - bicubic, 16Sum and random mask sampling, ordered from left to right. The frame number is shown in the figure. Zoom in for better visualization.

Qualitative comparison. Fig. 9 shows qualitative comparisons of the different sampling masks for single-sample-per-frame on Vimeo-90K-T and Vid4 datasets. Fig. 8 shows qualitative comparisons for two-samples-per-frame settings on an example from Vimeo-90K-T. The results demonstrate that higher temporal variability in masks improves visual reconstruction, aligning with quantitative findings. Fig. 9 shows that VSR models trained with bicubic or 16Sum compression struggle with fine details, especially in low-motion areas or heavily blurred inputs. In contrast, despite sampling only half the pixels, when combined with the highly dynamic random masks the models can better reconstruct fine details. Fig. 8 demonstrates the impact of mask dynamics on the reconstruction quality. The 8Sum alternate masks sampling already shows a substantial improvement from the blurry 8Sum reconstruction. While the outputs of the random masks sampling and the Tetromino masks sampling appear visually similar, subtle differences can be seen in the reconstruction of fine details. Additional results are provided in the supplementary material.

5. Conclusion

This work introduces a novel video compression framework drawn from image CS theory. We demonstrate that incorporating dynamics and non-regular sampling into the compression process enhances high resolution video reconstruction in the VSR task. Additionally, we show that using our proposed sampling masks improves reconstruction and argue that this represents only a minor modification to the existing image sensor binning flow. Our approach rethinks CS theory by employing a distinctive sampling scheme that overcomes a key limitation of conventional CS frameworks. Unlike traditional methods, which acquire measurements that cannot be directly interpreted as an image, our compressed signal remains an image. Although the information used for reconstruction is doubled, the model's computational complexity does not scale proportionally, as it is applied through concatenation with only a minor increase in inference time. Future work should address model robustness to mask ordering by cyclically shifting masks and explicitly incorporating them into the reconstruction flow, as demonstrated in the supplementary material for the LFCR outputs. Additionally, optimizing mask values through integration into the training process, as done in image CS neural networks, could further enhance performance.

References

[1] Vahid Abolghasemi, Saideh Ferdowsi, and Saeid Sanei. A gradient-based alternating minimization approach for optimization of the measurement matrix in compressive sensing. *Signal Processing*, 92(4):999–1009, 2012. 2

[2] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, pages 4778–4787, 2017. 3

[3] Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE SPM*, 25(2):21–30, 2008. 2, 5

[4] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *CVPR*, pages 4947–4956, 2021. 2, 3, 6, 7, 8

[5] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, pages 5972–5981, 2022. 3

[6] Bin Chen and Jian Zhang. Content-aware scalable deep compressed sensing. *IEEE TIP*, 31:5412–5426, 2022. 2

[7] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001. 2

[8] Wei Chen, Miguel RD Rodrigues, and Ian J Wassell. Projection design for statistical compressive sensing: A tight frame based approach. *IEEE TSP*, 61(8):2016–2029, 2013. 2

[9] Wenjun Chen, Chunling Yang, and Xin Yang. Fsoinet: feature-space optimization-inspired network for image compressive sensing. In *ICASSP*, pages 2460–2464, 2022. 2

[10] Sungsoo Choi, Seungjoon Lee, Taeheon Lee, Hochul Ji, Haeyong Park, Dongmo Im, Dongchul Lee, Jinyoung Kim, Sungyong You, Jaeho Choi, et al. World smallest 200MP CMOS image sensor with 0.56 μm pixel equipped with novel deep trench isolation structure for better sensitivity and higher CG. In *IISW*, pages 22–25, 2023. 1

[11] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004. 2

[12] David L Donoho. Compressed sensing. *IEEE TIT*, 52(4): 1289–1306, 2006. 2, 5, 7

[13] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003. 2

[14] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE TIT*, 52(1):6–18, 2006.

[15] Marco F Duarte and Yonina C Eldar. Structured compressed sensing: From theory to applications. *IEEE TSP*, 59(9): 4053–4085, 2011. 2

[16] Julio Martin Duarte-Carvajalino and Guillermo Sapiro. Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE TIP*, 18 (7):1395–1408, 2009. 2

[17] Michael Elad. Optimized projections for compressed sensing. *IEEE TSP*, 55(12):5695–5702, 2007. 2

[18] Mário AT Figueiredo, Robert D Nowak, and Stephen J Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE JSTSP*, 1(4):586–597, 2007. 2

[19] Laurent A Francis and Krzysztof Iniewski. *Novel advances in microsystems technologies and their applications*. CRC Press, 2017. 1

[20] Giovanni Del Galdo, Marcus Grossmann, Florian Römer, Harald Neubauer, Michael Schöberl, and Franz Uhrmann. Apparatus and method for providing an image, 2014. Patent application EP2985992A1. 3

[21] Hongping Gan, Minghe Shen, Yi Hua, Chunyan Ma, and Tao Zhang. From patch to pixel: A transformer-based hierarchical framework for compressive image sensing. *IEEE TCI*, 9: 133–146, 2023. 2

[22] Lu Gan. Block compressed sensing of natural images. In *DSP*, pages 403–406, 2007. 2, 4

[23] Simon Grosche, Fabian Brand, and André Kaup. A novel end-to-end network for reconstruction of non-regularly sampled image data using locally fully connected layers. In *IEEE MMSP*, pages 1–6, 2021. 2, 3, 6

[24] Simon Grosche, Andy Regensky, Jürgen Seiler, and André Kaup. Image super-resolution using T-Tetromino pixels. In *CVPR*, pages 9989–9998, 2023. 2, 3, 4, 5, 6

[25] Zachary T Harmany, Roummel F Marcia, and Rebecca M Willett. This is spiral-tap: Sparse poisson intensity reconstruction algorithms—theory and practice. *IEEE TIP*, 21(3): 1084–1096, 2011. 2

[26] James Janesick. *Scientific Charge-Coupled Devices*. SPIE Press, 2001. 1

[27] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration. In *ECCV*, pages 106–122, 2018. 3

[28] Hyunkeun Lee, Donghwan Seo, Woo-Tae Kim, and Byung-Geun Lee. A compressive sensing-based CMOS image sensor with second-order $\Sigma\Delta$ ADCs. *IEEE Sensors Journal*, 18 (6):2404–2410, 2018. 3, 5

[29] Hyunkeun Lee, Woo-Tae Kim, Jinho Kim, Myonglae Chu, and Byung-Geun Lee. A compressive sensing CMOS image sensor with partition sampling technique. *IEEE Transactions on Industrial Electronics*, 68(9):8874–8884, 2021. 3, 6

[30] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *NeurIPS*, 35:378–393, 2022. 2, 3, 6, 7, 8

[31] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE TPAMI*, 36(2):346–360, 2013. 6, 7, 8

[32] Hongying Liu, Zhubo Ruan, Peng Zhao, Chao Dong, Fanhua Shang, Yuanyuan Liu, Linlin Yang, and Radu Timofte. Video super-resolution based on deep learning: a comprehensive survey. *Artificial Intelligence Review*, 55(8):5981–6035, 2022. 1

[33] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. 2016. 6

[34] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE TSP*, 41(12):3397–3415, 1993. 2

[35] Moran Mordechay and Yoav Y Schechner. Matrix optimization for poisson compressed sensing. In *IEEE GlobalSIP*, pages 684–688, 2014. 2

[36] Yusuke Oike and Abbas El Gamal. CMOS image sensor with per-column $\Sigma\Delta$ ADC and programmable compressed sensing. *IEEE Journal of Solid-State Circuits*, 48(1):318–328, 2013. 3, 5

[37] Chanmin Park, Wenda Zhao, Injun Park, Nan Sun, and Youngcheol Chae. A 51-pJ/pixel 33.7-dB PSNR 4× compressive CMOS image sensor with column-parallel single-shot compressive sensing. *IEEE Journal of Solid-State Circuits*, 56(8):2503–2515, 2021. 3, 5, 6

[38] Maxim Raginsky, Rebecca M Willett, Zachary T Harmany, and Roummel F Marcia. Compressed sensing performance bounds under poisson noise. *IEEE TSP*, 58(8):3990–4002, 2010. 2, 4, 5

[39] Michael Schöberl, Jürgen Seiler, Siegfried Foessel, and André Kaup. Increasing imaging resolution by covering your sensor. In *ICIP*, pages 1897–1900, 2011. 2

[40] Jürgen Seiler, Markus Jonscher, Thomas Ussmueller, and André Kaup. Increasing imaging resolution by non-regular sampling and joint sparse deconvolution and extrapolation. *IEEE TCSVT*, 29(2):308–322, 2018. 2

[41] Minghe Shen, Hongping Gan, Chao Ning, Yi Hua, and Tao Zhang. Transcs: a transformer-based hybrid architecture for image compressed sensing. *IEEE TIP*, 31:6991–7005, 2022. 4

[42] Shuwei Shi, Jinjin Gu, Liangbin Xie, Xintao Wang, Yujiu Yang, and Chao Dong. Rethinking alignment in video super-resolution transformers. In *NeurIPS*, 2022. 3

[43] Wuzhen Shi, Feng Jiang, Shaohui Liu, and Debin Zhao. Image compressed sensing using convolutional neural network. *IEEE TIP*, 29:375–388, 2019. 2

[44] Jiechong Song, Chong Mou, Shiqi Wang, Siwei Ma, and Jian Zhang. Optimization-inspired cross-attention transformer for compressive sensing. In *CVPR*, pages 6174–6184, 2023. 2

[45] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE TIT*, 53(12):4655–4666, 2007. 2

[46] Masayuki Uchiyama, Geunsook Park, Sangjoo Lee, Tomoyasu Tate, Masashi Minagawa, Shino Shimoyamada, Zhiqiang Lin, King W Yeung, Lien Tu, Wu-Zang Yang, et al. A 40/22 nm 200MP stacked CMOS image sensor with 0.61 μm pixel. In *IISW*, 2021. 1

[47] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 7

[48] Stephen J Wright, Robert D Nowak, and Mário AT Figueiredo. Sparse reconstruction by separable approximation. *IEEE TSP*, 57(7):2479–2493, 2009. 2

[49] Kai Xu, Ziwei Yu, Xin Wang, Michael Bi Mi, and Angela Yao. Enhancing video super-resolution via implicit resampling-based alignment. In *CVPR*, pages 2546–2555, 2024. 2, 3, 6, 7, 8

[50] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *IJCV*, 127:1106–1125, 2019. 5, 6, 7, 8

[51] Jian Zhang, Chen Zhao, and Wen Gao. Optimization-inspired compact deep compressive sensing. *IEEE JSTSP*, 14(4):765–774, 2020. 2

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7