Supplementary Material

Bowen Zhang*, Zhijin Qin†, and Geoffrey Ye Li*

* Department of Electrical and Electronic Engineering, Imperial College London, London, UK

† Department of Electronic Engineering, Tsinghua University, Beijing, China

{k.zhang21, geoffrey.li}@imperial.ac.uk, qinzhijin@tsinghua.edu.cn

I. Spatially-variant Ratios by Differentiable Functions

In this material, we design a sensing method with spatially-variant compression ratios by defining differentiable functions.

A. System overview

We show the overall pipeline of the proposed sensing system in Fig. 1. Denote H and W as the height and width of video frames. We first generate a compression ratio map, $\boldsymbol{M} \in \mathcal{R}^{H\times W\times 1}$, from a small trainable matrix. We then generate a binary mask $\boldsymbol{Z} \in \mathcal{R}^{H\times W\times (T\times 4)}$ from M. The generation of the ratio map and mask will be introduced later. At the same time, the programmable sensor will capture a scene four times using four compression ratios, i.e., $1/T$, $2/T$, $4/T$, $8/T$. The ratios are the same for different spatial locations. After that, an initial reconstruction will be conducted on these sensed data, generating one result of shape $(H\times W\times T)$ for each ratio, and all the results are concatenated along the channel dimension, generating $\boldsymbol{V}_0 \in \mathcal{R}^{H\times W\times (T\times 4)}$. The sensor model and initial generation method for each ratio has been introduced in the main text. Please refer to Sec.II. C and D for more details. Next, $\boldsymbol{Z}$ and $\boldsymbol{V_0}$ are multiplied element-wisely. In this process, only the initial reconstruction result from the ratio specified by M will be kept while others are masked by 0. The remaining results are then input to a video reconstruction network. A training loss, $\mathcal{L}_1$, based on rate-distortion theory, is introduced to train the whole network end-to-end.

B. Ratio and Mask Generation

The ratio generation network is similar to the one in the main text. The only difference is that the feature map generated by the ratio generation network is designed to have one channel and a sigmoid function is applied to the feature map to ensure each pixel in $\boldsymbol{M}$ is in [0,1].

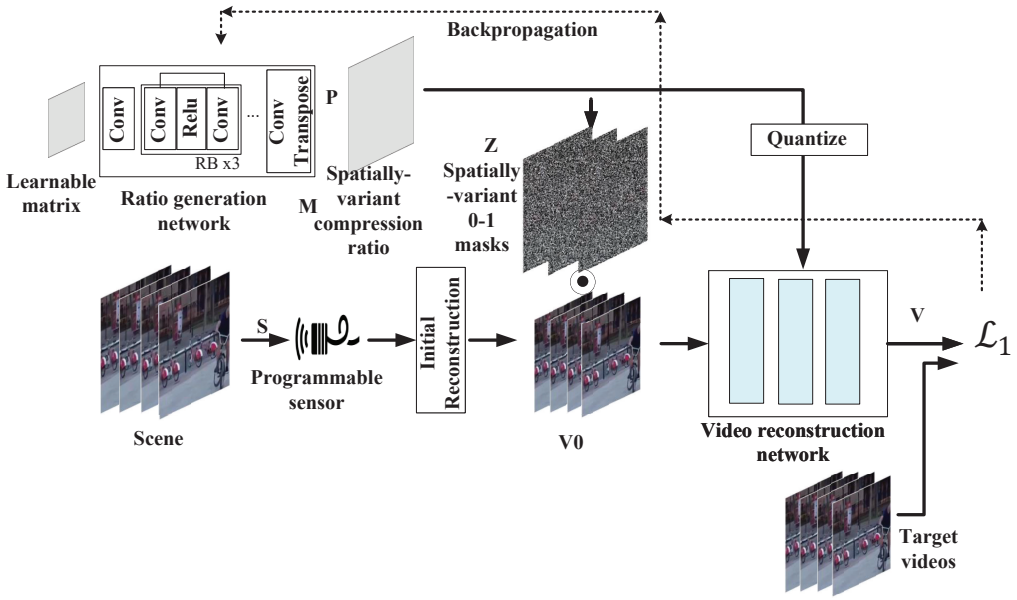

Fig. 1: The overview architecture of the proposed video compressed sensing system with spatially-variant ratios.

$\boldsymbol{Z}$ is expected to have the following property: for each spatial location (i, j) in M_{ij}, if its value is r, $\boldsymbol{Z}_{ij} \in \mathcal{R}^{4T}$ should satisfy,

$$Z_{ij} = \begin{cases} [0, \cdots, 0, 0, \cdots, 0, 0, \cdots, 0, 0, \cdots, 0], & \text{if } r \in [0, 1/5] \\ [1, \cdots, 1, 0, \cdots, 0, 0, \cdots, 0, 0, \cdots, 0], & \text{if } r \in [1/5, 2/5] \\ [0, \cdots, 0, 1, \cdots, 1, 0, \cdots, 0, 0, \cdots, 0], & \text{if } r \in [2/5, 3/5] \\ [0, \cdots, 0, 0, \cdots, 0, 1, \cdots, 1, 0, \cdots, 0], & \text{if } r \in [3/5, 4/5] \\ [0, \cdots, 0, 0, \cdots, 0, 0, \cdots, 0, 1, \cdots, 1], & \text{if } r \in [4/5, 1] \end{cases} \tag{1}$$

In this way, if $r \in [0, 1/5]$, $\boldsymbol{Z}_{ij}$ is a zero vector and will mask out all the values in $\boldsymbol{V_0}$; if $r \in [1/5, 2/5]$, $\boldsymbol{Z}_{ij}$ will only keep results from $1/T$ in $\boldsymbol{V_0}$ while masking out others, denoting the compression ratio for (i,j) is $1/T$; if $r \in [2/5, 3/5]$, $\boldsymbol{Z}_{ij}$ will only keep results from $2/T$ in $\boldsymbol{V_0}$ while masking out others; etc.

To generate Z, we first define a vector $t = [1, 1, \cdots, 1, 2, 2, \cdots, 2, 3, 3, \cdots, 3, 4, 4, \cdots, 4] \in \mathcal{R}^{4T}$, where each element repeat for T times. Next, we get $Z^{(1)} \in \mathcal{R}^{H \times W \times 4T}$, $Z^{(2)} \in \mathcal{R}^{H \times W \times 4T}$

by,

$$Z_{ijk}^{(1)} = \begin{cases} 1, & \text{if } Z_{ijk} < 5M_{ij} \\ 0, & \text{else} \end{cases} \tag{2}$$

for $i = 1, 2, \cdots H; j = 1, 2, \cdots W; k = 1, 2, \cdots, 4T$.

$$Z_{ijk}^{(2)} = \begin{cases} 1, & \text{if } Z_{ijk} > 5M_{ij} - 1 \\ 0, & \text{else} \end{cases} \tag{3}$$

for $i = 1, 2, \cdots H; j = 1, 2, \cdots W; k = 1, 2, \cdots, 4T$.

Z is then obtained by $Z = Z^1 Z^2$. Note that Eq. (2) and Eq. (3) define the forward model from M to Z. The backward gradient from Z to M is defined as 1.

C. Video reconstruction network

The video reconstruction network is nearly the same as Fig. 2 in the main text. The only difference is that, as the initial reconstruction has been conducted, initial reconstruction is no longer in the network. The fusion network will take $\boldsymbol{V_0} \in \mathcal{R}^{H \times W \times 4T}$ as input and generate $\hat{\mathcal{V}}_1 \in \mathcal{R}^{H \times W \times T}$.

D. Training losses

The training loss, $\mathcal{L}_1$, is still defined based on the rate-distortion theory

$$\mathcal{L}_1 = ||V - \hat{V}||^2 + \lambda \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} M_{ij}, \tag{4}$$

where $||V - \hat{V}||^2$ denotes the distortion of reconstructed video, $\frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} M_{ij}$ represents the average value of M and minimizing this value will reduce the average compression ratio. λ is the introduced trade-off parameter between distortion and compression ratio.

Note that we did not scale M in Eq. (4) according to the real number of measurements for simplicity. Scaling here means, if $M_{i,j} \in [1/5, 2/5]$, $1/T$ ratio is implemented for (i, j), resulting in one measurement; thus we should multiple $M_{i,j}$ in Eq. (4) by 1; if $M_{i,j} \in [3/5, 4/5]$, $4/T$ ratio is implemented for (i, j), resulting in four measurements; thus we should multiple $M_{i,j}$ in Eq. (4) by 4. We find the experimental results are still good without scaling, and this is because for most locations, $M_{i,j}$ are chosen from [0,3/5]. In this range, the number of measurements increases linearly, i.e., 0 for $M_{ij} \in [0, 1/5]$, 1 for $M_{ij} \in [1/5, 2/5]$, and 2 for $M_{ij} \in [2/5, 3/5]$.