Video Rescaling Networks with Joint Optimization Strategies
for Downscaling and Upscaling

Yan-Cheng Huang¹,*  Yi-Hsin Chen¹,*  Cheng-You Lu¹
Hui-Po Wang²  Wen-Hsiao Peng¹  Ching-Chun Huang¹
¹ National Yang Ming Chiao Tung University, Taiwan
² CISPA Helmholtz Center for Information Security

{so756722.iie07g, yhchen.iie07g, johnny305.cs04}@nctu.edu.tw
hui.wang@cispa.saarland, {wpeng, chingchun}@cs.nctu.edu.tw

Abstract

This paper addresses the video rescaling task, which arises from the needs of adapting the video spatial resolution to suit individual viewing devices. We aim to jointly optimize video downscaling and upscaling as a combined task. Most recent studies focus on image-based solutions, which do not consider temporal information. We present two joint optimization approaches based on invertible neural networks with coupling layers. Our Long Short-Term Memory Video Rescaling Network (LSTM-VRN) leverages temporal information in the low-resolution video to form an explicit prediction of the missing high-frequency information for upscaling. Our Multi-input Multi-output Video Rescaling Network (MIMO-VRN) proposes a new strategy for downscaling and upscaling a group of video frames simultaneously. Not only do they outperform the image-based invertible model in terms of quantitative and qualitative results, but also show much improved upscaling quality than the video rescaling methods without joint optimization. To our best knowledge, this work is the first attempt at the joint optimization of video downscaling and upscaling.

1. Introduction

With the increasing popularity of video capturing devices, a tremendous amount of high-resolution (HR) videos are shot every day. These HR videos are often downscaled to save storage space and streaming bandwidth, or to fit screens with lower resolutions. It is also common that the downscaled videos need to be upscaled for display on HR monitors [11, 16, 22, 2, 29].

In this paper, we address the joint optimization of video downscaling and upscaling as a combined task, which is referred to as video rescaling. This task involves downscaling an HR video into a low-resolution (LR) one, followed by upscaling the resulting LR video back to HR. Our aim is to optimize the HR reconstruction quality while regularizing the LR video to offer comparable visual quality to the bicubic-downscaled video for human perception. It is to be noted that the rescaling task differs from the super-resolution task; at inference time, the former has access to the HR video while the latter has no such information.

One straightforward solution to video rescaling is to downscale an HR video by predefined kernels and upscale the LR video with super-resolution methods [14, 17, 32, 6, 1, 23, 20, 10, 25, 31, 8, 15, 9]. With this solution,
The downscaling is operated independently of the upscaling although the upscaling can be optimized for the chosen downscaling kernels. The commonly used downscaling (e.g. bicubic) kernels suffer from losing the high-frequency information [21] inherent in the HR video, thus creating a many-to-one mapping between the HR and LR videos. Reconstructing the HR video by upscaling its LR representation becomes an ill-posed problem. The independently-operated downscaling misses the opportunity of optimizing the downscaled video to mitigate the ill-posedness.

The idea of jointly optimizing downscaling and upscaling was first proposed for image rescaling [11, 16, 22, 2]. It adds a new dimension of thinking to the studies of learning specifically to upscale for a given downscaling method [14, 17, 32, 26, 3, 6]. Recognizing the reciprocality of the downscaling and upscaling operations, IRN [29] recently introduced a coupling layer-based invertible model, which shows much improved HR reconstruction quality than the non-invertible models.

These jointly optimized image-based solutions (Fig. 1(a)) are not ideal for video rescaling. For example, a large number of prior works [1, 23, 20, 10, 25, 31, 8, 15, 9] for video upscaling have adopted the Multi-Input Single-Output (MISO) strategy to reconstruct one HR frame from multiple LR frames and/or previously reconstructed HR frames (Fig. 1(b)). They demonstrate the potential for recovering the missing high-frequency component of a video frame from temporal information. However, image-based solutions do not consider temporal information. In addition, two issues remain widely open as (1) how video downscaling and upscaling could be jointly optimized and (2) how temporal information could be utilized in the joint optimization framework to benefit both operations.

In this paper, we present two joint optimization approaches to video rescaling: Long Short-Term Memory Video Rescaling Network (LSTM-VRN) and Multi-Input Multi-Output Video Rescaling Network (MIMO-VRN). The LSTM-VRN downscals an HR video frame-by-frame using a similar coupling architecture to [29], but fuses multiple downscaled LR video frames via LSTM to estimate the missing high-frequency component of an LR video frame for upscaling (Fig. 1(b)). LSTM-VRN shares similar downscaling and upscaling strategies to the traditional video rescaling framework. In contrast, our MIMO-VRN introduces a completely new paradigm by adopting the MIMO strategy for both video downscaling and upscaling (Fig. 1(c)). We develop a group-of-frames-based (GoF) coupling architecture that downscals multiple HR video frames simultaneously, with their high-frequency components being estimated also simultaneously in the upscaling process. Our contributions include the following:

- To the best of our knowledge, this work is the first attempt at jointly optimizing video downscaling and upscaling with invertible coupling architectures.
- Our LSTM-VRN and MIMO-VRN outperform the image-based invertible model [29], showing significantly improved HR reconstruction quality and offering LR videos comparable to the bicubic-downscaled video in terms of visual quality.
- Our MIMO-VRN is the first scheme to introduce the MIMO strategy for video upscaling and downscaling, achieving the state-of-the-art performance.

2. Related Work

This section surveys video rescaling methods, with a particular focus on their downscaling and upscaling strategies. We regard the image-based rescaling methods as possible solutions for video rescaling. Fig. 2 is a taxonomy of these prior works.

2.1. Upscaling with Predefined Downscaling

The traditional image super-resolution [14, 17, 32, 26, 3, 6] or video super-resolution [1, 23, 20, 10, 25, 31, 8, 15, 9] methods are candidate solutions to video upscaling. The former is naturally a single-input single-output (SISO) upscaling strategy, which generates one HR image from one LR image. The latter usually involves more than one LR video frame in the upscaling process, i.e. the MISO upscaling strategy, in order to leverage temporal information for better HR reconstruction quality. Most of the approaches in this category adopt a SISO downscaling strategy with a pre-defined kernel (e.g. bicubic) chosen independently of the upscaling process. Therefore, they are unable to adapt the downscaled images/videos to the upscaling.

2.2. Upscaling with Jointly Learned Downscaling

To mitigate the ill-posedness of the image upscaling task, some works learn upscaling and downscaling jointly by
encoder-decoder architectures [11, 16, 22, 2]. They turn the fixed downscaling method into a learnable model in order to adapt the LR image to the upscaling process that is learned jointly. The training objective usually requires the LR image to be also suitable for human perception. Recently, IRN [29] introduces an invertible model [4, 5, 13] to this joint optimization task. It is able to perform image downscaling and upscaling by the same set of neural networks configured in the reciprocal manner. It provides a means to model explicitly the missing high-frequency information due to downscaling by a Gaussian noise.

### 2.3. Invertible Rescaling Network

IRN [29] is an invertible model designed specifically for image rescaling. The forward model of IRN comprises a 2-D Haar transform and eight coupling layers, as shown in Fig. 3. By applying the 2-D Haar transform, an input image \( x \in \mathbb{R}^{C \times H \times W} \) is first decomposed into one low-frequency band \( y' \in \mathbb{R}^{C \times \frac{H}{2} \times W} \) and three other high-frequency bands \( z' \in \mathbb{R}^{3C \times \frac{H}{2} \times W} \). These two components \( y', z' \) are subsequently processed via the coupling layers in a way that the output \( y \) becomes a visually-pleasing LR image and the \( z \) encodes the complementary high-frequency information inherent in the input HR image \( x \). In theory, the inverse coupling layers can recover \( x \) losslessly from \( y \) and \( z \) because the model is invertible. In practice, \( z \) is unavailable for upscaling at inference time. The training of IRN requires \( z \) to follow a Gaussian distribution so that at inference time, a Gaussian sample \( \tilde{z} \) can be drawn as a substitute for the missing high-frequency component.

Although IRN achieves superior results on the image rescaling task, it is not optimal for video rescaling. Essentially, IRN is an image-based method. This work presents the first attempt at jointly optimizing video downscaling and upscaling with an invertible coupling architecture (Fig. 3).

### 3. Proposed Method

Given an HR video composed of \( N \) video frames \( \{x_t\}_{t=1}^{N} \), where \( x_t \in \mathbb{R}^{C \times H \times W} \), the video rescaling task involves (1) downscaling every video frame \( x_t \) to its LR counterpart \( y_t \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}} \), where the quantized version \( \hat{y}_t \) of which forms collectively an LR video \( \{\hat{y}_t\}_{t=1}^{N} \) and (2) upscaling the LR video to arrive at the reconstructed HR video \( \{\hat{x}_t\}_{t=1}^{N} \). Unlike most video super-resolution tasks, which focus primarily on learning upscaling for a given downscaling method, this work optimizes jointly the downscaling and upscaling as a combined task. It has been shown in many traditional video super-resolution works [1, 23, 20, 7, 24, 25, 10, 31, 9, 15, 8] that the extra temporal information in videos allows the lost high-frequency component of a video frame due to downscaling to be recovered to some extent. This work makes the first attempt to explore how such temporal information could assist downscaling in producing an LR video that can be up-scaled to offer better super-resolution quality in an end-to-end fashion. In a sense, our focus is on both downscaling and upscaling. The objective is to minimize the distortion between \( \{\hat{x}_t\}_{t=1}^{N} \) and \( \{x_t\}_{t=1}^{N} \) in such a combined task while the LR video \( \{\hat{y}_t\}_{t=1}^{N} \) is regularized to offer comparable visual quality to the bicubic-downscaled video for human perception. It is to be noted that the LR video is not meant to be exactly the same as the bicubic-downscaled video since doing so may not lead to the optimal downscaling and upscaling in our task.

The reciprocality of the downscaling and upscaling operations motivates us to choose an invertible network for our task. With the superior performance of coupling layer architectures in recovering high-frequency details of LR images [29], we develop our downscaling and upscaling networks, especially for video, using a similar invertible architecture (Sec. 2.3) as the basic building block. We propose two approaches, LSTM-VRN and MIMO-VRN, to configure or extend these building blocks for joint learning of video downscaling and upscaling. Their overall architectures are depicted in Fig. 4, with detailed operations given in the following sections.

### 3.1. LSTM-based Video Rescaling Network

Like most video super-resolution techniques, the LSTM-VRN (Fig. 4(a)) adopts the SISO strategy to downscale HR video frames \( \{x_t\}_{t=1}^{N} \) individually to their LR ones \( \{\hat{y}_t\}_{t=1}^{N} \) by the forward model of the invertible network. The operation is followed by the MISO-based upscaling, which departs from the idea of drawing an input-agnostic Gaussian noise [29] for complementary high-frequency information. Specifically, we fuse the current LR frame \( \hat{y}_t \) and its neighbouring frames \( \{\hat{y}_{t-i}, \hat{y}_{t+i}\}_{i=1}^{C} \) by a LSTM-based predictive module to form an estimate \( \tilde{z}_t \) of the missing high-frequency component \( z_t \) at inference time. The resulting
\[ \hat{z}_t \text{ is fed to the inverse model together with the } \hat{y}_t \text{ for reconstructing the HR video frame } \hat{x}_t. \text{ The fact that } z_t \text{ needs to be estimated from multiple LR frames } \{\hat{y}_i\}_{i=t-L}^{t+L} \text{ determines what information should remain in the LR video to facilitate the prediction. This connects the upsampling process tightly to the downscaling process, stressing the importance of their joint optimization. In addition, we rely on the inter-branch pathways of the coupling layer in the forward model to correlate } z_t \text{ and } y_t \text{ in such a way that } z_t \text{ could be better predicted from } y_t \text{ and its neighbors } \{\hat{y}_{t-i}, \hat{y}_{t+i}\}_{i=1}^L. \]

The predictive module plays a key role in fusing information from \( \hat{y}_t \) and \( \{\hat{y}_{t-i}, \hat{y}_{t+i}\}_{i=1}^L \). We incorporate Spatiotemporal-LSTM (ST-LSTM) [27] for propagating temporal information in both forward and backward directions, in view of its recent success in video extrapolation tasks. Eq. (1) details the forward mode of the predictive module for time instance \( t \):

\[ h_t^L = ST-LSTM(f_{t-1}, h_{t-1}^L) \]
\[ h_t^U = ResidualBlock(\hat{y}_t) \]
\[ a_t = \sigma(W \odot h_t^U) \]
\[ f_t = (1 - a_t) \odot h_t^L + a_t \odot h_t^U \]

where \( \sigma \) is a sigmoid function, \( \odot \) is the standard convolution, and \( \odot \) is Hadamard product. Note that an attention signal \( a_t \) guided by the current LR frame \( \hat{y}_t \) combines the temporally-propagated hidden information \( h_t^L \) and the features \( h_t^U \) of \( \hat{y}_t \) to yield the output \( f_t \). As Fig. 4(a) shows, the forward propagated \( f_t \) is further combined with the backward propagated \( b_t \) to predict \( \hat{z}_t \) through a 1x1 convolution.

For upsampling every LR video frame \( \hat{y}_t \), the proposed predictive module works in a sliding-window manner with a window size of \( 2L + 1 \). That is, the forward (respectively, backward) ST-LSTM always starts with a reset state \( 0 \) when accepting the input \( \hat{y}_{t-L} \) (respectively, \( \hat{y}_{t+L} \)). This design choice is out of generalization and buffering considerations. We avoid running a long ST-LSTM at inference time because the training videos are rather short. Moreover, the backward ST-LSTM introduces delay and buffering requirements.

Finally, we note in passing that LSTM-VRN exploits temporal information across LR video frames only for up-sampling while its down-sampling is still a SISO-based scheme, which does not take advantage of temporal information in HR video frames for down-sampling.

### 3.2. MIMO-based Video Rescaling Network

Our MIMO-VRN (Fig. 4(b)) is a new attempt that adopts a MIMO strategy for both upscaling and downscaling, making explicit use of temporal information in these operations. Here, we propose a new basic processing unit, called Group of Frames (GoF). To begin with, the HR input video \( \{x_t\}_{t=1}^N \) is decomposed into non-overlapping groups of frames, with each group including \( g \) frames, namely \( \{x_t\}_{t=1}^g, \{x_t\}_{t=g+1}^{2g}, \ldots \). The downscaling pro-
ceeds on a group-by-group basis; each GoF is downscaled independently of each other. Within a GoF, every HR video frame \( x_t \in \mathbb{R}^{C \times H \times W} \) is first transformed individually using 2-D Haar Wavelet, to arrive at its low-frequency \( y_t' \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}} \) and high-frequency \( z_t' \in \mathbb{R}^{3C \times \frac{H}{2} \times \frac{W}{2}} \) components. We then group the low-frequency components \( \{y_t'\}_{i=1}^{t+g} \) in a GoF as one group-type input \( Y_t \in \mathbb{R}^{gC \times \frac{H}{2} \times \frac{W}{2}} \) to the coupling layers (i.e. replacing \( y_t' \) in Fig. 3 with \( Y_t \)) and the remaining high-frequency components \( \{z_t'\}_{i=1}^{t+g} \) as the other group-type input \( Z_t \in \mathbb{R}^{3gC \times \frac{H}{2} \times \frac{W}{2}} \) (i.e. replacing \( z_t' \) in Fig. 3 with \( Z_t \)). Because each group-type input contains information from multiple video frames, the coupling layers are able to utilize temporal information inherent in one group-type input to update the other. With two downscaling modules, the results are a group of quantized LR \( \hat{Y}_t \) and the group high-frequency component \( \hat{Z}_t \). It is worth noting that due to the nature of group-based coupling, there is no one-to-one correspondence between the signals in \( \hat{Y}_t \in \mathbb{R}^{gC \times \frac{H}{2} \times \frac{W}{2}} \) and \( \hat{Z}_t \in \mathbb{R}^{3gC \times \frac{H}{2} \times \frac{W}{2}} \).

The upscaling proceeds also on a group-by-group basis, with the group size \( g \) and the group formation fully aligned with those used for downscaling. As depicted in Fig. 4(b), we employ a residual block-based predictive module to form a prediction of the missing high-frequency components \( \{\hat{z}_i\}_{i=1}^{t+g} \) from the corresponding group of LR frames \( \{\hat{y}_i\}_{i=1}^{t+g} \). Similar to the notion of the group-type inputs for downscaling, the LR frames \( \{\hat{y}_i\}_{i=1}^{t+g} \) and the estimated high-frequency components \( \{\hat{z}_i\}_{i=1}^{t+g} \) comprise respectively the two group-type inputs \( \hat{Y}_t \) and \( \hat{Z}_t \) to the invertible network operated in inverse mode. With this MIMO-based upscaling, a group of HR frames \( \{\hat{x}_i\}_{i=1}^{t+g} \) are reconstructed simultaneously.

3.3. Training Objectives

LSTM-VRN. The training of LSTM-VRN involves two loss functions to reflect our objectives. First, to ensure that the LR video \( \{\hat{y}_i\}_{i=1}^{N} \) is visually pleasing, we follow common practice to require that \( \{\hat{y}_i\}_{i=1}^{N} \) have similar visual quality to the bicubic-downscaled video \( \{x^{bic}_i\}_{i=1}^{N} \); to this end, we define the LR loss as

\[
\mathcal{L}_{LR} = \frac{1}{N} \sum_{t=1}^{N} \|x^{bic}_t - \hat{y}_t\|^2. \tag{2}
\]

Second, to maximize the HR reconstruction quality, we minimize the Charbonnier loss \([14]\) between the original HR video \( \{x_t\}_{t=1}^{N} \) and its reconstructed version \( \{\hat{x}_t\}_{t=1}^{N} \) subject to downscaling and upscaling:

\[
\mathcal{L}_{HR} = \frac{1}{N} \sum_{t=1}^{N} \sqrt{\|x_t - \hat{x}_t\|^2 + \epsilon^2}, \tag{3}
\]

where \( \epsilon \) is set to \( 1 \times 10^{-3} \). The total loss is \( \mathcal{L}_{total} = \mathcal{L}_{HR} + \lambda \mathcal{L}_{LR} \), where \( \lambda \) is a hyper-parameter used to trade-off between the quality of the LR and HR videos.

MIMO-VRN. The training of MIMO-VRN shares the same \( \mathcal{L}_{LR} \) and \( \mathcal{L}_{HR} \) losses as LSTM-VRN because they have common optimization objectives. We however notice that MIMO-VRN tends to have uneven HR reconstruction quality over video frames in a GoF (Sec. 4.4). To mitigate the quality fluctuation in a GoF, we additionally introduce the following center loss for MIMO-VRN:

\[
\mathcal{L}_{center} = \frac{1}{M \times g} \sum_{m=1}^{M} \sum_{t=(m-1)g+1}^{mg} \|x_t - \hat{x}_t\|^2 - c_m, \tag{4}
\]

where \( g \) is the group size, \( c_m = \sum_{t=(m-1)g+1}^{mg} \|x_t - \hat{x}_t\|^2 / g \) denotes the average HR reconstruction error in a GoF, and \( M \) is the number of GoF’s in a sequence. Eq. (4) encourages the HR reconstruction error of every video frame in a GoF to approximate the average level \( c_m \).

4. Experimental Results

4.1. Setup

Datasets. For a fair comparison, we follow the common test protocol to train our models on Vimeo-90K dataset [30]. It has 91,701 video sequences, each is 7 frames long. Among them, 64,612 sequences are for training and 7,824 are for test. Each sequence has a fixed spatial resolution of 448 \( \times \) 256. The performance evaluation is done on two standard test datasets, Vimeo-90K-T and Vid4 [18]. Vid4 includes 4 video clips, each having around 40 frames.

Implementation and Training Details. Our proposed models adopt the settings from IRN [29], which consists of two downscaling modules (Fig. 3). Each module is composed of one 2-D Haar transform and eight coupling layers. Both LSTM-VRN and MIMO-VRN have eight predictive modules (Fig. 4) replicated and stacked for a better prediction of the missing high-frequency component. The sliding window size for LSTM-VRN is set to 7, which includes the current LR video frame together with 6 neighbouring LR frames (3 from the past and 3 from the future). The GoF size \( g \) for MIMO-VRN is set to 5. For data augmentation, we randomly crop training videos to 144 \( \times \) 144 as HR inputs and use their bicubic-downscaled versions (of size 36 \( \times \) 36) as LR ground-truths. We also apply random horizontal and vertical flipping. LSTM-VRN and MIMO-VRN share the same LR and HR training objectives (Eq. (2) and Eq. (3)), with the \( \lambda \) for \( \mathcal{L}_{LR} \) set to 64. The training of MIMO-VRN additionally includes the center loss (Eq. (4)), the hyper-parameter of which is chosen to be 16. We use Adam optimizer [12], with \( \beta_1 = 0.9 \) and \( \beta_2 = 0.5 \) and a batch size of 16. The weight decay is set to \( 1 \times 10^{-12} \). We use an
Table 1. PSNR-Y / SSIM-Y comparison on Vid4 for $\times 4$ upscaling. ‘†’ represents the model adopting the joint optimization for downscaling and upscaling. Red, green, and blue indicate the best, the second best, and the third best performance, respectively.

| Downscale | Upscale | Method                  | Calendar | City | Foliage | Walk | Average |
|-----------|---------|-------------------------|----------|------|---------|------|---------|
| SISO      | SISO    | DRN-L [6]               | 22.47 / 0.7289 | 26.25 / 0.7011 | 24.88 / 0.6681 | 28.84 / 0.8752 | 25.61 / 0.7433 |
|           | SISO    | CAR† [22]               | 24.48 / 0.8143 | 30.19 / 0.8444 | 26.98 / 0.7841 | 31.59 / 0.9250 | 28.28 / 0.8421 |
|           | SISO    | IRN† [29]               | 26.62 / 0.8850 | 33.48 / 0.9337 | 29.71 / 0.8871 | 35.36 / 0.9696 | 31.29 / 0.9188 |
| MISO      | MISO    | DUF [10]                | 24.04 / 0.8110 | 28.27 / 0.8313 | 26.41 / 0.7709 | 30.60 / 0.9141 | 27.33 / 0.8318 |
|           | MISO    | EDVR-L [25]             | 24.05 / 0.8147 | 28.00 / 0.8122 | 26.34 / 0.7635 | 31.02 / 0.9152 | 27.35 / 0.8264 |
|           | MISO    | PFNL [31]               | 24.37 / 0.8246 | 28.09 / 0.8385 | 26.51 / 0.7768 | 30.65 / 0.9135 | 27.40 / 0.8384 |
|           | MISO    | TGA [9]                 | 24.47 / 0.8286 | 28.37 / 0.8419 | 26.59 / 0.7793 | 30.96 / 0.9181 | 27.59 / 0.8419 |
|           | MISO    | RSDN [8]                | 24.60 / 0.8355 | 29.20 / 0.8527 | 26.84 / 0.7931 | 31.04 / 0.9210 | 27.92 / 0.8505 |
|           | MISO    | LSTM-VRN†               | 27.31 / 0.9039 | 34.36 / 0.9482 | 31.13 / 0.9213 | 36.18 / 0.9742 | 32.24 / 0.9369 |
| MIMO      | MIMO    | MIMO-VRN†               | 29.23 / 0.9389 | 35.49 / 0.9573 | 33.25 / 0.9535 | 37.17 / 0.9812 | 33.79 / 0.9577 |
|           | MIMO    | MIMO-VRN-C†             | 28.83 / 0.9322 | 35.13 / 0.9544 | 32.72 / 0.9476 | 36.93 / 0.9808 | 33.40 / 0.9537 |

Table 2. PSNR-Y / SSIM-Y comparison on Vimeo-90K-T for $\times 4$ upscaling. ‘†’ represents the model adopting the joint optimization for downscaling. Red, green, and blue indicate the best, the second best, and the third best performance, respectively.

| Downscale | Upscale | Method                  | Average |
|-----------|---------|-------------------------|---------|
| SISO      | SISO    | DRN-L [6]               | 35.63 / 0.9262 |
|           | SISO    | CAR† [22]               | 37.69 / 0.9493 |
|           | SISO    | IRN† [29]               | 40.83 / 0.9734 |
| MISO      | MISO    | DUF [10]                | 36.37 / 0.9387 |
|           | MISO    | EDVR-L [25]             | 37.63 / 0.9487 |
|           | MISO    | TGA [9]                 | 37.59 / 0.9516 |
|           | MISO    | RSDN [8]                | 37.23 / 0.9471 |
|           | MISO    | LSTM-VRN†               | 41.42 / 0.9764 |
| MIMO      | MIMO    | MIMO-VRN†               | 43.26 / 0.9846 |
|           | MIMO    | MIMO-VRN-C†             | 42.53 / 0.9820 |

initial learning rate of $1 \times 10^{-4}$, which is decreased by half for every 30k iterations. Our code is available online.  

Baselines. We include three categories of baselines for comparison: (1) SISO-down-SISO-up with predefined downsampling kernels (e.g. DRN-L [6]), (2) SISO-down-SISO-up with jointly optimized downsampling and upscaling (e.g. CAR [22] and IRN [29]), and (3) SISO-down-MISO-up with predefined downsampling kernels (e.g. DUF [10], EDVR-L [25], PFNL [31], TGA [9], and RSDN [8]). The first two categories perform video downsampling and upsampling on a frame-by-frame basis. The third category includes video super-resolution methods, where the predefined downsampling is done frame-by-frame and the learned upsampling is MISO-based. The predefined downsampling uses the bicubic interpolation method. It is to be noted that the methods adopting the learned downsampling perform upsampling based on their respective LR videos, which would not be the same as the bicubic-downscaled videos. The results for the methods in categories (1) and (2) are produced using the pre-trained models released by the authors. Those in category (3) are taken from the papers since these baselines share exactly the same setting as ours. We report results for a downsampling/upsampling factor of 4 only, following the common setting for video rescaling.

Metrics. For quantitative comparison, we adopt the standard test protocol in the super-resolution tasks to evaluate Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) on the Y channel, denoted respectively PSNR-Y and SSIM-Y.

4.2. Comparison of Quantitative Results

Tables 1 and 2 report the PSNR-Y and SSIM-Y results of the reconstructed HR videos on Vid4 and Vimeo-90K-T. Table 3 summarizes the results for the downsampled videos. The following observations are immediate:

(1) Optimizing jointly video downsampling and upscaling improves the HR reconstruction quality. This is confirmed by the fact that LSTM-VRN achieves considerably higher PSNR-Y (32.24dB on Vid4 and 41.42dB on Vimeo-90K-T) than the baselines with video super-resolution methods for upsampling (27.33-27.92dB on Vid4 and 36.37-37.59dB on Vimeo-90K-T) [25, 10, 31, 9, 8], which adopt the same SISO-down-MISO-up strategy yet with a predefined downsampling kernel. We note that the image-based joint optimization schemes, e.g. IRN [29] and CAR [22], achieve better HR reconstruction quality than the traditional video-based baselines, even without using temporal information for upsampling. The superior performance of joint optimization schemes is attributed to the fact that they can better embed HR information in LR frames for upsampling.

(2) Incorporating temporal information in the LR video improves further on the HR reconstruction quality. The result is evidenced by the 0.95dB and 0.59dB PSNR-Y gains of LSTM-VRN over IRN [29] on Vid4 and Vimeo-90K-T. Both share a similar invertible network for downsampling, but our LSTM-VRN additionally leverages information from multiple LR video frames to predict the high-
frequency component of a video frame during upscaling.

(3) **MIMO-VRN achieves the best PSNR-Y/SSIM-Y results.** It outperforms LSTM-VRN by 1.55dB and 1.84dB in PSNR-Y on Vid4 and Vimeo-90K-T, respectively, while LSTM-VRN already shows a significant improvement over the other baselines. The inclusion of the center loss (see MIMO-VRN-C) causes a modest decrease in PSNR-Y/SSIM-Y but helps to alleviate the quality fluctuation in both the resulting LR and HR videos (Sec. 4.4). These results highlight the benefits of incorporating temporal information into both downscaling and upscaling in an end-to-end optimized manner.

(4) **Both LSTM-VRN and MIMO-VRN produce visually-pleasing LR videos.** Table 3 shows that the LR videos produced by our models have a PSNR-Y of more than 40dB when compared against the bicubic-downscaled videos. This together with the SSIM-Y results suggests that they are visually comparable to the bicubic-downscaled videos, as is also confirmed by the subjective quality comparison in Fig. 5 and the supplementary document.

### 4.3. Comparison of Qualitative Results

Figs. 6 presents a qualitative comparison on Vid4. As shown, our models produce higher-quality HR video frames with much sharper edges and finer details. The other methods show blurry image quality and fail to recover image details. From Fig. 5, our downscaling models produce visually comparable results to the bicubic downscaling method, which indicates the visually-pleasing property of our LR videos. The reader is referred to our project page[^1] for more results.

### 4.4. Ablation Experiments

**Temporal Propagation Methods in LSTM-VRN.** Table 4 presents results for three temporal propagation schemes in LSTM-VRN. The first runs LSTM in forward direction without reset. The second and the third implement the proposed method with uni- or bi-directional propagation, respectively. We see that the sliding window-based reset is advantageous to the HR reconstruction quality. This may be attributed to the fact that the training videos in Vimeo-90K are rather short. When trained on Vimeo-90K, the first variant may not generalize well to unseen long videos in Vid4. As expected, with the access to both the past and future LR frames, the bi-directional propagation performs better than the uni-directional one (i.e. Fig. 4(a) without the backward path).

**GoF Size.** Table 5 studies the effect of the GoF size on MIMO-VRN’s performance. The setting GoF1 reduces to the SISO-up-SISO-down method, which is similar to IRN [29] except that it introduces a prediction of the high-frequency component from the LR video frame. For a fair comparison, we re-train IRN [29] on Vimeo-90K and denote the re-trained model by IRN_Ret. Note that the pre-trained IRN [29] performs better than IRN_Ret since it is trained on a different (image-based) dataset. We see that GoF1 and IRN_Ret show comparable performance, especially on the HR videos. This suggests that without additional temporal information, the prediction of the high-frequency component from the LR video is ineffective. However, increasing the GoF size, which involves more temporal information in downscaling and upscaling, improves the quality of the HR video significantly. GoF5 is seen to be the best setting.

---

[^1]: Refer to our project page for more results.
Figure 6. Qualitative comparison on Vid4 for 4× upscaling. Zoom in for better visualization.

Figure 7. The impact of the center loss on the quality of the HR and LR videos. The per-frame PSNR-Y is visualized as a function of frame indices. The GoF size is 5. MIMO-VRN: no center loss. HR: the center loss imposed on the HR video only. LR: the center loss imposed on the LR video only. HR&LR: the center loss imposed on both the HR and LR videos.

Table 6. PSNR-Y of MIMO-VRN with and without the center loss on Vid4. The mean absolute deviation (MAD) indicates the average absolute deviation of the per-frame PSNR-Y from the GoF mean.

| Center loss | PSNR-Y | MAD |
|-------------|--------|-----|
| HR LR       | HR LR  |     |
| √           | 33.79  | 45.05 | 0.88  | 0.55  |
| √           | 33.40  | 45.54 | 0.28  | 0.63  |
| √           | 33.55  | 42.42 | 0.86  | 0.38  |
| √           | 33.13  | 43.32 | 0.31  | 0.29  |

Center Loss. Fig. 7 visualizes the PSNR-Y of the HR and LR videos produced by MIMO-VRN as functions of time. Without the center loss (see MIMO-VRN), the PSNR-Y of both the HR and LR videos fluctuates periodically by as much as 2dB. Observe that the crest points of the HR video occur roughly at the GoF centers while the trough points are at the GoF boundaries. Table 6 performs an ablation study of how this center loss would affect the HR and/or LR videos when it is imposed on these videos. We observe that introducing the center loss largely mitigates the quality fluctuation in the corresponding HR and/or LR video (see the MAD results in Table 6 and Fig. 7). It however degrades the HR and/or LR quality in terms of PSNR-Y, as compared to the case without the loss. We make the choice of imposing the center loss on the HR video only for two reasons. First, this leads to a minimal impact on the HR reconstruction quality. The second is that the quality fluctuation in the LR video is less problematic in terms of subjective quality because the PSNR-Y measured against the bicubic-downscaled video is way above 40dB. On closer visual inspection, these LR videos hardly show any artifacts in the temporal dimension.

4.5. Complexity-performance Trade-offs

LSTM-VRN and MIMO-VRN present different complexity-performance trade-offs. (1) LSTM-VRN is relatively lightweight, having 9M network parameters as compared to 19M with MIMO-VRN. (2) LSTM-VRN does not require additional buffering/delay and storage for downscaling as is necessary for MIMO-VRN. (3) MIMO-VRN has better LR/HR quality while LSTM-VRN has more consistent LR/HR quality temporally. They use depends on the complexity constraints and performance requirements of the application.

5. Conclusion

This work presents two joint optimization approaches to video rescaling. Both incorporate an invertible network with coupling layer architectures to model explicitly the high-frequency component inherent in the HR video. While our LSTM-VRN shows that the temporal information in the LR video can be utilized to good advantage for better upscaling, our MIMO-VRN demonstrates that the GoF-based rescaling is able to make full use of temporal information to benefit both upscaling and downscaling. Our models demonstrate superior quantitative and qualitative performance to the image-based invertible model. They outperform, by a significant margin, the video rescaling framework without joint optimization.

Acknowledgements. This work is supported by Qualcomm technologies, Inc. (NAT-439543), Ministry of Science and Technology, Taiwan (109-2634-F-009-020) and National Center for High-performance Computing, Taiwan.
References

[1] 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 1, 2, 3

[2] Yuzhao Chen, Xi Xiao, Tao Dai, and Shu-Tao Xia. Hrnet: Hamiltonian rescaling network for image downsampling. In IEEE International Conference on Image Processing (ICIP), 2020. 1, 2, 3

[3] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 1, 2

[4] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. In Proceedings of International Conference on Learning Representations (ICLR), 2015. 3

[5] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In Proceedings of International Conference on Learning Representations (ICLR), 2017. 3

[6] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jie Zhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. Closed-loop matters: Dual regression networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 1, 2, 6

[7] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 3

[8] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. In Proceedings of European Conference on Computer Vision (ECCV), 2020. 1, 2, 3, 6, 11

[9] Takashi Isobe, Songjiang Li, Xu Jia, Shanxin Yuan, Gregory Slabaugh, Chunjing Xu, Ya-Li Li, Shengjin Wang, and Qi Tian. Video super-resolution with temporal group attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 1, 2, 3, 6

[10] Younghyun Jo, Seoung Wug Oh, Jaeeyon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 1, 2, 3, 6

[11] Heewon Kim, Myungsun Choi, Bee Lim, and Kyoung Mu Lee. Task-aware image downsampling. In Proceedings of European Conference on Computer Vision (ECCV), 2018. 1, 2, 3

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Proceedings of International Conference on Learning Representations (ICLR), 2015. 5

[13] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In Advances in Neural Information Processing Systems (NIPS), 2018. 3

[14] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 1, 2, 5

[15] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In Proceedings of European Conference on Computer Vision (ECCV), 2020. 1, 2, 3

[16] Yue Li, Dong Liu, Houqiang Li, Li Li, Zhu Li, and Feng Wu. Learning a convolutional neural network for image compact-resolution. IEEE Transactions on Image Processing (TIP), 2018. 1, 2, 3

[17] Bee Lim, Sanghyun Son, Heewon Kim, Seunghun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017. 1, 2

[18] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2013. 5

[19] R. Rassool. Vmaf reproducibility: Validating a perceptual practical video quality metric. In 2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), 2017. 11

[20] Mehdi SM Saajjadi, Raviteja Vemulpalali, and Matthew Brown. Frame-recurrent video super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 1, 2, 3

[21] Claude Elwood Shannon. Communication in the presence of noise. Proceedings of the IRE, 1949. 2

[22] Wanjie Sun and Zhenzhong Chen. Learned image downsampling for upsampling using content adaptive resampler. IEEE Transactions on Image Processing (TIP), 2020. 1, 2, 3, 6

[23] Xin Tao, Hongyun Gao, Renjie Liao, Yue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 1, 2, 3

[24] Yaping Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 3

[25] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019. 1, 2, 3, 6, 11

[26] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In The European Conference on Computer Vision Workshops, 2018. 1, 2

[27] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and S Yu Philip. Predrnn: Recurrent neural networks
for predictive learning using spatiotemporal lstms. In Advances in Neural Information Processing Systems (NIPS), 2017.

[28] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (TIP), 2004.

[29] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In Proceedings of European Conference on Computer Vision (ECCV), 2020.

[30] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. International Journal of Computer Vision (IJCV), 2019.

[31] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In Proceedings of IEEE International Conference on Computer Vision (ICCV), 2019.

[32] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In Proceedings of European Conference on Computer Vision (ECCV), 2018.
Appendix

This supplementary document provides additional results to validate the proposed methods. These include (1) quantitative results based on the Video Multi-Method Assessment Fusion (VMAF) metric [19], a video quality metric that is shown to correlate highly with human perception; (2) an examination of temporal consistency on the downscaled and upscaled videos following the method in RSDN [8]; (3) an ablation study of the predictive module; and (4) more qualitative results of the downscaled and upscaled videos.

A. Quantitative results based on VMAF

This section presents quantitative results for upscaled and downscaled videos in Vid4 based on the Video Multi-Method Assessment Fusion (VMAF) [19] metric. VMAF is an objective video quality metric that is shown to correlate highly with human perception. It accepts as inputs a reference video and a distorted video. Its output is a score between 0 and 100. The higher the VMAF score, the more closely the two input videos match each other. Compared in Table 7 are the VMAF scores of the reconstructed high-resolution (HR) videos produced by a few baselines (whose code is available) and our schemes. The reference videos are the original HR videos. Likewise, Table 8 shows the results for the downscaled videos, where the reference videos are the bicubic-downscaled videos.

From Table 7, we observe that our LSTM-VRN, MIMO-VRN and MIMO-VRN-C outperform IRN [29] (image-based joint optimization scheme) consistently. As compared with EDVR-L [25], a traditional video super-resolution approach, the proposed methods show significantly improved VMAF scores. These observations are in line with the PSNR/SSIM results reported in Table 1 of the main paper. A similar observation can be made regarding the results of the downscaled LR videos in Table 8, which agrees with the PSNR/SSIM results in Table 3 of the main paper. Remarkably, all three proposed methods have VMAF scores very close to 100, suggesting that their LR videos are visually similar to bicubic-downscaled videos.

B. Temporal consistency

In this section, the temporal consistency of the downscaled and upscaled videos is examined. We follow RSDN [8] to extract a row or a column of pixels at the co-located positions in consecutive video frames. We then stitch vertically (or horizontally) these extracted rows (or columns) of pixels to form an image, in order to visualize their variations in the temporal dimension. From Fig. 8, we are not aware of any noticeable inconsistency across reconstructed HR video frames. Moreover, the resulting images of our methods resemble closely the ground-truths.

In Fig. 9, IRN [29] and LSTM-VRN result in slight temporal inconsistency in some areas of the low-resolution (LR) videos, particularly the Calendar and City sequences. This is evidenced by the aliasing artifact especially appeared on the alphabet of Calendar sequence and the building of City sequence. Nevertheless, such inconsistency does not appear in the LR videos produced by MIMO-VRN and MIMO-VRN-C.

C. The effectiveness of the predictive module

Table 9 provides quantitative results to justify the effectiveness of the proposed predictive module. Recall that the predictive module forms a prediction $\hat{z}$ of the missing high-frequency component $z$ from the LR video frames $\tilde{y}$. This new feature distinguishes our schemes from IRN [29], the image-based joint optimization scheme, which uses a Gaussian noise for $\hat{z}$. The upper section of Table 9 compares the results of LSTM-VRN with IRN_Ret (the re-trained IRN that uses the same training dataset as LSTM-VRN), validating that the predictive module is effective for reconstructing better HR videos. The lower section of Table 9 conducts the same analysis for MIMO-VRN-C, where we replace $\hat{z}$ produced by the predictive module with a fixed zero tensor, a scheme termed MIMO-VRN-C-Zero. MIMO-VRN-C-Zero is trained in the same way as MIMO-VRN-C. We see that the predictive module is still effective, even though the gain is less significant than the case in LSTM-VRN.

### Table 7. Comparison of VMAF scores for $\times 4$ upscaled HR videos on Vid4. The reference videos are the original HR videos. The higher the VMAF scores, the better the HR reconstruction quality. Red and green indicate the best and the second best performance, respectively.

| Method   | Calendar | City | Foliage | Walk | Average |
|----------|----------|------|---------|------|---------|
| EDVR-L [25] | 64.11    | 47.22| 77.12   | 89.86| 69.58   |
| IRN [29]  | 78.03    | 85.16| 91.14   | 97.81| 88.04   |
| LSTM-VRN | 81.99    | 87.35| 95.31   | 98.57| 90.81   |
| MIMO-VRN | 80.44    | 85.47| 94.24   | 98.42| 89.64   |
| MIMO-VRN-C| 94.18    | 94.92| 95.38   | 97.82| 95.55   |
| LSTM-VRN | 95.86    | 95.31| 96.52   | 99.47| 96.79   |
| MIMO-VRN | 97.37    | 96.84| 97.88   | 99.83| 97.98   |
| MIMO-VRN-C| 97.94    | 97.17| 98.34   | 99.87| 98.33   |

### Table 8. Comparison of VMAF scores for $\times 4$ downscaled LR videos on Vid4. The reference videos are generated by the bicubic downscaling method. The higher the VMAF scores, the more closely the LR videos resemble the bicubic-downscaled ones. Red and green indicate the best and the second best performance, respectively.

| Method   | Calendar | City | Foliage | Walk | Average |
|----------|----------|------|---------|------|---------|
| IRN [29]  | 94.18    | 94.92| 95.38   | 97.82| 95.55   |
| LSTM-VRN | 95.86    | 95.31| 96.52   | 99.47| 96.79   |
| MIMO-VRN | 97.37    | 96.84| 97.88   | 99.83| 97.98   |
| MIMO-VRN-C| 97.94    | 97.17| 98.34   | 99.87| 98.33   |
Table 9. Ablation study of the predictive module. IRN_Ret is trained using the same dataset as LSTM-VRN. MIMO-VRN-C-Zero is an implementation of MIMO-VRN-C without the predictive module. All the presented results are evaluated on Vid4 and they show that the proposed predictive module can improve the quality of reconstructed HR videos.

| Method             | Predictive Module | HR (PSNR-Y / SSIM-Y / VMAF) |
|--------------------|-------------------|------------------------------|
| IRN_Ret            |                   | 30.72 / 0.9087 / 86.37       |
| LSTM-VRN           | √                 | 32.24 / 0.9369 / 88.04       |
| MIMO-VRN-C-Zero    |                   | 33.04 / 0.9575 / 89.00       |
| MIMO-VRN-C         | √                 | 33.40 / 0.9609 / 89.64       |

D. More qualitative results

Figs. 10 and 11 provide more qualitative results, comparing the reconstructed HR videos of different models. They again suggest that our models can recover fine details and sharp edges.

Fig. 12 presents a frame-by-frame qualitative comparison between MIMO-VRN and MIMO-VRN-C, with a GoF size of 5. MIMO-VRN-C comes with an additional center loss to ensure temporal consistency. It is seen that both MIMO-VRN and MIMO-VRN-C can successfully reconstruct image details. Comparing MIMO-VRN with MIMO-VRN-C, we are not aware of any significant quality variation in the temporal dimension, even though Fig. 7 of the main paper suggests that the HR quality of MIMO-VRN may fluctuate more significantly than MIMO-VRN-C. Fig. 13 displays more LR video frames from Vid4, showing that our models offer comparable visual quality to the bicubic-downscaled videos.
Figure 8. Visualization of temporal consistency across the reconstructed HR video frames. The images shown are formed by vertically (or horizontally) stitching rows (or columns) of pixels extracted separately from consecutive video frames at co-located positions (indicated by the red lines).
Figure 9. Visualization of temporal consistency across the downscaled LR video frames. The images shown are formed by vertically (or horizontally) stitching rows (or columns) of pixels extracted separately from consecutive video frames at co-located positions (indicated by the red lines).
Figure 10. Qualitative comparison on Vid4 for $\times 4$ upscaling. Zoom in for better visualization.

Figure 11. Qualitative comparison on Vimeo-90K-T for $\times 4$ upscaling. Zoom in for better visualization.
Figure 12. Qualitative frame-by-frame comparison between MIMO-VRN and MIMO-VRN-C, with a GoF size of 5. The images shown are ×4 upscaled HR video frames of Vid4.
Figure 13. Sample LR video frames from Vid4. Our models show comparable visual quality to the bicubic method.