# A Study on Video Superresolution Using Reference Frames

**Xi Huang**

**Tomo Miyazaki** **Yoshihiro Sugaya** **Shinichiro Omachi**

## Abstract

In this paper, we proposed a reference-based super resolution method to tackle the super resolution task. The existing super resolution methods could be roughly divided into three types: Single Image Super Resolution (SISR), Multi-frame Super Resolution (MFSR) and the Reference-based Super Resolution (Reference-based SR). Recently, SISR and MFSR have been widely studied. However, these SR methods have limited performance due to the loss of the information in the low-resolution images. This limitation will degrade performance when the under-resolved images are at a very low-resolution level. As a solution of this limitation, the proposed reference-based SR method has three advantages: 1) With the support of the high-resolution reference frame, the performance of the SR model is able to be further improved; 2) With only two input images, the calculation cost can be cut off compared with the MFSR methods; 3) With the help of the reference, it allows the proposed model to obtain superior results even when dealing with the very poor quality images, which beneficial to the data reduction. To evaluate the proposed model, we did a lot of experiments and compared our model to the baselines with both quantita-

tive and visual estimations. We experimented under both scaling factors of $\times 2$ and $\times 4$, and found that the proposed method could recover high quality super-resolved images from a very low resolution level. This experiment results showed the superiority of the proposed model on the very poor quality images. In the analysis, we assumed a video compression system and conducted video super resolution experiments using the proposed method. Finally, we found a better fashion for the video super resolution and verified the practicability of this video compression system.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

## 1.1 Background

Super resolution (SR) technology is widely known in image processing. Image super resolution aims at converting a low-resolution (LR) image into a high-resolution (HR) level. Recently, super resolution has been widely studied. Single image super resolution (SISR) is the most popular task of the SR techniques. Dong [5] realized an end-to-end mapping from low-resolution image to high-resolution image by using a simple three-layer network called SRCNN. Since then, learning-based super resolution method using Convolution neural network (CNN) has attracted attention of many researchers due to its outstanding performance. Kim [9] improved the SR performance by increasing the depth of the network to 20 layers. It can be considered that the depth of the network has a great affect on the super resolution. Such observation has been seen in the image recognition field. Lim [14] proposed EDSR and MDSR respectively and showed the importance of network depth on the super resolution research.

Multi-frame super resolution (MFSR) is another kind of super resolution task, which utilizes a series of LR frames as input to super-resolve one of them. Different from SISR, MFSR takes advantage of time information which is suitable for the video super resolution. Huang [8] proposed to combine the bidirectional recurrent neural network (RNN) and the 3D convolutional network. Such architecture successfully made use of the bidirectional time information and had a good progress on the MFSR methods. However, such methods suffer from high computational cost especially when using the 3D convolution.

However, there exist a limitation either for SISR or MFSR. Since LR images are inherently devoid of information, such methods can hardly recover these images to a real high-resolution level. Therefore, the proposed method focused on the SR method using reference frame.

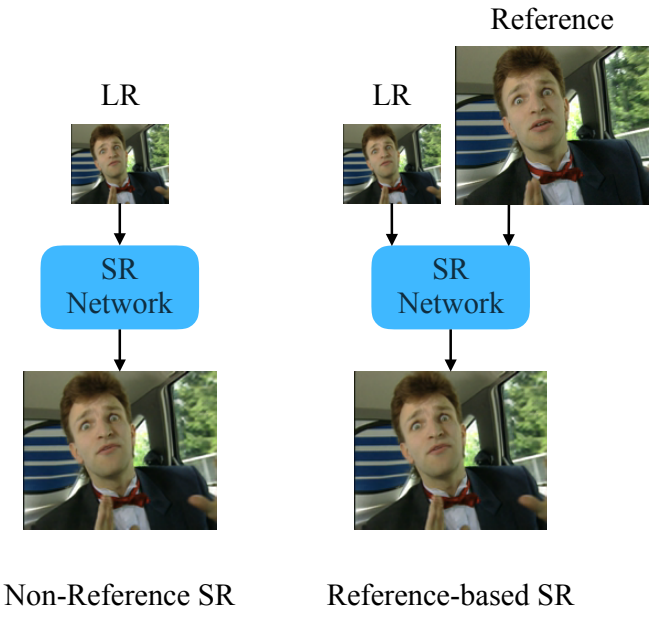



Figure 1: Simple illustration of the reference-based super resolution.

Reference frame is an high-resolution adjacent image of the target frame to be super-resolved. In the proposed method, we input it as a reference to help the super resolution. A simple illustration of our reference-based super resolution is showed in Figure 1.

The main advantages of our work are three-folds:

1. With the support of the reference frame, most of the information in the target frame is kept. Thus the SR performance can be expected to be further improved.

2. Compared with the MFSR, reference-based method has less input images, which cuts off the computational cost.

3. Since the reference frame maintains rich information, we can recover the target frame even from a lower level, which is beneficial to the data reduction.

## 1.2 Purpose

In conclusion, our research aims at realizing the video super resolution using the reference frames. While utilizing the motion vector between the target and the reference, we successfully applied the high-frequency details of the reference frame to the reconstructed target frame.

## 1.3 Definition

For the convenience of writing, we define the following specific terms.

1. Target frame: The frame image to be super resolved.
2. Reference frame: high-resolution adjacent frame of the target frame, which is different from but shares abundant similar information with the target frame.
3. LR: Low-resolution image of target frame.

## 1.4 Composition

The composition of this paper is as follows.

1. **Introduction**: Described the background and purpose of this research.
2. **Related Works**: Studies related to the super resolution.
3. **Proposal**: Details of the proposed method.
4. **Experiments**: Comparison experiments on the baseline model and the proposed model.
5. **Conclusion**: Conclusion of the paper and the future work.

# 2 Related works

## 2.1 Super Resolution

Recently, super resolution has been widely studied. The current super resolution methods could be roughly divided into 3 types: Single image SR, Multi-frame SR and Reference-based SR. As the most widely studied SR method, SISR got a significant progress during past 5 years. A lot of well designed networks have been proposed which boosted the performance of the super resolution. Kim [10] developed a deeply-recursive network. Adding a loss to each recursive unit, this method successfully reduced the loss and improved the performance of the network. Inspired from the Laplacian Pyramid, Lai [11] proposed to operate the super resolution level by level. Different from the above approaches, there is another kind of GAN-based method, which aims at generating photo-realistic images instead of recovering the LR images to the original ones. From this perspective, Ledig proposed SRGAN [12] which could generate texture details when super-resolving at a large upscaling factors.

While SISR methods only deal with a single image, Multi-frame SR methods deal with a series of consequent frames, which hold more useful information, e.g., time and motion. Huang [8] proposed a bidirectional network to make use of the time information for the super resolution. Liu [15] introduced the optical flow in the network to make an alignment of the frames. However, the common weak point of these methods is the high- computational cost.

## 2.2 Reference-based SR

In addition of the above two types of SR methods, Reference-based SR method as a minor approach of the super resolution has attracted attentions in recent years. In the Reference-based methods, a high-resolution reference image is utilized when super-resolving the LR

image. Normally, these HR reference images are different images from the LR ones but share some similar relationship with each others.

The reference images could be roughly divided into two types: 1) images that share similar texture features with the target images; 2) images similar to the targets but taken from different viewpoints. For the first category, [20, 19] proposed a Natural Texture Transfer. They extracted features from both LR and HR images and let them get through a reference-conditioned texture transfer. In the transfer, corresponding texture features will be adopted into the LR images. With different references, different textures will be adopted.

For the references taken from different viewpoints, most of the Reference-based methods adopted a patch-based [2, 21] approach, which had a match between the low-resolution patches and the high-resolution ones. For matching the LR and HR patches, [2] first down-sampled the HR patches then calculated the features, which did not make full use of the high frequency features of the HR references. In addition, patch-based approaches inherently lack the flexibility for the non-rigid deformation of the super resolution. To tackle this problem, Zheng [7] proposed to make use of the optical flow instead of the patch matching manner. However, this method operated on the super-resolved images, which we think might be a waste of calculation. In this paper, we leverage the adjacent frames as the reference.

### 2.3 Optical Flow

To make use of the motion vector between the LR and the reference frame, we introduced optical flow in our proposed model. Optical flow is the instantaneous velocity of the relatively moving pixel in an observed image. Under sequences of ordered images, optical flow can be calculated from the relationships between the adjacent frames. Given a pair of frames, the position of pixel $A$ is $(x_1, y_1)$ in the $t$-th frame, while the position changes to $(x_2, y_2)$ in the $(t + 1)$-th frame,

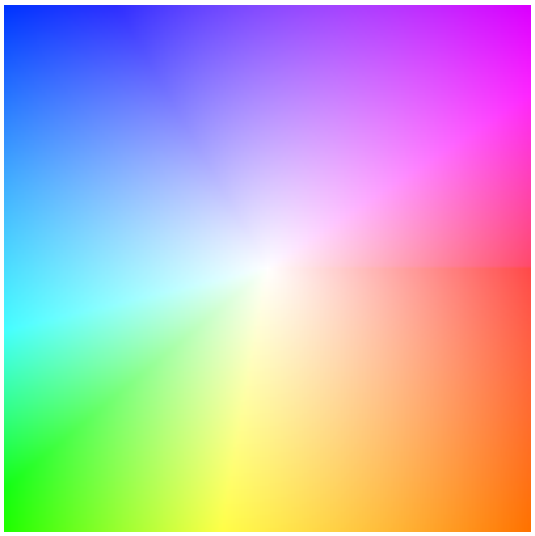


Figure 2: Color coding of the optical flow. Different colors represent different directions and the intensity represents the magnitude of the velocity.

i.e.,

$$I_t(x_1, y_1) = I_{t+1}(x_2, y_2) = I_{t+1}(x_1 + u, y_1 + v) \tag{1}$$

The optical flow between the two frames could be calculated as $(u, v)$, so we can obtain a 2-channel optical flow image with the same size of the frames.

Optical flow is usually visualized using color coding. As showed in Figure 2, every pixel in the image indicates the displacements of two directions. Different colors represent different directions and the intensity represents the magnitude of the velocity.

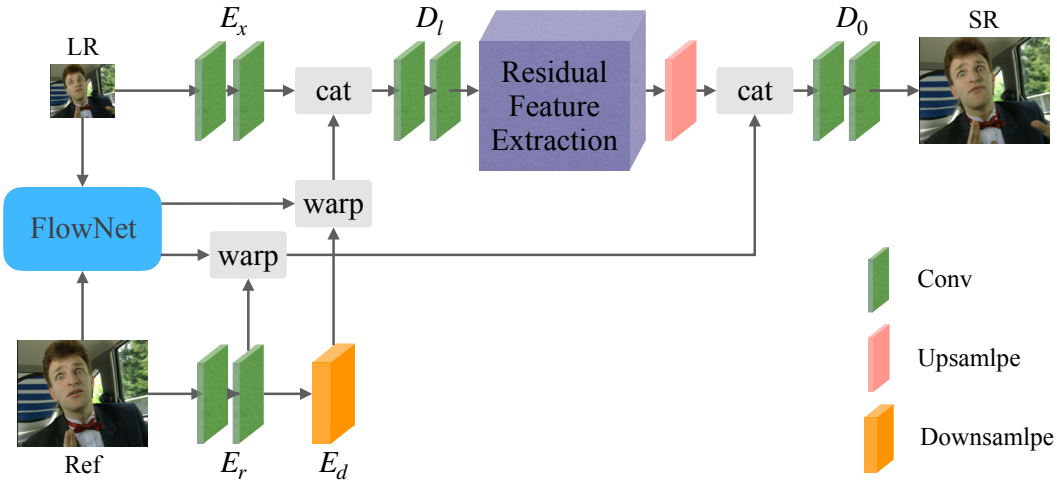



Figure 3: Architecture of our proposed method.

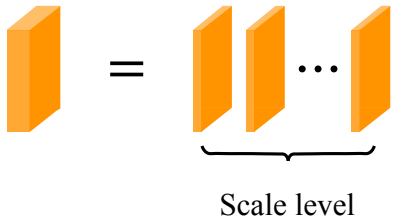



Figure 4: Downsample Encoder. The number of the downsample layers is determined by the scale level of the LR frame, which is computed as $l = log_2 L$ where $L$ denotes the scaling factor.

# 3 Proposal

In this section, we will explain the details of the proposed method.

## 3.1 Overview

The main architecture of the proposed network is displayed in Figure 3. First, we define the mathematical system as follows. As we can see from the Figure 3, there are two inputs in our model. One is LR which

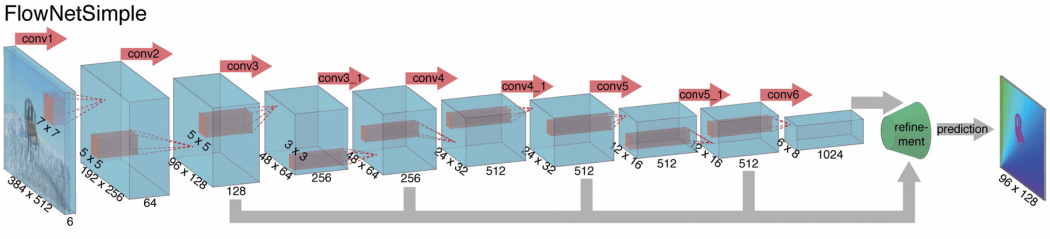



Figure 5: Architecture of the FlowNetS.

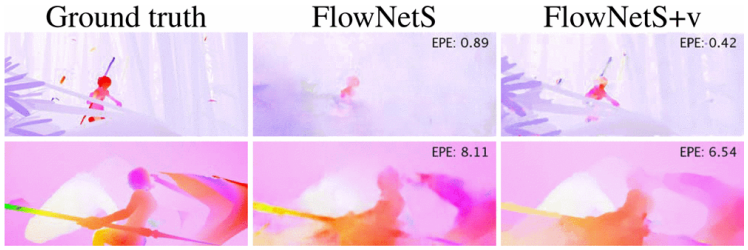



Figure 6: Examples of the optical flow prediction in the FlowNet.

refers to the low-resolution image $\mathbf{x}$ downsampled from the original high-resolution ground truth $\mathbf{y}$ with the scaling factor $L$. Another is the reference high-resolution frame $\mathbf{r}$, which refers to the adjacent frame of $\mathbf{x}$. SR refers the super-resolved image which we denote it as $\hat{\mathbf{y}}$.

First we calculate the optical flow between the LR and the reference frames via FlowNet, and both of the two inputs frames should get through their corresponding encoders. Then the encoded feature maps of the reference frame will be warped with different scales of optical flows. These warped features, which contain abundant high-frequency details, will be concatenated to the encoded and extracted feature maps respectively to introduce high-frequency reference. Finally, the super-resolved target frame can be obtained by getting through a decoder.

Thus the proposed model can be described as

$$\hat{\mathbf{y}} = F(\mathbf{x}|(\theta, \mathbf{r})) \tag{2}$$

where $F$ denotes the proposed model and $\theta$ denotes the parameter set of $F$.

## 3.2 Optical flow and Warp

### 3.2.1 FlowNet

To make full use of the reference frame, we introduced optical flow in the network. By warping with optical flow, we successfully applied those high-frequency details to the LR image. To calculate the optical flow between $\mathbf{x}$ and $\mathbf{r}$, we exploit FlowNet[6] in our model.

FlowNet is a neural network model that was proposed to calculate the optical flow of two images. Here we utilized the FlowNetS model in our method which is illustrated in Figure 5. The two images are concatenated before input into the network. It first extracts the features of the two concatenated images by decreasing the size of the feature maps and increasing the feature channels gradually. Then with the refinement network, the optical flow will be predicted from low size feature maps to the upsampled original size of the images.

To make it applicable for our model, we replace the last linear interpolation layer with the convolution layer to calculate the corresponding size of the optical flow. Two examples of the performance in the FlowNet[6] are showed in Figure 6. The first column displays the ground truth of the optical flow, and the second and third columns are the predicted optical flows calculated using the FlowNetS and FlowNetS+v respectively.

To predict the optical flow between the LR and reference frames, we first upscale the LR to the original size using the bicubic kernel. By input the two frames into the FlowNet, we can have predicted flows as follow:

$$(f_l, f_0) = FlowNet(\mathbf{x}^{up}, \mathbf{r}) \tag{3}$$

where $\mathbf{x}^{up}$ denotes the upscaled LR frame. $f_l$ denotes the predicted flow of the $l$-th scale level.

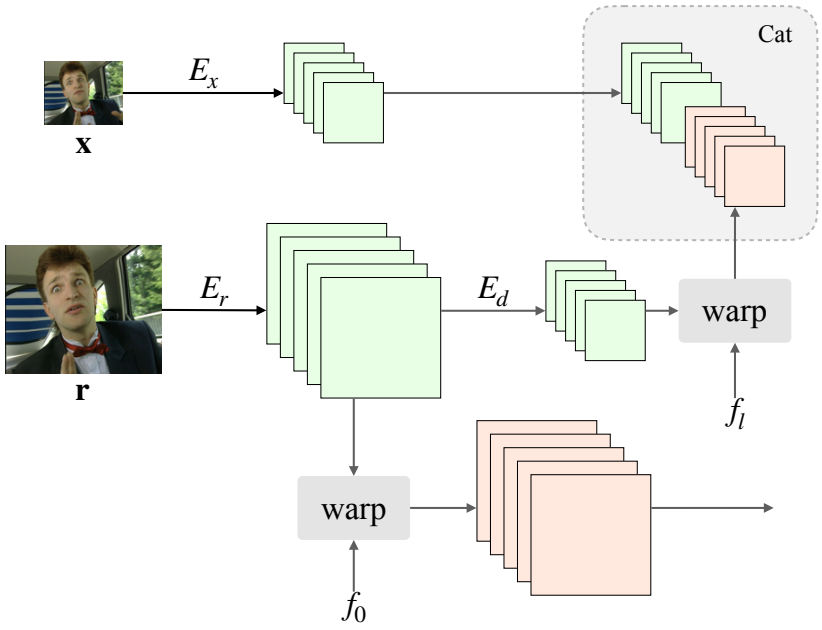



Figure 7: Details of the Encoding and Warping phase.

### 3.2.2 Encoder and Decoder

Since the LR and reference frames are of different size, we first encode the two input frames as the preparation for warping. Image warping is a kind of image processing of image distortion and affine, i.e., with the calculated optical flows, we can distort the image to the after-movement image based on the motion vector.

The illustration of the encoding phase is displayed in Figure 7. LR frame is directly input a 2-convolution-layer encoder, while the reference frame has to get through two encoders. The first one is the same as LR frame and the second one is an additional downsample encoder. The encoded data can be described as follow:

$$X_e = E_x(\mathbf{x}) \tag{4}$$

$$R_e = E_r(\mathbf{r}) \tag{5}$$

$$R_D = E_d(R_e) \tag{6}$$

where $E_x$ and $E_r$ denote the convolution encoders of LR and reference frames respectively. $E_d$ denotes the downsample encoder. After encoding, the encoded data of the reference frame $\mathbf{r}_D$ will be warped with the predicted flow $f_l$

$$R_w^l = warp(f_l, R_D) \tag{7}$$

Finally, we concatenate the encoded $\mathbf{x}_e$ and the warped $\mathbf{r}_D$ as the input of the decoder

$$X_d^l = D_l([X_e, R_D]) \tag{8}$$

where $[\cdot]$ denotes the concatenation operation and $D_l$ denotes the decoder for the $l$-th scale level data. $X_d^l$ here refers to the decoded data of the $l$-th scale level.

### 3.3 Feature Extraction

When finished warping at the low scale level, we will enter the feature extraction part. As we mentioned in the introduction, the depth of the network is essential for the performance of super resolution, so we choose a Residual in Residual (RIR) [18] backbone as our feature extractor.

The details of the extractor architecture is illustrated in Figure 8. A basic Residual Block is consist of two convolution layers with a skip connection. The RIR architecture stacks these blocks by adding several short skip connection and a long skip connection. These skip connections could help to simplify the task of the layers inside the connection, which leads to a very deep trainable network. We simply form the backbone by stacking the Residual Blocks (RB) with short and long skip connections and have

$$F_x = H_f(X_d^l) \tag{9}$$

where $H_f$ denotes the residual feature extractor and $F_x$ denotes the extracted feature maps.

Note that [18] proposed a Channel Attention (CA) mechanism in the RBs, but we only utilize the simple stacked RBs for simplicity. As for

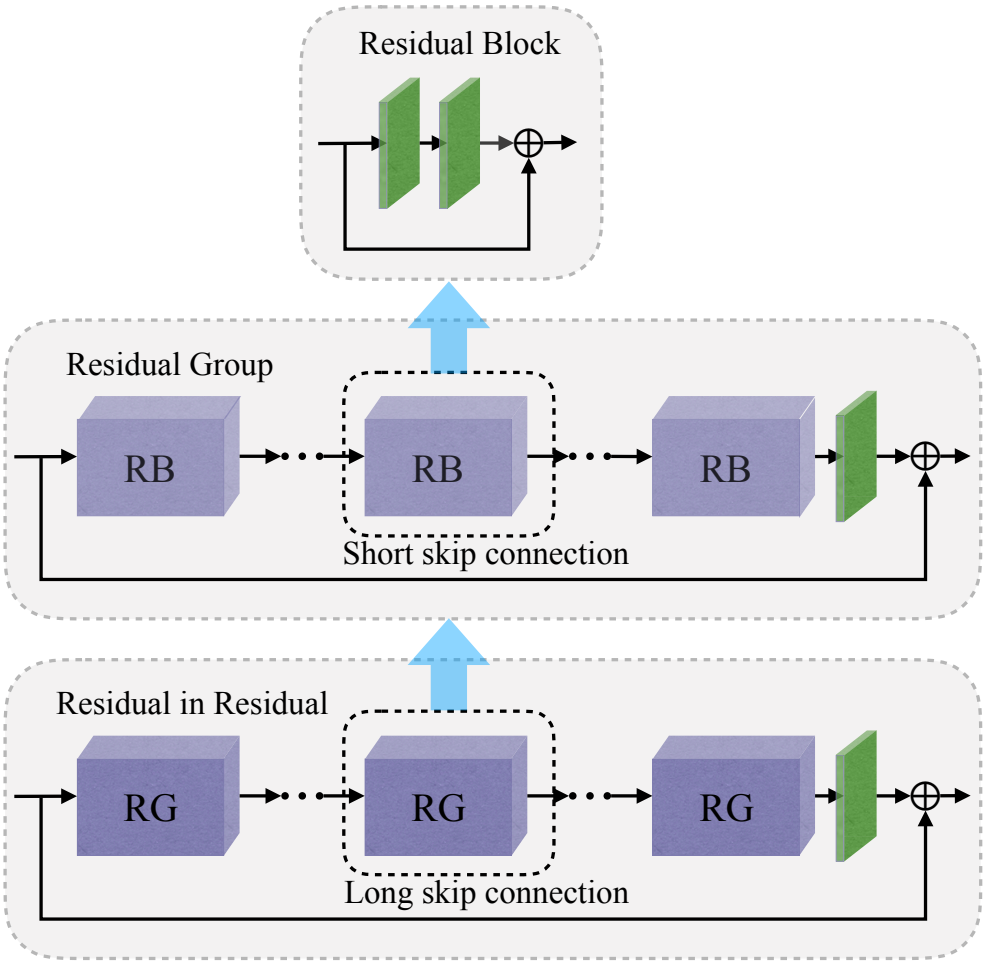



Figure 8: Details of the Residual in Residual feature extractor.

a better performance, we also conducted additional experiments using CA.

### 3.4 Multi-scale Warping

In our proposed model, we consider conducting warping at two different scale levels. First is at the low resolution scale level $l$, and the other is at the original high-resolution scale level $0$. After the feature extraction part, the extracted features will go through an upsample

layer

$$F_U = H_u(F_x) \tag{10}$$

where $F_U$ denotes the upsampled features and $H_u$ denotes the upsample layer. Here we used the Sub-Pixel Convolution Network [3] as the kernel of the upsample layer.

Then the same as in the low scale level $l$, the warping operation will be conducted at the high scale level $l = 0$ like

$$R_w^0 = warp(f_0, R_e) \tag{11}$$

Note that here we take $R_e$, which denotes the encoded data before the downsample encoder, as the under-warping data maps instead of $R_D$ since the extracted features have been upscaled.

Finally, the super-resolved target frame $\hat{\mathbf{y}}$ will be output after going through the 2-layer decoder $D_0$

$$\hat{\mathbf{y}} = D_0([F_U, R_w^0]) \tag{12}$$

The above is the whole main flow of the proposed network model.

# 4 Experiments

In this section, we will talk about the experiments conducted to show the performance of the proposed model.

## 4.1 Datasets

We first describe the dataset using in our work. We utilize two video datasets for training, which are *YUV21* [1] and *Hollywood II* [16]. *YUV21* is a popular video dataset and was used in many video super resolution tasks and all video sequences are in the uncompressed YUV4MPEG format. *Hollywood II* is a video dataset that contains 12 classes of human actions and 10 classes of scenes distributed over 3669 video clips and approximately 20.1 hours of video in total. In our experiments, we used 44 video clips of these two datasets in total for the training phase and 6 clips from *YUV21* as the test set.

For the experiments, we first extracted frames from these videos. For the purpose of learning large motion in the SR model, we extracted one frame every two frames and make two adjacent extracted frames as one frame set. The first frame in the frame set is used as the reference and the second one is the target. Thus we in total extracted 1174 frame sets from the training videos and 15 sets from the test videos as training and test data respectively. For the data augmentation, we conducted 3 types of rotation, i.e., $90°$, $180°$, $270°$, and horizontal flip.

## 4.2 Experimental Settings

Here we specify the details of the implement setting of the experiments.

### 4.2.1 Baseline Models

To evaluate the effectiveness of the proposed model, we compared its performance with two kinds of baseline model. As mentioned in Section 3.3, we utilized the models in RCAN [18] as our compared baseline. We introduced a model that simply stacked the Residual

Blocks as the first baseline. We followed the architecture settings in [18], i.e., the number of the Residual Group and the Residual Block are 10 and 20 respectively. We also set the kernel size of the convolution layers with $3 \times 3$ and each layer holds a filter number of 64. Note that this is a single-image super resolution method, and we validated whether the under-resolved LR frames could recover those high-frequency details via the proposed model. For a better performance, we also exploited the RCAN model that applied a channel attention mechanism in the Residual Blocks as the second baseline.

### 4.2.2 Training Settings

The experiments were conducted under the scaling factors of $\times 2$ and $\times 4$ respectively. For the training phase, we set the initial learning rate with $1 \times 10^{-4}$ and let it decay to the half of itself every 4 epochs. The batch size $N$ was set to 16, and in each batch, we extracted patches with the sizes of $96 \times 96$ and $128 \times 128$ for $\times 2$ and $\times 4$ scaling factors respectively. For the fairness of comparison, we utilized the same training loss function as the [18], i.e., $L_1$ loss function

$$L_1(\theta) = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_1 \tag{13}$$

To train the proposed network, we used the ADAM optimizer where $\beta_1$ was set to 0.9 and $\beta_2$ was set to 0.999. We implemented the model using Pytorch [17].

## 4.3 Experimental Results

Under the settings above, we trained the models conditioned on the scaling factors of $\times 2$ and $\times 4$. We evaluated the these models during the training phase using the test dataset. The averaged PSNR results on the test dataset are showed in Figure 9 and Figure 10.

Under the scaling of $\times 2$, we trained 2 kinds of baseline models and their corresponding proposals. *Baseline* denotes the simply stacked

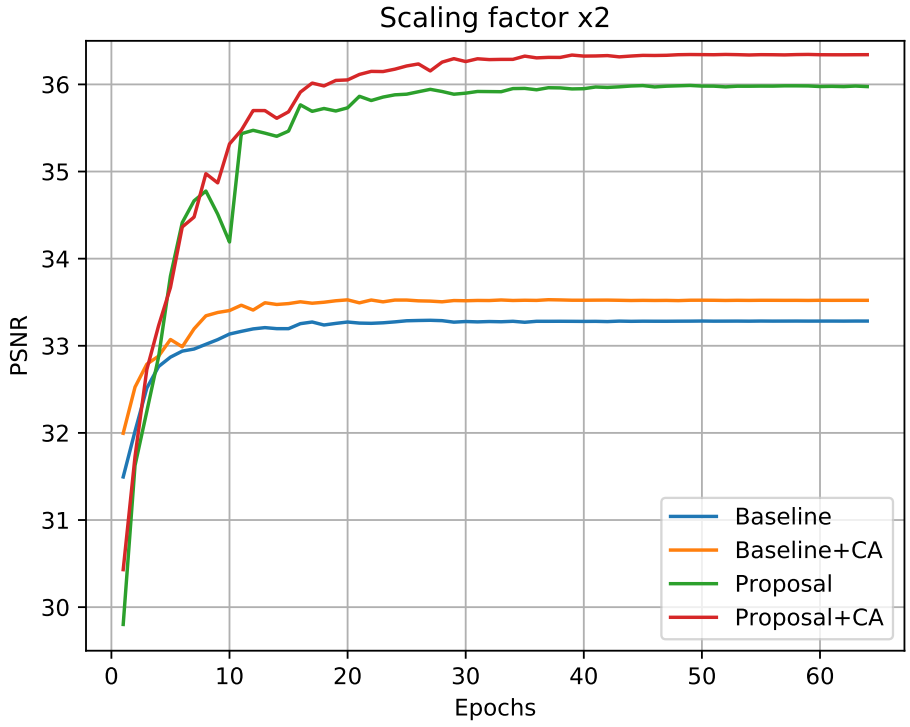



Figure 9: PSNR evaluation on the test dataset during training under the scaling factor of 2. CA here denotes the channel attention used model.

Residual Block model while *Baseline+CA* denotes the channel attention model. As we can observe from the Figure 9, the baselins are obviously lower than the propsals no matter the channel attention is used or not. Nevertheless, we still could confirm the effectiveness of the channel attention mechanism that it led the *Baseline* and the *Proposal* to better performances.

Due to the limit of the storage, we only trained the *Baseline* and the *Proposal* models, where the channel attention was not used, under the scaling of $\times 4$. The gap between these two models becomes more obvious as observed. We can see that averaged PSNR of the baseline is only about 27.6 while the proposed model can reach over 32.

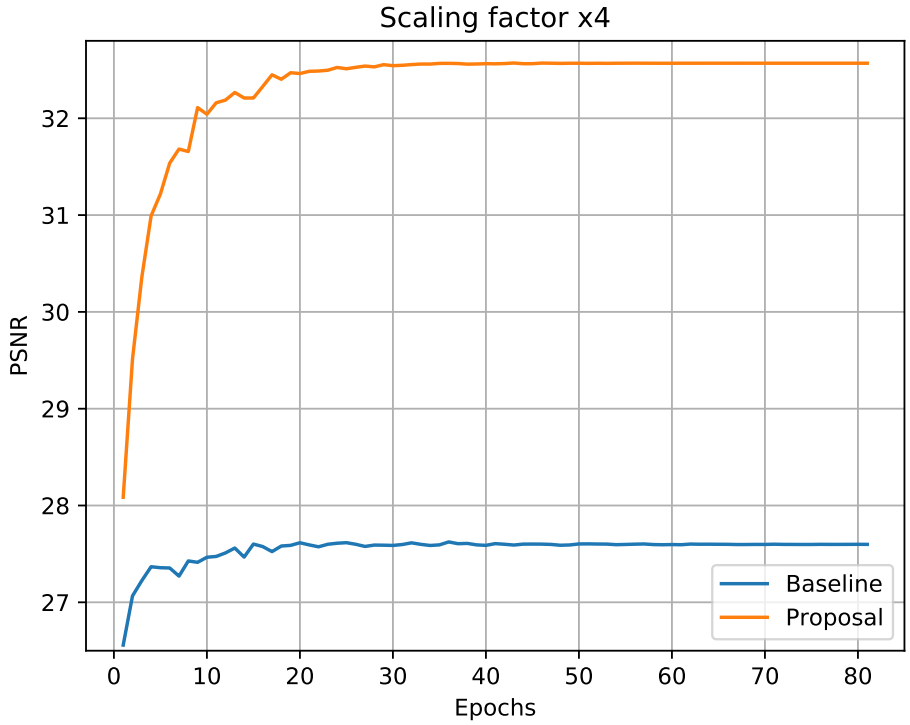



Figure 10: PSNR evaluation on the test dataset during training under the scaling factor of 4.

The results indicates the intuitively show the superiority of the proposed reference-based super resolution model.

### 4.3.1 Quantitative Evaluation

For a more clear evaluation of the proposal, we report the PSNR and SSIM scores of every test frame in Table 1 and Table 2.

In Table 1, the black bold represents the highest score in the comparison while the blue color represents the second one. Consistent with the averaged evaluation, most of the test frames obtained highest scores when super-resolved using the *Proposal+CA* model and obtained the second when using the *Proposal* model. However, some of the them were given the opposite results, i.e., they got higher scores when us-

ing the baseline models. We will discuss about these results in the following sections.

Though some results using the baseline moedls showed in Table 1 are better than the proposals, such kind of results does not appear in the Table 2. Table 2 reports the quantitative results of the test frames under the scaling factor of $\times 4$, i.e., the LR frames will have lower resolution and are more difficult to reconstructed. As observed from the table, all PSNR scores of the *Proposal* are higher than that of the baseline. Similar results are also seen from the SSIM scores except the scores of *stefan9_2*, which are not much different.

From the above observations, we can find that our reference-based model perform better than those non-reference models especially with very low-resolution LR images. Such discovery is important for the video super resolution, i.e., we can super-resolve those very low-resolution frames to decent high resolution ones only using an additional reference frame.

Table 1: Quantitative comparison with the baseline models on test dataset under the scaling factor of x2. CA refers the use of Channel Attention mechanism proposed in [18]. The black bold represents the highest score in the comparison while the blue color represents the second one.

| Method | Baseline | | Baseline+CA | | Proposal | | Proposal+CA | |
|---|---|---|---|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| carphone10_4 | 33.08 | 0.9599 | 33.42 | 0.9633 | 34.19 | 0.9660 | **34.45** | **0.9674** |
| carphone7_2 | 33.12 | 0.9626 | 33.35 | 0.9646 | 34.26 | 0.9714 | **34.42** | **0.9725** |
| carphone7_4 | 33.83 | 0.9619 | 34.60 | 0.9669 | 35.01 | 0.9702 | **35.31** | **0.9711** |
| grandma5_2 | 36.19 | 0.9509 | 36.36 | 0.9525 | 42.53 | 0.9865 | **42.76** | **0.9871** |
| grandma5_4 | 35.86 | 0.9495 | 35.99 | 0.9511 | 41.45 | 0.9843 | **41.71** | **0.9851** |
| miss am4_2 | 41.50 | 0.9843 | 41.73 | 0.9854 | 43.96 | **0.9884** | **44.01** | **0.9884** |
| miss am4_4 | 41.80 | 0.9844 | 42.46 | 0.9867 | 46.02 | 0.9899 | **46.27** | **0.9903** |
| salesman12_2 | 31.67 | 0.9075 | 31.65 | 0.9090 | 40.97 | 0.9880 | **41.42** | **0.9888** |
| salesman12_4 | 31.59 | 0.9057 | 31.52 | 0.9066 | 41.74 | 0.9906 | **42.13** | **0.9910** |
| stefan4_2 | 25.41 | 0.8594 | **25.60** | **0.8675** | 25.33 | 0.8580 | 25.45 | 0.8613 |
| stefan4_4 | 25.43 | 0.8553 | 25.67 | 0.8656 | 25.52 | 0.8818 | **27.39** | **0.9131** |
| stefan9_2 | 25.84 | 0.8409 | **25.97** | **0.8467** | 25.32 | 0.8130 | 25.61 | 0.8278 |
| stefan9_4 | 25.41 | 0.8371 | **25.60** | **0.8456** | 25.18 | 0.8216 | 25.39 | 0.8323 |
| suzie2_2 | 38.58 | 0.9624 | 38.82 | 0.9643 | 39.38 | 0.9692 | **39.80** | **0.9709** |
| suzie2_4 | 39.95 | 0.9718 | **40.09** | **0.9727** | 39.78 | 0.9707 | 40.07 | 0.9724 |
| Average | 33.28 | 0.9262 | 33.52 | 0.9299 | 36.04 | 0.9433 | **36.41** | **0.9480** |

Table 2: Quantitative comparison with the baseline model on test dataset under the scaling factor of x4.

| Method | Baseline | | Proposal | |
|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM |
| carphone10_4 | 26.07 | 0.8174 | **29.64** | **0.9102** |
| carphone7_2 | 26.12 | 0.8231 | **28.83** | **0.9165** |
| carphone7_4 | 26.35 | 0.8225 | **29.81** | **0.9081** |
| grandma5_2 | 29.95 | 0.8359 | **40.21** | **0.9796** |
| grandma5_4 | 29.79 | 0.8357 | **39.13** | **0.9773** |
| miss am4_2 | 34.20 | 0.9274 | **40.22** | **0.9782** |
| miss am4_4 | 34.17 | 0.9273 | **43.38** | **0.9833** |
| salesman12_2 | 26.95 | 0.7094 | **37.83** | **0.9739** |
| salesman12_4 | 26.93 | 0.7100 | **39.31** | **0.9842** |
| stefan4_2 | 21.24 | 0.6043 | **21.46** | **0.6315** |
| stefan4_4 | 21.24 | 0.6030 | **25.41** | **0.8623** |
| stefan9_2 | 22.26 | **0.6194** | **22.32** | 0.6146 |
| stefan9_4 | 21.78 | 0.6050 | **21.98** | **0.6194** |
| suzie2_2 | 33.04 | 0.8885 | **35.50** | **0.9408** |
| suzie2_4 | 33.88 | 0.8969 | **34.82** | **0.9177** |
| Average | 27.60 | 0.7751 | **32.66** | **0.8798** |

### 4.3.2 Visual Evaluation

In addition to the quantitative results, we provide visual evaluations on the test frames to show the high quality performance of our proposed model.

Figure 11 shows the visual comparison results under the scaling factor of $\times 2$. The first row displays the ground truth of the target frames and rest 3 rows display the LR frames, SR frames super-resolved by baseline and proposal, respectively.

High performance of the proposed model can be confirmed from the zoomed patches. We can observe from the patches that the proposed model recovers more high-frequency details than the baseline. Observing the forth row of the Figure 11, we can see that the baseline model failed to reconstruct the stripes of the tie on the man. Though it recovered some stripe-like appearances in the picture, the direction of which is wrong compared with the ground truth. In the contrast, the proposed model successfully recover the stripes. Similar case can be observed from the fifth row of the Figure 11, where the stripes of the font are super-resolved by the proposed model. Also in the last row, the proposal recovers the double-fold eyelid of the woman while the baseline reconstructs it as a single-edged eyelid.

In spite of the above, there still are some undesirable results of the proposed model when experimented under the scaling factor of $\times 2$. Such results are also reported in the quantitative evaluation of Table 1. The visual evaluations of these test samples are showed in Figure 12. When paying attention on the zoomed patches of first row, we can observe that both the ground truth and the baseline result only have two lines, while an additional blurred line appears under the two lines in the proposal patch. Similar result could be obseved from the second row, either. This reason can be inferred from the reference frame, where there is a line at the same location of the blurred line. This means that some appearances in the reference, which do not exist in the target, did not disappear after the super resolution. Since this is a

test sample that contains a large motion and camera movement, it may be hard for the proposed model to deal with such large movements.

Next we showed the visual evaluation results of the case under $\times 4$ scaling. Even without the help of the zoomed patches, we can clearly distinguish the differences between the baseline and the proposal. We observed that there are obvious artifacts in the baseline results. This is not surprising since our test frames themselves are difficult samples, which means it is very hard to conduct super resolution under such high scaling condition. However, benefit from the reference frame, the proposed model tackled such problem by referring to the reference frame. As can be seen from the Figure 13, the proposed model generated very clear high-resolution results compared to the baseline.

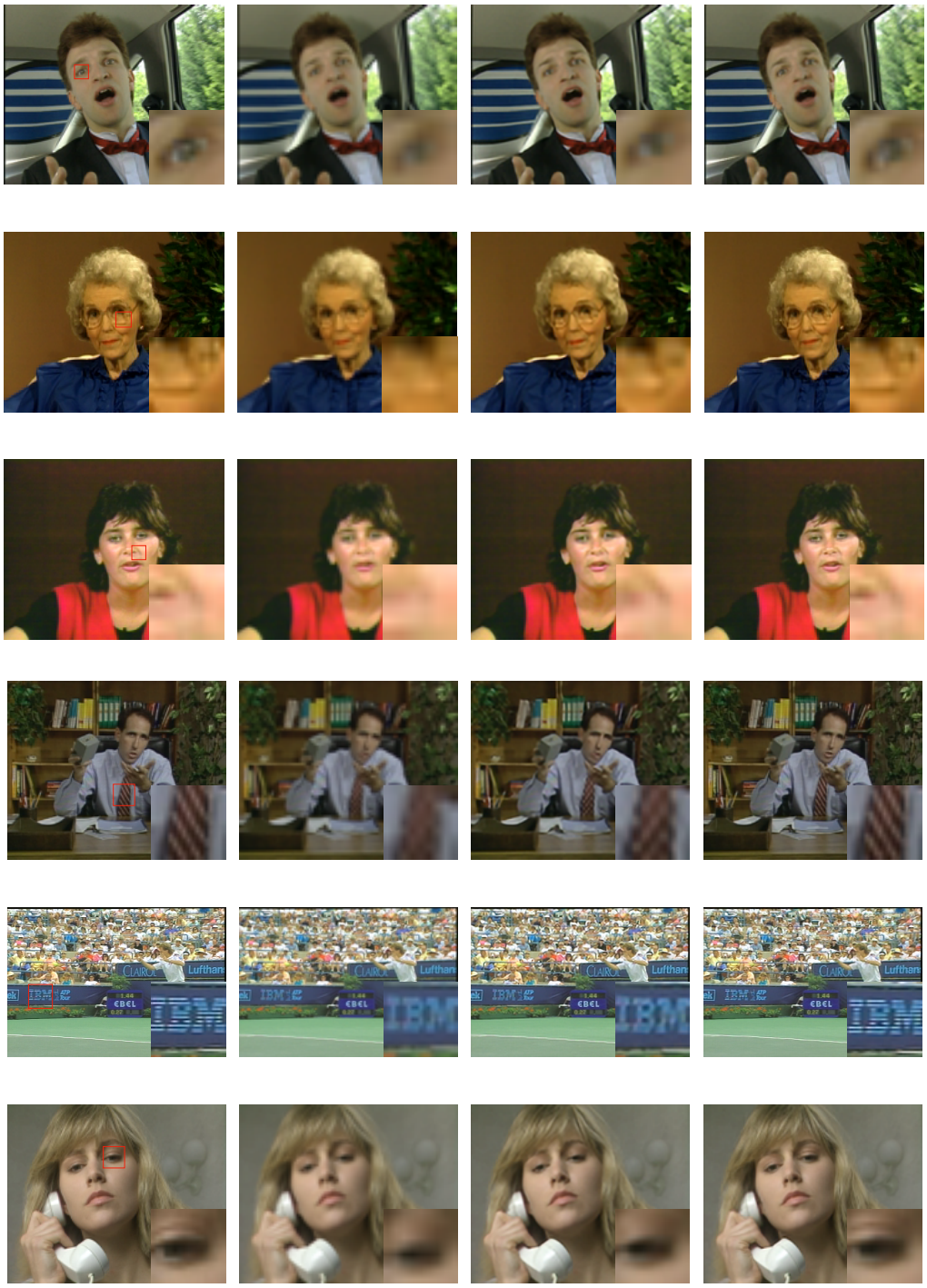



Figure 11: Visual comparison on the test dataset under the scaling factor of $\times 2$.

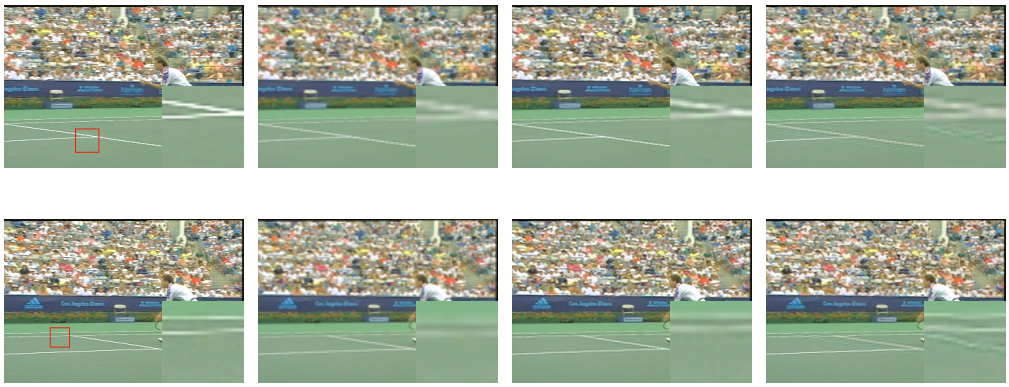



Figure 12: Undesirable test results under the scaling factor of $\times 2$.

# 5 Analysis

In this section, we conducted analysis experiments to further explore the practicability of the proposed model.

## 5.1 Video Super-Resolution with Reference

Since the proposed model could well perform even from a very low-resolution condition, this advantage could contribute to the data reduction during the video super resolution. Similar to the normal video compression, we assumed a video compression system using video super resolution and explored the feasibility of proposed model. Here we use the reference frame as the I frame and super-resolve the following consistent frames with the proposed model. We proposed two fashions to conduct the experiment.

1. **SR Reference**: Super-resolve the first target frame using the previous high-resolution frame as the reference. Then super-resolve the following LR frame using the super-resolved frame as the reference.

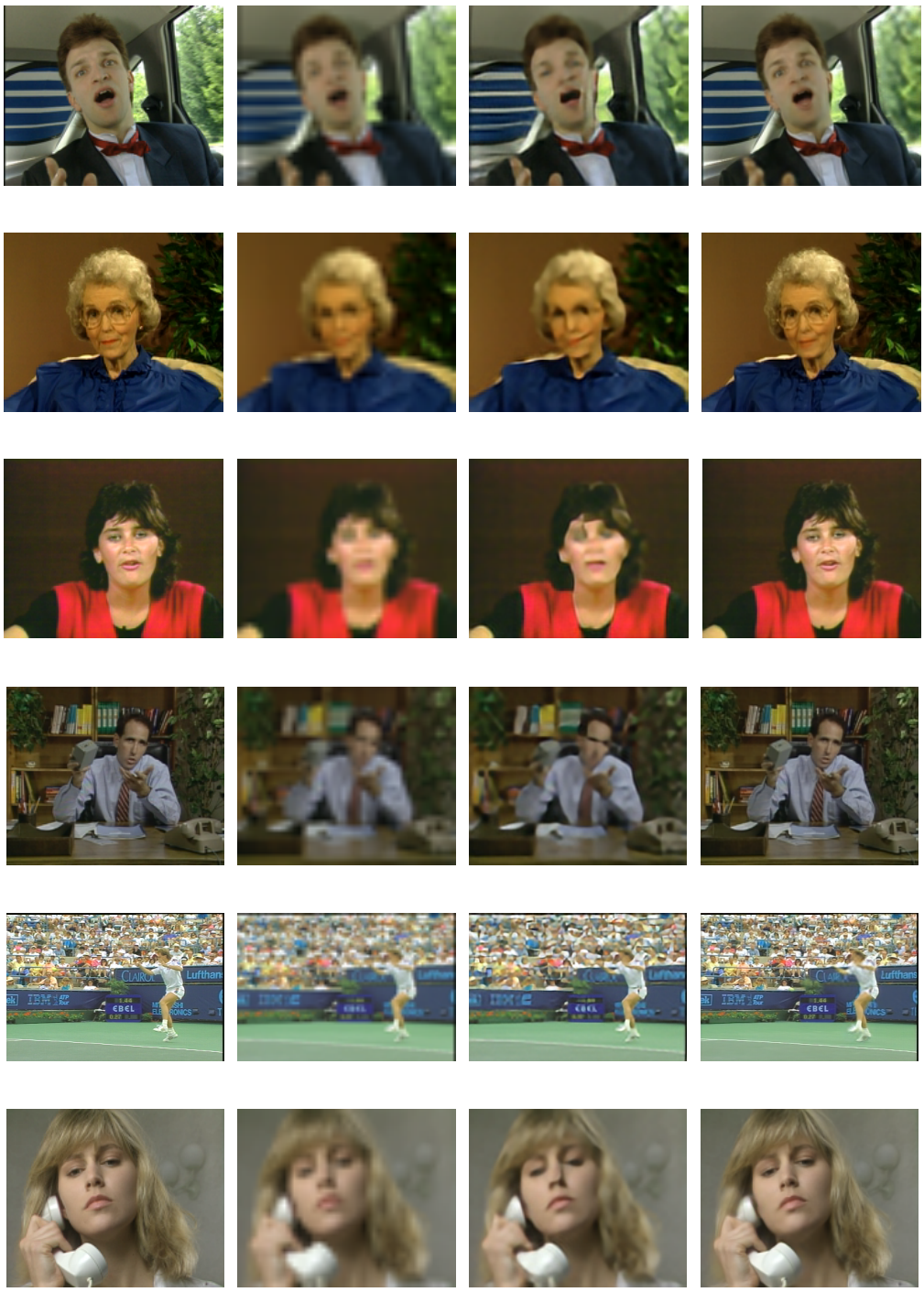



Figure 13: Visual comparison on the test dataset under the scaling factor of $\times 4$.

2. **Single Reference**: Super-resolve a series of LR frames using the same high-resolution frame as the reference.

We experiment via the above two fashions to explore the better way to conduct video super resolution.

### 5.1.1 Settings

We evaluate the proposed model with the scaling factor of $\times 4$ only. As the standard comparative value, we conducted additional experiments using the baseline under the scaling factor of $\times 2$ to estimate the performance of the proposed model. For the comparison, we also experimented using the baseline under the scaling factor of $\times 4$. Different from the above experiments, where we extracted one frame every two frames, we extracted a series of consistent frames as for this test. In the experiment, we utilized two videos from the test data, and extracted 11 consistent frames of each video as the test data.

### 5.1.2 Results

The experiment results for the quantitative evaluation are reported in the Table 3. To our observation, the performance of the Single Reference fashion is better than the SR Reference fashion since the scores of the former are all higher than the latter. From the table we can know that those LR frames, which are close to the reference, got the best scores by super-resolving using the proposed model even the baseline experimented under the $\times 2$ scaling factor. However, the results gradually deteriorate as the motion between the LR and the reference becomes large. In the *carphone* sample, the PSNR of the *Single Reference* results becomes lower than the *Baselinex2* after the sixth frame. In the *salesman*, though the PSNRs of all *Single Reference* are higher than the *Baselinex2*, the SSIM score becomes lower after the forth frame.

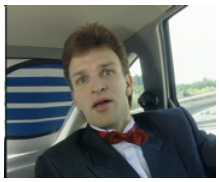
The visual results of the *Single Reference* fashion are showed in Figure 14. The frame at the upper left corner refers to the first frame

and the frame at the bottom right corner refers to the last frame. We observed that the frames are gradually blurred. When it comes to the last frame, the degree of the deterioration becomes quit large. Trough the visual estimation, for the both videos, we can say that the frames are well super-resolved until the fifth frame visually.

Since our proposed SR method refers to the reference frame based on the optical flow between the LR and the reference, the frames before the reference could also be super-resolved using the reference. Thus with a single reference, we can obtain about 10 well super-resolved frames both before and after. From this exploration, we validated the practicability for the video super resolution.

Table 3: Quantitative comparison for the two fashions of video super resolution.

| Method | Baselinex4 | | Baselinex2 | | SR Reference | | Single Reference | |
|---|---|---|---|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| carphone_01 | 27.54 | 0.8608 | 35.57 | 0.9735 | **41.81** | **0.9905** | **41.81** | **0.9905** |
| carphone_02 | 27.41 | 0.8592 | 35.49 | 0.9734 | 36.83 | 0.9789 | **37.58** | **0.9823** |
| carphone_03 | 27.41 | 0.8558 | 35.56 | 0.9738 | 34.92 | 0.9689 | **37.07** | **0.9796** |
| carphone_04 | 27.48 | 0.8563 | 35.57 | 0.9734 | 34.34 | 0.9627 | **36.83** | **0.9777** |
| carphone_05 | 27.44 | 0.8578 | 35.56 | 0.9735 | 32.95 | 0.9508 | **36.65** | **0.9770** |
| carphone_06 | 27.45 | 0.8564 | 35.66 | **0.9741** | 32.24 | 0.9417 | **35.96** | 0.9726 |
| carphone_07 | 27.54 | 0.8543 | **35.76** | **0.9744** | 31.32 | 0.9253 | 33.83 | 0.9534 |
| carphone_08 | 27.62 | 0.8570 | **35.84** | **0.9743** | 30.49 | 0.9103 | 32.91 | 0.9366 |
| carphone_09 | 27.73 | 0.8621 | **36.02** | **0.9755** | 30.31 | 0.9034 | 33.00 | 0.9311 |
| carphone_10 | 27.72 | 0.8609 | **35.90** | **0.9749** | 30.09 | 0.8971 | 32.57 | 0.9242 |
| Average | 27.53 | 0.8581 | 35.70 | **0.9741** | 33.53 | 0.9430 | **35.82** | 0.9625 |
| salesman_01 | 26.85 | 0.7050 | 31.58 | 0.9048 | **37.40** | **0.9732** | **37.40** | **0.9732** |
| salesman_02 | 26.89 | 0.7063 | 31.58 | 0.9049 | 34.27 | 0.9461 | **34.67** | **0.9496** |
| salesman_03 | 26.89 | 0.7039 | 31.62 | 0.9045 | 32.73 | 0.9213 | **33.25** | **0.9283** |
| salesman_04 | 26.93 | 0.7032 | 31.59 | 0.9035 | 32.05 | 0.9073 | **32.75** | **0.9191** |
| salesman_05 | 26.94 | 0.7020 | 31.64 | **0.9046** | 31.34 | 0.8890 | **32.15** | 0.9040 |
| salesman_06 | 27.03 | 0.7076 | 31.67 | **0.9050** | 30.76 | 0.8781 | **32.17** | 0.9022 |
| salesman_07 | 27.00 | 0.7093 | 31.69 | **0.9053** | 30.13 | 0.8655 | **31.98** | 0.8991 |
| salesman_08 | 27.02 | 0.7117 | 31.62 | **0.9045** | 29.63 | 0.8564 | **31.82** | 0.8969 |
| salesman_09 | 27.00 | 0.7120 | 31.63 | **0.9049** | 29.12 | 0.8441 | **31.70** | 0.8948 |
| salesman_10 | 27.08 | 0.7180 | 31.59 | **0.9045** | 28.67 | 0.8339 | **31.63** | 0.8953 |
| Average | 26.96 | 0.7079 | 31.62 | 0.9047 | 31.61 | 0.8915 | **32.95** | **0.9162** |



Reference

SR

Reference

SR

Figure 14: Visual results of the video super resolution with the Single Reference fashion.

## 5.2 Discussion

We first illustrated the averaged PSNR of the test data during the training procedure. For the $\times 2$ scaling, the proposed models improved the PSNR scores about 2.7dB from the baselines. We also confirmed the the effectiveness of the channel attention mechanism, and noticed that it brought a better improvement on the proposed model than on the baseline. For the $\times 4$ scaling, proposed model showed its superiority over the baseline.

In the quantitative comparison, we observed that most of the test frames had best scores using proposed models, but there still were some undesirable results under the scaling factor of $\times 2$. From the visual evaluation we found that the reference sometimes could obstruct the performance of the super resolution. Several reasons could be considered:

1. Intense movement of the camera will bring intense visual change, which leads to a large motion. Large motions may bring difficulty for the optical flow calculation.
2. The performance of the FlowNet may not be good enough. Since we utilized the FLowNetS, which is a simple one, a poor performance of the optical flow calculation may restrict the performance of super resolution.
3. For the simplicity, we utilized a very simple 2-convolution-layer decoder. With a more powerful decoder, the interference of the inaccurate optical flow may be able to be avoided.

Finally, we assumed a video compression system and explored the practicability of the proposed model on video super resolution. In the experiment, we verified the Single Reference fashion is the better way to conduct video super resolution. With the both quantitative and visual estimation, we verified that well super-resolved frames could be obtained within 5 frames to the reference frame, i.e., 10 well super-resolved could be obtained. A large degree of deterioration appears

when the motion becomes large. This may be for the inaccurate optical flow. We used an upscaled image of the LR frame to calculate the optical flow through the FlowNet. However, with a high scaling factor of $\times 4$, the upscaled image will be very vague resulting in a blurred optical flow map. In spite of this, we still confirmed the feasibility and effectiveness of the video super resolution.

# 6 Conclusion

## 6.1 Conclusion of the paper

In this paper, we proposed a reference-based super resolution method. By introducing the optical flow, we successfully made use of the reference frame in the super resolution.

The proposed model could be divided into three parts. The first is the optical flow calculation and multi-scale warping, the second is the resisual feature extraction and the third is the encoding and decoding. We calculate the multi-scale optical flows via the FlowNet and warp the multi-scale encoded feature maps with the optical flows. We exploit the Residual in Residual architecture as the feature extractor to extract deep features. Finally, with an upsample layer and the decoder, the super-resolved target frame will be obtained from the output of the network.

In the experiments, we evaluated the proposed model with both quantitative and visual estimation under the scaling factor of $\times 2$ and $\times 4$ respectively. We found that the proposed model has superior performance under the very low resolution condition. In the experiment of the video super resolution, we verified the practicability of the proposed model.

## 6.2 Future Works

There are several under-resolved problems of the proposed method.

1. The proposed method still could not well applicable with the large motions. The following solutions can be considered: 1) Improve the performance of the FlowNet. Since the FlowNetS is not the best model to calculate the optical flow and more high performance models have been proposed, using a better model may help to obtain more accurate optical flow maps. 2) Improve the decoder. The same as the above, we used a very simple

decoder in the network. Since there are many situations that could lead to the inaccurate optical flows, simple decoder may not able to recover such error. With a more powerful decoder, the interference of the inaccurate optical flow may be able to be avoided.

2. Due to the limitation of storage, we could not train the *Proposal+CA* model under the scaling factor of $\times 4$. Though the training speed is not very slow, a lower computation cost means that we can add additional mechanism to improve the performance of the model. Hence we are considering reducing the computation cost of the model. Here is a solution of using Depthwise Separable Convolution [4] instead of the convolution.
3. Improvement of the model. In the comparison experiment under the scaling factor of $\times 2$, we confirmed the effectiveness of channel attention mechanism. Hence the attention mechanism could be considered to improve the model. The channel attention proposed in [18] is a kind of self-attention mechanism. Since we have a high-resolution reference frame, a better attention may be able to improve the model better than the self-attention.
4. Comparison with the existing reference-based SR methods. In the experiments, we only compared the proposed method with the baseline method, which is a single image super resolution method. For a better comparison, we should compare the proposed method with those reference-based methods. However, with different task settings and implement reasons, we did not conduct the comparisons. We hope to find a fair way to conduct the comparison with existing reference-based methods in the future.

# References


[1] "https://media.xiph.org/video/derf/".

[2] V. Boominathan, K. Mitra, and A. Veeraraghavan. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In *2014 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10, May 2014.

[3] 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. 11 2016.

[4] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[6] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[7] Haoqian Wang Yebin Liu Lu Fang Haitian Zheng, Mengqi Ji. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. 2018.

[8] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *Advances in Neural Information Processing Systems*, pages 235–243, 2015.

[9] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[12] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[13] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[14] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun 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*, pages 136–144, 2017.

[15] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, Xinchao Wang, and Thomas S Huang. Learning temporal dynamics for video super-resolution:

A deep learning approach. *IEEE Transactions on Image Processing*, 27(7):3432–3445, 2018.

[16] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[18] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018.

[19] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *arXiv:1903.00834v1*, 2019.

[20] Zhifei Zhang, Zhaowen Wang, Zhe L. Lin, and Hairong Qi. Reference-conditioned super-resolution by neural texture transfer. *ArXiv*, abs/1804.03360, 2018.

[21] Haitian Zheng, Mengqi Ji, Lei Han, Ziwei Xu, Haoqian Wang, Yebin Liu, and Lu Fang. Learning cross-scale correspondence and patch-based synthesis for reference-based super-resolution. In *BMVC*, 2017.

## Acknowledgment

First of all, I would like to extend my sincere gratitude to my supervisor, Prof. Shinichiro Omachi, Graduate School of Engineering, Tohoku University, for his instructive advice and useful suggestions on my thesis. I am deeply grateful of his help in the completion of this thesis.

Also, I would like to express my gratitude to Prof. Akinori Ito, Graduate School of Engineering, Tohoku University, and Prof. Ayumi Shinohara, Graduate School of Information Science, Tohoku University, for their valuable comments on this thesis.

I would like to express my sincere thanks to Associate Professor Yoshihiro Sugaya, Graduate School of Engineering, Tohoku University, and Assistant Professor Tomo Miyazaki for giving me polite and enthusiastic guidance in carrying out this research.

Special thanks should go to the members of Omachi Lab who gave me a lot of discussions, advice, and enjoyable daily life.

Finally, I am deeply indebted to all my friends and family members for their continuous support and encouragement.

# Achievement

## Conference Presentation

Xi Huang, Tomo Miyazaki, Yoshihiro Sugaya, Omachi Shinichiro
*Multi-Frame Super Resolution Using 3D Convolution and RNN Prediction.*
2018 Tohoku-Section Joint Convention of Institutes of Electrical and Informatin Engineers, Japan

Xi Huang, Tomo Miyazaki, Yoshihiro Sugaya, Omachi Shinichiro
*Super Resolution for Multi Frames with 3D Feature Extraction and RNN Prediction.*
International Academic Conference 2019 International Symposium on Signal Processing Systems

# Awards

IEEE Sendai Section Student Awards 2018, The Encouragement Prize.