MetaVSR: A Novel Approach to Video Super-Resolution for Arbitrary Magnification

Zixuan Hong[1,2], Weipeng Cao[2(✉)], Zhiwu Xu[1], Zhenru Chen[2], Xi Tao[2],
Zhong Ming[1,2], Chuqing Cao[3], and Liang Zheng[3]

[1] College of Computer Science and Software Engineering, Shenzhen University,
Shenzhen 518060, China

[2] Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen),
Shenzhen 518107, China
caoweipeng@gml.ac.cn

[3] Anhui Province Key Laboratory of Machine Vision Inspection,
Yangtze River Delta HIT Robot Technology Research Institute, Wuhu 241000, China

Abstract. Video super-resolution is a pivotal task that involves the recovery of high-resolution video frames from their low-resolution counterparts, possessing a multitude of applications in real-world scenarios. Within the domain of prevailing video super-resolution models, a majority of these models are tailored to specific magnification factors, thereby lacking a cohesive architecture capable of accommodating arbitrary magnifications. In response to this lacuna, this study introduces "MetaVSR", a novel video super-resolution model devised to handle arbitrary magnifications. This model is structured around three distinct modules: inter-frame alignment, feature extraction, and upsampling. In the inter-frame alignment module, a bidirectional propagation technique is employed to attain the alignment of adjacent frames. The feature extraction module amalgamates superficial and profound video features to enhance the model's representational prowess. The upsampling module serves to establish a mapping correlation between the desired target resolution and the input provided in lower resolution. An array of empirical findings attests to the efficacy of the proposed MetaVSR model in addressing this challenge.

Keywords: Video super-resolution · Deep learning · Neural network

1 Introduction

The domain of computer vision extensively employs super-resolution technology to achieve the transformation of low-resolution images into high-resolution counterparts. This technology finds widespread application across various domains, including but not limited to medical imaging [20], surveillance [2], and security [1]. As it stands, it has solidified its position as a cornerstone within the realm of computer vision methodologies.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
S. Rudinac et al. (Eds.): MMM 2024, LNCS 14554, pp. 300–313, 2024.
https://doi.org/10.1007/978-3-031-53305-1_23

Conventional up-sampling methods predominantly rely on interpolation techniques to enhance image resolution. Examples include nearest-neighbor, bilateral, and bicubic interpolations. However, these traditional methodologies often engender the predicament of image blurring, thereby constraining their overall efficacy within practical applications. In light of these limitations, the emergence of deep learning has ushered in a revolutionary shift in super-resolution techniques.

Deep learning methodologies have accomplished more than simply mitigating the blurring issue that plagues conventional techniques. These methods have introduced a realm of visually arresting effects. Consequently, deep learning-based super-resolution models exhibit the proficiency not only to rectify the challenge of image blurring endemic to traditional techniques but also to engender breathtaking visual enhancements. It follows that these models underpinned by deep learning can effectively and accurately restore high-resolution images, thus effecting a substantial and noteworthy enhancement in the visual quality of images.

The application of neural networks in the field of imaging provides efficient solutions to computer vision problems. SRCNN [6] and VSRnet [12] perform well in Single-Image Super-Resolution (SISR) and Video Super-Resolution (VSR) tasks, laying the foundation for subsequent research. The difference between video and image lies in the fact that video contains more spatiotemporal information, and simply super-resolution video frames may lead to artifacts. Therefore, designing an effective frame alignment method has a significant impact on the final video super-resolution results. In recent years, many VSR methods [3–5,8,10,23,27–29] have been proposed. They focus on video super-resolution at specific magnification and require retraining the network when other magnification super-resolution is required. This process is time-consuming and limits the practical application in different scenarios. In addition, existing VSR methods usually adopt a simple convolutional layer stacking approach for feature extraction, which fails to fully utilize shallow and deep feature information, thus affecting performance.

Therefore, in an attempt to overcome the above mentioned limitations of existing models, this paper proposes MetaVSR, a model that can realize video super-resolution at arbitrary magnification. It mainly consists of inter-frame alignment, feature extraction, and upsampling modules. In the inter-frame alignment module, we use a bidirectional propagation method to achieve frame alignment of neighboring frames in a pyramid structure, which provides a better frame alignment effect to improve the model's performance. In the feature extraction module, we utilize a dense residual block to fuse the shallow and deep feature information to obtain a more detailed representation of video frames for better feature learning capability. The upsampling module generates the final super-resolution results by establishing the mapping relationship between the target and low resolutions and calculating the corresponding weights. We evaluate the proposed model through extensive experiments and test it on widely used benchmark datasets. The experimental results show that our method is remarkably effective and produces excellent visual results.

In summation, this study presents a collection of substantive contributions, encapsulated as follows:

– We introduce a novel model named MetaVSR, adept in effectuating video super-resolution for arbitrary magnification scenarios. Comprising inter-frame alignment, feature extraction, and upsampling modules, this model constitutes a pioneering stride toward enhancing video quality.
– To refine the alignment process, we proffer a novel approach-bidirectional frame alignment employing a pyramid structure. By amplifying information interchange among neighboring frames, this innovative methodology bolsters alignment precision, thereby serving as a cornerstone for optimizing video super-segmentation outcomes.
– We advocate the deployment of dense residual blocks to amalgamate shallow and profound feature information, thereby engendering a more intricate feature representation. The ultimate super-resolution outputs materialize through the establishment of a mapping nexus between target and low resolutions. Extensive experimental results on benchmarks demonstrate the effectiveness of the proposed MetaVSR model.

2 Related Works

2.1 Single Image Super Resolution

The application of neural networks to image super-resolution helps to learn image features to overcome the image blurring problem caused by the traditional up-sampling method. SRCNN [6] represents a turning point within the filed of single image super-resolution, laying the groundwork for subsequent research endeavors. Existing research in the field of single image super-resolution can be categorized primarily into three groups: methods based on residual networks [15,16,31] which leverage their robust feature learning capabilities for performance enhancement; methods employing attention mechanisms [25,30], which elevate the significance of attended regions; and other approaches [7,9,22,26]. These techniques have made remarkable progress in the domain of single-image super-resolution, continually advancing image quality through the exploration of diverse network architectures and technical modalities.

Residual-Based Methods. The approaches based on residual networks integrate image features through skip connections to enhance the network's capacity to learn image characteristics. For instance, in [16], the authors addressed image super-resolution by combining residual learning and sub-sampling layers [24]. However, this method falls short of effectively merging shallow and deep-level image information. As a result, in [31], the authors introduced Residual Dense Blocks (RDBs), which concatenate shallow and deep-level image information along the channel dimension to enhance image learning. Nonetheless, these methods are constrained to feature extraction on a single scale, inevitably leading to the loss of certain local information. Hence, in [15], the authors proposed a method to adaptively extract image features across different scales to achieve superior visual effects.

Attention-Based Methods. Incorporating attention mechanisms into the single-image super-resolution task enables networks to selectively focus on crucial information within features. In [30], the authors introduced the RCAN network, incorporating designed Residual Channel Attention Blocks (RCABs) to introduce channel-wise attention mechanisms, allowing neural networks to better exchange information across different channels, thereby enhancing network performance. In [25], the authors presented a novel architecture that combines RCAN with LSTM to enhance network performance. However, due to attention mechanisms potentially disregarding inter-dimensional correlations, the approach proposed by [19] effectively addressed this issue. This method merges layer-wise attention modules with channel-spatial attention modules and employs a residual block approach to tackle this concern.

Others Methods. In recent years, with the surge in popularity of stable diffusion models, numerous scholars have incorporated such models into the realm of single-image super-resolution tasks [7,22]. These approaches commonly utilize the DDPM and U-Net architectures to accomplish image super-resolution tasks. Nevertheless, all of the aforementioned methods are confined to specific magnification factors. Therefore, in an attempt to achieve arbitrary magnification in single-image super-resolution, meta-learning techniques are used to implement arbitrary magnification super-resolution [9,26].

2.2 Video Super Resolution

With the rapid development of image super-resolution technology, the field of video super-resolution is also indirectly driven. VSRnet [12], based on the concept of SRCNN [6], introduced deep learning to video super-resolution tasks for the first time, which laid the foundation of the field and spawned a series of evolved video super-resolution work. However, videos have more information in the spatio-temporal dimension. Therefore, fully utilizing this spatio-temporal information to achieve better frame alignment results will have a profound impact on the final video super-resolution. Currently, video super-resolution work can be categorized into two main groups: Motion Estimation and Motion Compensation (MEMC) methods [3–5,8,10,23,29] and Deformable Convolution [27,28].

MEMC-Based Methods. The motion information between video frames is estimated using optical flow and applied with corresponding distortions. FRVSR [23] employs a recurrent framework, utilizing motion estimates from previous frames for the current frame. This allows the network to handle information propagation over extensive temporal ranges, resulting in continuous video super-resolution outcomes. TOFlow [29] explores task-oriented motion and utilizes self-supervised and task-specific approaches for motion representation learning. TecoGAN proposed by [5] incorporated spatiotemporal discriminators and a Ping-Pong loss function to achieve coherent video super-resolution results. BasicVSR [3] and BasicVSR++ [4] designed a bidirectional propagation mechanism, enabling each frame to acquire alignment information from a wide range of frames. This approach facilitates the super-resolution of target frames, yielding significantly enhanced visual effects. STARnet [8] leveraged the temporal-spatial correlations to provide more accurate motion compensation information, leading to more precise alignment results. In [10], the authors designed recurrent detail structure blocks and hidden state adaptation blocks and then combined them through unidirectional propagation to finally enhance the quality of video super-resolution.

Deformable-Convolution-Based Methods. Due to MEMC's shortcomings in dealing with large-scale motion and occluded regions, deformable convolution has been introduced. By calculating offset values, the network can more accurately capture local information, thereby enhancing the performance of traditional convolutions. EDVR [28] combined pyramid structures with deformable convolutions to achieve video frame alignment across different scales, resulting in more accurate alignment outcomes and improved performance. TDAN [27] utilized deformable convolutions and offset values at the feature level of video frames within the same scale for alignment.

3 Proposed Method

3.1 Overview

Given (2N+1) video frames $I_t^{LR}, t \in [-N, N]$, as an example, we notate the middle frame as the target frame I_0^{LR} and the rest of the frames as neighboring frames. Our goal is to utilize the proposed video super-resolution model to reconstruct the high-resolution target frame I_0^{HR} by using the information of the neighboring frames, which is given to our target frame by bidirectional propagation strategy. The proposed high-resolution network architecture MetaVSR can be seen shown in Fig. 1, which is mainly composed of inter-frame alignment, feature extraction, and upscale modules. In the inter-frame alignment module, we use a bidirectional propagation strategy to achieve frame alignment of neighboring frames to provide better super-resolution results. The feature extraction module utilizes dense residual blocks to form a continuous memory mechanism to form multilevel features and fuses shallow and deep information to obtain better video frame feature learning capability. The upsampling module generates the final super-resolution results by establishing the mapping relationship between the target resolution and the lower resolution and calculating the corresponding positional weights.

3.2 Inter-frames Alignment

In the inter-frame alignment module, we introduce bidirectional propagation for information transfer between neighboring and target frames, improving alignment results. For optical flow computation, we use the pre-trained SpyNet [21] module denoted as S, combining convolutional neural networks with pyramid structures to efficiently compute accurate flow information across scales. Such an optical flow computation is not only more accurate, but also capable of capturing motion information at different scales, thus providing more comprehensive information to support the frame alignment process.

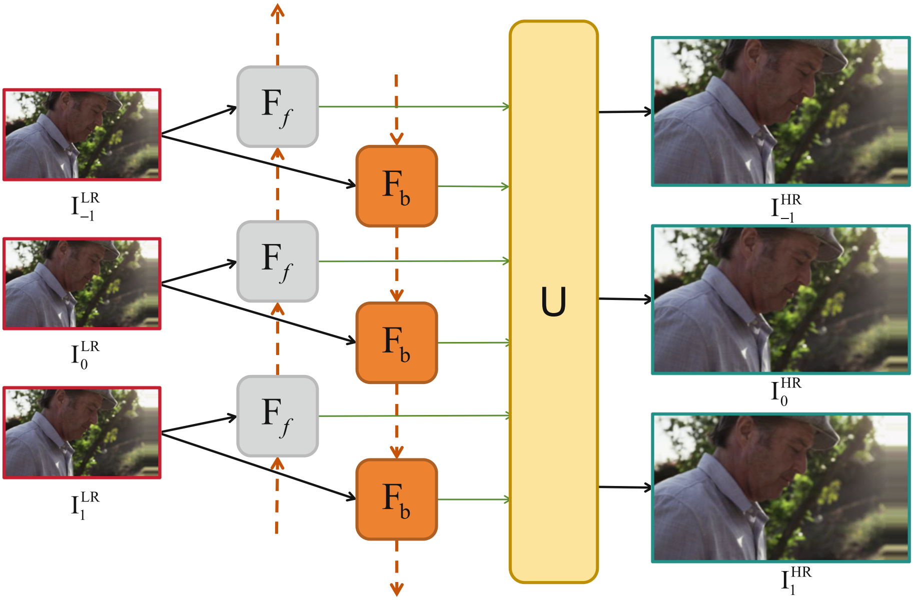

Fig. 1. This figure shows the network architecture of MetaVSR. It is a bidirectional propagation architecture, where F_f, F_b, and U are the forward propagation, backpropagation, and upscale modules, respectively. The architecture of forward propagation and backpropagation is the same, which mainly consists of inter-frame alignment module and feature extraction module, and they differ only in the input information of the modules.

In the propagation process, it is divided into two directions, forward propagation and backward propagation, which are denoted as F_f and F_b, respectively. Taking the forward propagation process of the target frame I_0^{LR} as an example, in the forward propagation process, we will process the feature information of the previous t frame $t \in [-N, -1]$ through layer-by-layer propagation to generate the feature h_{-1}^{LR} of the previous frame I_{-1}^{LR}, which is then inputted into the forward propagation module. At the same time, the target frames I_0^{LR} and I_{-1}^{LR}

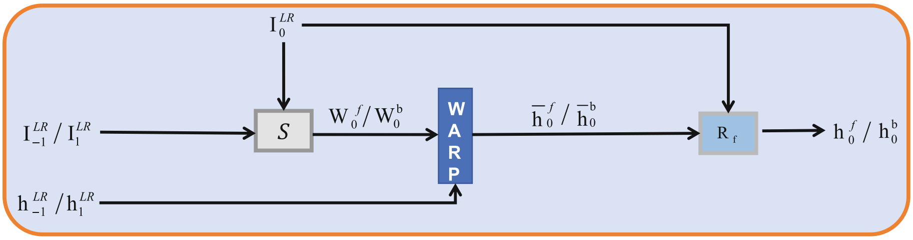
I$_0^{LR}$
I$_{-1}^{LR}$ / I$_1^{LR}$
S
W$_0^f$/W$_0^b$
WARP
$\bar{h}_0^f$ / $\bar{h}_0^b$
R$_f$
h$_0^f$ / h$_0^b$
h$_{-1}^{LR}$ /h$_1^{LR}$

Fig. 2. This figure shows the structure of the bidirectional propagation F_f and F_b, which are identical. The features h_{-1} and h_1 from the previous frame propagation are input into the forward and backward propagation, respectively, and the corresponding low-resolution frames are input into module S for optical flow computation to generate the optical flow estimate $W_0^{f,b}$. Subsequently, $W_0^{f,b}$ and $h_{0\pm1}^{LR}$ are utilized for warping to generate $\bar{h}_0^{f,b}$ results. This is followed by R_f for feature extraction.

are inputted into the SpyNet network to predict the optical flow estimation W_0^f between them.

$$W_0^f = S(I_0^{LR}, I_{-1}^{LR}) \tag{1}$$

The optical flow estimates W_0^f and h_{-1}^{LR} are subsequently utilized to obtain the feature $\bar{h}_0^f$ aligned with the current frame after a warping operation.

$$\bar{h}_0^f = warp(W_0^f, h_{-1}^{LR}) \tag{2}$$

Finally, it is input to the module R_f for feature extraction, and the generated features of the current frame, h_0^f, are input to the forward propagation process of the next target frame. As for the process of backward propagation, its main input is the features of the latter t frames after propagation, $t \in [1, N]$, and its overall process is the same as that of forward propagation. For the flowchart of the propagation process, it can be seen as shown in Fig. 2.

$$h_o^f = R_f(I_0^{LR}, \bar{h}_0^f) \tag{3}$$

3.3 Feature Extraction Module

In the feature extraction module R_f, we fully borrow the design idea of Residual Dense Blocks (RDBs) from [31]. The uniqueness of this block is that it can efficiently fuse shallow and deep information to effectively capture multi-level features of an image. By interrelating multiple residual blocks in a densely connected manner, we achieve dense extraction of local features and also establish a continuous information transfer mechanism. This allows the current RDB block to be organically linked with the previous blocks, which in turn stabilizes the entire network training process. However, since the feature extraction module inputs the current frame I_0^{LR} together with the feature $\bar{h}_0^f$ aligned to it, we need to stitch them in the channel dimension beforehand, followed by fusion through

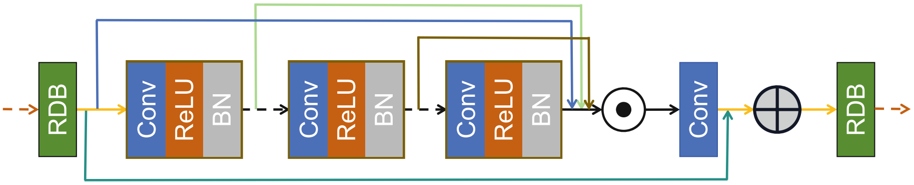

Fig. 3. This figure shows the structure of the feature extraction module. With the help of RDBs block, the features at all levels are fused to utilize the shallow and deep information to form a continuous memory extreme for the network's richer features for learning.

a convolutional layer to maintain feature consistency and integration. In addition, the introduction of RDB increases the depth of the network, which may bring about the problem that the model is difficult to converge. For this reason, we introduce Batch Normalization after each convolutional block to enhance the training speed and stability of the model. For more details on the feature extraction module, see Fig. 3.

3.4 Upscale Module

After bidirectional propagation and feature extraction, we obtain $\bar{h}_0^f$ and $\bar{h}_0^b$, respectively. In order to further fuse these features, we concat them according to the channel dimensions and then fuse them by convolutional layers to obtain the feature map F^{LR}. To realize arbitrary magnification super-resolution, inspired by MetaSR [9], we design a mapping relationship between target resolution and low resolution. By calculating the corresponding weights, we are able to obtain the final target super-resolution results. Its structure can be seen as shown in Fig. 4.

In the target resolution-low resolution mapping relationship, for each pixel (i, j) in the target resolution, the coordinates $(i^{'}, j^{'}) = (floor(\frac{i}{r}), floor(\frac{j}{r}))$ in the low-resolution image are generated by inverse mapping according to the selected scaling factor r and constructed into a position matrix by using its offset.

$$PositionMatrix = \begin{pmatrix} R(0) & R(0) \\ R(0) & R(1) \\ R(0) & R(2) \\ \ldots & \ldots \\ R(i) & R(j-1) \\ R(i) & R(j) \\ R(i) & R(j+1) \\ \ldots & \ldots \end{pmatrix} \tag{4}$$

where $R(i) = \frac{i}{r} - floor(\frac{i}{r}), R(j) = \frac{j}{r} - floor(\frac{j}{r})$, and the size of position matrix is $HW \times 2$. Then, Position matrix is input to the weight prediction network ϕ, and the weight $W(i, j)$ corresponding to each position of the target resolution image can be obtained.

$$W(i,j) = \phi(PositionMatrix, \theta) \tag{5}$$

where θ is the weight value of the weight prediction network. After obtaining the corresponding weights, the feature results can be obtained by multiplying the weights with the corresponding positions on the low-resolution image.

$$I_0^{HR}(i,j) = F^{LR}(i',j')\,W(i,j) \tag{6}$$

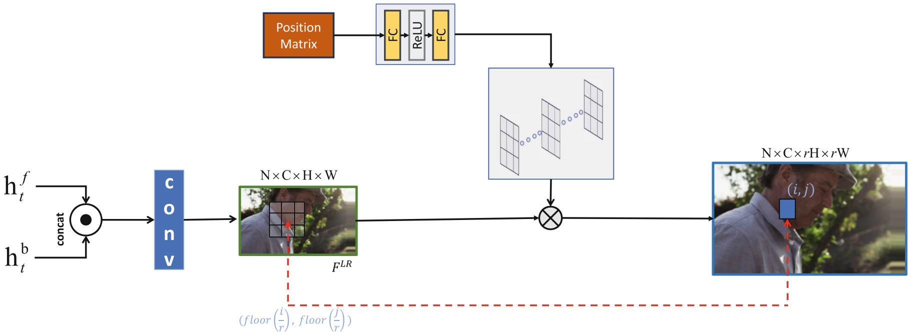

Fig. 4. Inspired by MetaSR's meta-upsampling module [9]. This figure illustrates the structure of the upsampling module. This involves establishing a mapping between target and low resolutions, and calculating positional weights to derive feature values for each pixel in the target resolution.

3.5 Loss Function

In the training process, we mainly use two loss functions, Charbonnier Loss [14] and VGG Loss [11], denoted as L_{ch} and L_{vgg}, respectively, so the overall loss function is shown in Eq. 7:

$$L(I_0^{HR}, I_0^{GT}) = L_{ch}(I_0^{HR}, I_0^{GT}) + \lambda L_{vgg}(I_0^{HR}, I_0^{GT}) \tag{7}$$

where $\mathcal{L}_c(x,y) = \sqrt{\|x-y\|^2 + \epsilon^2}$, ϵ and λ are constants. In our all experiments, we set ϵ and λ to 1e−6 and 1e−4, respectively.

4 Experiments

4.1 Implementation Details

Training Datasets. During the training phase, we employed the Vimeo90K [29] dataset with a resolution of 256 × 448 as our training set. To assess the performance of our model, we conducted testing using the Vimeo90K, REDS4 [18], and Vid4 [17] datasets. We quantified the performance using metrics such as PSNR and SSIM.

Training Details. We generate a set of scaling factors ranging from 1 to 4 with an increment of 0.1 and set the batch size to 8, while keeping the dimensions of the low-resolution frames fixed at 50×50. During each iteration, a target scaling factor is randomly selected from the set of scaling factors, and the target cropping size is obtained by multiplying the selected scaling factor with the pre-set low-resolution size, and then the video frame is randomly cropping according to the target cropping size, and the cropping video frame is downsampled by bicubic. For optimization, we use the Adam optimizer [13], setting the parameters β_1 and β_2 to 0.9 and 0.99, respectively. The initial learning rate is set at 1e−4, and we employ PyTorch's CosineAnnealing function to adjust the learning rate every 300 epochs, with a minimum rate of 1e−6. In our experiments, the number of blocks in our RDBs is 16, with 8 layers in each block.

4.2 Comparisons with State-of-the-Art Methods

We compare our model with SOTA models, e.g., EDVR [28], STARnet [8], RSDN [10], and calculate the corresponding PSNR and SSIM evaluation metrics. All methods are tested and compared on three datasets Vimeo90K [29], REDS4 [18], and Vid4 [17]. And all of them use the Bicubic downsampling method to generate low-resolution images and perform experiments with 4x super-resolution. The quantitative metrics of the experimental results can be seen as shown in Table 1. Note that in the experimental results, the values in red and blue represent the best and the second-best results, respectively. PSNR and SSIM metrics are represented by PSNR/SSIM. The above experimental results demonstrate that compared to EDVR, STARnet, and RSDN, our method shows a small improvement in both PSNR and SSIM on the Vimeo90K dataset. On the REDS4 and Vid4 datasets, the quantization results are comparable to their results. Thus, in general, our method is comparable to the results of the compared methods in terms of metrics computation for Vimeo90K and REDS4 compared to the three chosen methods, with a small improvement especially on the Vimeo90K dataset. In addition, our method is effective in providing a better description of the contours of individual objects in the scene compared to other methods. And to better illustrate the effectiveness of the methods, Fig. 5 shows the corresponding visual comparison results, and it can be seen that our method is able to obtain better visual results in general.

Table 1. The table shows the quantitative metrics results obtained by performing 4x super-resolution video on the Vimeo90K, REDS4, and Vid4 datasets.

		bicubic	EDVR	RSDN	STARnet	Our
Vimeo90K	RGB	29.97/0.867	33.32/0.928	33.80/0.923	33.99/0.923	34.90/0.936
	Y-channel	31.32/0.887	35.10/0.942	35.81/0.946	35.80/0.945	36.16/0.948
REDS4	RGB	24.95/0.718	27.24/0.811	27.93/0.812	27.58/0.804	27.70/0.809
	Y-channel	26.28/0.743	29.58/0.829	29.27/0.821	29.95/0.830	29.59/0.819
Vid4	RGB	20.67/0.604	25.19/0.795	25.52/0.804	25.57/0.805	24.66/0.788
	Y-channel	22.05/0.628	26.62/0.802	27.01/0.821	26.87/0.813	25.80/0.787

In addition, our method is able to realize the video super-resolution task at arbitrary magnification, rather than a super-resolution network for a specific magnification, and thus it is possible to realize video super-resolution at arbitrary magnification using only our trained network. Therefore, in order to better illustrate the effectiveness of our method for arbitrary magnification super-resolution, we performed the video super-resolution task at 2x and 3x respectively, and the experimental results can be seen as shown in Figs. 6 and 7.

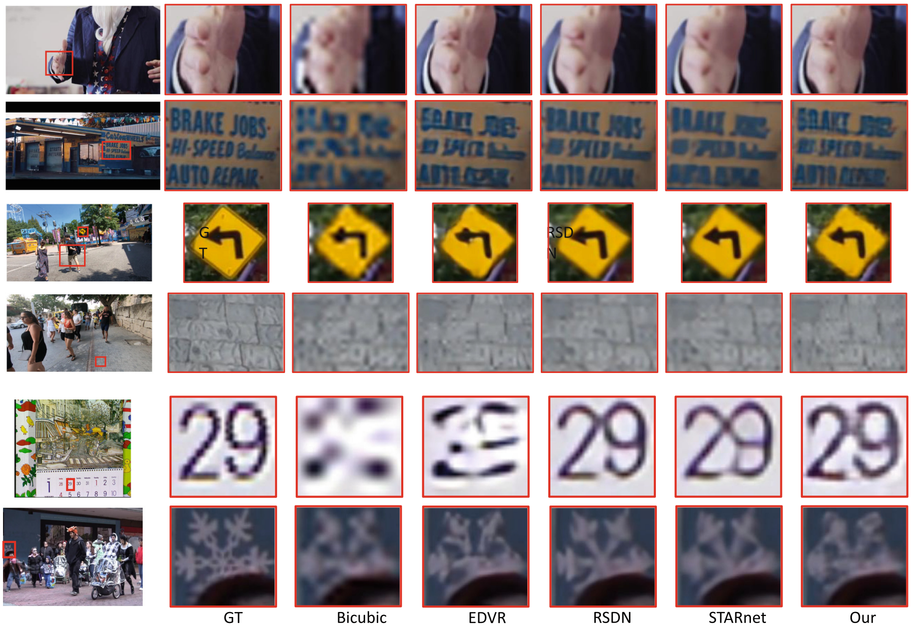

Fig. 5. This figure shows the visualization of our method with 4x super-resolution on the dataset. From top to bottom are Vimeo90K, REDS4 and Vid4, two representative video frames were selected for each dataset. In the first, second and fourth rows, our method shows better restoration for text and texture. In the fifth line, there is a better description of the contours of the object contours.

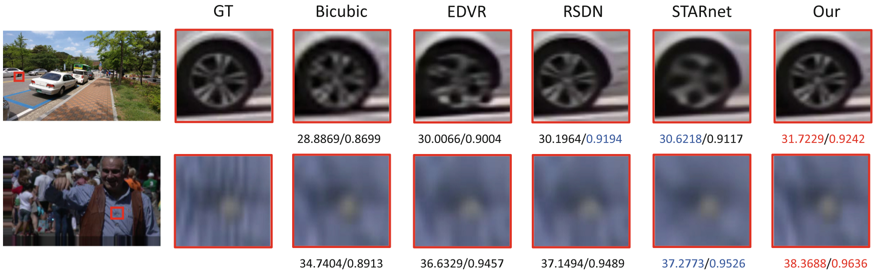

Fig. 6. This figure shows the results of performing a 2x super-resolution experiment and calculating the corresponding PSNR and SSIM metrics. Our method is able to obtain better metrics results and has a better description of the contours of the objects, e.g., the restoration of the tire contours and details in the first row.

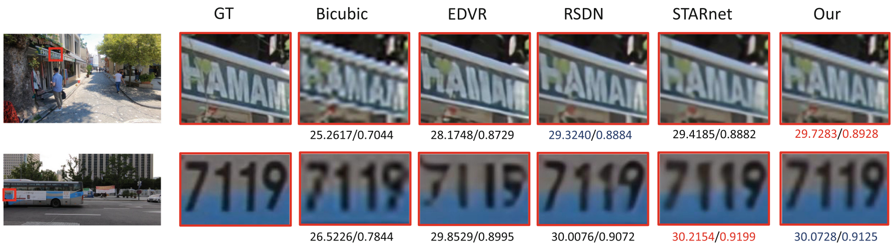

Fig. 7. This figure shows the results of performing a 3x super-resolution experiment and calculating the corresponding PSNR and SSIM metrics. Our method yields relatively smooth and textured results for text reduction.

5 Conclusions

To rectify the issue of necessitating network retraining for specific super-resolution magnifications-resulting in substantial temporal expenditures and confining the method's applicability within practical contexts—this paper introduces an innovative arbitrary magnification video super-resolution model named MetaVSR. The MetaVSR model is designed to establish a robust video super-resolution framework through the realization of inter-frame alignment via bidirectional propagation, the fusion of deep and shallow feature information, and computations grounded in the mapping relationship between the target and low-resolution frames. Extensive experimental validation substantiates our approach, revealing commendable quantitative metrics and visually appealing outcomes across widely accessible benchmark datasets. Additionally, our technique boasts expeditious inference speeds, versatility across various prevalent scenarios, and the attainment of favorable visual results.

Nevertheless, it is worth noting that the visual efficacy of the method still requires enhancement, particularly when confronted with intricate scenes featuring distant subjects. The recognized limitations of this model pave the way for meaningful avenues of research in the forthcoming endeavors. Indeed, we hold a conviction that the MetaVSR model is poised to exhibit even more remarkable performance in the realm of video reconstruction, offering promising prospects for the future.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (62106150, 61836005, 62372304) and the Open Research Fund of Anhui Province Key Laboratory of Machine Vision Inspection (KLMVI-2023-HIT-01).

References

1. Cao, W., Wu, Y., Chakraborty, C., Li, D., Zhao, L., Ghosh, S.K.: Sustainable and transferable traffic sign recognition for intelligent transportation systems. IEEE Trans. Intell. Transp. Syst. **24**, 15784–15794 (2022)

2. Cao, W., Zhou, C., Wu, Y., Ming, Z., Xu, Z., Zhang, J.: Research progress of zero-shot learning beyond computer vision. In: Qiu, M. (ed.) ICA3PP 2020. LNCS, vol. 12453, pp. 538–551. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60239-0_36
3. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: BasicVSR: the search for essential components in video super-resolution and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4947–4956 (2021)
4. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: BasicVSR++: improving video super-resolution with enhanced propagation and alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5972–5981 (2022)
5. Chu, M., Xie, Y., Leal-Taixé, L., Thuerey, N.: Temporally coherent GANs for video super-resolution (tecogan). arXiv preprint arXiv:1811.09393 (2018)
6. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)
7. Gao, S., et al.: Implicit diffusion models for continuous super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10021–10030 (2023)
8. Haris, M., Shakhnarovich, G., Ukita, N.: Space-time-aware multi-resolution video enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2859–2868 (2020)
9. Hu, X., Mu, H., Zhang, X., Wang, Z., Tan, T., Sun, J.: Meta-SR: a magnification-arbitrary network for super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1575–1584 (2019)
10. Isobe, T., Jia, X., Gu, S., Li, S., Wang, S., Tian, Q.: Video super-resolution with recurrent structure-detail network. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 645–660. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_38
11. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
12. Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A.K.: Video super-resolution with convolutional neural networks. IEEE Trans. Comput. Imaging **2**(2), 109–122 (2016)
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 624–632 (2017)
15. Li, J., Fang, F., Mei, K., Zhang, G.: Multi-scale residual network for image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 517–532 (2018)
16. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144 (2017)
17. Liu, C., Sun, D.: On Bayesian adaptive video super resolution. IEEE Trans. Pattern Anal. Mach. Intell. **36**(2), 346–360 (2013)
18. Nah, S., et al.: Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)

19. Niu, B., et al.: Single image super-resolution via a holistic attention network. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 191–207. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_12
20. Patwary, M.J., Cao, W., Wang, X.Z., Haque, M.A.: Fuzziness based semi-supervised multimodal learning for patient's activity recognition using RGBDT videos. Appl. Soft Comput. **120**, 108655 (2022)
21. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4161–4170 (2017)
22. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Trans. Pattern Anal. Mach. Intell. **45**(4), 4713–4726 (2022)
23. Sajjadi, M.S., Vemulapalli, R., Brown, M.: Frame-recurrent video super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6626–6634 (2018)
24. Shi, W., et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874–1883 (2016)
25. Shoeiby, M., Armin, A., Aliakbarian, S., Anwar, S., Petersson, L.: Mosaic super-resolution via sequential feature pyramid networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 84–85 (2020)
26. Soh, J.W., Cho, S., Cho, N.I.: Meta-transfer learning for zero-shot super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3516–3525 (2020)
27. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: TDAN: temporally-deformable alignment network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3360–3369 (2020)
28. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: EDVR: video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)
29. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. Int. J. Comput. Vision **127**, 1106–1125 (2019)
30. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301 (2018)
31. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2472–2481 (2018)