




DSCVSR: A Lightweight Video Super-Resolution for Arbitrary Magnification

Zixuan Hong^{1,2}, Weipeng Cao^{2,3} , Zhiwu Xu¹, Zhong Ming^{1,2},
Chuqing Cao³, and Liang Zheng³

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, a fundamental task in the field of computer vision, has gained much attention and performance in recent years. However, since deep learning introduces a large number of parameters, which can result in a large resource overhead, the model cannot be deployed on edge devices. Therefore, in this paper, we design a lightweight video super-resolution model, named Depthwise Separable Convolutional Video Super-Resolution (DSCVSR), which utilizes a continuous memory mechanism by constructing a dense depthwise separable convolutional residual block to fuse the deep and shallow feature information in order to enable the network to better learn the details in the video, and also constructs an information-filling module to solve the problem of information loss brought about by the depthwise separable convolution, as well as designing an information-filling module to solve the problem of information loss brought about by the depthwise separable convolution information loss problem caused by deep separable convolution, and a knowledge distillation loss is designed to migrate the knowledge from the teacher's model to the model to achieve the superscoring results with arbitrary multiplicity. In the experiments, the method is tested on common video datasets, and it is verified that the proposed method can achieve good results with a small number of parameters.

Keywords: Deep learning · Video super-resolution · Knowledge distillation · Lightweight network · Edge device

1 Introduction

As one of the expressions of multimedia, video is widely used in the fields of target detection, video surveillance, and semantic segmentation [31]. Due to the limitations of capture devices, the captured video often suffers from problems

such as distortion, noise and low resolution. And super-resolution is one of its solutions and a fundamental task in the field of computer vision, which has received wide attention and achieved corresponding performance in recent years.

Traditional super-resolution methods primarily use interpolation to enhance resolution, which often leads to image and edge blurring. The advent of deep learning has provided effective solutions for video super-resolution. However, most current video super-resolution models focus on fixed magnification tasks. When other magnification scale are needed, the network must be retrained, resulting in significant time costs. Currently, there are relatively few algorithms that address arbitrary video super-resolution, with notable examples including MetaVSR [12] and VideoINR [8].

Although deep learning is effective in providing outstanding performance for video super-resolution tasks, its computational effort due to a large number of parameters results in a large resource overhead and thus cannot be deployed on a limited number of edge devices [2, 3]. To be able to apply such deep learning models to edge devices, the first and foremost method is to compress the models. Among the common model compression methods, the main ones are: pruning, quantization, low-rank decomposition, knowledge distillation, compact networks, and a mixture of the above methods. Although the above methods are effective in reducing the number of parameters in the model, they result in a loss of model performance, so additional strategies need to be designed to recover the model performance.

To achieve a video super-resolution model with fewer parameters and the ability to handle arbitrary scaling, this paper proposes the Depthwise Separable Convolutional Video Super-Resolution (DSCVSR) model based on MetaVSR. By designing Dense Depthwise Separable Convolution Blocks (DSCBs), the model integrates both shallow and deep feature information to better learn local and global details in video frames. Additionally, an information filling module is designed to compensate for information loss caused by depthwise separable convolutions. Knowledge distillation techniques are employed, using feature-based distillation loss to transfer knowledge from the teacher model, enabling DSCVSR to achieve performance comparable to MetaVSR. Validation results shows that DSCVSR reduces the parameter count by 71.36% compared to MetaVSR while maintaining similar performance.

In summary, the main contribution of this work can be concluded as follows:

- The Depthwise Separable Convolutional Video Super-Resolution (DSCVSR) model is proposed, which is capable of achieving super-resolution performance with arbitrary scale on a small number of parameters.
- A Dense Depthwise Separable Convolutional Residual Block (DSCB) is designed to effectively fuse the deep and shallow feature information with the help of a continuous memory mechanism for the network to learn the local and global information in the video frames.
- Using the knowledge distillation method, a feature-layer-based knowledge distillation loss is designed to effectively transfer the knowledge of the teacher model to the DSCVSR and achieve better quantitative indicators and visual effects.

2 Related Works

2.1 Video Super Resolution

VSRnet [18], based on SRCNN [9], marked the initial application of deep learning techniques in video super-resolution, laying the groundwork for the field and serving as a reference for subsequent developments. VESPCN [1] and DRVSR [26] utilized convolutional neural networks to design corresponding motion compensation modules, enabling precise optical flow calculation for inter-frame alignment. However, this category of methods often overlooks global video information. Therefore, BasicVSR [4] and RealBasicVSR [6] propagate motion information throughout the entire video sequence, allowing each frame to acquire global video information for more accurate optical flow calculation, achieving more precise video alignment effects. Additionally, they leverage this global motion information to address occlusions and lighting variations in video scenes. TOFlow [29] explores task-oriented motion and employs self-supervision and task-specific approaches for motion representation learning. TGA [16] merges temporal information in a hierarchical manner, dividing the video sequence into multiple groups and processing them with attention modules and deep intra-group fusion modules to provide complementary information for reference frames to compensate for lost details. Although the above methods can solve most of the video scenarios, the performance is not good enough for the final video super-segmentation when there is a large motion or occlusion in the video. The deformable convolution-based approach can effectively solve this problem. EDVR [28], the first to introduce deformable convolution into video super-resolution tasks, effectively tackles motion blur and designs a temporal and spatial attention fusion module to address fusion issues. [14] enhances EDVR by introducing a preprocessing module composed of rigid convolution sub-modules and feature enhancement sub-modules, a temporal 3D convolution fusion module, and a reconstruction module based on channel attention, effectively enhancing feature learning and overall performance. DNLN [27] designs a non-local attention module based on deformable convolution and a non-local network alignment module. BasicVSR++ [5], an extension of BasicVSR [4], incorporates second-order network propagation and flow-guided deformable alignment, enhancing the model’s performance.

2.2 Knowledge Distillation

The goal of knowledge distillation is to improve the performance and generalization of student models by transferring knowledge from a complex teacher model to a simplified student model. In [25], it trains the student model using distillation loss, which is primarily composed of cross-entropy controlled by the temperature parameter T . However, since in most previous distillation methods has been mainly focused on the features and loss functions between the same layers, ignoring the neighboring layers, a review mechanism was proposed in [7] that able to utilize the multi-layer information of the teacher model to

guide the student model. The process of review mechanism was also enhanced by utilizing residual learning, Attention-Based fusion module and hierarchical content layering loss function. And in most of today's knowledge distillation, all tend to ignore the impact of other factors on the distillation performance, [10] and [30] proposed their respective schemes to focus on image background, and experiment shown that by adding attention to the image background, the performance of model will increased in the student model. In [21], it proposed to utilize Rank Mimicking and Prediction-guided Feature Imitation to distill a single-stage detector, where the former distills the anchor ordering of the teacher model as a kind of knowledge, and the latter utilizes the prediction discrepancy to guide the effective feature imitation, so that the feature imitation is close to the accurate prediction. Besides, MobileVOS [23] proposed a pixel-by-pixel representation of distillation loss and a simple boundary-aware sampling scheme for accelerating the training convergence, along with a natural generalization utilized the labels of ground-truth as structural information, and interpolated the representation distillation and contrastive learning through hyperparameters to compute the pixel loss, which ultimately can obtain better accuracy.

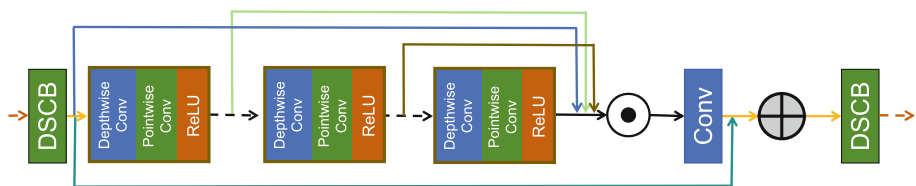


Fig. 1. This is designed for the feature extraction module of the DSCVSR model, which changes all the convolutional layers in the residual dense network into deeply separable convolutions.

3 Proposed Method

3.1 Dense Depthwise Separable Convolutional Residual Block

Depthwise separable convolution has been in the public eye since the introduction of MobileNet [13], a lightweight object detection model. Depthwise separable convolution is mainly composed of depthwise convolution and pointwise convolution, which can better reduce the number of parameters of the traditional convolution. Assuming that the input feature map size is $H \times W \times C_{input}$ and the output feature map size is $H \times W \times C_{output}$, and for the traditional convolution it is necessary to use a convolution kernel size of $D_K \times D_K \times C_{input} \times C_{output}$, the computation of which is shown in Eq. 1.

$$D_K \cdot D_K \cdot M \cdot D_F \cdot D_F \quad (1)$$

where D_K is the size of the convolution kernel, C_{input} is the number of input channels of the feature map, C_{output} is the number of output channels of the feature map, H is the height of the feature map, and W is the width of the feature map. For depthwise separable convolution, on the other hand, the depthwise convolution kernel size is $D_K \times D_K \times 1 \times C_{input}$, and the pointwise convolution kernel size is $1 \times 1 \times C_{input} \times C_{output}$, and thus the computational overhead of depthwise separable convolution can be seen in Eq. 2

$$D_K \cdot D_K \cdot C_{input} \cdot C_{output} \cdot H \cdot W + C_{input} \cdot C_{output} \cdot H \cdot W \quad (2)$$

As a result, the depthwise separable convolution has an overhead reduction in the number of parameters as in Eq. 3 compared to the conventional convolution.

$$\frac{D_K \cdot D_K \cdot C_{input} \cdot C_{output} \cdot H \cdot W + C_{input} \cdot C_{output} \cdot H \cdot W}{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F} = \frac{1}{C_{output}} + \frac{1}{D_K^2} \quad (3)$$

Therefore, we optimized the feature extraction module of the teacher model MetaVSR using depthwise separable convolution blocks, and proposed the Dense Depthwise Separable Convolution Block (DSCB) to reduce the model's parameter count. The designed DSCB not only maintains the original residual dense block's capability for local feature extraction but also establishes a continuous information transfer mechanism, efficiently enabling the network to learn both shallow and deep feature information (as shown in Fig. 1).

3.2 Information Filling Module

Although deep separable convolution can effectively reduce the number of parameters of a model and thus obtain a lightweight model with fewer parameters, the reduction of the number of parameters will inevitably lead to the loss of information caused by the model during the learning process. Inspired by BasicVSR [4], we design an information filling module to provide additional information to compensate for the information loss problem. This module mainly fills additional feature information in an interval manner during the processing of video sequence frames, which can be represented by Eq. 4.

$$h_i^{f,b} = \begin{cases} R_f \left(\left[R \left[h_{i\pm 1}^{f,b}, \bar{h}_i^{f,b} \right], I_i^{LR} \right] \right) & i = 2n + 1 \\ R_f \left(\left[\bar{h}_i^{f,b}, I_i^{LR} \right] \right) & i \neq 2n + 1 \end{cases} \quad (4)$$

where R_f denotes the feature extraction module composed of depthwise separable convolution, R represents the feature extraction module consisting of 15 residual blocks, n is a positive integer, $h_{i\pm 1}^{f,b}$ is the feature propagated in the bidirectional propagation process, $\bar{h}_i^{f,b}$ is the aligned feature obtained from the bidirectional propagation module after the motion estimation and warp operations. I_i^{LR} represents the original low-resolution image of the i frame.

It can be seen that in the process of information filling, the features coming from bidirectional propagation and the features after alignment are spliced and

fused, followed by additional feature extraction through stacked residual blocks, and the resulting features are spliced and fused with the currently processed low-resolution frames through dense depthwise separable convolution residual blocks for feature extraction. The overall framework can be seen as shown in Fig. 2.

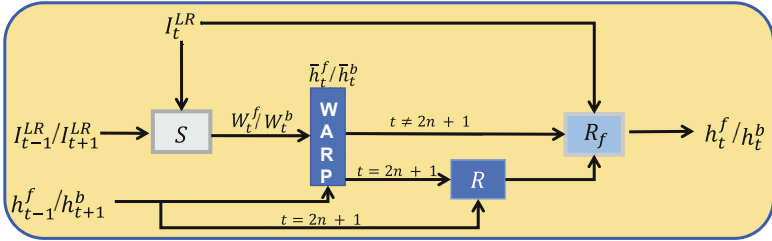


Fig. 2. This figure shows the structure of the bidirectional propagation module of the DSCVSR model and also shows how the information filling module is used.

3.3 Knowledge Distillation Loss

Since in common knowledge distillation, some approaches only compute the corresponding loss between the output of the teacher model and the output of the student model, in order to make the student model have the same performance as that of the teacher model, but in some works it is considered necessary to focus on the output of the intermediate layer of features as well, in order to better migrate the knowledge from the teacher model to the student model. Therefore, we have designed a feature-based knowledge distillation loss that aims to compute the L_1 loss between features in a layer-by-layer way in the feature extraction module of the teacher model and the student model, as shown in the general framework in Fig. 3. It will be able to effectively reproduce the super-resolution performance of the teacher model in the student model at a lower number of parameters by means of this knowledge distillation. Therefore, the final knowledge distillation loss is shown in Eq. 5, where F_T and F_S represent the feature maps of the teacher model and the student model, respectively. n represents the number of the feature extraction blocks in the feature extraction module, and i denotes the output features of the i -th feature extraction block of the two models.

$$\mathcal{L}_{distill}(F_T, F_S) = \frac{1}{n} \sum_{i=1}^n \|F_T^i - F_S^i\| \quad (5)$$

In addition, in order to be able to have visually comparable performance to the teacher model MetaVSR, pixel-level based losses are also introduced by

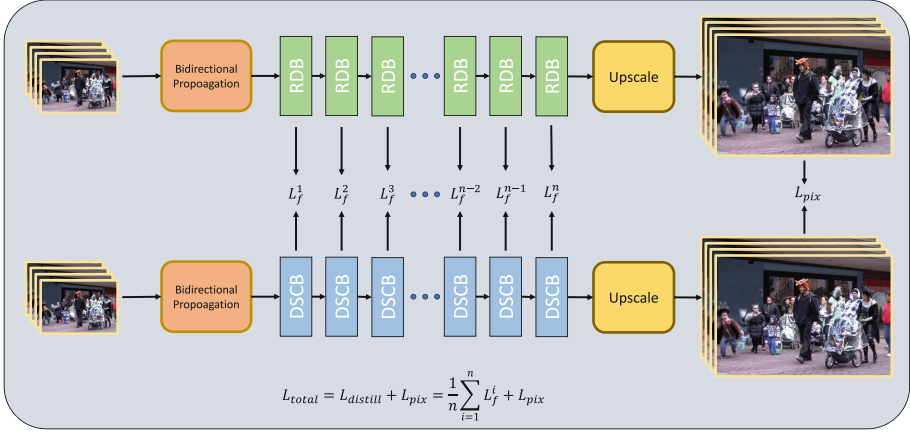


Fig. 3. The figure shows an architectural diagram of knowledge distillation and illustrates the scope of use of the distillation function.

introducing the Charbonnier loss [20] and the VGG loss [17]. As a result, the training loss function is shown in Eq. 6,

$$\mathcal{L}_{total}(F_T, F_S, I_T^{SR}, I_S^{SR}) = \mathcal{L}_{distill}(F_T, F_S) + \mathcal{L}_{ch}(I_T^{SR}, I_S^{SR}) + \lambda \mathcal{L}_{vgg}(I_T^{SR}, I_S^{SR}) \quad (6)$$

where $\mathcal{L}_{distill}$, \mathcal{L}_{ch} , and \mathcal{L}_{vgg} are the distillation loss, the Charbonnier loss, and the VGG loss, respectively. $\mathcal{L}_{ch}(x, y) = \sqrt{\|x - y\|^2 + \epsilon^2}$, where ϵ is a constant set to 1×10^{-6} during training, and I_T^{SR} and I_S^{SR} are the reconstructed high-resolution images of the teacher model and the student model, respectively. λ is a constant set to 1×10^{-4} .

Table 1. The table shows the results of the x4 quantitative metrics of the method on the three datasets, and the proposed DSCVSR is comparable to the MetaVSR performance in terms of performance, with the number of covariates being 71.36% of the MetaVSR model. However, it still performs mediocre on the Vid4 dataset.

| Algorithm | Params(M) | Runtimes(s) | Y-channel | | | RGB-channel | | |
|-----------|-----------|-------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | | Vimeo90K | Vid4 | REDS4 | Vimeo90K | Vid4 | REDS4 |
| | | | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM |
| bicubic | - | - | 31.32/0.8869 | 22.05/0.6280 | 26.28/0.7427 | 29.97/0.8669 | 20.67/0.6045 | 24.95/0.7175 |
| EDVR | 11.46 | 3.01 | 35.10/0.9422 | 26.62/0.8020 | 29.58/0.8290 | 33.32/0.9279 | 25.19/0.7948 | 27.24/0.8105 |
| RSDN | 23.58 | 6.15 | 35.81/0.9451 | 27.01/0.8208 | 29.27/0.8210 | 33.80/0.9230 | 25.52/0.8039 | 27.93/0.8123 |
| STARnet | 408.26 | 4.56 | 35.80/0.9446 | 26.87/0.7933 | 29.95/0.8239 | 33.99/0.9226 | 25.57/0.8047 | 27.58/0.8042 |
| MetaVSR | 91.19 | 3.21 | 36.16/0.9476 | 25.80/0.7874 | 29.59/0.8189 | 34.90/0.9362 | 24.66/0.7779 | 27.70/0.8033 |
| DSCVSR | 25.87 | 6.12 | 35.89/0.9418 | 25.37/0.7889 | 29.63/0.8255 | 34.03/0.9248 | 24.00/0.7713 | 28.01/0.8184 |

4 Experiment

4.1 Implementation Details

In the training phase, we use the Vimeo90K [29] dataset with a resolution size of 256×448 as the training set. The performance of our model is tested using Vimeo90K, REDS4 [24] and Vid4 [22], and quantitative metrics are calculated using PSNR and SSIM.

Firstly, a set of scaling factors from 1 to 4 with a step size of 0.1 is generated, and the batch size for one iteration of training is set to 8, while the resolution frame size for training is fixed to 50×50 . In each iteration of training, a target scaling factor r is randomly selected from the set of scaling factors, and the selected scaling factor is multiplied by the pre-set low-resolution size $50 \times r$ to obtain the target cut size, then the video frames used for training are randomly cut according to this target cut size, and the cut video frames are downsampled using the bicubic interpolation method to obtain the low- and high-resolution training pairs, the low-resolution video frames are inputted into the model, and the output of the over-scored video frames is extrapolated by the model, and the output is compared to the corresponding real high-resolution video frames in a. The loss is calculated to update the weights of our model, in which we utilize the Adam [19] optimizer to update our model with parameters β_1 and β_2 set to 0.9 and 0.99, respectively the initial learning rate is set to $1e-4$, and the learning rate is updated using the CosineAnnealing function in PyTorch with a period of 300 iterations updated with a period of 300 iterations and a minimum learning rate of $1e-6$ and in our experiments, the number of DSCB blocks is the same as the number of RDB blocks in MetaVSR, which is 16, and there are 8 layers in each block.

4.2 Comparisons with State-of-the-Art Methods

In order to more intuitively visualize the difference between DSCVSR and MetaVSR in terms of visual and quantitative results, DSCVSR is compared visually and quantitatively with other methods, such as EDVR [28], RSDN [15], STARnet [11] and MetaVSR [12]. Note that in all the experimental results, red and blue colors are used to denote the first and second best results, respectively. The quantitative metrics of the experimental results can be seen as shown in Table 1, and the quantitative metrics are denoted by PSNR/SSIM. From the experimental results in Table 1, it is known that the proposed DSCVSR is 71.36% of the MetaVSR model and is able to obtain comparable performance, and Fig. 4 illustrates the visual comparison experiment of the x4 super-resolution task.

In addition, to better illustrate that the DSCVSR model has the same performance as MetaVSR for arbitrary magnification super-resolution, we are also do some experiments about x2 and x3 super-resolution experiments, which visual comparison results are displayed in Fig. 5 and Fig. 6, respectively.

Overall, it can be seen by conducting x2, x3, and x4 super-resolution experiments that the proposed DSCVSR model with fewer number of parameters then exhibits comparable performance to MetaVSR.

4.3 Ablation Study

In order to prove the effectiveness of the proposed distillation loss, the corresponding ablation experiments are designed, which are mainly divided into two groups: only pixel loss L_{pix} and two groups of experiments with pixel loss and distillation loss $L_{distill}$, and the video supershooting experiments with x4 super-resolution are carried out on the Vimeo90K, REDS4, and Vid4 datasets, respectively, and corresponding quantization metrics are computed, and the final quantization results can be seen in the Table 2 is shown. It can be seen through this table that the proposed knowledge distillation loss has an average of 2.25% and 2.55% improvement on the three datasets when performing quadruple super-scoring experiments relative to the pixel loss only approach. Thus, the result effectively demonstrates that the performance of the distilled student model will

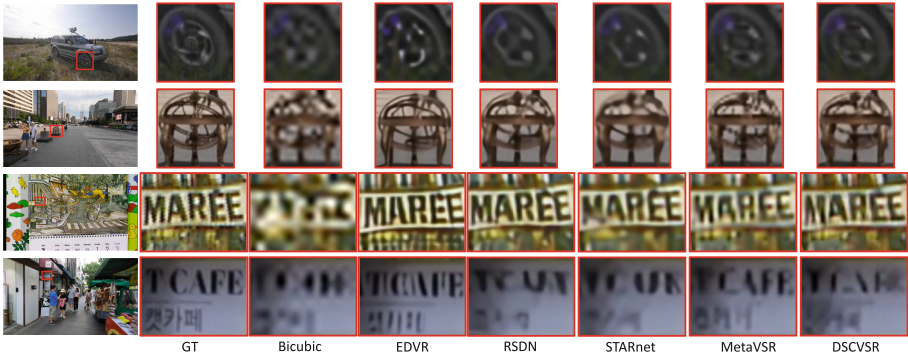


Fig. 4. This figure shows the x4 visual experiment comparison. In the first and fourth rows of visual experiments, it can be seen that DSCVSR is able to be similar to MetaVSR, and indicates better results for the reconstruction of scene object contours compared to other methods. In the third row, the reconstruction of the font “MAREE” shows no over-smoothing effect.

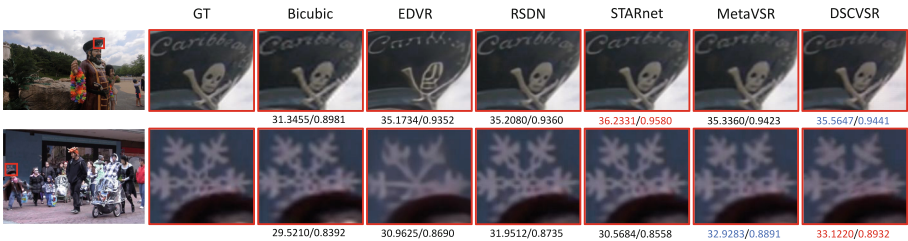


Fig. 5. This figure shows the results of the visual comparison experiment for the x2 super-resolution. In the x2 visual comparison experiments, the reduction principle for distant fonts in the scene can effectively restore the outline and content of the fonts, e.g., in the first line of the experiment. The DSCVSR model, on the other hand, shows a slight advantage in the metrics in scenes with complexity.

be able to be enhanced by considering the feature loss in the intermediate process during knowledge distillation.

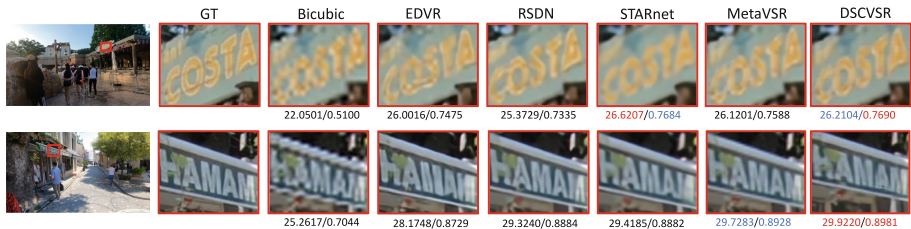


Fig. 6. This figure shows the results of the visual comparison experiment for the x3 super-resolution. In the triple experimental results, DSCVSR’s reconstruction of the fonts “COSTA” and “HAMAM” is more effective in restoring the scene’s font surfaces than the other methods, and shows a slight advantage in the calculation of metrics.

Table 2. This table shows the impact of the distillation loss. As shown in this table, combine the distill loss and pixel loss, the model DSCVSR can get more fined result on x4 super-resolution experiments.

| $L_{distill}$ | L_{pix} | Vimeo90K | Vid4 | REDS4 |
|---------------|-----------|---------------------|---------------------|---------------------|
| ✗ | ✓ | 33.21/0.9129 | 23.51/0.7593 | 27.41/0.7812 |
| ✓ | ✓ | 34.03/0.9248 | 24.00/0.7713 | 28.01/0.8184 |

5 Conclusions

To deploy a video super-resolution network with arbitrary magnification on edge devices, we optimized the teacher model MetaVSR. Leveraging depthwise separable convolution and knowledge distillation techniques, we designed the Depthwise Separable Convolutional Video Super-Resolution Network (DSCVSR). Experimental results validate the effectiveness of the DSCVSR method, achieving comparable performance to the MetaVSR model with fewer parameters.

Although the DSCVSR model achieves performance on par with the MetaVSR model, the introduction of depthwise separable convolution technology results in a decrease in model inference speed. Therefore, improving model inference speed will be one of the future optimization directions. We believe that optimizing the DSCVSR model will facilitate the application of lightweight video super-resolution networks with arbitrary scaling functionality across various domains, providing an effective solution for deployment on edge devices in different fields.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (62106150), the Open Research Fund of Anhui Province Key Laboratory of Machine Vision Inspection (KLMVI-2023-HIT-01), and the Director Fund of Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen) (24420001).

References

1. Caballero, J., et al.: Real-time video super-resolution with spatio-temporal networks and motion compensation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4778–4787 (2017)
2. Cao, W., Li, D., Zhang, X., Qiu, M., Liu, Y.: BLSHF: broad learning system with hybrid features. In: Memmi, G., Yang, B., Kong, L., Zhang, T., Qiu, M. (eds.) KSEM 2022. LNCS, vol. 13369, pp. 655–666. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-10986-7_53
3. Cao, W., et al.: A review on multimodal zero-shot learning. Wiley Interdisc. Rev. Data Mining Knowl. Discov. **13**(2), e1488 (2023)
4. 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)
5. 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)
6. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Investigating tradeoffs in real-world video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5962–5971 (2022)
7. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5008–5017 (2021)
8. Chen, Z., et al.: Videoinr: learning video implicit neural representation for continuous space-time super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2047–2057 (2022)
9. 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)
10. Guo, J., et al.: Distilling object detectors via decoupled features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2154–2164 (2021)
11. 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)
12. Hong, Z., et al.: MetaVSR: a novel approach to video super-resolution for arbitrary magnification. In: Rudinac, S., et al. (eds.) MMM 2024. LNCS, vol. 14554, pp. 300–313. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-53305-1_23
13. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint [arXiv:1704.04861](https://arxiv.org/abs/1704.04861) (2017)
14. Huang, Y., Chen, J.: Improved EDVR model for robust and efficient video super-resolution. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 103–111 (2022)

15. 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
16. Isobe, T., et al.: Video super-resolution with temporal group attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8008–8017 (2020)
17. 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
18. 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)
19. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
20. 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)
21. Li, G., Li, X., Wang, Y., Zhang, S., Wu, Y., Liang, D.: Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 1306–1313 (2022)
22. Liu, C., Sun, D.: On Bayesian adaptive video super resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(2), 346–360 (2013)
23. Miles, R., Yucel, M.K., Manganelli, B., Saà-Garriga, A.: Mobilevos: real-time video object segmentation contrastive learning meets knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10480–10490 (2023)
24. 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 (2019)
25. Polino, A., Pascanu, R., Alistarh, D.: Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668* (2018)
26. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4472–4480 (2017)
27. Wang, H., Su, D., Liu, C., Jin, L., Sun, X., Peng, X.: Deformable non-local network for video super-resolution. *IEEE Access* **7**, 177734–177744 (2019)
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 (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, L., Ma, K.: Improve object detection with feature-based knowledge distillation: towards accurate and efficient detectors. In: International Conference on Learning Representations (2020)
31. Zhou, X., Cao, W., Gao, H., Ming, Z., Zhang, J.: STI-Net: spatiotemporal integration network for video saliency detection. *Inf. Sci.* **628**, 134–147 (2023)