

Multi-Frame Alignment for Video Super-Resolution Using Attention

Marco Di Rienzo[1], Matteo Bruni[2], Leonardo Galteri[3], Federico Becattini[4(✉)], and Marco Bertini[1]

[1] University of Florence, Florence, Italy
{marco.dirienzo,marco.bertini}@unifi.it
[2] Small Pixels, Florence, Italy
matteo.bruni@small-pixels.com
[3] Pegaso Telematic University, Naples, Italy
leonardo.galteri@unipegaso.it
[4] University of Siena, Siena, Italy
federico.becattini@unisi.it

Abstract. This paper presents a novel attention-based video super-resolution (VSR) method that avoids costly optical flow estimation while effectively exploiting temporal correlations between frames. We propose an aligner module that utilizes cross-attention to blend relevant patches from adjacent frames, gathering information from multiple frames simultaneously. This method improves upon traditional flow-based approaches by working at a block level and enabling the blending of several pixels, yielding better alignment for larger motions. The proposed VSR technique can upscale videos up to 4x while simultaneously removing compression artifacts, enhancing both resolution and quality. Experimental results demonstrate the effectiveness of this approach compared to classic flow-based methods, particularly in handling compressed videos where compression artifacts can severely impact optical flow estimation.

Keywords: Video super-resolution · Attention · Artifact Removal

1 Introduction

Video super-resolution (VSR) techniques aim to improve the quality of a sequence of low-resolution frames by generating a high-resolution sequence. Single-Image Super-Resolution (SISR) techniques can be trivially applied to an entire video, obtaining a frame from each individual low-resolution input frame; however, in this way the temporal correlations that exist between frames are not exploited to further improve the produced output. Thus, the exploitation of temporal dependencies turns out to be essential for the proper functioning of VSR methods. Several VSR methods [3,6,14,23,25,36] rely on the alignment of temporal features within the pixel space using optical flow, which is used to estimate the motion between frames and warp them accordingly. Their standard

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
E. Rodolà et al. (Eds.): ICIAP 2025, LNCS 16167, pp. 608–620, 2026.
https://doi.org/10.1007/978-3-032-10185-3_48

pipeline involves feature alignment as the first step, followed by the actual super resolution task, where high-resolution details are reconstructed from the aligned frames. However, optical flow techniques are usually computationally expensive and thus are not viable for applications where real-time processing represents a hard constraint. Furthermore, optical flow limits alignment to adjacent frames: even though it is possible to align temporally distant frames, with large time offsets the method is prone to errors. In this case, temporal inaccuracies can significantly impact the motion estimation and the overall perceived quality of the super-resolved video, making it challenging to maintain high fidelity across video frames. Another serious complicating factor emerges when trying to increase the resolution of compressed videos [31], as the presence of compression artifacts and the removal of important information at the pixel level often become detrimental to optical flow estimation. In particular, the estimated motion can be severely distorted due to different noise patterns, resulting in poor performance in the alignment phase.

In this work, we adhere to the two-step paradigm involving an alignment module followed by a super-resolution convolutional model. However, instead of relying on costly flow-estimation steps, we improve the alignment module with an attention-based mechanism, capable of blending relevant patches from adjacent frames. In this way, we can easily gather information from multiple adjacent frames. The proposed solution is general and can replace any flow-based alignment strategy found in super-resolution pipelines. In addition, since we train the whole pipeline end-to-end, we can jointly tackle super-resolution and video artifact removal, which proves to help obtain better alignments and overall better visual quality in the reconstructed frames.

Our main contributions are the following:

- We propose an attention-based video upscaling method, that retrieves similar patches to gather information and improve the visual quality.
- Our solution avoids costly flow-alignment by building aligning patches via cross-attention. This removes the overhead of estimating a flow field and allows us to align information from multiple frames. In principle, our approach can be plugged-in in any existing method that relies on flow-alignment.
- The proposed method can upscale videos up to 4x and simultaneously removes compression artifacts, enhancing both the resolution and the quality of the videos. Experimental results demonstrate the effectiveness of our approach against classic flow-based approaches.

2 Related Work

A great amount of image super-resolution techniques have been developed, including SRCNN [10], FSRCNN [11], VDSR [15], ESPCN [28], and SRGAN [18]. These models leverage convolutional layers to extract features from the input low-resolution (LR) images. By incorporating deconvolutional layers in the final

stages of the network, the image is upscaled directly within the model. Significant improvements have been achieved with recent transformer-based methods. Notable methods include IPT [7], SwinIR [21], UFormer [34], ART [37], CAT [9] and HAT [8]. VSR techniques instead enhance video quality by leveraging temporal information across multiple frames. Kappeler et al. [14] introduced VSRnet, a pioneering VSR technique using CNNs. This model aligns and processes consecutive frames to improve resolution. GAN-based techniques as in [32], employ adversarial training to produce high-resolution videos with enhanced perceptual quality. RNN-based methods such as [30], utilize temporal dependencies in video sequences to better capture motion and maintain consistency across frames. The typical VSR pipeline includes alignment, feature extraction and fusion, and reconstruction modules. Alignment is crucial to ensure that frames are accurately aligned before the super-resolution step. However, optical flow methods like SpyNet [26] can struggle with large motions. Attention-based alignment methods, such as the one proposed in this work, offer a more flexible approach by blending multiple reference blocks. Feature extraction and fusion from multiple frames allow the model to gather comprehensive information, improving the quality of the upscaled video, and the final reconstruction module refines the output, enhancing details and reducing artifacts. Recent innovations like BasicVSR [5] employ a bidirectional recurrent network to integrate backward and forward propagation, significantly improving temporal coherence and detail reconstruction, while EGVSR [4] introduces an end-to-end trainable system that relies on optical flow and warping for frame alignment, enhancing computational efficiency. Other recent strategies have shifted to use the flow to guide deformable convolutions [6,20,29] and deformable attention schemes [22]. To further enhance robustness against unaccurate optical flow, the approach proposed in [27] introduces patch alignment, which averages the motion vectors within predefined grid areas.

3 Method

We aim to predict the high-resolution frame SR_t using an input combination of the low-resolution frame I_t and a sequence of neighboring low-resolution frames $(\ldots, I_{t-2}, I_{t-1}, I_{t+1}, I_{t+2}, \ldots)$. The network is parameterized by weights θ where each θ^l denotes the weights for a certain layer l. For a training set of input low-resolution frames I_i, and the corresponding target high-resolution frames T_i, with $i \in \{1, ..., N\}$, where i represents a specific training sample index, the model is trained for the parameters $\hat{\theta}$ that minimize the error between the target and the prediction. The model learns the function h_θ where $SR_i = h_\theta(I_i)$, and approximates SR_i as close as possible to the target reference frame T_i.

Video super-resolution pipelines often involve a first step of alignment where the neighboring frames are aligned towards the current frame. In this work, we propose a novel alignment method solely based on cross-attention, avoiding typical optical flow estimation and warping operations.

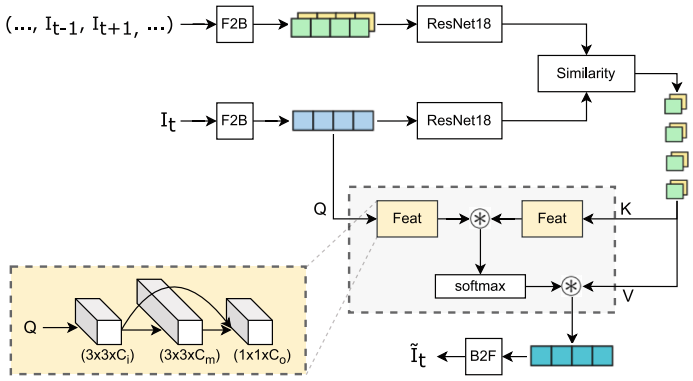


Fig. 1. Aligner module. Blocks inside the shaded box compute the cross-attention. The features for keys and queries are extracted by the convolutional layers shown in the yellow box, C indicates hyper-parameters for the channels dimension. (Color figure online)

Aligner Module. The proposed alignment method is based on cross-attention and its scheme is shown in Fig. 1. The first step of our method is to divide the reference frames $(\ldots, I_{t-2}, I_{t-1}, I_{t+1}, I_{t+2}, \ldots)$ and the target frame (I_t) into non-overlapping blocks (F2B operation in the scheme) and stacking them along the batch dimension. Then, for each block of the target frame, we find the most similar blocks in each of the reference frames: to achieve this, features of the blocks are extracted by passing them through a pre-trained ResNet18, and then a similarity between target and reference blocks is computed to only keep the k most similar blocks. This step allows computing the cross attention on a limited number of blocks, significantly reducing the computational burden.

Once we have the blocks from the target frame and the top-k blocks for each reference frame, we compare them via cross-attention. We use the reference blocks as V (value), while we use cascaded residual blocks, shown in the yellow box in Fig. 1, to extract features from the two sets of target and reference blocks, which we use as Q (query) and K (key), respectively. Q and K are matrix multiplied to create a dot-product attention matrix $A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$. Note that the matrix A maps images from the K domain to the Q domain via an adaptive linear combination; that is, the matrix represents a linear combination map that allows pixel values from the reference blocks to be blended together to reconstruct pixels of the target block. This linear combination map is used to align references to the target. The new set of aligned blocks is then reorganized back to compose an image (B2F operation in the scheme), i.e., producing the aligned frame $\tilde{I}_t$. This method is advantageous over the typical optical flow and warping because optical flow computes the planar displacement between frames at a pixel level, while our method works at a block level and permits the blending of several pixels in the reference blocks to reconstruct a single pixel in the target block. This yields a better alignment on larger motions. By keeping the top-k

blocks, our method can also find similar blocks in distant areas of the reference frame, and use that in the subsequent cross-attention. Moreover, more reference frames can be considered and aligned at once.

Upsample Module. Once the aligned frame has been obtained, we concatenate it with the low-resolution frame along the channel dimension and send the resulting tensor as input to the upsample module. The upsample module draws inspiration from that of ESPCN [28]: it is composed of several consecutive residual blocks and a pixel-shuffle [28] layer at the end that performs the upsampling.

The output of the upsample module is then added to the original low-resolution target frame I_t upsampled via bicubic interpolation. The effectiveness of this operation is due to the high correlation between the input and the output frame, allowing the network to focus on predicting high frequencies instead of the whole image. The result is the estimated high-resolution frame SR_t. The overall scheme of the proposed architecture is shown in Fig. 2.

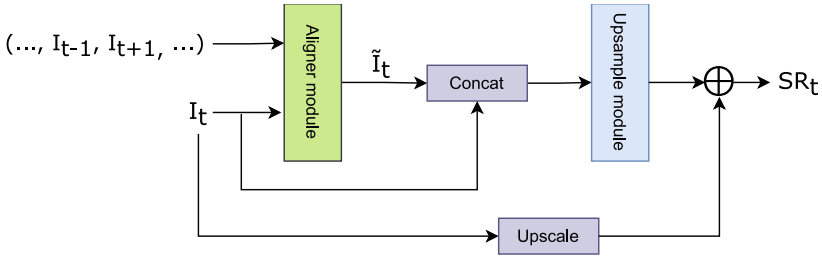


Fig. 2. Full scheme of our architecture.

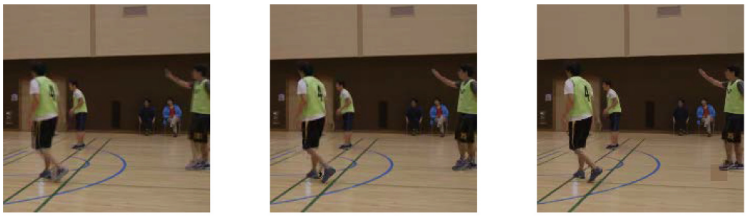
Objective Function. During the training process, we aim to recover high-resolution signals from the low-resolution input images. Our objective function comprises several loss terms, which we outline in detail.

Pixel Loss. This term is a signal-based loss and is the most commonly used in SISR and VSR; it computes the L1 or Charbonnier distance [2] between each generated pixel in the predicted HR frame and the corresponding pixel in the ground truth HR frame. We use the Charbonnier-loss for better stability: $\mathcal{L}_{pixel} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(SR_i - T_i)^2 + \epsilon^2}$ with N the batch size, ϵ 10^{-3}, SR the predicted HR frame and T the HR frame.

Perceptual Loss. This term measures the perceptual similarity between the predicted HR frame and the ground truth HR frame. We use the LPIPS loss [35] with AlexNet [17] as backbone: $\mathcal{L}_{perceptual} = \frac{1}{N}\sum_{i=1}^{N} d_{\text{LPIPS}}(SR_i, T_i)$.

Alignment Loss. This term is used to push the aligner module toward a better reconstructed low-resolution frame and is similar to the loss used in optical flow-based methods, usually called the warping loss. It is a pixel loss with Charbonnier-distance between the aligned LR frame and the target LR frame:

$\mathcal{L}_{align} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(\tilde{I}_i - I_i)^2 + \epsilon^2}$ where N is the batch size, ϵ is 10^{-3}, $\tilde{I}$ is the aligned LR frame and I is the target LR frame. The total loss is a weighted sum of the aforementioned loss terms: $\mathcal{L}_{total} = \alpha\mathcal{L}_{pixel} + \beta\mathcal{L}_{perceptual} + \gamma\mathcal{L}_{align}$ where α, β and γ are the weights for each term.



(a) Motion image (b) SpyNet+warp [26] (c) Attention-based 2f

Fig. 3. Different alignments. (a) Average between reference and target; (b) alignment of SpyNet+Warp [26]; (c) attention-based alignment.

4 Experiments

In this section, we present the results of different experiments. All experiments are done on a single NVIDIA GeForce RTX 2080 Ti GPU with 10GB of memory.

Dataset and Metrics. We conduct all experiments on the *BVI-DVC* dataset [24]. The dataset consists of 200 sequences of frames, each truncated at the 64th frame, regardless of the frame rate, which ranges from 25 to 120. These sequences encompass various content types, including natural scenes, man-made objects, and city scenes. We created downscaled versions at 1080p and 270p resolutions from the 2160p originals, which will be our ground-truth and training data, respectively. From the 1080p downscaled sequences, we removed 4 sequences as test set[1]. The remaining sequences serve as the ground truth for our training data. To generate the compressed input data, each 270p sequence is compressed using the H.264 codec with a Constant Rate Factor (CRF) of 22. We chose a fixed CRF to mitigate the mode-collapse issue discussed in [12]. However, it is important to note that a fixed CRF does not ensure consistent quality across all frames; frames with motion may have lower quality compared to static frames. Given our objective of applying super-resolution to compressed videos (which involves both artifact reduction and super-resolution), it is crucial to train the models on the compressed videos rather than just downscaled versions of high-quality videos. Training the model solely for super-resolution could result in failure to detect features to super-resolve or even enlarge compression artifacts,

[1] *AmericanFootballS3Harmonics*, *BasketballS3YonseiUniversity*, *CityStreetS7IRIS*, *SquareS2IRIS*.

thereby diminishing the overall quality. In this study, we aim to produce results centered on both reconstruction accuracy and perceptual naturalness. Therefore, we employ metrics such as PSNR [13], SSIM [33], LPIPS [35]. These are full-reference metrics, meaning that they require both the target and the reconstructed output for quality assessment. Furthermore, we consider VMAF [19] which is a widely used full reference video metric, and ARNIQA [1], which is a no-reference metric, therefore not requiring the target image for the assessment. All metric evaluations are performed on the 3 RGB channels. Image quality metrics are individually calculated for each frame and averaged throughout the test video. On the other hand, the VMAF value is calculated by encoding the generated frames back into a video and evaluated against the target video directly.

Alignment Performance. In this experiment, we compare the alignment capabilities of the motion estimation/motion compensation with optical flow and warp, with our attention-based aligner module described in Sect. 3 with various temporal window sizes. We first warp the reference (e.g., the previous frame) to the target frame (i.e., the current frame) with optical flow estimated by SpyNet [26]. Then we perform the same alignment with our attention-based method. All experiments are performed with a block size of 16 with a stride of 16, and $k = 1$. An example pair and alignment results of both methods can be seen in Fig. 3.

The effectiveness of the optical flow-based techniques diminishes when confronted with significant motion. This decline primarily stems from the fact that this method constrains the changes between frames to planar geometric motion with a one-to-one correspondence of pixels. From the perspective of the attention mechanism, these constraints are equivalent to restricting the attention matrix to a permutation matrix where each row has only one entry as 1. However, in our approach, we can loosen this constraint by ensuring that the sum of each row equals 1 (achieved through softmax), instead of only one element. This relaxation permits the involvement of multiple pixels and blending, facilitating the alignment of the reference image to the target image which yields enhanced performance as shown in Table 1. For our attention-based alignment, we report the results taking into account a variable number N of reference images, where $\lceil N/2 \rceil$ images are taken from the previous frames and $\lfloor N/2 \rfloor$ images are taken from the subsequent frames. Interestingly, increasing the number of reference frames provides more relevant information to the model, yet even the single-reference aligner proves to be much more effective than the flow-based aligner.

Table 1. Alignment varying the amount of reference frames.

Method	#Ref	PSNR	SSIM	LPIPS
SpyNet+warp	1	23.21	0.873	0.096
Attention-based	1	29.47	0.966	0.042
Attention-based	2	36.95	0.973	0.034
Attention-based	4	**37.17**	**0.976**	**0.026**

Upscaling Videos by 4x. We upsample the input sequences of size ($480 \times 270 \times 3$) to a size of ($1920 \times 1080 \times 3$), for an upscale factor of $4\times$. Both lossless and compressed videos with a CRF of 22 are considered. Under our training setup, because we have access to both the input and the target data, the model is trained in a fully supervised fashion. Hence, the network is trained to super-resolve the input frame, for which the ground truth frame is available.

We randomly crop patches of input size ($64 \times 64 \times 3$) from the input frames for efficient training. The patches are randomly augmented with mirror flips and rotations. The entire pipeline is trained end-to-end using the loss described in Sect. 3. Adam optimizer [16] with default settings is used to train the model for 100 epochs with a batch size of 8. The learning rate is initialized to $5e-4$. The model is compared against a classic bicubic upsampling, i.e., the upsample module described in Sect. 3 without the aligner module (which we name SRNet - Super Resolution Network), and an optical flow based network similar to EGVSR [4], trained with similar hyperparameters and the same upsample module of our network for a fair comparison.

Table 2. Evaluation of $4\times$ upsampling on videos without compression. Each metric is calculated on RGB channels.

Model	#Params	PSNR	SSIM	LPIPS	VMAF	ARNIQA
Bicubic	–	29.98	0.829	0.316	68.32	0.254
SRNet	1.3M	31.02	0.837	0.170	84.03	0.393
Flow+SRNet	2.7M	32.52	0.866	0.095	87.05	0.411
XA+SRNet	1.4M	**32.84**	**0.865**	**0.086**	**88.26**	**0.424**

Uncompressed Videos. We first train and test our model on uncompressed videos.The average metric values measured for the test videos are shown in Table 2. It can be seen that the SRNet model alone is better than the bicubic approach on PSNR and large improvements can be observed on perceptual metrics. VSR models outperform the SRNet method on every metric, underlining the importance of utilizing the additional information contained in the temporal dimension of videos. Our method also gains an advantage over the optical flow method. Examples of generated SR frames using the different models and the bicubic method are shown in Fig. 4 together with the input and target frames. The bicubic method only generates averaged pixel values and has issues generating the edges and the finer details, leading to the degradation of high-frequency information and thus a blurry image. In contrast, the frames generated by the models are clearer and have sharper edges compared to the bicubic method. Additionally, VSR models improve over the SR model by utilizing additional information from other frames, leading to an even clearer image and improved consistency between frames, according to the VMAF metric. Our method can recover more details than the flow-based model.



Fig. 4. 4× upsample on test clip *basket* and *square* not compressed.

Compressed Videos. In this experiment, we train our network on compressed video with H.264 codec and CRF = 22. These lossy/compressed images lack rich detail and often contain visual artifacts and aliasing. The average metric values measured for the test videos are shown in Table 3. It can be seen that on lossy images, the bicubic method does worse by a significant margin even on PSNR, due to the lack of ability to reduce compression artifacts and create sharp details. Similarly to the experiment on uncompressed videos, the VSR models outperform the SR model, especially on perceptual metrics. Samples of generated SR frames are shown in Fig. 5. Again, the bicubic method reconstructs a blurry image, but this time it also amplifies compression artifacts like ringing. On the other hand, the models can reconstruct a sharper image, and also reduce the compression artifacts. As before, VSR models improve over the SR model by utilizing additional information from other frames. In this case, the effect of leveraging temporal information is bigger due to the additional loss of information due to compression.

Table 3. Evaluation of 4× upsampling on videos compressed with CRF = 22. Each metric is calculated on RGB channels.

Model	#Params	PSNR	SSIM	LPIPS	VMAF	ARNIQA
Bicubic	-	28.64	0.787	0.395	49.16	0.235
SRNet	1.3M	29.26	0.799	0.229	61.17	0.336
Flow+SRNet	2.7M	29.67	0.795	0.176	64.69	0.366
XA+SRNet	1.4M	**29.99**	**0.802**	**0.155**	**65.11**	**0.372**



Fig. 5. 4× upsample on test clip *basket* and *square* compressed with CRF 22.



(a) no LL. (b) 1 LL frames. (c) 2 LL frames. (d) All LL. (e) GT

Fig. 6. 4× upsample with varying number of lossless (LL) frames. Improvements can be seen when lossless frames are available, especially between the upsample with no lossless frames and 1 lossless frame.

Table 4. 4× upsampling on videos compressed with CRF 22, varying the lossless frames in the aligner window.

Frames	PSNR	SSIM	LPIPS	VMAF	ARNIQA
1	30.83	0.838	0.126	68.88	0.374
2	31.12	0.830	0.110	71.03	0.378
3	31.15	0.831	0.105	71.93	0.380
4	31.16	0.831	0.104	72.13	0.381
5	31.16	0.832	0.102	**72.21**	0.381
6	**31.16**	**0.832**	**0.102**	72.20	**0.381**

Compressed Videos with Lossless Frames Available. In this experiment, we simulate a scenario in which a compressed video has to be upscaled, but also some lossless frames are available. A possible application scenario is a streaming setting where a subset of the frames are streamed at a high bitrate and the rest are streamed after being compressed. We use these additional lossless frames by inserting them in the window of the aligner module.

The resulting metrics on the test videos are shown in Table 4. It can be seen that between the model with all compressed frames and the model with just 1 lossless frame, there is a considerable increase of 0.84dB in PSNR and 3.77 points of VMAF. The increase is still sensible in PSNR and VMAF with an additional frame, to decrease to a negligible amount with additional lossless frames.

Table 5. Performance evaluation of 4× upsampling when the temporal window size is changed.

Frames	PSNR	SSIM	LPIPS	VMAF	ARNIQA
2	31.88	0.855	0.118	85.69	0.409
3	32.71	0.863	0.090	87.92	0.420
4	32.80	0.864	0.088	88.06	0.422
5	32.83	0.865	0.087	88.22	0.424
6	32.84	0.865	0.087	88.25	0.424
7	**32.84**	**0.865**	**0.086**	**88.26**	**0.424**

A visual comparison between upsampled compressed frames with a varying number of lossless frames in the temporal window is shown in Fig. 6. The difference is especially visible between Fig. 6 awhere no lossless frames are available and Fig. 6 bwhere 1 lossless frame is available. We can see that some color is added back to the image, e.g. the red stripes on the pole, the shapes are clearer and some background noise is removed. In Fig. 6 cwe can see the result with 2 lossless frames, where the image is even clearer and the shapes more defined. For comparison, the result using only lossless frames is shown in Fig. 6 d. Note that we use a window of 3 past frames and 3 future frames. The main difference lies in the shapes in the image. Compared with the ground truth in Fig. 6 e, the image using only lossless frames has consistent shapes, while the image using only two lossless frames still maintains the deformed shapes of the compressed source. This stems from the fact that the compressed image is still used as the query in our alignment module, making the lossless frames help in improving the general quality of the image, but leaving the deformed shapes as they are.

Ablation Studies. We first vary the temporal length of the input sequence from 2 (fusing 1 previous and reference frames) to 7 (fusing 3 previous, reference, and 3 future frames) to super-resolve one frame. The metrics on test videos are reported in Table 5. If we compare these results with those in Table 2, we see that with only 2 frames (previous and reference) the metrics are higher by a considerable margin than the upsample module alone (single frame), denoted as SRNet in the table, with an improvement of 0.86dB in PSNR and about 1.6 points of VMAF, but lower than the Flow+SRNet due to the additional information. With 3 frames (1 previous, reference, 1 future) the metrics surpass the flow method by about 0.2dB of PSNR and 0.87 points of VMAF, underlining the importance of considering the future frame to predict the reference one.

Then, we compare the results varying the number of top-k blocks in the aligner module. The idea is that also considering blocks that are not the exact same object but have similar content, can help the aligner. Results are shown in Table 6. For the considered test videos, only minimal gains are achieved when changing from $k = 1$ to $k = 2$, and no improvement in visual quality can be noticed.

Table 6. Evaluation of 4× upsampling when $k = 1$ and $k = 2$.

k	PSNR	SSIM	LPIPS	VMAF	ARNIQA
1	29.99	0.802	0.155	65.11	0.372
2	**30.08**	**0.803**	**0.148**	**65.28**	**0.374**

5 Conclusions

In this paper, we introduced an attention-based method for patch alignment in a video super-resolution pipeline. The advantage of our proposed aligner is that it avoids costly optical flow computations, commonly used in state of the art methods, and can be easily extended to take into account multiple frames. We empirically demonstrated the effectiveness of the proposed approach for video super-resolution up to 4x, while also addressing the issue of compression artifact removal. The proposed method is model agnostic and can be plugged into any super-resolution pipeline that leverages an alignment module, potentially improving any flow-based method.

References

1. Agnolucci, L., Galteri, L., Bertini, M., Del Bimbo, A.: Arniqa: learning distortion manifold for image quality assessment. In: Proc. of WACV (2024)
2. Barron, J.T.: A general and adaptive robust loss function (2019)
3. Caballero, J., et al.: Real-time video super-resolution with spatio-temporal networks and motion compensation. In: Proc. of CVPR, pp. 4778–4787 (2017)
4. Cao, Y., Wang, C., Song, C., Tang, Y., Li, H.: Real-time super-resolution system of 4k-video based on deep learning (2021)
5. Chan, K.C.K., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: the search for essential components in video super-resolution and beyond (2021)
6. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: improving video super-resolution with enhanced propagation and alignment. In: Proc. of CVPR (2022)
7. Chen, H., et al.: Pre-trained image processing transformer. In: Proc. of CVPR (2021)
8. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proc. of CVPR, pp. 22367–22377 (2023)
9. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yuan, X., et al.: Cross aggregation transformer for image restoration. NeurIPS **35**, 25478–25490 (2022)
10. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks (2015)
11. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network (2016)
12. Galteri, L., Seidenari, L., Bertini, M., Bimbo, A.: Deep universal generative adversarial compression artifact removal. IEEE Trans. Multimedia **21** (2019)
13. Horé, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: 2010 20th International Conference on Pattern Recognition, pp. 2366–2369 (2010). https://api.semanticscholar.org/CorpusID:9506273
14. Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A.K.: Video super-resolution with convolutional neural networks. IEEE Trans. Comput. Imaging **2**, 109–122 (2016). https://api.semanticscholar.org/CorpusID:9356783
15. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks (2016)
16. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2017)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**, 84 – 90 (2012). https://api.semanticscholar.org/CorpusID:195908774

18. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network (2017)
19. Li, Z., Aaron, A., Katsavounidis, I., Moorthy, A., Manohara, M.: Toward a practical perceptual video quality metric (2016). https://netflixtechblog.com
20. Liang, J., et al.: VRT: a video restoration transformer. IEEE Trans. Image Process. (2024)
21. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: image restoration using swin transformer. In: Proc. of ICCV, pp. 1833–1844 (2021)
22. Liang, J., et al.: Recurrent video restoration transformer with guided deformable attention (2022)
23. Liu, H., et al.: Video super resolution based on deep learning: a comprehensive survey (2022)
24. Ma, D., Zhang, F., Bull, D.R.: BVI-DVC: a training database for deep video compression. IEEE Trans. Multimedia **24**, 3847–3858 (2022)
25. Qing, T., Ying, X., Sha, Z., Wu, J.: Video super-resolution with pyramid flow-guided deformable alignment network. In: Proc. of ICEEMT. IEEE (2023)
26. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network (2016)
27. Shi, S., Gu, J., Xie, L., Wang, X., Yang, Y., Dong, C.: Rethinking alignment in video super-resolution transformers. NeurIPS **35**, 36081–36093 (2022)
28. Shi, W., et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Proc. CVPR (2016)
29. Teed, Z., Deng, J.: RAFT: recurrent all-pairs field transforms for optical flow. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 402–419. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_24
30. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: TDAN: temporally deformable alignment network for video super-resolution (2018)
31. Wang, J., Shao, Z., Huang, X., Lu, T., Zhang, R., Li, Y.: From artifact removal to super-resolution. IEEE Trans. Geosci. Remote Sens. **60**, 1–15 (2022)
32. Wang, X., Chan, K.C.K., Yu, K., Dong, C., Loy, C.C.: Edvr: video restoration with enhanced deformable convolutional networks (2019)
33. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**, 600–612 (2004). https://api.semanticscholar.org/CorpusID:207761262
34. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: efficient transformer for high-resolution image restoration. In: Proc. of CVPR, pp. 5728–5739 (2022)
35. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric (2018)
36. Zhu, Q., Chen, F., Zhu, S., Liu, Y., Zhou, X., Xiong, R., Zeng, B.: Dvsrnet: deep video super-resolution based on progressive deformable alignment and temporal-sparse enhancement. IEEE Trans. Neural Netw. Learn. Syst. (2024)
37. Zhu, Q., Li, P., Li, Q.: Attention retractable frequency fusion transformer for image super resolution. In: Proc. of CVPR, pp. 1756–1763 (2023)