Quantized Warping and Residual Temporal Integration for Video Super-Resolution on Fast Motions

Konstantinos Karageorgos[0000−0002−5426−447], Kassiani Zafeirouli[0000−0002−9496−6329], Konstantinos Konstantoudakis[0000−0001−5092−8796], Anastasios Dimou[0000−0003−2763−4217], and Petros Daras[0000−0003−3814−6710]

Visual Computing Lab
Information Technologies Institute, Centre for Research and Technology Hellas, Thessaloniki, Greece
{konstantinkarage,cassie.zaf,k.konstantoudakis,dimou,daras}@iti.gr
vcl.iti.gr

Abstract. In recent years, numerous deep learning approaches to video super resolution have been proposed, increasing the resolution of one frame using information found in neighboring frames. Such methods either warp frames into alignment using optical flow, or else forgo warping and use optical flow as an additional network input. In this work we point out the disadvantages inherent in these two approaches and propose one that inherits the best features of both, warping with the integer part of the flow and using the fractional part as network input. Moreover, an iterative residual super-resolution approach is proposed to incrementally improve quality as more neighboring frames are provided. Incorporating the above in a recurrent architecture, we train, evaluate and compare the proposed network to the SotA, and note its superior performance in faster motion sequences.

Keywords: super resolution, motion compensation

1 Introduction

Super resolution (SR) refers to a group of algorithms that aim to upsample a low resolution input (LR) in order to produce a higher resolution (HR) output. The challenge lies in reproducing the missing high frequency details of the input. SR is an ill-posed problem as there is no unique relationship between a LR and a HR image. SR methods can utilize either a single image (SISR), or multiple images (MISR) as an input. Given the abundance of video data streams, it became evident that MISR methods can be used to improve the resolution of a video due to the temporal consistency of successive frames. Video SR methods (VSR) effectively recover high frequency content, using information from neighboring frames.

VSR methods aim to exploit unique information from each one of the neighbouring frames in order to produce a true HR result. For this purpose, the frames, which contain the same content shifted in an arbitrary way, must be accurately registered to the examined one. The biggest challenge here is to account for the inter-frame motion in the sequence. Realistic videos can contain arbitrary motion due to the camera object movement, making the registration a tedious task.

A naive approach to the problem is to concatenate all inputs and let a deep Convolutional Neural Network (CNN) implicitly model the spatial relationship between useful features. Although the increased depth and pooling operations of modern CNNs have quite big effective receptive fields, their convolutional nature remains local. While local correspondences may get captured at higher layers, additional complexity is added to the model making it more difficult to train and generalize. This is especially relevant in faster motion sequences, where object displacements between neighboring frames are larger and a correspondingly wider receptive field is required to capture them.

A common way to alleviate the inter-frame motion is to explicitly compensate this disparity by using warping to spatially align the neighboring images to a common reference location. Despite the intuitive merit of explicit warping, it constitutes a resampling operation using interpolation, which inherently causes blurring, lowered contrast and loss of information, reducing the super-resolution's effectiveness.

Another important parameter of the VSR methods is the strategy used to incorporate information from the neighboring frames, giving rise to different approaches. The number of frames used and the input sequence are important choices that bound the application of the methods proposed in literature. Most recent convolutional methods have to be trained and tested on a fixed number of neighbors, regardless of the early or late fusion scheme used.

This work focuses on improving super-resolution quality on video sequences with larger motions. Towards this goal, we propose a two step approach to neighboring frame registration: to warp neighboring frames only by the integer part of the optical flow, thus avoiding interpolation and the associated quality degradation; and to use the fractional part of the flow as an input to the neural network, letting it model the sub-pixel correspondences. We incorporate this approach into a recurrent residual architecture that fuses information from neighboring frames using a shared reconstruction branch. The resulting network progressively enhances the output quality with each processed input, offering the flexibility to adapt inference speed and quality by using more or less neighboring frames as input.

The contributions of this work can be summarized in the following aspects:

- An explicit quantized motion compensation methodology, that preserves detail at the input level. The proposed method significantly improves the results and the generalization capacity of a baseline network, especially on complex videos with high inter-frame motion.

– An implicit modeling of sub-pixel motions, using the fractional part of the optical flow as an additional input to the network.
– A recurrent CNN architecture that progressively enhances the produced output with each input frame using residuals is proposed. It can handle frame sequences of arbitrary length, offering unique flexibility.

The proposed methodologies are thoroughly analysed and our claims are firmly supported by extensive experiments and ablation studies.

The rest of this paper is organized as follows: Section 2 discusses recent and relevant advances in deep learning-based SR; Section 3 considers the advantages and disadvantages of different registration strategies and explains the reasoning behind the proposed hybrid approach; The proposed network architecture and its constituent modules are presented in Section 4; Section 5 presents and discusses experimental results, comparisons with state of the art VSR methods, and ablation studies. Lastly, 6 provides a conclusion.

2 Related Work

Single-Image Super-Resolution: In 2014, for the first time, Dong et al. exploited the power of convolutional neural networks, by proposing SRCNN [2], a lightweight 3-layer convolutional model, to address the single image super resolution (SISR) problem. Later, based on SRCNN, deeper and more complex models, such as VDSR [10] with 20 stacked layers, DRCN [11] and DRRN [19] with recursive leaning and parameters sharing and MemNet [20] with memory block, were introduced and achieved higher reconstruction performance. Inspired by DenseNet [6], Tong et al. suggested SRDenseNet [22], by removing the pooling layers, and the RDN model [26] improved SRDenseNet's performance by exploiting local and global residual skip connections. Yang et al. proposed the Deep Edge Guided Recurrent Residual Network (DEGREE) [25], motivated by the fact that edge features can provide valuable guidance for SISR. Based on the conventional back-projection method [7], Haris et al. proposed DBPN [3], a network with iterative upsampling and downsampling modules. Finally, for more photo-realistic results a combination of generative adversarial networks with perceptual and texture matching losses was used in SRGAN [13] and ENEt [17] models, respectively.

Video Super-Resolution Most of the deep learning based Video Super Resolution (VSR) approaches address the VSR task by combining a motion estimation module with an image warping module. Kappeler et al. proposed VSRnet [9], a model that exploits the temporal information by jointly processing multiple consecutive frames. The neighboring frames are warped towards the reference frame by a conventional optical flow algorithm before they are fed to the model. Based to this premise, VESPCN [1] introduced a spatial transformer network, in order to efficiently encode motion information between frames. The aforementioned model is jointly trained with a SISR sub-pixel convolution network for fast and accurate reconstruction. Tao et al. [21] used the same motion

compensation transformer as in VESPCN to produce the motion field and an SPMC layer for simultaneous sub-pixel motion compensation and resolution enhancement. More recently, Sajjadi [18] proposed an end-to-end recurrent video super resolution model, which exploits the information of the previously inferred super-resolved HR frame to reconstruct the subsequent frame. The flow estimation and the SR network sub-modules are trained simultaneously.

The aforementioned methods have as common core element the image warping module that performs alignment by estimating optical flow information between the reference and its neighboring frames. Unlike this approach, Jo et al. [8] proposed DUF, a network that avoids explicit motion estimation and compensation by generating dynamic upsampling filters. The EDVR architecture [23] follows the logic of implicit alignment, introducing a Pyramid, Cascading and Deformable (PCD) alignment module, where alignment is done in a coarse-to-fine manner without the classic image warping technique. Haris et al. extended the SISR DBPN architecture to video super resolution with the RBPN [4], a recurrent model that treats each neighboring frame as a separate source of information that iteratively refines the HR features through multiple up and down projections. Our work introduces a novel approach to register neighboring frames using a quantized warping method, which models the subpixel displacements, to treat efficiently the flow information. Moreover we employ a recurrent residual reconstruction module to refine the SR output with an arbitrary number of input frames.

3 Optical Flow and Spatial Alignment

In video super resolution, the resolution of a reference frame is increased using information from neighboring frames, which are assumed to depict the same scene at different points in time. In each frame, the same object may occupy a different position, due to its own movement or global camera movement. For this reason, VSR methods must take into consideration the relative displacement of objects in neighboring frames and, explicitly or implicitly, align the information therein with the core information contained in the reference frame. Spatial correspondence between frames is usually expressed with optical flow. Based on optical flow, most deep learning VSR methods either opt for explicit motion compensation, or forgo compensation and use the flow as an additional input, implicitly letting the network compensate. Both approaches, however, have drawbacks.

Explicit motion compensation approaches [1,9,18,21] use optical flow to warp each neighboring frame, producing a warped frame that is spatially aligned with the reference frame. As optical flow values are floating-point, pixel values in the warped frame are calculated by interpolation, based on the original pixels of the neighboring frame. Interpolation, however, degrades image quality by introducing irreversible error, resulting in lowered contrast and blur [12].

No warping approaches [4,8,23], instead, use optical flow as an additional input layer, and let the neural network learn to align information based on this input. This approach, although effective, necessitates a large receptive field in

order to capture faster motions, resulting in an increased number of parameters and correspondingly slower training and execution times. Moreover, as the placement of the region of interest within the large receptive field varies, it becomes harder for the network to adapt and focus only in the relevant information.

Driven by the above observations, the present work proposes to split optical flow into an integer and a fractional part, using the former for interpolation-free "quantized" warping, and the latter as an additional input to the network. This approach combines the advantages of both warping and no-warping, retaining a small relative displacement, while avoiding interpolation errors, and making efficient use of all available information.

Fig. 1 shows pixel correspondence and information usage in the two approaches mentioned above as well as the proposed approach, showcasing the pros and cons of each with a simplified example. Taking a 5×5 pixel area, the example focuses on calculating a new SR value for the green pixel of interest. The corresponding position in the neighboring frame, according to optical flow, lies 1.7 pixels to the right and 1.4 pixels down.

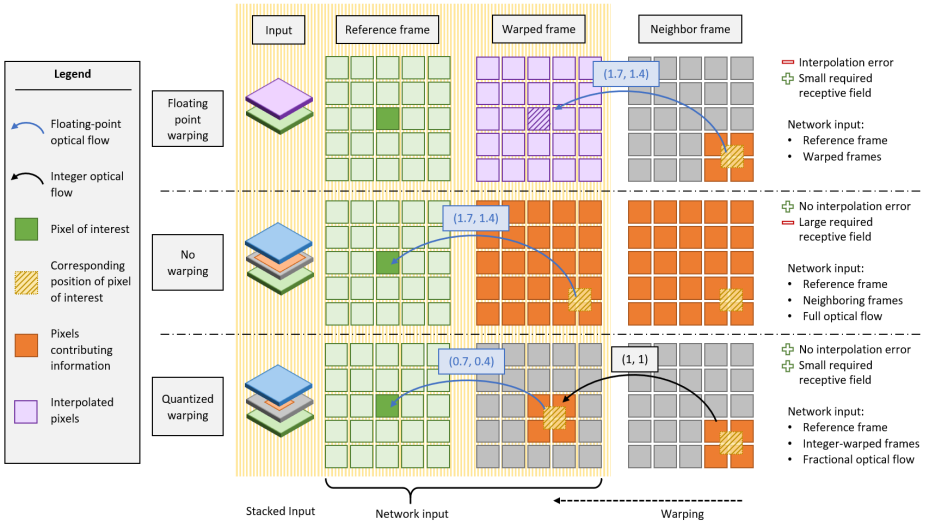

Fig. 1: A visual overview of different approaches to optical flow warping for super resolution. The top row describes floating point warping, where interpolation is used to warp the neighboring frame into exact spatial alignment with the reference. The middle row depicts foregoing warping and alignment, using instead a larger receptive field and the optical flow as an additional input to the network. The bottom row describes the proposed approach, constructing a roughly aligned warped image using only the integer part of the optical flow, and providing the fractional part as an additional network input. The last column summarizes the pros and cons of each approach, along with their network input.

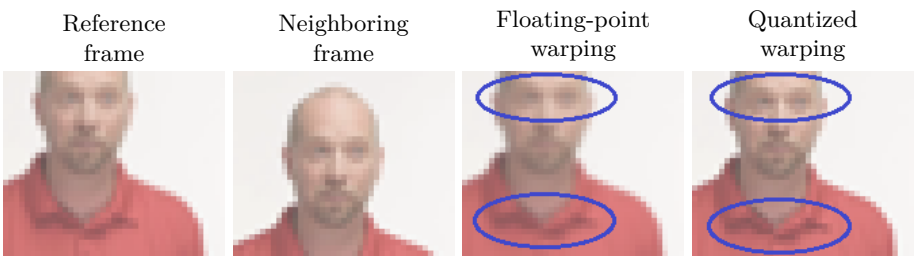

Fig. 2: Illustration of the the blurring effect of traditional alignment warping. In the image warped with floating-point optical flow, note the loss of detail in the eyes and the reduced contrast in the shadows of the shirt's collar. Compare with the corresponding regions in the image produced with quantized warping

In floating-point warping, the warped frame is calculated by taking the optical flow vectors and interpolating between pixels. The purple interpolated pixel corresponds exactly to the green pixel of interest. As all information is spatially aligned between referenced and warped frame, the network need not have a large receptive field; in theory, even a 1×1 receptive field could be enough, though usually a somehow wider field is used to also extract information from neighboring pixels. The reference and warped frames are then stacked and used as input to the network.

In the case of no warping, the neighbor frame itself is used as an input, along the reference frame. The optical flow field provides an additional input. Here, the receptive field must be at least as large as the maximum motion vector length. A large number of pixels from the neighboring frame will contribute to the end result, and the network must learn to focus on the most relevant pixels according to the optical flow input.

Finally, in the proposed approach of quantized warping, the warped frame is computed by shifting pixels in the neighbor frame by the integer part of the optical flow vectors. Hence, the spatial correspondence between the reference and warped frame is not perfect, as in floating-point warping, but displacements are confined to the [0, 1) range, for arbitrary large displacements. The minimum required receptive field here is 2x2 pixels, although again this can be widened to extract additional information from neighboring pixels. The network input consists of the reference and warped frames, as well as the fractional part of the optical flow field, which was not used for warping. Therefore, the network must learn to take into account the displacement between the reference and warped frames, but this is now only a sub-pixel displacement. Displacements larger than 1 pixel are offset during the warping phase, allowing the network to compensate for faster motions without a large receptive field.

Fig. 2 illustrates floating-point warping's interpolation error in a zoomed detail from a real video sequence. The third image, produced by warping the neighboring frame with the full optical flow, exhibits blur, loss of detail, and

reduced contrast. By contrast, the image produced by quantized warping retains the same level of sharpness as the original.

4 Network Architecture

Let $\{I_{t-N}, \ldots, I_t, \ldots, I_{t+N}\}$ be a sequence of $2N+1$ LR consecutive frames. We denote I_t as the reference frame, I_{t+n}, $n \in [-N, N]$, $n \neq 0$ as the neighboring frames and $F_{t->t+n}$ as the flow between I_t and I_{t+n}. The aim of VSR is to reconstruct a HR version of I_t, denoted by I_t^{SR}, by exploiting the information of the neighboring frames.

The proposed network architecture follows a recurrent structure that progressively reconstructs the I_t^{SR} image by adding, in each iteration, extra information from the neighboring frames using the back-projection process [7], inspired by the RBPN [4]. The proposed network comprises 3 processing stages. In the first stage, denoted as Shallow Feature Extraction, features are extracted from the available LR data. Next, in the Back-projection module, the basic processing to produce the respective HR features is performed and, finally, in the Reconstruction stage, the SR image is composed. This procedure is performed repetitively for each new frame used. The overall proposed architecture is depicted in Fig. 3.

Shallow Feature Extraction: The input LR I_t frame is passed through a convolutional layer to extract the initial LR features maps, S_t. Moreover, for each neighboring frame I_{t+n}, the corresponding warped frame I_{t+n}^{WARPED} w.r.t the reference frame is computed based on the proposed 'quantized' warping method. To warp a pixel of the neighboring frame to the reference, only the integer part of the optical flow $F_{t->t+n}$ is used. However, the integer part does not contain the precise displacement information and therefore, for each pixel, we utilize 4 warps with all 4 neighboring pixels in order to fully preserve the subpixel motion information, as shown in Fig. 4. Therefore, for one neighboring frame, 4 warped images are computed that are stacked to produce the corresponding warped frame I_{t+n}^{WARPED}.

Finally, the reference frame I_t, the neighboring warped frame I_{t+n}^{WARPED} and the fractional part of the of pre-computed flow map $F_{t+n}^{fractional}$, that was not used in warping process, are concatenated and given as input to a convolutional layer to produce feature maps M_{t+n}. The S_t and the M_{t+n} feature maps represent the single-scale and the multi-scale information, respectively.

Back-projection: The back-projection module combines the single and the multi-scale information by projecting the reference features S_t to each neighboring frame's features M_{t+n} in order to capture missing information. It takes as input the S_t and the M_{t+n} feature maps and outputs the refined HR feature maps H_{t+n} and the next LR features S_{t+n}, using convolutional structures.As shown in Fig. 5, the module, first, produces the refined HR H_{t+n} maps through the back projection to particular neighbor frame. Then downscales H_{t+n} to output the next LR features S_{t+n}. The whole process is described as follows:

$$H_{t+N-1}^{S} = \mathrm{Net}_S(S_{t+N-1}) \tag{1}$$

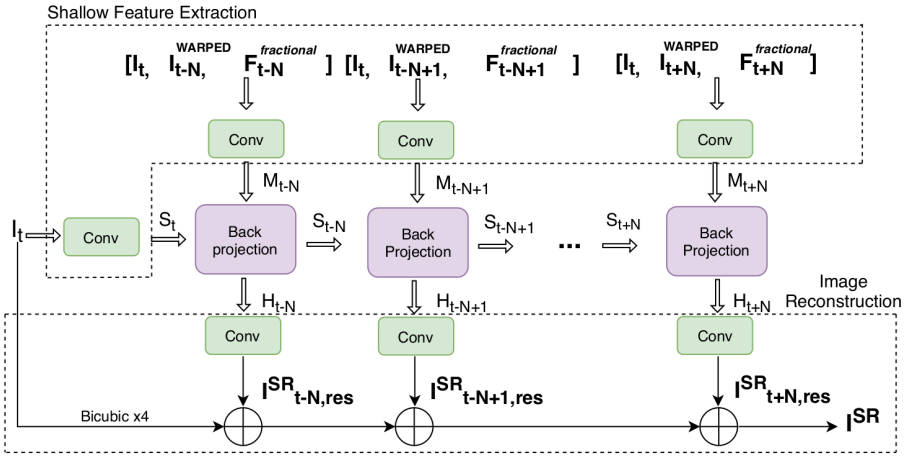

Fig. 3: Illustration of the unfolded architecture of the proposed recurrent network

$$H^M_{t+N} = \text{Net}_M(M_{t+N}) \tag{2}$$

$$H_{t+N_{res}} = \text{Net}_{res}(H^S_{t+N-1} - H^M_{t+N}) \tag{3}$$

$$H_{t+N} = H^S_{t+N-1} + H_{t+N_{res}} \tag{4}$$

$$S_{t+N} = \text{Net}_{downscale}(H_{t+N}), \tag{5}$$

where Net_S, Net_M, Net_{res} and $Net_{downscale}$ are the respective convolutional networks for each task.

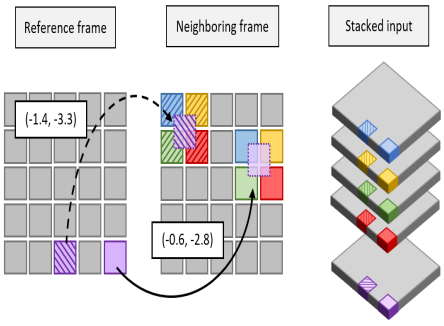

Fig. 4: The 4 warped images that are produced for each neighboring frame using our proposed quantized method

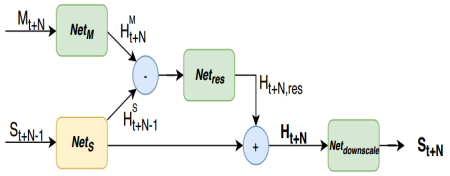

Fig. 5: Back-projection module

Image Reconstruction: In the proposed network, the image reconstruction module follows a temporal integration strategy in order to produce the final

super-resolved image I_t^{SR}. A recurrent residual reconstruction process has been developed that progressively enhances the produced I_{t+n}^{SR} at the image level by adding further information from each neighboring frame. This strategy exploits directly and efficiently the extra information from the neighbors and, in the last iteration, outputs a refined, detailed I_t^{SR} image.

Consequently, unlike the majority of VSR approaches that require distinct models for accepting a different number of frames as input, the proposed reconstruction architecture enables the same network, trained on fixed number of frames, to handle frame sequences of arbitrary length. The number of neighboring frames used in the inference phase depends on the desired inference speed and reconstruction quality.

It is evident, though, that the task is characterized by an inherent temporal locality, with the majority of useful information being on frames temporally adjacent to the reference one. The reconstruction process is formulated as:

$$I_t^{SR} = I_t^{bic} + I_{t-N,res}^{SR} + I_{t-N+1,res}^{SR} + \cdots + I_{t+N,res}^{SR} \tag{6}$$

where I_t^{bic} is the bicubic upscaled version of I_t.

This procedure is repeated until all available neighboring frames have been processed.

5 Experimental Results

5.1 Implementation and Training Details

All models are trained with the Vimeo-90k [24] dataset, which consists of 64612 7-frame sequences and contains diverse scenes and motions. For testing we use the standard benchmark datasets including Vid4 [14], and Vimeo-90k-T. The performance of the models is evaluated using the PSNR and SSIM quality metrics, both on the RGB color space and on the Y-channel (luminance) from YCbCr color space. By following [4], we crop $2s$ pixels around image boundary at testing phase, where s is the scale factor. Additionally, we remove the first and last 3 frames of the sequence. For our main model, we use a 3-stage DBPN [3] for Net_S and a 5-block ResNet [5] for Net_M, Net_{res}, $Net_{downscale}$, based on [4]. Each ResNet block consists of 2 convolutional layers with a 3×3 kernel and the up-sampling layer is a transposed convolutional layer with an 8×8 kernel, stride 4 and padding 2. The optical flow information is extracted using the implementation by [15].

During the training phase, RGB patches with size 64×64 are randomly cropped form the LR input images and the mini-batch size is set at 4. The extracted patches are augmented with vertical and horizontal flipping and rotation. The Adam optimizer is used for model's parameter update with $\beta_1 = 0.9$ and $\beta_1 = 0.999$. All proposed models are trained using L_1 norm as loss function with initial learning rate 10^{-4}, which decreases by a factor of 10 every 75 epochs.

Table 1: Quantitative evaluation of state-of-the-art VSR methods on Vimeo-90K and REDS dataset. Red and blue indicate the best and the second best performance (PSNR/SSIM)

Motion type	Bicubic (1 Frame)	DUF [8] (7 Frames)	EDVR [23] (7 Frames)	RBPN [4] (7 Frames)	Proposed (7 Frames)
Vimeo-90k-T (Y)	31.32/0.8684	36.37/0.9387	37.61/0.9489	37.16/0.9420	37.23/0.9445
REDS (RGB)	26.14/0.7292	28.63/0.8251	30.49/0.8700	29.84/0.8538	30.50/0.8698

5.2 Results on Large Motions and Generalization

We compare our proposed network with the 3 most prominent state-of-the-art methods in VSR, namely DUF [8], RBPN [4] and EDVR [23]. Testing is done on the most challenging and diverse datasets:Vimeo-90k-T and REDS [16].

Vimeo-90k-T is a large and commonly used dataset that contains diverse HQ data and a range of motion types. For evaluation on more challenging data, we test with REDS, that consists of high resolution HQ images, with larger and more complex motions. For the following results we trained our model on Vimeo-90k for upscaling x4 and using 6 neighboring frames, 3 past and 3 future ones.

In table 1, the quantitative evaluation of SoA VSR methods on the most challenging datasets, Vimeo-90k and REDS, is presented. For the results of this table we use networks trained on Vimeo-90k dataset for EDVR, RBPN and the proposed method. First of all, we can see that the proposed method presents a clear improvement over RBPN and DUF on both datasets, with the difference being bigger on REDS. Compared to EDVR, which is the current SoA, the performance of the proposed method is worse on Vimeo-90k but on par on REDS, despite having 8 million parameters less. These results indicate that our model generalizes better on unknown data, irrespective of the training data, and does not suffer from dataset overfitting issues. The fact that our model closes the performance gap with the SoA on the most complex and realistic dataset, indicates that our motion compensation strategy is successful. Qualitatively, the proposed model is capable to recover high frequency details and more accurate textures compared to existing methods, as shown in Fig. 8 on examples obtained from Vimeo-90k, REDS and Vid4.

To further illustrate the merit of our approach, we thoroughly compare our method with RBPN on different Vimeo-90k splits with different motion characteristics. RBPN's architecture is also based on back-projection modules and mainly differs from our model in motion information handling at the input. RBPN uses no warping or any other motion compensation for neighbouring frames. As can be seen on tables 2 and Fig. 6, the proposed method increasingly outperforms RBPN as the motion magnitude grows larger, with the difference reaching 0.19*db*.

5.3 Ablation Studies

In order to validate the additive value of each of our contributions, we implement them one by one and conduct relevant experiments and comparisons. The architectures mentioned throughout this section are smaller versions of the proposed model, to allow for shorter training duration. The total number of parameters is reduced from ≈ 12 to ≈ 1.8 million by reducing the ResNet blocks of each back-projection module from 5 to 2, the feature number of de-convolutional layers and DBPN from 64 to 32, as well as reducing the features of each convolutional layer from 256 to 128. If not mentioned explicitly otherwise, the ablation experiments are using 5 input frames in total.

Table 2: Quantitative comparison between RBPN and the proposed network, on Vimeo90k-T

Dataset \ Method	RBPN/F7	Proposed/F7	Diff
Vimeo-90k-T (Y)			
Slow	**34.18**/0.9200	**34.18/0.9221**	0.0
Medium	37.28/0.9470	**37.30/0.9496**	0.02
Fast	40.03/0.9600	**40.22/0.9626**	0.19
Avg.	37.16/0.9423	**37.23/0.9447**	0.07

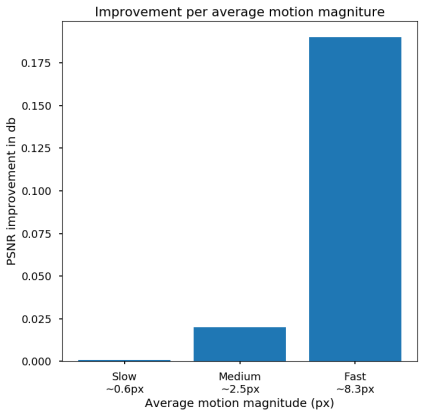

Fig. 6: Improvement over RBPN per motion magnitude

Quantized Warping Effectiveness: To strengthen our claim that the proposed quantized warping process is more suitable for the VSR task than the floating-point warping, we train two separate models with the same structure, parameters and neighbors but with different warping methods at the input level. The one model receives as input the concatenation of the reference frame, the integer warped frame and the fractional optical flow, whereas the other takes as input the concatenation of the reference frame and the floating warped frame. We also compare the above models with a third model, similar to RBPN[4], that uses no warping and relies on implicit motion estimation. The input of this model is a simple concatenation of the reference frame, the neighboring frame and the flow information between them.

Table 3 shows that our quantized warping method outperforms the floating-point one and increases the model's reconstruction performance by 0.44 dB on Vid4 and by more than 1 dB on Vimeo-90k-T, at all motion types. These results show that for the VSR task is more important to maintain the high frequency information of neighboring frames than to achieve a precise motion compensation using a interpolated warping method, which produces blurry input images.

Table 3: Effect of the warping method on the Vid4 and Vimeo-90k-T datasets for upscaling factor 4

Dataset \ Method	No warping	Floating-point warping	Proposed quantized warping
Vid4	**26.68/0.801**	26.24/0.786	**26.68/0.801**
Vimeo-90k-T			
Slow	**33.47/0.9120**	32.90/0.9038	**33.47/0.9120**
Medium	36.48/0.9418	35.30/0.9273	**36.53/0.9423**
Fast	39.26/0.9551	38.16/0.9431	**39.54/0.9580**
Avg.	36.40/0.9363	35.45/0.9247	**36.51/0.9374**

Notice that floating point warping causes blur even for motions smaller than 1 pixel, which explains the reduced performance on the slow split.

Compared to no warping method, we notice that the two approaches provide similar results for slow motions -Vid4, Vimeo90k-slow- and the effectiveness of our method is more clear at medium and fast motions. These results are expected as for small displacements the network with no warping is capable to focus on the relevant information between the reference and the neighboring frame. However, for faster motions is difficult for the network to capture the larger region of interest and align implicitly the useful information, despite the final large receptive field. Our method provides the network with warped detailed images that make easier and more efficient the reconstruction process.

Usage of Subpixel Motion Information: As described in Section 4, our warping method utilizes all 4 neighboring pixels to fully preserve subpixel motion information. Using a simple, nearest-neighbor interpolation, for each neighboring frame is produced only one warped image based on the information of the nearest pixel. This method would result in inputs with lost displacement information, but crisp details. We evaluated the added value of warping the entirety of the neighborhood by training a network using each method. As seen on table 4, the additional information from all neighboring pixels improves the reconstruction ability of the network.

Sequence Length: We train our model with video sequences of different lengths to investigate how the number of the neighboring frames affects the reconstruction performance. As shown in table 5, the model's performance improves with longer sequences, as the network benefits from the extra relevant information. The F7 model, with 3 past and 3 future neighbors outperforms the F5 model, with 2 past and 2 future neighbors, by more than 0.2 dB on both datasets.

Table 4: Effect of neighbors per pixel on Vid4 and Vimeo-90k-T datasets

Dataset \ Method	1 Neighbor	4 Neighbors
Vid4	26.66/0.8005	**26.68/0.8006**
Vimeo-90k-T (avg.)	36.48/0.9374	**36.51/0.9374**

Table 5: Effect of sequence length on Vid4 and Vimeo-90k-T dataset

Dataset \ Method	Proposed F5	Proposed F7
Vid4	26.68/0.8006	**26.90/0.8104**
Vimeo-90k-T (avg.)	36.51/0.9374	**36.76/0.9402**

Residual Temporal Integration Module: We validate the efficiency and the flexibility of our model to handle arbitrary number of frames. We observed that by training our recurrent model with a fixed number of input frames, its performance sharply deteriorates when presented with inputs of different length. To overcome this, we trained a model with sequences of varying length. As can be seen in Fig. 7, the resulting model achieves more stable performance across a wider range of input lengths.

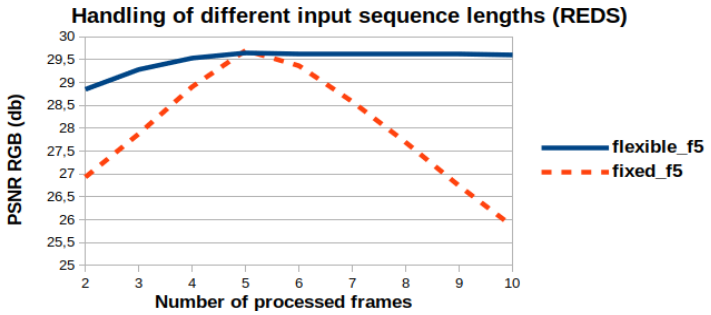

Fig. 7: Performance curves on REDS for models trained using different input lengths (varying between 2–5 frames for the solid-blue, 5 for the dashed-orange).

6 Conclusion

In this paper, we propose a quantized warping method and a residual temporal integration module combined to a flexible VSR framework that generates high quality results and outperforms most of the previous approaches on big and complex motions. In an extensive set of experiments, we show that the proposed warping method is more suitable for the VSR task compared to floating-point warping or no warping and boosts the model's performance. The quantized warping method is a general algorithm that could be used to other tasks, which focus

more on detailed warped images than precise alignment, beyond VSR. Moreover, the residual temporal integration module allows the network to be flexible to frame sequences with arbitrary length without extra training.

The proposed method is better suited to take advantage of neighboring frames with big displacements, due to rapid motion or temporal frame distance, as it explicitly compensates the relative displacements and lets the network model only the remaining, subpixel displacements. Consequently, it generalizes better across different datasets with diverse motion content, showing performance competitive to the SoA, despite using a significantly smaller network and generic optical flow.

Acknowledgements: This research was funded by the European Commission under contract H2020–787061 ANITA

References

1. Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., Shi, W.: 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) 3, 4
2. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015) 3
3. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1664–1673 (2018) 3, 9
4. Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3897–3906 (2019) 4, 7, 9, 10, 11
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 9
6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017) 3
7. Irani, M., Peleg, S.: Improving resolution by image registration. CVGIP: Graphical models and image processing **53**(3), 231–239 (1991) 3, 7
8. Jo, Y., Wug Oh, S., Kang, J., Joo Kim, S.: Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3224–3232 (2018) 4, 10
9. Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A.K.: Video super-resolution with convolutional neural networks. IEEE Transactions on Computational Imaging **2**(2), 109–122 (2016) 3, 4
10. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016) 3

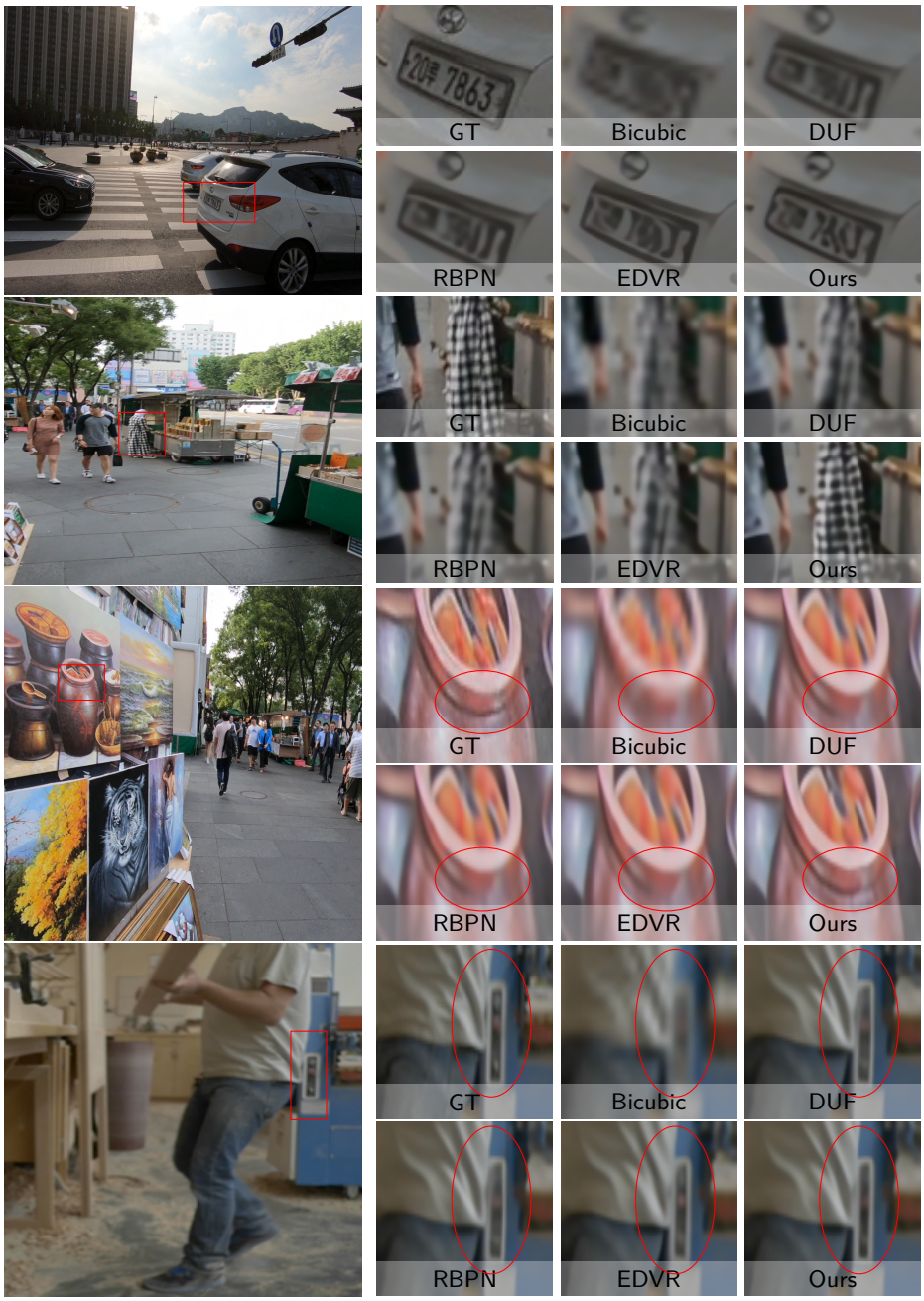

Fig. 8: Visual results on REDS and Vimeo-90k. Zoom in to see better visualization

11. Kim, J., Kwon Lee, J., Mu Lee, K.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1637–1645 (2016) 3
12. Konstantoudakis, K., Vrysis, L., Tsipas, N., Dimoulas, C.: Block unshifting high-accuracy motion estimation: A new method adapted to super-resolution enhancement. Signal Processing: Image Communication **65**, 81–93 (2018) 4
13. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017) 3
14. Liu, C., Sun, D.: A bayesian approach to adaptive video super resolution. In: CVPR 2011. pp. 209–216. IEEE (2011) 9
15. Liu, C., et al.: Beyond pixels: exploring new representations and applications for motion analysis. Ph.D. thesis, Massachusetts Institute of Technology (2009) 9
16. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Mu Lee, K.: Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019) 10
17. Sajjadi, M.S., Scholkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4491–4500 (2017) 3
18. 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) 4
19. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3147–3155 (2017) 3
20. Tai, Y., Yang, J., Liu, X., Xu, C.: Memnet: A persistent memory network for image restoration. In: Proceedings of the IEEE international conference on computer vision. pp. 4539–4547 (2017) 3
21. 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) 3, 4
22. Tong, T., Li, G., Liu, X., Gao, Q.: Image super-resolution using dense skip connections. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4799–4807 (2017) 3
23. 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 Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019) 4, 10
24. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. International Journal of Computer Vision **127**(8), 1106–1125 (2019) 9
25. Yang, W., Feng, J., Yang, J., Zhao, F., Liu, J., Guo, Z., Yan, S.: Deep edge guided recurrent residual learning for image super-resolution. In: IEEE Transactions on Image Processing. pp. 5895–5907 (2017) 3
26. 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) 3