SN Computer Science (2024) 5:323
https://doi.org/10.1007/s42979-024-02710-x

ORIGINAL RESEARCH

Frame Selection Using Spatiotemporal Dynamics and Key Features as Input Pre-processing for Video Super-Resolution Models

Arbind Agrahari Baniya[1] · Tsz-Kwan Lee[1] · Peter Eklund[1] · Sunil Aryal[1]

Received: 27 January 2023 / Accepted: 13 February 2024

Abstract

This paper presents a novel approach to video super-resolution (VSR) by focusing on the selection of input frames, a process critical to VSR. VSR methods typically rely on deep learning techniques, those that are able to learn features from a large dataset of low-resolution (LR) and corresponding high-resolution (HR) videos and generate high-quality HR frames from any new LR input frames using the learned features. However, these methods often use as input the immediate neighbouring frames to a given target frame without considering the importance and dynamics of the frames across the temporal dimension of a video. This work aims to address the limitations of the conventional sliding-window mechanisms by developing input frame selection algorithms. By dynamically selecting the most representative neighbouring frames based on content-aware selection measures, our proposed algorithms enable VSR models to extract more informative and accurate features that are better aligned with the target frame, leading to improved performance and higher-quality HR frames. Through an empirical study, we demonstrate that the proposed dynamic content-aware selection mechanism improves super-resolution results without any additional architectural overhead, offering a counter-intuitive yet effective alternative to the long-established trend of increasing architectural complexity to improve VSR results.

Keywords Video super-resolution · Deep learning · Image similarity metrics · Input selection · Data pre-processing · Video processing · Video quality

Introduction

Super-Resolution for generating high-resolution visuals from low-resolution inputs is a classic problem in computer vision. Its initial solution was provided by Image Super-Resolution (ISR), which only utilises spatial information of a single image or multiple discrete images to produce fundamental visual quality improvements [1, 2]. Extending the target super-resolving subject from single images to video signals, the adoption of super-resolution approaches used in conventional ISR to Video Super-Resolution (VSR) fails to capture the temporal information present in videos [3, 4]. VSR aims to adopt several temporally correlated low-resolution frames within a video sequence to super-resolve the frame series. The cross-consideration of spatial and temporal dimensions across multiple input frames induces VSR as a highly non-linear multi-dimensional problem that remains an active research field.

In recent years, Deep Neural Networks (DNN) have been widely adopted in the VSR to leverage highly non-linear multi-dimensional characteristics and features in the input video frames with promising results [5]. These learning-based VSR approaches [6–13] utilize temporal information in a video as a learning feature followed by stages of frame alignment and fusion to reconstruct and up-sample the resulting pixels. However, their use of commonly adopted frame alignment techniques, conventional Motion

This article is part of the topical collection "Advances on Signal Processing and Multimedia Applications" guest edited by Andrew Sung and Simone Santini.

✉ Arbind Agrahari Baniya
a.agraharibaniya@deakin.edu.au

Tsz-Kwan Lee
glory.lee@deakin.edu.au

Peter Eklund
peter.eklund@deakin.edu.au

Sunil Aryal
sunil.aryal@deakin.edu.au

[1] School of Information Technology, Deakin University, Geelong, Australia

Published online: 14 March 2024

Estimation and Motion Compensation (MEMC) using optical flow and warping [12], or modern machine learning technologies such as deformable convolution [14] may be sensitive to large changes in luminance [15, 16] and motion [5]. To counter, 2D/3D and recurrent convolutions have been used to learn inter-frame correlation without any implicit or explicit frame alignment [17].

To reveal inter-frame correlation along a video sequence without any implicit or explicit frame alignment, the input frames adopted to be learned are commonly based on a sliding window mechanism, including a fixed number of consecutive frames from either past and/or future timestamps to the target frame [18]. Most VSR models using such a sliding-window mechanism treat all neighbouring input frames as equally important without rank or preference. However, each neighbouring frame in a sliding window may express a different correlation because of the context and content changes across the time domain. As a result, a fixed selection of consecutive frame(s) from the target frame in a sliding window may not be optimal for learning spatiotemporal correlation [7].

A fixed number of consecutive neighbouring frames to/from a given target frame in video super-resolution (VSR) models can impede the ability of these models to capture the temporal context of the video sequences. This limitation can result in a lack of information about motion and changes in the scene, negatively impacting the performance of VSR models. To address this issue, recent VSR models have employed all-frame-in bidirectional neural networks, which benefit from the information available from a larger temporal window for each given timestamp. However, these models are complex and may not be applicable to real-world applications because of the need for all frames to be available simultaneously with considerable time and memory requirements [17].

A practical alternative is to apply an efficient frame selection mechanism to the conventional sliding window mechanism in VSR models. By comparing frames within the sliding window and selecting the most relevant frames to the target frame, VSR models can extract more discriminate features required for the super-resolution task. The relevance of frames can be defined in various measures of similarity based on properties such as features, visual appearance, luminance, and element structure.

In this study, we aim to investigate the potential impact of using image comparison measures for input frame selection in VSR. Despite the intuitive reasoning behind this approach, it has yet to be explored in the literature in a comprehensive way. To address this gap, we conduct an analysis of image similarity measures and develop two dynamic content-based input frame selection algorithms for VSR: the SpatioTemporal Input Frame Selection (STIFS) algorithm and the Feature-based Input Frame Selection (FIFS) algorithm. Through an empirical study, we evaluate the performance of these algorithms compared to conventional sliding window methods. Additionally, we extend the applicability of the best-performing selection algorithm to a state-of-the-art 360° video super-resolution model. Overall, the key contributions of this paper are:

1. an analysis of the effectiveness of widely used image similarity measures for input frame selection in VSR.
2. the development of two dynamic content-aware input frame selection algorithms for VSR, namely STIFS [19] and FIFS.
3. an empirical study evaluating the impact of input selection using the proposed algorithms for VSR compared to the conventional sliding window method.
4. an application of the best-performing frame selection algorithm to the state-of-the-art 360° video super-resolution model.

Background

Sliding Window-Based VSR and its Limitations

Using frame alignment in VSR, Motion Estimation and Motion Compensation (MEMC) [6, 9, 20, 21] remains challenging, particularly when inter-frame motion or luminance variance is evident across neighbouring frames [22]. Alternatively, deformable convolutions proposed by Dai et al. [14] have been used for frame alignment by enhancing DNN's capacity to model the transformation of geometric variations of objects. Although deformable convolution is more tolerant to variance in luminance or motion, it involves higher computational overhead [7, 10, 23]. Recently, more VSR methods have been proposed that do not rely on frame alignment techniques to alleviate the above-mentioned limitations. These methods promote 2D convolution [24], 3D convolution [8, 25], or Recurrent Neural Networks (RNN) [17, 26, 27] to exploit spatial or spatiotemporal information in a video.

However, most VSR models simply use a fixed set of consecutive frames for super-resolving each target frame in the video. Some recent methods have introduced variations of the model architectures to extract different features from the given consecutive frames attempting to capture the unique temporal characteristics between video frames. Enhanced Deformable Convolution Networks (EDVR) [7] make use of a Temporal-Spatial Attention (TSA) mechanism where convolution-based similarity distance is used to generate temporal attention maps in element-wise multiplication with the original feature maps of the frame and compute a spatial attention mask by a fusion process. Even after being incorporated with complex components like TSA, the information feed via input frames to these models remains the same. This

implies that the models' learning relies only upon the same inputs to map low-resolution frames to a higher-resolution output, even when the operations applied to extract features from the input might vary.

The literature shows that the field lacks a mechanism to effectively select the input frames for either alignment or non-alignment-based VSR models. Non-frame alignment models suffer more from monotony in the input space resulting from the conventional sliding-window mechanism, with the exception of RNN-based models, which commonly use one consecutive frame in addition to the target frame and the hidden state propagated from super-resolving frames from past timestamps. Two of the non-frame alignment-based methods are VSRResFeatGAN [24], and Dynamic Upsampling Filters (DUF) [8], which use 2D and 3D convolution, respectively. Both methods use a sliding window mechanism to select a fixed number of frames from both past and future temporal dimensions and rely on either 2D convolution to extract the spatial correlation or 3D convolution to extract the spatiotemporal correlation. The capability of convolution layers in these models to learn the optimal spatial features in a frame or spatiotemporal features between frames could be limited since the temporal proximity insinuating the cross-correlation, relevance, and mutual information between video frames may not be fully utilised.

VSR Challenges

Although the evaluation and comparison of a new VSR model with the current state-of-the-art VSR models are beyond the scope of this paper, our intention with this discussion is to acknowledge the fierce competition in VSR research and the relatively minor gains achieved via modelling and addressing the complex problem of video super-resolution. As an example, IconVSR [12] harnessed the sequential modelling ability of bidirectional recurrent neural networks in combination with MEMC to obtain peak signal-to-noise ratio (PSNR) improvement of only 0.03 dB over the previously best-performing model, EDVR [7] on the Vimeo90k [21] test set. This exemplifies the challenges in improving the performance of existing VSR models. Interesting to mention is the extent of the changes made to the model to obtain this modest improvement. Similarly, despite the complexity of the model proposed, the recent BasicVSR model can only improve the PSNR on Vid4 by 0.04 dB compared to the previously best-performing model Recurrent Structure-Detail Network (RSDN) [13]. RSDN, in turn, was only able to improve the super-resolution outcome on Vid4 [28], in PSNR terms, by 0.07 dB compared to the EDVR model, the best-performing model preceding RSDN. These examples further emphasize the trend of increasing architectural complexities in the VSR literature to achieve only marginal improvements in super-resolution results. Therefore, any such improvements in super-resolution results without any added architectural overhead would be considered cost-effective, efficient and practical alternatives to increasing architectural complexities. The proposed input frame selection strategy offers such an alternative.

Scope of this Work

Our literature study concludes that, although limited attempts have been made to treat frames at different timestamps differently in some alignment-based VSR methods, no work has been proposed to effectively select the input frames in current VSR models, despite the hypothesis that such an approach will likely benefit the feature space, and may achieve improved super-resolution outcomes, especially for non-frame-alignment based VSR models. At the same time, it is hypothesised that selecting the most relevant input frames will improve VSR results at a lower computational cost compared to models with increased learnable parameters formulating a more computationally expensive approach.

By leveraging temporal information and considering pixel-level and feature-level comparisons of neighbouring frames with the target frame, our proposed input selection algorithms aim to determine the most relevant frames for super-resolution reconstruction. This approach allows prioritising frames with relevant content and visual patterns, thereby capturing and leveraging temporal information effectively. By integrating these algorithms into VSR models, significant improvements in super-resolution outcomes are anticipated, particularly for non-frame-alignment-based methods, while also reducing computational complexity. Therefore, in this study, we aim to explore the effectiveness of employing frame comparison matrices for input selection in VSR and investigate its impact on VSR performance, specifically through the four major contributions that have been highlighted in "Introduction".

Input Selection Mechanisms

Based on the properties used to define relevance and facilitate selection, frame selection mechanisms can be broadly categorized into three types:

1. Pixel-based similarity measures compare the similarity between a given target frame and its neighbouring frames based on the pixel values. These methods are computationally efficient, but they may not be as effective in cases with significant motion or luminance changes between frames. Examples of pixel-based measures include Mean Pixel Value Difference (MPVD), Normalized Cross Correlation (NCC), Correlation Coef-

ficient, and Mutual Information (MI) [29, 30]. Among these methods, MPVD is one of the simplest yet most effective measures for pixel-based comparisons.

2. Quality-based similarity measures compare the similarity between a given target frame and its neighbouring frames based on the quality or visual appearance of the frames. These methods consider factors such as luminance and contrast and are expected to be more effective than pixel-based measures in cases with significant noise or compression artefacts. Examples of quality-based measures include Peak Signal to Noise Ratio (PSNR) [31], Structural Similarity Index (SSIM) [31], and Learned Perceptual Image Patch Similarity (LPIPS) [32].
3. Feature-based similarity measures compare the similarity between a given target frame and its neighbouring frames based on the feature points or descriptors extracted from the frames. These methods can be effective in cases with distinct features in the video frames, such as text or objects. However, these methods are the most computationally intensive among the three approaches. Examples of feature-based measures include shallow features using SIFT (Scale-Invariant Feature Transform) [33], FAST (Features from Accelerated Segment Test) [34], BRIEF (Binary Robust Independent Elementary Features) [35], ORB (Oriented FAST and Rotated BRIEF) [36], BRISK (Binary Robust Invariant Scalable Keypoints) [37] or deep features using VGG16 [38] or ResNet [39]. Using deep features for selection can prove to be computationally expensive in a VSR model. Among the conventional shallow feature detection methods, ORB has been identified as one of the most efficient and robust methods [40].

Analysis of Selection Measures

Pixel-Based vs. Quality-Based

We perform a frame-to-frame comparison between example target frames and their neighbours using the pixel-based method—MPVD and quality-based methods—PSNR and SSIM for all four clips of the Vid4 dataset. For this analysis, we consider the target frame at timestamp $t = 12$ and its 11 neighbours in each temporal direction. From the graphs shown in Fig. 1a and b, it is evident that MPVD is highly correlated with both PSNR and SSIM, justifying the ability of MPVD to capture similarity/difference between frames at a similar level as quality-based metrics.

The computational cost of the MPVD method for selecting input frames in video super-resolution (VSR) is significantly lower compared to the PSNR and SSIM methods, as shown in Table 1. The ORB method, on the other hand, is the most computationally expensive among all the methods presented in Table 1 due to the need to extract features explicitly for each frame before making comparisons and selections. The time computation presented in Table 1 was performed on a machine with an Intel(R) Core(TM) i7-8665U CPU @ 1.90 GHz (2.11 GHz) processor, 16 GB of installed RAM (15.8 GB usable) and running a 64-bit Windows 10 Enterprise operating system.

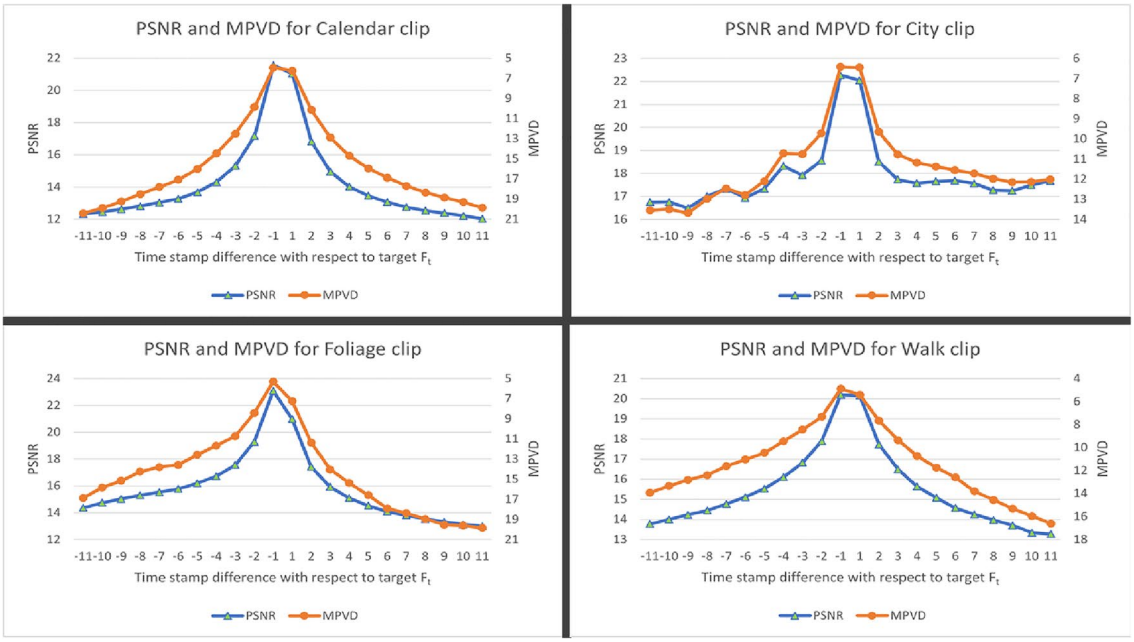
This highlights the need to consider the trade-offs between computation cost and effectiveness of the selection measures for selecting neighbouring frames for a given target frame in VSR, as it is done repeatedly using a sliding window over the entire video. It is important to note that despite the higher computational need for PSNR-based and SSIM-based selections, the nature of selection is highly correlated to MPVD-based selection, as demonstrated in Fig. 1a and b. Therefore, for developing the selection algorithm in the following sections, we consider the MPVD method as an optimal selection measure.

Pixel-Based vs. Feature-Based

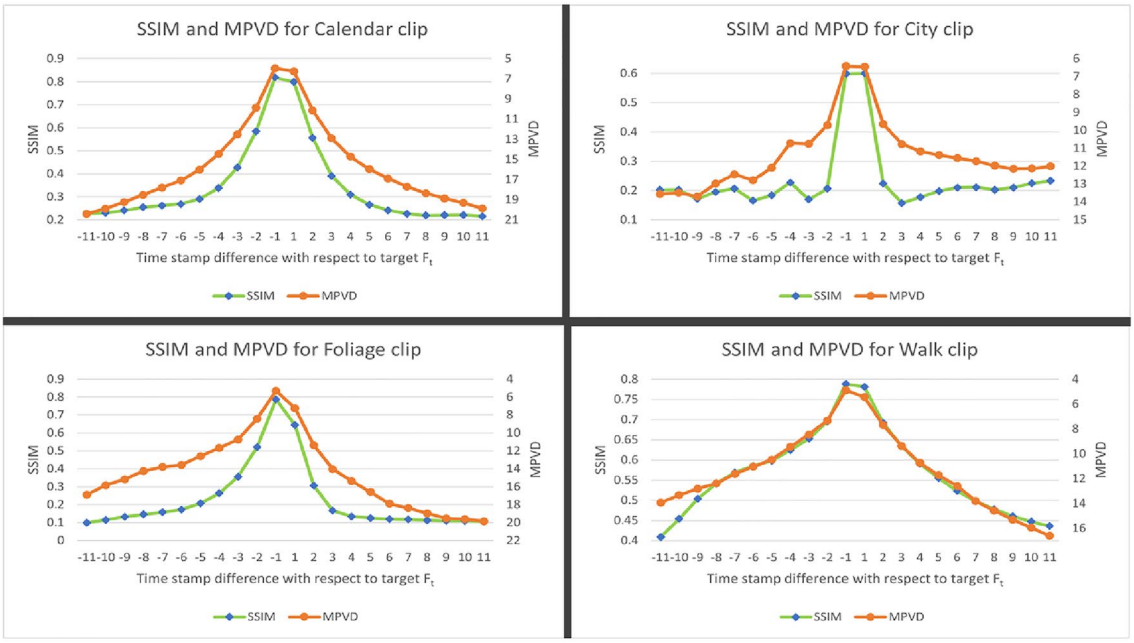
Despite the higher computational cost associated with feature-based selection, it is important to note that the nature of comparison and level of sophistication varies from other approaches. Feature-based methods are well suited for deep learning VSR models for several reasons:

1. Robustness: Feature-based methods can be more robust to changes in luminance and motion, as they extract distinct and consistent features across frames, regardless of the appearance of the frames.
2. Spatial and temporal information: Feature-based methods can extract both spatial and temporal information from the video frames. This can be useful for deep learning models that must capture both information types to generate high-quality HR frames.
3. Scale-invariance: Feature-based methods like ORB are scale-invariant, meaning they can detect feature points across different scales. This can be beneficial in cases where the scale of the objects or the details in the scene change between the frames.

As demonstrated in Fig. 2, the frequency in terms of percentage (%) of target frames for which non-consecutive neighbouring frames were selected varies significantly when using MPVD versus ORB. This illustrates the diversity of these two selection measures. Therefore, in order to study the impact of these measures individually on VSR results, we have also chosen to use ORB for our input selection algorithm development in the following sections.



(a) PSNR and MPVD correlation in 4 clips of Vid4 Dataset.



(b) SSIM and MPVD correlation in 4 clips of Vid4 Dataset

Fig. 1 PSNR, SSIM and MPVD Correlation between target frame F_t, where $t = 12$ and its 11 neighbours in each temporal direction in 4 clips of the benchmark Vid4 Dataset

Spatial vs. Spatiotemporal

MPVD is computationally efficient and provides a mechanism for a spatial comparison between frames. However, MPVD does not yet optimally consider the spatiotemporal inter-dependencies among video frames. We have analysed the sensitivity of using MPVD when used alone for selection and depicted the result in Fig. 2. The figure shows that for most of the clips selected from the Vimeo90k septuplet dataset, MPVD selected more non-consecutive neighbouring frames compared to the ORB method. The higher frequency of selecting non-consecutive neighbouring frames by MPVD

Table 1 Time taken in seconds to perform PSNR, SSIM, MPVD and ORB-based selection for 34 frames of the City clip of Vid4 Dataset

Selection type	Selection measure	Average time (s)
Pixel-based	MPVD	0.06131
Quality-based	PSNR	0.14508
Quality-based	SSIM	1.19330
Feature-based	ORB	3.80651

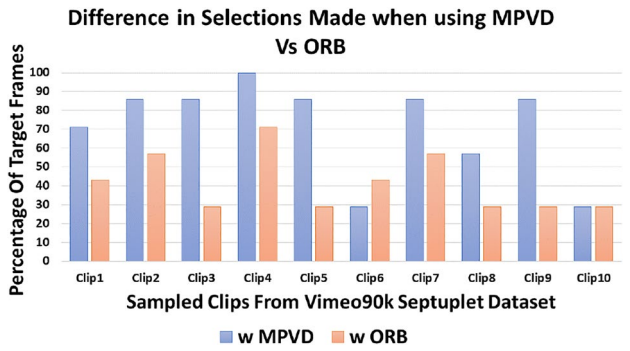


Fig. 2 Comparison between the percentage of target frames for which immediate consecutive past and future neighbouring frames were not selected when using MPVD and ORB as selection measures on ten randomly selected clips of the Vimeo90k [21] dataset

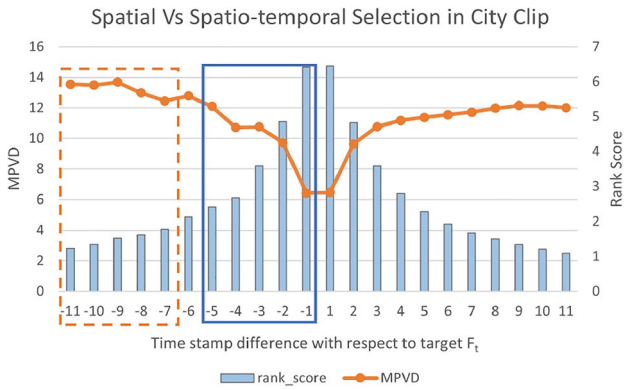


Fig. 3 Comparison between spatial and spatiotemporal selection. The dashed bounding box represents frame selection based on a spatial metric (MPVD) alone. The solid bounding box represents frame selection based on the spatiotemporal metric (MPVD/TD)

may be due to its heightened sensitivity to noise and pixel value variations in the frames.

Additionally, we investigated the impact of MPVD (pixel-based) ranking and selection by considering factors beyond the spatial factors and revealed the spatiotemporal relationship for that. As shown in Fig. 3, if we considered the selection of five out of eleven frames with reference to target frame F_t, where $t = 12$ for the City clip, based on the spatial metric MPVD only, the most distant five frames from the target frame are selected because they exhibit the largest spatial differences, as highlighted by the dotted bounding box in Fig. 3. However, when Temporal Distance (TD) is considered, the most distant frames rank lowest, despite having the largest MPVD with F_t; thus, the nearest five frames are selected, as highlighted by the solid bounding box in Fig. 3. Considering spatial dimension alone inverts the VSR to multi-image super-resolution, which is undesirable. Both spatial and temporal dimensions must be considered to capture true spatiotemporal interdependence between the target frame and its neighbouring frames.

The Proposed Input Frame Selection Algorithms

STIFS for Pixel-Based Selection

To mitigate the shortcomings of the sliding-window approach in current VSR models, our novel SpatioTemporal Input Frame Selection (STIFS) algorithm uses the frame-wise spatiotemporal correlation between neighbouring frames and the target frame to capture their relationship in the input space to a VSR network. The frame-wise spatiotemporal correlation comprises spatial differences and temporal differences between frames. To compute the spatial difference, we make use of the Mean Pixel Value Difference (MPVD) between the target frame F_t and the neighbouring frames $F_{t\pm\delta}$, where $\delta \in \{\pm 1, \ldots, \pm n-1\}$, where n is the total number of frames in the video. MPVD is defined as:

$$MPVD(F_t, F_{t\pm\delta}) = \frac{1}{h \times w} \sum_{j=1}^{h\times w} \|p_j(F_t) - p_j(F_{t\pm\delta})\| \quad (1)$$

where h and w are the height and width of the frames in terms of pixels, respectively; $p_j(\cdot)$ is the value of j^{th} pixel of a given frame.

Algorithm 1 STIFS Algorithm

Input: List of n low-resolution frames for a given video denoted as $frames[1,..,n]$
Output: List of selected frames with three frames for each target frame F_t in the list $final\_input$
Initialisation: $rank\_scores \leftarrow [], final\_input \leftarrow []$
while $i \in \{1,\cdots,n\}$ **do**
  $target(F_t) \leftarrow frames[i]$
  while $j \in \{1,\cdots,n\}$ **do**
    if $i \neq j$ **then**
      $neighbour(F_{t\pm\delta}) \leftarrow frames[j]$
      $MPVD(F_t, F_{t\pm\delta})$ using eqn. (1)
      $TD(F_t, F_{t\pm\delta})$ using eqn. (2)
      $r(F_{t\pm\delta})$ using eqn. (3)
      rank_scores.append($r(F_{t\pm\delta})$)
    else
      rank_scores.append(0)
    end if
    $j \leftarrow j+1$
  end while
  $sorted\_indices \leftarrow argsort(rank\_scores)$
  $selected\_indices \leftarrow sorted\_indices[-2:]$
  $final\_input.append(frames[i])$
  $final\_input.append(frames[selected\_indices[0]])$
  $final\_input.append(frames[selected\_indices[1]])$
  $i \leftarrow i+1$
end while
return $final\_input$

The temporal component of the spatiotemporal correlation is the Temporal Distance (TD) between a target frame F_t and neighbour $F_{t\pm\delta}$ calculated as,

$$TD(F_t, F_{t\pm\delta}) = \|\delta\|. \tag{2}$$

The rank score for each frame $F_{t\pm\delta}$ in the neighbouring space of target frame F_t is then computed as,

$$r(F_{t\pm\delta}) = \frac{\text{MPVD}\ (F_t, F_{t\pm\delta})}{\text{TD}\ (F_t, F_{t\pm\delta})}. \tag{3}$$

The STIFS algorithm then uses the rank scores of neighbouring frames to select two neighbouring frames from the given $n-1$ frames that could belong to either past or future dimensions in reference to the target frame F_t. The overall algorithm for the frame selection to an input space of a VSR model for a given video sequence with the total number of frames n, where each frame is of size $h \times w$, is presented in Algorithm 1. Based on the proposed STIFS Algorithm 1, the selection is repeated for each target frame F_t in a video sequence, finally giving an input space of size $n \times 3$, with two neighbouring frames $F_{t\pm\delta}$ and one target frame F_t for each LR frame in *frames*[1, ..., n]. The algorithm selects neighbouring frames by ranking them while capturing both the spatial and temporal correlation between F_t and each neighbouring frame $F_{t\pm\delta}$.

By considering the rank scores of neighbouring frames and capturing both spatial and temporal correlations, the algorithm dynamically chooses two neighbouring frames, either from the past or future relative to the target frame, for each low-resolution frame in the video sequence. The resulting input space for the VSR model is a collection of the selected frames, which exhibit higher spatial and temporal correlation with respect to the target frame. This frame selection mechanism optimises the utilisation of relevant information for super-resolution, potentially leading to improved reconstruction quality and enhanced visual fidelity in the resulting high-resolution videos.

Algorithm 2 FIFS Algorithm

Input: List of n low-resolution frames for a given video denoted as $frames[1,...,n]$.
Output: List of selected frames with three frames for each target frame F_t in the list $final\_input$.
Initialisation: $final\_input \leftarrow []$, $frame\_descriptors \leftarrow []$
while $i \in \{1,\cdots,n\}$ **do**
    $feature\_points \leftarrow keypoints(frames[i])$
    $descriptors \leftarrow descriptor(feature\_points)$
    $frame\_descriptors.append(descriptors)$
    $i \leftarrow i+1$
end while
while $i \in \{1,\cdots,n\}$ **do**
    $target(F_t) \leftarrow frames[i]$
    while $j \in \{1,\cdots,n\}$ **do**
        if $i \neq j$ **then**
            $neighbour(F_{t\pm\delta}) \leftarrow frames[j]$
            $num\_matches = matching(frame\_descriptors[i], frame\_descriptors[j])$
            $matches.append(num\_matches)$
        else
            $matches.append(0)$
        end if
        $j \leftarrow j+1$
    end while
    $sorted\_indices \leftarrow argsort(matches)$
    $selected\_indices \leftarrow sorted\_indices[-2:]$
    $final\_input.append(frames[i])$
    $final\_input.append(frames[selected\_indices[0]])$
    $final\_input.append(frames[selected\_indices[1]])$
    $i \leftarrow i+1$
end while
$return final\_input$

function KEYPOINTS($frame$)
    $feature\_points \leftarrow []$
    for $pixel$ **in** $frame$ **do**
        $patch \leftarrow select\_pixels(pixel, radius \leftarrow 3)$
        if $brightness(patch)$**not in** $range(brightness(pixel) \pm threshold \leftarrow 31)$ **then**
            $feature\_points.append(pixel)$
        end if
    end for
    $feature\_points \leftarrow sort\_by\_harris\_response(feature\_points)$
    $feature\_points \leftarrow add\_scale\_invariance(feature\_points)$
    $feature\_points \leftarrow add\_rotational\_invariance(feature\_points)$
    return $feature\_points$
end function

function DESCRIPTOR($feature\_points$)
    for $point$ in $feature\_points$ **do**
        $neighbourhood\_window \leftarrow creat\_window(central\_point \leftarrow point, window\_size \leftarrow 31 \times 31)$
        $selected\_pixels \leftarrow random\_selection(neighbourhood\_window, required\_pixels\_number \leftarrow 256)$
        $descriptor\_vector \leftarrow BRIEF(selected\_pixels)$
        $descriptor\_vector \leftarrow add\_rotational\_invariance(descriptor\_vector)$
        $descriptors.apppend(descriptor\_vector)$
    end for
    return $descriptors$
end function

function MATCHING($descriptor\_F_t, descriptor\_F_{t\pm\delta}$)
    $num\_matches \leftarrow BRUTE\_force\_matcher(descriptor\_F_t, descriptor\_F_{t\pm\delta})$
    return $num\_matches$
end function

(a) Conventional 2D frame ($t = 11$) of Walk clip of Vid4 dataset [28].

(b) Equirectangular 360° frame ($t = 15$) of Cafe_Scene_1 clip of 360VDS dataset [41].

Fig. 4 Visualisation of key points identified as part of the FIFS algorithm

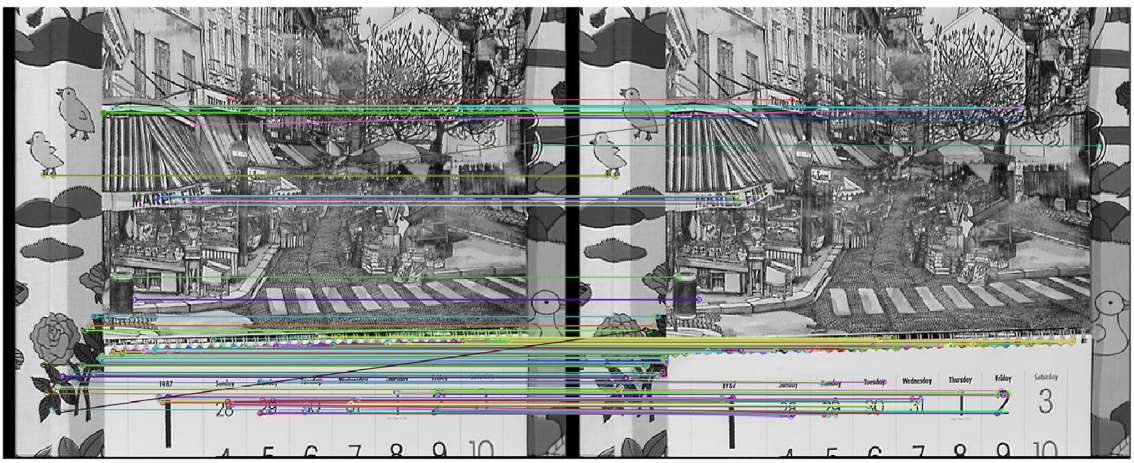


Fig. 5 Visual example of feature matching obtained from FIFS algorithm between frame 7 and 6 of Calendar clip of Vid4 dataset

FIFS for Feature-Based Selection

Feature Descriptor in FIFS

Image features can be broadly categorized as deep features and shallow features. Deep features are extracted using deep learning techniques such as convolution neural networks like ResNet [39] and VGG16 [38]. They are able to capture high-level, semantic information about the content of an image. Shallow features, on the other hand, are extracted using conventional image processing techniques such as edge detection, colour histograms, and texture analysis. These features capture low-level information about the image, such as edges, colour distribution, and texture patterns. Despite the sophistication of information representation in deep features, these are not suitable for selection measures in VSR because of their heavy computational complexity and high memory requirements. They are also not rotation-invariant, which will be easily affected by changes in the orientation across the video frames.

In contrast, shallow features, such as those extracted using ORB [36], are more suited for use as a selection measure in VSR due to their computational efficiency and robustness to changes in image scale and orientation [40]. ORB extractor identifies shallow features by using the Harris corner detector to find key points in a video frame and then extracts binary descriptors based on intensity gradients in a neighbourhood around each point. Figure 4 illustrates the resultant key points identified by ORB feature extractor for conventional 2D video and equirectangular 360° video frames, respectively, by using the proposed FIFS algorithm. By adopting ORB as the ideal method for feature extraction,

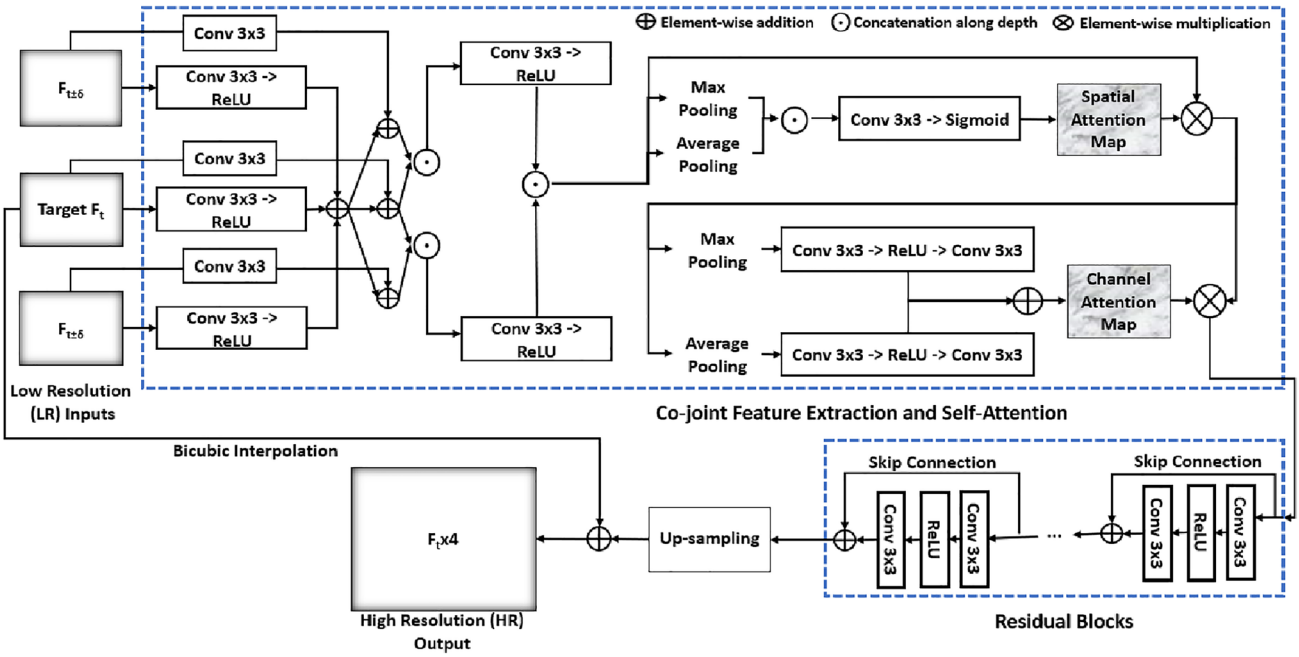


Fig. 6 Prototype VSR model architecture with three input frames

our proposed Feature-based Input Frame Selection (FIFS) is able to offer rotation-invariant and scale-invariant capabilities, implying that it is robust to changes in the frame's orientation and size of the objects in video. ORB also has a relatively low computational cost compared to other shallow and deep feature extraction methods, making it better suited for real-time processing as part of the proposed FIFS algorithm.

Matching the Descriptors in FIFS

Figure 5 illustrates a visual example of a feature mapping result from the proposed FIFS algorithm. The FIFS algorithm successfully identifies relevance between the two frames by adopting the Brute Force Matching technique [42] to match the ORB key point descriptors between two frames. The Brute Force Matching technique enables the proposed FIFS algorithm to compare each descriptor in the target frame (F_t) to every descriptor in the neighbouring frame ($F_{t\pm\delta}$) and finds the closest match. The process is repeated for every descriptor in the target frame, resulting in a set of matches between each pair of frames, as shown in Fig. 5.

The FIFS algorithm adopts the Brute Force matching technique as it is a general-purpose method that does not incorporate any assumptions about data structures and distributions [42]. As a result, it works well on any feature descriptors and can compute the distances between them. The proposed FIFS algorithm enables feature descriptor

Table 2 Super-resolution results in terms of PSNR/SSIM on benchmark Vid4 [28] dataset from Prototype VSR model with varied input frame selection approaches

Clip Name	w FIFS	w STIFS	No Select
Calendar	**22.9902/0.7562**	22.9641/0.7549	22.8012/0.7459
City	**27.0417/0.7609**	26.9856/0.7567	26.8662/0.7491
Foliage	**25.6242/0.7240**	25.6033/0.7225	25.5228/0.7175
Walk	**29.5690/0.8908**	29.4569/0.8884	29.3463/0.8864
Average	**26.3063/0.7830**	26.2525/0.7806	26.1341/0.7747

Bold highlights the highest result value

Table 3 Super-resolution results in terms of PSNR/SSIM on benchmark UDM10 [43] dataset from Prototype VSR model with varied input frame selection approaches

Clip Name	w FIFS	w STIFS	No Select
Archpeople	**36.4118/0.9579**	36.2895/0.9569	36.1367/0.9554
Archwall	41.0347/0.9646	**41.3978/0.9679**	41.0512/0.9649
Auditorium	**29.8865/0.9159**	29.7817/0.9141	29.4926/0.9094
Band	**34.7408/0.9605**	34.5857/0.9590	34.5036/0.9587
Caffe	39.2256/0.9732	**39.4283/0.9738**	38.9971/0.9726
Camera	**46.4702/0.9930**	46.2378/0.9928	45.2736/0.9921
Lake	**31.1916/0.8393**	31.1397/0.8378	31.0689/0.8346
Clap	**36.7847/0.9670**	36.7160/0.9663	36.4288/0.9647
Photography	**37.6781/0.9693**	37.6419/0.9691	37.4740/0.9678
Polyflow	**38.9186/0.9595**	38.7239/0.9580	38.6392/0.9571
Average	**37.2343/0.9500**	37.1942/0.9496	36.9065/0.9477

Bold highlights the highest result value

mapping between a target frame(F_t) and a given neighbouring frame($F_{t\pm\delta}$) by applying the Brute Force matching technique following the ORB feature extraction process. The match function will finally return a number of matches, where each match represents a corresponding key point between the two frames. The steps involved in the proposed FIFS algorithm using ORB Features and Brute Force matching technique are outlined in Algorithm 2.

By leveraging feature-based techniques such as keypoint extraction, descriptor computation, and matching, the FIFS algorithm dynamically selects neighbouring frames for video super-resolution. The algorithm identifies distinctive keypoint features and computes descriptors to capture local structure and appearance information. By comparing these descriptors, the algorithm determines the number of matches between the target frame and other frames, indicating their similarity. The frames with the highest number of matches are selected as neighbours, resulting in an input space that incorporates frames exhibiting high relevance to the target frame. This feature-based frame selection approach employed by the FIFS algorithm enhances the utilisation of relevant information for super-resolution, potentially resulting in improved frame reconstruction accuracy and enhanced observable details in the resulting high-resolution videos.

Empirical Study

A two-staged empirical study is conducted to investigate the effects of proposed selection algorithms on super-resolution performance. The impacts of the proposed algorithms are firstly evaluated for a prototype VSR model as discussed in "Selection in Prototype VSR Model". A sophisticated state-of-the-art 360^{c}*irc* video super-resolution model is then considered in "Selection in State-of-the-Art VSR Model" to explore the applicability of proposed selection algorithms to a more complex and challenging task, specifically in the context of 360^{c}*irc* video super-resolution. The results of this study provide a deeper understanding of the potential benefits of using selection algorithms in VSR tasks.

Selection in Prototype VSR Model

A prototype VSR model is built to facilitate the empirical study of the impact of input frame selection algorithms. As shown in Fig. 6, this model is based on the residual convolution neural network architecture with a total of 5.4 million parameters. The model employs a feature extraction module composed of a series of convolution layers to extract the spatial and spatiotemporal information from the input video frames. It is a non-alignment model that uses co-joint feature extraction between a target frame (F_t) and its two neighbouring frames ($F_{t\pm\delta}$) to allow the extraction of unique temporal characteristics between them, even without any implicit or explicit frame alignment. Additionally, the module includes a self-attention mechanism, which uses spatial attention and channel attention to enhance the model's ability to focus on the most relevant features in the input frames.

The extracted features are then refined through a series of ten residual blocks, each consisting of two convolution operations, a ReLU activation layer and a skip connection, as shown in Fig. 6. The feature refinement step using residual blocks is then followed by up-sampling using a pixel-shuffle operation. The model is a residual learning model which learns the residue feature that is added element-wise to the bicubically interpolated target frame input (F_t) to generate

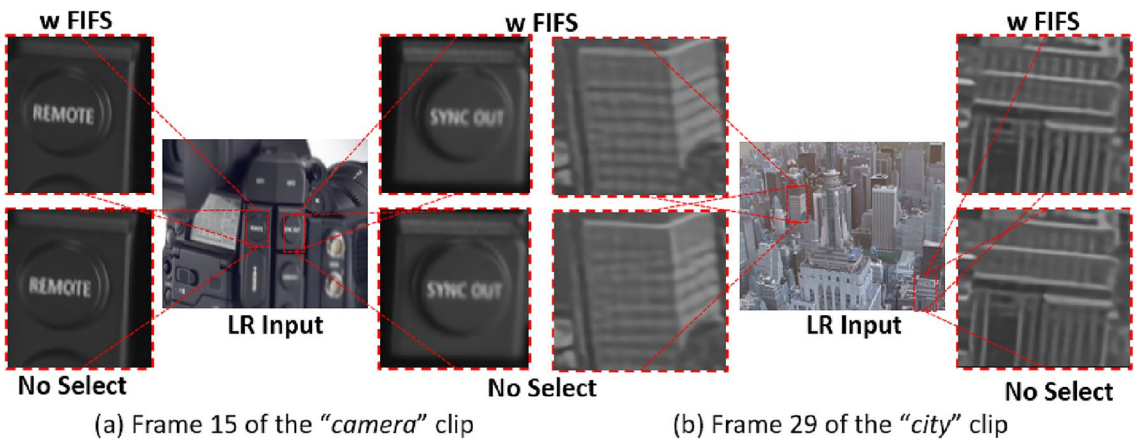


Fig. 7 Subjective inspection of visual quality generated by prototype model when using selection (w FIFS) compared to the conventional sliding window with no selection on the "camera" clip (**a**) and "city" clip (**b**)

Table 4 Comparative evaluation showcasing the impact of using selection in input space versus no selection on state-of-the-art 360° VSR model - S3PO with 360 Video Dataset [41]

Clip name	PSNR		SSIM		WS-PSNR		WS-SSIM	
	w FIFS	No select	w FIFS	No select	w FIFS	No select	w FIFS	No select
deHERAOrly1	**27.5318**	26.8693	**0.8804**	0.8657	**26.1049**	25.4074	**0.8505**	0.8297
deHERAOrly2	**26.9624**	26.4136	**0.8750**	0.8609	**25.4831**	24.9165	**0.8442**	0.8243
MoscowOlimp	**29.0714**	29.0498	**0.8816**	0.8810	**27.5669**	27.5410	**0.8506**	0.8502
360VR	**29.4168**	29.3857	**0.9074**	0.9072	**28.1489**	28.1265	**0.8899**	0.8896
BuzzLightyears	**30.8703**	29.8864	**0.9455**	0.9367	**29.1903**	28.1781	**0.9263**	0.9133

Bold highlights the highest result value

the corresponding high-resolution output frame($F_t \times 4$) as shown in Fig. 6.

Model Training

This prototype VSR model was trained on the Vimeo90k train set until learning saturation occurred at thirty epochs. Bicubic downsampling is used to generate the LR input frames. The prototype model was trained under three separate settings; each only varied in the input frame selection mechanism used. The three versions of prototype model training represent training with (i) FIFS as the input selection algorithm, (ii) STIFS as the input selection algorithm and (iii) a conventional sliding window with no selection. The corresponding input selection mechanisms are likewise used in the test phase as well.

Adam optimiser with SmoothL1 loss was used to train the model as it combines the advantages of L1-loss (steady gradients for large values) and L2-loss (less oscillation during update when values are small). Thus, it is less sensitive to outliers and sometimes prevents exploding gradients. The initial learning rate is set to 1×10^{-4} and decayed by a factor of 10 after every 10 epoch. Model training and testing are performed using two NVIDIA Tesla V100 GPUs.

Comparison

The empirical evaluation presented in Tables 2 and 3 provide valuable insights into the performance of the prototype VSR model with different input frame selection approaches. The results indicate that incorporating dynamic content-aware frame selection algorithms, namely FIFS and STIFS, significantly enhances the super-resolution performance compared to the model without frame selection.

In terms of quantitative metrics, both FIFS and STIFS consistently outperform the model without frame selection, as demonstrated by higher PSNR and SSIM scores across various video clips. The average PSNR/SSIM values in Tables 2 and 3 reinforce the superiority of FIFS and STIFS over the no-select approach. The FIFS algorithm, in particular, consistently achieves the best performance in terms of PSNR/SSIM, demonstrating its effectiveness in selecting frames with high spatial and temporal correlations to the target frames.

Moreover, the individual clip analysis reveals notable improvements achieved by the FIFS and STIFS algorithms over the no-select scenario. For instance, in Table 3, the "camera" clip exhibits a significant PSNR improvement of 1.1934 dB using FIFS and 0.9642 dB using STIFS. This demonstrates the capability of the proposed dynamic content-aware selection to capture essential details and enhance the overall visual quality in challenging scenarios.

The qualitative visualisations in Fig. 7 further reinforce the quantitative findings, showcasing the merit of FIFS and STIFS in restoring fine details, textures, and sharper edges in the high-resolution frames compared to the no-select scenario. The restored frames exhibit enhanced visual fidelity, indicating the ability of the frame selection algorithms to effectively leverage spatial and temporal correlations between frames, thereby improving the reconstruction quality.

Selection in State-of-the-Art VSR Model

We study the applicability of an input frame selection mechanism, specifically the proposed Feature-based Input Frame Selection (FIFS) algorithm, on 360° video super-resolution. 360° Video Super-Resolution (360VSR) is a challenging task, as conventional video processing methods are not well-suited for the distorted nature of equirectangular 360° video frames. However, the recently proposed Spherical Signal Super-resolution with Proportioned Optimization (S3PO) [41] model addresses these 360° specific requirements by incorporating strategic optimization and feature extraction while utilizing 2D convolution layers. Despite its non-alignment architecture, S3PO has been shown to outperform state-of-the-art VSR models in 360° video super-resolution.

Therefore, in this study, we aim to enhance the performance of the S3PO model by incorporating the FIFS algorithm for the input frame selection. The FIFS algorithm has proven to significantly improve super-resolution outcomes for 2D conventional videos, as discussed in "Comparison".

Furthermore, the scale and rotational invariant capability of FIFS makes it well-suited for feature-based selection in 360° videos as the likelihood of scale and rotation variations across the frames are even higher in these videos because of the distortions present in the equirectangular frames.

To carry out this investigation, we fine-tuned the pre-trained S3PO model by replacing the default selection of immediate consecutive past and future neighbouring frames as input with the proposed FIFS algorithm. We then evaluated the performance of the fine-tuned model through PSNR, SSIM, Weighted-Spherically PSNR (WS-PSNR), and Weighted Spherically SSIM (WS-SSIM) [44]. The evaluation results on five sampled clips from the test set of 360 Video Dataset (360VDS) [41] are presented in Table 4, demonstrating consistent improvement from the FIFS algorithm across all evaluation metrics. This further signifies the applicability and effectiveness of the proposed input frame selection mechanism even on 360° video super-resolution.

Conclusion and Future Directions

In this study, we investigate the impact of using image comparison measures for input frame selection in video super-resolution. Despite the potential of this approach, it has yet to be explored in the VSR literature. Addressing this gap, we conduct an extensive analysis of image similarity measures and develop two dynamic content-aware input frame selection algorithms for VSR: the SpatioTemporal Input Frame Selection (STIFS) algorithm and the Feature-based Input Frame Selection (FIFS) algorithm. Our empirical study shows that these algorithms outperform conventional sliding window methods in terms of both PSNR and SSIM quality metrics on benchmark datasets. Furthermore, we extend the applicability of the best-performing selection algorithm to a state-of-the-art 360° video super-resolution model, resulting in even greater improvement. Our key contributions include the development of cost-effective, efficient, and practical alternatives compared to the increasingly complex architectures that drive the VSR literature. This study opens up a new avenue of research and has the potential to revolutionize the field of VSR.

Based on the result of the empirical study of our proposed dynamic content-aware input frame selection algorithms with feature-based selection capability, the shallow feature-based approach could be further enhanced for the VSR task by not only using the identified features for selection process but also adding the shallow features as additional input feeds to the VSR model. Shallow features extracted in this way would contain low-level information about the frames, such as edges, colour distributions, and texture patterns, and could complement the deep features being extracted by VSR models to learn the super-resolution task. The proposed STIFS and FIFS algorithms could also be extended to select a varied number of input frames from varied sizes of selection windows. This cements STIFS and FIFS as adjunct techniques that could be adopted in conjunction with any VSR model in order to enhance super-resolution performance.

The dynamic content-aware input frame selection mechanism proposed in this study opens up promising pathways for future work in the field of VSR. The integration of shallow features in the VSR model that are extracted during the selection process will allow low-level information, such as edges, colour distributions, and texture patterns, to be comprehensively represented and, thus, a potential direction to enhance the learning capabilities of the model and improve the overall super-resolution performance. Additionally, STIFS and FIFIS currently select a fixed number of input frames from fixed-size selection windows. However, the adaptability of the algorithms could be further enhanced by allowing the selection of a varied number of input frames and varying sizes of selection windows. This flexibility would enable fine-tuning of the selection process based on different video characteristics, dataset properties, and application requirements. Such an extension would reinforce STIFS and FIFS as versatile techniques that can be seamlessly integrated with various VSR models to achieve superior super-resolution results. Finally, the scalability and efficiency of the input frame selection algorithms can be improved to accommodate real-time or near-real-time applications. Investigating techniques such as parallelisation, hardware acceleration, or optimization algorithms could help reduce the computational complexity and enable a faster selection of input frames.

Acknowledgements This research was supported by the Faculty of Science, Engineering and Built Environment (SEBE), Deakin University's Faculty Start-up Scholarship for PhD.

Author Contributions Conceptualization, A.A.B; methodology, A.A.B; software, A.A.B; validation, A.A.B, T-K.L., P.W.E, S.A.; formal analysis, A.A.B; investigation, A.A.B; data curation, A.A.B; writing—original draft preparation, A.A.B; writing—review and editing, A.A.B, T-K.L, P.W.E. and S.A.; visualization, A.A.B, T-K.L, P.W.E; supervision, T-K.L, P.W.E. and S.A. All authors have read and agreed to the published version of the manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data Availability Not Applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Research Involving Human and/or Animals Not applicable.

References

1. Wang Z, Chen J, Hoi SC. Deep learning for image super-resolution: A survey. IEEE transactions on pattern analysis and machine intelligence; 2020.
2. Arefin MR, Michalski V, St-Charles P-L, Kalaitzis A, Kim S, Kahou SE, Bengio Y. Multi-image super-resolution for remote sensing using deep recurrent networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020;pp. 206–207.
3. Liang M, Du J, Li L, Xue Z, Wang X, Kou F, Wang X. Video super-resolution reconstruction based on deep learning and spatio-temporal feature self-similarity. IEEE Transactions on Knowledge and Data Engineering. 2020;1–1. https://doi.org/10.1109/TKDE.2020.3034261.
4. Liu Z-S, Siu W-C, Chan Y-L. Efficient video super-resolution via hierarchical temporal residual networks. IEEE Access. 2021;9:106049–64. https://doi.org/10.1109/ACCESS.2021.3098326.
5. Liu H, Ruan Z, Zhao P, Dong C, Shang F, Liu Y, Yang L. Video super resolution based on deep learning: A comprehensive survey. arXiv preprint arXiv:2007.12928 2020.
6. Haris M, Shakhnarovich G, Ukita N. Recurrent back-projection network for video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019;pp. 3897–3906.
7. Wang X, Chan KC, 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;pp. 0–0.
8. Jo Y, Oh SW, Kang J, Kim SJ. 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. 2018;pp. 3224–3232.
9. Bao W, Lai W-S, Zhang X, Gao Z, Yang M-H. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. IEEE Trans Pattern Anal Mach Intell. 2021;43(3):933–48. https://doi.org/10.1109/TPAMI.2019.2941941.
10. Tian Y, Zhang Y, Fu Y, Xu C. Tdan: Temporally-deformable alignment network for video super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020.
11. Isobe T, Li S, Jia X, Yuan S, Slabaugh G, Xu C, Li Y-L, Wang S, Tian Q. Video super-resolution with temporal group attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020.
12. Chan KCK, Wang X, Yu K, Dong C, Loy CC. 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 (CVPR). 2021;pp. 4947–4956.
13. 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, editors. Computer Vision - ECCV 2020. Cham: Springer; 2020. p. 645–60.
14. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision. 2017;pp. 764–773.
15. Wu Y, Zhang L, Gu Z, Lu H, Wan S. Edge-ai-driven framework with efficient mobile network design for facial expression recognition. ACM Trans Embed Comput Syst. 2023;22(3). https://doi.org/10.1145/3587038.
16. Wu Y, Guo H, Chakraborty C, Khosravi M, Berretti S, Wan S. Edge computing driven low-light image dynamic enhancement for object detection. IEEE Transactions on Network Science and Engineering. 2022;1–1. https://doi.org/10.1109/TNSE.2022.3151502.
17. Agrahari Baniya A, Lee, T-K, Eklund P, Aryal S, Robles-Kelly A. Online video super-resolution using unidirectional recurrent model. 2022. https://doi.org/10.36227/techrxiv.21500235.v1.
18. Sajjadi MS, Vemulapalli R, Brown M. Frame-recurrent video super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018;pp. 6626–6634.
19. Agrahari Baniya A, Lee, T-K, Eklund, PW, Aryal S. STIFS: spatio-temporal input frame selection for learning-based video super-resolution models. In: Proceedings of the 19th international conference on signal processing and multimedia applications - SIGMAP. SciTePress; 2022. p. 48–58. https://doi.org/10.5220/0011339900003289.
20. 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 (CVPR); 2020.
21. Xue T, Chen B, Wu J, Wei D, Freeman WT. Video enhancement with task-oriented flow. Int J Comput Vision. 2019;127(8):1106–25.
22. Hung K-W, Qiu C, Jiang J. Video super resolution via deep global-aware network. IEEE Access. 2019;7:74711–20. https://doi.org/10.1109/ACCESS.2019.2920774.
23. Wang H, Su D, Liu C, Jin L, Sun X, Peng X. Deformable non-local network for video super-resolution. IEEE Access. 2019;7:177734–44. https://doi.org/10.1109/ACCESS.2019.2958030.
24. Lucas A, Lopez-Tapia S, Molina R, Katsaggelos AK. Generative adversarial networks and perceptual losses for video super-resolution. IEEE Trans Image Process. 2019;28(7):3312–27. https://doi.org/10.1109/tip.2019.2895768.
25. Kim SY, Lim J, Na T, Kim M. 3dsrnet: Video super-resolution using 3d convolutional neural networks. arXiv preprint arXiv:1812.09079 2018.
26. Isobe T, Zhu F, Jia X, Wang S. Revisiting temporal modeling for video super-resolution. arXiv preprint arXiv:2008.05765 2020.
27. Zhu X, Li Z, Zhang X-Y, Li C, Liu Y, Xue Z. Residual invertible spatio-temporal network for video super-resolution. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019;pp. 5981–5988.
28. Liu C, Sun D. On bayesian adaptive video super resolution. IEEE Trans Pattern Anal Mach Intell. 2013;36(2):346–60.
29. Russakoff DB, Tomasi C, Rohlfing T, Maurer CR. Image similarity using mutual information of regions. In: Pajdla T, Matas

J, editors. Computer Vision - ECCV 2004. Berlin, Heidelberg: Springer; 2004. p. 596–607.
30. Avcibas I, Sankur B, Sayood K. Statistical evaluation of image quality measures. J Electron Imaging. 2002;11(2):206–23. https://doi.org/10.1117/1.1455011.
31. Hore A, Ziou D. Image quality metrics: Psnr vs. ssim. In 2010 20th International Conference on Pattern Recognition. 2010;pp. 2366–2369. IEEE.
32. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018;pp. 586–595.
33. Lindeberg T. Scale invariant feature transform; 2012.
34. Viswanathan DG. Features from accelerated segment test (fast). In Proceedings of the 10th Workshop on Image Analysis for Multimedia Interactive Services, London, UK. 2009;pp. 6–8.
35. Calonder M, Lepetit V, Strecha C, Fua P. Brief: Binary robust independent elementary features. In European Conference on Computer Vision. 2010;pp. 778–792. Springer.
36. Rublee E, Rabaud V, Konolige K, Bradski G. Orb: An efficient alternative to sift or surf. In 2011 International Conference on Computer Vision. 2011;pp. 2564–2571. Ieee.
37. Leutenegger S, Chli M, Siegwart RY. Brisk: Binary robust invariant scalable keypoints. In 2011 International Conference on Computer Vision. 2011;pp. 2548–2555. Ieee.
38. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 2014.
39. 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. 2016,pp. 770–778.
40. Tareen SAK, Saleem Z. A comparative analysis of sift, surf, kaze, akaze, orb, and brisk. In 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). 2018;pp. 1–10. https://doi.org/10.1109/ICOMET.2018.8346440.
41. Agrahari Baniya A, Lee T-K, Eklund PW, Aryal S. Omnidirectional video super-resolution using deep learning. 2022. https://doi.org/10.36227/techrxiv.20494851.v1.
42. Noble FK. Comparison of opencv's feature detectors and feature matchers. In 2016 23rd International Conference on Mechatronics and Machine Vision in Practice (M2VIP). 2016;pp. 1–6. https://doi.org/10.1109/M2VIP.2016.7827292.
43. Yi P, Wang Z, Jiang K, Shao Z, Ma J. Multi-temporal ultra dense memory network for video super-resolution. IEEE Trans Circuits Syst Video Technol. 2020;30(8):2503–16. https://doi.org/10.1109/TCSVT.2019.2925844.
44. Sun Y, Lu A, Yu L. Ahg8: Ws-psnr for 360 video objective quality evaluation. In Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0040, 4th Meeting; 2016.