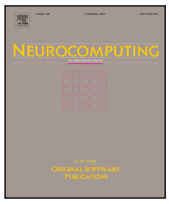
Neurocomputing 614 (2025) 128728

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Enhancing video frame interpolation with region of motion loss and self-attention mechanisms: A dual approach to address large, nonlinear motions

Yeongjoon Kim [a,1], Sunkyu Kwon [a,1], Donggoo Kang [b], Hyunmin Lee [a], Joonki Paik [a,b,*]

[a] Department of Artificial Intelligence, Chung-Ang University, Seoul, 06974, South Korea
[b] Department of Image, Chung-Ang University, Seoul, 06974, South Korea

ARTICLE INFO

Communicated by Z. Guan

Keywords:
Video frame interpolation (VFI)
Attention mechanism
Region of motion loss
Optical flow
Deep learning

ABSTRACT

Video frame interpolation is particularly challenging when dealing with large and non-linear object motions, often resulting in poor frame quality and motion artifacts. In this study, we introduce a novel dual-approach methodology for video frame interpolation that effectively addresses these complexities. Our method consists of two key components: a Region of Motion (RoM) loss and self-attention mechanisms. The RoM loss is designed to spotlight significant movements within frames. This is achieved by employing feature-matching techniques that assign tailored weights during the training process, ensuring that areas of intense motion are given priority. This is facilitated by the computation of optical flow, which identifies crucial feature points and highlights regions of significant motion for targeted enhancement. Our method incorporates self-attention mechanisms to maintain inter-frame continuity while emphasizing the unique attributes of individual frames. The self-attention scores reduce motion discrepancies and enhance the distinctiveness and texture quality of each frame. We validate the efficacy of our approach through extensive evaluations on benchmark datasets, including Vimeo-90K, Middlebury, UCF101, and SNU-Film.

1. Introduction

Video Frame Interpolation (VFI) is an essential technique for generating intermediate frames between two consecutive video frames, playing a crucial role in enhancing video processing applications. This technique enhances video quality and performance by creating smoother transitions between frames and improving temporal resolution. VFI is particularly important for applications like frame rate enhancement, allowing videos to be played at higher refresh rates, which improves viewing experiences [1]. Additionally, VFI facilitates the creation of slow-motion videos [2] by generating additional frames, offering more detailed visual content. VFI is also used to restore high-definition frames in low-resolution videos [3–5], thereby enhancing visual clarity. These advancements are crucial in meeting the growing demand for high-quality video content on modern platforms, where superior playback quality is increasingly expected. Traditional VFI methods involve a series of steps, including motion estimation, motion compensation, occlusion detection, and motion-compensated frame synthesis. While these approaches are foundational in the field, they are often prone to challenges such as blurring, visual artifacts caused by a motion discrepancy, and ghosting effects, which detract from the overall quality of the interpolated video frames. These issues are especially pronounced in videos with large or nonlinear object movements, resulting in reduced interpolation performance.

The integration of deep learning techniques into the realm of computer vision has significantly put forward the field of Video Frame Interpolation (VFI), catalyzing breakthroughs and expanding research efforts. Deep learning approaches to VFI typically merge motion estimation with pixel synthesis, where accurately predicting the motion of objects across frames is crucial for enhancing interpolation outcomes. Convolutional Neural Networks (CNNs) are at the forefront of motion estimation, with methods broadly categorized into kernel-based [6–10] and flow-based algorithms [11–17]. Despite their effectiveness, CNN-based strategies encounter challenges in handling large motions due to the localized nature of convolution operations and the limitations imposed by the size of their receptive fields.

Given the limitations faced by CNNs, the exploration of transformers has gained momentum, leveraging their adaptability across diverse

* Corresponding author at: Department of Artificial Intelligence, Chung-Ang University, Seoul, 06974, South Korea.
E-mail addresses: yjkim@ipis.cau.ac.kr (Y. Kim), kwonsk@ipis.cau.ac.kr (S. Kwon), dgkang@ipis.cau.ac.kr (D. Kang), dl218218@ipis.cau.ac.kr (H. Lee), paikj@cau.ac.kr (J. Paik).
[1] These authors equally contributed to this manuscript.

https://doi.org/10.1016/j.neucom.2024.128728
Received 13 February 2024; Received in revised form 28 August 2024; Accepted 12 October 2024
Available online 7 November 2024

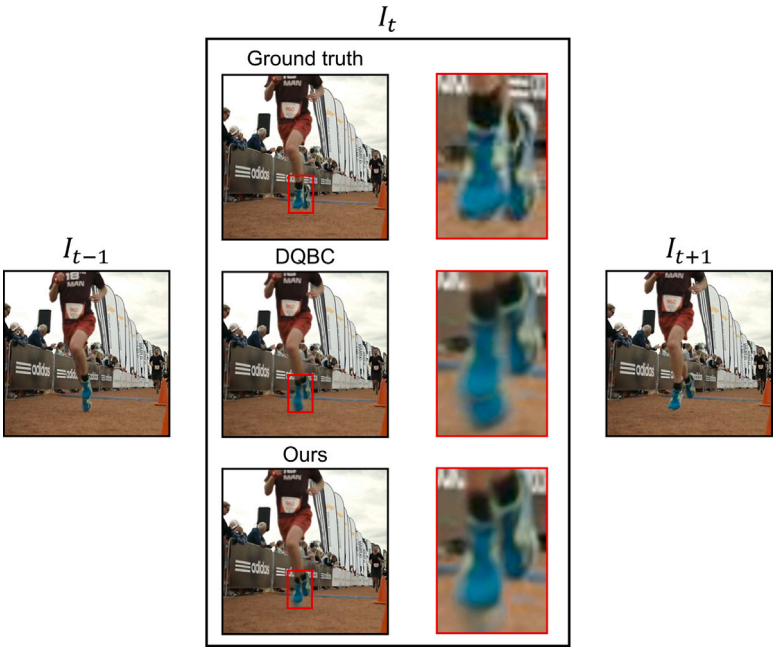


Fig. 1. Comparison of VFI techniques for pronounced motions: While the DQBC method results in artifacts within the shoe region, the proposed method achieves precise rendering of the shoe region.

applications [18–20]. Unlike CNNs, which are constrained by their inherent locality, transformers excel through their ability to compute attention scores globally across all pixels. This global perspective allows transformer-based VFI models to overcome CNNs' difficulties in capturing long-range dependencies [21,22], facilitating more precise motion estimation via the self-attention mechanism [18]. Despite the promising advancements achieved with these transformer-based approaches, challenges remain, especially in accurately interpolating videos with extensive large-scale motion.

The challenges in handling significant object motion in Video Frame Interpolation (VFI) can be summarized as follows:

- The presence of pronounced non-linear object movements introduces considerable challenges in accurately interpolating video frames.
- The dependency on receptive fields for motion estimation often leads to instability and inaccuracies in scenarios involving rapidly moving objects.
- The task of capturing the intricate dynamics of real-world motion poses significant obstacles, taking away from the effectiveness of motion interpolation in replicating true motion behaviors.

In this paper, we present a novel video frame interpolation (VFI) technique that leverages Region of Motion (RoM) loss and attention scores, employing an efficient and robust method that utilizes RoM loss and transformer attention scores. The proposed method enhances the modeling of relationships between temporally adjacent frames through a bidirectional correlation technique [23], which not only mitigates the issues associated with receptive field dependence but also aids in generating a more accurate motion field. To address the interpolation problem of large-scale motion, our approach utilizes a feature-matching algorithm across two consecutive frames. This algorithm identifies motion intensity, based on which bounding boxes are defined—termed as RoM. We then compare these motion regions with those from the ground truth, formulating a RoM loss to minimize the differences between the predicted and actual motion fields. Moreover, we incorporate self-attention scores within the foundational encoder of the CNN architecture, enhancing correlation modeling. This enhancement is further applied to ContextNet, developed specifically for refining the motion field. The resulting feature maps are finely tuned to contexts marked by self-attention, particularly benefiting the depiction of rapidly moving objects.

Through this advanced technique, our research effectively addresses the challenges posed by substantial motion, focusing on frame regions with pronounced movement. By combining RoM loss and self-attention mechanisms, our method significantly enhances motion estimation accuracy, especially in cases involving large and rapid object movements, thereby overcoming the limitations of traditional CNN-based methods. As shown in Fig. 1, the proposed VFI method demonstrates superior performance in generating intermediate frames for videos featuring dynamic activities, such as running, significantly surpassing the capabilities of existing methods [23].

The key contributions of this paper are as follows:

- Introduction of the RoM loss: This loss leverages feature-matching algorithms to identify features indicative of pronounced motion. During the training phase, these features are emphasized using tailored weights.
- Implementation of self-attention scores: Self-attention scores are incorporated to focus on each individual frame. This highlights the importance of both inter-frame correlation and the extraction of self-attentive features within individual frames.
- Superior performance in frame interpolation: Our model exhibits exceptional performance in frame interpolation for VFI tasks, particularly on benchmarks characterized by substantial motion.

In summary, this study presents a robust and innovative solution for video frame interpolation, delivering significant improvements in interpolation quality. Our contributions include the introduction of the RoM loss, which emphasizes significant frame movements, the implementation of self-attention scores to enhance inter-frame continuity, and the demonstration of superior performance in interpolation tasks, especially on benchmarks characterized by substantial motion. These advancements have important implications for various video processing applications.

The structure of the paper is as follows: Section 2 provides an analysis of video frame interpolation techniques and reviews self-attention mechanisms. Section 3 describes the proposed model, which addresses large movements and improves robustness compared to existing methods. Section 4 presents the experimental results, including the

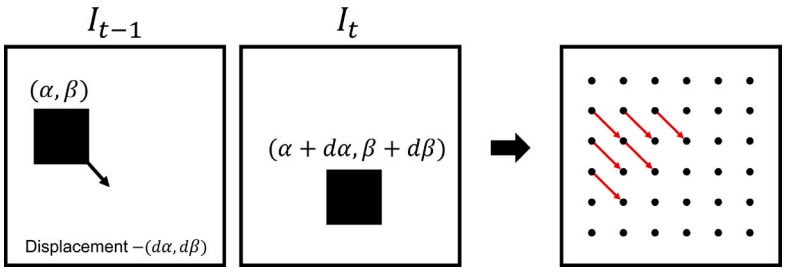


Fig. 2. Pixel-based motion estimation via optical flow.

datasets used and various test scenarios. Finally, Section 5 concludes the study by summarizing the findings and suggesting directions for future research.

2. Related work

In the field of video processing, the advancement of video frame interpolation (VFI), motion estimation, and self-attention mechanisms stand as pivotal areas of research that have significantly contributed to the development of sophisticated video enhancement techniques. This section provides an overview of the related work in these domains, setting the stage for our proposed methodology that leverages Region of Motion (RoM) loss and self-attention mechanisms to address the challenges of large and complex motion patterns in video frame interpolation. Through this exploration, we aim to contextualize our contributions within the broader landscape of video processing advancements.

2.1. Video frame interpolation (VFI)

Video Frame Interpolation (VFI) is a research area dedicated to creating additional intermediate frames within video sequences. In the field of deep learning-based VFI, methods are generally divided into two main groups: *kernel-based* [6–10] and *flow-based* [11–17]. Kernel-based approaches estimate spatially adaptive convolution kernels for synthesizing pixels, whereas flow-based strategies excel in capturing larger movements without the constraints of kernel sizes. Consequently, flow-based techniques have gained prominence in the field of VFI, offering a more flexible solution for accurately representing motion.

2.2. Motion estimation

The distinction in motion between background elements and objects poses a significant challenge in the creation of intermediate frames, making precise motion estimation a critical component in Video Frame Interpolation (VFI). Motion estimation involves identifying motion vectors that delineate the temporal shift from initial to subsequent frames within a sequence. Motion vectors can be determined through three primary methodologies: pixel-based [24,25], feature-based [26,27], and deep-learning-based methods [28,29].

Pixel-based methods. Optical flow [30–34] is a widely used technique in pixel-based motion estimation, capturing the motion dynamics of individual pixels across a sequence. As shown in Fig. 2, it calculates both the direction and magnitude of pixel displacements between consecutive frames.

Another remarkable pixel-based approach is block-matching [35–37], which identifies macroblocks in consistent locations across successive frames for motion estimation. Fig. 3 illustrates the block-matching technique, where the red box indicates a macroblock and the red arrow represents the motion vector.

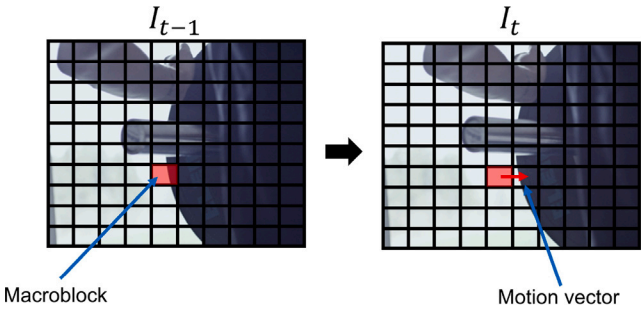


Fig. 3. Pixel-based motion estimation through block matching. The red boxes indicate macroblocks and the red arrow represents the motion vector.

Feature-based methods. This method employs visual features such as corners and textured areas to estimate motion by comparing features between frames. Algorithms like SIFT [38], SURF [39], and ORB [40] are crucial for feature extraction, detecting points and generating descriptors around them. Feature matching involves comparing these descriptors across images to find corresponding feature point pairs, with nearest neighbor matching [41] and RANSAC [42] being primary techniques for this purpose. Fischler and Bolles proposed an iterative approach to model estimation from randomly sampled feature points, identifying inliers that fit the model appropriately. Feature-based methods are employed in image retrieval, object recognition, image classification, and VFI, where they discern feature points in adjacent frames.

Deep-learning-based methods. Deep learning methods have introduced networks that automate the feature extraction process by training on specific frame features. In this subsection, we will omit basic CNN-based motion estimation methods and briefly introduce techniques that utilize the attention mechanism and transformers. Sarlin et al. developed a context aggregation mechanism using attention [28], which assigns scores to feature points and filters out mismatches based on these scores for precise feature point detection. Sun et al. integrated self and cross-attention layers within a transformer to generate feature descriptors influenced by both images [29], enhancing the extraction of quality feature points in challenging conditions. Moreover, recent advancements in 3D vision have shown that integrating monocular 3D object detection techniques [43,44] with deep learning can significantly improve the accuracy of motion estimation. Deep learning approaches are noted for their accuracy and reliability in motion estimation.

2.3. Self-attention mechanism

The attention mechanism, originally introduced for neural machine translation [45], enables the capture of long-range dependencies in data [18]. Self-attention, a key component of the transformer model, assesses the relevance of different positions within a sequence, facilitating the identification of contextual relationships. This mechanism's

scalability has led to its application across various domains, including vision tasks with the Vision Transformer (ViT) [19].

The self-attention mechanism, also known as scaled dot-product attention, operates using key, query, and value sequences. It calculates attention scores and probability distributions to assess the significance of each element, producing an output as a weighted sum of the value vectors.

Building on this mechanism, recent developments aim to improve VFI performance. Zhang *et al.* re-examine inter-frame attention processing to refine appearance features and extract motion data [46]. Additionally, Danier *et al.* employ a self-attention approach for efficient inference on high-resolution videos, utilizing a multi-axis self-attention layer that maintains linear complexity relative to input dimensions [47]. Furthermore, the self-attention mechanism is suitable for high-speed data processing and is particularly effective when combined with event-based cameras [48,49] that offer high temporal resolution. This combination is expected to significantly enhance interpolation quality by accurately capturing object movements. Moreover, the effectiveness of machine learning algorithms in predicting complex patterns, as demonstrated by Nisa et al. [50], underscores the potential of integrating such techniques with self-attention models to further enhance video analysis systems. The self-attention mechanism is not limited to simple image processing but has also proven to be effective in other domains such as few-shot object detection [51] and vision-language interaction [52], highlighting its versatility.

3. Proposed methods

From a generative standpoint in Video Frame Interpolation (VFI), our objective is to accurately generate intermediate frames within videos, especially in scenarios where both background and objects exhibit significant motion. To solve this problem, we introduce a cutting-edge VFI model designed to excel in conditions of substantial motion, facilitating the creation of precise intermediate frames. The proposed model, termed "Nonlinear Video Frame Interpolation with RoM Loss and Attention Score", features two primary components: **(1) Self-attention score** mechanism dynamically adjusts to emphasize regions of the frame with significant motion. This mechanism is integrated with features from both the Basic Encoder and ContextNet, enabling the model to generate a self-attention feature map that captures complex motion patterns. By focusing on these regions, the model improves the prediction of motion trajectories for objects and backgrounds across frames. **(2) RoM loss function** focuses on large motion features, enhancing the model's performance by selectively minimizing prediction errors in regions of significant motion. This prioritization is crucial for maintaining visual coherence in the interpolated frames. Fig. 4 provides a schematic representation of the proposed methodology, illustrating how these components interact to produce high-quality intermediate frames in scenarios with significant motion.

3.1. Self-attention score

To enhance the interpolation quality of video frames characterized by significant motions, our approach integrates a self-attention mechanism [18,53], with a particular emphasis on key data points involved in substantial movements. The self-attention mechanism has proven effective in addressing challenges associated with large motions and refining the quality of interpolated frames. Its ability to dynamically adjust weights allows prioritization of areas with intense motion within the input frame by assigning them higher importance, while areas with less motion receive lower weights. This strategy aligns with our method's focus on addressing large motions in Video Frame Interpolation (VFI). The self-attention mechanism offers several advantages over traditional interpolation methods. Traditional methods often struggle to accurately capture and represent large motions within video frames, leading to artifacts and blurred results. In contrast, self-attention dynamically weights different areas within the frame, prioritizing regions with significant motion. This approach enhances motion estimation and produces higher-quality interpolated frames. Moreover, the flexibility and scalability of self-attention make it effective in handling complex motion patterns that traditional convolutional methods may not adequately address. As shown in Fig. 4, self-attention scores are applied to the feature maps obtained from both the Basic Encoder and ContextNet. Once these feature maps are refined, a CNN-based frame synthesis network generates the final interpolated frame, calculating both image and occlusion map residuals to enhance interpolation accuracy.

The self-attention scores play a crucial role in maintaining inter-frame continuity by selectively focusing on regions with high motion, ensuring these areas are consistently tracked and accurately represented across frames. Additionally, by highlighting unique attributes of individual frames, such as areas of sudden or complex motion, the self-attention scores help preserve the distinct characteristics of these regions while seamlessly integrating them into the interpolated frames.

The architecture of the self-attention mechanism is illustrated in Fig. 5, and the procedure to compute the attention map is detailed below:

$$S_n = \frac{\exp(d_n)}{\sum_{k=1}^{n} \exp(d_n)}, \quad d_n = f_K(X_n)^T f_Q(X_n), \tag{1}$$

where X_n denotes the feature map derived from the input image by the backbone. d_n is derived by taking the transpose of the key from f_K and subsequently computing its dot-product with the query from f_Q. The attention map S_n is obtained by applying a softmax function to d_n. The previously computed attention weights are then employed to determine a weighted sum of the value across all sequence elements.

$$A_n = \sum_{n=1}^{N} S_n(f_V(X_n)), \tag{2}$$

where A_n denotes the attention weight, and N represents the number of positions derived from the feature of the previous hidden layer.

As illustrated in Fig. 5, two consecutive input image frames I_0 and I_1 are processed through the backbone network to produce respective feature maps. Bilinear correlation is applied to the feature maps extracted from two input frames, serving as a significant prior in the VFI task. This enhances the accuracy of motion field generation through the Densely Quried Bilateral Correlation (DQBC) method [23]. The self-attention scores improve texture quality by effectively capturing and emphasizing fine details in areas of significant motion, which traditional methods might overlook or blur. This ensures that intricate textures and subtle motion nuances are well-preserved in the interpolated frames, resulting in sharper and more realistic outcomes. In the proposed method, the utilization of self-attention is key to extracting vital information from feature maps. This leads to the generation of a self-attention feature map, which emphasizes critical areas, particularly those with significant object movement. This approach plays a crucial role in producing high-quality intermediate frames, particularly effective in situations with considerable motion.

3.2. Region of motion (RoM) loss

Conventional loss functions, especially ℓ_1 and ℓ_2 losses, are widely used in image processing tasks due to their inherent simplicity and computational advantages. However, in scenarios characterized by extensive motion, these mechanisms struggle, as they uniformly assign equal importance to both stationary and moving regions within a frame, failing to differentiate between the two. As shown in Fig. 4, our proposed RoM loss addresses the issue of creating inconsistent intermediate frames in videos with significant object movement. It handles large motions by identifying the high-motion areas in a frame and reducing the difference between the predicted area and the actual scene. To develop this function, we employed the Shi-Tomasi corner

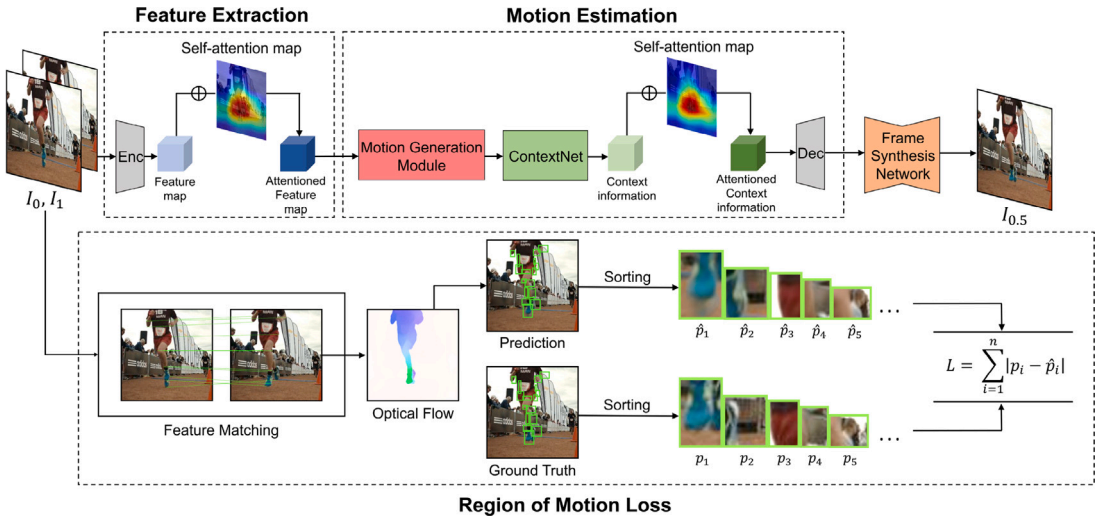


Fig. 4. Schematic representation of the proposed framework. Consecutive frames are processed through self-attention in the Basic Encoder and ContextNet, generating a motion-focused feature map. The proposed RoM loss captures motion magnitude using feature matching and optical flow, leading to the creation of a region of motion box. This box is then placed on both the ground truth and predicted intermediate frames, facilitating the progressive learning process by minimizing differences between the frames.

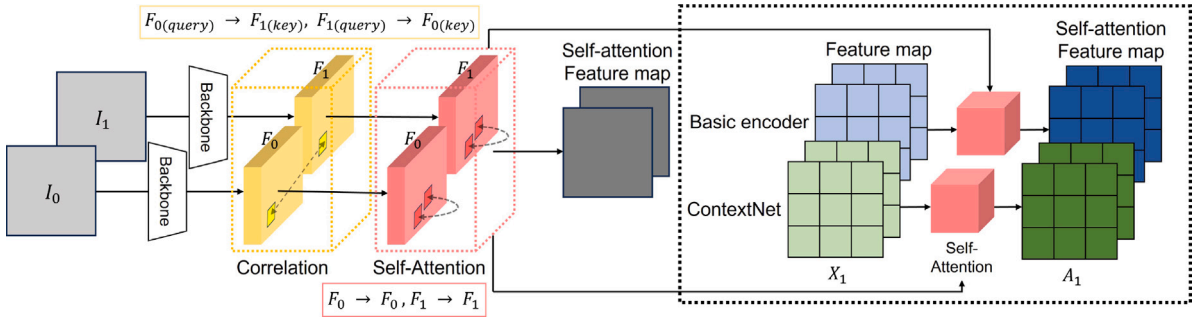


Fig. 5. Adapted self-attention architecture in our proposed method. This illustrates the process where two consecutive frames are fed into the Basic Encoder and ContextNet. Here, a self-attention mechanism is utilized on the extracted feature maps, leading to the creation of a feature map that specifically highlights areas of motion within the frames.

detection [54] and the Lucas-Kanade optical flow estimation [55] algorithms. These techniques are simpler and more resource-efficient than other deep learning approaches that rely on vast datasets and intricate structures [56].

3.2.1. *Motion estimation-based feature point matching*

Feature point extraction involves identifying key points within an image or video frame. This technique finds applications in a range of vision tasks, including object segmentation [57–59], video classification [60–62], and video frame interpolation [6,8,11,22,23]. In the proposed method, we employ the Shi-Tomasi corner detection algorithm for this purpose. This algorithm builds upon the foundational Harris corner detection method [63]. At its core, the Shi-Tomasi approach assumes that significant changes in an image during window movement indicate that the region beneath the original window likely represents a corner, which can be mathematically expressed as

$$E(u,v) = \sum_{x,y} w(x,y)[I(x+u, y+v) - I(x,y)]^2, \tag{3}$$

where $E(u,v)$ represents the difference in intensity between the original window $I(x,y)$ and the moving window $I(x+u, y+v)$ when moving along the x or y axis. The function $w(x,y)$ is a rectangular or Gaussian window function. A matrix representing the intensity change around a specific pixel for corner detection must be defined. The equation for calculating matrix M is as follows:

$$M = \sum_{x,y} w(x,y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}, \tag{4}$$

where I_x and I_y are the intensity changes in the x and y directions at the corresponding pixel location in the image. The set of eigenvalues $E_M = \{\lambda_1, \lambda_2\}$ of the matrix M play a pivotal role in detecting corners. In Shi-tomasi corner detection algorithm, the minimum value of the eigenvalues R is defined as the corner score:

$$R = \min(\lambda_1, \lambda_2), \qquad CornerPoint(F) = \{(x,y) | R(x,y) > R\}, \tag{5}$$

where R serves as a threshold, differentiating whether a point is deemed a corner. Pixels with coordinates (x,y) exceeding the R value are designated as corner points F. These identified corner points F in the image subsequently play a role in the optical flow computation.

Optical flow analysis is crucial for understanding the movement of objects or pixels across successive video frames, providing essential insights into motion dynamics. In our approach, we utilize the Lucas-Kanade method for optical flow estimation [55], a well-regarded and widely-used technique in various visual computing tasks. This method relies on the assumption of brightness constancy around a pixel to compute the optical flow vector, which indicates the direction and speed of motion for an object or pixel. The Lucas-Kanade technique begins by identifying key feature or corner points in the image, followed by calculating motion vectors based on the brightness levels of nearby pixels.

While the Lucas-Kanade method is sometimes criticized for its sensitivity to large-scale motions, it is important to contextualize what constitutes "large motion" within the realm of Video Frame Interpolation (VFI). In VFI, a significant motion typically spans the gap between 1 and 15 frames, equivalent to half a second at a standard video

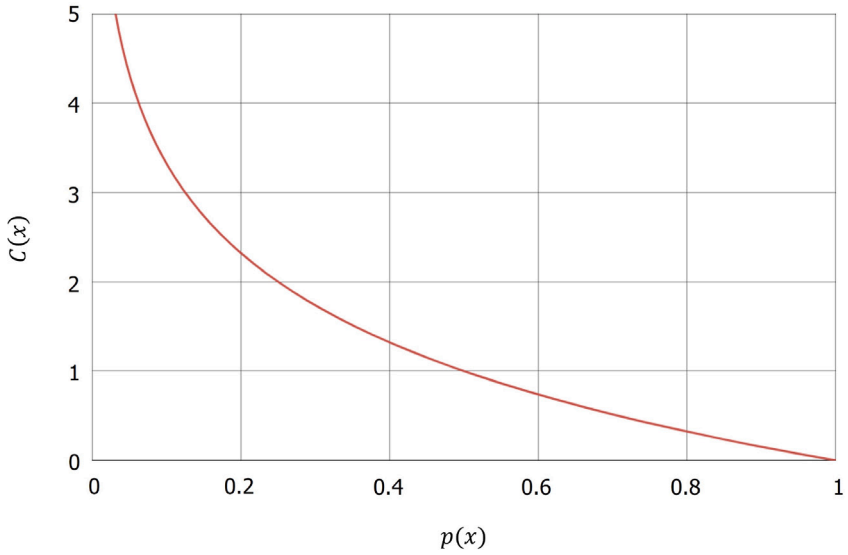


Fig. 6. Illustration of information content. This diagram encapsulates a key principle of information theory: events with a low probability $p(x)$ are associated with a high information content $C(x)$.

frame rate of 30 frames per second. Such a duration generally does not result in the type of extensive object or background movement that would compromise the effectiveness of Lucas-Kanade's optical flow estimation, thus maintaining its viability and accuracy in our proposed method.

The mathematical formulation of optical flow is described by the equation

$$I_x(F) \cdot u + I_y(F) \cdot v = -I_t(F), \tag{6}$$

where $F = x_i, y_i | i = 0, \ldots, n-1$ denotes the set of feature points, $I_x(F)$ and $I_y(F)$ represent the spatial gradients in the x and y directions, respectively, for pixel intensity at a feature point F, identified using the Shi-Tomasi corner detection method. The variables u and v denote the components of the velocity vector at the feature points F. $I_t(F)$ captures the temporal intensity change at the feature point. This equation serves to compute the velocity vectors (u and v) at each feature point, thereby quantifying the motion at those locations. By applying optical flow calculations, we can accurately estimate motion between video frames, an essential aspect of our interpolation strategy.

3.2.2. Leveraging partial information to optimize video frame interpolation

Information theory is a framework for quantifying the information content inherent in an event [64]. At the heart of this theory lies the principle that events with lower occurrence probabilities (i.e., rarer events) possess greater informational value than more common ones. The information content of an event is determined by the following equation:

$$C(x) = \log_a \left(\frac{1}{p(x)} \right) = -\log_a p(x), \tag{7}$$

where $p(x)$ is the probability of event x occurring. This equation outputs a value that quantifies the amount of information, with the base of the logarithm, a, defining the unit of measure for this information. Specifically, when a is set to 2, the information is measured in "bits".

As shown in Fig. 6, which visually illustrates the information content, it becomes evident that as the probability $p(x)$ approaches 0, the value of $C(x)$ tends towards infinity, while conversely, as $p(x)$ approaches 1, $C(x)$ approaches 0.

Eq. (7) delineates that events occurring with lower probabilities possess higher information content, whereas events occurring more frequently have lesser information content. This principle underscores a fundamental principle of information theory: the greater the information content, the more significant the information conveyed. Information theory establishes a method for quantitatively assessing the information content of an event [64], positing that less probable (rarer) events are more informative than common occurrences. This theory articulates the relationship between event probability and information content through the equation $-\log_a p(x)$, where $p(x)$ is the probability of an event x and a denotes the logarithm's base. According to this relationship, events with lower probabilities are characterized by higher information content, suggesting that in the realm of information theory, a larger amount of information content signifies more substantial information.

We have developed a loss function inspired by information theory to effectively address challenges associated with large motions. Diverging from the traditional application of information theory, which evaluates information based on event probabilities, our approach in Video Frame Interpolation (VFI) assesses it based on the extent of regions within a video frame. Thus, in scenarios involving significant object movements, such motions are deemed to carry high informational value. Additionally, we opt to either overlook or assign lesser importance to areas exhibiting minimal motion, such as backgrounds. As shown in Fig. 7, this strategy of reallocating focus enhances emphasis on regions experiencing substantial motion—highlighted in green to signify the object's motion (i.e., crucial information)—and reduces attention on less critical areas, marked in blue to denote the background (i.e., non-essential information).

The proposed technique effectively filters out redundant details with minimal information content from the frame, focusing solely on crucial details with high informational value for interpolation. As shown in the right panel of Fig. 7, the cumulative information content of the frame is significantly reduced, while the informational value of key elements is preserved. This selective filtering process enables the model to allocate computational resources more efficiently and produce more accurate intermediate frames. By selectively focusing on regions of interest, the attention-based filtering process plays a crucial role in enhancing the performance of video frame interpolation models. This technique has shown to be particularly effective in scenarios involving complex motion patterns and challenging visual elements.

The essential informational content $\hat{p}_i$ within a frame, focusing solely on crucial areas, is calculated using the equation:

$$\hat{p}_i = \sum_{i=0}^{m-1} w \cdot (\alpha \cdot f(\alpha)) + (1 - \alpha \cdot f(1-\alpha)), \tag{8}$$

where i signifies the frame permeated with essential information, and α denotes the fraction of critical information regions retained after

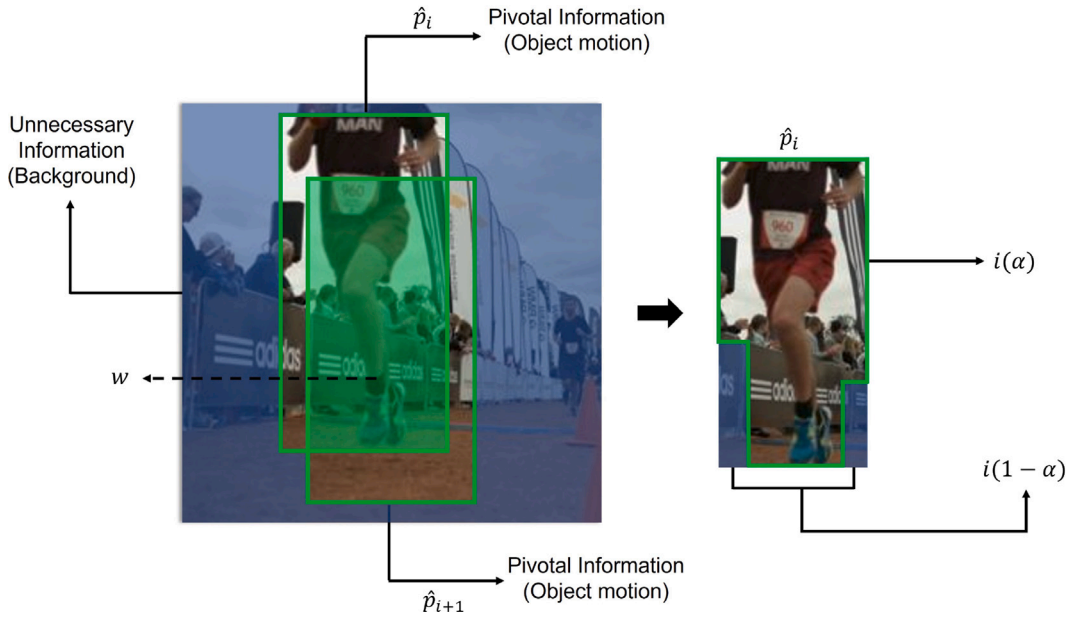


Fig. 7. Illustration of large motion within a frame. This example highlights the significant motion of an object, where the primary information content, denoted by $\hat{p}_i$, is centered on areas of extensive motion. The region marked by w represents the weight assigned to this crucial information, accentuating the importance of movement.

excluding non-essential details. This formula computes the overall information content by applying the weighting of both essential and non-essential information across $i \in 0, \ldots, m-1$ frames to their corresponding regional values.

The parameter w acts as a weighting factor, accentuating the focus on essential information. Drawing inspiration from information theory, our methodology prioritizes areas abundant in motion, pivotal for the interpolation process. As depicted in Fig. 7, the segment where two frames intersect, denoting essential information, is labeled as w. This intersection area w between the images $\hat{p}_i$ and $\hat{p}_{i+1}$ signifies motion within the entire frame. By prioritizing the weighting of essential informational content, RoM boxes are generated, thereby amplifying the frame's motion emphasis.

3.2.3. Region of motion (RoM) box

Once feature points have been detected and optical flow has been computed, the region of pivotal motion is identified using the principle of information theory. This RoM is then encapsulated within a bounding box.

The procedure for creating the RoM box is outlined as follows:

$$f_k^{set} = (f_k^{F_X},\ f_k^{F_Y}), \quad f_{k+1}^{F_{set}} = (f_{k+1}^{F_X},\ f_{k+1}^{F_Y}), \tag{9}$$

$$f_{(W,H)} = f_{k+1}^{F_{set}} - f_k^{F_{set}}, \tag{10}$$

$$\hat{P}_{rom} = \sum_{j=0}^{m-1} \max_{w \in W,\ h \in H}(\|u_i, v_i\|), \tag{11}$$

where the detected feature points$(f_k^{F_{set}},\ f_{k+1}^{F_{set}})$ represent the change in position between two consecutive video frames, with each feature point moving from $(f_k^{F_X}, f_k^{F_Y})$ to $(f_{k+1}^{F_X}, f_{k+1}^{F_Y})$. Here F_{set} is defined as $F_{set} = F_{(X,Y)} = \{x_i, y_i | i = 0, \ldots, n-1\}$, where n represents the total number of feature points. Following the initial steps, RoM boxes are constructed by determining their width f_w and height f_h based on the x and y coordinates of identified feature points. These feature points also serve to calculate the magnitude of motion vectors (u, v), employing the Lucas-Kanade method for precision. Our main goal is to boost interpolation accuracy in scenarios characterized by intense motion, necessitating a focus on isolating significant motion values.

To achieve this, we select only the top m motion vectors, which are then used to construct RoM boxes $\hat{P}^{RoM\,box}j$ for $j \in 0, \ldots, m-1$, specifically targeting areas with the most pronounced motion. This approach ensures that regions experiencing considerable movement are accurately represented. Consequently, the critical information discussed in Section 3.2.2 is aggregated from these RoM boxes ($\alpha = \hat{P}rom$), capturing the essence of significant motion within the frame. Upon forming the RoM boxes, they are organized according to their motion magnitude, streamlining the focus on areas of utmost relevance to the interpolation task.

As shown in Fig. 8 and (12), this prioritization strategy ensures that regions with substantial motion are tackled first, in line with our goal of focusing on pivotal movements.

$$\hat{P}_{pred} = S_{\text{desc}}(\hat{P}_{rom}), \tag{12}$$

where $S_{\text{desc}}(\cdot)$ represents the descending order sorting operation.

3.3. Loss function

Our model is trained end-to-end with four distinct loss functions: reconstruction loss $\mathcal{L}_{rec}$, teacher reconstruction loss $\mathcal{L}_{tea}$, distillation loss $\mathcal{L}_{distill}$ as per DQBC [23], and region of motion(RoM) loss $\mathcal{L}_{rom}$.

The reconstruction loss, denoted by ℓ_{rec}, is the ℓ_1 loss between the predicted intermediate frame $I_{k+0.5}$ and its ground truth $I_{k+0.5}^{gt}$:

$$\ell_{rec} = \|I_{k+0.5} - I_{k+0.5}^{gt}\|_1. \tag{13}$$

Adopting a strategy from Huang et al., we employ a teacher model for superior guidance in learning motion estimation. This approach involves using an auxiliary IFBlock [13], which, having access to $I_{k+0.5}^{gt}$, processes the initially proposed motion fields $M_{t\to0}$, $M_{t\to1}$ to produce refined versions $M_{t\to0}^{tea}$, $M_{t\to1}^{tea}$. The teacher loss is calculated as follows:

$$\ell_{tea} = \|O \cdot w(I_k, M_{t\to k}^{tea}) + (1-O) \cdot w(I_{k+1}, M_{t\to k+1}^{tea}) - I_{(2k+1)/2}^{gt}\|_1, \tag{14}$$

where O represents the occlusion map generated by the proposed model.

The distillation loss, $\ell_{distill}$, accumulates the ℓ_2 losses from the refined motion fields $M_{t\to0}^{tea}$, $M_{t\to1}^{tea}$, and all intermediary motion fields produced by the proposed model throughout the motion wstimation and refinement stages.

The Region of Motion (RoM) loss, denoted as $\mathcal{L}_{rom}$, is designed to encapsulate a bounding box that aligns with the location of significant motion in the ground truth. It evaluates the difference in values within

Fig. 8. Arranging RoM boxes in descending order of motion magnitude optimizes the interpolation process by prioritizing areas with the most substantial motion, ensuring these critical regions are addressed first.

this designated box across both the ground truth and our model's predictions, with the objective of minimizing discrepancies. By doing so, it directs the model's attention towards key areas of motion, enhancing the fidelity of interpolated frames in regions characterized by substantial movement. The formulation of the RoM loss is as follows:

$$\mathcal{L}_{rom} = \sum_{i=0}^{n-1} |I^{\hat{P}_{pred}}_{(2k+1)/2} - \hat{I}^{\hat{P}_{pred}}_{(2k+1)/2}|. \tag{15}$$

The RoM loss is specifically designed to address challenges in scenes with extensive motion by targeting regions exhibiting significant movement. By guiding the model to emphasize critical information crucial for effective interpolation, this loss function significantly enhances the accuracy and visual quality of the generated intermediate frames. The overall loss function, denoted as $\mathcal{L}_{total}$, is formulated as the aggregation of the distinct loss components:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{tea} + w_1\mathcal{L}_{distill} + w_2\mathcal{L}_{rom}, \tag{16}$$

where w_1 and w_2 serve as weight hyper-parameters to fine-tune each individual loss term.

4. Experimental results

This section presents the experimental assessment of the proposed model, evaluating its effectiveness across multiple datasets and performance metrics to compare it against current methodologies. We aim to establish the model's baseline performance with a focus on its ability to handle challenging scenarios, such as large, nonlinear object motions. To optimize the model, we conduct parameter fine-tuning, examining the effects of various factors, such as learning rates and network configurations, on overall performance. This analysis identifies key elements that contribute to the model's effectiveness and highlights areas for potential improvement. Performance is assessed using both quantitative metrics, such as PSNR and SSIM, and qualitative methods, including visual inspections. These evaluations demonstrate the model's strengths in specific areas, providing insights into its robustness and adaptability. The experiments are designed to comprehensively test the model, offering a clear comparison with existing methods and exploring its potential for practical applications.

4.1. Evaluation datasets and metrics

We assess the proposed algorithm across various video datasets featuring different image resolutions, providing an overview of the datasets and metrics used for evaluation:

Vimeo90K [17]. A popular benchmark in the field of video frame interpolation (VFI), the Vimeo90K test dataset comprises 3782 triplets, each with a resolution of 448 × 256, making it a standard choice for VFI evaluation.

UCF-101 [65]. Featuring a wide array of human actions, the UCF-101 dataset includes 379 triplets, each sized at 256 × 256. This diversity makes it an excellent resource for testing VFI models across various dynamic scenes.

Middlebury [66]. Known for its detailed evaluation framework, the Middlebury benchmark is divided into two sets: Other and Evaluation. For our study, we engaged with the Other set, which is composed of 12 triplets that include ground truth intermediate frames, offering precise benchmarks for VFI performance.

SNU-FILM [67]. The SNU-FILM dataset features 1240 frame triplets with resolutions ranging from 368 to 720 in width and from 384 to 1280 in height. It is divided into four categories based on the level of motion difficulty: Easy, Medium, Hard, and Extreme, providing a comprehensive testbed for evaluating VFI models under varying motion conditions.

Metrics. We compute the average Interpolation Error (IE) on the Middlebury dataset. Lower IEs indicate better performance. We evaluate the PSNR and SSIM on the Vimeo90K, UCF101, and the SNU-FILM datasets for comparisons.

4.2. Model analysis

In this section, we detail the steps involved in carrying out the proposed model's experiments, including the parameters used, the procedural steps followed, and the specific conditions under various scenarios. Additionally, we explain how the model computations are configured and executed.

Experimental procedure. The experimental procedure to evaluate the proposed model involves several key steps. The process begins with model training using the Vimeo90K training set. The training is conducted on four Nvidia RTX 4090 GPUs with batch sizes 64 and handles patches of 256 × 256. Data augmentation techniques such as random cropping, flipping, rotating and inverting are extensively used to improve the robustness of the model and its ability to generalize across different video sequences.

After initial training, parameter fine-tuning is performed to optimize the performance of the model. This includes testing configurations with varying number of feature points, number of RoM boxes, and batch size. The model is evaluated under various motion complexities and scene dynamics, in particular categorized by the SNU-FILM dataset as easy, moderate, hard, and extreme motion difficulties. From different datasets, we select the best configurations based on performance metrics including PSNR, SSIM, and IE.

In the model evaluation stage, the performance of the trained model is quantitatively evaluated using the aforementioned metrics. These metrics are computed for each dataset to provide a comprehensive understanding of how well the model interpolates video frames. The results are compared with those of existing state-of-the-art VFI models, with an emphasis on performance at different levels of motion complexity.

Computational setup. The computational setup is crucial for effectively managing the intensive training and evaluation processes required for video frame interpolation. The experiments are conducted on a high-performance computing environment consisting of 4 Nvidia RTX 4090 GPUs, which provide the necessary computational power for handling large datasets and performing complex calculations.

Table 1
Performance evaluation of the proposed VFI model versus established methods across Vimeo90K, UCF-101, Middlebury, and SNU-FILM datasets. This analysis utilizes average Interpolation Error (IE) for Middlebury and PSNR/SSIM metrics for the remaining datasets to measure performance: lower IE values and higher PSNR/SSIM scores reflect enhanced performance. The highest-ranking outcomes are highlighted in **RED** for the top performer and BLUE for the runner-up. The proposed method's performance is shown in two configurations: **Set 1** (4096 feature points, 16 boxes) and **Set 4** (8192 feature points, 24 boxes).

Method	Year	Vimeo90K	UCF-101	Middlebury	SNU-FILM			
					Easy	Medium	Hard	Extreme
SepConv	ICCV'17	33.79/0.9702	34.78/0.9669	2.27	39.41/0.9900	34.97/0.9762	29.36/0.9253	24.31/0.8448
ToFlow	IJCV'19	33.73/0.9682	34.58/0.9667	2.15	39.08/0.9890	34.39/0.9740	28.44/0.9180	23.39/0.8310
CyclicGen	AAAI'19	32.09/0.9490	35.11/0.9684	–	37.72/0.9840	32.47/0.9554	26.95/0.8871	22.70/0.8083
DAIN	CVPR'19	34.71/0.9756	34.99/0.9683	2.04	39.73/0.9902	35.46/0.9780	30.17/0.9335	25.09/0.8584
CAIN	AAAI'20	34.65/0.9730	34.91/0.9690	2.28	39.89/0.9900	35.61/0.9776	29.90/0.9292	24.78/0.8507
AdacoF	CVPR'20	34.47/0.9730	34.90/0.9680	2.24	39.80/0.9900	35.05/0.9753	29.46/0.9244	24.31/0.8439
BMBC	ECCV'20	35.01/0.9764	35.15/0.9689	2.04	39.90/0.9902	35.31/0.9774	29.33/0.9270	23.92/0.8432
ABME	ICCV'21	36.18/0.9805	35.38/**0.9698**	2.01	39.59/0.9901	35.77/0.9789	30.58/0.9363	25.42/0.8639
RIFE	ECCV'22	35.61/0.9779	35.29/0.9690	1.96	40.02/0.9904	35.72/0.9786	30.07/0.9326	24.82/0.8529
DQBC	IJCAI'23	36.37/**0.9812**	35.35/0.9696	1.86	40.15/0.9907	36.10/**0.9796**	30.78/**0.9371**	25.41/0.8628
Proposed Set 1	–	**36.38**/**0.9812**	35.32/0.9696	1.84	40.18/0.9906	36.13/0.9795	**30.80**/**0.9371**	**25.46**/**0.8629**
Proposed Set 4	–	36.33/0.9811	**35.39**/**0.9698**	**1.81**	**40.28**/**0.9908**	**36.16**/**0.9796**	30.76/0.9369	25.44/0.8627

Table 2
Performance comparison of augmented dataset: Evaluating the proposed enhanced method (Proposed-Aug) against various video frame interpolation (VFI) techniques.

Method	Year	Vimeo90K	UCF-101	Middlebury	SNU-FILM			
					Easy	Medium	Hard	Extreme
VFIformer	CVPR'22	36.50/0.9816	35.43/0.9700	1.82	40.13/0.9907	36.09/0.9799	30.67/0.9378	25.43/0.8643
IFRNet	CVPR'22	36.20/0.9808	35.42/0.9698	–	40.10/0.9906	36.12/0.9797	30.63/0.9368	25.27/0.8609
UPR-Net	CVPR'23	36.42/0.9815	**35.47**/0.9700	–	**40.44**/**0.9911**	36.29/**0.9801**	30.86/0.9377	**25.63**/0.8641
DQBC-Aug	IJCAI'23	36.57/0.9817	35.44/0.9700	1.78	40.31/0.9909	36.25/0.9799	**30.94**/0.9378	25.61/**0.8648**
Proposed Set 1-Aug	–	**36.58**/**0.9818**	35.42/0.9700	1.79	40.33/0.9908	36.26/0.9798	**30.94**/**0.9379**	**25.63**/0.8647
Proposed Set 4-Aug	–	36.53/0.9816	**35.47**/**0.9701**	**1.76**	40.40/0.9909	**36.30**/0.9799	30.93/0.9377	25.60/0.8645

During training, the AdamW optimizer is used, with a weight decay of 10^{-4} to regularize the model and prevent overfitting. The learning rate follows a cosine annealing schedule, starting at 2×10^{-4} and gradually reducing to 2×10^{-6} over 540 epochs. This approach ensures that the model learns efficiently, with a controlled reduction in the learning rate to achieve better convergence.

Data is processed in batches of 64, where each batch consists of 256×256 image patches. Given the significant memory demands of such large-scale processing, the batch size is chosen to maximize the use of available GPU memory while maintaining computational efficiency. The model's loss function is a combination of distillation loss and RoM loss, with weights of 0.01 and 0.02, respectively. This loss function ensures that the model learns to interpolate frames accurately while preserving the integrity of motion regions.

Memory management is carefully handled to optimize the use of GPU resources. The number of feature points and RoM boxes is adjusted based on the available GPU memory, with feature points set to either 4096 or 8192 and RoM boxes ranging from 16 to 24. Techniques such as gradient accumulation and mixed precision training are employed to manage memory usage effectively, allowing for larger models or higher resolutions to be processed without exceeding the GPU memory limits. These optimizations are essential for ensuring that the model performs well under the constraints of the available hardware, enabling the efficient training and evaluation of the proposed video frame interpolation model.

4.3. Comparison with existing methods

We conducted extensive evaluations against a range of leading Video Frame Interpolation (VFI) models, such as SepConv [7], ToFlow [17], CyclicGen [68], DAIN [11], CAIN [67], AdaCoF [10], BMBC [69], ABME [70], RIFE [13], and DQBC [23]. The results, shown in Table 1, indicate that on average, the proposed technique outperforms other models in terms of PSNR, SSIM [71], and IE metrics. Particularly, the proposed approach achieved the highest PSNR scores on the Vimeo90K and SNU-FILM datasets, with SSIM results on par with DQBC, recognized for its efficacy in VFI.

To enhance the performance of the propose method, we implemented a data augmentation scheme that leverages temporal reversal and spatial rotation of 90°, effectively quadrupling the dataset for testing. The final interpolation results were derived by averaging the outputs from these four diversified sets, yielding a marginal improvement over the initial approach. This augmented iteration of the proposed model is identified as Proposed-Aug in Table 2, where it is compared against recent models like VFIformer [21], IFRNet [72], UPR-Net [73], and DQBC [23].

We provide visual comparisons of our interpolated results in Fig. 9, showing the qualitative performance of our model against other notable methods such as ABME, RIFE, VFIformer, and DQBC. The figure illustrates the interpolation challenges and achievements in a video sequence characterized by nonlinear motion and significant variations.

For instance, in a frame capturing a golf swing, common challenges observed with other methods include blurring or loss of detail in the golf club's portrayal. The proposed method, however, reproduces the scene with exceptional fidelity, particularly in capturing the golf club with enhanced clarity.

In another example of a flying bird, although our method slightly misrepresents the right wing's position relative to the original, it notably circumvents the instability in wing depiction seen with other methods. This results in a more accurate representation of the bird's wing structure.

Lastly, in a scene depicting children running, competing methods struggle to accurately capture the structure of the hands. In contrast, the proposed model provides a clearer and more precise representation

Fig. 9. Visual assessment of nonlinear and large motion video. The proposed method outperforms competing approaches in producing sharp and precise imagery, adeptly handling scenes characterized by nonlinear movements and significant variations.

Table 3
Hyperparameter tuning analysis: Evaluation of the impact of three key hyperparameters including feature points, RoM boxes, and batch size on the performance of the proposed VFI model.

	Set 1	Set 2	Set 3	Set 4	Set 5
Feature point	2048	4096	4096	4096	8192
RoM box	8	16	20	24	24
Batch size	16	16	22	22	16

of the fingers, enhancing the overall detail and realism of the scene's subjects.

4.4. Parameter fine-tuning

We conducted a parameter fine-tuning to determine the optimal performance settings by adjusting hyperparameters such as batch size, number of feature points, and the number of RoM boxes. We tested four different settings, which are summarized in Table 3.

To design the proposed RoM loss, we first created a RoM box by determining the number of feature points to extract during the feature point extraction process. By calculating optical flow and detecting feature points, we can create a moving area that is ordered by the amount of movement. We tested different numbers of feature points to determine the optimal setting. We optimized the model's performance by adjusting memory usage. Specifically, we expanded the number of feature points and RoM boxes, and in Settings 3 and 4, we increased the batch size to 22. After experimentation, we determined that the most favorable results were achieved with 16 batches. Additionally, we observed improved performance when augmenting the number of feature points in Settings 2 and 5, along with the utilization of more RoM boxes. However, it is important to note that we faced memory constraints, which limited our ability to further increase the number of RoM boxes.

Table 4 shows the evaluation metrics (PSNR, SSIM, IE) on the Vimeo90K, Middlebury, UCF-101, and SNU-FILM datasets for each setting. Setting 2 provided the best performance on Vimeo90K, while Setting 5 showed the best performance on the other datasets.

The results suggest that improvements in performance can be attained through increasing the number of extracted feature points and the quantity of generated boxes. Nonetheless, such enhancements did not lead to parallel gains in performance on the Vimeo90K dataset. Moving forward, our research will focus on developing approaches to consistently optimize performance across all datasets within a uniform framework.

4.5. Discussions and limitations

Experimental results reveal several key phenomena related to the performance of the proposed model. In particular, increasing the number of extracted feature points and the amount of RoM boxes generated improved performance on certain datasets such as UCF-101 and SNU-FILM, but these improvements did not lead to parallel gains on the Vimeo90K dataset. This discrepancy suggests that the sensitivity of the model to these parameters may depend on the specific characteristics of the dataset, such as motion complexity or scene diversity.

The underlying cause of this phenomenon may lie in inherent differences between datasets. For example, Vimeo90K with certain motion patterns and scene structures may not benefit from additional feature points and RoM boxes as much as other datasets. This indicates that while the current configuration of the model is usually effective, further adjustments or adjustments may be required to maximize performance on different types of data.

Based on these results, we consider a more adaptive approach that allows us to dynamically adjust the model's parameters based on the characteristics of the dataset being processed. This can include implementing a framework that uniformly optimizes performance across all datasets by analyzing the characteristics of the dataset and

Table 4
Performance evaluation of settings 2 to 5 on Vimeo90K, UCF-101, Middlebury, and SNU-FILM datasets. The best results for each dataset are indicated by underlined values.

	Vimeo90K	UCF-101	Middlebury	SNU-FILM(Mean)	Average
Setting1	36.53/0.9815/1.92	35.45/0.9701/2.70	38.63/0.9880/1.80	33.27/0.9434/3.90	35.97/0.9708/2.58
Setting2	**36.58/0.9818/1.91**	35.42/0.9700/2.71	38.68/0.9881/1.79	33.29/0.9433/3.90	36.00/0.9708/2.58
Setting3	36.49/0.9816/1.93	35.41/0.9700/2.71	38.48/0.9881/1.81	33.15/0.9432/3.91	35.88/0.9707/2.59
Setting4	36.52/0.9816/1.92	35.42/0.9700/2.71	38.51/0.9873/1.81	33.25/0.9435/3.90	35.93/0.9706/2.59
Setting5	36.53/0.9817/1.92	**35.47/0.9701/2.70**	**38.82/0.9883/1.76**	**33.31/0.9433/3.90**	**36.03/0.9709/2.57**

automatically adjusting the settings of the model. Furthermore, other ways to improve the robustness of the model must be studied, especially if the efficiency of the current configuration is limited. This includes investigating more sophisticated data augmentation techniques or loss functions that can better capture the nuances of various video sequences.

In summary, while the proposed model shows strong performance on various datasets, there are clear limitations to generalizing improvements uniformly. To address these limitations, it is important to increase the efficiency of the model across a wide range of applications through targeted adaptation and improvement.

5. Conclusion

In this paper, we present an innovative approach to video frame interpolation that leverages Region of Motion (RoM) loss and self-attention mechanisms. This technique is specifically designed to overcome the complexities of interpolating frames involving large and nonlinear movements of objects. By utilizing RoM loss, the proposed model identifies and prioritizes regions of interest based on motion intensity through optical flow analysis between feature points. This enables precise interpolation in scenes characterized by significant motion by aligning the interpolated regions closely with the ground truth.

Furthermore, the integration of self-attention scores with features extracted from the Basic Encoder and ContextNet directs the model's focus towards critical motion areas within the frame, facilitating more accurate predictions.

Our approach advances video frame interpolation by introducing a loss function specifically tailored for complex motions. The method's capability to handle challenging scenarios makes it applicable to real-world tasks such as video enhancement, compression, and surveillance.

Our evaluation of this method across renowned benchmark datasets such as Vimeo-90K, Middlebury, UCF101, and SNU-Film demonstrates its superior performance, both quantitatively and qualitatively, against existing state-of-the-art methods.

Key contributions of our research include:

- Introducing the novel application of RoM loss in a video frame interpolation framework, enabling focused processing on significant motion regions for enhanced accuracy, particularly in videos featuring large and nonlinear object movements.
- Implementing self-attention scores to refine the model's understanding of frame dynamics, resulting in more coherent and lifelike interpolated frames.
- Achieving a balance between efficacy and efficiency, allowing for training and deployment on a single GPU, making it accessible for practical applications.

Our proposed method demonstrates strong interpolation capabilities for large object motions in video frame interpolation, leveraging artificial intelligence technology to efficiently deliver high-quality frames, making it suitable for real-time processing and large-scale applications. However, the research does have some limitations. While increasing the number of feature points and generated boxes can improve performance, these adjustments did not consistently yield gains on the Vimeo-90K dataset. Future work will focus on optimizing performance uniformly across all datasets.

Future research directions include exploring hybrid loss functions for better handling of subtle motions, optimizing self-attention mechanisms to reduce computational complexity, and extending the method to 3D interpolation or GAN-based enhancements.

CRediT authorship contribution statement

Yeongjoon Kim: Methodology, Conceptualization. **Sunkyu Kwon:** Writing – original draft, Data curation. **Donggoo Kang:** Resources. **Hyunmin Lee:** Resources. **Joonki Paik:** Writing – review & editing.

Code availability

The codes for using the dataset, and generating the experimental results are available in a public repository at https://github.com/VFI-FIRMA/FIRMA.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported partly by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by Korea government (MSIT) [2021-0-01341, Artificial Intelligent Graduate School Program(Chung-Ang University)], and partly by Field-oriented Technology Development Project for Customs Administration through National Research Foundation of Korea(NRF) funded by the Ministry of Science & ICT and Korea Customs Service (2021M3I1A1097911)

Appendix A. Region of motion loss weight

The Region of Motion (RoM) Loss does not simply minimize the overall loss between ground truth and predictions; instead, it selectively minimizes the loss by concentrating on regions with significant motion during training. This is accomplished by refining these high-motion regions to effectively differentiate and prioritize them. The selected regions, referred to as RoM boxes, are divided into union and intersection components, as shown in Fig. A.10. The overlapping portions of the RoM boxes are used as weights for the regions with significant motion.

To reformulate the equation to calculate the total area covered by k boxes, considering their overlapping regions. This formula accounts for the weights of the overlapping areas, providing a comprehensive calculation of the total area.

- Area of Each Box: Simply sum the areas of all individual boxes.

$$A\left(\bigcup_{i=1}^{k}\text{RoMBox}_i\right)$$

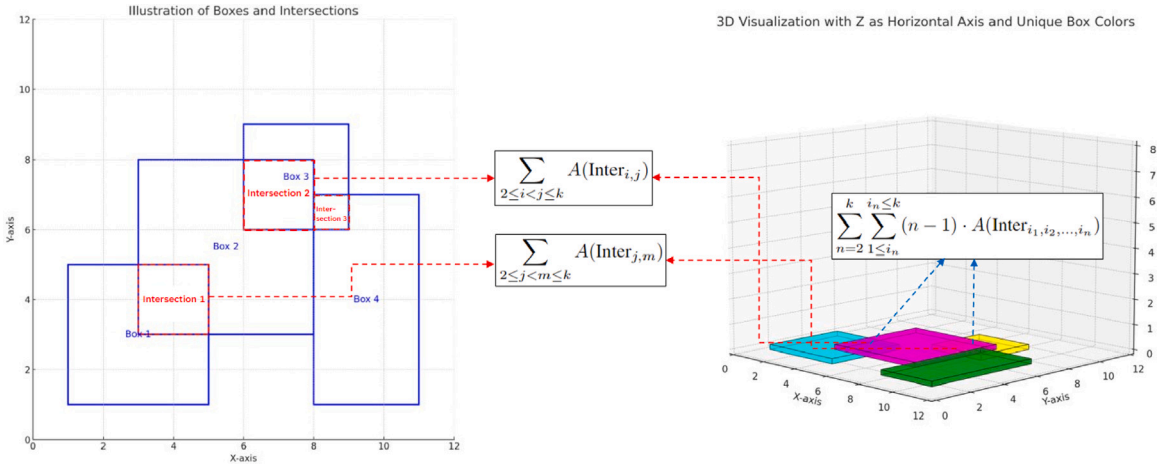


Fig. A.10. The image illustrates the overlap of bounding boxes and how these regions correspond to the equations for calculating the total covered area. The arrows indicate specific overlapping regions: one arrow marks the overlap between two boxes (denoted as $A(\text{Inter}_{i,j})$), and another marks the overlap among three boxes (denoted as $A(\text{Inter}_{j,m})$). The equation near the top represents the total area calculation, accounting for both individual and overlapping areas with appropriate weights.

- Area of Overlapping Between Two Boxes: Sum the areas of overlapping regions between two boxes, considering the weight of each intersection.

$$\sum_{2\le i<j\le k} A(\text{Inter}_{i,j})$$

- Area of Overlapping Among Three Boxes: Sum the areas of overlapping regions among three boxes, considering the weight of each intersection.

$$\sum_{3\le i<j<m\le k} A(\text{Inter}_{i,j,m})$$

- Area of Overlapping Among Multiple Boxes: Sum the areas of overlapping regions among multiple boxes, considering the weight of each intersection.

$$\sum_{n=2}^{k}\sum_{1\le i_n}^{i_n\le k}(n-1)\cdot A(\text{Inter}_{i_1,i_2,\ldots,i_n})$$

This formula ensures that the overlapping areas are appropriately weighted, allowing for a precise calculation of the total area covered by k boxes in a scenario where boxes may overlap in various configurations.

We consider a set of k bounding boxes, each defined by its coordinates (x_i, y_i) and dimensions (w_i, h_i). These boxes may overlap, forming complex intersection regions. Our goal is to calculate the union of all these areas, reflecting the true extent of coverage without duplication from overlaps.

The total area, A_{total}, is computed not by simply summing the areas of all boxes, but by considering the union of all boxes, thereby adjusting for any overlaps:

$$RoM_{total}^{weighted} = A\left(\bigcup_{i=1}^{k}\text{RoMBox}_i\right) + \sum_{n=2}^{k}\sum_{1\le i_n}^{i_n\le k}(n-1)\cdot A(\text{Inter}_{i_1,i_2,\ldots,i_n})$$

where, A_{total}, is calculated by considering the union of all bounding boxes, $A\left(\bigcup_{i=1}^{k}\text{RoMBox}_i\right)$, which represents the area without double-counting overlaps, and adding the weighted areas of intersections, $\sum_{n=2}^{k}\sum_{1\le i_n}^{i_n\le k}(n-1)\cdot A(\text{Inter}_{i_1,i_2,\ldots,i_n})$, where k is the total number of bounding boxes, $A(\text{RoMBox}_i)$ is the area of the ith box, $\text{Inter}_{i_1,i_2,\ldots,i_n}$ represents the intersection region where n boxes overlap, and $(n-1)$ serves as a weighting factor to adjust for overlapping redundancy and prevent multiple counting.

Appendix B. Additional experiments

We conducted additional experiments comparing our proposed interpolation method with the existing interpolation method. Fig. B.11 shows results in a video containing general linear movement. As expected, in linear motion, any model accurately predicts intermediate frames, closely resembling the ground truth.

In addition to the general video interpolation results tested in the paper, we also conducted experiments on complex videos with long distances between frames. Fig. B.12 shows the experimental results of interpolating frame 3 between frame 1 and frame 5. In the initial frame depicting a backflip, none of the models accurately generated the shoe's position. However, while the existing method does not properly understand the shape of the shoe, the proposed method can be seen to express the shape of the shoe well. The second frame shows a sword fight scene. where RIFE and VFIformer fail to shape the sword and DQBC succeeds, but the tip appears blurry. On the other hand, our proposed method faithfully renders the complete shape of the knife without any blurring. The concluding frame depicts a fish at a fishing spot, with our proposed method naturally reproducing the fish's shape.

To perform a slightly more difficult experiment, Fig. B.13 shows the results of an experiment with frame 4 interpolated between frames 1 and 7. The first frame shows a dog jumping, but in experiments where the period between frames 1 and 7 is long, the shape of the long tail is not properly recognized in all methods and is generated short. Nevertheless, the proposed method is significant in that it attempts to capture finely elongated shapes. In the second frame, where two girls are dancing, our model generates one girl's shoelaces in an accurate shape similar to ground truth.

Fig. B.14 is an enlarged version of some of the experimental results of Figs. B.12 and B.13. Specifically, in comparison to DQBC, which serves as the baseline model for our proposed method, our model demonstrates superior performance. This is evidenced by our method's ability to generate intermediate frames that closely resemble the ground truth more effectively than DQBC.

Data availability

Data will be made available on request.

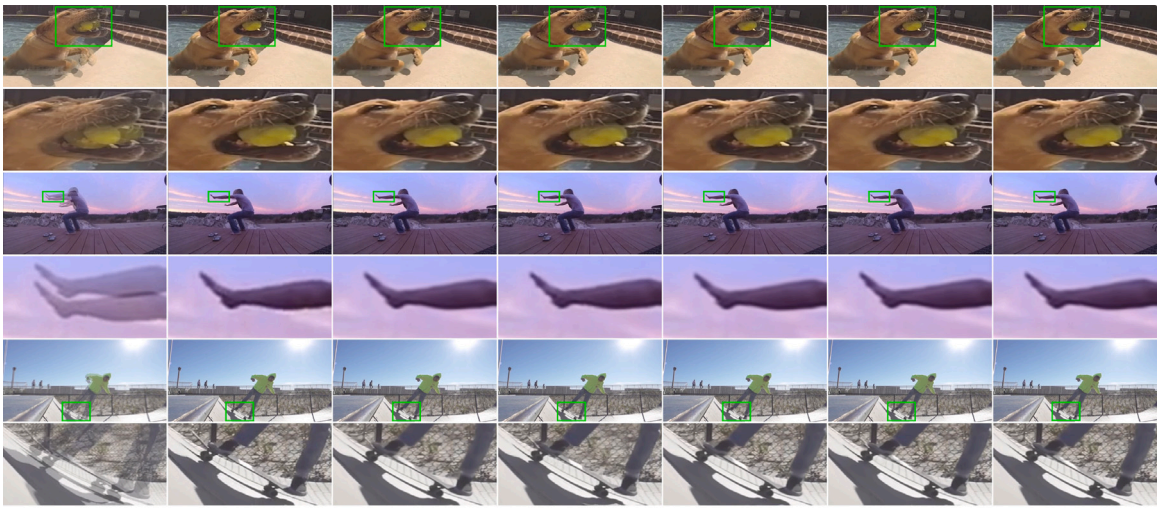


Fig. B.11. Visual comparison of the proposed method and other methods on a video containing linear and small movements. All methods produce sharp visual results that are similar to the ground truth.

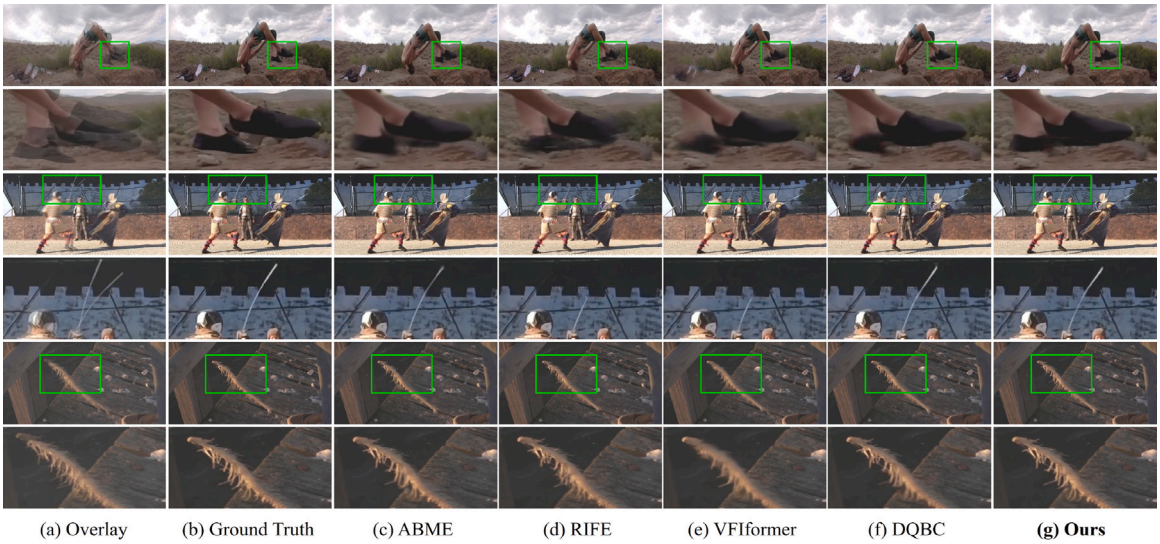


Fig. B.12. More visual evaluation on a video containing nonlinear and large movements. Experiment with interpolating frame 3 from frame 1 to frame 5.

Fig. B.13. More visual evaluation on a video containing nonlinear and large movements. Experiment with interpolating frame 4 from frame 1 to frame 7.

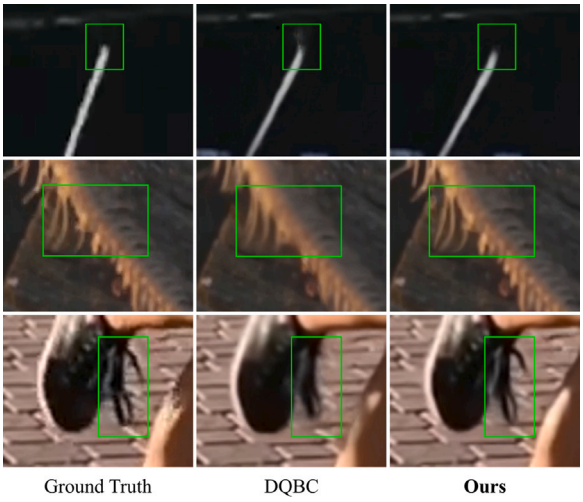


Fig. B.14. Comparison of the proposed model with DQBC.

References

[1] K. Xu, F. Ren, CSVideoNet: A real-time end-to-end learning framework for high-frame-rate video compressive sensing, in: 2018 IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE, 2018, pp. 1680–1688.
[2] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, J. Kautz, Super slomo: High quality estimation of multiple intermediate frames for video interpolation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9000–9008.
[3] M. Haris, G. Shakhnarovich, N. Ukita, Space-time-aware multi-resolution video enhancement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2859–2868.
[4] S.Y. Kim, J. Oh, M. Kim, Fisr: Deep joint frame interpolation and super-resolution with a multi-scale temporal loss, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11278–11286.
[5] X. Xiang, Y. Tian, Y. Zhang, Y. Fu, J.P. Allebach, C. Xu, Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3370–3379.
[6] S. Niklaus, L. Mai, F. Liu, Video frame interpolation via adaptive convolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 670–679.
[7] S. Niklaus, L. Mai, F. Liu, Video frame interpolation via adaptive separable convolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 261–270.
[8] X. Cheng, Z. Chen, Video frame interpolation via deformable separable convolution, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 10607–10614.
[9] X. Cheng, Z. Chen, Multiple video frame interpolation via enhanced deformable separable convolution, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (2021) 7029–7045.
[10] H. Lee, T. Kim, T.-y. Chung, D. Pak, Y. Ban, S. Lee, Adacof: Adaptive collaboration of flows for video frame interpolation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5316–5325.
[11] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, M.-H. Yang, Depth-aware video frame interpolation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3703–3712.
[12] W. Bao, W.-S. Lai, X. Zhang, Z. Gao, M.-H. Yang, Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement, IEEE Trans. Pattern Anal. Mach. Intell. 43 (3) (2019) 933–948.
[13] Z. Huang, T. Zhang, W. Heng, B. Shi, S. Zhou, Real-time intermediate flow estimation for video frame interpolation, in: European Conference on Computer Vision, Springer, 2022, pp. 624–642.
[14] Z. Liu, R.A. Yeh, X. Tang, Y. Liu, A. Agarwala, Video frame synthesis using deep voxel flow, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4463–4471.
[15] S. Niklaus, F. Liu, Context-aware synthesis for video frame interpolation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1701–1710.
[16] X. Xu, L. Siyao, W. Sun, Q. Yin, M.-H. Yang, Quadratic video interpolation, Adv. Neural Inf. Process. Syst. 32 (2019).
[17] T. Xue, B. Chen, J. Wu, D. Wei, W.T. Freeman, Video enhancement with task-oriented flow, Int. J. Comput. Vis. 127 (2019) 1106–1125.
[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[21] L. Lu, R. Wu, H. Lin, J. Lu, J. Jia, Video frame interpolation with transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3532–3542.

[22] Z. Shi, X. Xu, X. Liu, J. Chen, M.-H. Yang, Video frame interpolation transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17482–17491.
[23] C. Zhou, J. Liu, J. Tang, G. Wu, Video frame interpolation with densely queried bilateral correlation, 2023, arXiv preprint arXiv:2304.13596.
[24] Z. Pan, J. Lei, Y. Zhang, F.L. Wang, Adaptive fractional-pixel motion estimation skipped algorithm for efficient HEVC motion estimation, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 14 (1) (2018) 1–19.
[25] J.R. Bergen, P. Anandan, K.J. Hanna, R. Hingorani, Hierarchical model-based motion estimation, in: Computer Vision—ECCV'92: Second European Conference on Computer Vision Santa Margherita Ligure, Italy, May 19–22, 1992 Proceedings 2, Springer, 1992, pp. 237–252.
[26] D. Farin, et al., Evaluation of a feature-based global-motion estimation system, in: Visual Communications and Image Processing 2005, Vol. 5960, SPIE, 2005, pp. 1331–1342.
[27] J. Xu, H.-w. Chang, S. Yang, M. Wang, Fast feature-based video stabilization without accumulative global motion estimation, IEEE Trans. Consum. Electron. 58 (3) (2012) 993–999.
[28] P.-E. Sarlin, D. DeTone, T. Malisiewicz, A. Rabinovich, Superglue: Learning feature matching with graph neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4938–4947.
[29] J. Sun, Z. Shen, Y. Wang, H. Bao, X. Zhou, LoFTR: Detector-free local feature matching with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8922–8931.
[30] N.A. Mohd, S.A. Mostafa, A. Mustapha, A.A. Ramli, M.A. Mohammed, N.M. Kumar, Vehicles counting from video stream for automatic traffic flow analysis systems, 2020.
[31] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, T. Brox, Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.
[32] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, Flownet 2.0: Evolution of optical flow estimation with deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2462–2470.
[33] D. Sun, X. Yang, M.-Y. Liu, J. Kautz, Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8934–8943.
[34] Z. Teed, J. Deng, Raft: Recurrent all-pairs field transforms for optical flow, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer, 2020, pp. 402–419.
[35] E. Cuevas, D. Zaldívar, M. Pérez-Cisneros, H. Sossa, V. Osuna, Block matching algorithm for motion estimation based on artificial bee colony (ABC), Appl. Soft Comput. 13 (6) (2013) 3047–3059.
[36] S.T. Khawase, S.D. Kamble, N.V. Thakur, A.S. Patharkar, An overview of block matching algorithms for motion vector estimation, Rice (2017) 217–222.
[37] G. Senbagavalli, R. Manjunath, Motion estimation using variable size block matching with cross square search pattern, SN Appl. Sci. 2 (2020) 1–9.
[38] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[39] H. Bay, T. Tuytelaars, L. Van Gool, Surf: Speeded up robust features, in: Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9, Springer, 2006, pp. 404–417.
[40] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2564–2571.
[41] M. Muja, D.G. Lowe, Scalable nearest neighbor algorithms for high dimensional data, IEEE Trans. Pattern Anal. Mach. Intell. 36 (11) (2014) 2227–2240.
[42] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.
[43] H. Gao, X. Yu, Y. Xu, J.Y. Kim, Y. Wang, Monoli: Precise monocular 3d object detection for next-generation consumer electronics for autonomous electric vehicles, IEEE Trans. Consum. Electron. (2024).
[44] Z. Cao, L. Xu, D.Z. Chen, H. Gao, J. Wu, A robust shape-aware rib fracture detection and segmentation framework with contrastive learning, IEEE Trans. Multimed. 25 (2023) 1584–1591.
[45] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014, arXiv preprint arXiv:1409.0473.
[46] G. Zhang, Y. Zhu, H. Wang, Y. Chen, G. Wu, L. Wang, Extracting motion and appearance via inter-frame attention for efficient video frame interpolation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5682–5692.
[47] D. Danier, F. Zhang, D. Bull, LDMVFI: Video frame interpolation with latent diffusion models, 2023, arXiv preprint arXiv:2303.09508.
[48] X. Li, S. Yu, Y. Lei, N. Li, B. Yang, Intelligent machinery fault diagnosis with event-based camera, IEEE Trans. Ind. Inform. 20 (1) (2023) 380–389.
[49] X. Chen, X. Li, S. Yu, Y. Lei, N. Li, B. Yang, Dynamic vision enabled contactless cross-domain machine fault diagnosis with neuromorphic computing, IEEE/CAA J. Autom. Sin. 11 (3) (2024) 788–790.
[50] M.U. Nisa, D. Mahmood, G. Ahmed, S. Khan, M.A. Mohammed, R. Damaševičius, Optimizing prediction of YouTube video popularity using XGBoost, Electronics 10 (23) (2021) 2962.
[51] H. Gao, S. Wu, Y. Wang, J.Y. Kim, Y. Xu, FSOD4RSI: Few-shot object detection for remote sensing images via features aggregation and scale attention, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. (2024).
[52] H. Gao, W. Jiang, Q. Ran, Y. Wang, Vision-language interaction via contrastive learning for surface anomaly detection in consumer electronics manufacturing, IEEE Trans. Consum. Electron. (2024).
[53] Z. Chen, L. Tong, B. Qian, J. Yu, C. Xiao, Self-attention-based conditional variational auto-encoder generative adversarial networks for hyperspectral classification, Remote Sens. 13 (16) (2021) 3316.
[54] J. Shi, et al., Good features to track, in: 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 1994, pp. 593–600.
[55] B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: IJCAI'81: 7th International Joint Conference on Artificial Intelligence, Vol. 2, 1981, pp. 674–679.
[56] P. Simon, V. Uma, Deep learning based feature extraction for texture classification, Procedia Comput. Sci. 171 (2020) 1680–1687.
[57] Y. Liang, X. Li, N. Jafari, J. Chen, Video object segmentation with adaptive feature bank and uncertain-region refinement, Adv. Neural Inf. Process. Syst. 33 (2020) 3430–3441.
[58] G. Pei, F. Shen, Y. Yao, G.-S. Xie, Z. Tang, J. Tang, Hierarchical feature alignment network for unsupervised video object segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 596–613.
[59] M. Zhen, S. Li, L. Zhou, J. Shang, H. Feng, T. Fang, L. Quan, Learning discriminative feature with crf for unsupervised video object segmentation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16, Springer, 2020, pp. 445–462.
[60] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, X. Xue, Exploring inter-feature and inter-class relationships with deep neural networks for video classification, in: Proceedings of the 22nd ACM International Conference on Multimedia, 2014, pp. 167–176.
[61] L. Shao, Z. Cai, L. Liu, K. Lu, Performance evaluation of deep feature learning for RGB-D image/video classification, Inform. Sci. 385 (2017) 266–283.
[62] S. Xie, C. Sun, J. Huang, Z. Tu, K. Murphy, Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 305–321.
[63] C. Harris, M. Stephens, et al., A combined corner and edge detector, in: Alvey Vision Conference, Vol. 15, Citeseer, 1988, pp. 10–5244.
[64] J.V. Stone, Information theory: a tutorial introduction, 2015.
[65] K. Soomro, A.R. Zamir, M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, 2012, arXiv preprint arXiv:1212.0402.
[66] S. Baker, D. Scharstein, J. Lewis, S. Roth, M.J. Black, R. Szeliski, A database and evaluation methodology for optical flow, Int. J. Comput. Vis. 92 (2011) 1–31.
[67] M. Choi, H. Kim, B. Han, N. Xu, K.M. Lee, Channel attention is all you need for video frame interpolation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 10663–10671.
[68] Y.-L. Liu, Y.-T. Liao, Y.-Y. Lin, Y.-Y. Chuang, Deep video frame interpolation using cyclic frame generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 8794–8802.
[69] J. Park, K. Ko, C. Lee, C.-S. Kim, Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, Springer, 2020, pp. 109–125.
[70] J. Park, C. Lee, C.-S. Kim, Asymmetric bilateral motion estimation for video frame interpolation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14539–14548.
[71] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[72] L. Kong, B. Jiang, D. Luo, W. Chu, X. Huang, Y. Tai, C. Wang, J. Yang, Ifrnet: Intermediate feature refine network for efficient frame interpolation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1969–1978.
[73] X. Jin, L. Wu, J. Chen, Y. Chen, J. Koo, C.-h. Hahm, A unified pyramid recurrent network for video frame interpolation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1578–1587.

Yeongjoon Kim received his BS in Business Administration from Kookmin University and his MS in Artificial Intelligence Engineering from Chung-Ang University. He is currently working as a Principal Researcher in the AI field at Farmhannong, and is interested in the development of smart farm industry using AI. His main interests include computer vision, LLM, image segmentation, and meta-learning (zero/few-shot learning) for domain adaptation.

Sunkyu Kwon was born in Seoul, Korea, in 1997. He received a B.S degree in Software Engineering from Catholic Kwandong University, South Korea, in 2022. Also, he received the M.S. degree in AI Imaging at Chung-Ang University, South Korea, in 2024. His research interests include video stabilization and video frame interpolation.

Donggoo Kang was born in Seoul, Korea, in 1992. He received the B.S. degree in Financial Economics from Seokyeong University, South Korea, in 2018. Also, he received the M.S. degree in AI Imaging at Chung-Ang University, South Korea, in 2020. Currently, he is pursuing a Ph.D. degree in AI Imaging at Chung-Ang University. His research interests include computational photography and human–object interaction discovery.

Hyunmin Lee was born in Changwon, Korea, in 1997. He received a B.S degree in Industrial Systems Engineering from Gyeongsang National University in 2022. Currently, he is pursuing a M.S degree in Artificial Intelligence at Chung-Ang University. His research interests include the discovery of human–object interactions.

Joonki Paik was born in Seoul in 1960, earned his BS from Seoul National University and MS/Ph.D. from Northwestern University. He began his career at Samsung Electronics, designing image stabilization chipsets. Since 1993, he has been a professor at Chung-Ang University. Dr. Paik has held roles as Dean, technical consultant, and Project Manager for Korea's Military AI Education Program. He is a two-time Chester-Sall Award recipient and served as President of the Institute of Electronics and Information Engineers. Since 2020, he has been a graduate school director at Chung-Ang University.