# Complex Motion-aware Splatting for Video Frame Interpolation

Minho Park, Yuseok Bae
*Electronics and Telecommunications Research Institute(ETRI)*
Daejeon, Korea
{roger618, baeys}@etri.re.kr

***Abstract*—Video frame interpolation, a crucial component of computer vision, synthesizes additional frames to enhance the frame rate of a video, leading to improved performance with minimal additional cost. Despite recent advancements with deep learning and convolutional neural networks (CNNs), it still remains a challenge to generate precise intermediate frames, especially when complex and fast motions are involved. This paper presents a novel deep learning-based framework for video frame interpolation that incorporates a complex motion detection module and proposes a complex motion-aware splatting (CMS) method. We employ a forward warping approach that uses a complex motion map as a weight map in splatting. The framework further leverages a module that embeds temporal and spatial information from the frame sequence to acquire motion information. The effectiveness of our proposed model is demonstrated through qualitative and quantitative results on a public dataset.**

***Index Terms*—Video frame interpolation, Deep learning, Motion estimation**

## I. Introduction

Video frame interpolation has long been a subject of great interest in the field of computer vision. It synthesizes one or more frames between consecutive frames in a video to increase the overall frame rate. This method offers the advantage of improving low framerate videos to high framerate ones without additional costs. Frame interpolation typically involves two distinct stages [1]–[3]. Firstly, the pixel-level motion between two frames is estimated, resulting in flow maps that represent the motion of pixels. Secondly, based on the obtained flow maps, intermediate frames are predicted. The quality of the final intermediate frames depends on the performance of the first stage. There are various factors that make motion prediction challenging in video frame interpolation. Large and nonlinear motion patterns, occlusion, and changes in lighting conditions all contribute to the difficulty of accurately predicting motion [4], [5]. Consequently, as the predicted motion flow map becomes less accurate, the generation of precise intermediate frames becomes increasingly challenging. Therefore, achieving accurate motion prediction under such challenging conditions remains a key and formidable problem in video frame interpolation.

In recent years, deep learning approaches, particularly convolutional neural networks (CNNs), have led to the emergence of various deep learning-based frame interpolation methods [6]–[8]. These methods have achieved state-of-the-art performance compared to traditional hand-crafted approaches.

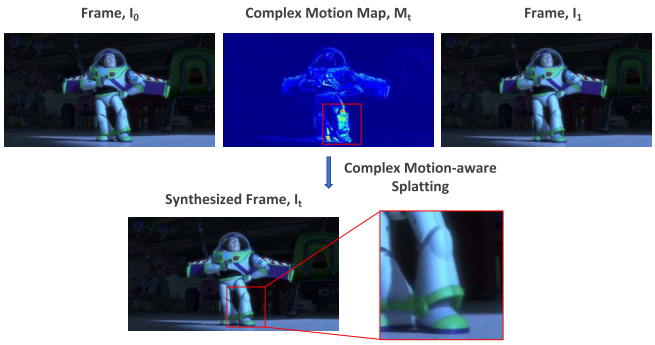


Fig. 1. An overview illustrating the method of generating intermediate frames using splatting with the Complex Motion Map.

CNNs have been employed to estimate reliable motion vectors by estimating optical flow and to synthesize intermediate frames by warping with the optical flow. In [6], a deep neural network was trained to synthesize intermediate frames by estimating flowing pixel values named deep voxel flow. Unlike existing optical flow estimation networks, this proposed network did not require optical flow supervision. Furthermore, in [7], a novel context-aware synthesis approach was introduced, which involved warping not only the input frames but also their pixel-wise contextual information. Prior to being fed into the frame synthesis network, both the input frames and contextual information were pre-warped. In [8], network for fully differentiable forward warping, which is named softmax splatting is proposed. It utilizes task-specific feature pyramids and resolves the issue of multiple source pixels being warped to a single target pixel.

Despite the advancements made by previous methods in video frame interpolation, synthesizing intermediate frames remains challenging, particularly when dealing with complex and fast motions. This paper aims to propose a novel deep learning-based framework for video frame interpolation that effectively synthesizes intermediate frames, even in sequences with complex motion. To achieve this goal, the proposed method incorporates complex motion detection module to identify specific regions where complex motions are likely to occur. We propose a complex motion-aware splatting (CMS) method that enables forward warping based on the acquired motion information. In the proposed method, we process the

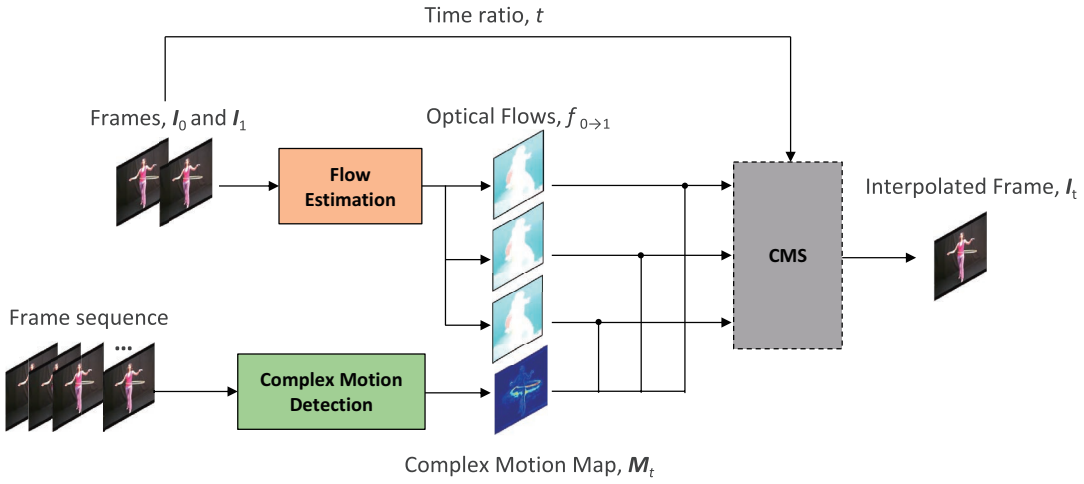


Fig. 2. The overview of the proposed Complex Motion-aware Splatting (CMS) for video frame interpolation. Our proposed method comprises three main modules: 1) Flow estimation, which takes two frames as input and predicts multiple optical flows between the consecutive frames. 2) Complex motion detection, which takes a frame sequence as input and acquires a complex motion map from within the sequence. 3) Complex motion-aware splatting (CMS), which performs splatting based on the optical flow and complex motion map.

acquired motion information by converting it into a complex motion map for the forward warping. More specifically, we focus on pixels that have undergone complex motion by assigning them higher weights within the map. This strategic weighting enables the production of accurate intermediate frames, even in the presence of sophisticated motion patterns. The key contributions of this paper are summarized as follows:

- We propose a novel forward warping approach, CMS based on detecting complex motion in frame sequences. We acquire a complex motion map that contains motion information and use this as a weight map in splatting.
- We propose a framework that employs a module to embed the temporal and spatial information of the frame sequence for acquiring motion information, and subsequently, derives a complex motion map based on this.
- To validate the effectiveness of our proposed model, we show both qualitative and quantitative results on a public dataset.

## II. Method

### A. Complex Motion Map

We utilize a complex motion detection module for the acquisition of the complex motion map. Our approach is inspired by the concept of the exceptional motion estimator proposed in study A, leading us to propose an LSTM [10]-based video frame reconstruction module. We adhere to the same learning settings as in study [1], initially training this module with videos composed solely of linear and slow movements. Upon completion of this initial training phase, we fix the parameters, thereby preventing any further adjustments during subsequent learning. The reconstructed frame can be represented by

$$\hat{\mathbf{I}}_t = g(\mathbf{I}_{t-2N-2}, \mathbf{I}_{t-2N}, \ldots, \mathbf{I}_t), \tag{1}$$

where $g$ refers to LSTM-based reconstruction model.

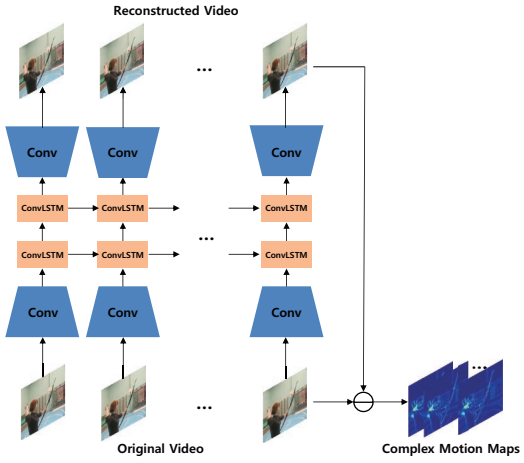


Fig. 3. The architecture of the complex motion detection module. This module is trained to reconstruct the input frame sequence, and during inference, it outputs the difference between the reconstructed frames and the ground-truth frames as the complex motion map.

When a video containing non-linear movements is input into this module, it becomes evident that the pixels experiencing complex motions are not accurately restored. We quantify the difference between these inaccurately restored frames and the actual ground-truth frames, designating this discrepancy as the complex motion map. The extent of inaccuracy in the generated frame correlates with the complexity and non-linearity of the object's movement within the video, resulting in larger errors for the corresponding pixels. Leveraging this observation, we effectively utilize this error map as a complex motion map. The complex motion map can be expressed by

$$\begin{aligned} \mathbf{M}_t &= |\mathbf{I}_t - f(\mathbf{I}_{t-2N-2}, \mathbf{I}_{t-2N}, \ldots, \mathbf{I}_t)| \\ &= |\mathbf{I}_t - \hat{\mathbf{I}}_t|. \end{aligned} \tag{2}$$

Figure 3 shows the architecture of the proposed complex motion detection module.

### B. Complex Motion-aware Splatting

Interpolation techniques using the warping approach via optical flow can be primarily bifurcated into two types: forward warping and backward warping. Forward warping benefits from the ability to leverage well-trained optical flow estimation techniques and holds considerable potential for performance improvement. However, it requires calculations for z-buffering and necessitates subsequent hole filling, which is a notable drawback. On the other hand, backward warping techniques operate by approximating flows, such as through the quadratic equation method, rather than computing precise flows. As a result, the generated outputs often contain significant errors, which is a disadvantage. Nevertheless, backward warping techniques are comparatively easier to implement because they do not require additional hole filling processes. Due to the relative simplicity of the backward warping approach, many studies have focused on developing frame interpolation technologies based on this method. However, the need for additional modules and computations to generate backward flow introduces extra cost and can contribute to making the overall model heavier.

To address this issue, the 'softmax splatting' [8] approach leverages splatting, a method within the forward warping technique. It utilizes PWC-Net [11], a state-of-the-art model that delivers high performance in optical flow estimation. Building on this softmax splatting approach, we propose complex motion-aware splatting (CMS) that assigns weights to important pixels during splatting. The CMS is defined as follows:

$$\vec{\epsilon}\left(\mathbf{I}_0, f_{0\to t}\right) = \frac{\vec{\Sigma}\left(\mathbf{M}_0 \cdot \mathbf{I}_0, f_{0\to t}\right)}{\vec{\Sigma}\left(\mathbf{M}_0, f_{0\to t}\right)}, \tag{3}$$

where $\vec{\Sigma}$ is summation splatting defined in [8]. $\mathbf{M}_0$ represents the complex motion map when $t$ is 0.

### C. Implementation Details

We adopt the pre-trained motion refinement module from [12] and fine-tune it using CMS. In order to acquire the complex motion map, multiple input frames are required, which we have configured as a hyper-parameter. We set and train with a constant number of 11 input frames. Within the motion refinement module, the number of optical flows extracted is fixed at four. We use the sum of Charbonnier loss [13] and the census loss [14] as the training loss.

## III. Experiments

### A. Dataset

We utilize two separate datasets for training and validation. Initially, we use a non-exceptional motion dataset [9], which is composed of videos with linear and slow movements, to train the complex motion detection module. This video set consists of car driving with a steady speed and drone footage capturing natural landscapes, among others. We train the complex motion detection module using these video frames and do not use them for subsequent training (all parameters are fixed). We employ the UCF-101 dataset [15] to train the flow estimation module. The UCF-101 dataset includes various action categories that contain complex motions, and it consists of a total of 101 action categories. We separate the videos included in the UCF-101 dataset into frame units and resize them to a size of 224×224.

TABLE I
Performance Comparison on the UCF101 Dataset

| | **PSNR** |
|---|---|
| SepConv | 34.78 |
| DAIN | 35.00 |
| CAIN | 34.98 |
| AdaCoF | 34.90 |
| SoftSplat | 35.39 |
| M2M-PWC | 35.17 |
| CMS(Ours) | **36.59** |

### B. Quantitative Results

To validate the effectiveness of CMS, we compare the performance of several state-of-the-art models on the UCF-101 dataset. The models to be compared are SepConv [5], DAIN [3], CAIN [7], AdaCoF [16], SoftSplat [8], and M2M-PWC [12]. SepConv and AdaCoF are interpolation method based on convolution kernel estimation. The remaining methods, DAIN, CAIN, SoftSplat, and M2M-PWC, focus on how to accurately estimate the optical flow between two consecutive frames, and how to improve the quality of the intermediate frame based on it. Typically, the PSNR metric is used to measure the quality of the generated intermediate frames. The results of the intermediate frame quality produced by the SOTA models and our proposed CMS are presented in Table 1. It can be seen that the performance of our proposed method outperforms that of other models.

### C. Qualitative Results

Figure 4 shows the qualitative results of frame interpolation using our proposed model. The first two columns are the input images, and the third column represents the complex motion map. The fourth and fifth columns are the ground-truth frame and the generated frame respectively. As one can observe from the two input frames, the UCF-101 dataset contains a substantial amount of relatively dynamic movements. Therefore, one can identify areas with large motion between the two frames. These areas are highly activated in the complex motion map. We leveraged this information to train the CMS, which enabled us to generate frames that closely resemble the ground-truth.

## IV. Conclusion

In this paper, we have proposed a novel deep learning-based framework for video frame interpolation which successfully

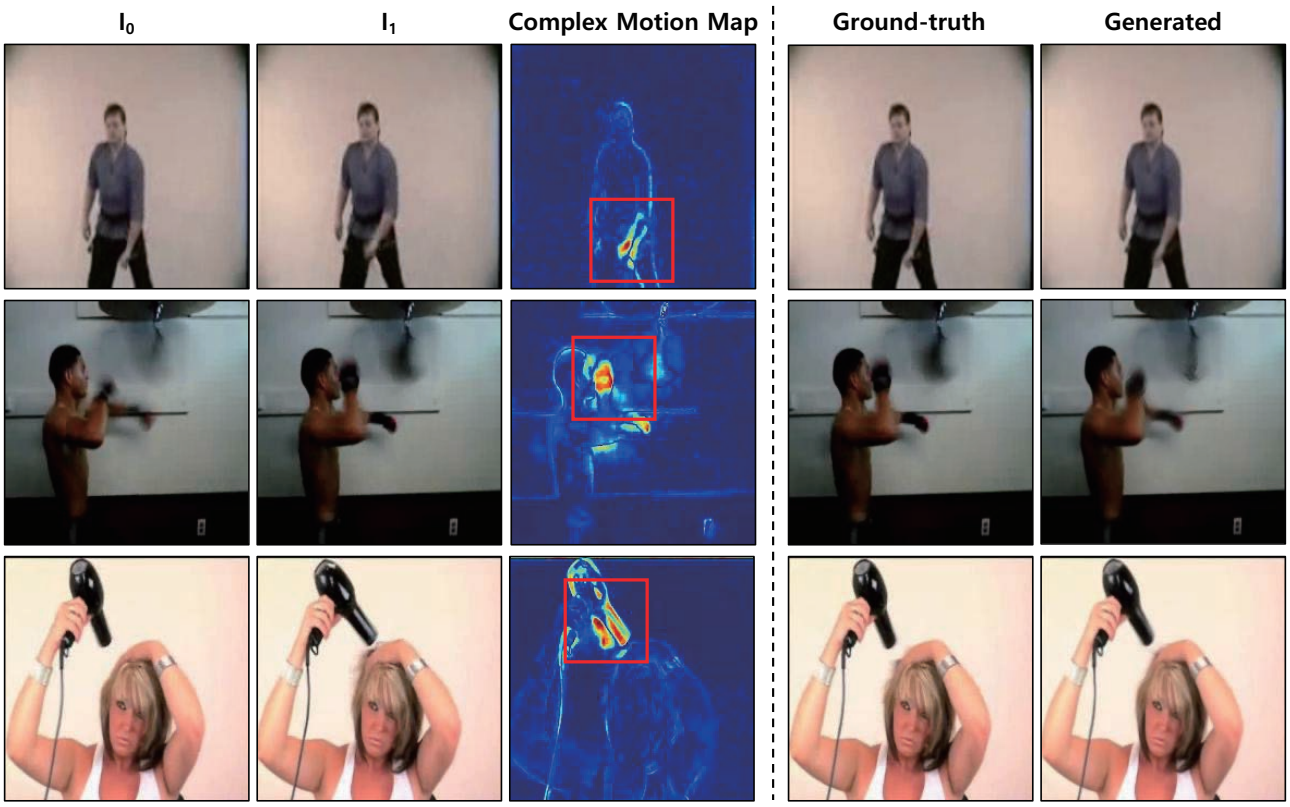


Fig. 4. Qualitative results on the UCF101 Dataset

manages complex and fast motions. This framework introduces a complex motion detection module and a complex motion-aware splatting (CMS) method, which leverages the notion of a complex motion map to enhance the accuracy of frame interpolation. The motion map, derived from the error differences in a LSTM-based video frame reconstruction module, serves as a useful tool for prioritizing complex motion during the splatting process. We have also presented a novel forward warping approach, which effectively synthesizes intermediate frames in videos that contain complex motion. By focusing on pixels undergoing complex motion and assigning them higher weights within the map, our CMS method provides a way to generate more accurate intermediate frames even in the presence of sophisticated motion patterns. The effectiveness of our proposed model has been demonstrated through both qualitative and quantitative results on the UCF-101 public dataset, with our model outperforming state-of-the-art methods. The introduction of the complex motion map, and its integration within the splatting process, has been a key contributing factor in the superior performance of our proposed method.

## ACKNOWLEDGMENT

This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2014-3-00123, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis)

## REFERENCES

[1] Park, Minho, et al. "Robust video frame interpolation with exceptional motion map." IEEE Transactions on Circuits and Systems for Video Technology 31.2 (2020): 754-764.
[2] Jiang, Huaizu, et al. "Super slomo: High quality estimation of multiple intermediate frames for video interpolation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[3] Bao, Wenbo, et al. "Depth-aware video frame interpolation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[4] Niklaus, Simon, Long Mai, and Feng Liu. "Video frame interpolation via adaptive convolution." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
[5] Niklaus, Simon, Long Mai, and Feng Liu. "Video frame interpolation via adaptive separable convolution." Proceedings of the IEEE international conference on computer vision. 2017.
[6] Liu, Ziwei, et al. "Video frame synthesis using deep voxel flow." Proceedings of the IEEE international conference on computer vision. 2017.
[7] Niklaus, Simon, and Feng Liu. "Context-aware synthesis for video frame interpolation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[8] Niklaus, Simon, and Feng Liu. "Softmax splatting for video frame interpolation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[9] Kim, Hak Gu, et al. "Vrsa net: Vr sickness assessment considering exceptional motion for 360 vr video." IEEE transactions on image processing 28.4 (2018): 1646-1660.
[10] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.
[11] Sun, Deqing, et al. "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[12] Hu, Ping, et al. ”Many-to-many splatting for efficient video frame interpolation.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[13] Charbonnier, Pierre, et al. ”Two deterministic half-quadratic regularization algorithms for computed imaging.” Proceedings of 1st international conference on image processing. Vol. 2. IEEE, 1994.
[14] Meister, Simon, Junhwa Hur, and Stefan Roth. ”Unflow: Unsupervised learning of optical flow with a bidirectional census loss.” Proceedings of the AAAI conference on artificial intelligence. Vol. 32. No. 1. 2018.
[15] Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah. ”UCF101: A dataset of 101 human actions classes from videos in the wild.” arXiv preprint arXiv:1212.0402 (2012).
[16] Lee, Hyeongmin, et al. ”Adacof: Adaptive collaboration of flows for video frame interpolation.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.