

Enhanced Correlation Matching based Video Frame Interpolation

Sungho Lee* Narae Choi* Woong Il Choi
Samsung Research, South Korea

Abstract

We propose a novel DNN based framework called the Enhanced Correlation Matching based Video Frame Interpolation Network to support high resolution like 4K, which has a large scale of motion and occlusion. Considering the extensibility of the network model according to resolution, the proposed scheme employs the recurrent pyramid architecture that shares the parameters among each pyramid layer for the optical flow estimation. In the proposed flow estimation, the optical flows are recursively refined by tracing the location with maximum correlation. The forward warping based correlation matching enables to improve the accuracy of flow update by excluding incorrectly warped features around the occlusion area. Based on the final bi-directional flows, the intermediate frame at arbitrary temporal position is synthesized using the warping and blending network and it is further improved by refinement network. Experiment results demonstrate that the proposed scheme outperforms the previous works at 4K video data and low-resolution benchmark datasets as well in terms of objective and subjective quality with the smallest number of model parameters.

1. Introduction

Video frame interpolation (VFI) is a task to generate intermediate frames between consecutive frames and its main application is to increase frame-rate [14] or to represent slow-motion video [9]. It is also applicable to synthesize novel views of 3D contents [6]. Advanced deep learning model based approaches have been introduced to achieve a better quality of the interpolated frames by discovering knowledge from large-scale diverse video scenes.

However, those approaches have a limitation on the range of motion coverage because the network only explores the motion of pixels within the receptive field of the CNN model that is pre-defined by their kernel size. As the most of capturing and display devices already support 4K resolution, the support of higher resolution is an inevitable feature of the practical application of VFI. The limitation of

handling large motion will become more pronounced as the image resolution increases and will be a major problem of practical use.

We can simply consider increasing the receptive field of a model by increasing kernel size or depth of network which attend utilizing more parameters to overcome the limitation of motion range. However, it is an inefficient solution because the memory size and computational complexity will be increased accordingly. Instead of expanding network kernel, the pyramid based architecture [21, 22] has been proposed which progressively updates optical flow from coarse-scale to fine. Even though the network of each layer has a limited receptive field, the pyramid structure can extend the motion coverage by adding the number of pyramid layers. Recently, a recurrent model based pyramid network has been proposed to avoid the model parameters increasing according to the number of layers [18, 27]. It basically re-uses the fixed model parameters in each layer, and the optical flows are iteratively refined from residual flows using the same network.

We propose an enhanced correlation matching based video frame interpolation network (ECMNet) to support high-resolution video with the lightweight model. The proposed network consists of three sub-networks; The first one is the bi-directional optical flow estimation between two consecutive input frames. Inspired by recurrent pyramid structure, this module adopts the shared pyramid structure to extend the receptive field efficiently. Also, we present an enhanced correlation matching algorithm to improve the accuracy of flow estimation. The next part is a frame synthesis module to generate intermediate flows and frame at an arbitrary temporal position. The last is a refinement module to improve the interpolated frame. Unlike the first module, the last two modules operate once with the original scale of input frames. This leads to computational efficiency and reduction of network parameters. In summary, the major contributions of our paper are as follows:

- We propose an enhanced correlation computation method based on the forward warped feature for flow estimation.
- At each pyramid layer, flows are gradually updated us-

* indicates equal contribution.

ing the highest feature correlation based on a novel feature matching algorithm.

- The proposed network achieves the state-of-the-art performance with the fewest number of parameters at 4K dataset compared to the existing VFI methods.

2. Related works

2.1. Video frame interpolation

In this section, we focus on recently proposed learning based algorithms. It can be categorized as kernel based and flow based approaches. Long *et al.* [13] firstly introduced a generic CNN model to directly generate an intermediate frame. Since those generic models are not enough to exploit various motion characteristics of natural video, the result images are damaged by blur artifacts. It is enhanced by AdaConv and SepConv methods proposed by Niklaus *et al.* [16], which generate interpolated pixels by locally convolving the input frames. They formulated pixel interpolation as convolution over pixel patches instead of relying on optical flow, and it is able to synthesize pixels from relatively large neighborhoods than previous work. AdaCoF [11] proposed to apply a deformable convolution kernel and CAIN [5] proposed to utilize a feature reshaping operation, PixelShuffle with channel attention to synthesis image. These kernel based approaches tend to have a relatively simpler design of network model, while those have a limitation on the range of motion coverage because the pixels that are located out of range of the receptive field could not participate in interpolation. If it is tried to increase the coverage of large motion in high resolution, it requires a high memory footprint and heavy computational complexity.

The second approach is to estimate the motion flow of each pixel between consecutive frames and synthesize an image from the corresponding pixels guided by the estimated flow. Liu *et al.* [12] proposed the Deep Voxel Flow (DVF) to perform trilinear interpolation using the spatio-temporal CNN model. Bao *et al.* [9] proposed Super-Slomo which combines U-Net architecture of DVF and kernel based image synthesis between warped images. The main architecture in Super-Slomo [9] is the cascade model of bi-directional flow estimation between input frames and image synthesis to blend two warped images guided by estimated bi-directional flows and it becomes a standard framework of flow based frame interpolation methods. However, those flow estimation based on U-Net structure still has a limitation on handling large motion. Instead of U-Net [17], it has been proposed to employ an advanced flow estimation model like PWC-Net [21] applied iterative refinement using coarse-to-fine pyramid structure.

For the image synthesis part, a key issue is how to properly handle the occlusion area during blending warped pix-

els. MEMC-Net [3] implicitly handles occlusion by estimating occlusion masks from features of input frames, while DAIN [2] explicitly detects occlusion from estimated depth information. From depth, it is determined whether the object is foreground or background and it is possible to select pixels from the foreground object in the area where both objects are overlapped. The performance should depend on the accuracy of depth and it is still challenging to obtain accurate depth from a single image.

While most flow based approaches perform backward warping for image synthesis, it is proposed in Softmax splatting [15] to perform forward warping by solving the ambiguity on overlapping the pixels from different flows to the same pixel.

2.2. Flow Estimation Network for Large Motion

From traditional approaches of an optical flow estimation, multi-resolution based approaches were already introduced which refine the resolution of flows from coarse-scale to fine in order to avoid drastically increased computational complexity of full search [4]. In contrast to traditional coarse-to-fine approaches, PWC-Net [21] utilizes a learnable feature pyramid instead of image pyramid to overcome the variation of lighting changes or shadow. Since it has been the state-of-the-art method until RAFT [22] is introduced in 2020, PWC-Net has been mainly adopted in flow based VFI scheme [2, 15, 25]. RAFT [22] pointed out the drawbacks of the coarse-to-fine approach; the difficulty of recovering estimation error at a coarse resolution and the tendency to miss the small fast-moving object. It proposes a multi-scale correlation based estimation approach to overcome those issues. However, those approaches still have limitations in two aspects; one is that network size increases in proportion to the number of pyramid layers and the other is that it is required to re-train the network if the target resolution is different to the resolution of trained images. To overcome those issues, RRPN [27] proposed the recurrent pyramid model to share the same weights and base network among different pyramid layers. Compared to RRPN [27], XVFI-Net [18] proposes more generalized model to synthesis frames at the arbitrary temporal position because RRPN was only applicable to synthesize the center frame.

The network architecture of the proposed ECMNet is designed based on the recurrent pyramid approach as well. Compared to RRPN and XVFI-Net [18, 27], our ECMNet applied correlation matching based flow estimation instead of directly deriving flows from U-Net. Compared to PWC-Net and RAFT [21, 22], the correlation cost is constructed by comparing forward warped feature instead of backward warped one. Besides, while the existing methods [21, 22] derive the flows by the network with cost volume as an input, the proposed method determines the optical flows directly by checking the maximum correlation value.

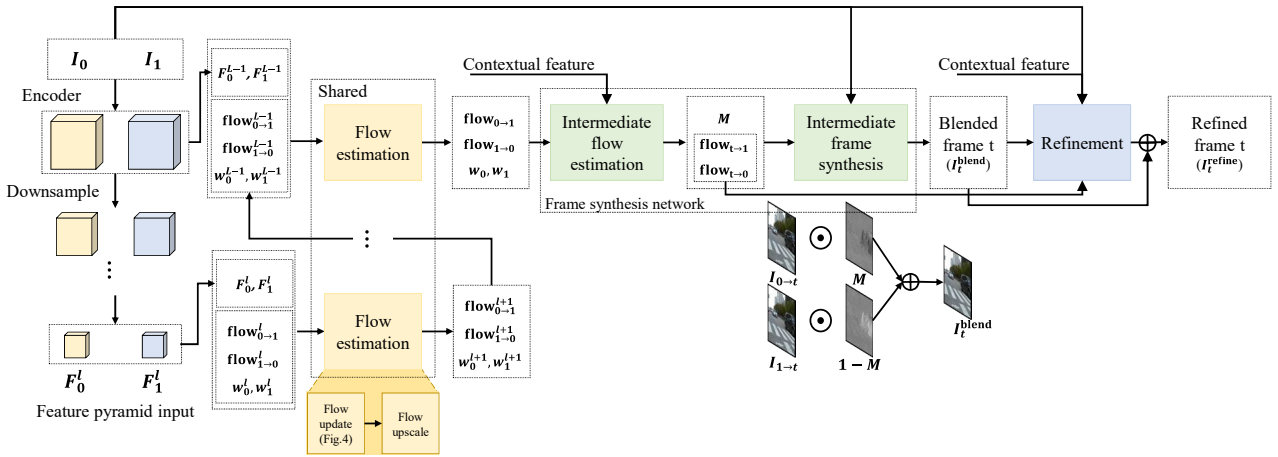


Figure 1: Overview of the network. Our framework consists of three sub-networks; Bi-directional flow estimation network, frame synthesis network, and refinement network. Shared pyramid network for flow estimation is denoted to yellow boxes. Green boxes indicates frame synthesis network which computes intermediate flow and initial intermediate frame for the arbitrary temporal position $t \in (0, 1)$. Finally, refinement network corrects the intermediate frame using estimated residual. Note that w_n ($n = 0$ or 1) is the learnable importance weight in F_n coordinate to resolve the forward warping issues, where multiple pixels from F_n map to the same target pixels in F_{1-n} [15], and learnable blending mask M is the confidence ratio of the warped image $I_{0 \rightarrow t}$.

3. Proposed algorithm

The overview of our framework is shown in figure 1. Our network takes consecutive images I_0 and I_1 , and the desired temporal position $t \in (0, 1)$ as inputs to synthesize the intermediate frame I_t .

In the flow estimation module, optical flows between input frames are estimated by comparing two encoded features. The intermediate flow estimation network then computes optical flows from the temporal position t to each input. Given the input images and intermediate flows, warping and blending are performed to synthesis the output frame at time t . Finally, the output image of the frame synthesis network is further improved at the following refinement network. The flow estimation module adopts a pyramid structure to compute the optical flow for input images from low to original resolution. After flow estimation, intermediate frame synthesis and refinement are performed only with the original scale of input images. This structure enables frame synthesis at any temporal position with estimated bi-directional optical flows. The details of the network architecture are described in the supplementary material.

3.1. Bi-directional optical flow estimation

Optical flow estimation for frame interpolation can be categorized as U-Net based and correlation based methods. U-Net based methods [18, 27] have a relatively simple structure but are more difficult to explicitly compare input

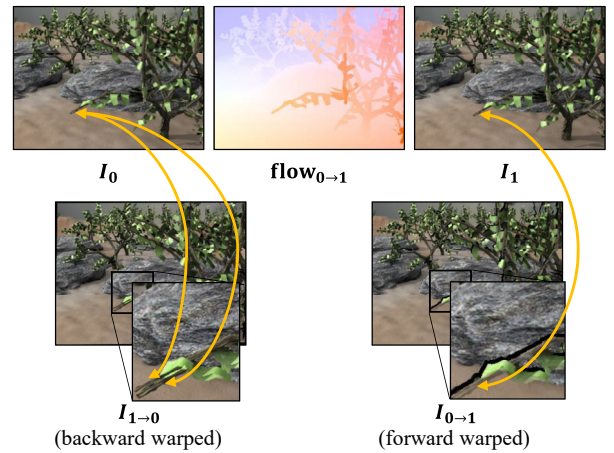


Figure 2: Backward and forward warping examples [1]. As shown in $I_{1 \rightarrow 0}$, backward warping creates duplication where the flow fail to map pixels to the shifted location such as occlusion region. The duplication makes difficulties to compare the similarity between the warped image and original image. On the other hand, forward warping doesn't cause confusion because it does not create unknown parts.

features than correlation based methods [21, 22, 24]. The correlation based one utilizes the feature correlation as an input of the flow estimation network to guide the warped

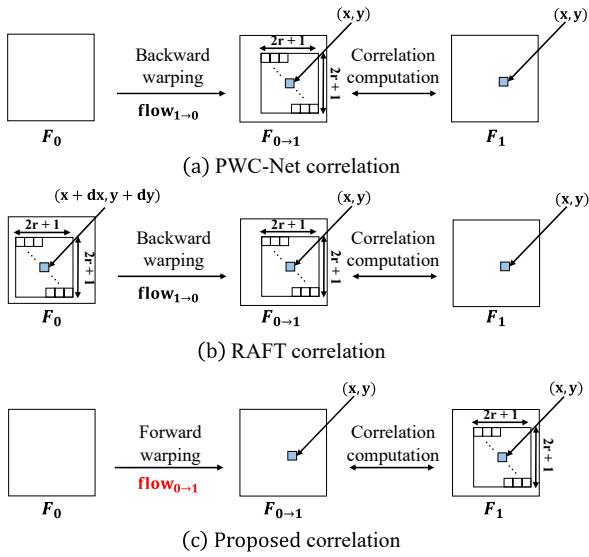


Figure 3: Comparison of correlation matching method between existing and the proposed one. Note that the local grid $(2r + 1) \times (2r + 1)$ in the features indicates the search range to find motion flows through correlation comparison. (a) PWC-Net [21] performs backward warping to compute $F_{0 \rightarrow 1}$ and correlation matching is followed. (b) RAFT [22] extracts the candidate from the corresponding location of (x, y) in the F_0 , candidates are warped using $F_{1 \rightarrow 0}(x, y)$ to align the spatial coordinate. (c) Our model performs forward warping to compute $F_{0 \rightarrow 1}$, and the correlation is derived with the candidates extracted from F_1 .

feature matches to the original feature. However, these algorithms are difficult to ensure that the output flow is corrected to a position with the highest correlation. Therefore, we propose the flow correction with the highest feature correlation based on novel feature matching algorithms in order to estimate bi-directional flows. To estimate more accurate flows, forward warping is performed to compute feature correlations. Inspired by softmax splatting approach [15], the learnable importance weights w_0^{l+1} and w_1^{l+1} for blending forward warped pixels are derived at each pyramid layer.

The bi-directional flow estimation network consists of an encoder, feature downsampling network, a flow update module, and a flow upscaling network.

Encoder and feature downsampling network Encoder network extracts the features F_0^l and F_1^l from each input images I_0 and I_1 . We denote the pyramid level as l . To generate the input feature for each pyramid layer, the strided convolution network is used to downsample the features by a factor of 2.

Flow update module The bi-directional flows are updated through each pyramid layer from the initial flows $\text{flow}_{0 \rightarrow 1}^1$ and $\text{flow}_{1 \rightarrow 0}^1$ which are set to zero. The module computes similarities between the input features as the correlation. The flows from the previous pyramid level are updated by the comparison of correlation values. To align the coordinates of two features for deriving correlation, forward warping is applied in contrast to the existing methods [21, 22] which use backward warping. Figure 2 illustrates backward warped image has duplicated region when it fails to map the corresponding pixels. These miss-aligned features make ambiguity in determining the optimal values of the correlation. Since the forward-warped feature does not include these miss-aligned features, it is possible to estimate more accurate flows. Figure 3 and 4 describe the correlation calculation and flow update process at a certain spatial location (x, y) . The proposed correlation matching method shown in figure 3(c) calculates feature correlation between forward warped feature $F_{0 \rightarrow 1}$ and the input feature F_1 . Since the location of warped feature $F_{0 \rightarrow 1}$ is aligned to the feature F_1 , the correlation is computed between the feature $F_{0 \rightarrow 1}$ at (x, y) and the feature F_1 at (x', y') , where $(x', y') \in \{x - r \leq x' \leq x + r, y - r \leq y' \leq y + r\}$ with search range r . The flows are corrected by selecting one having the highest correlation value from the candidates.

The details of the flow update process are shown in figure 4. The module calculates the similarity inside the search range by dot product according to the feature channel. This procedure is similar to derive the relevance between ‘query’ and ‘key’ in the attention mechanism [23]. Applying softmax operation enables the flow to be determined the highest correlation in a differential way. At each pyramid layer, the flow is gradually updated using the highest feature correlation within the limited search range. Since the updated flow is projected to the spatial coordination of F_1^l , it should be re-warped to the original coordination. We omit the re-warping process of the updated flow in figure 4 for compact expression. The optical flow of the opposite direction is also estimated by the same process to obtain bi-directional optical flows.

Flow upscaling network Flow upsampling is required to use the flows of the previous pyramid level estimated with lower resolution. Since simple bi-linear interpolation leads to degradation of flow details, we design the network to minimize the error that occurred in the upsampling process. The network takes inputs of the flow update module outputs, the importance weight outputs of the previous level, and the features of pyramid levels. The network consists of decoder network and two sub-networks to estimate the residuals and importance weights for the current pyramid level. The residuals contribute to improving the quality of the upsampled flows.

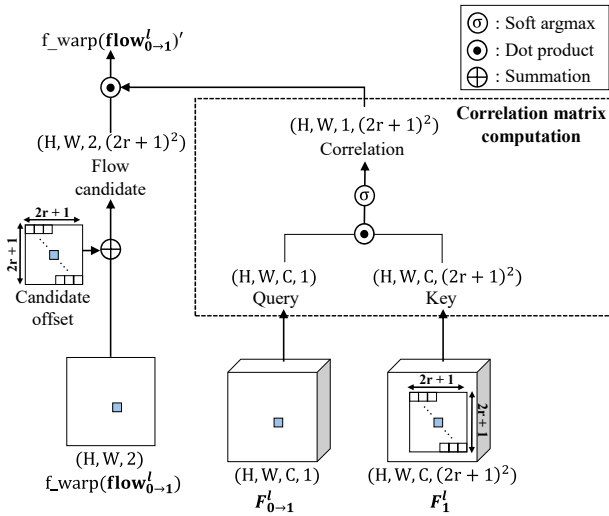


Figure 4: Flow update process. Correlations of candidates are calculated with dot product according to the feature channel. Based on the soft-argmax operation on the correlation matrix, it is possible to determine the candidate having the highest similarity. After that, flow is updated with the selected candidate.

3.2. Frame synthesis network

Intermediate frame is derived using two input frames and estimated bi-directional flows ($\mathbf{flow}_{0 \rightarrow 1}$ and $\mathbf{flow}_{1 \rightarrow 0}$). Green boxes of figure 1 describe frame synthesis network which consists of intermediate flow estimation network and frame synthesis layer. Frame synthesis network performs once at the original scale of input images, unlike flow estimation module in section 3.1 adopts pyramidal structure. We consider if the bi-directional flows are already estimated with enough receptive field which can be managed with a shared pyramidal structure, frame synthesis does not need to repeat with multi-scaled input images. We firstly calculate the intermediate flows $\mathbf{flow}_{t \rightarrow 0}$ and $\mathbf{flow}_{t \rightarrow 1}$ and then warping and blending are performed to interpolate the target frame I_t .

Feature extraction network Frame synthesis network uses contextual information to improve the details and to deal with occlusions. The contextual feature is generated using two kinds of features, one is extracted from the encoder of the flow estimation module, and the other is extracted from the *conv1* layer of the pre-trained VGG16 [19]. The final contextual feature presented in figure 1 is obtained through a convolution layer that has inputs of the aforementioned extracted features, two input images, and flows.

Intermediate flow estimation network We use backward warping to synthesis an intermediate frame to avoid hole generation. Since backward warping needs a flow that is starting from the temporal target position t to input images, we adopt the flow reversal approach [25] to estimate a flow aligned to coordinate at t . Since it is possible for flow reversal to generate holes in the reversed flow mostly around the occluded area due to forward warping, flow residuals are computed to compensate for each reversed flow. Additional refinement for the derived flows, such as the weighted average of corresponding flows, is not performed to maintain a characteristic of each flow.

$$\begin{aligned} \mathbf{flow}_{t \rightarrow 0} &= \text{rev}(t \cdot \mathbf{flow}_{0 \rightarrow 1}) + \Delta \mathbf{flow}_{t \rightarrow 0}, \\ \mathbf{flow}_{t \rightarrow 1} &= \text{rev}((1-t) \cdot \mathbf{flow}_{1 \rightarrow 0}) + \Delta \mathbf{flow}_{t \rightarrow 1}, \end{aligned} \quad (1)$$

where the function ‘rev’ denotes the flow reversal, and $\Delta \mathbf{flow}_{t \rightarrow 0}$ and $\Delta \mathbf{flow}_{t \rightarrow 1}$ are the flow residuals.

To estimate the flow residuals, we use the U-Net architecture composed of three layers of down-convolution and three layers of up-convolution. The network takes the bi-directional flows and the contextual feature as inputs, and generates residuals of intermediate flows and blending mask M .

Intermediate Frame Synthesis We simply perform frame synthesis with warping and blending. In the warping process, two input images are backward warped to time t using estimated intermediate flows. The warped images $I_{0 \rightarrow t}$ and $I_{1 \rightarrow t}$ usually contain artifacts especially the region which is unseen at the input images and appeared at the time t . To overcome the issue, we blend two warped images with the learnable blending mask M as follows.

$$I_t^{\text{blend}} = (M \odot I_{0 \rightarrow t}) + ((1-M) \odot I_{1 \rightarrow t}), \quad (2)$$

where \odot is the pixel-wise multiplication operation, and blending mask M denotes the confidence ratio of the warped image $I_{0 \rightarrow t}$. The confidence ratio has a value between 0 and 1.

3.3. Refinement network

The refinement network enhances the interpolated output given the blended intermediate frame, contextual feature, input images, and intermediate flows. Optical flow estimation can be failed in the occlusion, blurred, and brightness changed area. Most of the artifacts caused by flow estimation failure can be resolved by blending, but the artifacts may occur in the same area of $I_{0 \rightarrow t}$ and $I_{1 \rightarrow t}$. To solve this case, we design the network adding the dilated convolutional block to grow the receptive field inspired by image inpainting algorithm [8]. The network completes the occluded region and enforces the network to predict the residuals between the ground-truth and blended output.

3.4. Loss Functions

The proposed network illustrated in figure 4 is trained using the following two loss functions; One is to measure how close the blended intermediate frame I_t^{blend} to the ground truth frame I_t^{GT} , and the other is to compare the refined intermediate frame I_t^{refine} with the frame I_t^{GT} .

$$\begin{aligned} \mathcal{L}_{\text{blend}} &= \sum_x \|I_t^{\text{blend}}(x) - I_t^{GT}(x)\|_1, \\ \mathcal{L}_{\text{refine}} &= \sum_x \|I_t^{\text{refine}}(x) - I_t^{GT}(x)\|_1. \end{aligned} \quad (3)$$

The total-loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{blend}} + 2 \cdot \mathcal{L}_{\text{refine}}, \quad (4)$$

where the weight for each loss is empirically determined. All the networks are trained in an end-to-end manner.

3.5. Search range analysis in flow estimation

We can manage the range of receptive fields for flow estimation with the number of pyramid layers. At each pyramid level, the search range is determined with a radius r as mentioned in section 3.1. The totally covered search range is computed by

$$\begin{aligned} R_1 &= D_{\text{initial}} \cdot r, \\ R_L &= R_{L-1} \cdot D_L + r, \end{aligned} \quad (5)$$

where the R_L means the maximum range of motion coverage, D_L is a downsampling ratio according to the pyramid level in the encoder network, and L denotes the total number of pyramid levels.

The proposed model determines the variables to have enough total search range and to have an efficient number of parameters. Since we first reduce the image size as a quarter, the downsampling ratio is 2 except for the D_{initial} . The radius r is determined to 4. Therefore, the number of pyramid levels should be 6 to cover the 4K dataset, such as X4K1000FPS [18].

4. Experiments

We conducted experiments of video frame interpolation for four datasets: X4K1000FPS (4K) [18], VVC (Versatile Video Coding reference dataset, 4K), Vimeo90K (240p) [26], and UCF-101 (256 × 256) [20].

X4K1000FPS [18] The testset in X4K1000FPS contains 15 clips with a length of 33 consecutive frames and each frame is captured by 1000fps. We select 8 images with four frames interval in each sequence and conduct the 8 times interpolation to configure a similar test environment in the existing 4K VFI algorithm [18].

Method	X4K1000FPS		VVC		#P(M)
	PSNR	SSIM	PSNR	SSIM	
AdaCoF _o	24.35	0.7296	31.89	0.8909	21.8
AdaCoF _f	25.86	0.7691	32.67	0.8841	
CAIN _o	24.50	0.7458	32.43	0.8910	42.8
CAIN _f	27.36	0.8182	32.12	0.8958	
XVFI ₃	28.98	0.8573	33.17	0.8962	5.5
XVFI ₅	30.12	0.8707	33.08	0.8960	
Ours ₄ ^{x4k}	28.81	0.8430	33.15	0.8977	4.7
Ours ₅ ^{x4k}	30.47	0.8719	33.19	0.8979	
Ours ₆ ^{x4k}	30.51	0.8719	33.20	0.8979	

Table 1: Quantitative comparison on X4K1000FPS (4K) and VVC. Note that the numbers in red denote the best performance and numbers in blue mean the second best performance. Subscript o is official pre-trained weight provided from the author, subscript f denotes fine-tuned model with X4K1000FPS training dataset from the pre-trained weight, and subscripts of Ours and XVFI are the number of pyramid layers for the test. The number of parameters #P(M) is expressed in million units.

VVC VVC contains six 4K videos captured by 30 to 60fps. To configure a similar comparison environment with X4K1000FPS, we randomly selected 150 test triplets from VVC with the interval of the input images as 1/30s.

Vimeo90K and UCF-101 [20, 26] Both datasets are standard triplet (240p) sequences to measure the performance of video frame interpolation. To validate the performance in the low resolution and low frame-rate videos, we also evaluate for 3,782 triplets in Vimeo90K testset and randomly selected 350 triplets from UCF-101.

4.1. Implementation Details

We trained our model (Ours^{x4k}) for high resolution video using X4K1000FPS [18] dataset which contains 4,408 clips. Each clip is composed with a lengths of 65 consecutive frames and image resolution of 768 × 768 cropped from 4K size original image. During training, two input frames I_0 and I_1 , and one target intermediate frame I_t were grouped for a training sample. We designed the training data loader to randomly determine the temporal distance between two inputs as from 2 to 32, and the temporal position of the target frame within the temporal distance of inputs for arbitrary temporal frame interpolation. The 384 × 384 training image patches were randomly cropped from the original training data. We also performed flipping

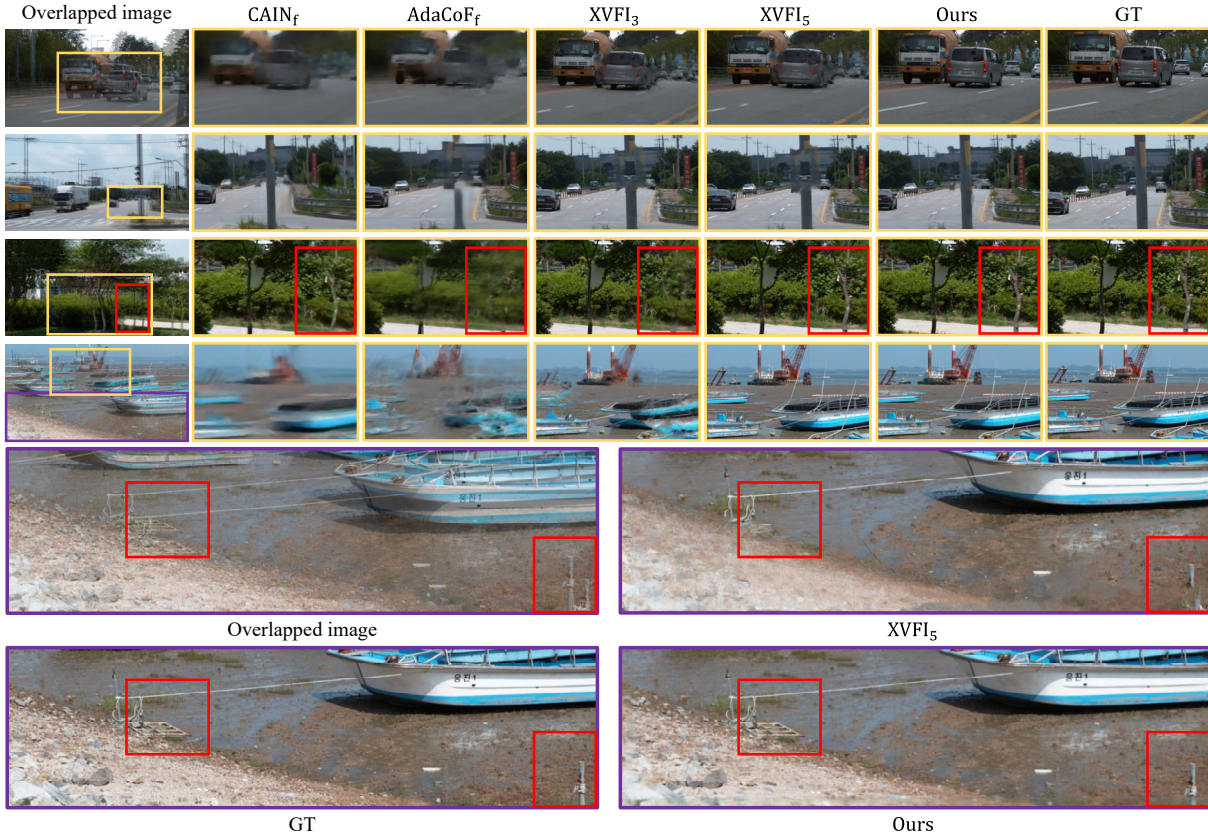


Figure 5: Visual comparisons of interpolation results from the X4K1000FPS dataset. Each result is zoomed from the yellow boxed part of the overlapped image. In particular, the purple boxed area is compared for the only two algorithms XVFI₅ and Ours. Prominent results are highlighted with red boxes.

images horizontally and vertically, and swapping the order of frames for data augmentation.

To evaluate the low resolution and low frame-rate dataset, we also trained our another model(Ours^{vimeo}) using Vimeo90K [26] dataset which contains 51,312 triplets of 256×448 video frames. We applied the random cropping as 256×256 , flipping, and swapping on the triplet for the augmentation. As shown in Equation 5, we determined the pyramid level L according to the investigated flow magnitude [18]. The L for Ours^{x4k} was set to 4, and for ours^{vimeo} was set to 1.

The weights of our model were initialized with the method of He initialization [7]. We adopted ADAM [10] optimizer with learning rate $1e - 4$, and reduced the learning rate factor of 4 at [100, 150, 180]-th epoch. We set the batch size as 14 for train Ours^{x4k}, and 50 for Ours^{vimeo}. Training takes about two days for Ours^{x4k}, and four days for Ours^{vimeo} to fully converge on NVIDIA Tesla V100 GPU.

Method	Vimeo90K		UCF-101		#P(M)
	PSNR	SSIM	PSNR	SSIM	
AdaCoF	34.27	0.9714	34.95	0.9504	21.8
CAIN	34.65	0.9730	34.83	0.9507	42.8
XVFI ^{vimeo}	35.07	0.9760	34.95	0.9505	5.5
Ours ^{vimeo}	34.95	0.9749	34.97	0.9506	4.7

Table 2: Quantitative comparison on Vimeo90K(240p) and UCF-101(256 × 256) with the number of parameters.

4.2. Quantitative and qualitative analysis

We conducted the quantitative evaluation for 4K video frame interpolation. For this experiment, we evaluate the methods on X4K1000FPS testset [18] and randomly selected 150 test triplets in VVC.

We compared our model (Ours^{x4k}) with previous video

frame interpolation models, CAIN [5], AdaCoF [11], and XVFI [18]. We evaluate both results of methods gathered by using pre-trained weights provided by the author and fine-tuned weight starting from the pre-trained model. Note that RRPN [27] is excluded in the experiment because the official code has not been published yet.

Table 1 compares the PSNR and the SSIM measures between our method and other ones. The results show that our model outperforms the other methods on both 4K datasets with the fewest # of parameters. Since the maximum flow magnitude in the X4K1000FPS testset is over 400, our method shows the best accuracy with the pyramid level 6 on the experiment. The qualitative results described in figure 5 for X4K1000FPS show that our interpolation results capture the objects which are failed in other ones.

We also evaluate our model(Ours^{vimeo}) compared to other previous methods on 3,782 Vimeo90K testset and randomly selected 350 test triplets from UCF-101.

In contrast to the 4K video frame interpolation, the test-set in Vimeo90K and UCF-101 contains similar statistics of optical flow magnitude compared to the training environment. Therefore, we evaluate our model with the pyramid level to 1 which is the same as the level in the training process.

For this experiment, we compare our method and other VFI methods pre-trained with vimeo90K dataset. Quantitative comparisons of the video frame interpolation results are shown in Table 2. These comparisons show our results are on par with the state-of-the-art method in [18], while showing much better results compared to the other video frame interpolation methods.

4.3. Ablation studies

Optical flow estimation The key contribution of our algorithm is accurate optical flow estimation for high-resolution with fast motion video. To validate the performance of our flow estimation module, We conduct the ablation study for the model using two existing flow estimation methods, PWC-Net [21] and RAFT [22]. Experiments are performed with two conditions for each algorithm. One is freezing the flow estimation layers while training the interpolation module, and the other is training the entire network from the pre-trained weights. The training is performed under the same conditions in X4K1000FPS dataset [18]. Table 3 shows the results of ablation studies. Since the existing methods are not targeting the large motion and are not familiar with the X4K1000FPS dataset, the fine-tuned model shows improved results compared to each frozen weight model. Although the results of fine-tuning have improved, the proposed forward warping based method has better results. From the results, accurate flow estimation and update modules with large receptive fields are necessary for the 4K VFI.

Method	PSNR	SSIM	#P(M)
PWC-Net _{fr} + synth	27.39	0.8090	11.7
PWC-Net _{fi} + synth	28.40	0.8183	
RAFT _{fr} + synth	29.17	0.8519	7.7
RAFT _{fi} + synth	29.67	0.8570	
Ours W.O.refine	30.02	0.8609	3.7
Ours _{full}	30.51	0.8719	4.7

Table 3: Ablation studies for optical flow estimation and refinement network. subscript ‘fr’ denotes freezed pre-trained weight for optical flow estimation network, subscript ‘fi’ is the model fine-tuned the network starting from the pre-trained weight in the training process, and ‘synth’ denotes ours frame synthesis network + refinement network.

Refinement module To verify the performance of the refinement module, we evaluate our algorithm with and without refinement module for the X4K1000FPS test dataset. The optical flow estimation can be failed in high-resolution video with fast object motion. Furthermore, our interpolation module is designed simply blending two warped frames than other kernel based methods [2, 3, 11, 16], which can cause missing regions in the blended output. We integrate the dilated convolution layer for the refinement module which is expected to fill the missing parts. As shown in Table 3, both PSNR and SSIM are noticeably improved with the refinement module.

5. Conclusion

We proposed ECMNet to support frame interpolation for high resolution with fast motion video. Our model proposed enhanced correlation matching based on bi-directional optical flow estimation from consecutive frames and adopts the advantage of shared weights at each pyramid layer which enables to adjust the number of pyramid levels according to the target video. Furthermore, the frame synthesis and refinement networks are designed to synthesize the intermediate frame efficiently. The proposed network is composed of the integration of convolution layer and differentiable operators which facilitates end-to-end training. Our experimental results support the robust performance of the proposed network qualitatively and quantitatively. Especially, for the 4K datasets that we targeted, ECMNet achieve outperformed results against the state-of-the-art algorithms with fewer parameters.

References

- [1] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and eval-

- uation methodology for optical flow. *International journal of computer vision*, 92(1):1–31, 2011. 3
- [2] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019. 2, 8
- [3] Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2, 8
- [4] Andrés Bruhn, Joachim Weickert, and Christoph Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International journal of computer vision*, 61(3):211–231, 2005. 2
- [5] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10663–10671, 2020. 2, 8
- [6] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world’s imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5515–5524, 2016. 1
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 7
- [8] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 5
- [9] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan 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*, pages 9000–9008, 2018. 1, 2
- [10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7
- [11] Hyeongmin Lee, Taeh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, and Sangyoung Lee. Adacof: Adaptive collaboration of flows for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5316–5325, 2020. 2, 8
- [12] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4463–4471, 2017. 2
- [13] Gucan Long, Laurent Kneip, Jose M Alvarez, Hongdong Li, Xiaohu Zhang, and Qifeng Yu. Learning image matching by simply watching video. In *European Conference on Computer Vision*, pages 434–450. Springer, 2016. 2
- [14] Simone Meyer, Oliver Wang, Henning Zimmer, Max Grosse, and Alexander Sorkine-Hornung. Phase-based frame interpolation for video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1410–1418, 2015. 1
- [15] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5437–5446, 2020. 2, 3, 4
- [16] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017. 2, 8
- [17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2
- [18] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. Xvfi: extreme video frame interpolation. *arXiv preprint arXiv:2103.16206*, 2021. 1, 2, 3, 6, 7, 8
- [19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5
- [20] Khurram Soomro, 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. 6
- [21] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan 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*, pages 8934–8943, 2018. 1, 2, 3, 4, 8
- [22] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 1, 2, 3, 4, 8
- [23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4
- [24] Jia Xu, René Ranftl, and Vladlen Koltun. Accurate optical flow via direct cost volume processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1289–1297, 2017. 3
- [25] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. *arXiv preprint arXiv:1911.00627*, 2019. 2, 5
- [26] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 6, 7
- [27] Haoxian Zhang, Yang Zhao, and Ronggang Wang. A flexible recurrent residual pyramid network for video frame interpolation. In *European Conference on Computer Vision*, pages 474–491. Springer, 2020. 1, 2, 3, 8