KIPS Trans. Softw. and Data Eng.
Vol.13, No.1 pp.17~27
https://doi.org/10.3745/KTSDE.2024.13.1.17
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

Kernel-Based Video Frame Interpolation Techniques Using Feature Map Differencing

Dong-Hyeok Seo† · Min-Seong Ko† · Seung-Hak Lee† · Jong-Hyuk Park††

ABSTRACT

Video frame interpolation is an important technique used in the field of video and media, as it increases the continuity of motion and enables smooth playback of videos. In the study of video frame interpolation using deep learning, Kernel Based Method captures local changes well, but has limitations in handling global changes. In this paper, we propose a new U-Net structure that applies feature map differentiation and two directions to focus on capturing major changes to generate intermediate frames more accurately while reducing the number of parameters. Experimental results show that the proposed structure outperforms the existing model by up to 0.3 in PSNR with about 61% fewer parameters on common datasets such as Vimeo, Middle-burry, and a new YouTube dataset. Code is available at https://github.com/Go-MinSeong/SF-AdaCoF.

Keywords : Deep Learning, Frame Differencing, Video Frame Interpolation, U-Net

특성맵 차분을 활용한 커널 기반 비디오 프레임 보간 기법

서 동 혁† · 고 민 성† · 이 승 학† · 박 종 혁††

요 약

비디오 프레임 보간(Video Frame Interpolation)은 움직임의 연속성을 증가시켜 영상을 부드럽게 재생할 수 있어 영상, 미디어 분야에서 사용되는 중요한 기술이다. 딥러닝 기반 비디오 프레임 보간 연구에서 널리 사용되는 방법 중 하나인 커널 기반 방법(Kernel Based Method)의 경우, 지역적인 변화를 잘 포착하지만 전체적인 변화를 처리하는 데 한계가 있었다. 이에 본 논문에서는 주요 변화 포착에 집중하기 위한 특성맵 차분, Two Direction을 적용한 새로운 U-Net 구조를 통해 파라미터 수를 줄이면서 중간 프레임을 보다 정확하게 생성하고자 한다. 실험 결과 제안한 구조가 기존보다 Vimeo, Middle-burry 등의 일반적인 데이터셋과 새로운 YouTube 데이터셋에서 기존 모델보다 약 61% 더 적은 파라미터로 PSNR 수치가 최대 0.3 우수한 성능을 달성하였다. 본 논문에서 사용한 코드는 https://github.com/Go-MinSeong/SF-AdaCoF에서 확인 가능하다.

키워드 : 딥러닝, 프레임 차분, 비디오 프레임 보간, U-Net

1. 서 론

비디오 프레임 보간(Video Frame Interpolation)은 두 개의 연속적인 프레임에서 기존에 존재하지 않은 중간 영역의 프레임을 만들어내는 방법으로, 컴퓨터 비전(Computer Vision) 영역에서 활발히 연구되고 있는 분야 중 하나이다[1]. 비디오 프레임 보간을 딥러닝으로 해결하게 되었을 때 지니게 되는 이점 중 하나는 고가의 장비와 전문적인 촬영 기술이 필요한 슬로우모션(Slow Motion)을 일반 사용자들도 저비용으로 손쉽게 만들어낼 수 있고, 정적인 사진 2장으로도 비디오처럼 만들 수 있다. 또한, 프레임을 보간할수록 프레임율(Frames Per Second)이 올라가 영상 자체가 부드러워지며 시청자가 비디오를 시청했을 때 시각적으로 향상된 품질의 비디오라고 느낄 수 있다.

비디오 프레임 보간에서의 딥러닝 모델은 크게 커널 기반 모델링(Kernel Based Modeling)과 흐름 기반 모델링(Flow Based Modeling)으로 2가지 방법론이 있지만 본 논문은 커널 기반 모델의 문제점을 보완하고자 한다. 커널 기반 모델은 흐름 기반 모델과 달리 입력 프레임의 픽셀값을 기반으로 중간 프레임을 생성하는데, 이때 커널을 사용하여 추출된 각 특성맵(Feature Map)을 합성에 이용한다. 이 방법은 종단간(End-To-End) 학습이 가능하고 흐름 기반 기법보다 학습 파라미터의 개수가 적고 입력 프레임의 픽셀들이 흐릿하더라도 보간 성능이 우수하다는 장점이 있다. 하지만 커널을 기반으로 변화를 유추하기 때문에 지정된 커널보다 변화가 크다면 완벽히 변화

† 비 회 원 : 국민대학교 AI빅데이터융합경영학과 학사과정
†† 종신회원 : 국민대학교 AI빅데이터융합경영학과 조교수
Manuscript Received : November 6, 2023
First Revision : December 12, 2023
Accepted : December 21, 2023
* Corresponding Author : Jong-Hyuk Park(jonghyuk@kookmin.ac.kr)

를 포착할 수 없다는 단점이 있다[2].

커널 기반 기법에서 대표적인 모델인 AdaCoF(Adaptive Collaboration of Flows for Video Frame Interpolation)[3]은 기존 커널 모델의 단점인 국한된 영역의 변화만 포착하는 것을 합성 모듈에서 Deformable Convolution[4]을 사용하여 일부 극복했다. 그러나, AdaCoF는 커널을 기반으로 변화를 유추하기 때문에 지정된 커널보다 변화가 크다면 여전히 특성 맵을 추출하는 과정에서 이미지의 전체적인 변화를 포착하는데 한계점이 있었다. 이에 본 논문에서는 AdaCoF의 U-Net[5] 구조에서 전체적인 변화를 포함한 특성 맵을 생성하고자 한다.

본 논문은 기존 AdaCoF 모델과 달리 각 프레임이 별도의 U-Net을 통과하여 해당 프레임에 집중하며 파라미터 개수도 감소시켰다. 또한, 커널에 국한되지 않고 전반적인 영역에서 변화를 추정하는 역할을 하는 특성맵 차분(Feature Map Differencing)을 사용해 전체적인 변화량을 잘 포착하도록 했다. 이와 더불어 풀링 레이어를 1×1 합성곱 레이어(Convolution Layer)로 바꾸어 추가적인 학습과 불필요한 정보 손실을 최대한 방지하고자 했다. 이를 기반으로 기존 AdaCoF의 파라미터 개수인 21.8M보다 약 3배 적은 8.5M의 파라미터 모델로 수정했다. 이때 Two Direction 구조에서 특성맵을 생성하는 커널이 각 프레임에서 동일한 특성을 뽑아낼 수 있도록 가중치를 재사용했다. 이러한 방식은 실험을 통해 PSNR 수치가 최대 0.3 더 우수하다는 것을 입증했다.

본 논문의 순서는 다음과 같다. 2장에서는 관련 연구를, 3장에서는 제안하는 모델에 대한 상세한 설명을 한다. 이후 4장에서 실험 결과 및 평가에 대해 언급하며 5장의 결론으로 마무리한다.

2. 관련 연구

2.1 비디오 프레임 보간

비디오 프레임 보간이란 영상에서 연속된 프레임 사이를 부드럽게 이어주는 새로운 프레임을 만드는 대표적인 컴퓨터 비전 과업 중 하나이다. Fig. 1은 커널 기반 비디오 프레임 기법과 흐름 기반 비디오 프레임 기법의 중간 프레임 생성 방식의 차이를 보여주며 자세한 내용은 다음과 같다.

커널 기반 비디오 프레임 보간 기법은 깊은 합성곱 신경망 층(Deep Convolutional Neural Network Layers)을 사용하여 프레임 보간을 수행한다. 흐름 추정 네트워크(Flow Estimation Network)를 사용하지 않기 때문에 종단간 학습이 가능하다. 다만, 두 개의 프레임을 통해서 중간 프레임을 추정하도록 훈련이 진행되기 때문에 입력 프레임 사이의 중간 시점만을 생성하게 된다. 그리고 커널 기반의 학습 방식은 지역적 공간 학습(Local Spatial Feature Learning)으로 인해서 커널의 크기에 따라 큰 변화(Large Motion)을 포착하는데 어려움이 존재한다. 이를 해결하기 위해 커널의 크기를 키워 큰 변화를 포착할 수 있으나 이는 모델의 연산량이 폭발적으로 증가될 수 있다[6]. 최근에는 렌더링(Rendering)된 비디오에서의 보간을 시도한 연구들도 있다[7].

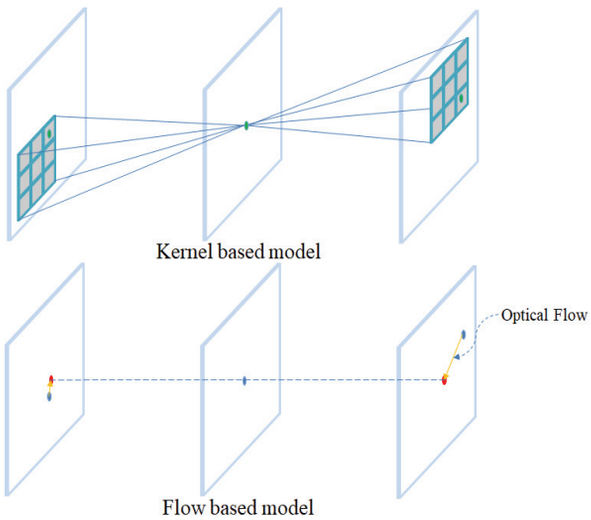

Fig. 1. Differences Between Kernel Based Model and Flow Based Model

흐름 기반 비디오 프레임 보간 기법은 비디오 프레임 보간에서 사용되는 가장 보편적이며 오래된 방식이다. 이전 프레임과 현재 프레임의 차이를 이용하고 픽셀 값과 주변 픽셀들의 관계를 통해 각 픽셀의 이동을 계산하여 변화를 구별해 내는 방법이다. 이는 일반적으로 광학적 흐름(Optical Flow)을 추정하는 흐름 추정 네트워크와 추정된 광학적 흐름을 기반으로 중간 프레임을 생성해 내는 합성 네트워크(Synthesis Network) 총 2가지로 구성되어 있다. 위와 같은 과정을 통해 흐름 추정 기반 비디오 프레임 보간 기법의 경우 각 픽셀마다 출력 픽셀에 대한 방향 벡터를 학습하게 된다.

광학적 흐름은 어떠한 인접한 두 장에서 나타나는 명암 변화를 고려하여 짧은 시간(물체 이동 거리를 몇 개 픽셀 정도로 작게 유지할 수 있을 정도의 시간)에 변화가 없다면 픽셀의 명암 값은 비슷할 것이라고 가정하고, 명암 값이 다르면 변화가 있다고 판단한다. 이러한 기법은 기존에 다양한 알고리즘 연산을 통해 해결되어 왔으나, 최근에는 합성곱 기반의 딥러닝 모델(RAFT, FlowFromer) 등을 이용하여 광학적 흐름을 연산하게 된다[8, 9].

흐름 추정 기반 비디오 프레임 보간 기법의 가장 큰 문제점은 이미지 보간에서 흐릿한 이미지를 생성하는 문제를 직면하고 있다. 더불어 2개의 딥러닝 구조를 사용하여 많은 연산량을 요구하기 때문에 종단간 학습은 대체적으로 불가능하다. 또한, 흐름 벡터(Flow Vector)에 영향을 크게 받아 흐름 추정 네트워크의 적절한 수행이 필요하다. 예를 들어, 조명의 갑작스러운 밝기 변화가 생겼을 때 흐름 추정 네트워크는 알맞은 흐름 벡터를 추정해 내지 못하고 이는 중간 프레임을 생성하

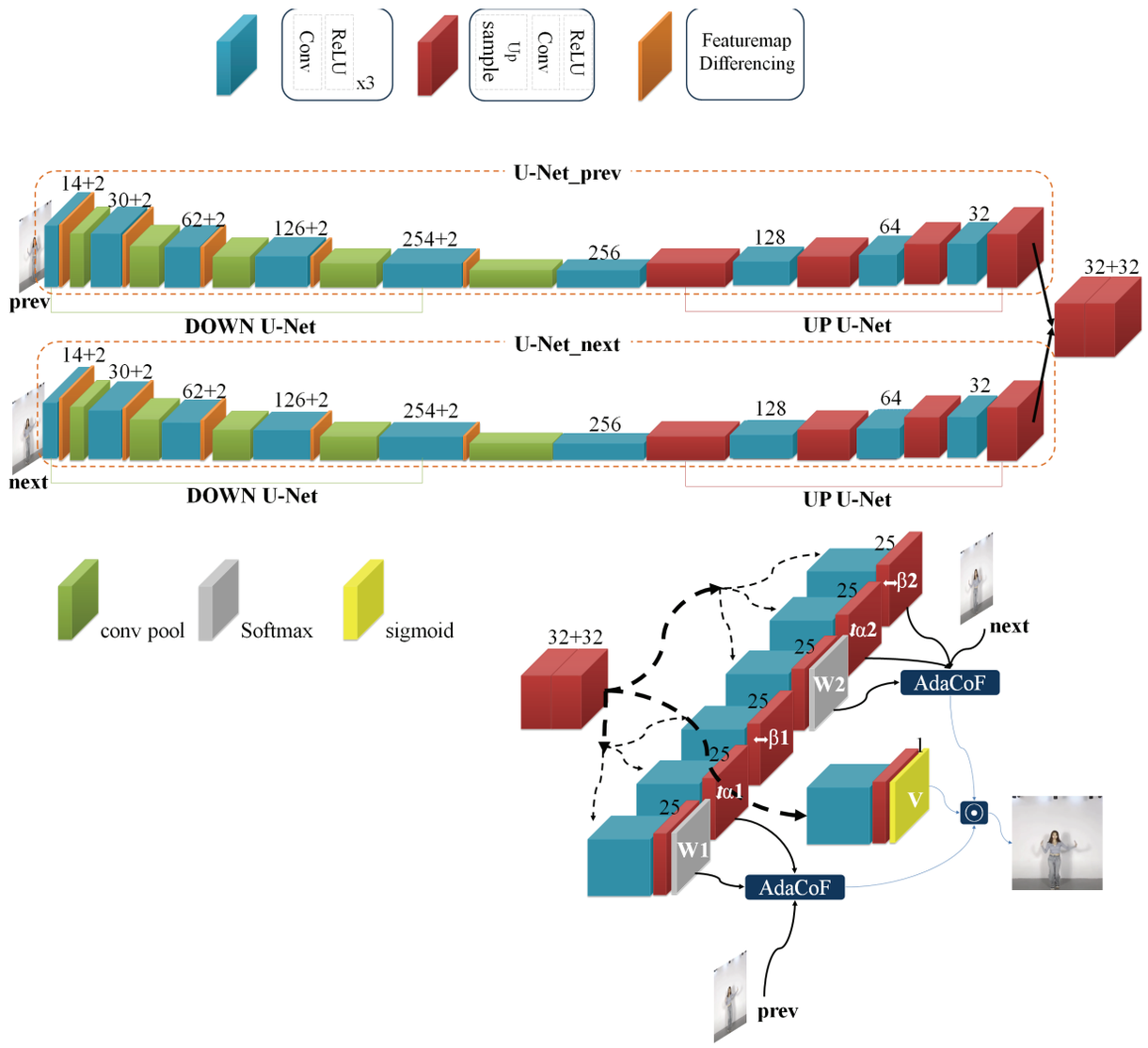

Fig. 2. Our SF-AdaCoF Model Architecture

는 과정에서 좋지 못한 결과를 유도하게 된다. 최근에는 이러한 문제를 해결하기 위한 연구들이 지속되고 있다[10].

2.2 AdaCoF

AdaCoF는 대표적인 커널 기반 모델 중 하나로 이전 프레임과 이후 프레임을 결합하여 특성맵 추출 네트워크(Feature Map Extraction Network)의 입력값을 구성한다.

특성맵 추출 네트워크에서 생성된 특성맵을 통해서 Equation (1)과 같은 총 7개의 특성맵을 생성해 낸다.

$$\alpha_1,\alpha_2,\beta_1,\beta_2,W_1,W_2,V \tag{1}$$

Equation (1)의 α와β는 이미지의 x, y축 방향으로의 오프셋(Offset) 값을 나타내는 특성맵이고, W는 가중치의 역할을 하는 특성맵을 의미하며 V는 폐색 영역(Occlusion)을 처리하기 위한 특성맵이다. 그리고 α_1,β_1,W_1은 이전 프레임, α_2,β_2,W_2는 이후 프레임과 Deformable Convolution을 기반으로 하는 합성 모듈에서 연산을 통해 학습이 이뤄진다.

Deformable Convolution은 기존 합성곱 연산과 달리 커널의 위치를 이미지에서 학습된 오프셋 정보를 통해 유동적으로 조정할 수 있다. 따라서 기존 합성곱 연산은 고정된 크기의 정보만을 얻게 되지만 합성 모듈에서 사용되는 Deformable Convolution은 커널 모양이 독립적으로 구성된다. 이는 커널 기반 기법의 단점인 커널 크기 제약을 일부 극복해 내며 보다 큰 동작 보간에 대한 좋은 성능을 얻게 된다. 이렇게 합성 모듈에서 이전과 이후 프레임에 대한 연산이 이루어진 이후, AdaCoF 모델은 만들어진 V 특성맵을 통해 이전 이후 특성맵 가중치 연산이 이루어져 최종적인 결과물을 도출해 낸다. 본 연구에서는 이러한 AdaCoF 모델을 기반으로 하는 새로운 구조를 제시하여 더 적은 파라미터 개수로 우수한 성능을 달성했다.

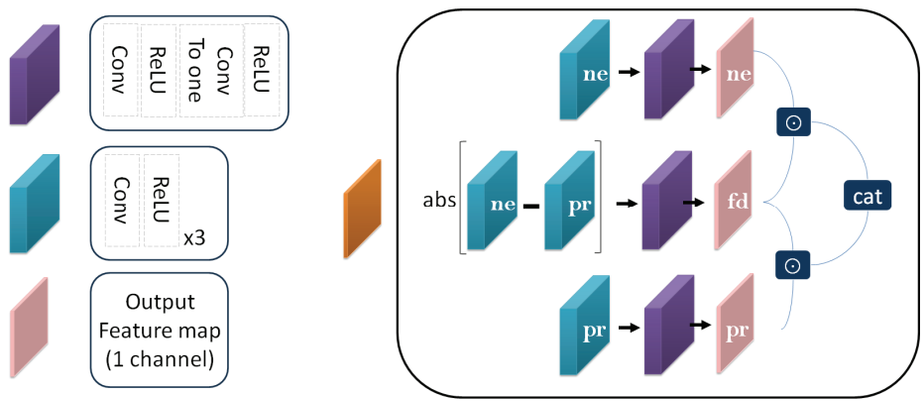

Fig. 3. Process of Feature Map Differencing

2.3 프레임 차분(Frame Differencing)

프레임 차분은 연속된 이미지 처리에서 사용되는 기법으로, 이후 이미지와 이전 이미지의 픽셀 값의 차이를 이용하는 기법을 의미한다. 움직임 추적, 객체 추적, 동작 인식 등에 사용할 수 있다[11]. 차이값이 0에 해당하는 픽셀의 경우 이전 이후의 이미지에서 변화가 없음을 뜻하며, 차이값이 0이 아니면 변화가 있음을 의미한다. 해당 기법은 간단한 연산으로 프레임 내 변화를 빠르게 탐지할 수 있으며 미세한 변화 또한 포착이 가능하다.

본 논문에서는 대표적인 커널 기반 모델인 AdaCoF에 프레임 차분에서 영감을 얻은 새로운 특성맵 차분 기법을 적용하여 제한된 커널 영역이 아닌 전반적인 변화를 파악할 수 있는 비디오 프레임 보간 모델을 제안하고자 한다.

3. 제안 모델

본 논문에서 제안한 구조는 Fig. 2와 같다. 기존 AdaCoF의 특성맵 추출 네트워크는 U-Net 구조이며, 이전 프레임과 이후 프레임을 결합하여 한번에 들어가는 구조로 구성되어 있다. 하지만 이러한 구조보다 각 프레임이 Two Direction으로 분리되어 서로 다른 U-Net을 통과하는 것이 해당 프레임에서의 특징을 보다 잘 추출해 낼 수 있다. 따라서 본 논문은 이전 프레임과 이후 프레임이 별도의 U-Net을 통과하는 Two Direction으로 구성하여 추후 병합하는 구조로 만들고, 특성맵 차분을 통해 커널 기반 모델에서 전체적인 변화를 포착할 수 있는 모델을 제안한다.

이때 해당 구조의 특성맵이 동일한 움직임의 변화를 파악한다는 점에서 SF(Simaese Feature Map)-AdaCoF라 칭한다. 구체적으로 기존 AdaCoF는 3채널(RGB)프레임 2개가 합쳐진 총 6채널의 입력으로 들어가며 하나의 U-Net만을 통과하게 된다. 그리고 각 Down U-Net 과정에서 Basic Conv Block(파란색 블록) 이후 최대 풀링 레이어(Maxpooling layer)를 지나게 된다.

하지만 SF-AdaCoF는 이전 시점의 프레임과 이후 시점의 프레임이 각각 U-Net_prev, U-Net_next라는 특성맵 추출 네트워크를 통과하여 특성맵을 추출하게 된다. 특히 서로 다른 U-Net 간의 상호 작용이 특성맵 차분을 통해 발생되며 이는 단순한 인코더(Encoder), 디코더(Decoder) 구조와 차이가 있다. Fig. 2의 Down U-Net 과정의 주황색 블록은 특성맵 차분 기법의 결과이며, Fig. 3은 특성맵 차분을 수행하는 과정을 의미한다. Fig. 3에서 각 시점의 특성맵간의 차이를 절댓값으로 추정한 결과와 각 시점의 특성맵을 보라색의 특성 압축 블록(Feature Map Compression Block)을 통해 1채널 특성맵을 얻게 된다. 특성맵 차분 압축 결과와 각 시점의 특성맵이 특성 압축 블록을 압축 결과를 아다마르 곱 연산을 진행하며, 이렇게 생성된 두 개의 특성맵은 다시 이전 이후 시점의 특성맵과 연결된다.

이후 U-Net_prev, U-Net_next에서 최종적으로 생성된 특성맵을 병합해 7개의 특수 목적의 특성맵을 생성하게 되고, 앞에서 언급한 Deformable Convolution 기반의 합성 모듈을 거친 뒤 V 특성 맵을 이용하여 최종적인 중간 프레임을 생성한다.

3.1 Two Direction

기존 커널 기반 모델은 입력으로 이전 시점과 이후 시점의 이미지를 붙여 한번에 넣는 구조를 주로 활용한다[3, 6, 12]. 하지만 이는 이전 시점과 이후 시점의 특징을 완전히 반영하지 못할 수 있다. 따라서 각 시점의 입력을 따로 받아 해당 프레임의 특징을 추출하고 추후 병합하는 구조를 만들고자 한다.

이전, 이후 시점의 프레임을 따로 입력값으로 활용하며 각각의 U-Net_prev, U-Net_next를 통해 특성맵을 추출함으로써 각 시점의 특징을 보다 잘 잡아낼 수 있다. 이때 이전과 이후 시점의 차이가 더욱 두드러지게 된다.

Two Directioin 구조는 2개의 U-Net으로 구성된다. 3채널(RGB)의 두 프레임을 합친 6채널의 입력에 대해 특성맵을 추

출하는 기존 U-Net 구조에서 이전, 이후 각각 3채널의 프레임을 입력받는 구조로 변경되었는데 이때 학습 파라미터 개수가 감소한다. 이는 기존 6채널 입력에 대해 특성맵을 추출할 때보다 적은 개수의 채널을 갖도록 특성맵을 추출하여 파라미터 개수를 약 50%로 줄일 수 있었다.

본 논문에서는 이전 이후 프레임에 대해서 각 U-Net_prev와 U-Net_next의 특성맵 압축 블록을 제외한 나머지 합성곱 연산 시 가중치를 재사용하여 진행한다. 가중치를 재사용시킴으로써 얻을 수 있는 이점은 다음과 같다.

모델의 순전파 과정에서 새로운 파라미터를 추가하는 것이 아닌 파라미터의 재사용으로 인해 파라미터 개수의 감소 효과를 얻을 수 있으며 이는 모델의 메모리 사용량 감소로 이어진다. 또한, 파라미터는 한 번의 역전파 과정에서 2번의 업데이트를 하게 되며 이는 비선형성의 증진을 유도한다[13]. 특성맵 차분을 하는 이유는 이전과 이후 프레임간의 관계를 고려하기 위함인데, 차분을 통해 관계를 확인하기 위해서는 이전과 이후에서 동일한 조건으로 만들어진 특성맵을 사용해야 한다. 이를 위해 가중치 재사용을 사용하여 이전과 이후 프레임에 대해 각각 동일한 의미를 가지는 특성맵을 생성한다. 이러한 가중치 재사용은 파라미터 개수를 더욱 감소시켜 최종적으로 기존보다 61% 감소시켰다.

3.2 특성맵 차분

본 논문이 제안한 Two Direction 방식은 이전, 이후 프레임이 서로 다른 U-Net을 거치기 때문에 마지막 레이어(Layer) 이전에는 두 프레임 간 관계를 학습하기 위한 단계가 존재하지 않는다. 이를 해결하기 위해 특성맵 차분 기법을 도입한다.

특성맵 차분 기법은 제안 구조에서 두 프레임의 전체적인 변화에 대한 포착 및 이전, 이후 프레임 간의 상호작용을 모델이 학습하게 하는 방법이다. 기존 커널 기반 비디오 프레임 보간 모델은 프레임간의 변화를 파악하기 위해 커널을 사용하였는데, 이를 사용하게 되면 지역적인 정보를 잘 파악하지만 한정된 수용 영역(Receptive Field)만을 통해 파라미터를 학습하기에 수용영역에 포함된 변화만을 추정할 수 있다[1, 2].

따라서 한계를 극복하고자 제안 연구에서는 이전 시점과 이후 시점의 프레임이 Two Direction 방식으로 입력된 특성맵 추출 네트워크에 특성맵 차분 기법을 추가하여 사용하였으며, 이를 통해 특성맵 간의 전체적인 변화를 탐지하고자 했다. 이후 특성맵 차분 결과에 합성곱 연산을 활용하여 변화가 발생한 지역을 잘 탐지할 수 있도록 하였고, 커널만으로 확인할 수 없는 영역의 변화도 모델이 잘 포착할 수 있도록 했다.

각 시점의 특성맵과 특성맵 차분 결과에서는 1채널의 특성맵을 추출할 수 있도록 특성 압축 블록을 구성하고 반복하는데, 특성 압축 블록의 구성은 Fig. 4와 같다.

Fig. 4에서 특성 압축 블록의 $Conv(K\times K)^n$에서 $K\times K$ 합성곱 연산을 의미하며, n은 출력 채널의 개수를 의미한다. 이전 시점과 이후 시점의 특성맵 차분 결과는 특성 압축 블록

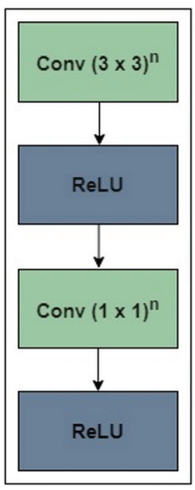

Fig. 4. Feature Map Compression Block

을 지나 1채널의 출력을 얻게 된다. 이때 차분된 특성맵의 특성 압축 블록 출력과 각 시점의 특성 압축 블록 출력 간의 아다마르(Hadamard)곱을 통해 최종 2가지의 결과물을 얻게 된다. 이후 이를 병합하여 각 시점의 기존 특성맵에 채널 기준으로 병합하여 다음 특징 추출을 진행하게 된다. 해당 과정은 Fig. 3에서 확인할 수 있다.

위 과정을 통해 앞에서 언급되었던 문제인 이전 시점과 이후 시점의 상호작용이 존재하지 않는 문제를 해결한다. 최종적으로 각 단계의 특성맵에는 이전, 이후 시점에 대한 정보와 두 시점의 변화 모두 포함하게 되어 프레임 간의 상호작용을 모델이 학습할 수 있다.

4. 실험 결과 및 평가

본 논문에서는 제안한 SF-AdaCoF의 우수함을 증명하기 위해 여러 데이터셋을 통해 다음과 같은 실험을 진행했다. 먼저 학습은 Vimeo-90k[14] 데이터셋을 활용하여 진행되었으며, 학습 모델의 평가를 위해 Vimeo-90k, Middle-burry[15], Davis[16], UCF101[17], K-POP YouTube 데이터에 대하여 실험을 진행했다. 또한, 평가지표의 평균값 이외에도 통계적 유의성을 확인하기 위해 각 데이터셋의 실험마다 표준편차를 구해 제시했다.

4.1 데이터셋

Vimeo-90k는 비디오 프레임 보간 과업에서 가장 보편적으로 학습 및 평가가 이루어지는 데이터셋으로 다양한 환경과 행동의 영상으로 구성되어 있다. 이는 (448, 256) 해상도의 15,000개의 영상으로 구성되었으며 총 73,171개의 Triplet 데이터셋으로 이루어져 있다. Triplet 데이터셋이란 이전, 중간, 이후 시점의 총 3개의 프레임이 하나의 세트로 묶인 것이며 이 중 중간 프레임을 정답값으로 활용하는 형식을 의미한다.

Middle-burry는 스테레오(Stero) 비전 연구에 사용하는 고

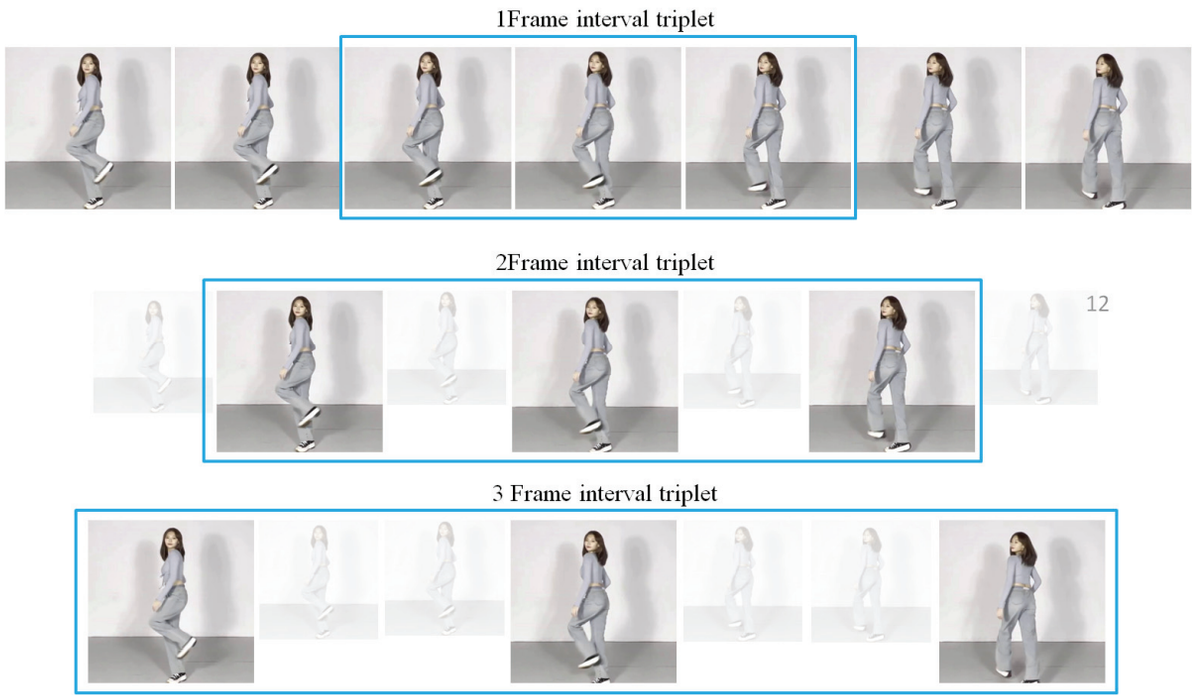

Fig. 5. Process of Build YouTube Triplet Dataset using Differences in Interval

정된 시점의 실내 장면 고해상도 데이터셋으로 다양한 환경 및 조명 조건에서의 고품질 이미지를 제공한다. 해당 데이터셋은 Vimeo-90k와 마찬가지로 비디오 프레임 보간 과업에서 보편적으로 사용되는 평가 데이터셋 중 하나이다. Davis는 객체 추적(Object Tracking) 및 세그멘테이션(Segmentation) 알고리즘을 평가하기 위한 고품질 비디오 데이터셋 중 하나이다. 특히 객체 간의 중첩이나 배경 변화 등의 다양한 조건이 있어 보간이 어려운 데이터셋임과 더불어 사람들이 상호작용하는 영상의 데이터로 이루어져 있어 일반화 능력을 평가하는데 유용하게 사용된다. UCF-101 데이터셋은 약 13,000개의 비디오 클립으로 구성되며 101가지 범주와 신체 동작, 사람 간의 상호작용, 물체와 인간의 상호작용, 예체능 등 5가지 유형으로 구성되어 있다. YouTube에서 수집된 데이터로 (320, 240)의 해상도를 지닌 비디오로 프레임은 25 FPS로 구성되어 있으며, 이는 주로 액션 인식(Action Recognition)에서 사용된다.

K-POP YouTube 데이터셋은 큰 움직임에 대한 이미지 보간을 실험하기 위해 YouTube에서 비교적 움직임이 큰 K-POP 안무 영상 10개를 선택하여 직접 데이터셋을 구축했다. 제작 방식은 다음과 같다. Pytube 라이브러리를 사용하여 유튜브에서 영상을 다운로드한다. 이후 Fig. 5에서의 첫 번째 행에서 볼 수 있듯이 유튜브 영상을 1 프레임 단위로 잘라 데이터셋을 제작했으며 파란색 박스는 하나의 Triplet 데이터셋을 의미한다. 이후 2 프레임, 3 프레임으로 프레임 간 간격을 넓혀 변화량의 정도가 각기 다른 3가지 유형의 데이터셋을 구축하였다. 이를 통해 같은 영상에서도 다양한 난이도에 대해서 평가하여 모델의 보편적인 성능을 확인하고자 했다. 최종적으로 K-POP YouTube 데이터셋은 총 125,220개의 Triplet 데이터셋을 가진다.

4.2 학습 방법 및 평가지표

제안 모델을 학습시킨 방법과 사용된 평가지표는 다음과 같다.

1) 학습 방법

제안 모델은 이미지의 해상도를(256, 256)으로 맞춰 사이즈를 재조정 한 뒤 Fig. 2의 구조를 통해 학습된다. Optimizer는 ADAMAX[18]를 사용하며 학습률은 0.001을 사용하며 매 20 에폭마다 절반으로 감소시키며 총 50 에폭을 학습시킨다. 자세한 내용은 github의 코드에서 확인할 수 있다.

2) 평가지표

피크 신호 대 잡음비(Peak Signal-to-noise ratio, PSNR)과 구조적 유사도(Structural Similarity Index Measure, SSIM)는 비디오 프레임 보간 과업에서 가장 많이 사용하는 평가 지표이다. 따라서 본 논문에서는 이 두 가지 지표를 이용하여 SF-AdaCoF에 대한 평가를 진행한다.

a) PSNR

$$PSNR = 10\log_{10}\left(\frac{R^2}{MSE}\right) \tag{2}$$

$$MSE = \frac{\sum_{M,N}[I_1(m,n)^2 - I_2(m,n)^2]^2}{M \times N} \quad (3)$$

Equation (2)에 나타나 있는 PSNR은 Equation (3)을 통해 계산되는 평균 제곱 오차(Mean Squared Error, MSE)를 활용하여 계산되는, 왜곡 강도 대 최대 픽셀 명암비를 나타낸다. 구체적으로, Equation (3)에서 I_1는 모델이 생성한 중간 프레임을 I_2는 정답값인 실제 중간 프레임을 의미하며 m,n은 각 프레임에서의 픽셀의 위치값을 나타낸다. 즉, 두 프레임 간의 동일한 위치에서의 픽셀값 간의 차이를 계산하는 의미이다. 또한, Equation (2)에서 R은 픽셀의 최댓값을 의미한다.

b) SSIM

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (4)$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (5)$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (6)$$

$$SSIM(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma} \quad (7)$$

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (8)$$

SSIM 평가지표는 국소 영상 구조, 휘도, 대비를 하나의 국소 품질 점수로 통합한 것으로 영상을 구성하는 주요 요소를 활용해 두 이미지 간의 유사도를 측정하기 위해 사용된다. 이 평가지표에서 구조는 픽셀 명암 패턴, 특히 휘도 및 대비를 위해 정규화한 이후에 이웃 픽셀에 나타나는 픽셀 명암 패턴을 의미한다[19].

Equation (4)의 $l(x,y)$은 휘도를 나타내며 이때, x,y는 각 각 모델이 생성한 중간 프레임과 정답값인 실제 중간 프레임을 의미한다. 또한 μ_x,μ_y는 각 프레임 내 픽셀의 평균을 의미한다. Equation (5)의 $c(x,y)$ 대비를 나타내며 σ_x,σ_y 는 각 프레임 내 픽셀의 표준편차를 의미한다. Equation (6)의 $s(x,y)$ 구조를 나타내며 σ_{xy}는 두 프레임의 공분산을 의미하고 각 식에서 사용된 C_1, C_2, C_3는 상수이다.

최종적으로 Equation (7) SSIM은 Equation (4), (5), (6)에 따라 계산된 휘도, 대비, 구조를 곱한 값으로 구할 수 있다.

4.3 손실함수

제안 모델은 아래의 2가지 손실함수를 조합하여 학습에 사용한다.

1) L1 Loss

정답 값인 중간 프레임 이미지와 합성한 이미지 간의 픽셀별 L1 Loss를 산정한다. 이를 통해 만들어진 이미지가 정답 이미지와 얼마나 차이가 나는가에 대해 파악한다. L2 Loss는 생성되는 중간 프레임을 흐릿하게 만드는 경향이 있기 때문에 L1 Loss를 사용한다[2].

2) Perceptual Loss

Perceptual Loss[20]은 VGG 네트워크의 중간 레이어에서 만들어지는 특성맵을 L2 Loss를 이용해 계산한 손실이다. 이는 현실적인 이미지를 만드는데 기여할 수 있다.

4.4 실험 결과

1) Vimeo, Middle-burry, Davis, UCF101 데이터셋

Table 1은 각 모듈 여부에 대한 실험으로 W/RW는 가중치 재사용을, CP는 풀링 레이어를 합성곱 레이어 구조로 변경한 것, TD는 Two Direction, FD는 특성맵 차분을 의미하며 괄호 안의 값들은 각 평가지표의 표준편차를 의미한다. 최종적인 제안 SF-AdaCoF는 기존보다 약 61% 적은 파라미터 개수를

Table 1. Vimeo, MiddleBurry, Davis, UCF-101 Dataset Experimental Results
W: with , RW : Reuse Weights, CP : Fully Convolutional, TD : Two Direction, FD: Feature map Differencing

Dataset \ Model	Metric	AdaCoF	SF-AdaCoF (W / RW, TD)	SF-AdaCoF (W/ RW, TD, CP)	SF-AdaCoF (W / RW, TD, CP, FD)
Vimeo	PSNR	34.241 (4.776)	33.233 (4.898)	33.289 (4.923)	**34.563 (4.671)**
	SSIM	0.955 (0.041)	0.943 (0.055)	0.944 (0.055)	**0.959 (0.034)**
Middle-burry	PSNR	35.452 **(3.900)**	34.176 (4.298)	34.411 (4.331)	**35.753** (4.084)
	SSIM	0.956 (0.022)	0.944 (0.029)	0.947 (0.025)	**0.959 (0.020)**
Davis	PSNR	26.548 (6.087)	25.343 **(5.823)**	25.244 (5.872)	**26.795** (5.949)
	SSIM	0.802 (0.178)	0.777 (0.181)	0.778 (0.184)	**0.814 (0.170)**
UCF-101	PSNR	35.162 (7.351)	34.892 (7.356)	34.920 (7.369)	**35.198 (7.338)**
	SSIM	0.950 (0.078)	0.947 (0.079)	0.947 (0.079)	**0.950 (0.077)**
Parameters	-	21.8M	5.9M	6.3M	8.5M

갖는다. 또한, W/RW, TD 모델의 경우 무려 5.9M의 적은 파라미터를 가짐에도 불구하고 기존 AdaCoF에 비해 적은 성능 감소를 보였다. 이를 통해 Two Direction 방법이 효과적으로 연산량을 줄여낼 수 있는 방법임을 확인할 수 있다.

Two Direction과 더불어 풀링 레이어를 합성곱 레이어로 변경함으로써 더 성능을 향상시켰다. 이는 정보 손실이 큰 풀링 레이어 대신 학습 가능한 합성곱 레이어를 사용함으로써 정보 손실의 감소가 일어났기 때문이다. 또한, 특성맵 차분을 통해 프레임간의 상호작용을 유도하였고 기존 AdaCoF에 비해 PSNR 수치가 최대 0.3 더 우수한 성능과 더불어 낮은 표준편차로 통계적 유의성을 보였다. 이와 같은 실험 결과를 통해 제안 모델의 실효성을 입증하였다.

2) 가중치 재사용 효과 실험

Table 2는 Two Driection 시 가중치 공유 여부에 따라 여러 데이터셋에 대하여 평가를 진행한 결과를 보여주는 표이다. Table 2의 W.O/RW은 가중치 재사용을 하지 않은 경우를, W/RW는 가중치 재사용을 한 것을 의미한다.

먼저 가중치 재사용이 성능 향상에 도움이 되는가에 대하여 살펴본다. Table 2의 4번째 열을 보고 확인할 수 있듯이 SF-AdaCoF에 대하여 가중치 재사용을 하는 경우 파라미터의 개수가 감소되지만 가중치 재사용하지 않았을 때 대비 모든 데이터셋에 대해 성능이 우수하다.

성능이 높게 나온 이유를 탐구하기 위해 Gradient Class Activation Map(GradCAM) 기법을 활용해 모델이 실제로 변화가 발생한 영역에 대해 얼마나 집중하는지 파악해 보았다. Fig. 6은 U-Net_prev과 U-Net_next에서 Down U-Net의 Upsample 이전 마지막 Basic Convolution Block에서의 GradCAM의 결과이다. 모델이 특정 영역에 집중할수록 붉은색을 띠며 집중하지 않을수록 파란색을 띤다. 해당 예시 프레임에서는 변화가 발생한 지역이 공이 있는 부분과, 손 모양 부분으로 실험 결과 가중치를 재사용 한 모델은 변화가 발생한 부분에 빨간색으로 잘 집중하여 포착하고 있는 것을 볼 수 있다. 또한, 동일한 영역을 포착하는 모습을 볼 수 있는데 이는 역전파 과정에서 재사용된 가중치가 지속적으로 학습되기 때문이다. 반면 가중치를 재사용하지 않은 모델은 변화가 발생한 부분에 집중하지 못하는 모습을 보였으며, 각 프레임이 서로 다른 가중치로 학습되기 때문에 포착하는 영역이 다른 것을 확인할 수 있다. 이러한 결과는 가중치 재사용이 성능 향상에 기여하는 것을 보여준다.

Table 2. Weights Reusing Experimental Results
W.O : With out , W: with , RW : Reuse Weights

Dataset \ Model	Metric	SF-AdaCoF (W.O / RW)	SF-AdaCoF (W / RW)
Vimeo	PSNR	34.017 (4.761)	**34.563 (4.671)**
	SSIM	0.954 (0.041)	**0.959 (0.034)**
Middle-burry	PSNR	34.877 (4.188)	**35.753 (4.084)**
	SSIM	0.952 (0.022)	**0.959 (0.020)**
Davis	PSNR	26.210 **(5.822)**	**26.795** (5.949)
	SSIM	0.802 (0.175)	**0.814 (0.170)**
UCF-101	PSNR	34.975 (7.338)	**35.198 (7.338)**
	SSIM	0.949 (0.077)	**0.950 (0.077)**
Parameters	-	14.1M	8.5M

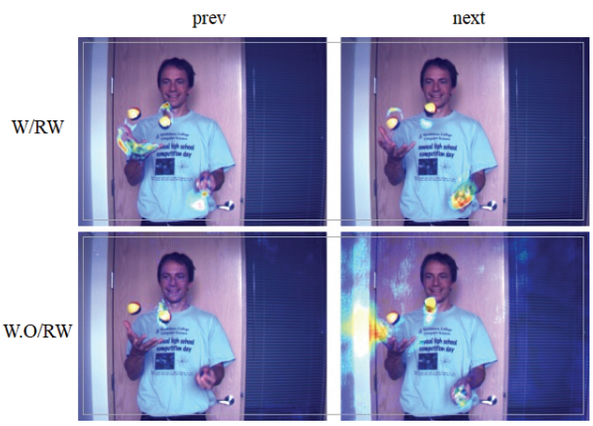

Fig. 6. GradCAM Results to See the Effect of Reusing Weights. Top: Reuse Weights, Bottom: Not Reuse Weights

3) Davis 데이터셋에 대한 평가

Davis 데이터셋은 다른 데이터셋에 비해 배경 변화, 객체 간 중첩으로 인해 보간이 어려운 데이터셋으로 본 논문에서는 이를 활용해 각 기법 간 결과물을 비교한다. 또한, 제시한 최종 SF-AdaCoF 모델을 다른 모델들의 결과물과 비교하며 SF-AdaCoF의 우수성을 입증한다.

해당 데이터에서의 실험 결과는 Fig. 7에서 제시한다. Fig. 7에서 가장 왼쪽에 있는 이미지는 정답 이미지이며 가장 우측 열에 있는 이미지가 제안된 최종 SF-AdaCoF의 결과물이다. 전반적으로 최종 SF-AdaCoF의 결과물이 다른 기법들에 비해 뚜렷한 품질을 보이는 것을 확인할 수 있다. 특히, 자동차와 관중 부분에서 기존 AdaCoF는 객체 부분이 겹쳐진 것처럼 흐릿하게 보간 되었지만 제안 모델의 결과물은 객체의 경계 부분이 뚜렷함을 확인할 수 있다.

4) YouTube 데이터셋(큰 움직임)에 대한 평가

비디오 프레임 보간 과업에서 변화가 큰 상황에서의 보간은 어려운 과제 중 하나이다. 또한, 많은 모델이 이와 같은 상황에 성능이 좋지 못한 경우가 대다수이다. 이에 따라 변화량이 큰 상황에서도 SF-AdaCoF가 강건한지 확인하기 위해 새로운 데이터셋으로 평가를 진행한다.

YouTube 데이터셋은 동작의 변화가 큰 데이터인 K-POP 안무 영상을 모아 구축한 데이터셋이다. 평가 데이터로 각각 1,000개의 Triplet을 사용하였고 실험 결과는 Table 3에서 확인할 수 있다. 각각의 Interval은 데이터 구축에 있어 연속적인

Fig. 7. Results of Davis Dataset from Top to Bottom, Davis Data : First Row : Juggle, Second Row : Burnout / The images show, from Left, Ground Truth, AdaCoF, SF-AdaCoF (W/RW, TD), SF-AdaCoF (W/RW, TD, CP), SF-AdaCoF (W.O/RW W/TD, CP, FD), and SF-AdaCoF (W//RW, TD, CP, FD)

Table 3. YouTube Dataset Experimental Results

W : with , RW : Reuse Weights, CP : Fully Convolutional, TD : Two Direction, FD : Feature map Differencing

Model / Dataset	Metric	AdaCoF	SF-AdaCoF (W / RW, TD)	SF-AdaCoF (W/ RW, TD, CP)	SF-AdaCoF (W / RW, TD, CP, FD)
YouTube Interval 1	PSNR	39.796 (4.648)	39.387 (4.748)	39.390 (4.767)	**39.848 (4.640)**
	SSIM	0.990 (0.008)	0.989 (0.008)	0.989 (0.008)	**0.990 (0.007)**
YouTube Interval 2	PSNR	34.476 (4.740)	34.195 (4.772)	34.165 (4.800)	**34.519 (4.734)**
	SSIM	0.979 (0.018)	0.978 (0.018)	0.978 (0.018)	**0.980 (0.018)**
YouTube Interval 3	PSNR	31.713 (4.453)	31.542 **(4.428)**	31.505 (4.450)	**31.757** (4.450)
	SSIM	0.971 (0.019)	0.970 (0.019)	0.970 (0.018)	**0.971 (0.018)**
Parameters	-	21.8M	5.9M	6.3M	8.5M

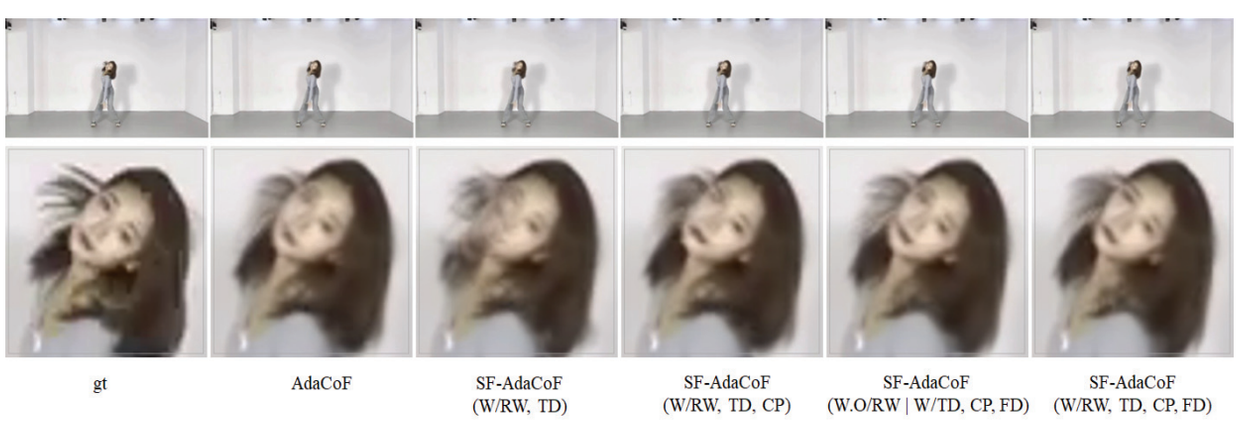

Fig. 8. Results of YouTube Dataset Experiment from Left to Right, The Images Show, from Left, Ground Truth, AdaCoF, SF-AdaCoF (W/RW, TD), SF-AdaCoF (W/RW, TD, CP), SF-AdaCoF (W.O/RW|W/TD, CP, FD) , and SF-AdaCoF (W//RW, TD, CP, FD)

데이터 간 간격을 얼마나 두었는지에 대한 것이다. 또한, 데이터는 안무 영상 특성상 Interval 1인 경우에도 다른 데이터셋에 비해 연속된 프레임 간 변화량이 크다. 이러한 YouTube 데이터셋을 활용하여 SF-AdaCoF 모델이 변화량이 큰 데이터에서도 성능이 우수함을 입증했다. Fig. 8은 YouTube Dataset Interval 2의 데이터셋 중 하나로 가장 왼쪽에 있는 이미지는 정답 이미지이며 가장 우측열에 있는 이미지가 제안된 최종 SF-AdaCoF의 결과물이다. Fig. 8에서 볼 수 있듯이 제안하는 최종 모델이 다른 모델에 비해 움직임이 큰 영역인 사람의 얼굴 부분을 더 정확하게 보간하는 것을 확인할 수 있다.

5. 결 론

비디오 프레임 보간 과업은 움직임의 연속성을 증가시켜 영상을 부드럽게 재생할 수 있어 영상, 미디어 분야에서 사용되는 중요한 기술이다.

본 논문에서는 커널 기반 모델 중 하나인 AdaCoF 모델을 기반으로 특성맵 차분 기법을 적용해 특성맵 간의 전체적인 변화를 포착할 수 있도록 하였다. 뿐만 아니라 Two Direction 과 가중치 재사용 방법론으로 모델의 파라미터를 23.1M에서 8.5M으로 약 61% 대폭 감소시키면서도 성능을 향상시켰다.

SF-AdaCoF는 Vimeo, Middle-burry 등의 다양한 데이터셋을 통한 실험으로 검증되었고, 모든 데이터셋에서 줄어든 파라미터에 대비해서 우수한 성능을 보였다. 또한, 동작의 변화가 큰 YouTube 데이터셋에 대해서도 다른 모델들에 비해 상대적으로 뚜렷한 결과물을 볼 수 있다. 결론적으로 이와 같은 실험 결과로 제안 모델의 우수성을 입증할 수 있었다.

본 논문에서는 변화량이 큰 상황에서의 한계를 보완하기 위한 방법에 대해 연구했고, 성능이 향상되긴 했으나 여전히 완벽한 결과물을 생성해내지는 못한다. 실험 결과 특성맵 차분을 사용해 기존보다 높은 품질의 결과물을 보였지만 여전히 변화량이 큰 객체 부분에 대해서 흐릿한 경우가 보였다. 따라서 사용된 특성맵 차분의 방법론을 고도화시키거나, 변화량이 큰 객체 부분에 집중해 보간할 수 있는 방법론에 대한 연구가 필요하다. 이러한 연구가 진행된다면 변화량이 큰 객체에 대해 보다 뚜렷하고 자연스러운 보간 결과물을 얻을 수 있을 것이다. 또한, SF-AdaCoF의 경우 모델의 구조 상 이전과 이후 프레임의 정확히 중간 시점의 이미지만을 생성할 수 있다. 만약 이전 혹은 이후 시점에 더 가까운 특정 시점의 보간을 원할 경우 반복적인 작업이 필요하며 이는 오류의 증가로 이어질 수 있다. 따라서 특정 시점에서의 보간을 위한 연구가 필요하다.

References

[1] J. Huh, K. Yoon, S. Kim, and J. Joung, “Research trends in deep learning-based video frame interpolation techniques,” *The Korean Institute of Broadcast and Media Engineers : Broadcast and Media Maganize*, Vol.18, No.2, pp.51-61, 2022.

[2] J. Dong, K. Ota, and M. Dong, “Video frame interpolation: A comprehensive survey,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol.19, No.78, pp.1-31, 2023.

[3] H. Lee, T. Kim, T.-y. Chung, D. Pak, Y. Ban, and 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.

[4] J. Dai et al., “Deformable convolutional networks,” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[5] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[6] S. Niklaus, L. Mai, and F. Liu, “Video frame interpolation via adaptive convolution,” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[7] K. M. Briedis, A. Djelouah, R. Ortiz, and M.Gross, “Kernel-based frame interpolation for spatio-temporallyadaptive rendering,” *SIGGRAPH '23: ACM SIGGRAPH 2023 Conference Proceedings*, 2023.

[8] Z. Teed and J. Deng, “RAFT: Recurrent all-pairs field transforms for optical flow,” *ECCV(European Conference on Computer Vision) 2020*, pp.402-419, 2020.

[9] Z. Huang et al., “FlowFormer: A transformer architecture for optical flow,” *ECCV(European Conference on Computer Vision) 2022*, pp.668-685, 2022.

[10] G. Zhang, Y. Zhu, H. Wang, Y. Chen, G. Wu, and L. Wang, “Extracting motion and appearance via inter-frame attention for efficient video frame interpolation,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.5682-5692, 2023.

[11] N. Singla, “Motion detection based on frame difference method,” *International Journal of Information & Computation Technology*, Vol.4, No.15, pp.1559-1565, 2014.

[12] S. Niklaus, L. Mai, and F. Liu. “Video frame interpolation via adaptive separable convolution,” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[13] O. Köpüklü, M. Babaee, and G. Rigoll, “Convolutional neural networks with layer reuse,” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[14] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, “Video Enhancement with Task-Oriented Flow,” Vimeo 90k [Data set], http://toflow.csail.mit.edu

[15] D. Scharstein et al., “High-resolution stereo datasets with subpixel-accurate ground truth,” In German Conference on Pattern Recognition (GCPR 2014), Münster, Germany, September, 2014.

[16] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. van Gool, M. Gross and A. Sorkine-Hornung, “A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation,” [Data set], https://davischallenge.org

[17] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," [Data set], https://www.crcv.ucf.edu/data/UCF101.php
[18] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Published as a Conference Paper at the 3rd International Conference for Learning Representations*, San Diego, 2015.
[19] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," *2010 20th International Conference on Pattern Recognition*, 2010.
[20] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

서 동 혁

https://orcid.org/0009-0005-8185-399X
e-mail : hyeoks1856@gmail.com
2018년 ~ 현 재 국민대학교
AI빅데이터융합경영학과 학사과정
관심분야 : Deep Learning, Computer Vision & BigData

고 민 성

https://orcid.org/0009-0004-7719-1682
e-mail : kms990321@gmail.com
2018년 ~ 현 재 국민대학교
AI빅데이터융합경영학과 학사과정
관심분야 : Deep Learning, Computer Vision & BigData

이 승 학

https://orcid.org/0009-0009-3089-2795
e-mail : dltmdgkr95@gmail.com
2018년 ~ 현 재 국민대학교
AI빅데이터융합경영학과 학사과정
관심분야 : Deep Learning, Computer Vision & BigData

박 종 혁

https://orcid.org/0000-0003-4283-1155
e-mail : jonghyuk@kookmin.ac.kr
2015년 서울대학교 산업공학과(학사)
2021년 서울대학교 산업공학과(박사)
2015년 ~ 2016년 삼성전자 메모리사업부
2021년 ~ 현 재 국민대학교
AI빅데이터융합경영학과 조교수
관심분야 : Deep Learning, Machine Learning, Computer Vision