NAFNet 기반 개선된 비디오 프레임 보간 기법

윤기환, 정진우, 김성제, 허진강
한국전자기술연구원
{rlghksdbs, jw.jeong, sungjei.kim, jingang4394} @keti.re.kr

Enhanced video frame interpolation based on NAFNet

Kihwan Yoon Jinwoo Jeong Sungjei Kim Jingang Huh
Korea Electronic Technology Institute

요 약

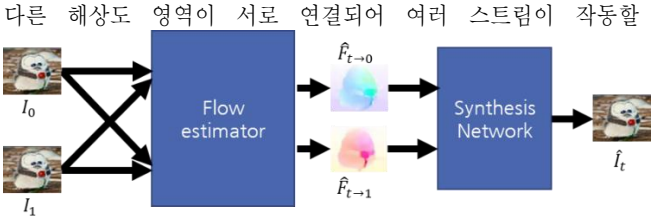
최근 딥러닝은 다양한 컴퓨터 비전에 적용되어 높은 성능을 제공하고 있고 이에 따라 중간 프레임을 생성하는 비디오 프레임 보간 기법에도 딥러닝이 적용되고 있다. 많은 딥러닝 기반의 비디오 프레임 보간 기법은 크게 옵티컬 플로우를 추정하는 플로우 추정 네트워크와 합성 네트워크로 구성되며 본 논문에서는 합성 네트워크 부분의 성능향상을 위한 네트워크에 대하여 다룬다. 합성 네트워크에 주로 사용되는 UNet 구조와 GridNet 구조의 장단점과 네트워크에 따른 보간 결과의 차이에 대해서 알아보고 영상 복원에서 제안된 NAFNet 을 비디오 보간 기법에 맞게 변형시켜 합성 네트워크에 적용한 보간 결과의 차이를 보였다. 실험결과는 기존 네트워크 대비 Vimeo90K 데이터셋에 대하여 PSNR 값이 0.63dB 개선됨을 보여준다.

1. 서론

비디오 프레임 보간 기법(Video Frame Interpolation)은 연속된 두 개의 프레임을 사용하여 두 프레임 사이의 중간 프레임을 생성하는 것이다. 이것은 비디오 프레임율을 늘려 비디오 재생 시 영상이 부드럽게 재생되도록 만들어주거나 슬로우 (slow) 모션을 생성하게 한다. 최근 다양한 딥러닝 네트워크를 사용한 비디오 보간 기법이 제안되어 전통적인 방법에 비해 월등한 성능을 나타내고 있다.

비디오 프레임 보간 기법은 커널 기반 기법과 플로우 기반 기법으로 구분되며 플로우 기반 기법은 <그림 1>과 같이 두개의 입력 프레임{ I_0, I_1 } 사이의 플로우{ $F_{t\to0}$, $F_{t\to1}$ }를 추정하는 플로우 네트워크와 추정된 플로우 및 문맥 정보, 마스크 등을 이용하여 중간 프레임 $\hat{I}_t$ 을 생성하는 합성 네트워크(synthesis network)로 구성이 되어있다. [4,6,7]. 비디오 프레임 보간 기법에서 합성 네트워크의 구성에 따라 성능 향상에 큰 영향을 미치는 것으로 보고되고 있으며[3-9], 따라서 본 논문에서는 합성 네트워크의 성능을 개선하기 위한 네트워크를 제안하도록 한다.

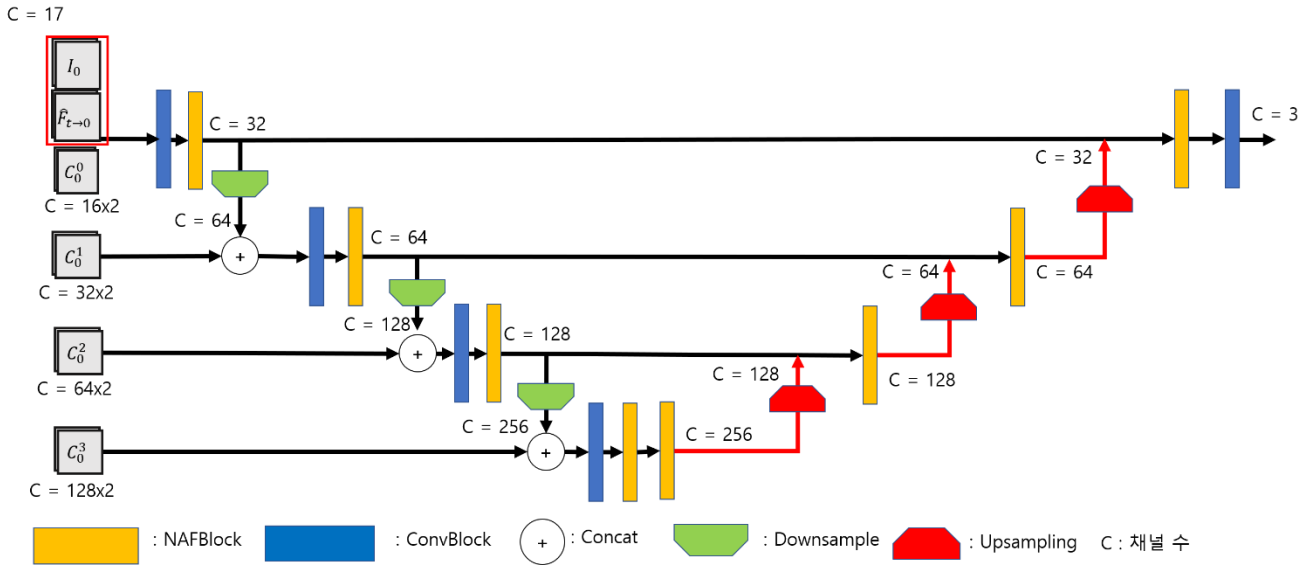
기존의 보간 기법에서 합성 네트워크는 UNet[1] 구조와 GridNet[2] 구조가 주로 사용되었다[3-9]. UNet 구조는 인코더-디코더 구조로 인코더 부분은 서브 샘플링 연산자를 사용하여 수용 영역(receptive field)를 증가시키고 특징 맵(feature map) 사이즈를 줄이며 디코더 부분에서는 다시 기존 해상도 사이즈로 복원을 시켜준다. UNet 구조는 단순하고 파라미터의 수가 적은 장점이 있지만 하나의 스트림으로 특징 맵의 사이즈를 줄이면 출력 예측에서 해상도 손실이 발생한다는 단점이 있다. GridNet 의 경우 그리드 패턴으로 서로 다른 해상도 영역이 서로 연결되어 여러 스트림이 작동할 수 있어 더 높은 성능을 제공할 수 있지만 연산량이 UNet 에 비해 크다는 단점이 있다[2].



<그림 1. 비디오 프레임 보간 기법 네트워크 구조>

기존의 연구들에서 동일한 플로우 네트워크를 사용한 후 UNet 과 GridNet 의 두 네트워크의 직접적인 성능을 비교, 평가한 적은 없다. 본 논문에서는 UNet 과 GridNet 을 직접적으로 비교하여 비디오 프레임 보간 기법에 더 우수한 합성 네트워크를 제시한다. 또한 최근 제안된 영상 복원 네트워크인 NAFNet(Non-leaner Activation Funtions Network)[10]을 이용하여 비디오 프레임 보간을 위한 합성 네트워크를 구성하는 방법에 대하여 제안한다.

본 논문의 구성은 다음과 같다. 2 절에서는 기존에 사용되던 합성 네트워크를 간략히 설명하고 NAFNet 에서 제안한 NAFBlock 을 비디오 프레임 보간 기법에 적용시킨 방법에 대하여 설명한다. 3 절에서는 성능 평가를 위한 실험 환경과 결과를 서술하여 제안 기법의 우수성을 보여준다. 4 절에서는 결론을 서술한다.



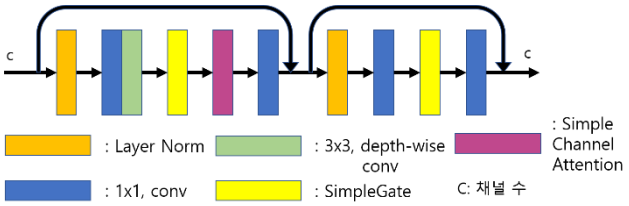
<그림 3. 비디오 보간 기법을 위한 NAFNet 기반 합성 네트워크 구조도>

2. NAFNet 기반 합성 네트워크

영상 복원을 위해 딥러닝을 이용한 다양한 기법이제안되고 있으며 그 성능이 지속적으로 향상되고 있다. 최근 네트워크 복잡도도 기존 복원 방법에 비해 감소함에도 복원 성능은 더욱 우수한 NAFNet 기법이 Chen 에 의해 제안되었다 [10]. 여기서 Chen 은 최근 연구되는 네트워크들의 구조가 너무 복잡하다는 문제를 제시하였고 ReLU, GeLU, Sigmoid 와 같은 비선형 연산을 사용하지 않는 NAFBlock 을 제안하였다 <그림 2>. 이 논문에서는 NAFBlock 을 단순한 UNet 구조에 적용하였음에도 불구하고 영상 복원 분야에서 기존 방법들에 비해 우수한 성능을 보였다[10].

비디오 프레임 보간 기법의 합성 네트워크는 일종의 영상 복원으로 볼 수 있으며 본 논문에서는 합성 네트워크로 NAFBlock 을 적용하여 보간 성능을 개선하였다. NAFBlock 을 적용한 네트워크 구조는 <그림 3>과 같이 UNet 구조를 사용하였다. 제안한 정제 네트워크의 입력으로 입력 프레임 $\{I_0, I_1\}$, 플로우 $\{\hat{F}_{t\to0}, \hat{F}_{t\to1}\}$, 입력에 플로우를 적용하여 와핑한 와핑 프레임 및 마스크(mask)를 사용하였다. 또한 최근 연구 결과에 따르면 문맥 정보를 합성 네트워크에 입력으로 사용할 경우 더욱 높은 성능을 제공한다고 보고되었으므로 문맥 정보$\{C_0^m, C_1^m\}$를 입력으로 사용하였다[3, 8]. 이때 m=0 일 때 해상도가 입력 프레임과 같고 m 이 1 씩 증가할수록 해상도는 1/2 씩 감소한다.

네트워크의 세부 사항은 다음과 같다. 네트워크의 인코더 부분에서 입력 프레임과 플로우는 총 17 채널, 문맥 정보들$\{C_0^n, C_1^n\}$의 채널을 각각 [16, 32, 64, 128]를 사용하여 총 문맥정보에 대한 채널은 [32, 64, 128, 256] 이다. 문맥 정보는 각각의 NAFBlock 을 통과한 후 다운 샘플링된 특징 데이터에 연결 (concatenation) 된다. NAFBlock 의 입력과 출력에 대한 채널의 수는 동일해야 하기 때문에 문맥 정보에 의해 늘어난 채널 수는 NAFBlock 을 통과하기 전에 ConvBlock 을 추가하여 채널을 일치시키도록 하였다. 또한 NAFBlock 을 통과한 후 다운 샘플링은 kernel size=2, stride=2 인 컨볼루션을 사용하여 다운 샘플링을 해주었다. 네트워크의 디코더 부분에서 업 샘플링을 하는 경우 픽셀 셔플을 사용해 업 샘플링을 해주었다.



<그림 2. NAFBlock 구조 [10]>

<표 1. NAFNet 내부 Block 세부사항>

ConvBlock	Conv2D, kernel size=3, stride=1
DownSample	Conv2D kerner size=2, stride=2
UpSampling	PixelShuffle, upscale factor =2

<표 1>을 통해 더 자세한 블록들의 구조를 알 수 있다.

3. 실험결과 및 분석

본 논문은 제안한 합성 네트워크의 성능을 평가하기 위해 플로우 추정 기법은 RIFE 기법에서 제안된 IFNet 을 사용하였다 [7]. RIFE 의 합성 네트워크는 UNet 으로 구성되었으며 제안 방법과 동일하게 입력으로 17 채널을 사용하였으며 문맥 정보를 사용하였다. 그러나 문맥 정보를 구성하는 방법은 제안 방법과 일치하지 않는다. 본 절에서 RIFE 네트워크를 기반으로 합성 네트워크를 UNet, GridNet 과 NAFNet 구조를 사용했을 경우에 대한 보간 결과 차이를 먼저 보이고 각 네트워크를 사용할 경우 보간 시간 및 메모리 소모량에 대하여 보인다.

실험은 Vimeo90K 데이터를 패치 사이즈 224x224 크기로 사용하여 학습하였고 검증 데이터로 Vimeo90k , UCF101, HD 를 사용하였다. 성능 비교를 위하여 Peak to Noise Ratio(PSNR)을 사용하였다. 모든 학습은 그래픽 카드 RTX A6000 을 사용하였고 에폭 300, 배치 사이즈 32, 옵티마이저는 AdamW, Cosine Annealing 을 사용하여 최대 학습율

0.0003 에서 최소 학습율 0.00003 까지 점진적으로 줄이며 실험을 하였다. 성능 검증에는 그래픽 카드는 RTX Titan 을 사용하였다

<표 2>는 각각의 검증 데이터에 대하여 각각의 정제 네트워크에 대한 PSNR 측정결과를 비교하여 보여준다. <표 2>의 결과를 통해 기존 UNet 구조가 하나의 흐름으로학습하기 때문에 출력에서 해상도 손실이 있다는 문제점이다중 흐름을 가지는 GridNet 을 통해 개선이 되는 것을 알 수있다. 제안한 NAFNet 기반 합성 네트워크의 경우UNet 구조임에도 불구하고 GridNet 에 비해 성능이 개선이되는 것을 알 수 있다. 이것은 합성 네트워크를 구성할 시네트워크의 전체 구조도 중요하지만 블록의 구조도 성능에 큰영향을 끼친다는 것을 보여준다. NAFNet 기반 합성네트워크에서 Vimeo90K 와 HD 데이터의 PSNR 값이 가장 높은것을 확인할 수 있고 UCF 데이터에 대해서는 두번째로 높은것을 확인할 수 있다. 평균적으로 NAFNet 을 학습 네트워크로사용하였을 경우 기존의 정제 네트워크 보다 성능이 개선되는것을 알 수 있다. 특히 Vimeo90K 에서 0.65dB 의 큰 성능향상을 보이고 있다.

<표 2. 각각의 정제 네트워크 PSNR 비교>

Datasets	UNet	GridNet	NAFNet (제안 방법)
Vimeo90K	35.50	36.04	36.15
UCF101	35.26	35.35	35.33
HD	32.15	32.26	32.35

<표 3> 과 <표 4> 는 1280x720 크기의 영상 한 장을 처리하기 위해 각각 합성 네트워크에서 소요되는 추론 시간과 GPU 메모리를 보여준다. 결과를 살펴보면 UNet 이 가장 빠르고 적은 메모리를 사용함을 알 수 있다. 반면에 NAFNet 기반 합성 네트워크는 GridNet 에 비해 더욱 짧은 추론 시간을 가지면서 적은 GPU 메모리를 사용함을 알 수 있다. 이는 성능, 추론 시간, 메모리 사용량 측면에서 제안 방법이 GridNet 기반 합성 네트워크 보다 우수하다는 것을 보여준다.

<표 3. 추론 시 합성 네트워크의 연산 시간 비교 (ms)>

	UNet	GridNet	NAFNet (제안 방법)
Running time	76.8	158.7	137.3

<표 4. 추론 시 GPU 메모리 비교>

	UNet	GridNet	NAFNet (제안 방법)
Memory(MiB)	1787	3545	2967

4. 결론

본 논문에서는 정제 네트워크에 따른 보간 결과를 실험을 통하여 비교하였고 NAFBlock 을 활용한 합성 네트워크를 제안하였다. 제안한 방법은 기존 방법의 결과보다 향상된 결과를 얻었었고 GridNet 에 비해 연산 속도가 빠르고 성능이 좋지만 UNet 에 비해 연산속도가 느린 문제점이 있다. 따라서 NAFNet 의 내부 구조 변화를 통한 추가적인 연구가 필요하다.

감사의 글

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임. (No.2021-0-00087, SD/HD 급 저화질 미디어의 고품질 변환 기술 개발).

참고문헌

[1] Olaf Ronnerberger, Philipp Fischer, Thomas Brox. “UNet: Convolution Networks for Biomedical Image Segmentation” In International Conference on Medical image computing and computer-assisted intervention, MICCAI. 2015.

[2] Damien Fourure, Remi Emonet, et al. “Residual Conv-Deconv Grid Network for Semantic Segmantation.” Proceedings of the British Machine Vision Conference. 2017

[3] Niklaus, Simon, and Feng Liu. "Context-aware synthesis for video frame interpolation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[4] 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, CVPR. 2018.

[5] Bao, Wenbo, et al. "Depth-aware video frame interpolation." Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition, CVPR. 2019.

[6] Haopeng Li, Yuan Yuan, Qiwang, et al. “Video Frame Interpolation via Residual Refinement.” Proceedings of the IEEE conference on Acoustics, Speech, and signal Processing, ICASSP. 2020.

[7] Huang, Zhewei, et al. "Rife: Real-time intermediate flow estimation for video frame interpolation." arXiv preprint arXiv:2011.06294 (2020).

[8] Simon Niklaus, Feng Liu. “Softmax Splatting for Video Frame Interpolation.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. 2020.

[9] Junheum Park, Chul Lee and Chang-Su Kim. “Asymmetric Bilateral Motion Estimation for Video Frame Interpolation.” Proceedings of International Conference on Computer Vision, ICCV. 2021.

[10] Liangyu Chen, Xiaojie Chu, et al. “Simple Baseline for Image Restoration.” 1st place on the NTIRE 2022: CVPR 2022 New Trend in Image Restoration and Enhancement workshop and challenge