비스플라인과 푸리에 기저 함수를 이용한 연속적인 시공간 비디오 초해상도 복원 연구

김은진 ○, 김현진, 유재준
울산과학기술원 인공지능대학원
{ eunjin.kim, hyeonjin.kim, jaejun.yoo}@unist.ac.kr

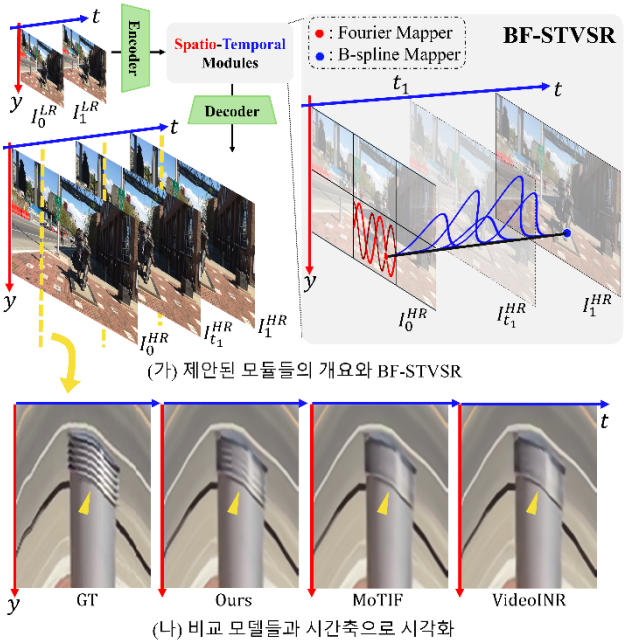
요 약

저해상도 및 저프레임률 비디오를 고해상도 및 고프레임률 비디오로 복원하는 것은 고품질 비디오를 위해 필수적이며, 이를 위해 연속적인 시공간 비디오 초해상도(Continuous Spatial-Temporal Video Super Resolution, C-STVSR) 복원 연구가 대두되었다. 하지만 Implicit Neural Representation (INR)을 활용하여 연속적인 모델링을 진행하는 기존의 연구들은 단순한 좌표 연결 또는 사전 학습된 광학 흐름 네트워크를 기반으로 움직임을 예측하고 있다. 이와 같은 방식은 복잡한 비디오 데이터를 효과적으로 포착하는 데 한계를 보인다. 본 연구에서는 일반적인 연구 결과와 달리 위치 인코딩(position encoding)을 단순하게 적용하는 것이, 특히 광학 흐름 네트워크와 결합되었을 때, 고해상도 복원 성능을 저하시킨다는 것을 발견했다. 이를 해결하기 위해, 본 연구에서는 **BF-STVSR** 이라는 새로운 C-STVSR 프레임워크를 제안한다. 이 프레임워크는 비디오의 시공간적 특성을 보다 효과적으로 표현하기 위한 두 가지 핵심 모듈을 포함한다: 1) 시간축으로의 부드러운 보간을 위한 B-spline Mapper, 2) 공간축으로의 고주파 정보 파악을 위한 Fourier Mapper. 제안된 모델은 기존의 C-STVSR 모델들 중에서 PSNR 과 SSIM 지표에 대해 가장 높은 성능을 달성하여, 향상된 고품질 비디오 복원을 보여준다.

1. 서론

저해상도 및 저프레임률 비디오를 고해상도 및 고프레임률 비디오로 복원하는 것은 고품질 비디오를 위해 필수적이다. 이를 위해, 많은 연구들이 딥러닝 기반의 비디오 초해상도 복원(Video Super-Resolution, VSR) 방법과 비디오 프레임 보간(Video Frame Interpolation, VFI) 방법을 제안하였다. VSR 은 이웃 프레임들의 정보를 활용해 목표 프레임의 공간 해상도를 향상시키고, VFI 는 비디오 데이터 내에서 목표 프레임까지의 움직임을 예측해 프레임률을 개선시킨다. 하지만 기존 방법들은 대부분 학습 데이터셋 해상도에 제한되었다.

이런 한계를 극복하기 위해, 이산 신호를 다층 퍼셉트론(MLP)으로 매핑시켜 연속표현으로 나타낼 수 있는 Implicit Neural Representation (INR)을 적용한 연구들이 제안되고 있다. 최근에는 INR 을 비디오에 적용하여 연속적인 시공간 비디오 초해상도 복원(Continuous Spatial-Temporal Video Super Resolution, C-STVSR)이 가능하게 하는 연구들이 제안되었다. 그 중에서, VideoINR[1]은 처음으로 시공간 좌표를 연속적인 후방 움직임 필드에 매핑하였다. 후방 움직임 벡터를 활용해 목표 프레임 좌표로 잠재 특징을



(가) 제안된 모듈들의 개요와 BF-STVSR

(나) 비교 모델들과 시간축으로 시각화

그림 1. (가) BF-STVSR 은 B-spline Mapper 와 Fourier Mapper 를 통해 향상된 시공간 비디오 초고해상도 복원이 가능. (나) 이전의 C-STVSR 모델들과 비교하기 위해 시간축으로 시각화한 결과.

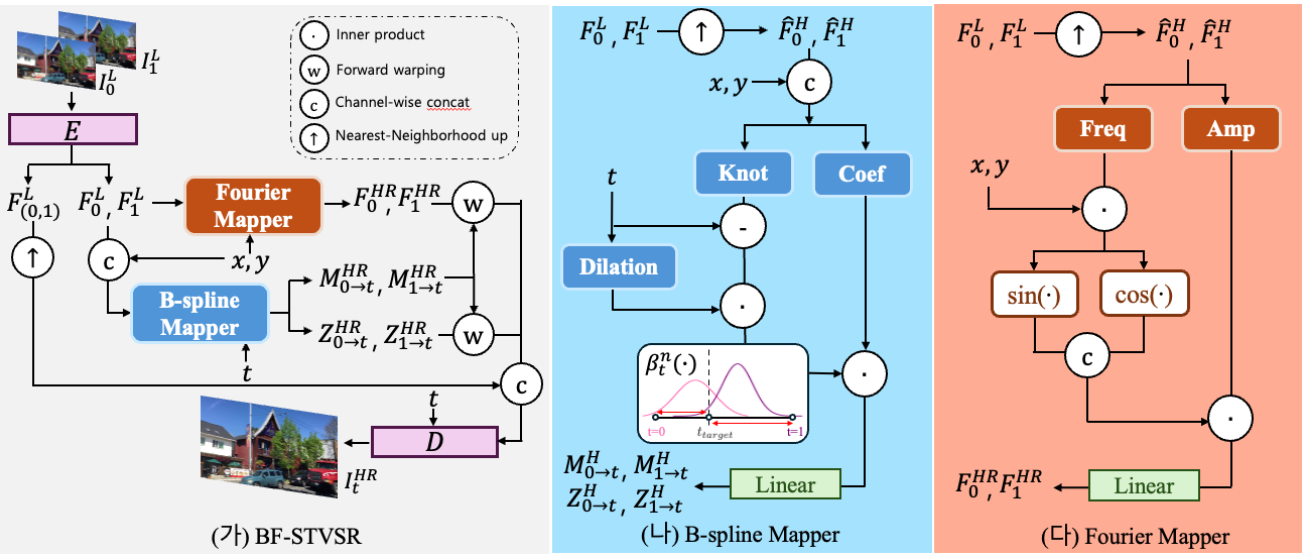


그림 2. (가) 제안된 프레임워크 BF-STVSR 의 전체적인 구조로 고품질의 고해상도 중간 프레임 복원을 목표로 함. (나) 제안된 B-spline Mapper 의 전체적인 구조. (다) 제안된 Fourier Mapper 의 전체적인 구조.

후방 와핑 시켜 목표 프레임을 복원할 수 있다. 이 방법을 확장해 MoTIF[3]는 사전 학습된 광학 흐름 네트워크를 추가로 사용하고, 소프트맥스 스플래팅 [5]을 활용한 전방 워핑을 적용함으로써 성능을 개선하였다.

그러나 기존 모델들은 좌표와 잠재 특징을 단순히 병합하는 방식을 사용하며, Fourier 인코딩과 같은[6,7,9] 위치 인코딩 기술을 활용하지 않고 있다. 이러한 방식은 특히 C-STVSR 을 진행할 때 비디오의 복잡하고 동적인 움직임을 표현하는 데 어려움을 겪으며, 결과적으로 높은 주파수 정보를 유지하지 못하고 품질이 낮은 프레임을 생성하는 문제를 초래한다. 추가적으로 본 연구에서는 단순히 위치 인코딩을 적용하는 것은 오히려 성능을 저하시킨다는 것을 발견했고, 이는 사전 학습된 옵티컬 플로우 네트워크와 결합되었을 때 더욱 두드러져서 모델의 유연한 움직임 예측을 제한할 가능성이 있다.

이를 해결하기 위해, 본 연구에서는 **BF-STVSR** 이라는 새로운 프레임워크를 제안하고, 두 가지 모듈을 통해 시공간축 고해상도 복원을 효과적으로 처리한다. 첫째로, 시간축으로는 연속적인 움직임을 표현하기 위해 비스플라인 기저 함수를 활용한 B-spline Mapper 를 제안한다. 둘째로, 공간축으로는 프레임 내에서의 높은 주파수 정보 파악을 위해 푸리에 기저 함수를 활용한 Fourier Mapper 를 제안한다. 또한, 사전 학습된 광학 흐름 네트워크를 학습할 때 직접적으로 입력으로 사용하는 것 대신, 간접적으로 네트워크의 움직임 예측을 지도할 수 있도록 한다. 이와 같은 방법을 통해 논문에서 제안하고 있는 BF-STVSR 은 C-STVSR 분야에서 가장 높은 성능을 달성했고, 다양한 실험을 통해 효과성을 입증하였다.

2. 시공간 기저 함수 변경을 통한 연속적인 비디오 초해상도 복원

2.1 BF-STVSR 작동 방식

본 논문에서 제안하고 있는 BF-STVSR 의 전체적인 구조는 그림 2(가)에 나타나 있다. 해당 모델의 목표는, 두개의 저해상도 프레임 $I_0^L, I_1^L \in \mathbb{R}^{3\times H\times W}$ 이 주어졌을 때, 임의의 시간 $t \in [0,1]$ 에서 임의의 스케일 s 로 고해상도 중간 프레임 $I_t^H \in \mathbb{R}^{3\times sH\times sW}$ 을 생성하는 것이다. 우선, 인코더 E 는 두 개의 저해상도 프레임을 입력으로 받아 세 개의 잠재 특징들 $F_0^L, F_{(0,1)}^L, F_1^L \in \mathbb{R}^{3\times H\times W}$ 을 생성한다. 여기서 F_0^L 과 F_1^L 은 각각 I_0^L 와 I_1^L 의 잠재 특징을 나타내며, $F_{(0,1)}^L$ 은 두 입력 프레임의 정보가 결합된 결과로 중간 프레임의 템플릿같은 잠재 특징으로 활용된다. 다음으로, 잠재 특징 F_0^L 과 F_1^L 은 두 개의 제안된 매퍼에 의해 처리된다. B-spline Mapper 는 목표 시간 t로의 고해상도 광학 흐름을 예측하여 $M_{0\to t}^H, M_{1\to t}^H \in \mathbb{R}^{2\times sH\times sW}$ 를 생성한다. 한편, Fourier Mapper 는 목표 스케일 s에서의 고해상도 공간 특징을 추정하여 $F_0^H, F_1^H \in \mathbb{R}^{C\times H\times W}$ 를 생성한다. 마지막으로, 고해상도 특징 F_0^H 와 F_1^H 는 예측된 광학 흐름 $\mathrm{M}_{0\to t}^{\mathrm{H}}, \mathrm{M}_{1\to t}^{\mathrm{H}}$ 를 기반으로 목표 시간 t로 전방 와핑을 수행하여 목표 좌표에서의 잠재 특징 F_t^H 를 생성한다. 와핑된 특징들은 목표 시간 t와 $F_{(0,1)}^H$ (이는 $F_{(0,1)}^L$ 의 최근접 이웃 업샘플링 버전)과 결합되며, 이를 디코더를 통해 최종적으로 고해상도 중간 프레임 I_t^H로 복원한다.

2.2 시간축 보간을 위한 B-spline Mapper

비디오에서 내재된 움직임을 정확히 예측하는 것은 이웃 프레임 두 개를 이용해 시각적으로 자연스러운 중간 프레임을 생성하는 데 있어 핵심적인 요소

이다. VFI 연구에서는 움직임을 정확히 예측하기 위해 다양한 기술이 제안되었지만, 대부분 고정된 스케일에서의 보간을 목표로 설계되었다. 따라서 임의의 스케일을 처리해야 하는 C-STVSR 에서 이러한 방법들을 적용하는 것은 간단하지 않다. 이를 극복하기 위해 VideoINR 과 MoTIF 는 시공간 좌표를 입력으로 받아 임의의 시간 t와 스케일 s에서 움직임을 모델링할 수 있는 다층 퍼셉트론(MLP)을 활용한 암시적 신경 표현(INR)을 사용했다. 이러한 INR 방법은 유연한 모델링이 가능하지만, 비디오의 복잡한 움직임을 포착하는 데에는 한계가 있다. 이를 극복하기 위해, 본 논문에서는 비스플라인 표현을 활용한 B-spline Mapper 모듈을 제안한다. 비스플라인 함수는 연속 신호를 모델링하는 데 효과적인 방법으로 잘 알려져 있으며[2], 부드럽고 연속적으로 움직이는 물체의 움직임을 효과적으로 포착할 수 있다. B-spline Mapper 의 세부적인 과정은 그림 2(나)에 나타나 있다.

본 논문에서는 MoTIF 의 Space-Time Local Implicit Neural Functions (ST-INF)를 수정하여 B-spline Mapper 를 구현했으며, ST-INF 와 유사하게 임의의 시간 t에서 고해상도 전방 움직임 벡터 $M^H_{0\to t}, M^H_{1\to t}$와 신뢰도 맵 $Z^H_{0\to t}, Z^H_{1\to t}$를 예측한다. 이때, B-spline Mapper 는 외부 네트워크(예: RAFT[4])에서 광학 흐름 정보를 가져오는 대신, 인코딩된 잠재 특징 F^L_0과 F^L_1을 입력으로 사용한다. 또한, 목표 시간 t에 대한 움직임 벡터를 직접적으로 예측하는 대신, B-spline Mapper p_ψ는 아래 수식과 같이 비스플라인 기저 함수의 coefficients 와 knots 를 예측하여 비디오의 내재된 움직임을 모델링한다:

$$p_\psi(z_r, \delta_r, \hat{t}) = c_r \odot \beta^n\left(\frac{\hat{t} - k_r}{d}\right). \quad (1)$$

여기서, $c_r = p_c(z_r, \delta_r)$, $k_r = p_k(z_r, \delta_r)$ 이고, $d = p_d(g)$이다. 각각 coefficients, knots, dilation 요소를 예측하는 함수이며, z_r은 최근접 좌표의 잠재 특징 벡터, $\hat{t}$는 참조 프레임과의 상대적 시간 거리, δ_r은 공간 상대 좌표를 나타낸다. 이렇게 얻어진 비스플라인 잠재 표현을 f_{θ_b}를 통해 선형적으로 투영함으로써, 쿼리 좌표 q에서의 움직임 벡터와 신뢰도 맵을 얻는다:

$$\{Z^H_{t_r\to t}(q), M^H_{t_r\to t}(q)\} = f_{\theta_b}(p_\psi(z_r, \delta_r, \hat{t})). \quad (2)$$

예측된 모션 벡터를 이용해, 잠재 특징 F^H_0, F^H_1와 신뢰도 맵은 소프트맥스 스플래팅[5]을 통해 목표 시간 t로 전방 와핑되고, 중간 잠재 특징 F^H_t와 이에 대응하는 신뢰도 맵 Z^H_t가 생성된다. B-spline Mapper 은 입력 프레임 특징맵에 내재된 움직임을 학습하여 더욱 강건하고 유연한 움직임 모델링 방식을 제공한다. 또한, 테스트 단계에서 광학 흐름 정보에 의존하지 않기 때문에 MoTIF[3]에 비해 효율적이다.

2.2 공간축 보간을 위한 Fourier Mapper

비록 B-spline Mapper 로 향상된 모션 모델링 덕분에 중간 특징 F^H_t의 품질이 개선되었지만, F^H_0과 F^H_1에서 전파된 특징의 품질이 여전히 중요한 역할을 한다. VideoINR[1]와 MoTIF[3]는 단순하게 MLP 를 사용하여 잠재 특징 F^L_0과 F^L_1을 보간한다. 그러나 INR 은 spectral bias 와 같이 고주파 세부 정보 파악에 한계가 있으며[6, 7, 8], 이는 보간된 중간 프레임 결과의 품질 저하로 이어질 수 있다. LTE[9]는 임의의 스케일 이미지 초해상도 태스크에서 이를 해결하기 위해, 푸리에 기저 함수를 활용해 잠재 특징을 표현하였다. 이는 향상된 성능을 보여주었고, 본 논문에서는 이러한 접근법에서 영감을 받아, 푸리에 표현을 활용한 Fourier Mapper 을 제안한다. Fourier Mapper 의 상세 과정은 그림 2(다)에 나타나 있다.

Fourier Mapper g_ϕ는 공간축으로 입력 프레임 내에서 고주파수 특징을 파악하기 위해, 푸리에 기저 함수의 frequency 와 amplitude 를 예측한다:

$$\{F^H_0(q), F^H_1(q)\} = f_{\theta_f}\left(g_\phi(z_r, \delta_r)\right), \quad (3)$$

$$where\ g_\phi(z_r, \delta_r) = A_r \odot \begin{bmatrix} cos(\pi F_r \delta_r) \\ sin(\pi F_r \delta_r) \end{bmatrix} \quad (4)$$

여기서 $A_r = g_a(z_r), F_r = g_f(z_r)$이고, $z_r = F^L_{t_r}(q_r)$은 쿼리 좌표 $q = (x, y)$에 가장 가까운 잠재 특징 벡터이고, $\delta_r (= q - q_r)$은 공간적 상대 좌표를 나타낸다. g_a와 g_f는 각각 amplitude 추정기($\mathbb{R}^C \mapsto \mathbb{R}^{2C}$)와 frequency 추정기($\mathbb{R}^C \mapsto \mathbb{R}^{2C}$)이다. 쿼리 포인트의 주요 주파수를 잠재 공간에서 예측함으로써, Fourier Mapper 는 보간된 특징 $\hat{F}^H_0$와 $\hat{F}^H_1$에서 고주파수 특징들을 더 잘 파악할 수 있다. 이렇게 얻어진 푸리에 잠재 표현을 f_{θ_f}를 사용해 F^H_0와 F^H_1로 선형 투영하면, 목표 좌표에서의 개선된 잠재 특징 F^H_t을 얻을 수 있다. Fourier Mapper 는 LTE 와 달리 저해상도 영역에서 계수들을 보간하지 않고, 고해상도 영역에서 각 계수를 직접 예측하고 있다.

2.3 손실함수

제안된 모델은 MoTIF[3]을 기반으로 end-to-end 방식으로 학습되고, 손실 함수는 아래와 같다:

$$\mathcal{L} = \mathcal{L}_{char}\left(I^H_t, \hat{I}^H_t\right) + \beta \sum_{i=0}^{1} \mathcal{L}_{char}\left(M^H_{i\to t}, \widehat{M}^H_{i\to t}\right) \quad (5)$$

여기서 $\mathcal{L}_{char}$는 Charbonnier 손실이며($\mathcal{L}_{char}(x, \hat{x}) = \sqrt{\|x - \hat{x}\|^2 + \varepsilon^2}$), β 는 0.01 로 설정된 하이퍼파라미터이다. I^H_t는 시간 t에서 예측된 고해상도 중간 프레임이고, $\hat{I}^H_t$는 정답 중간 프레임이다. $M^H_{i\to t}$은 시간 t에서 예측된 움직임 벡터이고, $\widehat{M}^H_{i\to t}$는 RAFT[4]로부터 예측된 움직임 벡터이다. 이때, RAFT 는 훈련 과정에서만 B-spline Mapper 을 지도하기 위해 사용되며, 추론 과정에서는 활용되지 않는다.

테이블 1. Gopro, Adobe240 데이터셋에 대해서 Fixed-STVSR 과의 성능 비교 결과. 성능 지표는 PSNR(dB)과 SSIM 을 사용. 모든 프레임은 공간 축으로는 $\times 4$, 시간 축으로는 $\times 8$ 크기로 보간됨. 최고 성능은 빨간색, 두 번째로 높은 성능은 파란색으로 표시.

VFI Method	VSR Method	Gopro	Adobe
SuperSloMo	BasicVSR	26.36 / 0.7977	25.94 / 0.7679
QVI	BasicVSR	26.27 / 0.7955	25.20 / 0.7421
DAIN	BasicVSR	26.43 / 0.7966	25.23 / 0.7725
TMNet		28.83 / 0.8514	28.30 / 0.8354
VideoINR		29.41 / 0.8669	29.27 / 0.8651
MoTIF		30.04 / 0.8773	29.82 / 0.8750
Ours		**30.20 / 0.8799**	**30.14 / 0.8808**

3. 결과 및 분석

3.1 실험 세부 사항

학습은 두 단계로 진행되며, 초기 450,000 번 반복 동안 공간 스케일링 팩터는 4 로 고정하며, 이후 150,000 번의 반복 동안은 $[1,4]$에서 균일하게 샘플링한다. Adam 옵티마이저($\beta_1 = 0.9, \beta_2 = 0.999$)를 사용하며, 학습률은 150,000 번 반복마다 10^{-4} 에서 10^{-7} 로 코사인 감쇠를 적용한다. 인코더로 ZoomingSlowMo[10]를 사용하고, 배치 크기는 32 이다. 또한, 훈련 초반 150,000 번 반복 동안 일정 확률로 예측된 전방 움직임을 실제 전방 움직임으로 대체한다. B-spline Mapper 는 세 개의 SIREN[11] 레이어를 사용해 계수와 매듭을 추정하며, Fourier Mapper 는 진폭 및 주파수 추정을 위한 세 개의 SIREN 레이어와 공간 인코딩을 위한 3×3 합성곱 레이어를 포함한다. 두 모듈 모두 잠재 차원은 64 로 설정된다.

데이터셋의 경우, 훈련시에는 720P 해상도의 Adobe240[12]을 사용하고, 총 133 개의 비디오 중 100 개는 훈련, 16 개는 검증, 17 개는 평가에 사용한다. 훈련 시 비디오에서 9 개의 연속 프레임을 선택해 1^{st} 와 9^{th} 프레임을 입력으로 사용하고, 중간의 세 프레임을 랜덤하게 샘플링하여 정답 데이터로 활용한다. 평가시에는 Adobe240[12], Gopro[14] 데이터셋을 사용한다. 기본 공간 스케일은 $\times 4$ 로 설정되고, 시간 스케일은 $\times 8$ 로 설정되어 다중 프레임 보간을 평가한다.

벤치마킹 모델의 경우, 연속적(Continuous)과 고정 스케일(Fixed-scale)로 구분해서 비교한다. Fixed-STVSR 모델은 학습 시 사용된 고정 스케일에서만 초해상도를 수행한다. 두 단계 방법으로는 고정 VSR 모델(e.g., BasicVSR[16]})과 VFI 모델(e.g., SuperSloMo[17], QVI[18], DAIN[19])을 사용한다. 연속적 방법에서는 MoTIF 와 제안된 BF-STVSR 을 비교한다.

테이블 2. Gopro 데이터셋에 대해서 C-STVSR 모델들과의 성능 비교 결과. 성능 지표는 PSNR(dB)과 SSIM 을 사용. 모든 프레임은 표에 명시된 크기로 보간되었고, 최고 성능은 빨간색으로 표기.

Temporal Scale	Spatial Scale	MoTIF	Ours
$\times 6$	$\times 6$	29.36 / 0.8505	**29.44 / 0.8516**
$\times 6$	$\times 12$	**25.81 / 0.7330**	25.78 / 0.7284
$\times 12$	$\times 6$	26.78 / 0.7908	**27.06 / 0.7961**
$\times 12$	$\times 12$	24.72 / **0.7108**	**24.87** / 0.7096
$\times 16$	$\times 6$	25.34 / 0.7527	**25.81 / 0.7621**
$\times 16$	$\times 12$	23.88 / 0.6923	**24.22 / 0.6950**

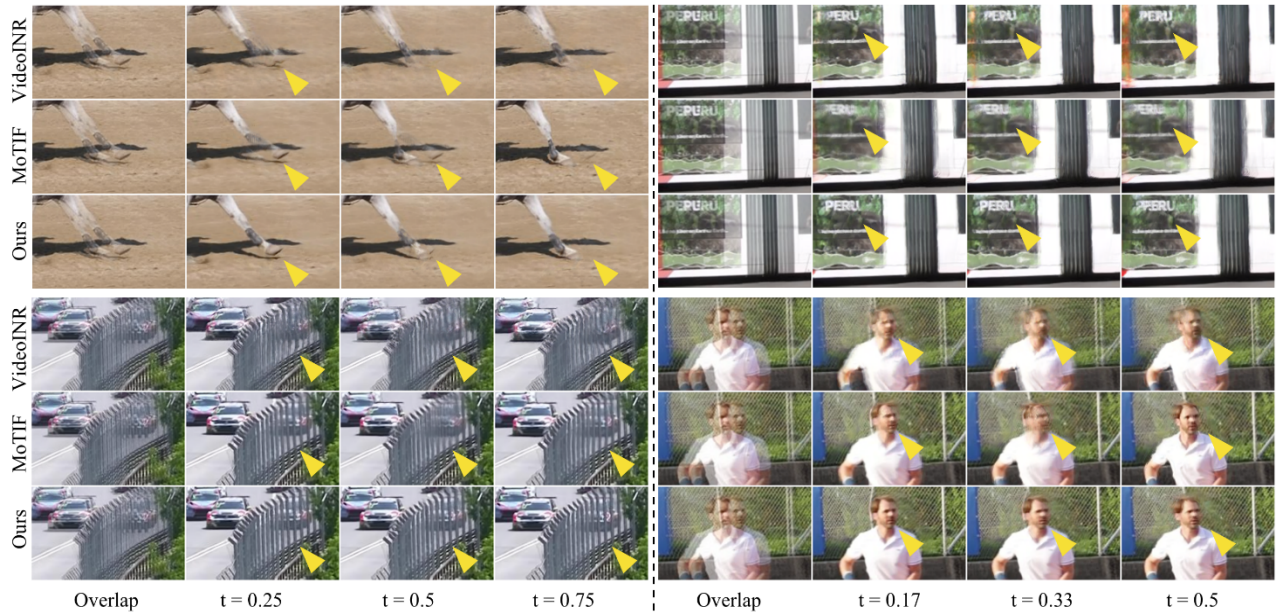
3.2 정량적 실험결과

테이블 1 은 Fixed-STVSR 방법들과 비교한 결과로 GoPro 와 Adobe240 에서 모두 BF-STVSR 이 가장 높은 성능을 보이며, 기존 SOTA 인 MoTIF 모델보다 높은 성능을 달성했다. 특히, MoTIF 는 사전 학습된 옵티컬 플로우 네트워크[4]를 활용해 시간적 특징을 생성하지만, 제안한 모델은 B-spline Mapper 과 Fourier Mapper 를 통해 더 강력한 시간적 및 공간적 특징 표현을 제공한다.

테이블 2 에서는 GoPro 데이터셋으로 임의의 스케일에 대해서 연속 스케일 STVSR(C-STVSR) 방법과 비교를 수행했다. 실험 결과, BF-STVSR 은 대부분의 테스트 케이스에서 최고의 성능을 보였다. 이는 B-spline Mapper 가 높은 스케일의 고해상도 복원에 대해서도 효과적으로 시간적 보간을 처리함을 시사한다. 특히, 동일한 공간축 스케일에서 시간축 스케일이 커질수록 ($\times 6$에서 $\times 12$ 와 $\times 16$) BF-STVSR 의 낮은 성능 저하를 볼 수 있다. 결과적으로, MoTIF 는 어려운 조건에서 더 큰 성능 저하를 경험하는 반면, 제안한 모델은 상대적으로 강건한 성능을 유지한다. 예를 들어, $\times 12$ 공간 스케일에서 시간 스케일이 $\times 6$에서 $\times 12$로 증가할 때, MoTIF 는 약 -4.02 dB 의 성능 저하를 보인 반면, 우리 모델은 약 -3.63 dB 로 약 0.4 dB 의 개선을 보인다. 이는 시간적 간격이 $\times 16$으로 확장되는 극한 조건에서도 제안한 방법이 높은 강건성을 유지함을 보여준다.

3.3 정성적 실험결과

그림 3 은 VideoINR, MoTIF 와 제안된 모델 BF-STVSR 을 정성적으로 비교한 결과이다. 결과는 학습 시 사용된 시간축 스케일($\times 8$, In-distribution)과 사용되지 않은 시간축 스케일($\times 6$, Out-of-distribution)의 보간된 프레임을 보여주고 있다. In-distribution 의 경우, 말의 발굽이나 난간의 줄무늬와 같은 고주파 세부 사항을 제안된 모델이 더 효과적으로 포착하고 있고, Out-of-distribution 의 경우에도 BF-STVSR 이 텍스트의 가장자리와 남성 얼굴 등의 역동적인 움직임이 포함된 장면에서 더 우수한 성능을 보여



(가) In-distribution (나) Out-of-distribution

그림 3. "Overlap"은 두 입력 프레임($t = 0, 1$)의 평균 이미지를 나타내며, 옆의 이미지는 $t = (0, 1)$ 에서 보간된 결과를 보여준다. (가) 학습 시 사용된 시간축 스케일($\times$ **6**)에서의 보간 결과를 나타낸다. (나) 학습 시 사용되지 않은 시간축 스케일($\times$ **6**)에서의 보간 결과를 나타낸다.

준다. 이는 제안한 모델이 물체의 연속적인 움직임을 유연하게 보간하고, 프레임의 고주파 정보를 효과적으로 포착한다는 것을 증명한다.

3.4 Ablation 실험결과

테이블 3 은 사전 학습된 광학 흐름 네트워크 RAFT(**O.F**)와 B-spline Mapper(**B**), Fourier Mapper(**F**)의 사용 유무에 따른 모델 성능을 비교한 결과이다. 첫 번째 행은 MoTIF[3]의 기본 세팅으로, 두 번째와 세 번째 행을 보면 RAFT 를 제안된 모듈과 함께 사용할 경우 성능이 오히려 저하되는 것을 확인할 수 있다. 이는 광학 흐름 네트워크가 비디오 내 복잡한 시공간 정보를 효과적으로 활용하지 못한다는 점을 시사한다. 반면, 마지막 세 행에서는 제안된 모듈을 사용해 공간 및 시간 특징을 추출한 결과를 보여준다. 이 경우 모든 세팅에서 개선된 성능을 보이고 있다. 이는 제안된 모듈이 비디오 데이터에 내재된 풍부한 정보를 효과적으로 추출하고 활용하여 복잡한 시공간 특징을 더 잘 모델링할 수 있음을 보여준다. 또한, 마지막 행은 두 모듈을 통합한 BF-STVSR 의 결과로 가장 우수한 성능을 보여준다. 이는 B-spline Mapper 와 Fourier Mapper 가 서로 보완적으로 작용해 효과적인 C-STVSR 을 가능하게 하는 것을 입증한다. 본 실험 결과는 제안된 모듈들이 기존 광학 흐름 네트워크에 의존하지 않고도 복잡한 비디오 데이터를 처리하는 데 있어 효과적임을 보여주고 있다.

테이블 3. 제안된 위치 임베딩(B, F)과 사전 학습된 광학 흐름 네트워크(O.F)의 영향을 GoPro 와 Adobe 데이터셋에서 비교한 결과. 성능 지표는 PSNR(dB)과 SSIM 사용.

O.F	B	F	GoPro	Adobe
√			30.04 / 0.8773	29.82 / 0.8750
√		√	29.94 / 0.8764	29.73 / 0.8741
√	√		30.03 / 0.8774	29.81 / 0.8756
		√	30.12 / 0.8783	30.02 / 0.8784
	√		30.16 / 0.8792	30.11 / 0.8801
	√	√	**30.20 / 0.8799**	**30.14 / 0.8808**

4. 결론

본 논문에서는 연속 시공간 비디오 초해상도(Continuous Spatial-Temporal Video Super Resolution, C-STVSR)를 위한 새로운 프레임워크인 BF-STVSR 을 제안했다. 구체적으로는, 두 개의 정교한 위치 인코딩 모듈을 제안한다. B-spline Mapper 는 비스플라인 기저 함수를 활용하여 시간축으로 부드럽고 연속적인 보간을 수행하며, Fourier Mapper 는 주요 공간 주파수를 포착하여 세밀한 공간적 디테일을 효과적으로 모델링한다. 또한, 실험 결과로 단순한 위치 인코딩은 C-STVSR 의 성능을 저하시킬 수 있다는 것을 시사하고, 시공간 축에 맞춘 정교한 위치 인코딩의 중요성을 강조한다. BF-STVSR 은 다양한 데이터셋에서 PSNR 과 SSIM 지표에 대해 가장 높은 성능을 달성하였고, 다양한 실험을 통해 해당 모델의 효과성을 보여주었다.

감사의 글

이 논문은 과학기술정보통신부·광주광역시가 공동 지원한 '인공지능 중심 산업융합 집적단지 조성사업'으로 지원을 받아 수행된 연구임. 이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-II220264, 지식기반 심층논리 신경망을 활용한 통합적 비디오 해석과 생성 연구). 이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2020-II201336, 인공지능대학원지원(울산과학기술원)). 이 (성과)는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2022R1C1C100849612, 강력하고 해석 가능한 딥러닝 생성 모델을 위한 신호처리 기반 분석 도구 및 학습 알고리즘 개발: 일반, 바이오, 의료 영상으로의 응용).

참고문헌

[1] Zeyuan Chen, Yinbo Chen, Jingwen Liu, Xingqian Xu, and Xiaolong Wang. Videoinr: Learning video implicit neural representation for continuous space-time super-resolution. In CVPR, 2022.

[2] Byeonghyun Pak and Kyong Hwan Jin. B-spline texture coefficients estimator for screen content image super resolution. In CVPR, 2023.

[3] Yi-Hsin Chen, Si-Cun Chen, Yen-Yu Lin, and Wen-Hsiao Peng. Motif: Learning motion trajectories with local implicit neural functions for continuous space-time video super resolution. In ICCV, 2023.

[4] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In ECCV, 2020.

[5] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In CVPR, 2020.

[6] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.

[7] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In International conference on machine learning, pages 5301–5310. PMLR, 2019.

[8] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. Advances in neural information processing systems, 33:7537–7547, 2020.

[9] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In CVPR, pages 1929–1938, 2022.

[10] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P. Allebach, and Chenliang Xu. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In CVPR, 2020.

[11] V. Sitzmann, J.N. Martel, A.W.Bergman, D. B. Lindell, and G.Wetzstein. Implicit neural representations with periodic activation functions, in NeurIPS, 2020.

[12] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In CVPR, 2017.

[13] Ce Liu and Deqing Sun. A bayesian approach to adaptive video super resolution. In CVPR 2011.

[14] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In CVPR, 2017.

[15] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In CVPRW, 2019.

[16] Kelvin C.K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In CVPR, 2021.

[17] Huaizu Jiang, Deqing Sun and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In CVPR, 2018.

[18] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In NeurIPS, 2019.

[19] Wenbo Bao, Wei-Sheng Lai, Chao Ma and Ming-Hsuan Yang. Depth-aware video frame interpolation. In CVPR, 2019.

[20] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In CVPR, 2021.

[21] Tianfan Xue, Baian Chen, Jiajun Wu and William T Freeman. Video enhancement with task-oriented flow. International Journal of Computer Vision (IJCV), 2019.