특집논문 (Special Paper)
방송공학회논문지 제30권 제3호, 2025년 5월 (JBE Vol.30, No.3, May 2025)
https://doi.org/10.5909/JBE.2025.30.3.344
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

FCTM 6.0의 신경망 기반 특징맵 변환 기술 분석

정 혜 원[a], 임 달 홍[a], 유 장 현[a], 이 주 영[b], 김 휘 용[a]‡

Analysis of NN-Based Feature Transformation Technology in FCTM 6.0

Hyewon Jeong[a], Dalhong Lim[a], Janghyun Yu[a], Jooyoung Lee[b], and Hui Yong Kim[a]‡

요 약

본 논문은 FCM (Feature Coding for Machines)에서 사용되는 핵심 기술인 특징맵 변환 기법들을 고찰하고, 최신 구조인 LightFCTM을 중심으로 경량화된 특징맵 변환 네트워크의 설계 철학과 구조를 설명한다. LightFCTM은 기존 FCTM의 연산 및 메모리 복잡도를 줄이기 위해 설계된 구조로, LightFENet과 LightDRNet으로 구성되며 구조 개선, 순차적 복원 구조 적용, 채널 수 축소의 세 가지 핵심 변경을 포함한다. 본 논문은 계층 수에 따라 확장 가능한 LightFCTM의 일반화된 구조와 핵심 구성 요소를 설명하고, ablation study를 통해 각 설계 요소의 효용성과 선택 근거를 실증하였다. LightFCTM은 이전 버전의 FCTM 대비 50% 이상의 복잡도 절감과 30% 이상의 BD-rate 개선을 달성하였다. 또한, 특정 조건에서의 성능 저하 요인을 분석하고 이를 개선하기 위한 향후 연구 방향을 제시한다.

Abstract

This paper provides an overview of neural network-based feature transformation techniques used in FCM (Feature Coding for Machines), with a particular focus on the lightweight structure adopted in FCTM 6.0, known as LightFCTM. LightFCTM is designed to reduce the computational and memory complexity of the original FCTM architecture. It comprises two main components－LightFENet and LightDRNet－and incorporates three key modifications: structural simplification, a sequential reconstruction mechanism, and reduced channel dimensions. The paper outlines the general structure of LightFCTM, which is scalable by layer depth, and explains its core components and underlying design philosophy. Through ablation studies, the impact and justification of each design element are examined. LightFCTM demonstrates more than 50% reduction in complexity and over 30% improvement in BD-rate compared to previous versions. The paper also analyzes performance degradation observed under certain conditions and discusses potential directions for future improvement.

Keyword : Feature compression, Feature coding, Coding for machines, Feature transform

a) 경희대학교 컴퓨터공학부(Kyung Hee University)
b) 한국전자통신연구원(Electronics and Telecommunications Research Institute)
‡ Corresponding Author : 김휘용(Hui Yong Kim)
E-mail: hykim.v@khu.ac.kr
Tel: +82-2-6405-5430
ORCID: https://orcid.org/0000-0001-7308-133X
※ 이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2020-0-00011, (전문연구실)기계를 위한 영상부호화 기술).
※ 이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00155911, 인공지능융합혁신인재양성(경희대학교)).
· Manuscript April 7, 2025; Revised May 7, 2025; Accepted May 8, 2025.

Ⅰ. 서 론

이미지 및 비디오 코딩의 중요성이 증가함에 따라, ISO/IEC JTC1/SC29 MPEG에서는 HEVC[1], VVC[2]와 같은 다양한 비디오 코덱 표준을 개발해 왔다. 이러한 전통적인 압축 방식은 영상 데이터를 높은 압축률로 줄이면서도, 인간이 인지하는 시각적 품질을 최대한 유지하는 것을 목표로 설계되어 있다. 이는 대부분의 경우, 인간이 최종적으로 영상을 소비한다는 전제를 바탕으로 한 접근이다.

하지만 최근에는 딥러닝 기반 컴퓨터 비전 기술의 빠른 발전과 더불어, 스마트 시티, 영상 감시 시스템, 커넥티드 차량 등 다양한 지능형 플랫폼이 등장하고 있다. 이들 시스템에서는 이미지나 비디오 데이터가 대량으로 지속적으로 생성되며, 이를 기반으로 기계가 직접 판단하고 의사결정을 수행한다. 이러한 환경에서는 영상이 인간에게 보여지기보다는, 기계 간 통신을 통해 처리되고 활용되는 경우가 많아지고 있다. 이에 따라, 인간의 시각적 품질보다는 기계 학습 및 추론 성능을 고려한, 새로운 형태의 영상 압축 방식이 요구되고 있다.

이러한 변화에 대응하여, ISO/IEC JTC1/SC29 MPEG WG4에서는 FCM (Feature Coding for Machines)이라는 새로운 표준을 개발 중이다. FCM은 기계 시각 기반 태스크를 수행하는 기계 작업 네트워크를 전반부와 후반부로 나누고, 전반부는 센서 또는 엣지 디바이스에서 실행한 뒤, 그 중간에서 추출된 중간 특징맵을 압축하여 서버로 전송하는 방식을 채택하고 있다. 서버에서는 복원된 중간 특징맵을 입력으로 하여 후반부 기계 작업 네트워크를 수행한다.

이 접근법은 기존의 픽셀 기반 영상 대신 특징맵 자체를 압축 대상으로 삼으며, 기계 학습 태스크 성능 저하를 최소화하면서도 데이터 전송량을 줄이는 것을 목표로 한다. 또한, 영상 전체를 서버로 전송하지 않기 때문에 서버 측 연산 부담을 경감시키고, 원본 영상의 시각 정보가 포함되지 않는다는 점에서 개인정보 보호 측면에서도 장점이 있다.

FCM 표준의 성능 평가는 FCTM (Feature Compression Test Model)이라는 참조 소프트웨어를 통해 진행되고 있다. FCTM의 인코더는 고차원의 중간 특징맵을 보다 저차원으로 변환하는 특징맵 변환 (feature transform) 과정을 수행하며, 디코더는 특징맵 역변환 (feature inverse transform) 과정을 통해 이를 다시 고차원으로 복원한다. 이 과정에서 인공신경망 기반의 특징맵 변환 및 역변환 기술이 적용된다.

본 논문에서는 FCTM에서 사용되고 있는 이러한 NN 기반 feature transform 기술에 대해 구체적으로 분석하고 설명하며, 그 구조와 설계 방식, 변환의 목적 및 기여에 대해 자세히 다룬다.

Ⅱ. FCTM의 feature transform 기술 발전 흐름

본 절에서는 현재 FCTM에서 사용되는 feature transform 및 inverse transform 기술이 개발되기까지의 기술적 흐름을 소개한다. 특히, FCM에서 사용되는 feature transform 네트워크의 입력이 해상도가 서로 다른 다계층 특징맵 (multiscale feature maps)이라는 점에 주목하여, 이들 특징맵을 효과적으로 저차원의 단일 특징맵으로 융합하는 다양한 접근법들을 중심으로 설명한다.

가장 초기의 FCTM 버전에서 사용되었던 feature transform 방식은 각 계층의 특징맵 해상도를 동일하게 맞춘 후, 이들을 순차적으로 융합하고 변환하는 구조였다[3][4]. 이 방식은 해상도 통일 과정과 특징맵 변환 과정이 별도로 수행되며, 변환된 저차원 특징맵은 압축 및 전송 후, 디코더에서 inverse transform을 통해 다시 해상도 통일된 특징맵으로 복원된다. 이후, 순차적으로 각 계층의 특징맵을 되살리는 방식으로 구성되었다. 이 구조는 높은 복원 성능을 보여 FCTM의 첫 번째 공식 버전 (버전 1.0)으로 채택되었다.

그러나, 이 방식은 해상도 통일과 특징맵 융합 및 변환이라는 일련의 과정이 분리되어 수행되기 때문에 네트워크 구조가 복잡하고 연산량이 많은 단점이 존재했다. 이러한 문제를 해결하기 위해, 이후 FCTM에서는 FENet (Feature Fusion and Encoding Network)[5][6]이라는 네트워크가 제안되었으며, 이는 해상도 정규화와 특징맵 융합을 동시에

수행할 수 있도록 설계되었다. 디코더 측에서는 DRNet (Feature Decoding and Reconstruction Network)을 사용하여 특징맵 역변환 및 복원을 수행하였다.

FENet과 DRNet 구조는 feature transform 과정을 보다 효율적으로 통합함으로써 전체 네트워크의 복잡도를 크게 줄였다. 그러나, DRNet에서 수행되는 feature inverse transform 과정은 여전히 높은 복잡도를 가지며, 연산량 및 파라미터 수 측면에서 부담이 존재했다.

이에 따라, FCTM 버전 6.0부터는 기존의 FENet 및 DRNet 구조를 경량화한 새로운 네트워크가 도입되었다. 경량화된 구조는 기존 대비 연산 복잡도를 줄이면서도 성능 저하를 최소화하는 것을 목표로 설계되었으며, 본 논문에서는 이 경량화 기술에 대한 자세한 내용을 제3장에서 설명한다.

Ⅲ. LightFCTM: 경량화된 feature transform 네트워크

본 장에서는 FCTM의 경량화 버전인 LightFCTM[7]의 구조와 설계 변경 사항에 대해 설명한다. LightFCTM은 LightFENet과 LightDRNet으로 구성되며, 각 네트워크의 구조는 그림 1에 제시된다. 본 네트워크는 다음의 세 가지 주요 변경을 통해 기존 대비 연산 및 메모리 복잡도를 현저히 감소시켰다. 표 1과 표 2는 LightFCTM의 복잡도와 성능을 나타내며 연산 복잡도 및 메모리 사용량을 약 50% 이상 절감하면서도, FCTM v5.0 대비 BD-rate를 30% 이상 개선하는 결과를 달성하였다. 또한, 입력 영상을 VVC 코덱으로 압축·복원한 뒤 기계 시각 작업을 수행했을 때의 성능인 remote-inference 대비 평균적으로 약 87%의 BD-rate 이

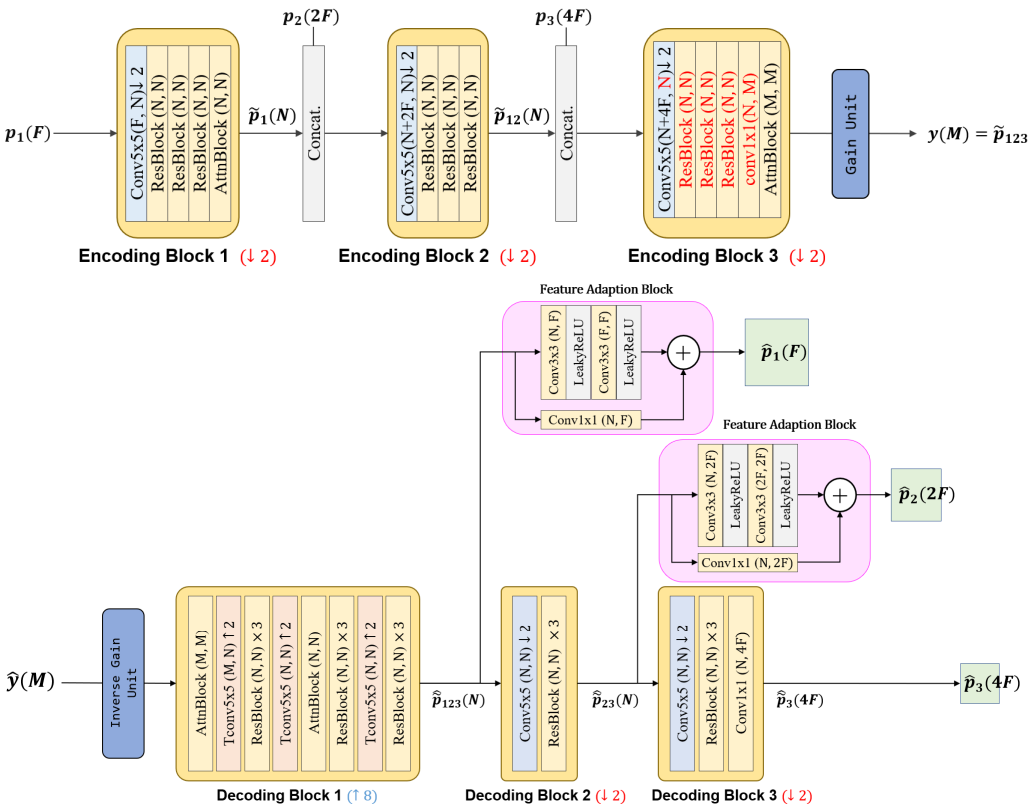


그림 1. 중간 특징맵의 계층 수가 3개일 때의 LightFCTM의 전체 구조 (위) 경량화된 feature transform 네트워크인 Light FENet (아래) 경량화된 feature inverse transform 네트워크인 Light DRNet

Fig. 1. Overall architecture of LightFCTM when the number of intermediate feature layers is three (Top) Lightweight feature transform network, Light FENet (Bottom) Lightweight feature inverse transform network, Light DRNet

표 1. 기존 feature transform 네트워크 (FENet)와 feature inverse transform 네트워크 (DRNet) 대비 LightFCTM의 연산 및 메모리 복잡도 비교
Table 1. Comparison of computational and memory complexity between the original feature transform/inverse transform networks (FENet/DRNet) and the proposed LightFCTM

		Encoder Complexity reduction rate (%)		Decoder Complexity reduction rate (%)	
		weights	kMac/pixel	weights	kMac/pixel
Instance Segmentation	OIV6	-51.92	-42.24	-70.62	-35.49
Object Detection	OIV6				
	SFU A/B SFU C SFU-D				
Tracking	TVD	-54.44	-46.42	-78.69	-47.51
	HiEve 1080p HiEve 720p	-55.78	-49.34	-71.41	-50.96
	OVERALL	-53.20	-43.40	-72.21	-37.66

표 2. FCTM 버전 5.0 대비 LightFCTM의 성능 비교
Table 2. Performance comparison between FCTM version 5.0 and LightFCTM

		proposal vs FCTM-v5.0 anchor			proposal vs Remote inference
		BD-rate	EncR	DecR	BD-rate
Instance Segmentation	OpenImageV6	-1.63%	90.879%	100.941%	-94.30%
Object detection	OpenImageV6	-24.74%	52.205%	99.179%	-94.86%
	SFU (Class A/B)	-45.55%	65.894%	87.848%	-49.92%
	SFU (Class C)	-39.53%	57.499%	74.599%	-88.59%
	SFU (Class D)	-63.58%	51.127%	87.542%	-90.18%
Object Tracking	TVD (OVERALL)	-38.21%	60.989%	59.570%	-94.81%
	HIEVE (1080p)	-7.72%	66.481%	61.279%	-91.53%
	HIEVE (720p)	-21.44%	66.382%	67.227%	-91.61%
	OVERALL	-30.30%	63.099%	84.947%	-86.97%

득을 얻었다. 다만, 객체 분할 작업에서는 이미 remote-inference 대비 매우 높은 성능을 달성하였기 때문에, 상대적으로 적은 성능 향상을 보였다.

표 1은 FCTM v5.0 대비 LightFCTM의 인코더와 디코더의 복잡도 감소량을 보여준다. 복잡도 측정 방식은 메모리 복잡도인 weights와 연산량 복잡도인 kMac/pixel이 있다. Weights는 모델의 총 파라미터 수를 나타내며, Mac (Multiply accumulate)은 곱셈-덧셈 연산의 수를 의미한다. 따라서 kMAC/pixel은 입력 이미지 한 픽셀당 모델이 몇 번의 곱셈-덧셈 연산을 수행하는지를 나타낸다. LightFCTM은 FCTM v5.0 대비 메모리 및 연산 복잡도를 약 50% 감소시켰다.

그림 3은 FCTM v5.0, LightFCTM, 그리고 remote-inference의 인코딩 및 디코딩 시간 비교 결과를 보여준다. 인코딩 시간 측면에서 remote-inference는 원본 영상을 VVC 코덱으로 압축하므로 상당한 시간이 소요되며, 이는 축소된 특징맵만을 VVC로 압축하는 FCTM 기반 방식들과 비교할 때 현저히 긴 시간임을 확인할 수 있다. 특히 LightFCTM은 연산 복잡도 감소에 따라 FCTM v5.0 대비 인코딩 시간이 추가로 감소한 것으로 나타난다. 반면, 디코

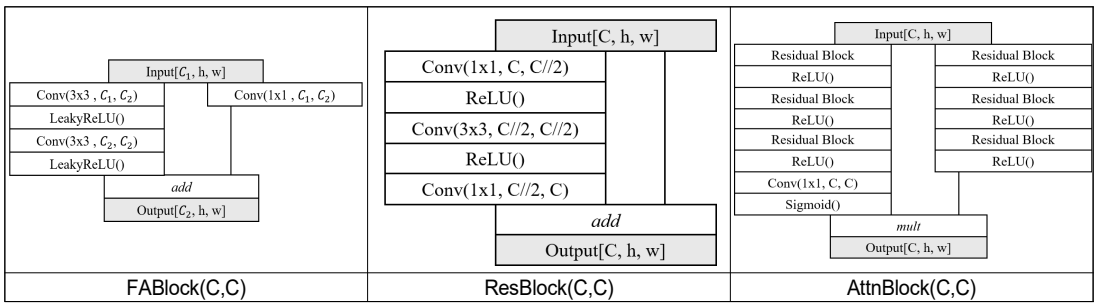


그림 2. LightFCTM에서 사용되는 기본 블록 구조들
Fig. 2. Basic block architectures used in LightFCTM

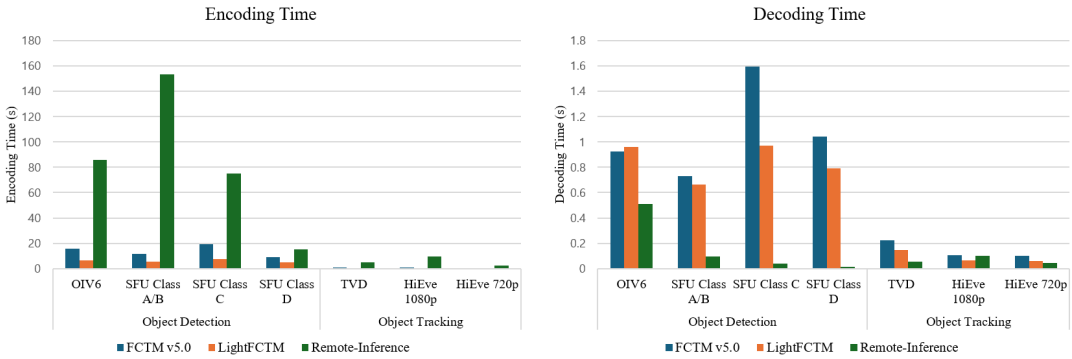


그림 3. FCTM v5.0, LightFCTM, Remote-inference의 인코딩 및 디코딩 소요 시간 비교
Fig. 3. Comparison of encoding and decoding times for FCTM v5.0, LightFCTM, and remote inference

딩 시간 측면에서는 remote-inference가 가장 적은 시간을 소요하며, FCTM v5.0보다 LightFCTM이 전반적으로 더 적은 디코딩 시간이 소요됨을 보인다. FCTM의 시나리오상 인코더는 에지 디바이스에, 디코더는 서버에 배치될 가능성이 크므로, 인코딩 시간의 절감이 더욱 중요한 요소임을 고려할 때 LightFCTM의 장점이 두드러진다.

1. 구조적 경량화: LightFENet의 마지막 블록 개선

LightFENet에서는 기존 FENet에서 사용되었던 변환 블록 구조 중 마지막 블록을 변경하였다. 기존 구조에서는 마지막 블록에만 residual block이 생략되어 있었으나, LightFENet에서는 일관성을 유지하기 위해 세 개의 residual block을 추가로 삽입하였다. 해당 변경은 다음과 같은 이점을 제공한다. 첫째, 마지막 블록의 깊이를 증가시킴으로써 다계층 특징맵의 융합 및 변환 성능을 향상시킨다. 둘째, 해당 블록은 입력 채널 수보다 출력 채널 수가 큰 구조를 가지므로, residual block은 중간 채널로의 변환 기능도 수행한다. 이를 통해 기존 convolution layer의 필터 수를 줄일 수 있었으며, 결과적으로 연산 복잡도는 소폭 감소하였다.

2. 복원 방식 변경: 병렬에서 순차 구조로의 전환

기존 DRNet은 각 계층의 특징맵을 병렬적으로 복원하는 구조였으나, LightDRNet에서는 FENet과 유사하게 순차적 복원 방식을 채택하였다. 공통된 중간 잠재 표현을 충분한 깊이의 블록을 통해 생성한 후, 이를 활용하여 각 계층의 특징맵을 순차적으로 복원한다. 또한, 각 계층 복원 시 특징맵 적응 블록 (Feature Adaptation Block)을 도입함으로써, 공통 잠재 표현으로부터 계층별 복원에 필요한 정보를 효과적으로 분리할 수 있도록 하였다. 이와 같은 구조 변경은 성능 저하 없이 DRNet의 전체 복잡도를 효과적으로 감소시켰다.

3. 채널 수 축소에 의한 복잡도 감소

FENet과 DRNet에서 공통적으로 사용되는 내부 채널 수와, FENet의 출력이자 DRNet의 입력으로 사용되는 최종 잠재 표현의 채널 수를 줄였다. 기존 구조에서는 내부 채널 수 (N) 192, 잠재 표현 채널 수 (M) 320을 사용하였으나, LightFCTM에서는 각각 128, 192로 축소하였다.

4. 중간 특징맵 계층 개수에 따른 LightFCTM의 구조

그림 1은 중간 특징맵의 계층 개수가 3개일 때의 LightFCTM 구조를 나타낸 예시이다. LightFCTM은 계층 수에 따라 유연하게 구조를 조정할 수 있으며, 본 절에서는 계층 수가 L, 입력 특징맵의 채널 수가 F의 배수, 내부 채널 수가 N, 잠재 표현 채널 수가 M일 때의 일반화된 구조를 기술한다. 특히, feature inverse transform 과정에서 사용되는 L개의 연속적인 Decoding Block으로 구성된다. 첫 번째 Decoding Block은 입력 텐서의 공간 해상도를 2^L배 상향 조정하는 역할을 한다. 이 블록은 채널 수 M의 입력에 대해 self-attention 연산을 수행하는 AttnBlock(M,M)으로 시작되며, 이후 5×5 커널을 사용하는 역합성곱 연산을 통해 채널 수를 N으로 변환하면서 해상도를 두 배로 확장하는 $CONV^{-1}(5\times5, M, N, 2\uparrow)$ 연산이 수행된다. 그다음으로는 잔차 학습을 위한 ResBlock(N,N)이 세 번 반복되어 특징 표현의 정제와 안정적인 학습을 도모한다. 나머지 Decoding Block들은 첫 번째 블록과는 달리 공간 해상도를 점진적으로 줄이도록 설계되었다. 각 블록은 5×5 커널을 사용하는 다운샘플링 합성곱 연산 $CONV(5\times5, N, N, 2\downarrow)$으로 시작되며, 이어서 ResBlock($N$,$N$)이 세 번 반복된다. 마지막 Decoding Block은 출력 채널 수에 맞추기 위해 $CONV(1\times1, N,$마지막 계층 채널 수$)$ 연산이 추가된다. 또한, 각 Decoding Block의 출력에는 Feature Adaptation Block (FABlock)이 연결되며, 최종 블록을 제외한 모든 블록에서 사용된다. l-번째 계층에 해당하는 FABlock은 N 채널의 공통 잠재 표현을 입력으로 받아, 해당 계층의 중간 특징맵을 복원하기 위해 필요한 정보로 변환한다. 앞선 설명들에서 사용된 기본 블록들의 구조는 그림 2에 나타나 있다. 이 과정을 통해 공통 잠재 표현으로부터 각 계층의 정보를 효과적으로 분리하고 복원할 수 있다. 이러한 구조는 계층 수에 관계없이 확장 가능하며, 계층 수가 증가하더라도 일정한 규칙에 따라 네트워크 구조가 조정되어 경량성과 성능을 동시에 유지할 수 있도록 설계되었다.

Ⅳ. LightFCTM의 다양한 구조 실험

본 장에서는 LightFCTM의 다양한 복원 방식에 따른 구조적 변형과 이에 따른 복잡도 및 성능 변화에 대해 분석한다. LightFCTM은 순차적 복원 방식을 기반으로 하며, 복원되는 해상도의 순서에 따라 상향식 (bottom-up)과 하향식 (top-down) 방식으로 나뉜다. 각 복원 단계 직전에 특징맵 적응 블록 (Feature Map Adaptation Block)을 삽입함으로써 정보 손실을 최소화하고, 효과적인 특징맵 복원이 가능하도록 설계되었다. 실험 결과, 상향식 복원 방식은 하향식 방식 대비 낮은 복잡도로 더 우수한 성능을 보였으며, 특징맵 적응 블록의 깊이가 증가할수록 복원 성능이 향상되는 경향을 보였다. 또한, 특징맵 적응 블록의 깊이를 선택적으로 조절함으로써 복잡도를 효율적으로 감소시키면서도 성능 향상을 달성할 수 있었다.

1. 상향식 및 하향식 복원 방식

순차적 복원 방식은 특징맵을 복원하는 순서에 따라 크게 상향식 복원 방식과 하향식 복원 방식으로 구분된다. 그림 4에서는 입력 특징맵의 채널 수가 F의 배수, 내부 채널 수가 N, 잠재 표현 채널 수가 M일 때의 상향식 및 하향식 복원 방식을 나타낸다.

상향식 복원 방식은 가장 큰 해상도의 특징맵부터 복원을 시작하며, 이후 점차 더 낮은 해상도의 특징맵을 순차적으로 복원해 나간다. 이 방식에서는 먼저 Decoding Block 1을 통해 가장 큰 해상도 크기만큼 특징맵을 복원한 후, Decoding Block 2와 3을 거치며 낮은 해상도 크기들의 특징맵들을 차례로 복원한다. 이때 각 복원 단계 앞에는 필요한 정보를 효과적으로 분리하기 위해 특징맵 적응 블록이 삽입된다. 반면, 하향식 복원 방식은 가장 낮은 해상도 크기

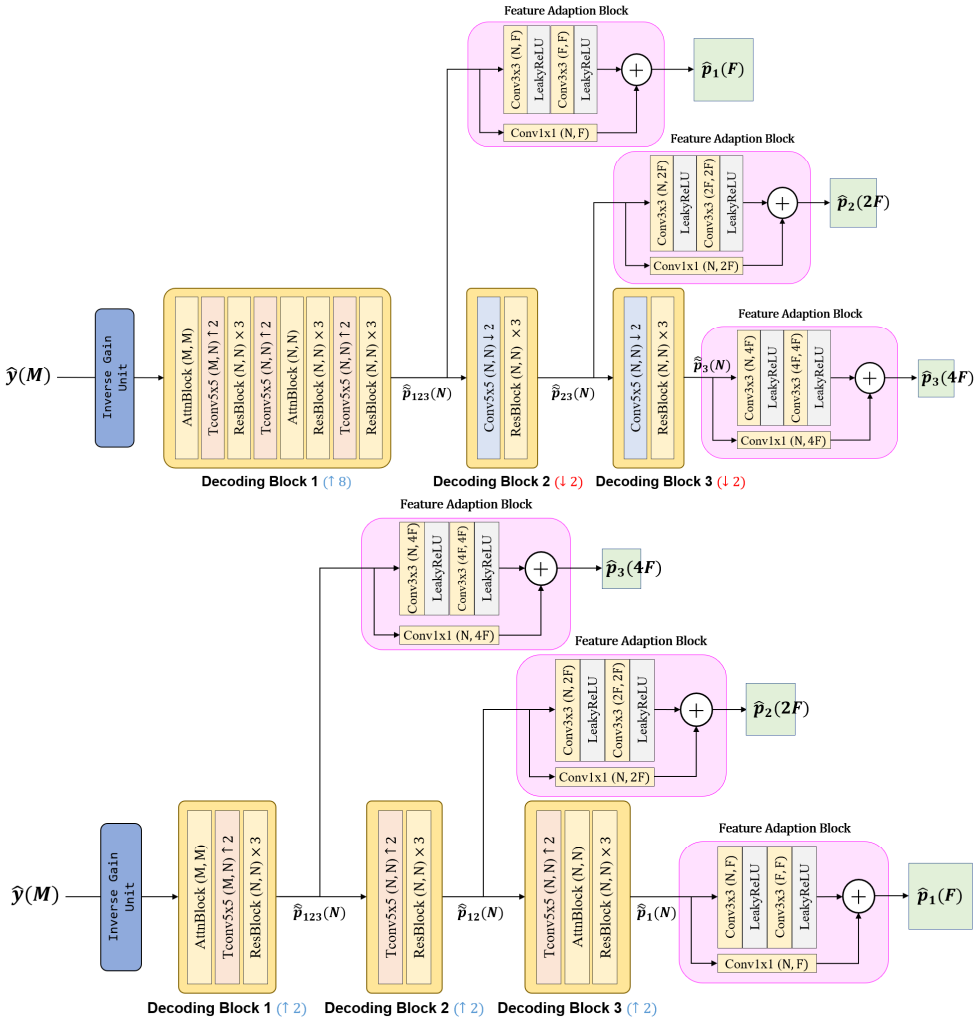


그림 4. 중간 특징맵의 계층 수가 3개일 때의 순차적 상향식 및 하향식 복원 블록 구조 (위) 순차적 상향식 복원 블록 구조 (아래) 순차적 하향식 복원 블록 구조

Fig. 4. Structure of the sequential top-down restoration block when the number of intermediate feature map layers is three (Top) structure of the sequential bottom-up restoration block (Bottom) structure of the sequential top-down restoration block

의 특징맵부터 복원을 시작하여 점차 높은 해상도로 진행된다. Decoding Block 1, 2, 3을 순차적으로 거치며 각 해상도의 특징맵을 복원하고, 이 역시 각 단계마다 특징맵 적응 블록을 통해 정보 손실 없이 복원이 이루어진다.

2. 가벼운 특징맵 적응 블록

복원 블록 내에서의 특징맵 적응은 복원 과정에서 정보 흐름을 원활하게 유지하기 위해 필수적인 요소이다. 그러나 그림 4에 제시된 기존의 특징맵 적응 블록은 비교적 높은 연산 복잡도를 가지는 구조이다. 이에 따라, 본 연구에서는 해당 블록을 단순한 채널 변환용 구조로 대체한 경량화 모델에 대한 실험도 함께 수행하였다. 예를 들어, 복잡한 구조 대신 1×1 convolution과 같은 최소한의 연산만을 적용하는 방식으로 복원 성능과 복잡도 간의 균형을 평가하였다. 이에 대한 구조는 그림 5에 나타나 있다.

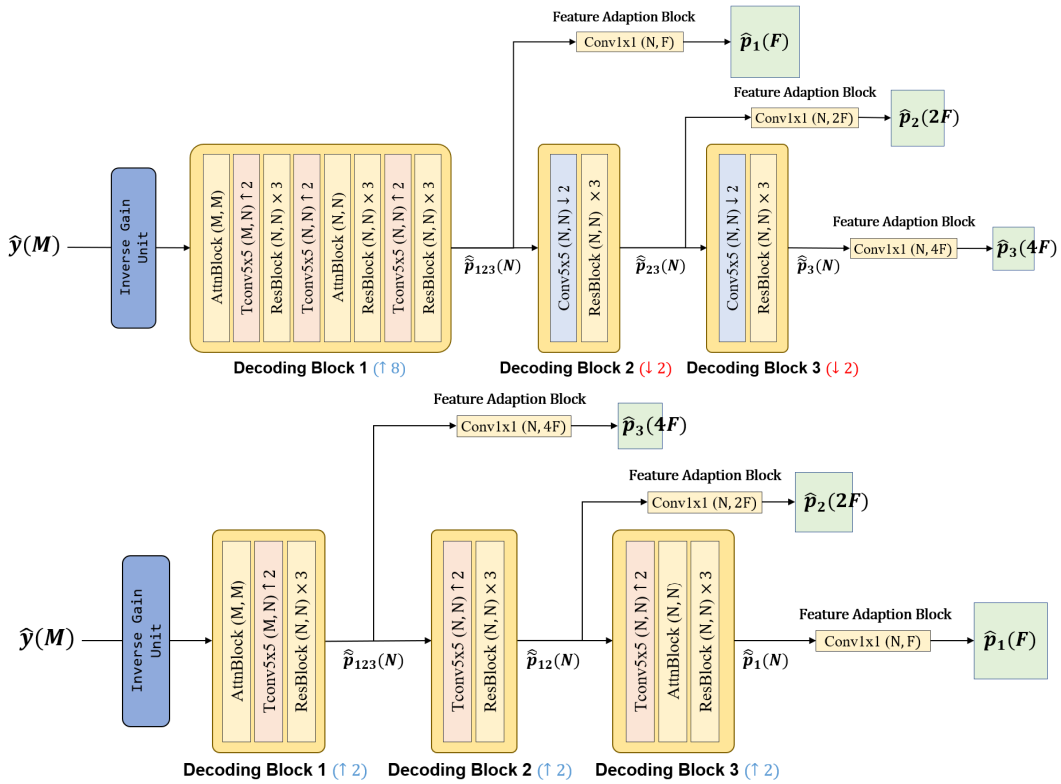


그림 5. 중간 특징맵의 계층 수가 3개일 때, 가벼운 특징맵 적응 블록을 적용한 순차적 복원 블록 구조. (위) 상향식 복원 블록. (아래) 하향식 복원 블록

Fig. 5. Structure of the sequential restoration block with lightweight feature map adaptation blocks when the number of intermediate feature map layers is three. (Top) Bottom-up restoration block. (Bottom) Top-down restoration block

3. 복잡도 및 성능 비교

다양한 비교 실험을 통해 LightFCTM의 복원 구조에는 상향식 복원 블록이 최종적으로 채택되었다. 다만, 모든 특징맵에 동일한 깊이의 특징맵 적응 블록이 적용된 것은 아니며, 그림 1의 하단 구조와 같이 마지막 특징맵에 대해서는 단순한 1×1 convolution 기반의 경량 적응 블록만이 적용되었다. 표 3은 기존 객체 추적 모델의 복원 구조와 비교하여, 제안된 모델들의 메모리 사용량 및 연산 복잡도 측면에서의 복잡도 감소율을 보여준다. 기존 복원 구조에서 내부 채널 수 및 잠재 표현 채널 수를 감소시킨 모델은 각각 19.13%, 27.69%의 감소율을 보였다. 기존 복원 구조는 병렬 복원 방식을 사용하므로, 복잡도 감소 효과는 제한적이다. 반면, 순차적 복원 방식을 적용한 모델에서는 상향식과 하향식 모두 약 50%의 메모리 사용량 감소, 약 30%의 연산

표 3. 중간 특징맵의 계층 수가 3개인 객체 추적 복원 모델의 복잡도 비교

Table 3. Complexity Comparison of Object Tracking Reconstruction Models with Three Intermediate Feature Map Layers

Restoration Models	Model size reduction rate (%, #params)	Computational cost reduction rate (%, GMac)
Original DRNet (128, 192)	-19.13	-27.69
Top-down DRNet	-54.19	-33.22
Top-down DRNet (Light FABlocks)	-91.61	-82.29
Bottom-Up DRNet	-51.24	-34.70
Bottom-Up DRNet (Light FABlocks)	-88.66	-78.85
Bottom-Up DRNet (Selective application of FABlocks, LightDRNet)	-78.73	-47.51

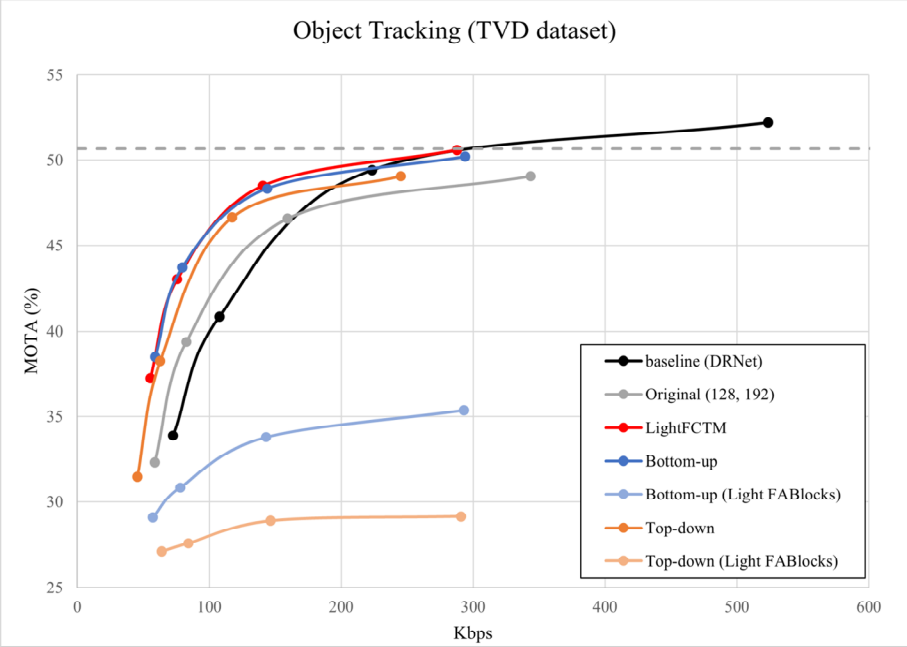


그림 6. 각 모델의 객체 추적에서의 Rate & Performance 그래프
Fig. 6. Rate and performance graph of each model on object tracking

량 감소를 나타냈다. 여기에 가벼운 특징맵 적응 블록을 적용한 경우, 두 방식은 각각 약 90%, 80%에 달하는 감소율을 보였다.

최종적으로, LightFCTM에 적용된 선택적 특징맵 적응 기반의 상향식 복원 모델은 78.73%의 메모리 사용량 감소와 47.51%의 연산량 감소를 달성하였다.

그림 6은 각 복원 모델에 따른 객체 추적 성능과 비트레이트 간의 관계를 나타낸 Rate & Performance 그래프를 보여준다. 이때 수평 점선은 미분할 성능 (unsplit performance)을 의미하며, 이는 기계 작업 네트워크의 원래 성능이자 FCM 구조에서 달성 가능한 성능의 상한선을 나타낸다. Baseline은 기존 DRNet 구조를 의미하며, 이외의 모든 모델들은 내부 채널 수 N과 잠재 표현 채널 수 M을 각각 128과 192로 감소시킨 경량화 모델들이다. 이 중, 기존 DRNet 구조에서 단순히 채널 수만 축소한 모델인 Original (128, 192)은 baseline 대비 일부 향상된 성능을 보였으나, 순차적 복원 방식이 적용된 상향식 (Bottom-up) 및 하향식 (Top-down) 복원 모델들은 구조적 변화를 통해 채널 수가 줄어든 상태에서도 추가적인 성능 향상을 달성하였다. 특히 상향식 복원 모델이 하향식보다 더 높은 성능을 보였는데, 이는 상향식 복원이 정보량이 가장 많은 복원된 잠재 표현으로부터 해상도가 가장 큰 특징맵을 우선 복원하기 때문으로 판단된다. 해상도가 높은 특징맵은 보다 정교한 객체 정보를 포함하고 있어, 객체 추적 성능에 긍정적인 영향을 미친다. 한편, 가벼운 특징맵 적응 블록이 적용된 모델들은 기존 모델 대비 복잡도가 가장 크게 감소하였으나, 깊은 특징맵 적응 블록이 적용된 모델에 비해 성능 저하가 상대적으로 크게 나타났다. 반면, LightFCTM에 적용된 선택적 특징맵 적응 기반의 상향식 복원 모델은 일반 상향식 모델과 유사한 성능을 유지하면서도, 복잡도를 더욱 효과적으로 감소시킨 것이 확인되었다.

V. 결론 및 향후 연구

본 논문에서는 기존 feature transform 및 inverse transform 네트워크의 구조적 복잡도를 해소하고 효율성을 개선하기 위한 경량화 구조인 LightFCTM을 자세히 설명했다.

LightFCTM은 기존 FENet과 DRNet의 구조를 분석하고, 세 가지 주요 변경 사항을 통해 연산량과 메모리 사용량을 효과적으로 감소시켰다. 첫째, FENet의 마지막 블록에 Residual block을 도입함으로써 성능을 유지하거나 향상시키는 동시에 연산 효율을 개선하였다. 둘째, DRNet은 기존 병렬 복원 방식 대신 순차적인 복원 방식으로 재설계되었으며, Feature Adaptation Block을 도입하여 복원 정확도를 향상시켰다. 셋째, 내부 채널 수와 잠재 표현 채널 수의 감소를 통해 전반적인 모델 경량화를 달성하였다. 실험을 통해 LightFCTM이 기존 FCTM 대비 연산 및 메모리 복잡도를 절반 수준으로 낮추면서도 평균 30% 이상의 BD-rate 절감 성능을 기록함을 확인하였다.

향후 연구에서는 LightFCTM이 작은 객체가 많은 영상에서 충분한 성능을 보이지 못하는 한계를 극복하는 데 중점을 둘 것으로 예상된다. 구체적으로, SFU Class A/B 중 Traffic 및 BQTerrace 시퀀스에서는 unsplit 성능이나 remote-inference 기반의 성능보다 낮은 결과를 나타내는 경우가 있었다. 여기서 unsplit 성능은 네트워크를 분할하지 않았을 때의 원래 성능을 의미하며, remote-inference 성능은 입력 영상을 VVC 코덱으로 압축·복원한 뒤 기계 시각 작업을 수행했을 때의 성능을 나타낸다. 이러한 조건에서도 LightFCTM이 안정적인 성능을 유지하도록 구조 개선 및 보조 모듈 도입에 대한 연구를 지속할 것으로 보인다.

참 고 문 헌 (References)

[1] J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1649 – 1668, Dec. 2012. doi: https://doi.org/10.1109/TCSVT.2012.2221191

[2] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, “Overview of the Versatile Video Coding (VVC) standard and its applications,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 10, pp. 3736 – 3764, Oct. 2021.

[3] Z. Zhang, M. Wang, M. Ma, J. Li, and X. Fan, “MSFC: Deep feature compression in multi-task network,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Shenzhen, China, Jul. 2021. doi: https://doi.org/10.1109/ICME51207.2021.9428258

[4] H. Han et al., “[FCVCM] Response to CfP : Enhanced Multi-scale Feature Compression for FCVCM”, ISO/IEC JTC 1 / SC 29 / WG 04 Doc. m65217, Oct. 2023.

[5] Y. Kim et al., “End-to-End Learnable Multi-Scale Feature Compression for VCM.” IEEE Transactions on Circuits and Systems for Video Technology, early access, Aug. 2023. doi: https://doi.org/10.1109/TCSVT.2023.3302858

[6] H. Jeong et al., “[FCVCM] Hybrid codec approach : Combination of L-MSFC-v2 Intra (m65200) with VVC” ISO/IEC JTC 1 / SC 29 / WG 04, Doc. m65202, Oct. 2023.

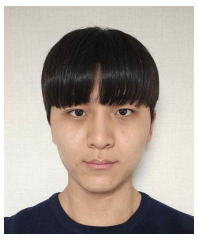
[7] D. Lim et al., “[FCM][CE1-related] LightFCTM: Small Modifications on FE/DRNet for lower complexity with better performance”, ISO/IEC JTC 1 / SC 29 / WG 04, Doc. m70122, Nov. 2024.

저 자 소 개

정 혜 원

- 2022년 2월 : 경희대학교 소프트웨어융합학과 학사
- 2024년 2월 : 경희대학교 컴퓨터공학과 석사
- 2024년 3월 ~ 현재 : 경희대학교 컴퓨터공학과 박사
- ORCID : http://orcid.org/0000-0003-1230-870X
- 주관심분야 : 영상처리, 비디오 부호화, 컴퓨터 비전, 머신러닝

저 자 소 개

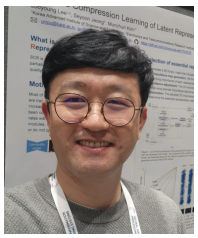


임 달 홍

- 2025년 2월 : 경희대학교 소프트웨어융합학과 학사
- 2025년 3월 ~ 현재 : 경희대학교 인공지능학과 석사
- ORCID : https://orcid.org/0009-0000-2733-5036
- 주관심분야 : 영상처리, 비디오 부호화, 컴퓨터 비전, 머신러닝

유 장 현

- 2023년 2월 : 경희대학교 전자공학과 학사
- 2025년 2월 : 경희대학교 컴퓨터공학부 석사
- 2025년 3월 ~ 현재 : 경희대학교 컴퓨터공학부 박사
- ORCID : https://orcid.org/0009-0009-1818-4315
- 주관심분야 : 비디오 부호화, 딥러닝, 컴퓨터 비전



이 주 영

- 2003년 2월 : 아주대학교 미디어학부 학사
- 2006년 2월 : KAIST 전산학과 공학석사
- 2024년 8월 : KAIST 전기및전자공학과 박사
- 2006년 ~ 현재 : 한국전자통신연구원(ETRI) 미디어부호화연구실 선임연구원
- ORCID : https://orcid.org/0000-0003-0753-0699
- 주관심분야 : 인공지능, 컴퓨터 비전, 생성 모델, 이미지/비디오 압축

김 휘 용

- 1994년 8월 : KAIST 전기및전자공학과 공학사
- 1998년 2월 : KAIST 전기및전자공학과 공학석사
- 2004년 2월 : KAIST 전기및전자공학과 공학박사
- 2003년 8월 ~ 2005년 8월 : ㈜애드팍테크놀러지 멀티미디어팀 팀장
- 2005년 11월 ~ 2019년 8월 : 한국전자통신연구원(ETRI) 실감AV연구그룹 그룹장
- 2013년 9월 ~ 2014년 8월 : Univ. of Southen California (USC) Visiting Scholar
- 2019년 9월 ~ 2020년 2월 : 숙명여자대학교 전자공학전공 부교수
- 2020년 3월 ~ 2024년 12월 : 경희대학교 컴퓨터공학과 부교수
- 2025년 1월 ~ 현재 : 경희대학교 컴퓨터공학과 정교수
- ORCID : https://orcid.org/0000-0001-7308-133X
- 주관심분야 : 비디오 부호화, 딥러닝, 영상처리, 디지털 홀로그램