특집논문 (Special Paper)
방송공학회논문지 제29권 제4호, 2024년 7월 (JBE Vol.29, No.4, July 2024)
https://doi.org/10.5909/JBE.2024.29.4.408
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

특징 축소 네트워크 및 VVC를 활용한 효율적인 3D INVR 모델 압축 기법

김 동 하[a], 문 현 철[a], 정 준 영[b], 이 광 순[b], 김 재 곤[a)‡]

An Efficient Compression Method of 3D INVR Model Utilizing Feature Reduction Network and VVC

Dong-Ha Kim[a], Hyeon-Cheol Moon[a], Junyoung Jeong[b], Gwangsoon Lee[b], and Jae-Gon Kim[a)‡]

요 약

암시적 신경망 표현(Implicit Neural Representation: INR)은 멀티뷰(multi-view) 영상이나 비디오로 MLP(Multi-Layer Perceptron) 신경망을 학습하고 새로운 뷰를 합성할 수 있도록 한다. 또한, 최근에는 렌더링(rendering) 성능을 개선하거나 학습 및 렌더링을 가속화하기 위해 명시적 표현인 복셀 그리드(voxel grid)를 혼용한 하이브리드(hybrid) 모델이 사용되고 있다. MPEG(Moving Picture Experts Group)의 INVR(Implicit Neural Visual Representation) AhG(Ad-hoc Group)은 암시적/명시적 신경망 표현 및 3DGS(3D Gaussian Splatting)의 명시적 표현 모델을 사용하여 효과적으로 3D 공간 영상/비디오를 표현 및 압축하기 위한 잠재적 표준기술을 탐색하고 있다. 본 논문은 하이브리드 정적 3D 모델인 TensoRF 모델의 압축 기법을 제시한다. 제안기법은 TensoRF의 명시적 표현인 텐서(tensor) 평면을 FRN(Feature Reduction Network)을 통해 잠재(latent) 평면으로 축소하고, 축소된 잠재 평면들로 구성된 특징맵을 VVC로 압축한다. 기존의 3D 비디오 압축 표준인 MIV(MPEG Immersive Video)의 시험모델 TMIV(Test Model for Immersive Video)의 DSDE(Decoder Side Depth Estimation) 모드 대비 우수한 압축 성능을 보인다.

Abstract

The Implicit Neural Representation (INR) enables training of Multi-Layer Perceptron (MLP) neural networks on multi-view images or videos, allowing synthesis of novel views. Recently, hybrid models, which combine explicit representation, voxel girds, with MLP have been used to improve rendering quality and/or accelerate learning and rendering processes. MPEG's Ad-hoc Group (AhG) of Implicit Neural Visual Representation (INVR) explores potential standardization technologies for efficiently representing and compressing 3D volumetric images/videos using implicit/explicit neural network representations and explicit representations models. This paper presents a compression method for the hybrid static 3D model, TensoRF. The proposed method compresses the explicit representation of TensoRF, tensor planes, through feature reduction network (FRN) into latent planes, and compresses the resulting feature maps using VVC. It demonstrates superior compression performance compared to the Decoder Side Depth Estimation (DSDE) mode of Test Model for Immersive Video (TMIV) of the existing 3D video compression standard, MPEG Immersive Video (MIV).

Keyword : Implicit Neural Visual Representation (INVR), Implicit Neural Representation (INR), TensoRF, Feature Reduction Network (FRN), Versatile Video Coding (VVC)

Ⅰ. 서 론

최근 영상/비디오를 신경망 학습을 통하여 표현하는 암시적 신경망 표현(Implicit Neural Representation: INR)[1]이 새로운 2D/3D 비디오의 표현 및 압축 기법으로 제시되고 있다. INR은 멀티뷰(multi-view) 영상이나 비디오를 MLP(Multi Layer Perceptron) 신경망으로 학습하고 학습된 신경망으로 새로운 뷰의 합성을 가능하게 한다. 또한, INR은 합성된 새로운 뷰의 렌더링(rendering) 성능을 개선하고 학습 및 렌더링을 가속화하기 위해 복셀 그리드(voxel grids)의 명시적(explict) 표현을 혼용한 하이브리드(hybrid) 방법을 사용한다[2][3]. 더 나아가 명시적 표현만을 사용해서 3D 비디오를 효과적으로 표현하는 3DGS (3D Gaussian Splatting)도 연구 중이다[4]. 한편, MPEG에서는 이러한 학습 기반의 명시적/암시적 신경망 표현 및 명시적 표현을 2D/3D 비디오를 표현/압축하는 새로운 접근방법으로 보고 INVR(Implicit Neural Visual Representation) AhG(Ad-hoc Group)[5]을 구성하고 그 표준화 가능성 및 후보 표준기술을 탐색하고 있다.

INVR AhG은 두 개의 탐색실험(Exploration Experiment: EE)[6]을 설정하고 학습 기반의 3D 비디오 표현/압축 기술 탐색을 진행하고 있다. EE2는 멀티뷰로 구성된 6DoF (Degree of Freedom) 몰입형 비디오를 3D INVR 모델로 학습하고 압축하는 기술에 대한 것이다. 탐색실험에서는 기존의 6DoF 몰입형 비디오 압축 표준인 MIV(MPEG Immersive Video)[7]의 시험모델인 TMIV(Test Model for Immersive Video)[8]의 압축 성능을 기준으로 더 높은 성능을 요구한다. EE2.1은 명시적 및 암시적 표현을 포함하는 하이브리드 INVR 모델[2][3]을, EE2.2는 3D Gaussian을 활용해 3D 비디오를 명시적으로 표현하는 기법인 3DGS를 범위로 하여 기술 탐색을 진행한다.

본 논문은 INVR EE2.1에 해당하는 하이브리드 3D INVR 모델인 TensoRF[3] 모델의 압축 기법을 제시한다. TensoRF는 명시적 표현인 복셀 그리드를 텐서 평면(tensor planes)과 텐서 벡터(tensor vectors)로 분해하여 텐서-평면과 벡터로 표현함으로써 메모리 사용량을 줄일 수 있다. 또한, MLP만을 활용한 암시적 신경망 표현 모델인 NeRF(Neural Radiance Field)[9]보다 우수한 렌더링 성능을 갖는다. 하지만, 멀티뷰 영상을 학습한 TensoRF[2]의 모델 크기는 여전히 180MB 정도로 큰 저장 공간을 요구한다. 또한, 텐서 평면이 차지하는 크기는 전체 모델 크기의 99.2% 정도로 학습된 모델의 대부분을 차지한다. 따라서, 명시적인 표현인 텐서 평면을 압축하여 학습된 TensoRF를 효율적으로 압축하는 기법이 요구된다.

본 논문에서는 그림 1의 전체 구조와 같이 특징 축소 네트워크(Feature Reduction Network: FRN)를 통하여 TensoRF의 명시적 표현인 텐서 평면을 잠재 평면(latent plane)으로 축소하고, 이로 구성된 특징맵을 VVC로 압축하는 효율적인 TensoRF 압축 기법을 제시한다[10]. 추가적으로 명시적 표현뿐만 아니라 TensoRF의 암시적 표현인 MLP 네트워크를 NNC(Neural Network Compression) 표준 코덱인 NNCodec[11]으로 압축한다. INVR의 EE2.1의 공통실험조건(CTC: Common Test Condition)에 따라 멀티뷰 비디오로 구성된 MIV 시퀀스 중 Mirror[7]를 활용하여 TensoRF를 학습하고, 제안기법으로 압축한 결과를 기존 3D 비디오 압축 표준인 MIV의 결과와 비교한다. 또한, TensoRF의 축소하지 않은 텐서 평면으로 특징맵을 구성하고 VVC로 압축한 기존의 TensoRF 압축 결과와 비교한다[12]. 제안기법은 비교한 기존의 방법들 보다 우수한 압축 대비 렌더링 성능을 보인다.

본 논문은 2장에서 3D 암시적 신경망 표현인 NeRF와 명시적 표현을 포함하는 하이브리드 모델인 TensoRF에 대해 소개하고, 3장에서 FRN과 VVC를 활용한 제안한 TensoRF 압축 기법을 상세히 설명한다. 4장에서 제안기법의 실험결과를 기술하고, 5장에서 결론을 맺는다.

a) 한국항공대학교 항공전자정보공학과(Department Electronics and Information Engineering, Korea Aerospace University)
b) 한국전자통신연구원(ETRI)
‡ Corresponding Author : 김재곤(Jae-Gon Kim)
E-mail: jgkim@kau.ac.kr
Tel: +82-2-300-0414
ORCID: https://orcid.org/0000-0003-3686-4786
※ This work was supported by IITP grant funded by the Korea government (MSIT) (No. 2018-0-00207, Immersive Media Research Laboratory).
· Manuscript May 12, 2024, 2024; Revised June 14, 2024; Accepted June 14, 2024.

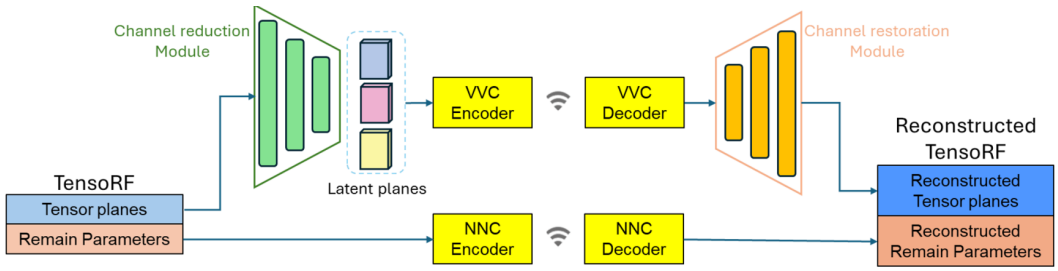

그림 1. TensoRF 압축을 위한 제안기법의 전체구조[10]
Fig. 1. Overall architecture of the proposed TensoRF compression method[10]

II. NeRF 및 TensoRF

1. 암시적 신경망 표현 NeRF

NeRF[9]는 3D 공간을 표현하는 대표적인 암시적 신경망 표현 모델이다. NeRF는 3D 공간의 여러 위치와 각도에서 촬영한 멀티뷰 영상을 활용하여 MLP 신경망 기반으로 3D 공간 영상을 표현한다. 학습된 MLP 신경망은 3D 공간 좌표와 카메라 방향에 해당하는 5개의 변수를 입력받아 해당 좌표의 색상값과 투명도를 나타내는 밀도값을 예측한다.

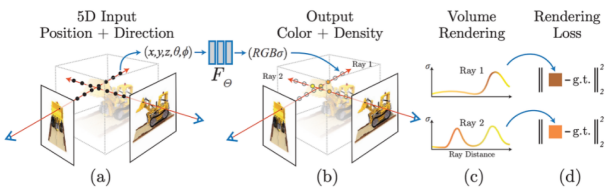

그림 2. NeRF의 학습 방법[9]
Fig. 2. Training method of NeRF[9]

그림 2는 멀티뷰 영상을 활용하여 NeRF의 신경망 학습 방법을 간략하게 나타낸 것이다[9]. 그림 2 (a)와 같이 멀티뷰 영상의 카메라 방향으로의 레이(ray) 위의 여러 샘플 포인트에 대해, 그림 2 (b)와 같이 신경망을 통하여 색상값과 밀도값을 예측한다. 이때 그림 2 (c)와 같이 빈 공간에 해당하는 포인트의 경우 밀도의 값은 낮고, 객체에 해당하는 포인트는 높은 밀도값을 가진다. 모델링된 함수를 통해 예측된 레이위의 샘플들의 색상값과 밀도값을 카메라 방향의 뷰로 투영(projection)하여 화소의 최종 RGB 값을 계산한다. 결국, 레이는 멀티뷰 영상을 구성하는 모든 화소에서 카메라 방향으로 나아가, 신경망을 통해 모든 레이 위의 샘플 포인트에 대해 색상값과 밀도값을 예측하고, 구성된 투영 계산식을 통해 각 화소 값을 도출한다. 이렇게 그림 2 (d)와 같이 생성된 뷰와 주어진 멀티뷰 영상 간의 차이를 최소화하도록 신경망을 학습한다.

2. 명시적 표현을 포함하는 TensoRF

NeRF는 암시적 신경망 표현을 사용하여 3D 영상 공간을 효과적으로 모델링하지만, MLP를 통해 모든 레이 위의 샘플 포인트에 대해 색상값과 밀도값을 예측해야 하는 복잡도 측면에서 단점이 있다. 즉, 새로운 뷰를 합성하기 위해 해당 영상을 구성하는 모든 화소에 대응하는 레이에 대해 복잡한 MLP 연산이 요구된다. 이러한 문제점을 해결하기 위해, 명시적 표현을 동시에 사용하는 하이브리드 모델 방법들이 연구되었으며 그림 3은 명시적 표현의 활용 방법의 예시이다[3].

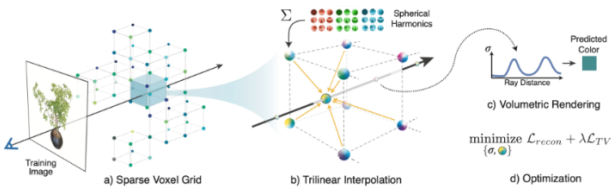

그림 3. 명시적 표현의 예시[3]
Fig. 3. An example of explicit representation[3]

명시적 표현은 주로 3D 공간을 일정한 크기의 정육면체로 분할하여 구성된 복셀 그리드를 통해 표현된다. 명시적 표현을 암시적 신경망 표현과 함께 활용하는 하이브리드

모델은 복셀 그리드에 색상 및 밀도의 정보를 대략적으로 저장해두고 레이를 복셀 그리드 상에 쏜다. 레이 위의 샘플 포인트에 대한 대략적인 색상값과 밀도값을 인접한 복셀로부터 보간(interpolation)하여 예측하고, 예측된 값을 얕은 MLP에 입력하여 최종적인 샘플 포인트의 색상값과 밀도값을 예측한다. 나머지 과정은 NeRF의 방법에 따라 레이 위의 샘플들을 뷰 영상에 투영하여 새로운 뷰를 합성한다. 하이브리드 모델의 학습은 복셀 그리드와 MLP를 동시에 진행한다.

하이브리드 모델은 복셀 그리드를 활용함으로써 렌더링 성능과 속도를 개선하지만, 이러한 복셀 그리드는 많은 메모리를 필요로 한다. TensoRF[2]는 이런 단점을 해결하기 위해 복셀 그리드를 텐서 평면과 텐서 벡터로 분해하여 연산 복잡도와 메모리 사용량을 개선했다.

그림 4는 TensoRF의 구성 및 학습 방법을 나타낸 것이다. TensoRF의 분해된 텐서 평면과 벡터를 외적하면 복셀 그리드를 복원할 수 있으며, 각 평면과 벡터는 다수의 랭크(rank)로 구성된다. 각 평면과 벡터를 외적하면, 랭크 수만큼의 복셀 그리드가 생성되고, 이를 합산하여 정밀한 형태의 복셀 그리드를 얻을 수 있다.

텐서 평면은 색상값을 예측하는 외관(appearance) 평면과 밀도값을 예측하는 밀도(density) 평면으로 구분되며, 외관 평면과 밀도 평면은 각각 xy, yz, zx를 나타내는 세 개의 평면과 벡터로 구성된다. TensoRF는 총 6개의 텐서 평면을 포함하며, 각각의 평면과 벡터를 외적하여 샘플 포인트의 값을 예측한다. 외관 평면으로부터 예측된 샘플 포인트의 값들은 도출된 값을 이어서(concatenation) 신경망에 입력해 최종적인 색상값을 예측하고, 밀도 평면에서 도출된 값들은 합산하여 최종적인 밀도의 값으로 사용한다. 학습에 사용하는 멀티뷰 영상으로 투영하는 과정은 NeRF와 같다.

Ⅲ. FRN과 VVC를 활용한 TensoRF 압축

하이브리드 모델인 TensoRF는 렌더링 성능과 속도를 개선하였고, 복셀 그리드를 텐서 평면과 벡터로 분해하여 메모리 사용량을 줄였다. 하지만 여전히 큰 메모리 용량이 요구되며 텐서 평면은 전체 모델 크기의 약 99% 를 차지한다. 따라서, 학습된 TensoRF를 효율적으로 압축하기 위해서는

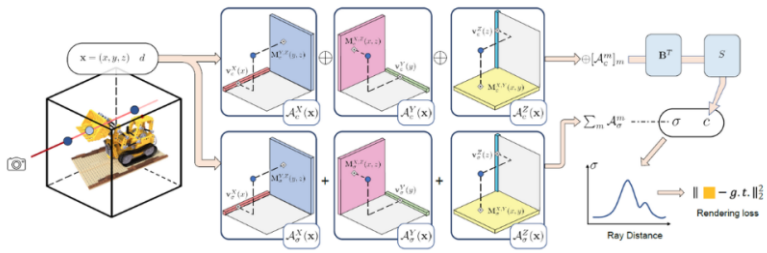

그림 4. TensoRF의 구성 및 학습 방법[2]
Fig. 4. The architecture and training method of TensoRF[2]

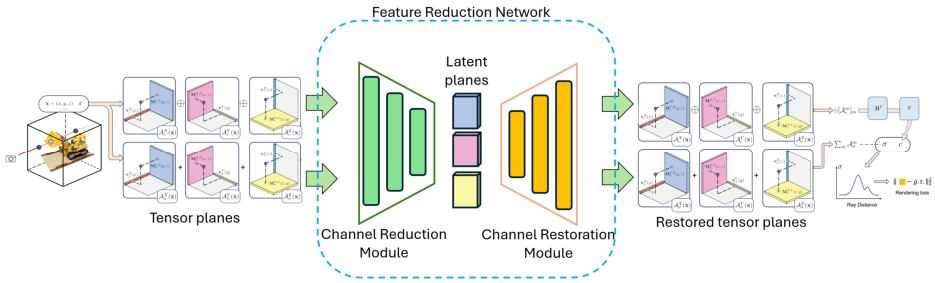

그림 5. FRN의 구조와 학습 과정[10]
Fig. 5. Structure and training process of FRN[10]

텐서 평면의 효율적인 압축이 필요하다. 본 논문은 FRN과 2D 비디오 코덱인 VVC를 활용한 TensoRF의 효율적인 압축 기법을 제안한다. 제안기법은 FRN을 활용하여 텐서 평면의 랭크 수를 축소하고, 축소된 텐서 평면으로 구성된 특징맵을 VVC로 압축한다.

1. Feature Reduction Network

그림 5는 본 논문에서 제시한 FRN의 구조와 훈련 과정을 보여준다[10]. FRN은 VVC로 압축될 특징맵의 해상도를 줄이기 위해 입력된 텐서 평면을 축소하고 복원하며, 복원된 텐서 평면은 실제 렌더링에 사용된다. FRN은 채널 축소 모듈(Channel Reduction Module)과 채널 복원 모듈(Channel Restoration Module)로 구성되어 있고, 이들 모듈은 간단한 CNN(Convolution Neural Network) 구조를 가진다. 채널 축소 모듈은 외관과 밀도 텐서 평면을 입력받아 텐서 평면의 랭크 수를 축소해 잠재 평면(latent plane)을 출력한다. 즉, 잠재 평면은 입력된 텐서 평면과 동일한 높이와 너비를 가지고 더 작은 랭크 수를 가진다. 또한, 잠재 평면은 채널 복원 모듈에 입력되어 축소 전 원래의 텐서 평면의 랭크 수로 복원된다. 마찬가지로 해당 모듈은 높이와 너비에 변화 없이 입력된 잠재 평면의 랭크 수를 본래의 텐서 평면의 랭크 수로 복원한다.

FRN은 TensoRF와 동시에 학습을 진행한다. 밀도와 외관 평면은 서로 다른 특성을 가지므로 각각 평면에 대한 FRN을 설계한다. 즉, 외관 평면과 밀도 평면에 대한 FRN은 해당 평면의 특성에 맞게 다른 구조를 가지며, 채널 축소 모듈을 통해 랭크 수가 각각 4와 2로 축소된다. 랭크 축소로 인한 오차를 최소화하기 위해 입력 및 복원된 텐서 평면 간의 MSE(Mean Square Error) 손실을 특징 손실(feature loss)로서 TensoRF의 학습을 위한 전체 손실함수에 포함한다.

2. 특징 손실함수

1절에서 언급한 바와 같이 MSE 손실은 FRN의 입력과 출력 텐서 평면 사이에 정의된다. 그러나 학습하는 멀티뷰 영상에 따라 텐서 평면의 중요도는 다를 수 있다. 예를 들면, LLFF(Local Light Field Fusion)[13] 멀티뷰 영상은 정면을 바라보는 영상만 포함하며, 이 경우 정면에 해당하는 xy 텐서 평면이 다른 텐서 평면에 비해 중요도가 높다. 따라서 TensoRF를 학습할 때, xy 텐서 평면에 더 많은 랭크 수를 할당한다. 랭크 수가 많은 텐서 평면은 렌더링 성능에 큰 영향을 미칠 수 있고, FRN은 텐서 평면을 복원할 때 랭크 수가 많은 텐서 평면을 더 정밀하게 복원해야 한다. 따라서 FRN 학습에서 중요한 텐서에 대한 특징 손실함수에 더 큰 가중치를 부여한다.

$$\begin{aligned} &feature\ loss \\ &= \omega_{xy} L_{mse}(P_{xy}, P'_{xy}) + \omega_{yz} L_{mse}(P_{yz}, P'_{yz}) + \\ &\quad \omega_{xz} L_{mse}(P_{xz}, P'_{xz}) \end{aligned} \tag{1}$$

$$1 = \omega_{xy} + \omega_{yz} + \omega_{xz},\ \omega_{xy} = \frac{R_{xy}}{R_{xy} + R_{yz} + R_{xz}} \tag{2}$$

식 (1)에서 P와 P'은 입력되는 텐서 평면과 복원된 텐서 평면을 나타낸다. L_{mse}는 입력 텐서 평면과 복원 텐서 평면 사이의 MSE 손실이다. ω_i는 각 텐서 평면의 손실에 대한 가중치를 나타내며 모든 ω_i의 합은 1이다. 식 (2)에서 R_i은 각 텐서 평면의 랭크 수를 나타낸다. 이때 각 평면의 랭크 수에 따라 가중치 ω_i의 값이 정해진다. 예를 들어 xy, yz, zx의 평면의 랭크 수가 각각 6, 2, 2라고 가정한다면 ω_{xy}, ω_{yz}, ω_{xz} 의 값은 0.6, 0.2, 0.2로 설정된다.

3. VVC를 활용한 잠재 평면 압축

텐서 평면과 FRN으로 축소된 잠재 평면은 여러 랭크로 구성되어 있고 일반적인 신경망에서 추출되는 특징과 같이 32비트의 부동 소수점의 형태이다. 본 논문에서는 잠재 평면으로 특징맵을 구성하고 이를 VVC로 압축한다. MPEG의 기계를 위한 특징 압축을 위해 개발 중인 FCM(Feature Coding for Machine)[14]의 VVC를 이용한 특징 압축 방법을 텐서 평면 및 잠재 평면 압축에 적용할 수 있다. 우선, 잠재 평면으로 주어진 특징을 VVC로 압축하기 위한 특징맵의 형태로 변환해야 한다. 특징맵 변환은 다음 두 가지 처리 과정을 포함한다. 첫 번째는 다중 채널로 구성된 특징

을 단일 프레임의 특징맵으로 패킹(packing)한다. 예를 들면 48개 채널로 각 채널을 8x6 타일 구조로 단일 프레임의 특징맵으로 구성한다. 두 번째는 패킹된 각 특징 값을 전체 특징값의 최소/최대값을 활용해 VVC 인코더의 입력 포맷인 10비트의 정수로 선형 양자화한다.

표 1은 외관 텐서 평면과 FRN을 거친 잠재 평면의 랭크 수 및 각 평면이 특징맵으로 변환될 때의 해상도를 나타낸다. 표 1과 같이, 텐서 평면의 랭크 수는 잠재 평면의 랭크 수보다 크기 때문에, 텐서 평면을 특징맵으로 변환하면 잠재 평면의 특징맵보다 더 높은 해상도를 가진다. 따라서, 그림 6의 예시와 같이, 텐서 평면의 특징맵은 여러개의 특징맵으로 구성되고 각각을 VVC로 압축할 수 있다[12]. 하지만 잠재 평면의 특징맵은 해상도가 매우 낮아 단일 프레임에 모든 특징맵을 패킹할 수 있으며 하나의 VVC 코덱으로 압축할 수 있다.

표 1. 텐서 평면과 잠재 평면의 랭크 수와 특징맵 해상도 비교
Table 1. Comparison of rank and feature map resolution between the tensor plane and latent plane

Appearance plane		Rank	Packed plane resolution (Width, Height)
Tensor plane	app_plane.0	48	5648, 4720
	app_plane.1	12	2824, 1416
	app_plane.2	12	3144, 1416
Latent plane	app_plane.0	4	1572, 1412
	app_plane.1	4	942, 1412
	app_plane.2	4	1572, 942

Ⅳ. 실험결과

본 논문의 실험에서는 INVR의 CTC에 따라 15개의 멀티뷰 정적 영상으로 구성된 Mirror 시퀀스를 사용하여 TensoRF를 학습한다. 이 과정에서, 두 개의 테스트 뷰(v06, v08)를 제외하고 나머지 영상을 사용하여 학습한다. 학습된 TensoRF 모델을 압축 및 복원 후 테스트 뷰를 렌더링하고 압축율-렌더링 화질의 압축 성능을 검증한다. 그림 1과 같이 학습된 FRN과 VVC를 활용하여 TensoRF의 명시적 표현인 텐서 평면을 잠재 평면으로 축소하고, 잠재 평면으로 구성된 특징맵을 VVC로 압축한다. 특징맵은 QP(Quantization Parameters)를 {22,27,32,37}로 설정하여 4개의 비트율 포인트에 대한 압축을 수행한다. 또한, 텐서 평면을 제외한 MLP 등 TensoRF의 나머지 매개변수를 NNCodec을 사용하여 압축한다. 성능평가는 압축된 비트열 데이터의 크기를 렌더링한 영상의 전체 픽셀로 나눈 BPP(Bits Per Pixel)와 테스트 뷰에 대한 렌더링 화질인 PSNR(Peak Signal-to-Noise Ratio)의 결과로 나타낸다.

제안기법의 BPP-PSNR 성능은 INVR CTC에 정의된 기준(anchor)인 기존 3D 비디오 압축 표준인 TMIV DSDE (Decoder Side Depth Estimation)의 성능과 비교한다. 또한 TensoRF의 텐서 평면을 VVC로 압축하고 나머지 TensoRF의 매개변수를 NNCodec으로 압축했을 때의 결과(TensoRF with VVC)와도 비교한다. 제안기법과 두 비교 방법은 동일한 멀티뷰 정적 영상을 압축하고 동일한 테스트 뷰에 대해 렌더링을 수행한다. 표 2는 제안기법과의 BPP-PSNR 성능 비교이다. 또한, 표 2는 특징맵 압축을 적용하지 않

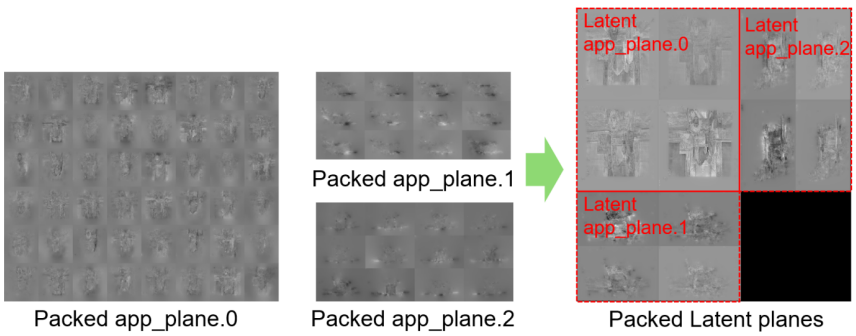

그림 6. 텐서 평면의 특징맵과 잠재 평면의 특징맵 비교
Fig. 6. Comparison of feature maps between the tensor plane and latent plane

표 2. 제안 방법의 BPP-PSNR 성능 비교
Table 2. Experimental results on the comparison of BPP-PSNR performances

	TMIV DSDE		Tensor plane compression with VVC (TensoRF with VVC)		Proposed method	
	BPP	PSNR	BPP	PSNR	BPP	PSNR
unc	-	-	-	32.02	-	29.51
QP1	12.82	28.29	7.29	31.41	2.73	29.11
QP2	5.87	28.17	4.74	30.65	2.29	28.07
QP3	3.77	28.10	3.24	28.54	2.02	26.39
QP4	2.07	27.82	2.44	25.41	1.88	24.52

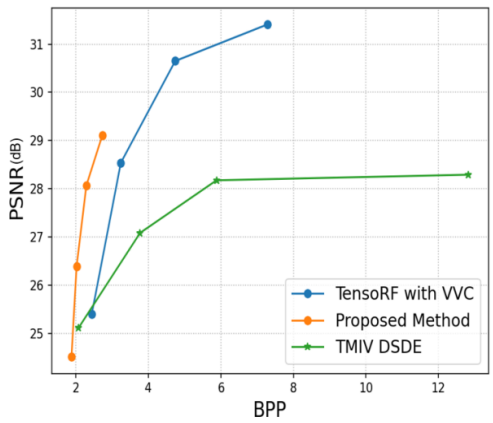

그림 7. BPP-PSNR 성능 비교
Fig. 7. BPP-PSNR performance comparison

았을 때의 학습된 TensoRF의 성능도 같이 보인다. 그림 7은 표 2의 결과를 율-왜곡(RD: Rate-Distortion) 곡선으로 나타낸 것이다. 제시된 결과와 같이 제안기법은 비교하는 기존 방법보다 우수한 BPP-PSNR 성능을 보인다.

그림 8은 유사한 BPP에서 VVC와 제안된 FRN과 VVC로 텐서 평면을 압축했을 때의 주관적 화질을 비교한 것이다. 유사한 BPP에서, 제안된 방법의 렌더링 성능이 VVC만을 사용했을 때보다 우수함을 확인할 수 있다.

실험결과에서 제안기법은 FRN을 적용하여 큰 정보 손실 없이 VVC로 압축하는 특징맵의 해상도를 줄임으로써 우수한 BPP-PSNR 성능을 얻은 것으로 볼 수 있다. 하지만 FRN을 적용하지 않고 간단히 텐서 평면의 랭크 수를 축소하고 TensoRF를 학습하면 제안기법과 같이 해상도가 축소된 특징맵을 압축할 수 있다. 따라서, 텐서 평면의 채널 축소에 비해 FRN이 효과를 검증하기 위해 다음과 같은 추가 실험을 진행하였다.

FRN은 각 외관 텐서 평면과 밀도 텐서 평면의 랭크 수를 4와 2로 축소한다. 유사하게, TensoRF를 학습할 때 FRN을 사용하지 않고 초기 학습부터 외관 텐서 평면과 밀도 텐서 평면의 랭크 수를 4와 2로 고정하여 학습한다. 이 실험에서는 FRN을 적용하지 않은 텐서 평면의 특징맵 해상도가 FRN이 적용된 잠재 평면 특징맵과 동일할 것이며, VVC로 압축했을 때 비슷한 크기의 압축 비트열이 나올 것이다. 따라서 두 방법을 비교할 때, 렌더링 성능이 높은 쪽이 더 효과적임을 나타낸다. 렌더링 PSNR 성능을 비교한 결과, 텐서 평면의 랭크 수를 4와 2로 고정하고 TensoRF를 학습한 렌더링 성능은 28.72(dB)이고, FRN을 적용했을 때는

그림 8. TensoRF 압축 주관적 화질 비교 (좌: VVC, 우: FRN 및 VVC)
Fig. 8. Comparison of Subjective Quality for TensoRF Compression (Left: VVC, Right: FRN and VVC)

29.51(dB)로 더 우수한 렌더링 화질 성능을 보인다. 따라서 FRN은 TensoRF의 텐서 평면의 랭크 수를 효과적으로 축소한다는 것을 알 수 있다.

V. 결 론

본 논문에서는 6DoF 몰입형 비디오의 압축을 위하여 하이브리드 3D INR 모델인 TensoRF 모델의 효과적인 압축 기법을 제시하였다. 제안기법은 정적 멀티뷰를 TensoRF로 학습한 다음 학습된 TensoRF의 텐서 평면을 FRN을 통하여 잠재 평면으로 축소하고, 축소된 잠재 평면으로 구성된 특징맵을 VVC로 압축한다. 제안기법은 INVR CTC의 기준(anchor)인 TMIV DSDE 및 TensoRF 텐서 평면의 특징맵을 VVC로 압축한 기존 방법 보다 우수한 BPP-PSNR 성능을 보였다. 또한, 추가적인 실험으로 제안한 FRN을 통한 평면 축소가 TensoRF의 랭크 수를 고정하여 텐서 평면을 축소한 것보다 효과적임을 보였다.

제안기법은 FRN을 TensoRF와 동시에 학습하여 원래의 TensoRF 보다 최대 렌더링 화질이 저하되는 한계가 있다. 하지만, FRN의 구조 확장 및 학습 기법의 개선으로 렌더링 화질을 개선할 수 있을 것으로 예상된다. 제안기법은 명시적 표현과 암시적 신경망 표현을 포함하는 다양한 하이브리드 NeRF 모델에 적용될 수 있으며, 하이브리드 NeRF 모델의 표현 및 압축을 통한 6DoF 몰입형 비디오 압축의 가능성을 제시하였다.

참 고 문 헌 (References)

[1] V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, “Implicit Neural Representations with Periodic Activation Functions,” *Advances in Neural Information Processing Systems*, vol. 13, pp. 7462-7473, 2020. doi: https://doi.org/10.48550/arXiv.2006.09661

[2] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, “Tensorf: Tensorial Radiance Fields,” *Proceedings of the European Conference on Computer Vision*, Arxiv, Israel, Oct. 2022. doi: https://doi.org/10.1007/978-3-031-19824-3_20

[3] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, “Plenoxels: Radiance Fields without Neural Networks,” *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5501 – 5510, 2022. doi: https://doi.org/10.1109/CVPR52688.2022.00542

[4] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3D Gaussian Splatting for Real-Time Radiance Field Rendering,” *ACM Transactions on Graphics*, vol.42, No.4, pp.1 – 14, 2023. doi: https://doi.org/10.1145/3592433

[5] G. Lafruit, Y. Liao, and G. Bang, *AhG on Implicit Neural Video Representations (INVR)*, ISO/IEC JTC1/SC 29/WG04, M60641, Oct. 2022.

[6] Y. Liao, and G. Bang, *BoG report on Implicit Neural Visual Representation (INVR)*, ISO/IEC JTC 1/SC 29/WG04, M68163, Apr. 2024.

[7] J. Jung, and B. Kroon, *Common Test Conditions for MPEG Immersive Video*, ISO/IEC JTC1/SC29/WG4, N0232, Jul. 2022.

[8] B. Salahieh, J. Jung, and A. Dziembowski, *Test Model 14 for MPEG immersive video*, ISO/IEC JTC1/SC29/WG4, N0242, Jul. 2022.

[9] B. Mildenhal, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing Scenes as Neural Radiance Fields for View Synthesis,” *Proceedings of the European Conference on Computer Vision*, pp. 405-421, Aug. 2020. doi: https://doi.org/10.1007/978-3-030-58452-8_24

[10] D. Kim, J. Lee, H. Moon, J. Kim, J. Jeong, and G. Lee, *[INVR] EE 2.2: Compression of TensoRF with Feature Compression Network and VVC*, ISO/IEC JTC1/SC29/WG4, M67819, Apr. 2024.

[11] MPEG Liaison and communication, *White paper on neural network coding*, ISO/IEC JTC1/SC29/AG3, N0057, Jan. 2022.

[12] D. Kim, J. Kim, J. Jeong, and G. Lee, *[INVR] EE2.2-Related: TensoRF compression using VVC*, ISO/IEC JTC1/SC29/WG4, M65139, Oct. 2023.

[13] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, “Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines,” *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1 – 14, 2019. doi: https://dl.acm.org/doi/10.1145/3306346.3322980

[14] C. Rosewarne and Y. Zhang, AHG on Feature Compression for *Video Coding for Machines,* ISO/IEC JTC 1/SC 29/WG 2, m61552, Jan. 2022.

저 자 소 개

김 동 하

- 2021년 8월 : 한국항공대학교 항공전자정보공학과 학사
- 2023년 8월 : 한국항공대학교 항공전자정보공학과 석사
- 2023년 9월 ~ 현재 : 한국항공대학교 항공전자정보공학과 박사과정
- ORCID : https://orcid.org/0009-0007-9918-571X
- 주관심분야 : 인공지능 기반 미디어 신호 처리, 암시적 신경망 표현, Immersive Video

문 현 철

- 2018년 2월 : 한국항공대학교 항공전자정보공학과 학사
- 2020년 8월 : 한국항공대학교 항공전자정보공학과 석사
- 2021년 9월 ~ 현재 : 한국전자기술연구원 연구원
- 2023년 9월 ~ 현재 : 한국항공대학교 항공전자정보공학과 박사과정
- ORCID : https://orcid.org/0000-0002-1672-2345
- 주관심분야 : 인공지능 기반 미디어 신호 처리, 인공지능 경량화

정 준 영

- 2013년 5월 : 퍼듀대학교 전자공학과 학사
- 2016년 5월 : 퍼듀대학교 전자공학과 석사
- 2016년 10월 ~ 2024년 3월 : 한국전자통신연구원(ETRI) 연구원
- 2024년 4월 ~ 현재 : 한국전자통신연구원 선임연구원
- ORCID : https://orcid.org/0000-0002-2457-1647
- 주관심분야 : Radiance-Field 모델 경량화 및 부호화

이 광 순

- 1993년 : 경북대학교 전자공학과 학사
- 1995년 : 경북대학교 전자공학과 석사
- 2001년 : 한국전자통신연구원 입사
- 2004년 : 경북대학교 전자공학과 박사
- 2013년 ~ 2015년 : 입체방송연구실장
- 2016년 ~ 현재 : 미디어연구본부 실감미디어연구실 책임연구원
- ORCID : http://orcid.org/0000-0001-6981-2099
- 주관심분야 : 이머시브 비디오 처리 및 부호화, Learning-based 3D representation 및 부호화

김 재 곤

- 1990년 2월 : 경북대학교 전자공학과 학사
- 1992년 2월 : KAIST 전기 및 전자공학과 석사
- 2005년 2월 : KAIST 전기 및 전자공학과 박사
- 1992년 3월 ~ 2007년 2월 : 한국전자통신연구원(ETRI) 선임연구원/팀장
- 2001년 9월 ~ 2002년 7월 : Columbia University 연구원
- 2015년 12월 ~ 2016년 1월 : UC San Diego, Visiting Scholar
- 2007년 9월 ~ 현재 : 한국항공대학교 항공전자정보공학부 교수
- ORCID : http://orcid.org/0000-0003-3686-4786
- 주관심분야 : 비디오 부호화 표준, 비디오 신호처리, Immersive Video, Deep Learning