(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0168843
(43) 공개일자 2024년12월02일

(51) 국제특허분류(Int. Cl.)
H04N 19/149 (2014.01) *G06N 3/045* (2023.01)
G06N 3/0464 (2023.01) *G06N 3/0985* (2023.01)
G06T 3/18 (2024.01) *H04N 19/124* (2014.01)
H04N 19/196 (2014.01) *H04N 19/46* (2014.01)
H04N 19/70 (2014.01)
(52) CPC특허분류
H04N 19/149 (2015.01)
G06N 3/045 (2023.01)
(21) 출원번호 10-2024-0061136
(22) 출원일자 2024년05월09일
심사청구일자 없음
(30) 우선권주장
1020230066090 2023년05월23일 대한민국(KR)

(71) 출원인
현대자동차주식회사
서울특별시 서초구 헌릉로 12 (양재동)
기아 주식회사
서울특별시 서초구 헌릉로 12 (양재동)
이화여자대학교 산학협력단
서울특별시 서대문구 이화여대길 52 (대현동, 이화여자대학교)
(72) 발명자
강제원
서울특별시 마포구 독막로 145, 104-904
이정경
서울특별시 양천구 목동동로 350 506동 1201호
(뒷면에 계속)
(74) 대리인
이철희

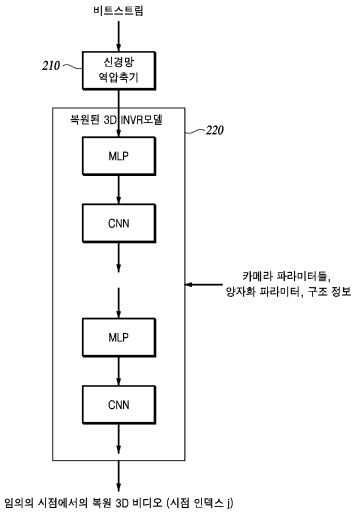
전체 청구항 수 : 총 18 항

(54) 발명의 명칭 **암시적 신경망 표현에 기반하는 3D 비디오 코딩방법 및 장치**

(57) 요 약

본 실시예는 암시적 신경망 표현에 기반하는 3D 비디오 코딩방법 및 장치를 개시한다. 본 실시예에서, 영상 복호화 장치는 비트스트림으로부터 3D INVR 모델의 파라미터들, 및 IPS를 복호화한다. 여기서, IPS는 양자화 파라미터, 및 3D INVR 모델의 구조에 관한 정보, 및 상기 3D INVR 모델의 입력 범위에 관한 정보를 포함한다. 영상 복호화 장치는 3D INVR 모델의 파라미터들, 및 양자화 파라미터를 이용하여 3D INVR 모델의 구조에 관한 정보에 부합하도록 3D INVR 모델을 복원한다. 영상 복호화 장치는 입력 범위 내에서 임의의 시점(view point)을 획득하고, 임의의 시점을 3D INVR 모델에 입력하여 현재 3D 비디오를 생성한다.

대 표 도 - 도2



(52) CPC특허분류
G06N 3/0464 (2023.01)
G06N 3/0985 (2023.01)
G06T 3/18 (2024.01)
H04N 19/124 (2015.01)
H04N 19/196 (2015.01)
H04N 19/46 (2015.01)
H04N 19/70 (2015.01)
(72) 발명자
이지후
서울특별시 성북구 보문사길 111, 102동 1102호
신주연
서울특별시 서대문구 성산로24길 26 301호
최정아
경기도 화성시 남양읍 현대연구소로 150
허진
경기도 화성시 남양읍 현대연구소로 150
박승욱
경기도 화성시 남양읍 현대연구소로 150

명 세 서

청구범위

청구항 1

영상 복호화 장치가 수행하는, 현재 3D(Dimensional) 비디오를 복원하는 방법에 있어서,

비트스트림으로부터 3D(Dimensional) INVR(Implicit Neural Video Representation) 모델의 파라미터들, 및 IPS(INVR Parameter Set)를 복호화하는 단계, 여기서, 상기 IPS는 양자화 파라미터, 및 3D INVR 모델의 구조에 관한 정보, 및 상기 3D INVR 모델의 입력 범위에 관한 정보를 포함함;

상기 3D INVR 모델의 파라미터들, 및 상기 양자화 파라미터를 이용하여 상기 3D INVR 모델의 구조에 관한 정보에 부합하도록 상기 3D INVR 모델을 복원하는 단계, 여기서, 상기 3D INVR 모델은 다층신경망(multi-layer perceptron, MLP)과 CNN(Convolutional Neural Network)를 포함함;

상기 입력 범위 내에서 임의의 시점(view point)을 획득하는 단계; 및

상기 임의의 시점을 상기 3D INVR 모델에 입력하여 상기 현재 3D 비디오를 생성하는 단계

를 포함하는 것을 특징으로 하는, 방법.

청구항 2

제1항에 있어서,

3D INVR 모델의 복원하는 단계는,

상기 3D INVR 모델의 구조에 관한 정보를 이용하여 상기 MLP과 CNN의 크기 및 개수를 결정하는 것을 특징으로 하는, 방법.

청구항 3

제1항에 있어서,

상기 3D INVR 모델의 구조에 관한 정보는,

상기 MLP와 CNN의 크기 및 개수를 정의하는 신택스, 및 상기 3D INVR 모델의 의 파라미터들의 형식 및 정밀도(precision)을 정의하는 신택스를 포함하는 것을 특징으로 하는, 방법.

청구항 4

제1항에 있어서,

상기 IPS는,

카메라 파라미터들을 추가로 포함하는 것을 특징으로 하는, 방법.

청구항 5

제4항에 있어서,

와핑을 이용하여 현재 시점에 대한 참조 이미지를 생성하는 단계, 여기서, 상기 와핑은, 상기 카메라 파라미터들을 이용하여 상기 참조 이미지의 좌표계를 월드 좌표계로 이동시기고, 상기 참조 이미지의 월드 좌표계를 상기 현재 3D 비디오의 좌표계로 변환하는 과정을 나타냄;

상기 와핑이 적용된 참조 이미지와 상기 현재 3D 비디오를 가중합함으로써, 개선된 현재 3D 비디오를 생성하는 단계

를 더 포함하는 것을 특징으로 하는, 방법.

청구항 6

제4항에 있어서,

와핑을 이용하여 현재 시점에 대한 참조 특징을 생성하는 단계, 여기서, 상기 와핑은, 상기 카메라 파라미터들을 이용하여 상기 참조 특징의 좌표계를 월드 좌표계로 이동시기고, 상기 참조 특징의 월드 좌표계를 상기 현재 3D 비디오의 특징의 좌표계로 변환하는 과정을 나타냄;

상기 와핑이 적용된 참조 특징과 상기 현재 3D 비디오의 특징을 가중합함으로써, 개선된 현재 3D 비디오를 생성하는 단계

를 더 포함하는 것을 특징으로 하는, 방법.

청구항 7

제1항에 있어서,

상기 IPS는,

깊이 정보(depth map)의 전달 여부를 지시하는 신택스를 추가로 포함하는 것을 특징으로 하는, 방법.

청구항 8

제7항에 있어서,

상기 신택스를 확인하는 단계를 더 포함하고,

상기 신택스가 상기 깊이 정보의 전달을 지시하는 경우, 상기 깊이 정보를 복호화하는 단계를 더 포함하는 것을 특징으로 하는, 방법.

청구항 9

제8항에 있어서,

와핑을 이용하여 현재 시점에 대한 참조 이미지를 생성하는 단계, 여기서, 상기 와핑은, 상기 깊이 정보를 이용하여 상기 참조 이미지의 좌표계를 월드 좌표계로 이동시기고, 상기 참조 이미지의 월드 좌표계를 상기 현재 3D 비디오의 좌표계로 변환하는 과정을 나타냄;

상기 와핑이 적용된 참조 이미지와 상기 현재 3D 비디오를 가중합함으로써, 개선된 현재 3D 비디오를 생성하는 단계

를 더 포함하는 것을 특징으로 하는, 방법.

청구항 10

제1항에 있어서,

상기 복호화하는 단계는,

상기 IPS를 GoI(Group of INVR)별로 복호화하되,

하나의 3D INVR 모델이 하나의 시간에서 복수의 시점들을 표현하도록 상기 GoI가 구성되는 것을 특징으로 하는, 방법.

청구항 11

영상 부호화 장치가 수행하는, 3D(Dimensional) 비디오들을 부호화하는 방법에 있어서,

상기 3D 비디오들, 및 상기 3D 비디오들에 대응하는 카메라 파라미터들을 획득하는 단계;

3D INVR(Implicit Neural Video Representation) 모델을 구성하는 단계, 여기서, 상기 3D INVR 모델은 다층신경망(multi-layer perceptron, MLP)과 CNN(Convolutional Neural Network)을 포함함;

상기 카메라 파라미터들에 대응하는 시점들 및 상기 카메라 파라미터들에 대응하는 3D 비디오들을 이용하여 상기 3D INVR 모델의 트레이닝을 수행하는 단계;

상기 트레이닝에 사용된 시점들에 기초하여 상기 3D INVR 모델의 입력 범위를 결정하는 단계;

상기 트레이닝의 결과에 기초하여 양자화 파라미터, 및 상기 3D INVR 모델의 구조에 관한 정보를 결정하는 단계;

상기 양자화 파라미터를 이용하여 상기 3D INVR 모델의 파라미터들을 부호화하는 단계; 및

상기 양자화 파라미터, 상기 3D INVR 모델의 구조에 관한 정보, 및 상기 3D INVR 모델의 입력 범위에 관한 정보를 포함하는 IPS(INVR Parameter Set)를 부호화하는 단계

를 포함하는 것을 특징으로 하는, 방법.

청구항 12

제11항에 있어서,

상기 IPS는,

상기 카메라 파라미터들, 및 깊이 정보(depth map)의 전달 여부를 지시하는 신택스를 추가로 포함하는 것을 특징으로 하는, 방법.

청구항 13

제11항에 있어서,

3D INVR 모델의 구조를 구성하는 단계는,

상기 MLP과 상기 CNN의 크기 및 개수에 기초하여 상기 3D INVR 모델의 구조를 상이하게 구성하고,

상기 3D INVR 모델의 구조에 관한 정보를 결정하는 단계는,

율왜곡 및 상기 트레이닝의 속도 측면에서 최적인 상기 3D INVR 모델의 구조를 선택하는 단계; 및

상기 3D INVR 모델의 구조를 지시하는 신택스 요소들을 상기 3D INVR 모델의 구조에 관한 정보에 포함시키는 단계

를 포함하는 것을 특징으로 하는, 방법.

청구항 14

제11항에 있어서,

상기 양자화 파라미터를 결정하는 단계는,

율왜곡 최적화 측면에서 상기 양자화 파라미터를 적응적으로 결정하는 것을 특징으로 하는, 방법.

청구항 15

제11항에 있어서,

상기 3D INVR 모델의 입력 범위를 결정하는 단계는,

상기 3D INVR 모델의 트레이닝에 사용된 시점들에 대해, 상기 시점들을 구성하는 요소들 각각의 상한과 하한을 결정하는 것을 특징으로 하는, 방법.

청구항 16

제11항에 있어서,

상기 3D INVR 모델의 트레이닝을 수행하는 단계는,

상기 3D INVR 모델의 출력과 GT(Ground Truth) 간 차이를 산정하는 단계, 여기서, 상기 GT로서 상기 3D 비디오들이 사용됨;

상기 차이, 및 희소 손실(sparsity loss)을 포함하는 손실 함수에 기초하여 상기 3D INVR 모델의 파라미터들을 업데이트하는 것을 특징으로 하는, 방법.

청구항 17

제11항에 있어서,

상기 3D INVR 모델의 트레이닝을 수행하는 단계는,

프루닝(pruning)을 이용하여 상기 3D INVR 모델의 파라미터들의 개수를 감소시키는 것을 특징으로 하는, 방법.

청구항 18

영상 복호화 장치에게 비디오 데이터를 제공하는 방법에 있어서,

상기 비디오 데이터를 비트스트림으로 부호화하는 단계; 및

상기 비트스트림을 상기 영상 복호화 장치로 전달하는 단계

를 포함하고,

상기 비디오 데이터를 부호화하는 단계는,

3D 비디오들, 및 상기 3D 비디오들에 대응하는 카메라 파라미터들을 획득하는 단계;

3D INVR(Implicit Neural Video Representation) 모델을 구성하는 단계, 여기서, 상기 3D INVR 모델은 다층신경망(multi-layer perceptron, MLP)과 CNN(Convolutional Neural Network)을 포함함;

상기 카메라 파라미터들에 대응하는 시점들 및 상기 카메라 파라미터들에 대응하는 3D 비디오들을 이용하여 상기 3D INVR 모델의 트레이닝을 수행하는 단계;

상기 트레이닝에 사용된 시점들에 기초하여 상기 3D INVR 모델의 입력 범위를 결정하는 단계;

상기 트레이닝의 결과에 기초하여 양자화 파라미터, 및 상기 3D INVR 모델의 구조에 관한 정보를 결정하는 단계;

상기 양자화 파라미터를 이용하여 상기 3D INVR 모델의 파라미터들을 부호화하는 단계; 및

상기 양자화 파라미터, 상기 3D INVR 모델의 구조에 관한 정보, 및 상기 3D INVR 모델의 입력 범위에 관한 정보를 포함하는 IPS(INVR Parameter Set)를 부호화하는 단계

를 포함하는 것을 특징으로 하는, 방법.

발명의 설명

기 술 분 야

[0001] 본 개시는 암시적 신경망 표현에 기반하는 3D 비디오 코딩방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 이하에 기술되는 내용은 단순히 본 실시예와 관련되는 배경 정보만을 제공할 뿐 종래기술을 구성하는 것이 아니다.

[0003] 최근 이미지를 비롯한 각종 데이터의 표현을 신경망 구조로 나타내는 암시적 신경망 표현 모델에 관한 연구가 활발하다. 기존의 비디오 표현 방식으로서, 픽셀 위치별 RGB 픽셀값을 갖는 명시적(explicit) 표현 방식이 사용된다. 이러한 명시적 표현 방식을 대체하기 위해 영상의 픽셀 위치인 (x,y) 좌표에서 (r,g,b) 값을 생성하는 함수를 신경망으로 표현하는 암시적 신경망 표현(implicit neural representation) 방식이 도입되고 있다.

[0004] 암시적 신경 표현 방식 중 3-Dimensional(3D) 비디오의 합성에 사용하는 대표 모델로서 NeRF(Neural Radiance Fields) 기술이 존재한다. 이미지 내의 모든 픽셀들을 월드 좌표계로 재구성하기 위한 학습 과정을 이용하여 NeRF는 임의의 시점(view point)에서 렌더링된 결과를 출력한다. 이때, NeRF는 월드 좌표계에서의 인덱스 및 월드 좌표계로의 투영에 필요한 카메라 파라미터들을 입력으로 사용한다. 따라서, 3D 비디오 부호화 효율을 향상시키고 3D 비디오 화질을 개선하기 위해, 암시적 신경망 표현 기반의 3D 비디오 표현 모델을 효율적으로 구성하

는 방안이 고려될 필요가 있다.

발명의 내용

해결하려는 과제

[0005] 본 개시는, 3D 암시적 비디오 표현(3D Implicit Neural Video Representation, 3D INVR) 모델의 파라미터들을 생성, 압축 및 전송하고, 디코더 측에서 요구하는 임의 시점의 품질을 개선하기 위한 카메라 파라미터들, 및 3D INVR 모델의 압축 방식을 개선하기 위한 부가 정보를 생성 및 전송하는 3D 비디오 코딩방법 및 장치를 제공하는 데 목적이 있다.

[0006] 또한, 본 개시는, 3D INVR 모델의 압축을 위한 양자화 파라미터, 및 3D INVR 모델의 구조에 따른 복잡도를 적응적으로 변화시켜 3D INVR 모델의 압축 방식을 개선하기 위한 부가 정보를 생성하는 3D 비디오 코딩방법 및 장치를 제공하는 데 목적이 있다.

과제의 해결 수단

[0007] 본 개시의 실시예에 따르면, 영상 복호화 장치가 수행하는, 현재 3D(Dimensional) 비디오를 복원하는 방법에 있어서, 비트스트림으로부터 3D(Dimensional) INVR(Implicit Neural Video Representation) 모델의 파라미터들, 및 IPS(INVR Parameter Set)를 복호화하는 단계, 여기서, 상기 IPS는 양자화 파라미터, 및 3D INVR 모델의 구조에 관한 정보, 및 상기 3D INVR 모델의 입력 범위에 관한 정보를 포함함; 상기 3D INVR 모델의 파라미터들, 및 상기 양자화 파라미터를 이용하여 상기 3D INVR 모델의 구조에 관한 정보에 부합하도록 상기 3D INVR 모델을 복원하는 단계, 여기서, 상기 3D INVR 모델은 다층신경망(multi-layer perceptron, MLP)과 CNN(Convolutional Neural Network)를 포함함; 상기 입력 범위 내에서 임의의 시점(view point)을 획득하는 단계; 및 상기 임의의 시점을 상기 3D INVR 모델에 입력하여 상기 현재 3D 비디오를 생성하는 단계를 포함하는 것을 특징으로 하는, 방법을 제공한다.

[0008] 본 개시의 다른 실시예에 따르면, 영상 부호화 장치가 수행하는, 3D(Dimensional) 비디오들을 부호화하는 방법에 있어서, 상기 3D 비디오들, 및 상기 3D 비디오들에 대응하는 카메라 파라미터들을 획득하는 단계; 3D INVR(Implicit Neural Video Representation) 모델을 구성하는 단계, 여기서, 상기 3D INVR 모델은 다층신경망(multi-layer perceptron, MLP)과 CNN(Convolutional Neural Network)을 포함함; 상기 카메라 파라미터들에 대응하는 시점들 및 상기 카메라 파라미터들에 대응하는 3D 비디오들을 이용하여 상기 3D INVR 모델의 트레이닝을 수행하는 단계; 상기 트레이닝에 사용된 시점들에 기초하여 상기 3D INVR 모델의 입력 범위를 결정하는 단계; 상기 트레이닝의 결과에 기초하여 양자화 파라미터, 및 상기 3D INVR 모델의 구조에 관한 정보를 결정하는 단계; 상기 양자화 파라미터를 이용하여 상기 3D INVR 모델의 파라미터들을 부호화하는 단계; 및 상기 양자화 파라미터, 상기 3D INVR 모델의 구조에 관한 정보, 및 상기 3D INVR 모델의 입력 범위에 관한 정보를 포함하는 IPS(INVR Parameter Set)를 부호화하는 단계를 포함하는 것을 특징으로 하는, 방법을 제공한다.

[0009] 본 개시의 다른 실시예에 따르면, 영상 복호화 장치에게 비디오 데이터를 제공하는 방법에 있어서, 상기 비디오 데이터를 비트스트림으로 부호화하는 단계; 및 상기 비트스트림을 상기 영상 복호화 장치로 전달하는 단계를 포함하고, 상기 비디오 데이터를 부호화하는 단계는, 3D 비디오들, 및 상기 3D 비디오들에 대응하는 카메라 파라미터들을 획득하는 단계; 3D INVR(Implicit Neural Video Representation) 모델을 구성하는 단계, 여기서, 상기 3D INVR 모델은 다층신경망(multi-layer perceptron, MLP)과 CNN(Convolutional Neural Network)을 포함함; 상기 카메라 파라미터들에 대응하는 시점들 및 상기 카메라 파라미터들에 대응하는 3D 비디오들을 이용하여 상기 3D INVR 모델의 트레이닝을 수행하는 단계; 상기 트레이닝에 사용된 시점들에 기초하여 상기 3D INVR 모델의 입력 범위를 결정하는 단계; 상기 트레이닝의 결과에 기초하여 양자화 파라미터, 및 상기 3D INVR 모델의 구조에 관한 정보를 결정하는 단계; 상기 양자화 파라미터를 이용하여 상기 3D INVR 모델의 파라미터들을 부호화하는 단계; 및 상기 양자화 파라미터, 상기 3D INVR 모델의 구조에 관한 정보, 및 상기 3D INVR 모델의 입력 범위에 관한 정보를 포함하는 IPS(INVR Parameter Set)를 부호화하는 단계를 포함하는 것을 특징으로 하는, 방법을 제공한다.

발명의 효과

[0010] 이상에서 설명한 바와 같이 본 실시예에 따르면, 3D 암시적 비디오 표현(3D Implicit Neural Video Representation, 3D INVR) 모델의 파라미터들을 생성, 압축 및 전송하고, 디코더 측에서 요구하는 임의 시점의

품질을 개선하기 위한 카메라 파라미터들, 및 3D INVR 모델의 압축 방식을 개선하기 위한 부가 정보를 생성 및 전송하는 3D 비디오 코딩방법 및 장치를 제공함으로써, 임의 시점의 비디오 생성 과정에서 3D 비디오 부호화 효율 및 3D 비디오 화질을 개선하는 것이 가능해지는 효과가 있다.

[0011] 또한 본 실시예에 따르면, 3D INVR 모델의 압축을 위한 양자화 파라미터, 및 3D INVR 모델의 구조에 따른 복잡도를 적응적으로 변화시켜 3D INVR 모델의 압축 방식을 개선하기 위한 부가 정보를 생성하는 3D 비디오 코딩방법 및 장치를 제공함으로써, 3D 비디오 부호화 효율 및 3D 비디오 화질을 개선하는 것이 가능해지는 효과가 있다.

도면의 간단한 설명

[0012] 도 1은 본 개시의 일 실시예에 따른, 3D INVR(Implicit Neural Video Representation) 기반 3D 영상 부호화 장치를 나타내는 블록도이다.

도 2는 본 개시의 일 실시예에 따른, 3D INVR 기반 3D 영상 복호화 장치를 나타내는 블록도이다.

도 3은 본 개시의 일 실시예에 따른 콘볼루션 레이어의 연산을 나타내는 예시도이다.

도 4는 SISR(Single Image Super Resolution) 네트워크를 나타내는 예시도이다.

도 5는 SISR에 이용되는 잔차 블록을 나타내는 예시도이다.

도 6은 NeRF(Neural Radiance Fields)의 임의의 시점의 추출과 관련된 좌표계들을 나타내는 예시도이다.

도 7은 NeRV(Neural Representations for Videos) 모델을 나타내는 예시도이다.

도 8a 및 도 8b는 본 개시의 일 실시예에 따른, NeRV 모델의 개념을 나타내는 예시도이다.

도 9는 본 개시의 일 실시예에 따른, NeRV 모델의 부호화 파이프라인을 나타내는 예시도이다.

도 10은 본 개시의 일 실시예에 따른, NeRV 모델의 복호화 파이프라인을 나타내는 예시도이다.

도 11은 본 개시의 일 실시예에 따른, 비트 마스크를 이용한 프루닝을 나타내는 예시도이다.

도 12는 본 개시의 일 실시예에 따른, 참조 이미지의 와핑을 나타내는 예시도이다.

도 13은 본 개시의 일 실시예에 따른, 참조 이미지/특징의 와핑을 나타내는 예시도이다.

도 14는 본 개시의 일 실시예에 따른, GoI의 정의를 나타내는 예시도이다.

도 15는 본 개시의 일 실시예에 따른, 영상 부호화 장치가 3D 비디오들을 부호화하는 방법을 나타내는 순서도이다.

도 16은 본 개시의 일 실시예에 따른, 영상 복호화 장치가 현재 3D 비디오를 복원하는 방법을 나타내는 순서도이다.

발명을 실시하기 위한 구체적인 내용

[0013] 이하, 본 개시의 일부 실시예들을 예시적인 도면을 참조하여 상세하게 설명한다. 각 도면의 구성요소들에 참조 부호를 부가함에 있어서, 동일한 구성요소들에 대해서는 비록 다른 도면상에 표시되더라도 가능한 한 동일한 부호를 가지도록 하고 있음에 유의해야 한다. 또한, 본 실시예들을 설명함에 있어, 관련된 공지 구성 또는 기능에 대한 구체적인 설명이 본 실시예들의 요지를 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명은 생략한다.

[0014] 본 실시예는 3D 영상(비디오)의 부호화 및 복호화에 관한 것이다. 보다 자세하게는, 3D 암시적 비디오 표현(3D Implicit Neural Video Representation, 3D INVR) 모델의 파라미터들을 생성, 압축 및 전송하고, 디코더 측에서 요구하는 임의 시점의 품질을 개선하기 위한 카메라 파라미터들, 및 3D INVR 모델의 압축 방식을 개선하기 위한 부가 정보를 생성 및 전송하는 3D 비디오 코딩방법 및 장치를 제공한다.

[0015] 도 1은 본 개시의 일 실시예에 따른, 3D INVR 기반 3D 영상 부호화 장치를 나타내는 블록도이다.

[0016] 도 1의 예시에서 3D INVR 기반 3D 영상 부호화 장치(이하, '영상 부호화 장치')는 다양한 시점에서 3D 비디오를 생성하는 것을 학습하도록 3D INVR 모델을 생성(즉, 트레이닝)하고, 학습된 3D INVR 모델을 압축하여 비트스트림을 생성하며, 생성된 비트스트림을 전송한다. 영상 부호화 장치는 3D INVR 모델을 포함하는 3D INVR 부호화부

(110) 및 신경망 압축기(Neural Network Compressor, 120)를 포함한다. 여기서, 본 개시에 따른 영상 부호화 장치에 포함되는 구성요소가 반드시 이에 한정되는 것은 아니다. 영상 부호화 장치는 3D INVR 모델을 트레이닝하기 위한 트레이닝부(미도시)를 추가로 구비하거나, 외부의 트레이닝부와 연동되는 형태로 구현될 수 있다.

[0017] 영상 부호화 장치의 각 구성요소는 하드웨어 또는 소프트웨어로 구현되거나, 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다. 또한, 각 구성요소의 기능이 소프트웨어로 구현되고 마이크로프로세서가 각 구성요소에 대응하는 소프트웨어의 기능을 실행하도록 구현될 수도 있다.

[0018] 영상 부호화 장치는 부호화된 비디오 데이터의 비트스트림을 비일시적인 기록매체에 저장하거나 통신 네트워크를 이용하여 3D INVR 기반 3D 영상 복호화 장치로 전송할 수 있다.

[0019] 도 2는 본 개시의 일 실시예에 따른, 3D INVR 기반 3D 영상 복호화 장치를 나타내는 블록도이다.

[0020] 도 2의 예시에서 3D INVR 기반 3D 영상 복호화 장치(이하, '영상 복호화 장치')는 비트스트림을 역압축(decompression)하여 3D INVR 모델을 복원하고, 복원된 3D INVR 모델을 이용하여 다양한 시점에서 3D 비디오를 생성한다. 영상 복호화 장치는 신경망 역압축기(Neural Network Decompressor, 210), 및 복원된 3차원 INVR 모델을 포함하는 3D INVR 복호화부(220)를 포함한다.

[0021] 도 1에 예시된 영상 부호화 장치와 마찬가지로, 영상 복호화 장치의 각 구성요소는 하드웨어 또는 소프트웨어로 구현되거나, 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다. 또한, 각 구성요소의 기능이 소프트웨어로 구현되고 마이크로프로세서가 각 구성요소에 대응하는 소프트웨어의 기능을 실행하도록 구현될 수도 있다.

[0022] 이하, 도 1 및 도 2에 예시된 구성요소들을 설명하기 전에, 본 개시에 사용되는 딥러닝 관련 요소 기술들을 설명한다.

[0023] **I. CNN(Convolutional Neural Network)**

[0024] CNN은 복수의 콘볼루션 레이어(convolution layer)와 풀링 레이어(pooling layer)로 구성된 신경망을 지칭하고, 영상 처리에 가장 적합한 것으로 알려진 딥러닝 기술이다. 콘볼루션 레이어는 다수의 커널(kernel) 또는 필터(filter)를 이용하여 특징맵(feature map, 또는 '특성'과 호환적으로 이용)을 추출한다. 이때 필터를 구성하는 커널 계수(kernel coefficient)가 학습 과정에서 결정되는 파라미터이다.

[0025] CNN의 콘볼루션 레이어 중에서 입력에 가까운 전단 레이어는 선, 점, 또는 면과 같은 단순하고 낮은(lower) 수준의 영상 특징에 반응하는 특징맵을 추출하고, 출력에 가까운 후단 계층은 텍스처(texture), 사물 일부(object parts) 등 높은(higher) 수준에 반응하는 특징맵을 추출한다.

[0026] 도 3은 본 개시의 일 실시예에 따른 콘볼루션 레이어의 연산을 나타내는 예시도이다.

[0027] 콘볼루션 레이어는 콘볼루션 연산을 이용하여 입력 영상으로부터 특징맵을 생성한다. 도 3의 예시에는, 커널 크기가 3×3인 커널(또는 필터)이 표현되어 있다. 커널 크기는 커널 사이즈(kernel size) 또는 필터 사이즈(filter size)로도 지칭된다. 커널은 가중치(weight)로도 불리우는 커널 파라미터(kernel parameter 또는 filter parameter)를 갖는다. 도 3에 예시된 커널은 총 9 개의 커널 파라미터를 갖는다. 커널 파라미터는 초기에 임의의 값으로 설정되고, 트레이닝(learning)을 기반으로 그 값이 업데이트될 수 있다.

[0028] 콘볼루션 레이어는 입력 영상에서 커널 사이즈만큼의 블록을 이용하여 콘볼루션 연산을 수행한다. 이때, 입력 영상 내의 커널 사이즈만큼의 블록을 윈도우(window)로 지칭한다.

[0029] 입력 영상에 대한 필터링을 래스터 스캔 순서(raster-scan order)로 수행할 때, 윈도우의 이동 크기를 스트라이드(stride)라고 한다. 도 3의 예시에서, 스트라이드는 1이다. 만약 스트라이드를 2로 설정하면, 윈도우를 2 샘플씩 띄어서 콘볼루션 연산을 수행하고, 결과적으로 특징맵의 가로와 세로의 크기는 입력 영상의 가로와 세로의 크기의 반이 된다.

[0030] 전술한 바와 같이, 하나의 콘볼루션 레이어는 다수의 필터를 포함할 수 있다. 필터의 개수(number of filter/filters) 또는 커널의 개수(number of kernel/kernels)를 채널(channel)이라고 한다. 즉, 채널의 개수는 필터의 개수와 동일하다. 또한 필터의 개수는 특징맵의 차원(dimension)의 크기를 결정한다.

[0031] 패딩(padding)은 콘볼루션 연산을 수행하기 전, 입력 데이터 주변을 특정값으로 채워서 입력 데이터를 확장하는 방법을 나타낸다. 패딩은 주로 출력 데이터의 공간적(spatial) 크기를 조절하기 위해 사용된다. 패딩 시 이용되는 값은 하이퍼파라미터에 의해 결정될 수 있으나, 주로 제로패딩(zero-padding)이 이용된다. 패딩을 사용하지

않을 경우, 콘볼루션 레이어를 거칠 때마다 출력 데이터의 공간적 크기가 감소하여, 경계의 정보들이 사라지는 문제가 발생할 수 있다. 따라서, 이러한 문제를 방지하기 위해, 패딩이 사용된다. 즉, 콘볼루션 레이어의 출력 데이터와 입력 데이터의 공간적 크기를 동일하게 맞추기 위해 패딩이 사용될 수 있다.

[0032] 디콘볼루션 레이어는 콘볼루션 레이어와 반대되는 연산을 수행한다. 디콘볼루션 레이어는 입력인 특징맵으로부터 원하는 데이터 영상을 출력으로 생성한다.

[0033] 풀링 레이어는 콘볼루션 레이어에 의해 생성된 특징맵을 서브샘플링(sub sampling)하는 과정인 풀링을 수행한다. 풀링 레이어는 2×2 크기의 윈도우를 이용하여 출력 결과가 입력의 가로와 세로 각각의 절반이 되도록 샘플을 선택한다. 즉, 풀링 레이어는 2×2 영역을 샘플 하나로 집약하여 입력 영상 또는 입력 특징맵의 크기를 축소하기 위해 이용된다.

[0034] 풀링 레이어의 반대 개념은 언풀링(unpooling) 레이어로 정의한다. 언풀링 레이어는 풀링 레이어와 반대로 차원을 확대하는 역할을 하고, 주로 디콘볼루션 레이어 이후에 사용된다.

[0035] 콘볼루션 인코더(convolutional encoder)-디코더(decoder) 구조는 콘볼루션 레이어들와 디콘볼루션 레이어들의 짝(pair)으로 구성되는 네트워크 구조이다. 콘볼루션 인코더는 콘볼루션 레이어와 풀링 레이어로 구성되어, 입력 영상으로부터 특징맵(또는 특징 벡터)을 출력한다. 콘볼루션 인코더의 최종 출력 벡터를 잠재 벡터(latent vector)로도 지칭한다. 콘볼루션 디코더(convolutional decoder)는 디콘볼루션 레이어와 언풀링 레이어로 구성되어, 특징맵 또는 잠재 벡터로부터 출력 영상을 생성한다.

[0036] 콘볼루션 인코더-디코더의 입력과 출력은 애플리케이션과 네트워크의 목적에 따라 다양하게 설정될 수 있다. 예컨대, 입력과 출력은 옵티컬 플로우 맵(optical flow map), 돌출 맵(saliency map), 영상 프레임(image frame) 등일 수 있다.

[0037] 도 4는 SISR 네트워크를 나타내는 예시도이다.

[0038] CNN이 적용되는 일 예로서 SISR(Single Image Super Resolution)이 있다. SISR 네트워크는 저해상도 입력 영상으로부터 고해상도 이미지를 출력으로 생성한다. SISR 네트워크는 도 4에 예시된 바와 같이, 다수의 콘볼루션 레이어들을 포함할 수 있다. 각 콘볼루션 레이어는 ReLU(Rectified Linear Unit)과 같은 활성화 함수(activation function)를 포함한다. SISR 네트워크의 파라미터들은 생성되는 SR(Super Resolution) 영상이 GT(Ground Truth)에 근접하도록 트레이닝될 수 있다.

[0039] CNN을 이용한 SR 방식들은 깊이를 증가시켜(예컨대, 콘볼루션 레이어의 개수를 증가시키는 것을 의미함) SR 성능을 향상시킬 수 있다. 깊이의 증가에 따라 발생할 수 있는 학습에서의 과적합(overfitting) 문제를 극복하기 위해, 스킵 연결(skip connection) 및 잔차 학습(residual learning)을 수행할 수 있는 잔차 블록(residual block)이 SISR 네트워크에 이용될 수 있다. 잔차 블록은, 도 5에 예시된 바와 같이, 입력 특징 x_l에 콘볼루션 연산을 적용하는 경로 외에 스킵 경로를 포함한다. 또한, 잔차 블록은 출력 x_{l+1}을 생성 시, 학습 효율에 기초하여 콘볼루션 연산을 적용하는 경로 또는 스킵 경로를 선택할 수도 있다. 도 5의 예시에서, 잔차 블록은 BN(Batch Normalization) 레이어를 포함한다.

[0040] 일 예로서, EDSR(Enhanced Deep residual networks for SISR)은 잔차 블록들을 연속적으로 연결하여 깊이를 증가시킴으로써, 네트워크의 성능을 증가시킨다. 다른 예로서, VDSR(Accurate Image Super-Resolution Using Very Deep Convolutional Networks)은 VGG(Visual Geometry Group) 네트워크 기반의 CNN 모델로서, 잔차 프레임을 최종 출력에 더해 주는 방식인 잔차신호 기반 학습(residual learning)을 사용한다. VDSR은 네트워크의 맨 마지막에 잔차 신호를 추가하여, 입력 신호에 잔차 신호를 더한다.

[0041] **II. 암시적 신경망 표현(implicit neural representation) 모델**

[0042] 최근 이미지를 비롯한 각종 데이터의 표현을 신경망 구조로 나타내는 암시적 신경망 표현 모델에 관한 연구가 활발하다. 기존의 비디오 표현 방식은 픽셀 위치별 RGB 픽셀값을 갖는 명시적(explicit) 표현 방식을 사용한다. 이러한 명시적 표현 방식을 대체하기 위해 영상의 픽셀 위치인 (x,y) 좌표에서 (r,g,b) 값을 생성하는 함수를 신경망으로 표현하는 암시적 신경망 표현 모델이 도입되고 있다. 명시적 표현 방식과 비교하여 암시적 신경망 표현 모델은 이미지의 해상도와 관계 없이 사용될 수 있다.

[0043] 암시적 신경망 표현 모델을 이용하는 3차원 비디오 코딩의 일 예로서, NeRF(Neural Radiance Fields) 기술이 존재한다.

[0044] 다시점 비디오의 다수 시점들(view point)에서 촬영한 이미지들을 학습함으로써, NeRF는 임의의 위치와 방향에서 바라본 대상의 모습을 합성한다. 카메라가 바라보는 방향의 레이 정보(ray information)를 샘플링한 3D 위치 (x, y, z) 및 시점 방향(viewing direction) (Θ, ϕ)로 구성된 5차원 좌표를 입력받은 후, NeRF는 입력 좌표에 대응하는 RGB 값 (r, g, b) 및 볼륨 밀도(volume density) σ를 출력한다. 여기서, 3D 위치 (x, y, z)는 월드 좌표계의 위치이고, 시점 방향을 구성하는 Θ, ϕ는 각각 방위각과 고도를 나타낼 수 있다.

[0045] 기존 암시적 신경망 표현 방식과 대비하여 고해상도의 복잡한 장면을 더 잘 표현하기 위해 NeRF는 다음과 같은 개선점을 갖는다. NeRF는 위치 인코딩(positional encoding)을 이용하여 위치 정보를 임베딩한 후, 수학식 1과 같이 임베딩된 인덱스를 입력으로 사용한다.

수학식 1

[0046] $$\gamma(p) = (\sin(2^0\pi p), \cos(2^0\pi p), \ldots, \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p))$$

[0047] 수학식 1의 예시와 같이, 각 공간 좌표는 2L 차원의 벡터 γ(p)로 매핑될 수 있다. 3D의 입력 좌표가 그대로 입력되는 것을 대신하여, sin, cos함수를 이용하여 각 공간 좌표는 2L 차원의 벡터 γ(p)로 매핑될 수 있다. 임베딩된 공간 인덱스를 NeRF 모델의 트레이닝에 이용함으로써, NeRF 모델은 고주파 변이를 포함하는 데이터를 더 잘 예측할 수 있다.

[0048] NeRF는 암시적 신경망 표현 모델로서, 다수의 전연결 레이어(Fully-connected Layer, FC)들을 포함하는 다층신경망(Multi-layer Perceptron, MLP) 네트워크를 사용한다. NeRF는 고차원의 입력 좌표에 대응하는 RGB 값 (r, g, b) 및 볼륨 밀도 σ를 출력함으로써 임의의 장면(scene)을 표현한다. 이후, RGB 값 (r, g, b) 및 볼륨 밀도 σ에 기초하는 볼륨 렌더링 기법에 따라, 장면을 지나는 임의의 레이 r의 색상 C(r)이 렌더링될 수 있다. C(r)은 미분가능(differentiable)하므로, GT(Ground Truth)와 함께, NeRF의 트레이닝에 이용된다. 이때, 각 5차원 좌표에 대해 GT로서, 시점 방향 (Θ, ϕ)을 따라 월드 좌표 (x,y,z)를 촬영한 영상이 사용된다.

[0049] 한편, 렌더링 효율을 향상시키기 위해 NeRF는 단일 MLP를 사용하는 것을 대신하여, 2 개의 망, 즉 coarse 네트워크 및 fine 네트워크를 포함한다. NeRF는 계층적인 샘플링(stratified sampling)을 이용하여 N_c 개의 위치를 샘플링한 후, N_c 개의 위치에서 coarse 네트워크의 출력을 산정한다. coarse 네트워크의 출력으로부터 fine 네트워크에 적용하기 위한 더 많은 샘플들을 생성하기 위해, N_c 개의 coarse 네트워크 출력에 기초하여 복합 색상 $C_c(r)$은 수학식 2와 같이 표현된다.

수학식 2

[0050] $$C_c(r) = \sum_{i=0}^{N_c} w_i c_i$$

[0051] 수학식 2에서, c_i는 N_c 개의 위치에서 산정된 coarse 네트워크의 색상이다. w_i는 대응하는 가중치로서 볼륨 및 샘플링된 위치들 간의 거리에 의존하는 값이다. 역변환 샘플링(inverse transform sampling)을 이용하여 전술한 복합 색상으로부터 높은 가중치를 갖는 N_f 개의 위치를 샘플링한 후, NeRF는 N_c+N_f 개의 위치에서 fine 네트워크의 출력을 산정하고, 산정된 fine 네트워크의 출력에 기초하는 색상 $C_f(r)$를 렌더링할 수 있다. 이때, coarse 네트워크 및 fine 네트워크의 출력에 따른 색상과 GT 간 MSE(Mean Square Error) 손실함수를 이용하여, NeRF가 트레이닝될 수 있다.

[0052] 전술한 바와 같이, 이미지 내의 모든 픽셀들을 월드 좌표계로 재구성하기 위한 학습 과정을 이용하여 NeRF는 임의의 시점에서 렌더링된 결과를 출력한다. 따라서 이미지 좌표계 (x, y)의 한 점으로 투영되는 모든 점들을 월드 좌표계로 변환하는 과정이 필요하므로, NeRF의 입력은 월드 좌표계에서의 인덱스와 더불어 월드 좌표계로의 투영에 필요한 카메라 파라미터들을 포함한다.

[0053] 카메라 파라미터들을 추출하는 과정은, 도 6의 예시와 같이 이미지 내 픽셀들을 초점거리 Z=-1를 갖는 좌표계 [X, Y, Z]를 거쳐서 월드 좌표계 [U, V, W]로 변환한다. 이때, 카메라에 대한 사전 정보가 필수적이며, 사전 정보로서 카메라 파라미터들은 내부(internal) 파라미터들와 외부(external) 파라미터들을 포함한다. 카메라 내부 파라미터들은 초점거리(focal length), 이미지 평면상의 주점(principal point), 및 비대칭 계수 (skew coefficient)를 포함한다, 카메라 내부 파라미터들에 따라 대상은 카메라 좌표계 [X, Y, Z]로 투영될 수 있다. 카메라 좌표계로 투영된 정보를 월드 좌표계 [U, V, W]로 변환하기 위해, 월드 좌표계 내에서 카메라가 어디에 위치하는지에 대한 평행이동(translation) 정보, 및 어디를 바라보고 있는지에 대한 회전(rotation) 정보가 필요하다. 카메라 외부 파라미터들은 전술한 평행이동(translation) 정보 및 회전 정보를 포함한다. 영상을 촬영할 때 회전 행렬(rotation matrix) R, 및 평행이동 벡터(translation vector) t가 결합된 행렬 [R|t]가 월드 좌표계에서 이미지 좌표계로의 투영계산에 사용된다. 따라서, 카메라 외부 파라미터들은 결합 행렬 [R|t]의 역행렬을 산정하여 추출될 수 있다.

[0054] 한편, NeRF의 입력인 5차원 좌표는 추출된 카메라 파라미터들로부터 유도될 수 있다. 카메라 파라미터들은 전술한 바와 같이, 촬영된 2D 영상 및 2D 영상에 대응하는 월드 좌표계 내의 위치들에 기초하여 유도될 수 있다. GT로 사용하기 위한 영상들을 촬영 시, 카메라 내부 파라미터들은 고정될 수 있다. 카메라 내부 파라미터들이 고정된 채로 사용되는 경우, 5차원 좌표는 카메라 외부 파라미터들로부터 유도될 수 있다.

[0055] 다른 예로서, 영상 촬영 시 사용되었던 카메라 파라미터들을 사용하는 방법 외에 colmap과 같이 다시점 이미지들의 특징점을 추출하여 월드 좌표계 내에서 카메라의 위치와 방향에 대한 정보를 재구성하는 방법도 이용될 수 있다.

[0056] 암시적 신경망 표현 모델을 이용하는 2차원 비디오 코딩의 일 예로서, 암시적 신경망 표현 기반의 비디오 코딩(Implicit Neural Representation for Video Coding, INRVC)이 존재한다. INRVC 기술은 영상의 픽셀 위치인 (x,y) 좌표에 대해 (r,g,b) 값을 생성한다. INRVC 기술의 일 예로서, NeRV(Neural Representation for Video) 모델이 존재한다. 전술한 바와 같이, 명시적 표현 방식은, 이미지 내 픽셀의 위치 (x,y)와 프레임의 시간 인덱스 t를 이용하여 (x,y,t) 그리드 상에 픽셀값을 표현한다. 암시적 신경망 표현 모델은, (x,y,t)의 위치를 나타내는 입력으로부터 RGB 픽셀값을 출력한다. 하지만, (x,y,t)의 모든 위치에서 신경망을 트레이닝하는 것은 연산 복잡도를 크게 증가시킬 수 있으므로, 도 7의 예시와 같이 NeRV 모델은 시간 인덱스 t만을 입력으로 하는 신경망 구조를 이용하여 시간 인덱스 t에서 전체 프레임의 RGB 이미지를 출력한다.

[0057] 도 8a 및 도 8b는 본 개시의 일 실시예에 따른, NeRV 모델의 개념을 나타내는 예시도이다.

[0058] 시간 인덱스 t 입력에 대해 이미지 출력을 용이하게 제공하기 위해, 기존 암시적 표현 모델의 MLP 네트워크를 이용하는 것보다 콘볼루션 레이어를 이용하는 것이 효과적일 수 있다. NeRV 모델은 다수의 콘볼루션 레이어들로 구성된 NeRV 블록이 적층된 구조를 이용하여 출력을 생성할 수 있다. 도 8a의 예시와 같이, MLP 기반 출력 및 NeRV 블록 기반 출력이 비교될 수 있다. NeRV 블록은 도 8b의 예시와 같이, 콘볼루션을 수행하는 콘볼루션 레이어들, 픽셀 셔플(pixel shuffle) 레이어, 및 활성화(activation) 레이어를 포함한다. 도 8b의 예시에서, C는 채널의 개수를 나타내고, W, H는 NeRV 블록의 입력의 너비, 높이를 나타내며, S는 업스케일링 인자를 나타낸다.

[0059] 또한, NeRV 모델은 수학식 3과 같이 시간 인덱스 t를 고차원 공간에 임베딩(embedding)한 후, 임베딩된 인덱스를 입력으로 사용한다.

수학식 3

[0060] $$\gamma(t) = (\sin(2^0\pi t), \cos(2^0\pi t), \ldots, \sin(2^{L-1}\pi t), \cos(2^{L-1}\pi t))$$

[0061] 수학식 3의 예시와 같이, 시간 인덱스 t는 2L 차원의 벡터 $\gamma(t)$로 매핑될 수 있다. 임베딩된 시간 인덱스 t를 NeRV 모델의 트레이닝에 이용함으로써, NeRV 모델은 고주파 변이를 포함하는 비디오 데이터를 더 잘 예측할 수 있다.

[0062] NeRV 모델의 손실함수로서, L1 손실과 SSIM(Structural Similarity Index) 손실의 조합이 사용될 수 있다. NeRV 모델의 트레이닝 시, 입력에 기초하여 NeRV 모델에 의해 추정된 이미지, 및 GT(Ground-Truth) 이미지(즉. 원본 이미지)의 모든 픽셀 위치들에서, 전술한 바와 같은 손실이 산정될 수 있다. L1 손실은 추정된 이미지와 GT 이미지에서 픽셀들 간 차이의 절대값을 이용하여 산정된다. SSIM 손실은 추정된 이미지와 GT 이미지에서 픽

셀들의 평균, 표준편차 및 상관도(correlation)에 기초하여 산정된다. 이후, 산정된 손실을 기반으로 트레이닝 과정에서 NeRV 모델을 구성하는 가중치들(또는 파라미터들)이 업데이트될 수 있다.

[0063] 도 9는 본 개시의 일 실시예에 따른, NeRV 모델의 부호화 파이프라인을 나타내는 예시도이다.

[0064] NeRV 모델을 이용하여 원본 비디오를 근사적으로 표현할 수 있으므로, NeRV 모델을 압축하여 전송함으로써, 신경망을 이용한 종단간(end-to-end) 압축 기술이 구현될 수 있다. NeRV 모델을 부호화하기 위한 부호화 파이프라인은 도 9의 예시와 같이, 비디오 오버피터(video overfitter, 910), 모델 프루너(model pruner, 920), 모델 양자화부(model quantizer, 930), 및 가중치 부호화부(weight encoder, 940)의 전부 또는 일부를 포함할 수 있다.

[0065] 비디오 오버피터(910)는 입력된 비디오 프레임을 NeRV 모델로 표현한다. 이때, 영상 부호화 장치는 전술한 손실함수를 이용하여 NeRV 모델을 트레이닝함으로써, NeRV 모델로 하여금 비디오 프레임을 표현하도록 한다. 모델 프루너(920)은 MLP 또는 MLP와 콘볼루션 레이어들이 혼합된 형태를 갖는 NeRV 모델 구조를 프루닝을 이용하여 단순화한다. 예컨대, 프루닝은 기설정된 임계치보다 작은 가중치를 영으로 설정한다. 모델 양자화부(930)은 프루닝이 적용된 가중치들(예컨대, 임계치 이상의 가중치들)을 양자화한다. 가중치 부호화부(940)은 양자화된 가중치들에 엔트로피 부호화를 적용하여 가중치들의 비트스트림을 생성한다.

[0066] 한편, 도 9에 예시된 부호화 파이프라인의 역순으로서, NeRV 모델을 복호화하기 위한 복호화 파이프라인은 도 10과 같이 예시될 수 있다. 복호화 파이프라인은 도 10의 예시와 같이, 가중치 복호화부(weight decoder, 1010), 모델 역양자화부(model dequantizer, 1020), 모델 복원부(model reconstructor, 1030), 및 비디오 생성부(video generator, 1040)의 전부 또는 일부를 포함할 수 있다.

[0067] 가중치 복호화부(1010)는 NeRV 모델의 가중치들의 비트스트림에 엔트로피 부호화를 적용하여 양자화된 가중치들을 생성한다. 모델 역양자화부(1020)는 양자화된 가중치들을 역양자화하여 프루닝이 적용된 가중치들을 생성한다. 모델 복원부(1030)는 프루닝이 적용된 가중치들에 기초하여 NeRV 모델을 복원한다. 비디오 생성부(1040)는 복원된 NeRV 모델을 이용하여 시간 인덱스에 따라 복원 비디오 프레임을 생성한다.

[0068] NeRV 모델은 전술한 바와 같이, 프레임의 시간 인덱스 t를 입력으로 이용한다. 따라서, t×h×w 크기의 픽셀들을 갖는 비디오가 입력되는 경우, NeRV 모델은 t×h×w 번이 아니라 t번만 입력 비디오를 샘플링하므로, 인코딩 속도 및 디코딩 속도 측면에서 모두 큰 이득이 기대될 수 있다.

[0069] 한편, 전술한 바와 같은 부호화 파이프라인 및 복호화 파이프라인은 3D INVR 모델의 압축 및 복원에도 이용될 수 있다.

[0070] 이하, 3D INVR 모델과 INVR 모델은 호환적으로 사용된다.

[0071] **III. 본 개시에 따른 실시예들**

[0072] 도 1에 예시된 영상 부호화 장치는 다양한 시점을 갖는 3D 비디오를 이용하여 3D INVR 부호화부(110) 내 3D INVR 모델을 트레이닝함으로써, 3D INVR 모델로 하여금 다양한 시점들의 3D 비디오를 생성할 수 있도록 한다. 이때, 3D INVR 모델의 입력으로는 다양한 시점들을 나타내는 정보가 사용되고, GT(Ground Truth)로는 각 시점에 대응하는 이미지가 사용된다. 트레이닝부는 3D INVR 모델의 출력과 GT 간 차이를 이용하여 3D INVR 모델의 파라미터들(가중치들)을 업데이트함으로써, INVR 모델로 하여금 다양한 시점의 3D 비디오를 생성하는 것을 학습하도록 한다. 학습이 완료된 3D INVR 모델은 모델 내 파라미터들에 기초하여 3D 비디오가 획득한 3D 공간 정보를 포함할 수 있다. 이하, 입력으로 사용되는 시점을 나타내는 정보를 시점 인덱스로 명칭한다.

[0073] 신경망 압축기(120)는 트레이닝된 3D INVR 모델의 파라미터들을 압축하여 비트스트림을 생성한다. 예컨대, 도 9에 부호화 파이프라인을 이용하여 3D INVR 모델이 압축될 수 있다. 영상 부호화 장치는 생성된 비트스트림을 영상 복호화 장치로 전송한다.

[0074] 3D INVR 모델의 신경망 구조는 NeRF를 이용하거나, NeRF에 부가적인 신경망 모델을 추가하거나, NeRF의 일부를 수정하여 구현될 수 있다. 예를 들어, NeRF는 전술한 바와 같이 MLP로 구현될 수 있다. 또는, NeRF는 MLP와 CNN의 조합으로 구성될 수 있다. 이때, NeRF는 MLP와 CNN의 조합을 반복적으로 포함할 수 있다.

[0075] NeRF 기반의 3D INVR 모델을 사용하는 경우, 영상 부호화 장치는 입력 인덱스로서, 전술한 바와 같은 5차원 좌표를 사용한다. 또한, 각 5차원 좌표에 대해 GT로서 시점 방향 (θ, ϕ)을 따라 월드 좌표 (x,y,z)를 촬영한 영상이 사용된다. 전술한 바와 같이, 5차원 좌표는 GT에 대응하는 카메라 파라미터들로부터 산정될 수 있다. 카메

라 내부 파라미터들이 고정된 채로 사용되는 경우, 5차원 좌표는 GT에 대응하는 카메라 외부 파라미터들로부터 산정될 수 있다. 3D INVR 모델의 출력에 해당하는 3D 비디오는, 5차원 좌표에 기초하여 3D INVR 모델이 월드 좌표계에 렌더링하는 영상, 또는 렌더링된 영상에 대응하는 이미지 좌표계에서의 영상을 나타낸다. 예컨대, 상이한 시점들을 갖는 복수의 렌더링된 영상들에 의해 3D 비디오가 구성될 수 있다.

[0076] 3D INVR 모델의 구현 시 MLP와 CNN의 조합의 개수를 증가시키거나 감소시킴으로써, 본 개시는 비트스트림의 크기를 결정하는 신경망의 크기와 3D 비디오의 공간 재현력의 충실도 간 균형(trade off)을 조정할 수 있다. 본 개시는, 해당 조합의 선택이 율왜곡에 어떠한 영향을 주는지를 검증하고, 검증 결과에 따라 해당 조합을 적응적으로 선택한다. 예컨대, MLP와 CNN의 조합의 개수가 증가할수록, 비트율이 증가하고 재현력의 충실도가 증가한다. 하지만, 조합의 개수가 특정 임계치 이상으로 지나치게 증가하면 재현력의 충실도가 급감할 수 있다. 따라서, 영상 부호화 장치는 다양한 조합을 포함하는 3D INVR 모델의 구조를 구현하여 최적의 성능을 제공하는 구조를 찾은 후, 최적의 성능을 제공하는 3D INVR 구조에 관한 정보를 영상 복호화 장치로 전달한다. 최적의 성능을 제공하는 3D INVR 구조로서, 전체 모델의 구조와 파라미터들이 전달될 수 있으며, 그외에 다음과 같은 방식도 사용될 수 있다.

[0077] 일 예로서, 3D INVR 구조는 영상 부호화 장치와 영상 복호화 장치에서 공유되고, 파라미터들이 별도로 전달될 수 있다. 예를 들어, 영상 부호화 장치와 영상 복호화 장치는, 임의의 M 개의 MLP와 임의의 N 개의 CNN를 포함하는 M+N 모듈을 갖는 3D INVR 모델의 구조로부터 파생된 M, M+1, M+2, … , M+N 개의 조합에 따라 상이하게 구성된 3D INVR 구조를 공유한다. 영상 부호화 장치는 어떠한 구조가 가장 최적의 성능을 제공하는지 선택하고, 선택된 3D INVR 구조의 인덱스를 결정한다. 영상 부호화 장치는 트레이닝된 파라미터들과 함께 인덱스를 영상 복호화 장치로 전송한다. 영상 복호화 장치는 전송받은 인덱스를 이용하여 3D INVR 구조를 결정하고, 복호화한 파라미터들을 적용하여 3D INVR 모델을 복원한다.

[0078] 다른 예로서, 신경망 압축기(120)는 양자화 파라미터를 이용하여 비트스트림의 크기와 재현력을 조정할 수 있다. 본 개시는 해당 양자화 파라미터의 선택이 율왜곡에 어떠한 영향을 주는 것인지를 검증하고, 검증 결과에 따라 해당 양자화 파라미터를 적응적으로 선택한다. 예컨대, 양자화 파라미터의 값이 커질수록 비트율이 감소한다. 하지만, 양자화 파라미터의 값이 특정 임계치 이상으로 커지면(즉, 양자화 스텝의 크기가 커지면), 재현력의 충실도가 급감할 수 있다. 따라서, 영상 부호화 장치는 다양한 양자화 파라미터들을 사용하는 3D INVR 모델을 구현하여 최적의 성능을 제공하는 양자화 파라미터를 찾은 후, 찾아진 양자화 파라미터를 영상 복호화 장치로 전달한다.

[0079] 3D INVR 모델은 시점과 시점 사이에 있는 임의의 중간 시점 비디오를 출력하기 위해, 카메라 파라미터들을 필요로 한다. 전술한 바와 같이, 카메라 파라미터들은 이미지 좌표계에서 카메라 좌표계로 변환하기 위한 카메라 내부 파라미터들과 월드 좌표계에서 카메라의 위치와 바라보는 방향을 명시하는 외부 파라미터들을 포함한다. 카메라 내부 파라미터들은 초점거리(focal length), 카메라 주점(principal point), 및 비대칭 계수(skew coefficient)를 포함한다. 카메라 외부 파라미터들은 회전(rotation) 계수 및 평행이동(translation) 계수를 포함한다.

[0080] 영상 부호화 장치는 3D INVR 모델의 트레이닝 및 부호화에 사용된, 주변 시점들에 대응하는 카메라 외부 파라미터들을 보간(interpolation)하고, 보간된 카메라 외부 파라미터들을 이용하여 임의의 중간 시점을 나타내는 5차원 좌표를 새로 생성할 수 있다. 이때, 카메라 내부 파라미터들은 고정된 것으로 간주한다. 영상 부호화 장치는 새로운 5차원 좌표, 및 대응하는 GT를 3D INVR 모델의 트레이닝에 추가로 사용할 수 있다. 이때, 새로운 5차원 좌표에 대응하는 GT는, 예컨대, 보간에 사용된 카메라 외부 파라미터들에 대응하는 GT들을 보간하여 생성될 수 있다. 구체적으로, 영상 부호화 장치는 주변 시점들에 대응하는 GT들을 카메라 파라미터들을 이용하여 월드 좌표계로 이동시킨 후, 이동된 GT들을 월드 좌표계에서 보간하여 새로운 GT를 생성한다. 영상 부호화 장치는 보간된 카메라 파라미터들을 이용하여 새로운 GT를 이미지 좌표계로 이동시킴으로써, 새로운 5차원 좌표에 대한 GT를 생성할 수 있다.

[0081] 영상 부호화 장치는 3D INVR 모델의 트레이닝 및 부호화에 사용된 5차원 좌표의 구성요소들의 상한 및 하한을 결정함으로써, 3D INVR 모델의 입력 범위를 결정한다. 즉, 학습이 완료된 3D INVR 모델은 입력 범위 내의 5차원 좌표에 대응하는 3차원 비디오를 생성할 수 있다. 영상 부호화 장치는 3D INVR 모델의 입력 범위를 영상 복호화 장치로 전송한다. 다른 예로서, 영상 부호화 장치와 영상 복호화 장치 간 약속에 따라, 기설정된 입력 범위가 사용될 수 있다.

[0082] 한편, 3D INVR이 3차원 공간 정보를 재현하기 위해 사용되나, 영상 부호화 장치는 추가적으로 기존의 깊이 정보

(depth map)를 함께 전달하고, 영상 부호화 장치와 영상 복호화 장치는 깊이 정보를 함께 사용할 수 있다. 영상 부호화 장치는 깊이 정보를 손실/무손실 압축한 후, 영상 복호화 장치로 전송할 수 있다. 또는, 영상 부호화 장치에 의한 깊이 정보의 전송 없이, 영상 복호화 장치가 깊이 정보를 생성/유도할 수 있다.

[0083] 결론적으로, 영상 부호화 장치는 부가 정보로서 INVR 모델의 트레이닝 및 부호화에 사용된 카메라 파라미터들, 및 3D INVR 모델의 압축에 적용된 양자화 파라미터를 전송한다. 또한, 영상 부호화 장치는 부가 정보로서 3D INVR 모델의 구조에 관한 정보, 3D INVR 모델의 입력 범위, 및 깊이 정보와 관련된 정보를 전송할 수 있다. 여기서, 깊이 정보와 관련된 정보는 깊이 정보의 전달 여부를 지시하는 신택스를 포함할 수 있다.

[0084] 도 2에 예시된 영상 복호화 장치는 신경망 역압축기(210)를 이용하여 비트스트림으로부터 3D INVR 모델의 파라미터들을 복호화한다. 예컨대, 도 10에 예시된 복호화 파이프라인을 이용하여 3D INVR 모델이 복원될 수 있다. 영상 복호화 장치는 부가 정보로서 INVR 모델의 트레이닝 및 부호화에 사용된 카메라 파라미터들, 및 INVR 모델의 압축에 적용된 양자화 파라미터를 복호화한다. 영상 복호화 장치는 부가 정보로서 3D INVR 모델의 구조에 관한 정보, 3D INVR 모델의 입력 범위, 및 깊이 정보와 관련된 정보를 복호화한다. 또한, 깊이 정보의 전달 여부를 지시하는 신택스에 따라 깊이 정보가 전송된 경우, 영상 복호화 장치는 추가적으로 깊이 정보를 복호화한다.

[0085] 3D INVR 복호화부(110)는 복원된 3D INVR 모델을 이용하여 다양한 시점 인덱스에서 3D 비디오를 생성한다. 영상 복호화 장치는 임의의 중간 시점을 획득한 후, 임의의 중간 시점을 3D INVR 모델에 입력하여 임의의 중간 시점에 대응하는 3D 비디오를 생성한다.

[0086] 예컨대, NeRF 기반의 3D INVR 모델을 사용하는 경우, 영상 복호화 장치는 임의의 중간 시점을 나타내는 입력 인덱스로서, 전술한 바와 같은 5차원 좌표를 사용한다. 각 5차원 좌표에 대해 3D INVR 모델은 시점 방향 (Θ, ϕ)을 따라 (x,y,z)를 바라보는 영상을 재생한다. 이때, 임의의 중간 시점을 나타내는 5차원 좌표는 입력 범위 내에 존재해야 한다. 예컨대, 5차원 좌표가 입력 범위 내에 존재하지 않는 경우, 영상 복호화 장치는 3D 비디오를 제공하지 않을 수 있다.

[0087] 전술한 부가 정보를 전달하기 위해, 다음과 같은 신택스 요소들을 포함하는 INVR 파라미터 셋(INVR Parameter Set, IPS)이 정의된다.

[0088] 3D INVR 모델의 구조와 관련된 정보는, 모델을 구성하는 MLP와 CNN의 크기 및 개수를 정의하는 신택스, 모델을 구성하고 있는 연결(edge)의 구조를 정의하는 신택스, 모델을 구성하는 파라미터들의 형식 및 정밀도(precision)을 정의하는 신택스, 모델을 구성하는 MLP와 CNN의 크기 및 개수에 관한 상한과 하한을 정의하는 신택스, 모델을 구성하는 연결의 개수에 관한 상한과 하한을 정의하는 신택스, 모델을 구성하는 파라미터들의 정밀도에 관한 상한과 하한을 정의하는 신택스 등을 포함한다.

[0089] 카메라 파라미터들과 관련된 정보는, 전술한 바와 같은 내부 파라미터들 및 카메라 외부 파라미터들을 포함한다.

[0090] 신경망 압축기(120)의 양자화 파라미터와 관련된 정보는, 양자화 파라미터의 사용 가능한 크기에 관한 상한과 하한을 정의하는 신택스를 포함한다.

[0091] 깊이 정보와 관련된 정보는 깊이 정보의 전달 여부를 지시하는 신택스를 포함한다.

[0092] 입력 범위와 관련된 정보는 시점의 상한과 하한을 정의하는 신택스를 포함한다.

[0093] 이하, MLP와 CNN의 조합의 개수를 증가시키거나 감소시킴으로써 비트스트림의 크기를 결정하는 신경망의 크기와 3차원 비디오의 공간 재현력의 충실도 간 균형(trade off)를 조정하는 방식을 설명한다. 다음에 설명된 압축 설명 방법의 선택, 또는 여러 압축 방법들의 조합에 따라, 영상 부호화 장치는 적응적으로 3D INVR 모델을 압축할 수 있다.

[0094] 영상 부호화 장치는, MLP 모듈 또는 CNN 모듈의 개수를 조정하여 INRV 모델의 전체적인 성능, 압축된 INRV 모델의 성능, 트레이닝 속도 등을 결정할 수 있다. 모듈의 개수를 증가시켜 3D INRV 모델의 크기가 커짐에 따라 재현력의 충실도가 콘벡스(convex)한 형태로 증가한다. 따라서, 영상 부호화 장치는 3D INRV 모델의 성능 및 크기에 대한 최적화 지점을 찾을 수 있다. 또한, 태스크에 따라 모듈의 개수를 적응적으로 조절함으로써, 영상 부호화 장치는 태스크별 최적을 결과를 도출할 수 있다. 출력되는 결과의 품질이 높아야 하는 경우, 영상 부호화 장치는 모듈의 개수를 증가시켜 3D INRV 모델을 구성한다. 하지만, 모듈의 개수가 증가하면, INRV 모델의 복잡도가 증가하여 INRV 모델의 트레이닝에 긴 시간이 소모될 수 있다. 모듈의 개수가 지나치게 증가되면, 재현력의 충실도가 급감할 수 있다. 또한, 가벼운 모델과 비교하여 3D INRV 모델이 복잡할수록 동일한 bpp(bits per

pixel)로 압축 시 3D INRV 모델의 성능이 감소한다. 따라서, 압축의 편이, 트레이닝 속도의 증가 등을 위해 비트율이 작아야 하는 경우, 영상 부호화 장치는 모듈의 개수를 감소시킨다. 여기서, bpp는 학습이 완료된 INVR 모델을 부호화하여 생성된 비트스트림을 영상 부호화 장치가 제공하는 3D 비디오의 해상도로 나눈 값으로 정의될 수 있다.

[0095] 일 예로서, 전술한 NeRF는 연속된 MLP 모듈 10 개로 구성될 수 있다. 다른 예로서, NeRV는 1 개의 MLP와 5 개의 CNN으로 구성될 수 있다. 다른 예로서, HNeRV(Hybrid Neural Representation for Videos)는 CNN과 MLP의 조합 5 개, 및 추가적인 CNN 6 개로 구성된다. 전술한 바와 같이, 3D INRV 모델의 크기가 증가할수록 성능은 콘벡스(convex)한 형태로 증가하므로, 영상 부호화 장치는 성능이 수렴할 때까지 모듈의 개수를 증가시킬 수 있다. 예를 들어, 3D INRV 모델의 중간에 CNN을 추가하여, 3D INRV 모델의 성능이 향상될 수 있다. 실험적으로, 3D INRV 모델의 중간에 추가되는 CNN의 개수가 10 개로 증가할 때까지, 3D INRV 모델의 성능이 향상된다.

[0096] 이하, MLP 및 CNN의 개수, 또는 파라미터들의 개수를 적응적으로 감소시키는 방법을 기술한다.

[0097] 도 11은 본 개시의 일 실시예에 따른, 비트 마스크를 이용한 프루닝을 나타내는 예시도이다.

[0098] 3D INRV 모델의 트레이닝 과정에서, 영상 부호화 장치는 프루닝(pruning)을 이용하여 파라미터들의 개수를 감소시킬 수 있다. 프루닝의 일 예로서, 비트 마스크(bit mask), 즉 0과 1로 이루어진 이진 마스크가 사용된다. 영상 부호화 장치는 비트 마스크를 학습한 후, 도 11의 예시와 같이 INRV 모델의 파라미터들에 승산한다. 비트 마스크는 트레이닝에 사용되는 가중치들을 결정하고, 손실함수의 역전파에 따라 업데이트된다.

[0099] 프루닝의 다른 예로서, 3D INRV 모델의 파라미터들 중에서 파라미터의 절대값이 기설정된 임계치 이하인 경우, 해당되는 파라미터를 0으로 설정하는 방식이 사용될 수 있다. 프루닝의 다른 예로서, 파라미터 행렬 내에서 기설정된 위치에서를 벗어나는 위치의 가중치들을 파라미터를 0으로 설정하는 방식이 사용될 수 있다. 예컨대, K×L 파라미터 행렬 내에서 S×T 영역에 위치하는 파라미터들만을 남기고 나머지 파라미터들은 0으로 프루닝될 수 있다(여기서, K, L, S, T는 임의의 양의 정수로서, S≤K, T≤L을 만족함).

[0100] 3D INRV 모델의 트레이닝 과정에서, 영상 부호화 장치는 비디오 공간 재현도 외에, 파라미터들의 크기를 조정하여 비트스트림의 크기를 조정하는 요소를 포함하는 손실 함수를 사용한다. 일 예로서, 파라미터들의 희소 손실(sparsity loss)를 이용하여 3D INRV 모델의 파라미터들이 조정될 수 있다. 희소 손실은 수학식 4와 같이 정의된다.

수학식 4

[0101] $$L_{sparsity}(w) = \frac{|w|_1}{|w|_2} + \frac{|w|_2^2}{|w|_1}$$

[0102] 수학식 4에서, $|w|_1$과 $|w|_2$는 L1 놈(norm)과 L2 놈을 나타낸다.

[0103] 희소 손실을 포함하는 손실 함수에 기초하여 트레이닝된 3D INRV 모델의 율왜곡을 측정하여, 영상 부호화 장치는 최적의 MLP 및 CNN의 개수, 또는 최적의 파라미터들의 개수를 결정한다. 수학식 4에 나타낸 희소 손실은 일 예일 뿐이고, 이 외에도 다양한 희소 손실의 산정 방법이 사용될 수 있다.

[0104] 3D INRV 모델의 트레이닝 후, 신경망 압축기(120)는, 예컨대 도 9에 예시된 부호화 파이프라인을 이용하여 3D INRV 모델의 파라미터들을 압축한다. 신경망 압축기(120)는, 희소화(sparsification), 프루닝(pruning), 단일화(unification), 국부적 스케일링(local scaling) 등을 위한 파라미터들을 이용하여, 비트스트림의 크기와 최종 INVR 모델의 성능 간의 트레이드오프(trade off)를 조정한다. 예컨대, 모델의 파라미터들이 행렬로 구성된 후, 파라미터 행렬은 임의의 다수의 하위 블록들로 분할될 수 있다. 단일화는, 일부 하위 블록들에 하나의 파라미터 값을 설정하는 동작을 나타낸다. 국부적 스케일링은 분할된 파라미터 행렬에서 일부 하위 블록들의 파라미터 값들을 스케일링하는 동작을 나타낸다.

[0105] 일 예로서, 영상 부호화 장치는 무손실 압축을 이용하여 IPS를 전송할 수 있다.

[0106] 이하, 디코더 측에서 요구하는 임의 시점의 비디오를 적응적 3D INVR 모델로 생성한 후, 그 품질을 향상시키기 위한 기술을 설명한다. 영상 복호화 장치는 다른 시점의 프레임들 간의 중복성을 이용하여 현재 시점을 기준으

로 다른 시점의 이미지를 참조함으로써 현재 시점의 화질을 개선한다. 본 개시에서는, RGB 픽셀들 간 참조와 3D INVR 모델의 특징(feature) 단에서의 참조가 제공된다. 일 예로서, 참조를 하기 위해, 영상 복호화 장치는 깊이 정보를 사용하여 참조 카메라 좌표계를 월드 좌표계로 위치시킨 후, 월드 좌표계를 원하는 시점의 카메라 좌표계로 변환한다. 이때, 영상 복호화 장치는 영상 부호화 장치에 의해 전송된 부가 정보인 깊이 정보를 복호화하고, 복호화된 깊이 정보를 사용한다. 또는, 영상 복호화 장치는 깊이 정보를 유도하고, 유도된 깊이 정보를 사용할 수 있다. 다른 예로서, 영상 복호화 장치는 카메라 파라미터들을 이용하여 참조 이미지의 시점을 변환할 수 있다.

[0107] 도 12는 본 개시의 일 실시예에 따른, 참조 이미지의 와핑을 나타내는 예시도이다.

[0108] 일 예로서, 영상 복호화 장치는 RGB 픽셀의 와핑(warping)을 이용하여 현재 시점의 화질을 개선한다. 도 12의 예시와 같이, RGB 픽셀의 와핑은 해당되는 깊이 정보(또는, 카메라 파라미터들)를 이용하여 참조 이미지의 좌표계를 월드 좌표계로 이동시키고, 월드 좌표계를 현재 이미지의 좌표계로 변환하는 과정을 나타낸다. 영상 복호화 장치는 전술한 와핑을 이용하여 현재 시점에 대한 참조 이미지를 생성한다. 영상 복호화 장치는 현재 시점으로 와핑된 참조 이미지와 현재 이미지의 가중합을 생성함으로써, 개선된 현재 이미지를 생성한다. 여기서, 현재 이미지는 현재 시점에 대응하는 이미지를 나타내고, 참조 이미지는 참조 시점에 대응하는 이미지를 나타낸다. 도 12의 예시에서, H는 와핑에 기초하는 참조 이미지 내 RGB 픽셀의 이동을 나타낸다.

[0109] 도 13은 본 개시의 일 실시예에 따른, 참조 이미지/특징의 와핑을 나타내는 예시도이다.

[0110] 다른 예로서, 영상 복호화 장치는 특징의 와핑을 이용하여 현재 시점의 화질을 개선할 수 있다. 이때, 영상 복호화 장치는 시점들 간 와핑을 3D INVR 모델의 특징 단에서 수행한다. 특징의 와핑 방법은 신경망(Neural Network, NN) 기반으로 MLP를 사용하거나, RGB 와핑과 동일하게 깊이 정보(또는, 카메라 파라미터들)을 기반으로 월드 좌표계와 카메라 좌표계 간 이동을 이용한다. 예컨대, NeRF 모델의 MLP 마지막 단이 참조 특징으로 이용될 수 있다. NeRV 모델의 마지막 단에 해당하는 NeRV 블록이 참조 특징으로 이용될 수 있다. 일 예로서, 도 13의 예시와 같이, CNN을 사용하여 와핑된 참조 특징/이미지와 현재 특징/이미지가 융합(fusion)될 수 있다. 도 13의 예시에서, 와핑(warping)은 참조 특징/이미지의 좌표를 현재 특징/이미지의 좌표로 변환하는 과정을 나타낸다. 영상 복호화 장치는 와핑된 참조 특징/이미지와 현재 특징/이미지를 결합(concatenation)시킨 후, 결합된 특징을 CNN에 입력할 수 있다.

[0111] GoI(Group of INVR)는 시점 및/또는 시간에 따라 3D INVR을 구성하는 방식을 나타낸다. 영상 부호화 장치는 GoI별로 IPS를 전달한다. 영상 복호화 장치는 GoI별로 IPS를 복호화하여 3D INVR 모델 및 관련된 파라미터들을 설정할 수 있다.

[0112] 도 14는 본 개시의 일 실시예에 따른, GoI의 정의를 나타내는 예시도이다.

[0113] 도 14에 예시된 바와 같이, 일 예로서, 하나의 3D INVR 모델이 하나의 시간 t에서 복수의 시점들을 표현하도록 GoI가 구성된다. 다른 예로서, 하나의 3D INVR 모델이 하나의 시점 v에서 복수의 시간들을 표현하도록 GoI가 구성될 수 있다. 다른 예로서, 하나의 3D INVR 모델이 복수의 시간들에서 복수의 시점들을(또는, 복수의 시점들에서 복수의 시간들을) 표현하도록 GoI가 구성될 수 있다.

[0114] 이하, 도 15 및 도 16의 도시를 이용하여, 3D INVR 기반 3D 비디오를 부호화하거나 복원하는 방법을 기술한다.

[0115] 도 15는 본 개시의 일 실시예에 따른, 영상 부호화 장치가 3D 비디오들을 부호화하는 방법을 나타내는 순서도이다.

[0116] 영상 부호화 장치는 3D 비디오들, 및 3D 비디오들에 대응하는 카메라 파라미터들을 획득한다(S1500).

[0117] 영상 부호화 장치는 3D INVR 모델을 구성한다(S1502).

[0118] 여기서, 3D INVR 모델은 다층신경망과 CNN을 포함한다. 영상 부호화 장치는, 예컨대, MLP과 CNN의 크기 및 개수에 기초하여 3D INVR 모델의 구조를 상이하게 구성할 수 있다.

[0119] 영상 부호화 장치는 카메라 파라미터들에 대응하는 시점들 및 카메라 파라미터들에 대응하는 3D 비디오들을 이용하여 3D INVR 모델의 트레이닝을 수행한다(S1504).

[0120] 일 예로서, 3D INVR 모델로서 NeRF가 이용되는 경우, 시점들을 구성하는 요소들은, 5차원 좌표를 나타낸다. 여기서, 5차원 좌표는 월드 좌표 (x,y,z) 및 시점 방향 (Θ, ϕ)를 포함한다. 영상 부호화 장치는 각 3D 비디오에 대응하는 카메라 파라미터들을 이용하여 해당되는 시점의 구성요소들을 추출할 수 있다.

[0121] 영상 부호화 장치는 시점들을 3D INVR 모델에 입력하여 출력을 생성한다. 영상 부호화 장치는 3D INVR 모델의 출력과 GT(Ground Truth) 간 차이를 산정한다. 여기서, GT로서 전술한 단계에서 획득된 3D 비디오들이 사용된다. 영상 부호화 장치는 전술한 차이, 및 희소 손실(예컨대, 수학식 4의 정의에 따른)을 포함하는 손실 함수에 기초하여 3D INVR 모델의 파라미터들을 업데이트한다.

[0122] 트레이닝 과정에서, 영상 부호화 장치는 프루닝(pruning)을 이용하여 3D INVR 모델의 파라미터들의 개수를 감소시킬 수 있다.

[0123] 영상 부호화 장치는 트레이닝에 사용된 시점들에 기초하여 3D INVR 모델의 입력 범위를 결정한다(S1506).

[0124] 3D INVR 모델의 트레이닝에 사용된 시점들에 대해, 전술한 시점들을 구성하는 요소들(예컨대, 5차원 좌표) 각각의 상한과 하한을 결정함으로써, 영상 부호화 장치는 3D INVR 모델의 입력 범위를 결정한다.

[0125] 영상 부호화 장치는 트레이닝의 결과에 기초하여 양자화 파라미터, 및 3D INVR 모델의 구조에 관한 정보를 결정한다(S1508).

[0126] 영상 부호화 장치는 율왜곡 및 트레이닝의 속도 측면에서 최적인 3D INVR 모델의 구조를 선택할 수 있다. 또한, 영상 부호화 장치는 율왜곡 최적화 측면에서 양자화 파라미터를 적응적으로 결정할 수 있다.

[0127] 영상 부호화 장치는 양자화 파라미터를 이용하여 3D INVR 모델의 파라미터들을 부호화한다(S1510). 예컨대, 영상 부호화 장치는 도 9에 예시된 바와 같은 부호화 파이프라인을 이용하여 3D INVR 모델의 파라미터들을 압축할 수 있다.

[0128] 영상 부호화 장치는 IPS를 부호화한다(S1512).

[0129] 여기서, IPS(INVR Parameter Set)는 부가 정보로서 양자화 파라미터, 3D INVR 모델의 구조에 관한 정보, 및 3D INVR 모델의 입력 범위에 관한 정보를 포함한다. 또한, IPS는 부가 정보로서 카메라 파라미터들, 및 깊이 정보(depth map)의 전달 여부를 지시하는 신택스를 추가로 포함할 수 있다.

[0130] 깊이 정보의 전달 여부를 지시하는 신택스가 깊이 정보의 전달을 지시하는 경우, 영상 부호화 장치는 깊이 정보를 추가로 부호화할 수 있다.

[0131] 도 16은 본 개시의 일 실시예에 따른, 영상 복호화 장치가 현재 3D 비디오를 복원하는 방법을 나타내는 순서도이다.

[0132] 영상 복호화 장치는 비트스트림으로부터 3D INVR 모델의 파라미터들, 및 IPS(INVR Parameter Set)를 복호화한다(S1600).

[0133] IPS(INVR Parameter Set)는 부가 정보로서 양자화 파라미터, 3D INVR 모델의 구조에 관한 정보, 및 3D INVR 모델의 입력 범위에 관한 정보를 포함한다. 또한, IPS는 부가 정보로서 카메라 파라미터들, 및 깊이 정보(depth map)의 전달 여부를 지시하는 신택스를 추가로 포함할 수 있다.

[0134] 영상 복호화 장치는 3D INVR 모델의 파라미터들, 및 양자화 파라미터를 이용하여 3D INVR 모델의 구조에 관한 정보에 부합하도록 3D INVR 모델을 복원한다(S1602).

[0135] 여기서, 3D INVR 모델은 다층신경망과 CNN을 포함한다. 영상 복호화 장치는, 예컨대, 3D INVR 모델의 구조에 관한 정보를 이용하여 MLP과 CNN의 크기 및 개수를 결정할 수 있다.

[0136] 영상 복호화 장치는 입력 범위 내에서 임의의 시점을 획득한다(S1604).

[0137] 영상 복호화 장치는 임의의 시점을 3D INVR 모델에 입력하여 현재 3D 비디오를 생성한다(S1606).

[0138] 일 예로서, 영상 복호화 장치는 와핑을 이용하여 현재 시점에 대한 참조 이미지/특징을 생성한다. 여기서, 와핑은, 카메라 파라미터들을 이용하여 참조 이미지/특징의 좌표계를 월드 좌표계로 이동시기고, 참조 이미지/특징의 월드 좌표계를 현재 3D 비디오의 좌표계로 변환하는 과정을 나타낸다. 이후, 영상 복호화 장치는 와핑이 적용된 참조 이미지/특징과 현재 3D 비디오/특징을 가중합함으로써, 개선된 현재 3D 비디오를 생성한다.

[0139] 다른 예로서, 깊이 정보(depth map)의 전달 여부를 지시하는 신택스가 깊이 정보의 전달을 지시하는 경우, 영상 복호화 장치는 깊이 정보를 복호화한다. 영상 복호화 장치는 와핑을 이용하여 현재 시점에 대한 참조 이미지/특징을 생성한다. 여기서, 와핑은, 깊이 정보를 이용하여 참조 이미지/특징의 좌표계를 월드 좌표계로 이동시기고, 참조 이미지/특징의 월드 좌표계를 현재 3D 비디오의 좌표계로 변환하는 과정을 나타낸다. 이후,

영상 복호화 장치는 와핑이 적용된 참조 이미지/특징과 현재 3D 비디오/특징을 가중합함으로써, 개선된 현재 3D 비디오를 생성한다.

[0140] 본 명세서의 흐름도/타이밍도에서는 각 과정들을 순차적으로 실행하는 것으로 기재하고 있으나, 이는 본 개시의 일 실시예의 기술 사상을 예시적으로 설명한 것에 불과한 것이다. 다시 말해, 본 개시의 일 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 개시의 일 실시예의 본질적인 특성에서 벗어나지 않는 범위에서 흐름도/타이밍도에 기재된 순서를 변경하여 실행하거나 각 과정들 중 하나 이상의 과정을 병렬적으로 실행하는 것으로 다양하게 수정 및 변형하여 적용 가능할 것이므로, 흐름도/타이밍도는 시계열적인 순서로 한정되는 것은 아니다.

[0141] 이상의 설명에서 예시적인 실시예들은 많은 다른 방식으로 구현될 수 있다는 것을 이해해야 한다. 하나 이상의 예시들에서 설명된 기능들 혹은 방법들은 하드웨어, 소프트웨어, 펌웨어 또는 이들의 임의의 조합으로 구현될 수 있다. 본 명세서에서 설명된 기능적 컴포넌트들은 그들의 구현 독립성을 특히 더 강조하기 위해 "...부(unit)" 로 라벨링되었음을 이해해야 한다.

[0142] 한편, 본 실시예에서 설명된 다양한 기능들 혹은 방법들은 하나 이상의 프로세서에 의해 판독되고 실행될 수 있는 비일시적 기록매체에 저장된 명령어들로 구현될 수도 있다. 비일시적 기록매체는, 예를 들어, 컴퓨터 시스템에 의하여 판독가능한 형태로 데이터가 저장되는 모든 종류의 기록장치를 포함한다. 예를 들어, 비일시적 기록매체는 EPROM(erasable programmable read only memory), 플래시 드라이브, 광학 드라이브, 자기 하드 드라이브, 솔리드 스테이트 드라이브(SSD)와 같은 저장매체를 포함한다.

[0143] 이상의 설명은 본 실시예의 기술 사상을 예시적으로 설명한 것에 불과한 것으로서, 본 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자라면 본 실시예의 본질적인 특성에서 벗어나지 않는 범위에서 다양한 수정 및 변형이 가능할 것이다. 따라서, 본 실시예들은 본 실시예의 기술 사상을 한정하기 위한 것이 아니라 설명하기 위한 것이고, 이러한 실시예에 의하여 본 실시예의 기술 사상의 범위가 한정되는 것은 아니다. 본 실시예의 보호 범위는 아래의 청구범위에 의하여 해석되어야 하며, 그와 동등한 범위 내에 있는 모든 기술 사상은 본 실시예의 권리범위에 포함되는 것으로 해석되어야 할 것이다.

부호의 설명

[0144] 110: 3D INVR 부호화부

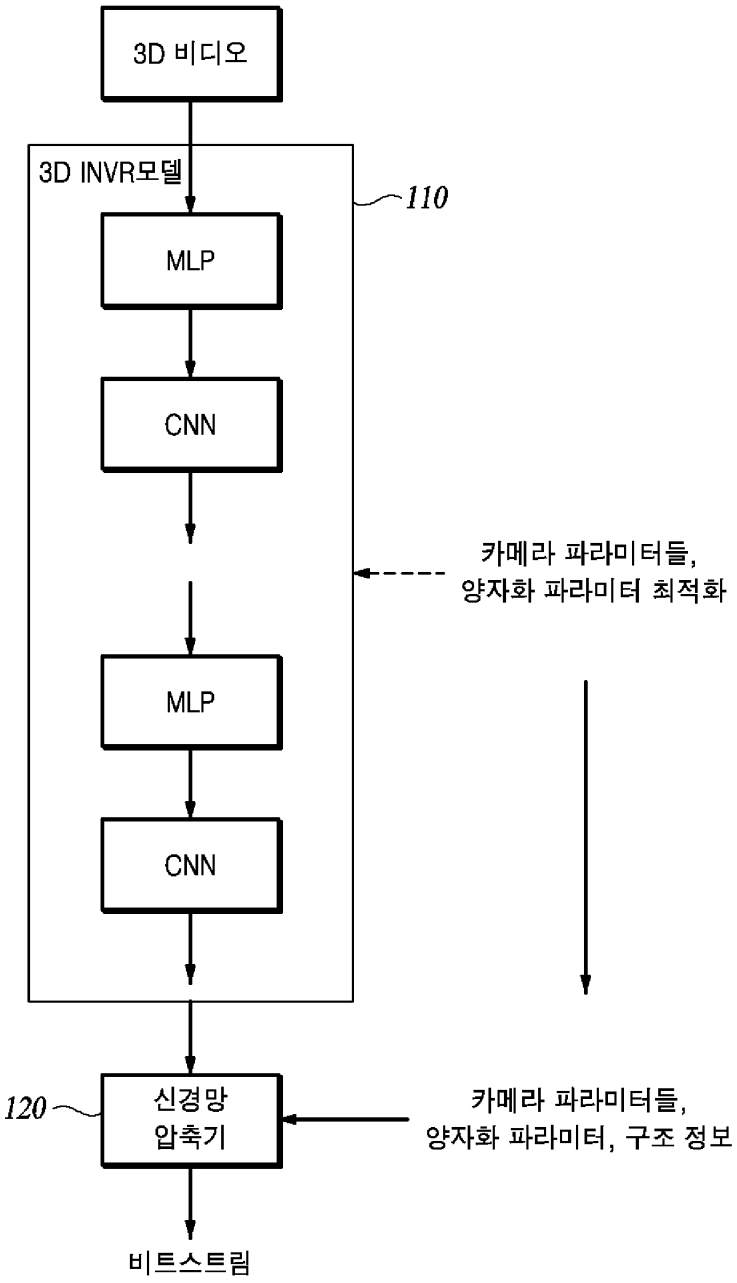
120: 신경망 압축기

210: 신경망 역압축기

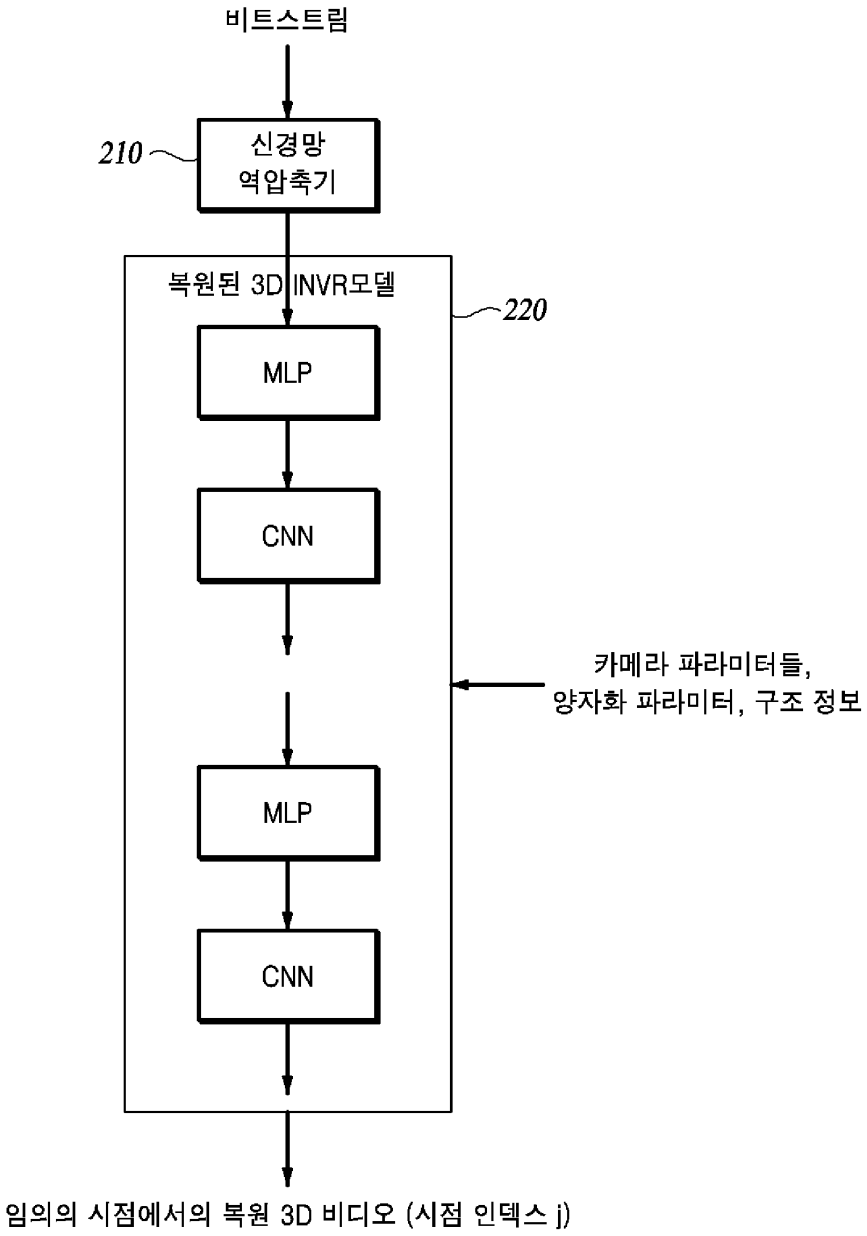
220: 3D INVR 복호화부

도면

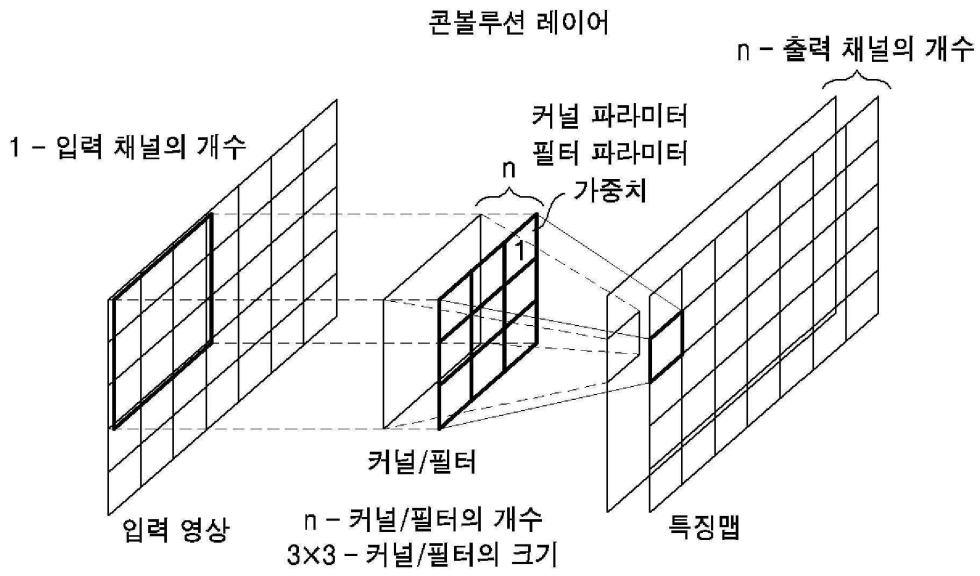
도면1


3D 비디오
3D INVR모델
110
MLP
CNN
카메라 파라미터들,
양자화 파라미터 최적화
MLP
CNN
120
신경망
압축기
카메라 파라미터들,
양자화 파라미터, 구조 정보
비트스트림

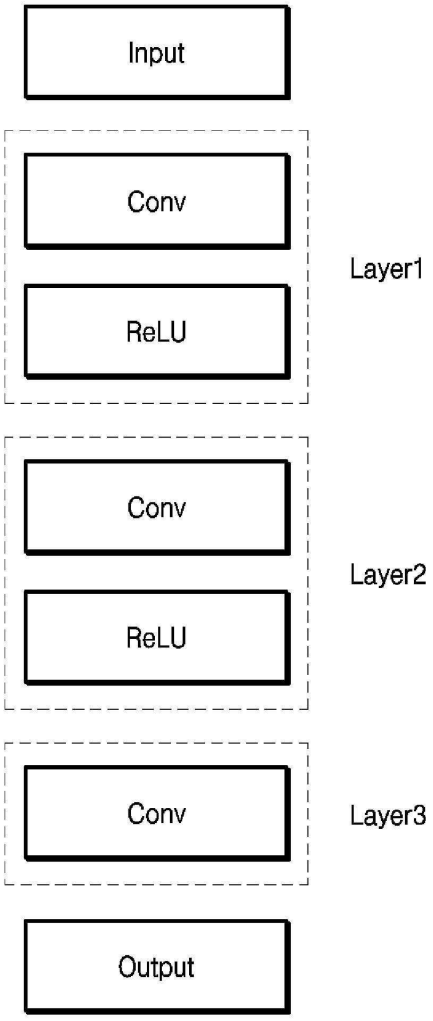
도면2


비트스트림
210
신경망
역압축기
복원된 3D INVR모델
220
MLP
CNN
카메라 파라미터들,
양자화 파라미터, 구조 정보
MLP
CNN
임의의 시점에서의 복원 3D 비디오 (시점 인덱스 j)

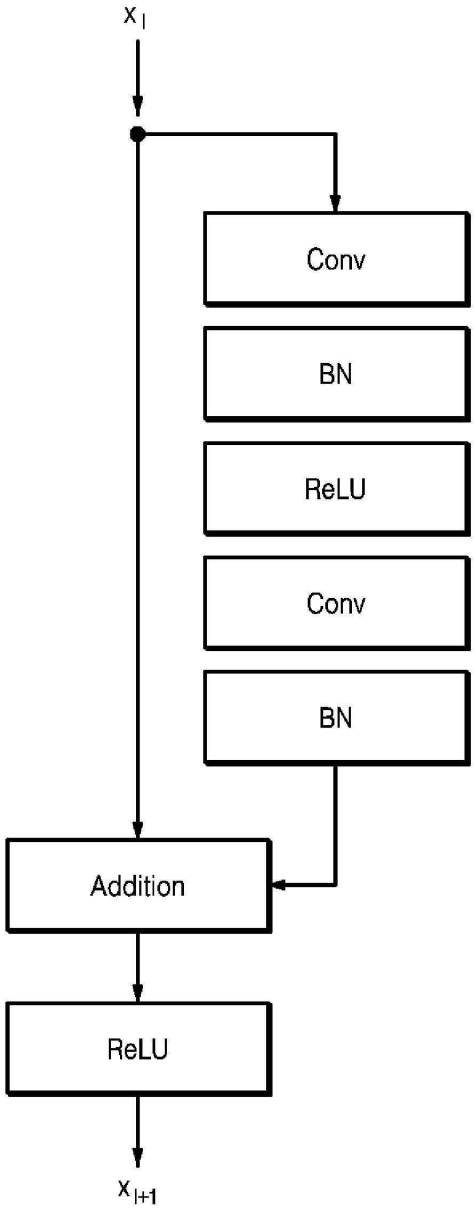
도면3


콘볼루션 레이어
n - 출력 채널의 개수
1 - 입력 채널의 개수
커널 파라미터
필터 파라미터
가중치
n
1
커널/필터
입력 영상
n - 커널/필터의 개수
3×3 - 커널/필터의 크기
특징맵

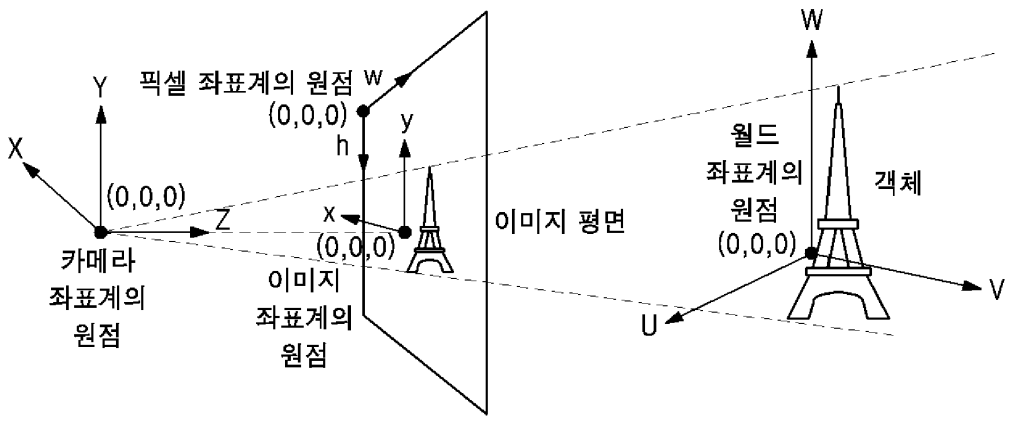
도면4


Input
Conv
ReLU
Layer1
Conv
ReLU
Layer2
Conv
Layer3
Output

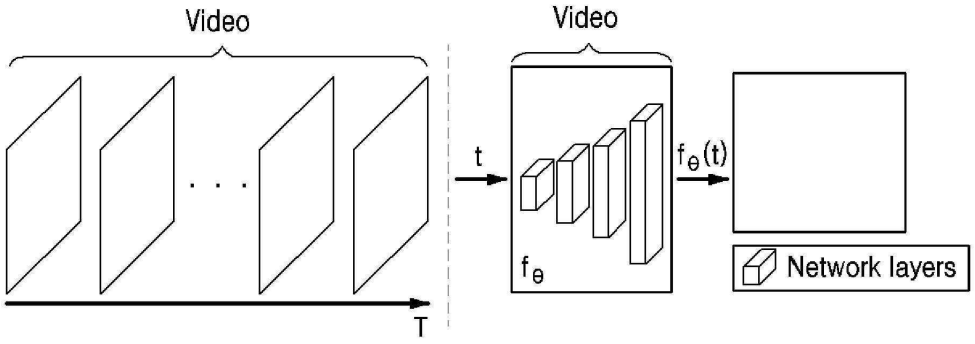
도면5



도면6

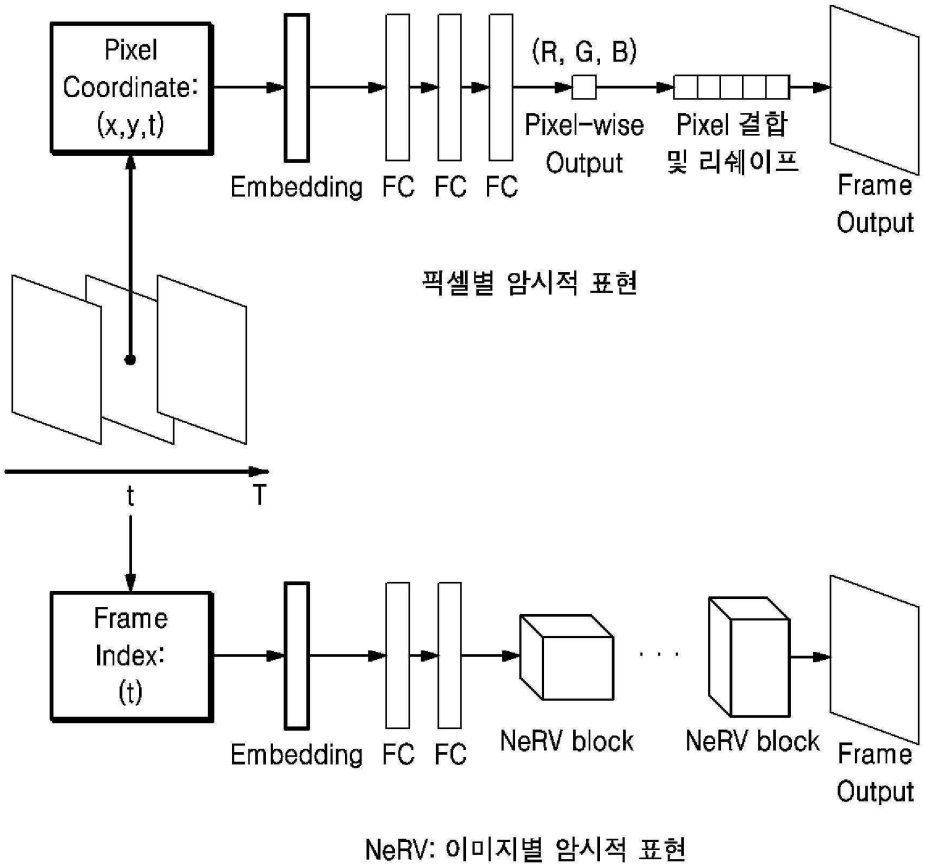


도면7

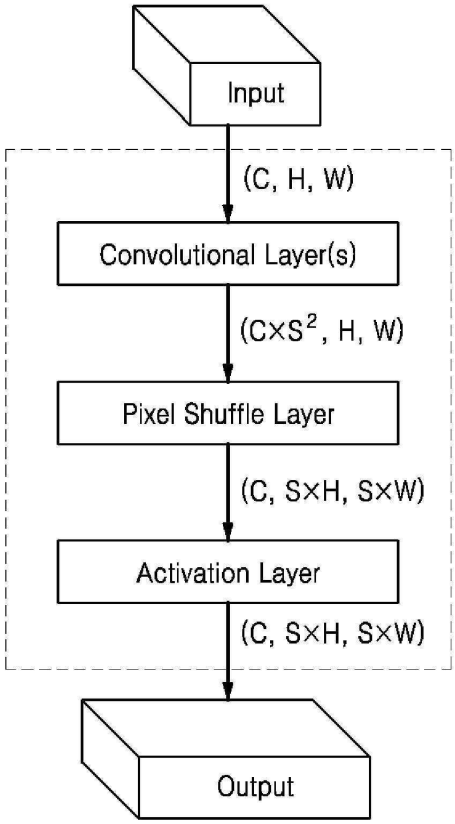


명시적 비디오 표현　　암시적 신경망 표현(e.g., NeRV)

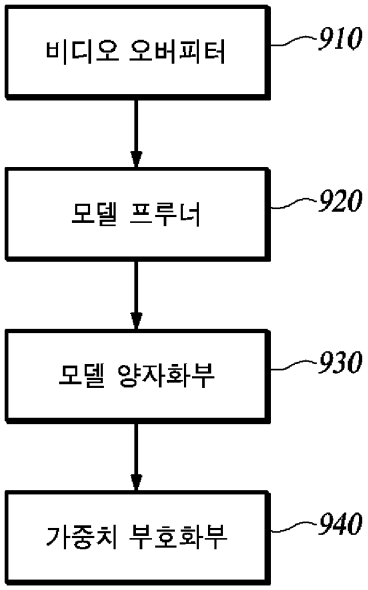
도면8a


Pixel
Coordinate:
(x,y,t)
Embedding
FC
FC
FC
(R, G, B)
Pixel-wise
Output
Pixel 결합
및 리쉐이프
Frame
Output
픽셀별 암시적 표현
t
T
Frame
Index:
(t)
Embedding
FC
FC
NeRV block
NeRV block
Frame
Output
NeRV: 이미지별 암시적 표현

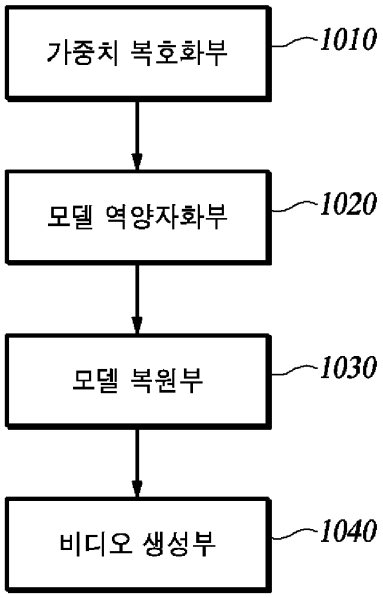
도면8b


Input
(C, H, W)
Convolutional Layer(s)
(C×S², H, W)
Pixel Shuffle Layer
(C, S×H, S×W)
Activation Layer
(C, S×H, S×W)
Output

도면9


비디오 오버피터
910
모델 프루너
920
모델 양자화부
930
가중치 부호화부
940

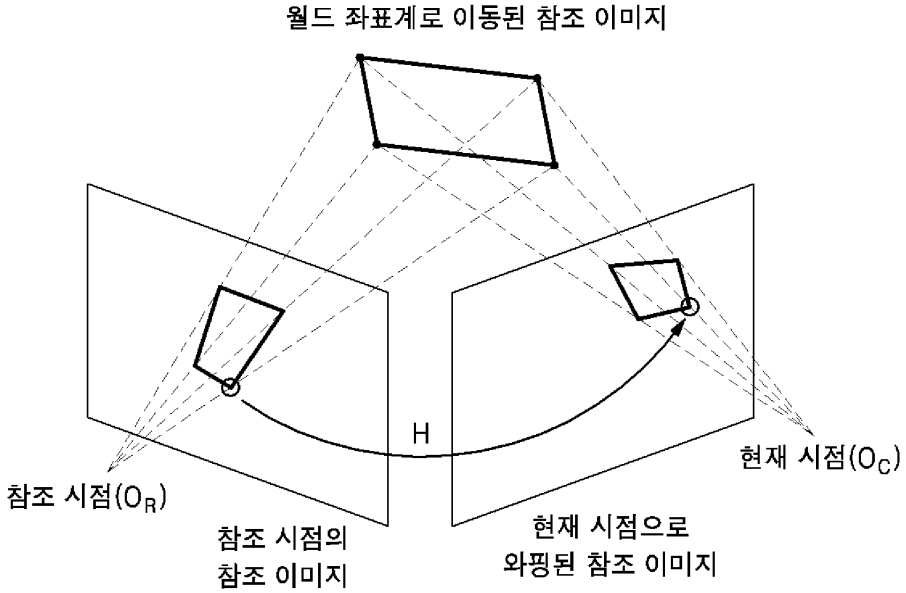
도면10



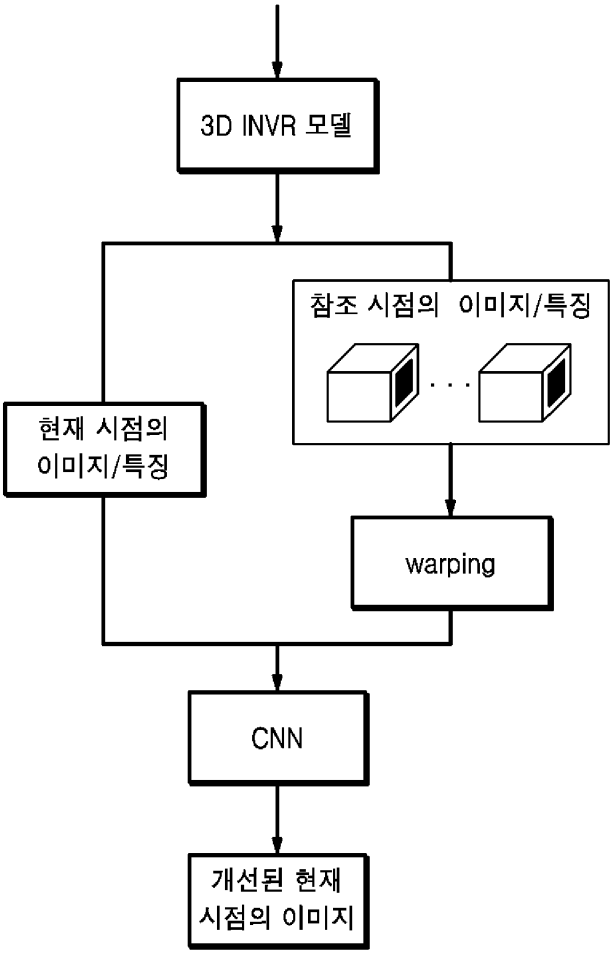
도면11

비트 마스크			×	가중치			=	프루닝된 가중치		
1	0	1		.7	.2	.1		.7	0	.1
1	0	1		−.2	.8	.9		−.2	0	.9
1	0	1		.2	.1	.3		.2	0	.3

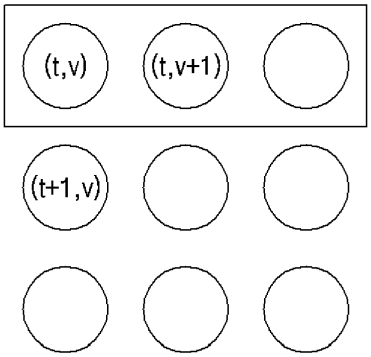
도면12



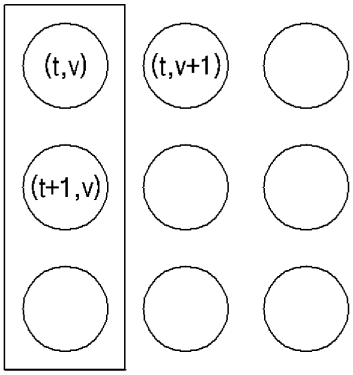
도면13



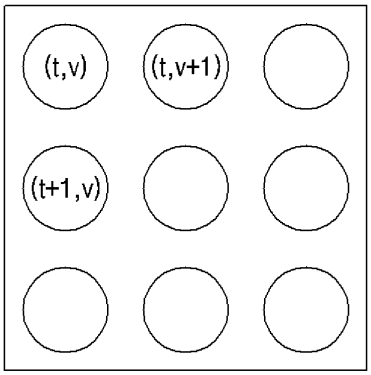
도면14



하나의 시간 t에서 복수의 시점들을 표현

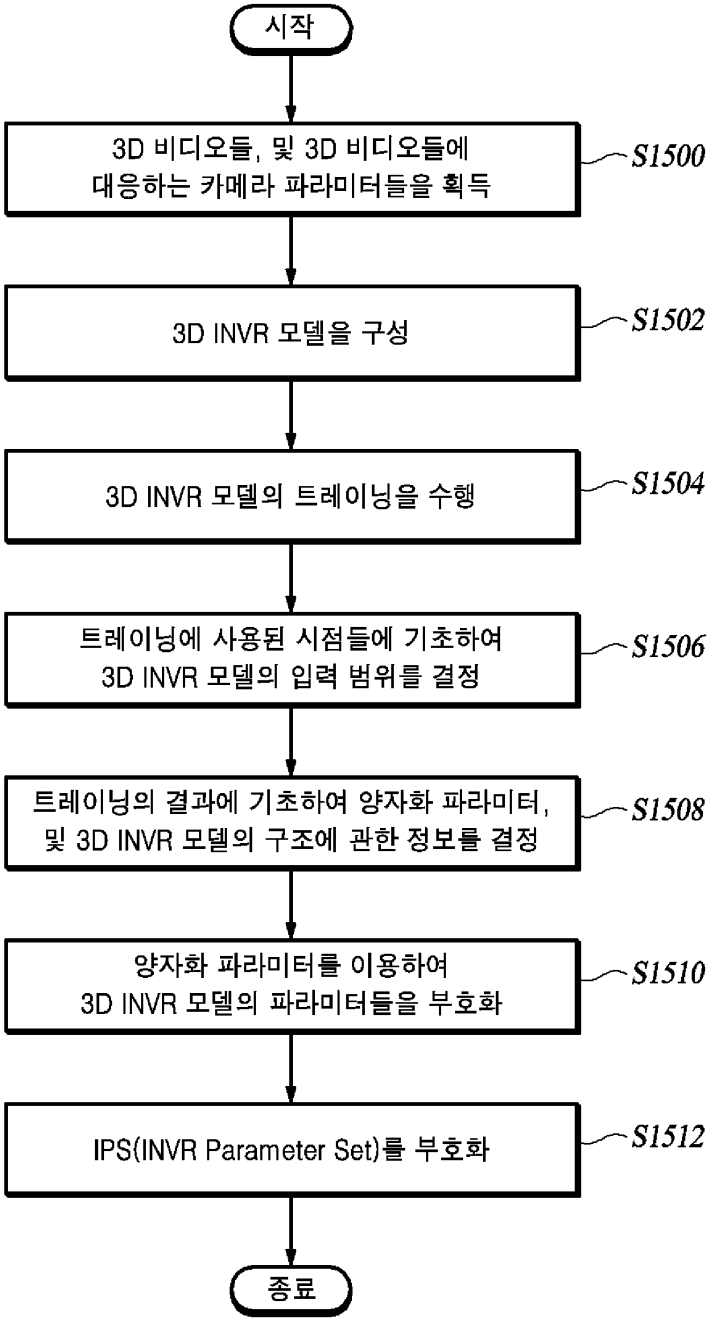


하나의 시점 v에서 복수의 시간들을 표현

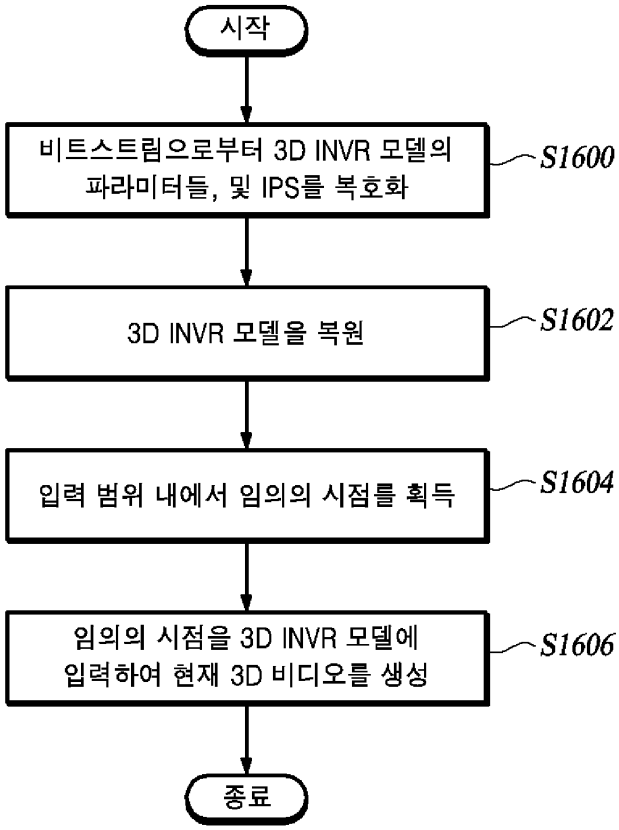


복수의 시간들에서 복수의 시점들을 표현

도면15


시작
3D 비디오들, 및 3D 비디오들에 대응하는 카메라 파라미터들을 획득
S1500
3D INVR 모델을 구성
S1502
3D INVR 모델의 트레이닝을 수행
S1504
트레이닝에 사용된 시점들에 기초하여 3D INVR 모델의 입력 범위를 결정
S1506
트레이닝의 결과에 기초하여 양자화 파라미터, 및 3D INVR 모델의 구조에 관한 정보를 결정
S1508
양자화 파라미터를 이용하여 3D INVR 모델의 파라미터들을 부호화
S1510
IPS(INVR Parameter Set)를 부호화
S1512
종료

도면16


시작
비트스트림으로부터 3D INVR 모델의 파라미터들, 및 IPS를 복호화
S1600
3D INVR 모델을 복원
S1602
입력 범위 내에서 임의의 시점를 획득
S1604
임의의 시점을 3D INVR 모델에 입력하여 현재 3D 비디오를 생성
S1606
종료