영상과 비디오로부터의 가상 시점 영상 생성 기술

□ 백형선, 박인규 / 인하대학교

요약

실감형 미디어를 구성하기 위해서는 다시점 영상 또는 비디오들로 구성된 대용량의 콘텐츠가 필수적이다. 이러한 콘텐츠는 다량의 카메라들을 목적에 따라 배치하여 획득하므로 영상 구성의 복잡성과 콘텐츠의 크기가 급격히 커진다는 문제점을 갖고 있다. 3D 미디어 환경에서 카메라의 개수를 최소화하면서도 목적에 맞게 다양한 시점을 제공할 수 있는 가상시점 영상 생성은 핵심적인 기술이다. 본 기고문에서는 다시점 영상과 비디오로부터 학습 기반의 가상 시점 영상 생성 연구들에 대해 체계적인 조사를 통해 그 결과를 다음과 같이 제시한다. 첫째, 가상 시점 영상 생성에 대한 배경 개념을 정의한다. 둘째, 제안하는 분류 방식에 따라 기존의 제안된 방법들을 상세하게 분석한다. 셋째, 가상 시점 영상 생성에 주로 사용되는 관련 데이터셋을 조사한다. 마지막으로는 각 연구들이 갖고 있는 특징들을 분석하고, 정량적, 정성적 평가 결과를 비교한다.

Ⅰ. 서 론

입력 영상으로부터 가상 시점 생성은 VR/AR, 홀로그램 등의 콘텐츠와 함께 컴퓨터 그래픽스, 3D 디스플레이, 게임 산업과 같은 실감 미디어에 다양한 적용 및 활용이 가능함에 따라 관련 연구들[5, 7, 10, 11, 14, 15, 16, 17, 18, 20, 21, 22, 24, 26, 29, 30, 33]이 많은 관심을 받고 있다. 최근 딥러닝을 활용한 가상 시점 영상 생성 기술들은 고전적인 기술에 비해 많은 성능 향상을 이루고 있다. 이를 크게 분류하면 다중뷰 스테레오 기반 최적화 접근법[7, 8, 26]과 학습 기반 접근법으로 분류할 수 있다. 또한 학습 기반의 접근법은 입력으로 들어오는 형태에 따라 세부적으로 단일 영상[16, 18, 24] 기반, 다중 시점 영상[5, 11, 6, 13, 30, 33] 기반 또는 비디오[8, 31]를 기반으로 분류한다.

※ 이 기고문은 삼성전자 미래기술육성센터의 지원을 받아 수행된 연구임(SRFC-IT1702-54). 이 기고문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(2020-0-01389, 인공지능융합연구 센터지원(인하대학교)).

학습 기반의 접근법은 기존의 장면 표현 방식과 학습 기반 네트워크와의 결합, 새로운 학습 기반의 모델을 제시, 학습 과정에 미분 가능한 형태의 렌더러를 포함하는 등의 다양한 연구들이 진행되고 있다. 본 기고문에서는 단일 또는 다시점 영상과 비디오를 입력으로 하는 가상 시점 영상 생성 기법들을 입력에 따라 기술들을 분류하고 제안된 논문들에서 사용된 모델링, 렌더링 기법들을 살펴본다. 또한, 정량적 및 정성적 결과를 포함한 연구들의 분석 결과를 제시한다.

Ⅱ. 문제 정의

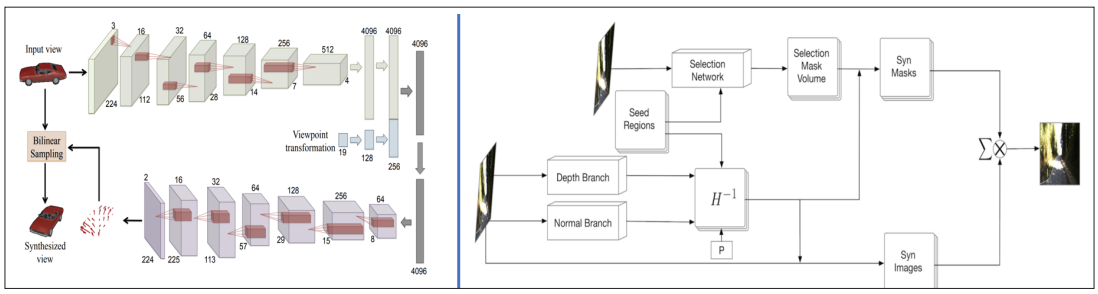
가상 시점 생성은 다음과 같이 정의한다. X_i를 임의 시점 i의 영상, P_i를 카메라 파라미터라 할 때, N개의 영상 $X_i = (X_1, \cdots, X_N)$과 카메라 파라미터 $P_i = (P_1, \cdots, P_N)$를 입력으로 하여 가상 시점 t의 영상 X_t를 생성한다. 일반적으로 가상 시점 생성은 IBR(image-based rendering)를 기반으로 입력 영상들의 화소들 간의 관계를 이용하여 생성하거나, IBR과 깊이 정보를 기반해 3D warping으로 생성한다. 본 기고문에서는 이와 같은 고전적인 방식과 결합한 학습 기반의 모델링 방식을 이용하여 하나 이상의 영상 또는 비디오를 입력으로 한 기술들에 관련된 연구들을 살펴본다.

Ⅲ. 영상으로부터의 가상 시점 영상 생성

학습 기반의 가상 시점 영상 생성은 주어진 하나 이상의 영상으로부터 3D 구조를 추정하여 학습된 장면 표현 방식을 렌더링 함으로써 얻을 수 있다. 렌더링 과정은 장면을 어떤 구조로 표현했는지에 따라 변화한다. 특히 입력이 제한적인 단일 영상으로부터 학습하는 방식은 3.1절에서 자세히 설명한다. 다시점의 영상을 입력으로 학습하는 방식은 풍부한 기하 정보를 바탕으로 단안 영상에 비해 높은 품질의 영상을 생성할 수 있으며 이는 3.2절에서 다양한 연구들을 통하여 설명한다.

1. 단안 영상 기반 기법

〈그림 1〉은 appearance 또는 3D 기하 정보를 이용한 가상 시점 영상 생성모델 구조를 보여준다. Tatarchenko 등은 가상 시점 영상의 화소 값들을 단순히 입력 영상과 가상 시점 정보로부터 직접적으로 추정하는 방식을 제안하였다[15]. 입력 영상으로부터 추출된 특징들을 3D



<그림 1> Appearance Flow 가상 시점 합성[29](왼쪽)과 [16](오른쪽)의 모델 구조

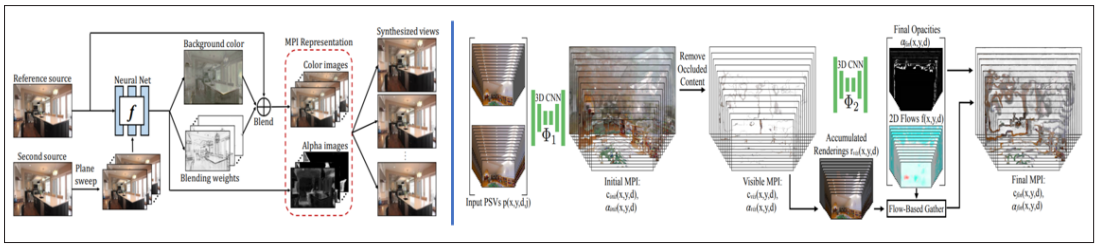
표현으로 간주하고, fully connected layer를 통해 인코딩된 시점 변화 정보와 결합하여 디코더 계층을 통과하여 결과 영상을 생성한다. 또한 Zhou 등은 단순히 네트워크를 통하여 화소 값을 생성하는 것이 아닌 입력으로 들어오는 영상의 화소 값을 활용하여 영상의 품질을 향상시키는 방법[29]을 제안하였다. 이는 시점은 다른 동일한 물체의 모습(텍스처, 모양, 컬러 등)들은 높은 상관관계를 갖고 있다는 관측을 바탕으로 인코더-디코더 구조를 활용하여 appearance flow를 추정한다. 추정된 flow 벡터와 입력 영상을 양선형 보간법을 통해 결과 영상을 생성한다.

Liu 등은 학습을 통해 호모그래피들을 추정하는 geometry-aware 네트워크를 제안하였다[16]. 제안된 방법은 두 개의 서브 네트워크로 구성되어 있다. 첫 번째 단계에서는 네트워크를 통해 입력 영상으로부터 깊이 맵과 법선 맵을 추정한다. 이후 예측한 두 개의 맵을 사용하여 입력 영상을 일정 개수의 super pixel로 분할한다. 앞서 예측한 깊이 정보, 법선 정보, super pixel과 함께 입력 영상과 가상 시점의 상대적인 자세 정보로부터 다수의 호모그래피를 추정하여 합성 영상 후보들을 생성한다. 두 번째 네트워크는 각 화소가 생성된 호모그래피들 중에서 하나를 통해 가상 시점을 생성하는지를 나타내는 선택 마스크를 추정한다. 최종적으로 선택 마스크에 따라 합성 영상들을 조합하여 결과 영상을 생성한다.

Wiles 등은 미분가능한 포인트 클라우드 렌더러를 통하여 가상 시점을 생성하는 SynSin을 제안하였다[18]. 해당 네트워크는 먼저 각각의 네트워크를 통하여 특징 맵과 깊이 맵을 추정하여 이를 포인트 클라우드로 생성한다. 생성된 포인트 클라우드는 입력으로 주어진 자세 변화를 통해 이동되며 제안된 미분가능한 포인트 클라우드 렌더러를 통해 해당 시점의 특징 맵을 생성한다. 마지막으로 결과 시점의 특징 맵으로부터 적대적 생성 네트워크를 기반으로 가상 시점 영상을 생성한다.

2. 다시점 영상 기반 기법

다시점 영상을 통한 가상 시점 생성 또한 장면 표현 방식을 정의하고, 이를 추정하여 가상 시점의 영상을 렌더링한다. 학습은 렌더링 생성 결과와 가상 시점의 참값 영상과의 손실 함수 계산을 통해 진행된다. 이러한 접근법은 단안 영상 기반의 접근법보다 많은 입력 영상들로 인해 풍부한 정보를 갖고 있어 뛰어난 품질의 영상을 생성할 수 있다는 점에서 많은 연구들이 진행되었다. 최근까지도 다수의 연구들[5, 10, 11, 17, 20, 21, 22, 30, 33, 6, 9, 13, 25, 28, 31]에서 다시점 영상 기반으로 가상 시점을 생성하는 좋은 성능을 보여주고 있다. 다시점 영상 기반의 접근법은 사용된 장면 표현 방식을 기준으



<그림 2> Stereo Magnification[30](왼쪽)과 [22](오른쪽)의 모델 구조

로 분류하였으며 본 기고문에서는 multi-plane images (MPI), neural implicit representation로 분류한다.

• **Multi-plane images**

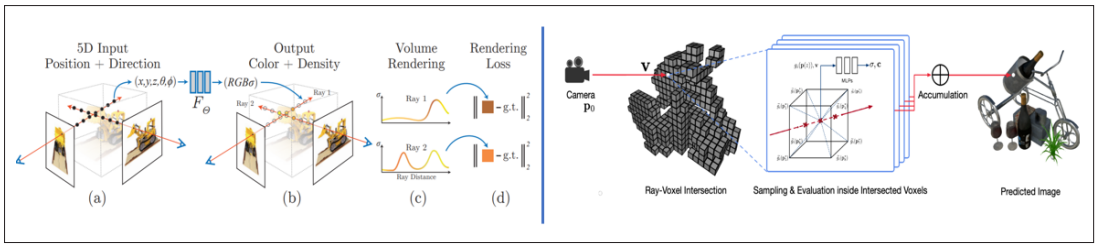
다중 평면 영상은 복수의 고정된 깊이 평면들을 기준으로 각 평면의 영상을 RGB와 alpha로 구성한 장면 표현 방식을 의미한다. [30]에서 Zhou 등은 딥러닝 프레임워크 상에 MPI를 활용한 가상 시점 생성을 최초로 제안하였다. 주요 아이디어는 다음과 같다. 먼저 입력은 좁은 베이스라인을 갖는 스테레오 영상이 사용되는데, 자세 정보를 따로 네트워크의 입력으로 주지 않고 스테레오 영상을 이용하여 계산한 plane sweep volume(PSV)을 통해 간접적으로 제공한다. 이와 같이 계산된 PSV와 입력 영상은 인코더-디코더 구조의 네트워크를 통해 컬러 영상과 alpha 영상을 얻을 수 있으며 호모그래피를 이용하여 가상 시점의 영상을 렌더링한다.

Srinivasan 등은 이론적인 분석을 통해 [30]에서 제안된 MPI의 가용한 렌더링 범위는 평면 영상의 개수에 대해 선형적으로 증가함을 보였다[22]. 이러한 분석을 바탕으로 3D CNN기반의 네트워크를 사용하여 무작위의 해상도로 훈련을 진행한다. 최소 32개에서 최대 128개의 평면의 개수가 선택되며, 제한된 GPU 메모리 크기를 고려하여 공간 해상도도 결정한다. 이와 같은 훈련법으로 추가적인 GPU 메모리의 사용 없이 고해상도와 평면의 개수를 확장한 성능을 갖도록 학습이 진행된다. 마지막으로 MPI 장면 표현 방식에서는 각 평면들의 중복되는 텍스처들이 많이 존재하여 렌더링 과정을 거친 영상에서 화소들이 반복되는 에러가 발생하였다. 이를 제거하기 위해 앞서 설명한 네트워크를 통해 추측한 초기 MPI에서 표면 요소들을 추출하여 중간 단계의 MPI를 생성한다. 이를 또다른 CNN 기반의 네트워크를 통하여 최종적인 alpha 맵과 flow 벡터들을 예측한다. 최종적인 컬러영상은 예측된 flow 벡터와 중간 단계의 MPI를 통해 계산한다.

기존의 MPI 기반 방식들은 스테레오 영상으로부터 내삽된(interpolated)시점의 영상을 생성하고 더욱 나아가 외삽된(extrapolated) 시점의 영상 생성을 가능하게 하였다[22, 30]. 그러나 앞선 방식들은 짧은 기준선을 갖는 스테레오 영상에 한정되어 있다. 한편 [5]에서 제안된 방법은 5개의 입력 영상으로부터 각각의 추정된 MPI들로부터 가중치에 따라 조합하여 넓은 시점 변화를 보여주는 결과 영상들을 생성하였다. 또한 [32]에서 제안된 기법은 MPI의 각 복셀을 RGBA 값 외에 잠재 특징 벡터를 도입하여 시간 변화에 따른 외양의 변화를 제어할 수 있게 확장하였다.

• **Neural Scene representation**

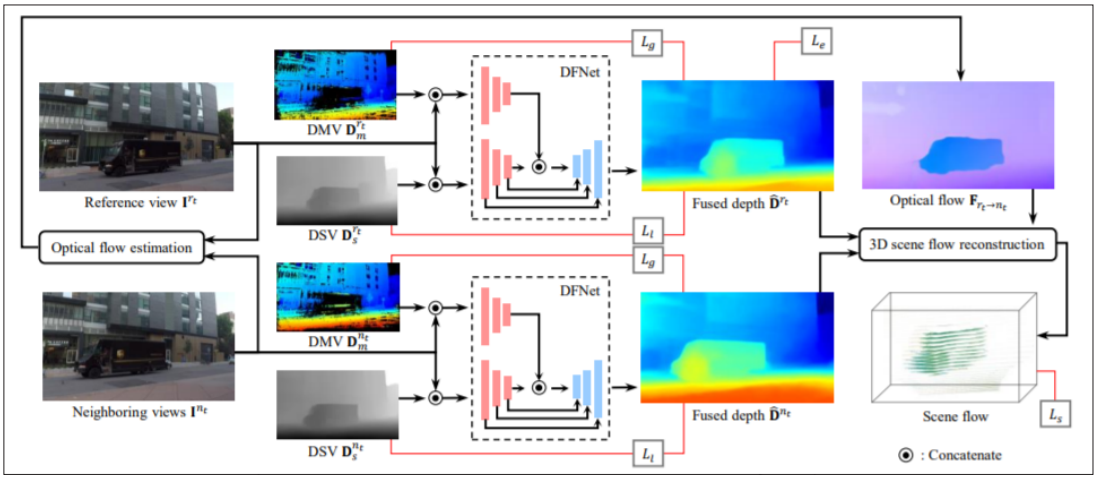
이 접근법은 2D 영상 혹은 3D 장면 정보를 복셀 그



<그림 3> NeRF[6](왼쪽)과 NSVF[13](오른쪽)의 모델 구조

리드, 포인트 클라우드, 메쉬와 같은 명시적인 기하 표현 방식이 아닌 신경망으로 파라미터화하여 장면을 표현하는 내재적인 표현을 사용한다. 이는 기존의 기법과 같이 3D 장면을 양자화하여 표현하는 것이 아닌 연속적인 값으로 나타내는 방식이라는 장점을 갖고 있다. 〈그림 3〉은 neural scene representation들의 파이프라인을 보여준다. Mildenhall 등에 의해 제안된 NeRF는 장면에 대한 neural radiance field를 최적화하는 방식으로 가상 시점 영상 생성의 인상적인 결과를 생성하였다[6]. NeRF는 다량의 입력 영상들의 카메라로부터 방사된 광선 상의 임의의 위치와 방향을 입력으로 받아 컬러 정보와 밀도를 출력하는 multi-layer perceptron(MLP) 네트워크를 통해 장면을 표현한다. 네트워크의 가중치는 다음과 같은 과정을 통해 최적화한다. 첫째, 입력 영상의 카메라로부터 방사되는 광선에서 계층적 볼륨 샘플링을 통해 샘플들을 얻는다. 계층적 볼륨 샘플링은 카메라 광선을 샘플링하는 과정에서 효율적인 샘플링을 위해 제안되었다. 단순히 많은 수의 샘플들을 생성하게 되면 렌더링에 영향을 미치지 않는 부분들이 다수 포함되어 비효율적이다. 이에 적은 수로 샘플링하여 계산된 결과를 활용하여 렌더링에 영향을 주는 샘플들을 추가적으로 선택한다. 둘째, 이와 같이 생성된 샘플들을 이용하여 볼륨 렌더링을 수행하고 생성된 컬러 정보를 참값을 나타내는 해당 화소와 비교함으로써 네트워크를 최적화한다.

[13]에 의해 제안된 neural sparse voxel fields (NSVF)는 단일 MLP 네트워크로 전체 공간을 모델링하는 것이 아닌 복셀 구조를 도입하여 NeRF[6]에 비해 10~20배 빠른 렌더링이 가능한 neural scene representation 방식을 제안하였다. NSVF는 카메라 광선과 교차하는 복셀을 검출함으로써 비효율적인 샘플들이 선택되는 것을 제한한다. 세부적으로는 광선이 어떠한 복셀을 통과하는지를 Octree와 축정렬 bounding box 테스트를 통해 빠르게 수행하며, 광선과 교차하는 복셀들은 작은 크기를 갖는 복셀들로 세분화되고 그렇지 않은 복셀들은 스스로 걸러진다. 세부적인 장면 표현이 가능하도록 복셀의 꼭지점들을 32차원의 임베딩을 통하여 표현한다. 이에 따라 임의의 복셀 내의 위치에



<그림 4> Novel View Synthesis of Dynamic Scenes[9]의 모델 구조

서 샘플링 될 경우에 해당 포인트의 위치 정보는 복셀들의 꼭지점들의 임베딩 값들로부터 삼선형 보간하여 취득한다. 임베딩된 위치 정보와 방향을 입력으로 컬러 정보와 밀도는 MLP를 이용하여 추정한다.

Martin-Brualla 등은 기존의 제한된 환경의 영상의 입력을 넘어 다량의 정제되지 않은 영상들을 입력으로 하여 3D장면을 복원하는 NeRF-W[23]를 제안하였다. 이는 기존의 방법들과 다르게 입력 영상들은 서로 다른 조명환경 또는 폐색을 발생시키는 일시적인 물체들이 존재하는 것을 의미한다. Appearance 임베딩 네트워크로 저차원 잠재 공간에서 노출, 조명, 날씨 등과 같은 영상별 모양 변화를 모델링한다. appearance embedding 벡터를 MLP의 입력으로 하여 정적인 부분의 RGBσ 를 추정한다. 일시적인 물체들 추정도 유사하게 임베딩 네트워크를 사용하여 구하였다. 일시적인 물체의 가림의 경우에는 항상 존재하는 경우가 아니기 때문에 추가적으로 uncertainty를 확률적으로 추정하여 최종 결과를 생성한다.

Yu 등은 적은 수의 입력 영상으로부터 가상 시점 영상을 생성하는 신경망으로 파라미터화된 함수 기반의 장면 표현 방식[1]을 제안하였다. Fully-convolutional 인코더를 통해 입력 영상의 특징들을 추출하고, 이를 MLP 네트워크에 사용하여 RGB 컬러와 밀도를 예측한다.

Ⅳ. 비디오로부터의 가상 시점 영상 생성

영상으로부터 가상 시점 생성과 비교할 때 비디오를 이용한 가상 시점 영상 합성은 시간에 따라 변하는 장면의 기하 정보 및 appearance에 대한 고려가 필요하므로 영상에서 사용된 일반적인 방법을 사용하면 좋지 않는 결과를 생성한다. 이에 기존의 제안되는 방법에서는 다중 카메라, 특수 하드웨어로 획득한 영상 또는 다중 시점의 동기화된 비디오를 이용하였다. 최근의 학습 기반의 2D영상 생성 모델들의 성능 향상으로 인하여 비디오로부터의 가상 시점 영상 생성 연구도 많은 관심을 받고 있다.

Yoon 등은 움직이는 물체가 존재하는 동적인 장면 환경 내에서 가상 시점을 생성하는 방법을 제안하였다[9]. 제안된 방법은 비디오로부터 인접한 프레임들로 구성된 다중 시점 영상임을 이용하여 정적인 장면에 대한 구조를 복원할 수 있다는 점과 각 프레임에서 취득한 깊이 맵으로부터 움직이는 물체에 대한 정보를 취득 가능하다는 점을 활용하였다. 이와 같이 추정된 정적인 배경과 동적인 물체를 결합함으로써 단안 카메라로부터 취득한 프레임들로부터 가상 시점의 영상을 생성한다.

Xian 등은 Neural Scene Representation을 기반으로 비디오로부터 가상 시점 생성이 가능한 Space-time Neural Irradiance Fields를 제안하였다[32]. 제안된 모델은 비디오로부터 각 프레임들의 깊이 맵을 추정하고, 이를 이용하여 신경망으로 파라미터화된 함수를 구성한다. 깊이 맵은 시간에 따라 변화하는 기하 정보와 appearance간의 모호성을 해결하는데 목적이 있다. 최적화된 신경망은 공간 도메인의 위치 x와 시간 t를 입력으로 받아 RGB 컬러와 밀도를 추정한다. 이에 따라 제안된 네트워크는 기존의 방식들과 달리 시점에 대한 정보를 사용하지 않으므로 시점 의존적 효과는 모델링 하지 못한다.

Ⅴ. 손실 함수

가상 시점 생성 모델은 일반적으로 L1 손실 함수를 복원 오차로 사용한다. 하지만 L1 손실 함수로 학습한 모

델의 렌더링 결과는 전체적으로 블러하게 생성되는 경향이 있다. 이는 가상 시점 영상 합성시 폐색 영역 혹은 입력 영상에서는 관측되지 않는 부분의 화소를 정확히 추정하지 못하고 주변 화소로부터 혼합되어 생성되기 때문이다. 이와 같은 문제를 해결하기 위해 특징 유사성을 활용한 VGG 손실 함수를 사용하면 L1으로 학습된 결과보다 인지적으로(perceptually) 우수한 결과 영상을 생성한다. VGG 손실 함수는 미리 학습된 VGG 네트워크를 이용하여 입력과 결과 영상의 특징들을 추출하고 이를 유클리디안 거리로 계산한다. 또한 생성 모델을 기반으로 가상 시점 생성 기법에 적대적 손실 함수를 사용하는 경우도 존재한다.

VI. 데이터셋

대부분의 가상 시점 생성 모델은 지도 학습(supervised learning)을 통해 학습되며, 데이터셋은 영상 또는 비디오로 구성되어 있다. 데이터를 가상 환경에서 취득한 경우 카메라 파라미터 참값이 존재하며 실제 환경에서의 경우에는 SfM(Structure from Motion), SLAM기반의 방법을 이용하여 생성한 카메라 파라미터를 사용한다. 본 기고문에서는 취득한 환경에 따라 데이터셋을 분류하였다.

일반적으로 실제 환경에서 카메라의 회전 및 이동을 다양하게 구성하여 대용량의 데이터셋 구축은 어려가지 어려움이 따른다. RealEstate10k는 동영상 스트리밍 사이트에 업로드된 비디오로부터 획득한 대용량의 데이터셋이다[30]. 약 10,000개의 비디오를 이용하여 80,000여 개의 짧은 클립으로 분할하고 이로부터 10,000,000개의 영상 데이터를 생성하였다. 또한 각 프레임의 자세 정보에 대해서는 SLAM 및 번들 조정 알고리즘을 사용하여 취득하였다. 카메라의 모션을 고려하여 비디오를 선정하였으며, 대부분이 건축물의 실내, 실외의 부드러운 모션 이동 비디오로 구성되어 있다.

Spaces 데이터셋은 16개를 카메라로 구성한 카메라 배열을 이용하여 취득한 영상들이다[11]. 각 카메라 간의 거리는 10cm이며, 가상 시점 생성시 입력과 결과 시점 간의 거리를 다양하게 결정할 수 있도록 설계되었다. 총 100개의 실내, 실외 장면을 촬영하였으며, 각 장

<표 1> RealEstate10k[30], NeRF Dataset[6], LLFF Dataset[5]에서의 가상 시점 생성 모델들의 정량적 평가 결과

알고리즘	RealEstate10k			Synthetic- NeRF			LLFF Dataset		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Stereo mag[30]	25.34	0.82	-	-	-	-	-	-	-
SynSin[18]	22.31	0.74	-	-	-	-	-	-	-
Single-MPI[24]	**26.4**	**0.859**	**0.103**	-	-	-	-	-	-
LLFF[5]	-	-	-	24.88	0.911	0.114	24.13	0,798	**0.212**
SRN[31]	-	-	-	22.26	0.846	0.170	22.84	0.668	0.378
NV[25]	-	-	-	26.05	0.893	0.160	-	-	-
NeRF[6]	-	-	-	31.01	0.947	0.081	**26.50**	**0.811**	0.250
NSVF[13]	-	-	-	**31.74**	**0.953**	**0.047**	-	-	-

면마다 5~10개의 영상들로 구성된다. LLFF데이터셋[5] 중 실제 환경에서 얻은 데이터들은 fine-tuning 목적으로 핸드폰 카메라를 이용하여 24개의 장면을 취득하였다. SfM을 이용하여 자세 정보를 추정하였다. Shiny 데이터셋[28]은 기존의 존재하는 데이터셋들과는 달리 복잡한 시점 의존성을 갖는 데이터로 만들었다. 빛의 반사 및 굴절 등이 포함된 8개의 장면에 대해 30~300개의 영상들로 구성되어 있다.

가상환경 기반의 데이터셋[2, 3, 6, 13]은 실제환경에서의 데이터셋과 달리 일반적으로 취득의 용이함에 의해 대규모로 구성이 가능하다. 자세 정보와 더불어 깊이맵, semantic 정보도 포함된 경우가 있다. 이는 3D 재구성과 같은 목적에 의한 데이터셋의 경우 다양한 주석 정보들이 포함되어 있다. 일반적으로 가상 시점 생성 모델의 경우에는 RGB 영상과 자세 정보만 사용된다.

앞서 설명한 바와 같이 실제환경 및 가상환경에서의 데이터 취득을 통해 가상 시점 영상 생성에 대한 연구들이 수행되고 있지만, 각 연구들의 목적에 맞게 데이터를 취득 및 가공하는 경우가 많다는 것을 확인할 수 있었다. 이는 아직 일반화된 가상 시점 생성 연구보다는 제한된 환경에서의 연구들이 수행되고 있음을 알 수 있다.

Ⅶ. 성능 평가

가상 시점 영상 생성을 평가하기 위해서는 하나 이상의 영상 또는 비디오로부터 추정한 결과 시점의 합성 영상과 실제 참값을 의미하는 영상과의 비교를 통하여 평가한다. 본 기고문에서는 영상 합성 평가에 주로 사용되는 세 가지 지표에 대해 살펴본다.

첫 번째로는 대표적인 영상 품질 평가지표인 PSNR (peak signal-to-noise ratio)이다. PSNR은 합성된

<그림 5> SynSin[18], LLFF[5], NeRF[6], NSVF[13]의 정성적 평가 결과

영상의 손실 정보를 평가하여 정량적으로 나타낸다. PSNR은 MSE를 기반으로 표현되며, MSE는 영상 간의 차이를 측정하는 수단으로 두 영상의 화소 값들의 차이로 표현된다. 손실되는 정보가 적을수록 MSE의 값이 작아지므로 PSNR의 값은 커지게 된다. 즉, 합성 영상이 참값에 가까울수록 PSNR 또한 큰 값을 갖는다.

다음으로는 SSIM(structural similarity index map)이다. SSIM은 두 영상 간의 차이를 구하는 방식이 아닌 사람의 시각 시스템이 영상의 구조 정보에 민감하게 반응한다는 것을 이용한 지표이다. 세부적으로는 SSIM은 두 영상의 평균 휘도, 표준편차를 통해 영상의 구조 정보를 표현하고 이를 비교함으로써 구조적 유사도 지수를 나타낸다.

마지막으로는 LPIPS(learned perceptual image patch similarity)이다. LPIPS는 앞서 설명한 PSNR, SSIM과 같은 지표는 사람의 지각적 과정에 비해 단순하고 지각적 유사성은 파악할 수 없다는 한계를 극복하기 위해 제안된 지표이다. Perceptual Loss와 유사하게 네트워크로 임베딩한 결과 간의 거리를 계산함으로써 지각적 유사성을 판단한다.

〈그림 5〉는 기고문에서 소개한 가상 시점 생성 기법들의 정성적 결과를 나타낸다. 세부적으로 살펴보면 SynSin의 경우에는 단안 2D 영상을 입력으로 하여 시점을 생성한다. 한 장의 영상만으로도 가상 시점을 생성할 수 있다는 장점이 있지만 제한된 입력으로 인하여 장면의 3D 구조를 추론하기가 어렵기 때문에 〈표 1〉에서도 단안 기반의 모델들의 성능이 낮음을 확인할 수 있다. Neural scene representation 기반의 NSVF, NeRF 기법들은 정량적, 정성적으로도 뛰어난 성능을 보여주고 있다. 그러나 이 방법은 새로운 장면에 대한 일반화 능력이 매우 적다는 점과 많은 수의 입력 영상이 필요하다는 단점이 있다. 마지막으로 MPI 기반의 방법들은 상대적으로 높은 일반화 능력과 준수한 성능을 보여준다. 하지만 다중 평면의 사용으로 인하여 메모리 사용량이 높고, 결과 영상의 텍스처들이 중복되어 나타나는 현상들이 발생한다.

Ⅷ. 결 론

본 기고문에서는 하나 이상의 영상 또는 비디오로부터 가상 시점 영상을 합성하는 방법들에 대해 다중뷰 스테레오 기반, 학습 기반 접근법으로 분류하여 기존의 기법들을 소개하고 특징을 분석하였다. 또한 학습 기반의 접근법은 단일 영상 기반, 다중 시점 영상 기반, 비디오 기반 기법으로 분류하여 비교하였다. 또한 성능 비교를 위해 일반적으로 통용되는 데이터셋, 손실 함수, 평가 지표를 소개하였다. 이를 기반으로 제안된 기법들의 정량적, 정성적 평가를 통해 각 접근법들에 대한 장단점을 분석하였다.

참 고 문 헌

[1] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa, pixelNeRF: Neural Radiance Fields from One or Few Images, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021), pp. 4578-4587.

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner, ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017), pp. 5828-5839.

[3] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu, Shapenet: An Information-Rich 3D Model Repository, arXiv preprint arXiv:1512.03012(2015).

[4] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun, Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction, ACM Trans. on Graphics (2017), Vol. 36, No. 4, pp. 1-13.

[5] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar, Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines, ACM Trans. on Graphics (2019), Vol. 38, No. 4, pp. 1-14.

[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, Proc. European Conference on Computer Vision (2020), pp. 405-421.

[7] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality Video View Interpolation Using a Layered Representation. ACM Trans. on Graphics (2004), Vol. 23, No. 3, pp. 600-608.

[8] Gaurav Chaurasia, Sylvain Duchene, Olga Sorkine-Hornung, and George Drettakis, Depth Synthesis and Local Warps for Plausible Image-based Navigation. ACM Trans. on Graphics (2013), Vol. 32, No. 3, pp. 1-12.

[9] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz, Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020), pp. 5336-5345.

[10] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely, DeepStereo: Learning to Predict New Views from the World' s Imagery, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2016), pp. 5515-5524.

[11] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker, DeepView: View Synthesis with Learned Gradient Descent, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 2367-2376.

[12] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski, Layered Depth Images, ACM Trans. on Graphics (1998), pp. 231-242.

[13] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt, Neural Sparse Voxel Fields, Proc. Advances in Neural Information Processing Systems (2020).

[14] Marc Levoy, and Pat Hanrahan, Light Field Rendering, ACM Trans. on Graphics (1996), pp. 31-42.

[15] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox, Multi-view 3D Models from Single Images with a Convolutional Network, Proc. European Conference on Computer Vision (2016), pp. 322-337.

[16] Miaomiao Liu, Xuming He, and Mathieu Salzmann, Geometry-Aware Deep Network for Single-Image Novel View Synthesis, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018), pp. 4616-4624.

[17] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based View Synthesis for Light Field Cameras, ACM Trans. on Graphics (2016), Vol. 35, No. 6, pp. 1-10.

[18] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson, SynSin: End-to-end View Synthesis from a Single Image, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020), pp. 7467-7477.

[19] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik, Modeling and Rendering Architecture from Photographs: a Hybrid Geometry- and Image-based Approach, ACM Trans. on Graphics (1996), pp. 11-20.

[20] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow, Deep Blending for Free-Viewpoint Image-based Rendering. ACM Trans. on Graphics (2018), pp. 1-15.
[21] Pratul P. Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng, Learning to Synthesize a 4D RGBD Light Field from a Single Image, Proc. IEEE/CVF International Conference on Computer Vision (2017), pp. 2243-2251.
[22] Pratul P. Srinivasan, Richard Tucker, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely, Pushing the Boundaries of View Extrapolation with Multiplane Images, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 175-184.
[23] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth, Nerf in the wild: Neural Radiance Fields for Unconstrained Photo Collections, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021), pp. 7210-7219.
[24] Richard Tucker, Noah Snavely, Single-View View Synthesis with Multiplane Images, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020), pp. 551-560.
[25] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh, Neural Volumes: Learning Dynamic Renderable Volumes from Images, ACM Trans. on Graphics (2019), Vol. 38, No. 4, pp. 1-14.
[26] Shenchang Eric Chen, and Lance Williams, View Interpolation for Image Synthesis, ACM Trans. on Graphics (1993), pp. 279-288.
[27] Steven M. Seitz, and Charles R. Dyer, View Morphing, ACM Trans. on Graphics (1996), pp. 21-30.
[28] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn, NeX: Real-Time View Synthesis with Neural Basis Expansion, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021), pp. 8534-8543.
[29] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A. Efros, View Synthesis by Appearance Flow, Proc. European Conference on Computer Vision (2016), pp. 286-301.
[30] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely, Stereo Magnification: Learning View Synthesis Using Multiplane Images, ACM Trans. on Graphics (2018), Vol. 37, No. 4, pp. 1-12.
[31] Vincent Sitzmann, Michael Zollhoefer, Gordon Wetzstein, Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations, Proc. Advances in Neural Information Processing Systems (2019).
[32] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim, Space-time Neural Irradiance Fields for Free-Viewpoint Video, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021), pp. 9421-9431.
[33] Zhengqi Li, Wenqi Xian, Abe Davis, Noah Snavely, Crowdsampling the Plenoptic Function, Proc. European Conference on Computer Vision (2020), pp. 178-196.

필자소개

백 형 선

- 2020년 2월 : 인하대학교 공간정보공학과 학사
- 2020년 2월 ~ 현재 : 인하대학교 전기컴퓨터공학과 석사과정
- 주관심분야 : 컴퓨터비전(가상 시점 영상 합성), 딥러닝

박 인 규

- 1995년 2월 : 서울대학교 제어계측공학과 학사
- 1997년 2월 : 서울대학교 제어계측공학과 석사
- 2001년 8월 : 서울대학교 전기컴퓨터공학부 박사
- 2001년 9월 ~ 2004년 2월 : 삼성종합기술원 전문연구원
- 2007년 1월 ~ 2008년 2월 : Mitsubishi Electric Research Laboratories 방문연구원
- 2014년 9월 ~ 2015년 8월 : MIT Media Lab 방문부교수
- 2018년 7월 ~ 2019년 6월 : University of California, San Diego (UCSD) 방문학자
- 2004년 3월 ~ 현재 : 인하대학교 정보통신공학과 교수
- ORCID : https://orcid.org/0000-0003-4774-7841
- 주관심분야 : 컴퓨터비전 및 그래픽스, deep learning, GPGPU