인공지능을 이용한 3D 콘텐츠 기술 동향 및 향후 전망

Recent Trends and Prospects of 3D Content Using Artificial Intelligence Technology

이승욱 (S.W. Lee, tajinet@etri.re.kr) CG/Vision연구실 책임연구원
황본우 (B.W. Hwang, bwhwang@etri.re.kr) CG/Vision연구실 책임연구원
임성재 (S.J. Lim, sjlimg@etri.re.kr) CG/Vision연구실 책임연구원
윤승욱 (S.U. Yoon, suyoon@etri.re.kr) CG/Vision연구실 선임연구원
김태준 (T.J. Kim, taejoonkim@etri.re.kr) CG/Vision연구실 선임연구원
김기남 (K.N. Kim, rlskal@etri.re.kr) CG/Vision연구실 선임기술원
김대희 (D.H Kim, kdh60243@etri.re.kr) CG/Vision연구실 위촉연구원
박창준 (C.J. Park, chjpark@etri.re.kr) CG/Vision연구실 책임연구원/실장

ABSTRACT

Recent technological advances in three-dimensional (3D) sensing devices and machine learning such as deep leaning has enabled data-driven 3D applications. Research on artificial intelligence has developed for the past few years and 3D deep learning has been introduced. This is the result of the availability of high-quality big data, increases in computing power, and development of new algorithms; before the introduction of 3D deep leaning, the main targets for deep learning were one-dimensional (1D) audio files and two-dimensional (2D) images. The research field of deep leaning has extended from discriminative models such as classification/segmentation/reconstruction models to generative models such as those including style transfer and generation of non-existing data. Unlike 2D learning, it is not easy to acquire 3D learning data. Although low-cost 3D data acquisition sensors have become increasingly popular owing to advances in 3D vision technology, the generation/acquisition of 3D data is still very difficult. Even if 3D data can be acquired, post-processing remains a significant problem. Moreover, it is not easy to directly apply existing network models such as convolution networks owing to the various ways in which 3D data is represented. In this paper, we summarize technological trends in AI-based 3D content generation.

KEYWORDS 딥러닝, 3D 콘텐츠, 3D 딥러닝, 3D 딥러닝 표준

Ⅰ. 서론

최근의 핫 이슈인 딥러닝은 여러 가지 응용 분야에서 기존의 전통적인 비전 기반의 접근방법(Hand Crafted Feature)보다 월등한 성능을 보여준다[1]. 2014년과 2015년은 인공지능을 통해 컴퓨터가 인

* DOI: https://doi.org/10.22648/ETRI.2019.J.340402
* 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2018년도 문화기술연구개발지원사업의 연구결과로 수행되었음[R2018030391, 게임 및 애니메이션을 위한 인공 지능 기반의 3D 캐릭터 생성 기술 개발].

간보다 사물의 인식을 더 잘하게 된 원년으로 기억될 것이다. 일반적으로 인간의 인식 오류율을 5% 정도로 가정하면, 이 당시 컴퓨터에 의한 인식 오류가 4%대로 떨어지게 되었다(인공지능의 경우 인식 오류율 측정 시 최대 근접값 상위 5개를 제공하기에 정확한 비교는 아닐 수 있음). 이 당시에 사용된 네트워크는 현재 가장 널리 사용되는 합성곱(Convolution) 기반의 네트워크[2]이며, 영상 인식, 자연어 처리, 게임 등의 많은 분야에서 우수한 성과를 이룩하였다. 합성곱 네트워크는 정형화된 데이터(본 고에서는 유클리디언 데이터로 정의, II장 참고)에 최적화되어 있으며, 특히 합성곱의 특성상 2차원 영상에 상당히 우수한 성과를 보인다. 합성곱의 과정을 통해 영상의 전역적인 특성과 세세한 특성을 파악하는 것이 가능하기 때문이다. 예를 들어, 사람 인식의 경우 기존의 Hand Craft Feature 방법은 알고리즘이 정의한 몇 가지 특성(미간 사이의 거리, 턱선의 모양 등)만으로 사람 여부를 판단하지만, 합성곱 기반 인공지능에서는 입력 영상의 모든 특징(예를 들어, 모든 방향의 외곽선)을 기반으로, 학습 데이터의 경향성에 의지하여 사람 여부를 판단하기에 상대적으로 더 정확한 결과를 도출한다.

또 다른 측면에서 초기의 인공지능은 식별모델(Discriminative Model)에 집중하였다. 영상 분석 및 영상 인식 등이 주요 응용분야이다. 그림 1을 보면, 식별모델은 입력 X가 주어질 때 출력 Y가 무엇인지 알아내는 것이다. 이는 수학적인 모델로 조건부 확률 P(Y|X), 즉 X라는 입력이 주어질 때 Y가 나올 확률을 구하는 문제이다. 사진을 입력하여 이것이 개인지 고양이인지 등을 알아내는 문제이다. 주어진 입력에서 개일 확률과 고양이일 확률 등을 계산하여 가장 높은 확률을 가지는 결과를 출력한다. 이와는 다른 생성 모델(Generative Model)이라는 것이 있는데, 이는 2014년 굿펠로우(Ian J. Goodfellow)[3]에 의해 GAN(Generative Adversarial Network)이라는 모델이 제안되면서 활발히 연구되는 분야이다. 이 경우는 출력된 결과물 Y를 기반으로 입력 X를 유추하는 것이다. 인식 결과에 의해 출력된 "고양이"라는 값으로 고양이 그림을 생성하는 것이 좋은 예시이다. 이 경우는 그림 1의 시스템의 전체 구성에 대한 확률 모델이 필요하여, 조건부 확률이 아닌 결합확률 P(X, Y)를 구하는 문제로 바뀐다. 쉽게 유추할 수 있는 바와 같이 X에서 Y로 가는 것은 많은 입력에서 하나로 가는 "Many to One Mapping"이지만, 반대의 경우는 "One to Many Mapping"이다. 앞의 경우와 비교해 상대적으로 수학적인 해를 찾기가 더 어려운 문제이다.

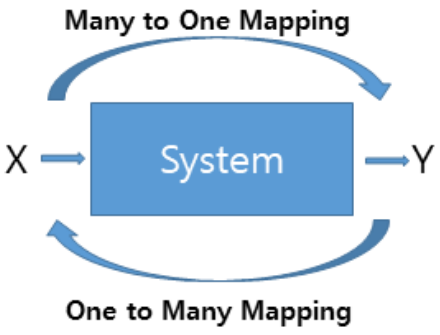

그림 1 식별모델과 생성모델의 차이

GitHub의 포스팅 그림을 보면(Cumulative number of named GAN papers by month, https://deephunt.in/the-gan-zoo-79597dc8c347) 최근 3년간 적대적 생성모델에 대한 연구 증가 추세를 알 수 있다. 2015년 월간 수 편에 달하던 논문이 2017년 기하급수적으로 증가하여, 2018년 월 350편 정도의 논문이 출간되고 있다.

3D 데이터 기반 학습은 기존의 인공지능 학습에 비해 몇 가지 차이점이 있다. 데이터 포맷의 형태의 차이와 학습 데이터 확보의 차이가 대표적인 예

이다. 본 장에서는 인공지능 기반 3D 콘텐츠 기술의 학습 데이터에 대해 논하고, II장에서는 데이터 포맷에 대한 논의를 진행한다.

1. 학습 데이터

음성 혹은 영상의 경우 쉽게 학습 데이터를 확보할 수 있다. 1차원 데이터인 음성은 녹음기 등을 이용하여 쉽게 확보할 수 있으며, 2차원 데이터인 영상의 경우도 카메라 등에 의해 쉽게 획득될 수 있다. 그러나 3차원 데이터의 경우는 최근 다양화된 저가 3D 센서에 의해 획득한다 하더라도, 후처리 작업이 필요하다. 예를 들어, 3D 스캐너로 확보된 데이터는 그림 2와 같은 후처리 작업이 필요하다.

본 고는 스캔 데이터를 정합하는 과정을 상세히 다루지는 않고 간단한 설명으로 대체한다. 그림 2의 (a)와 같이 다중 패치로 획득된 데이터를 비슷한 점의 위치를 찾아 이들 사이의 대응관계로 정합을 하고(b), 최종 정합이 완성된 결과물(c)에서 획득 과정에서 사라진 부분을 채우거나(홀 채움), 뒤집혀진 법선 벡터 방향을 정렬한다. 이런 과정을 거친 데이터는 특정한 요구사항이 있을 경우 외형의 리메쉬(remesh) 작업을 통해 토폴로지를 정규화하거나 해상도를 변경하는 등의 다양한 후처리가

그림 2 3D 스캔 데이터 후처리 과정

필요하다.

스카이 마인드[4]에서 정리한 위키를 참고하면 다양한 공개된 학습 데이터를 참고할 수 있다. 그러나 대부분의 학습 데이터는 2차원 영상위주이고, 3차원 데이터를 확보하는 것, 특히 후처리가 완료된 데이터는 확보가 어렵다. 그림 3은 데이터의 중요성에 대해 도시한다. 산업적 측면에서 보면 성장 동력은 하드웨어에서 소프트웨어로, 그리고 데이터로 이동되었다. 그러나 단지 데이터를 가지고 있다는 사실이 중요한 것이 아니라 정제된 데이터, 즉 학습에 유효하게 사용될 수 있는 데이터를 확보하는 것이 중요하다. 예를 들어, 3D 데이터의 경우 아무리 많은 데이터가 있다고 하여도 그림 2의 (a)와 같이 단일 패치만 있거나, 단일 패치의 캘리브레이션 정보 등이 없는 경우는 실제로 사용되기가 어렵다(최근 비지도 학습에 대한 관심 증가로 이러한 데이터에 대한 필요성도 어느 정도는 증가하고 있는 추세임). 다음은 대표적인 공개된 3D 학습 데이터 모델이다.

- ShapeNet: 가구, 자동차 등 태깅된(Annotated) 3D 모델을 제공. ShapeNetCor(55 카테고리, 51,300개 모델)와 ShapeNetStem(Shape-Net-Core의 부분 집합, 12,000개 모델)으로 구성 (https://www.shapenet.org/)
- Human3.6M: 남성 6인, 여성 5인의 배우에 의해 17가지 시나리오로 생성된 3.6백만 개의 자세 데이터 DB, 모션 정보 및 관련 태깅 포함(http://vision.imar.ro/human3.6m)
- ModelNet: 인체 포함하는 다양한 객체를 CAD 기반 3D 모델 제공(662 카테고리, 127,915개 모델) (http://modelnet.cs.princeton.edu/)
- CAESAR DB: 유럽과 미국인 5천 명에 대한 고품질 스캔 데이터(질감 및 모션 정보 없음) (http://store.sae.org/caesar/)

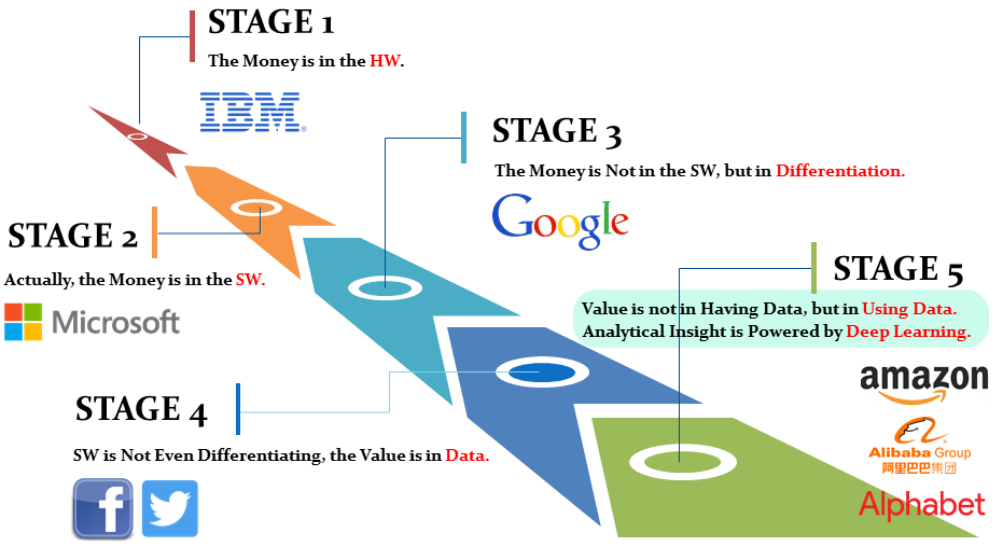

그림 3 산업에서의 데이터의 중요성

II. 인공지능 기반 3D 콘텐츠 기술

1. 3D 데이터 표현 방법

서론에서 잠시 논의한 바와 같이 2차원 인공지능 기술과 3차원 인공지능 기술의 차이점 중 하나는 데이터 포맷의 차이이다. 이 부분에 대해서는 [5,6]에 잘 설명되어 있다. 가장 근본적인 차이는 현재 합성곱 기반 학습 네트워크에 적합한 격자구조의 표현 여부이다. 그림 4는 전형적인 2차원 데이터 기반의 학습 방법이다.

다중 채널(RGB)로 구성된 입력영상을 이용하여 합성곱 기반 학습 수행 후 조건부 확률(그림 1 참고)값이 가장 높은 고양이를 출력하는 예제이다.

그림 6은 다양한 3D 그래픽 데이터의 표현 방식을 정의한다. 먼저 그림 5와 같은 차이에 의해 유클리디언 구조와 넌 유클리디언 구조로 나뉜다. 유클리디언 구조는 격자구조와 같이 정형화된 2차원 표현 방식으로 표현될 수 있고, 넌 유클리디언의 경우는 정형화되지 않은 표현에 의해 정의된다.

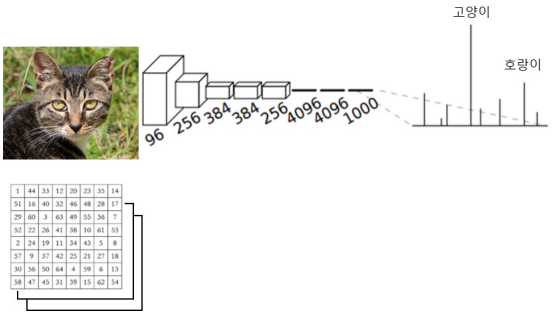

그림 4 유클리디언 데이터: 2D 영상의 학습 방법

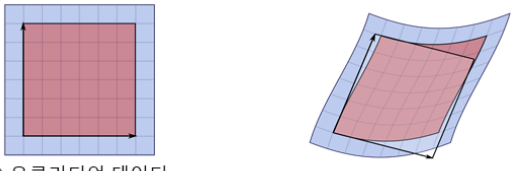

그림 5 유클리디언 vs 넌 유클리디언 데이터

프리미티브의 경우 정해진 구조(예를 들어, 컨트롤 포인트의 개수 고정 등)의 방식이면 유클리디언 표현으로 될 수 있고, 그렇지 않을 경우는 넌 유클리디언 표현으로 정의될 수 있다.

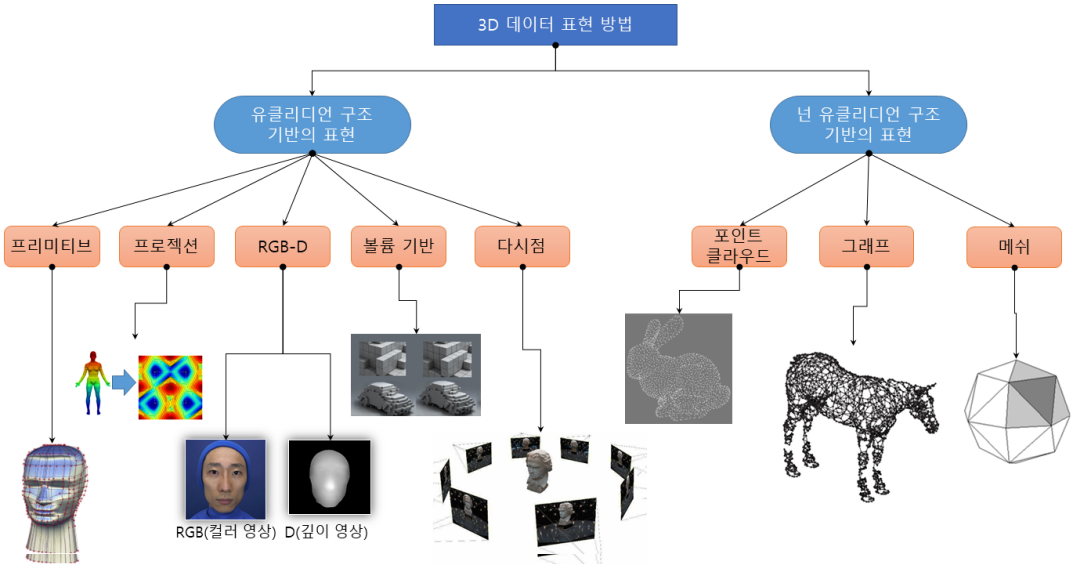

그림 6 다양한 3D 데이터 표현 방법

- 프리미티브: 수학적인 선, 면 등의 조합으로 데이터 생성, 각 프리미티브 데이터를 수학적으로 변형하는 컨트롤 포인트로 제어
- 프로젝션: 3D 데이터를 구 좌표계 혹은 실린더 좌표계로 맵핑하여 표현[7]
- RGB-D: 3D 데이터를 컬러영상과 깊이 정보로 분리하여 표현. 이 경우 360° 전체를 표현하기보다는 특정한 시점의 표현 가능
- 볼륨: 부피를 가지는 픽셀인 복셀을 정의하여, 복셀 값으로 3D 표현
- 다시점: 3D 데이터를 다시점으로 프로젝션하여 각 시점에서의 컬러 정보로 3D 표현
- 포인트 클라우드: 공간 위치 및 색상 정보를 포함하는 포인트 정보. 3D 스캐너의 일반적인 출력 형태로 사용되며 계측, 시각화 등에서 사용
- 그래프: 메쉬의 다른 형태의 표현으로, 그래프의 노드는 메쉬의 정점과 대응하고, 그래프의 에지는 메쉬의 연결정보에 대응하여 표현하는 방법
- 메쉬: 가장 일반적으로 산업계에서 사용되는 표현으로 3D 정점(Vertex), 다각형 면(삼각혹은 사각형의 면) 등으로 구성된 3D 데이터

2. 표현 방법에 따른 학습 방법

그림 7은 그림 6의 다양한 3D 데이터 표현 방식 중 유클리디언 구조의 표현에 따른 학습 방식을 간략화한 것이다. 각각의 표현 및 응용 분야에 따른 학습 방법은 참고문헌 [5,6]에 나타나 있다. 유클리디언 방식의 경우 전체적인 구조는 비슷하다. 합성곱 기반의 네트워크를 잘 사용할 수 있도록 3D 데이터를 가공하여 중간 단계 표현(예를 들어 프리미티브, 멀티뷰 등)으로 변형하는 것이 중요하다. 프리미티브 및 프로젝션 기반의 표현은 중간 단계의 표현이 2D 격자구조이기는 하지만, 일반적인 영상의 특징과는 다를 수 있기 때문에 합성곱 기반의 네트워크를 직접 사용하기보다는 기존의 네트워크를 다양한 방법으로 변형하여 사용한다. 이와는 다르게 RGB-D 혹은

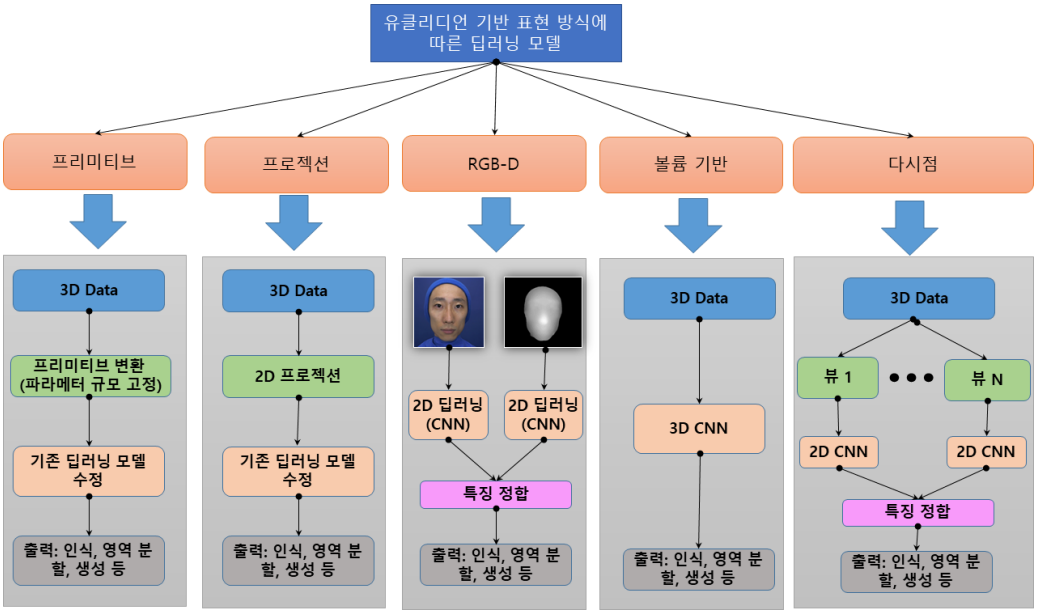

그림 7 다양한 3D 데이터 표현 방법

다시점 영상 등의 경우는 중간 단계의 표현이 영상과 같은 특성을 가지기 때문에 일반적으로 합성곱을 직접 이용하는 경우가 많다. 볼륨 기반의 데이터는 3차원 합성곱이 존재하므로 이를 직접 사용하기도 한다. 다시점 기반의 3D 딥러닝의 경우 가장 먼저 연구되기 시작한 방법으로 3D 데이터를 다양한 카메라 시점으로 랜더링하여 2차원 영상을 만들고, 각각에 대한 합성곱 방법을 수행한 후 각 뷰의 특징점을 정합하고(일반적으로 뷰 풀링을 이용[8]) 최종 결과물을 출력한다. 일반적으로 학습의 가속화를 위해 GPU를 사용한다.

다시점 기반의 방법은 다양한 응용에 적용될 수 있다. 예를 들어, 그림 8과 같은 학습 모델을 설계하여 2차원 영상을 입력하여 3차원 모델을 생성할 수 있다[9]. 3D 데이터를 16비트로 확장한 다시점 깊이 영상을 생성하여 지도학습 기반의 학습 데이터로 사용한다. 그림 8과 같이 RGB 영상과 대응하는 깊이 영상을 기반으로 도약 연결(skip connection)을 갖는 합성곱 기반 인코더-디코더 모델을 만들어 입력 영상에 대응하는 깊이 영상을 할 수 있도록 학습한다. 생성된 다시점 부분 깊이 영상 및 마스크 영상은 각 시점별 깊이 영상 생성을 돕는 역할을 수행하여 좀 더 향상된 결과를 생성할 수 있다. 입력 영상과 부분 깊이 영상 기반 깊이 영상 생성 모델은 인코딩 과정을 거쳐 나온 각 잠재 코드(latent code)와 시점별 변환 정보를 결합한 코드를 디코딩한다. 디코딩 단계에서 부분 깊이 영상과 예측된 깊이 영상 및 마스크 영역에 대한 L1 및 L2 loss 함수를 최소화하는 학습을 통해 최종 깊이 영상을 생성한다.

그림 9는 그림 8의 학습모델에 의해 생성된 결과를 보여준다. 사진 한 장을 입력하여 생성된 3D 모델로, 간단한 후처리를 통해 게임 및 애니메이션의 조연급 객체로 사용할 수 있다.

넌 유클리디언 기반의 3D 데이터 표현에 대한 학습 방법은 초기 연구가 진행 중이며, Eman[6]의 논문에 잘 요약되어 있다.

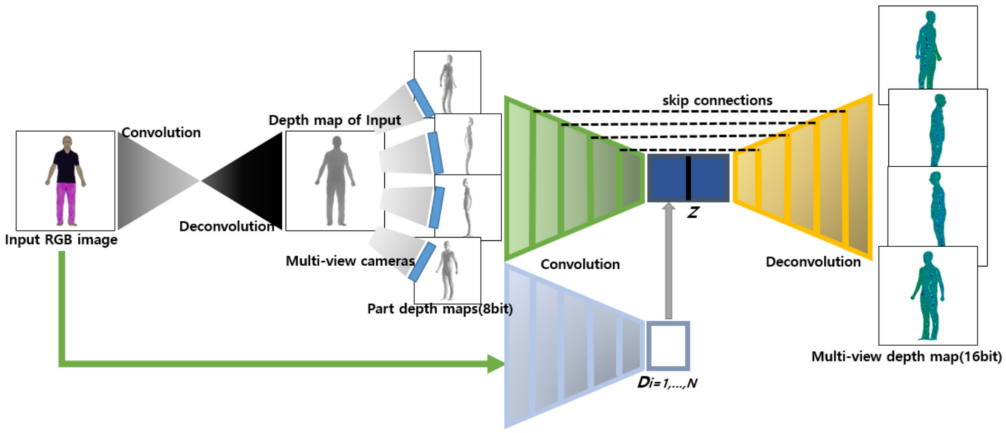

그림 8 다시점 기반 3D 데이터 표현방법에 의한 3D 데이터 생성 학습모델

3. 표준화 동향

이러한 기술 발전에 따라 관련된 표준화에 대한 논의가 진행 중이다[10]. 표준 분야는 크게 학습모델의 표현과 학습 데이터 표현에 대한 두 가지로 나눌 수 있다. 학습 방법의 경우 일종의 인코더 관련 기술임으로 표준의 대상은 아니지만, 참조모델 형태로 informative 표준이 개발되고는 있지만, 본고에서는 제외한다.

먼저 네트워크 포맷에 대한 표준이 있다. 가장 대표적인 것은 크로노스 그룹에서 정의하는 NNEF(Neural Network Exchange Format, https://www.khronos.org/nnef)이다. 현재 버전 1.0이 발표되었으며, 개발된 학습 모델이 다양한 프레임워크에서 동작할 수 있도록 다양한 오퍼레이션, 타입 정보 등을 메타데이터로 제공한다.

국내 학습 데이터에 대한 표준은 최근 TTA(Telecommunication Technology Association) PG 610에서 시작하였으며, "2D 이미지를 3D 모델로 생성하기 위한 3D 딥러닝 학습 파일 포맷"이라는 제목으로 표준화 작업을 진행 중이다(http://tta.or.kr/). 주요 내용은 3D 딥러닝 기술을 이용하여 2D 원화 영상을 3D 모델로 변경할 때 사용되는 학습 데이터의 파일 포맷을 XML 기반으로 정의하여 학습데이터의 재사용을 가능하게 한다.

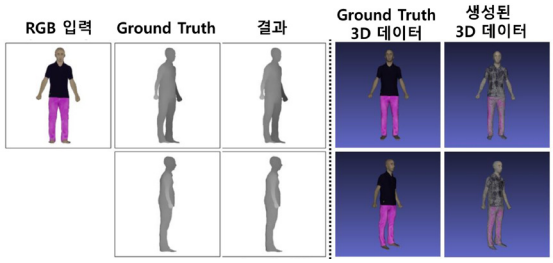

그림 9 한 장의 2D 영상 기반 3D 데이터 생성 결과

III. 결론

본 고는 최근 연구가 증가되는 인공지능을 이용한 3D 데이터에 대한 학습 방법 및 다양한 3D 데이터 표현 방법을 서술하였다. 식별 모델과 생성 모델의 발전에 따라 다양한 응용이 가능하며, 특히 3D 데이터 획득이 점차 용이해 지면서 관련 학습 데이터도 일부 공개되고 있어 3D 딥러닝에 대한 기술 개발이 가속화될 전망이다. 산업적인 측면에서 보면 관련된 표준기술의 발전으로 학습 데이터 및 학습 모델의 재사용이 가능하여 더 많은 응용서비스가 출현할 것으로 기대된다.

용어해설

Hand Crafted Feature 기존 영상인식 등에서 사용되는 특징점. 예를 들어, 얼굴 인식의 경우 외곽선의 모양 비율 등 알고리즘에서 정의한 몇 가지 요소

유클리디언 평면 고대의 수학자 유클리드가 정의한 평면으로 유클리드의 기하법칙(예를 들어, 임의의 한 점에서 다른 점으로 직선을 그을 수 있고, 직각은 모두 같다)이 적용되는 공간

Latent Code 잠재변수로 차원이 축소된 데이터 특징 벡터

약어 정리

CNN	Convolutional Neural Network
GPU	Graphics Processing Unit
RGB-D	Red Green Blue-Depth

참고문헌

[1] S. Khan and S.P. Yong, "A Comparison of Deep Learning and Hand Crafted Features in Medical Image Modality Classification," in *Proc. Int. Conf. Comput. Inform. Sci.*, Kuala Lumpur, Malaysia, Aug. 15-17, 2016, pp. 633-638.

[2] Taewan. Kim, "CNN, Convolution Neural Network 요약," Tawan.Kim Blog, Jan. 4, 2018, Available: http://taewan.kim/post/cnn/

[3] I.J. Goodfellow et al., "Generative Adversarial Nets." in *Proc. Adv. Neural Inform. Process. Syst.*, Montreal, Canada, Dec. 8-13, 2014, pp. 1-9.

[4] https://skymind.ai/wiki/open-datasets

[5] 이승욱 외, "3D 딥러닝 기술 동향," 전자통신동향분석 제33권 제5호, 2018,, pp. 103-110.

[6] E. Ahmed et al., "A survey on Deep Learning Advances on Different 3D Data Representations," 2018, arXive 1808.01462.

[7] Z. Cao et al., "3D Object Classifcation via Spherical Projections," 2017, arXiv 1712.04426.

[8] H. Su et al., "Multi-view Convolutional Neural Networks for 3D Shape Recognition," Sep. 2015, arXiv 1505.00880.

[9] 임성재 외, "한 장의 RGB 영상을 이용한 다시점 뎁스 맵 생성 기술," 대한전자공학회 2019년도 하계종합학술대회, 2019. 6.

[10] 강대기, "딥러닝을 위한 인공신경망 표준 포맷 동향," TTA 저널, vol. 179, 2018, pp. 85-90.