단일 이미지 기반 3D 모델 생성을 위한 딥-뉴럴 네트워크 분류 및 성능비교

김민경, 최유주
서울미디어대학원대학교 인공지능응용소프트웨어학과
muzzcats@naver.ac.kr, choirinayj@gmail.com

A Survey on Deep Neural Networks for 3D Reconstruction from a 2D Image

MinGeyung Kim Yoo-Joo Choi*
Department of AI Software Engineering, Seoul Media Institute of Technology
*Corresponding Author

요 약

단일 이미지로부터 3D 모델을 생성하는 방법은 메타버스와 가상현실 콘텐츠에 대한 필요성이 높아짐에 따라, 보다 효율적인 모델 생성방법으로서 관심이 높아지고 있다. 본 논문에서는 단일 이미지로부터 3D 모델을 자동 생성하는 기존 딥-뉴럴 네트워크들을 대상으로, 생성되는 3D 모델의 유형에 따라 기존 네트워크들을 분류하고, 주요 딥-뉴럴 네트워크의 형태와 특징, 그리고 모델 생성의 성능을 분석하고자 한다.

1. 서론

물체의 3D 형상표현(3D representation)은 일반적으로 과학적인 문제이며 컴퓨터 지원 기하학 설계(CAGD), 컴퓨터 그래픽스, 컴퓨터 애니메이션, 컴퓨터 비전, 의료 영상, 컴퓨터 과학, 가상 현실 ,디지털 미디어와 같은 다양한 분야의 핵심기술이다.

현재 3D 데이터를 확보하는 방법은 3D 스캐너를 사용하여 스캔 데이터를 정합하여 사용하는데, 이를 위해서는 스캔 장비가 추가로 필요하고 그리고 모델 완성을 위해서는 후처리 과정을 필요로 한다. 최근 메타버스와 가상현실 콘텐츠에 대한 필요성이 높아짐에 따라 효율적인 3D 가상 모델 구축 방법에 대한 관심과 연구가 활발히 이루어지고 있다.

본 논문에서는 효율적인 3D 모델 생성의 방법으로 2D 단일 이미지로부터 3D 모델을 생성하는 AI 기법들에 대한 소개와 분류, 그리고 기존 방법들의 주요 특징과 성능들을 분석 정리하고자 한다.

2. AI 기반 3D 모델 생성 기법의 분류

ShapeNet[1]과 Pix3D[2] 같은 2D 이미지와 이에 매칭되는 3D 모델에 대한 대규모 데이터셋이 구축되어 제공됨에 따라, 2D 이미지로부터 3D 모델을 생성하는 다양한 딥-뉴럴 네트워크에 대한 연구가 활발하게 진행되고 있다. 이러한 2D 이미지 기반 3D 모델 생성 딥-뉴럴 네트워크들은 3D 모델의 생성 형태에 따라 복셀표현(Voxel Representation)방법, 포인트 클라우드(Point Cloud), 메쉬 표현(Mesh Representation)으로 구분할 수 있고, 각 부류에 속하는 방법들은 <표 1>과 같다.

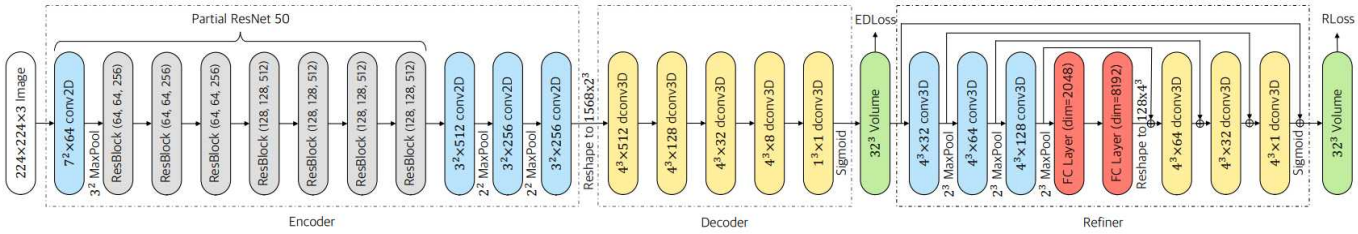
<표 1> 생성 3D 모델의 유형에 따른 2D 이미지 기반 3차원 모델 생성 딥-뉴럴 네트워크 방법들의 분류

3D Model Type	Deep Neural Networks
Voxel Reconstruction	Pix3D[2], 3D-EPN[3], 3D-R2N2[4], DRC[5], Pix2Vox++[6], Mem3D[7] IM-NET[17], OctNet[18]
Point Cloud	PointOutNet[13], 3D-LMNet[14], 3D-ARNet[15], 3D-PSRNET[16],
Polygonal Mesh Reconstruction	AtlasNet[8], Pixel2Mesh[9], 3DN[10], OccNet[11], IM-NET[12]

다음 절에서는 각각의 부류에 속하는 3D 모델 생성 딥-뉴럴 네트워크의 특성 및 제한점을 논한다.

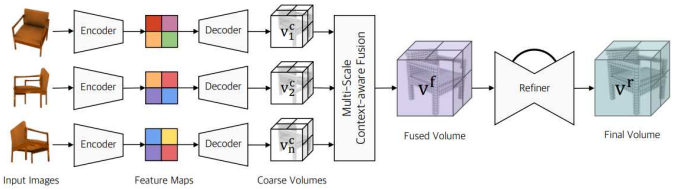
3. 복셀 생성 딥-뉴럴 네트워크 특성

3D 모델을 복셀 표현기법으로 생성하는 대부분의 제안들은 대부분 GPU의 메모리 제한 때문에 저



(그림 2) Pix2Vox++ 네트워크 구조[6]

해상도에서 처리된다.



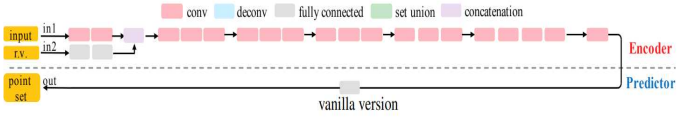
(그림 1) Pix2Vox++의 주요 구성요소 및 처리절차[6]

(그림 1)은 Pix2Vox++[6] 의네트워크의 구성을 보여주고 있다. Pix2Vox++ 처리 절차는 크게 4단계로 구분할 수 있다. 첫째, 인코더는 입력 이미지에서 특징 맵을 생성한다. 둘째 디코더는 각 기능 맵을 입력으로 사용하고 이에 따라 대략적인 3D 볼륨을 생성한다. 셋째, 단일 또는 여러 3D 볼륨이 다중 규모 컨텍스트 인식 융합 모듈로 전달되어 모든 거친 3D 볼륨에서 서로 다른 부분에 대한 고품질 재구성을 병렬로 선택하고 융합 3D 볼륨을 생성한다. 마지막으로, 리파이너는 융합된 3D 볼륨의 잘못 복구된 부분을 추가로 수정하여 최종 재구성 결과를 생성한다. (그림 2)는 정확한 3D 모델 생성을 위한 높은 계산 복잡도를 요하는 Pix2Vox++ 네트워크 구조를 보여주고 있다.

제한된 메모리를 가지고 고해상도 복셀 출력을 가지기 위하여 8진트리(Octree) 표현에 의한 접근 방법들[18]이 제시되었다. 8진트리를 이용하여 512^3 해상도까지 학습할 수 있는 능력을 확장하여 고밀도 복셀 방법의 계산 및 메모리 한계를 완화시켰지만 이 해상도조차도 시각적으로 자연스러운 형태를 만들지는 못하며 미세한 형상 디테일이 보존되지 않은 단점이 있다.

4. 포인트 클라우드 생성 딥-뉴럴 네트워크 특성

Fan 등은 단일 이미지로부터 포인트 클라우드를 생성하는 PointOutNet[13]을 제안하였다. 포인트 클라우드를 생성하는 방법들은 계산 및 메모리 한계를 갖는 복셀표현 방법의 대안 중 하나이며 LiDAR 센서나 깊이 카메라 등을 이용해 획득한 원시 데이터와 일치하는 정보를 가지며 복셀보다 낮은 밀도의 그리드를 사용하기 때문에 해상도에 따른 메모리 부하가 복셀표현 방법에 비해 적다. 하지만 점(points)들 간의 지역적인 연결성이 부족하고 포인트의 자유도가 높아 표면정보의 일부 또는 세부정보가 손실되어 물체표면(surface)를 표현하는데 한계를 가진다.

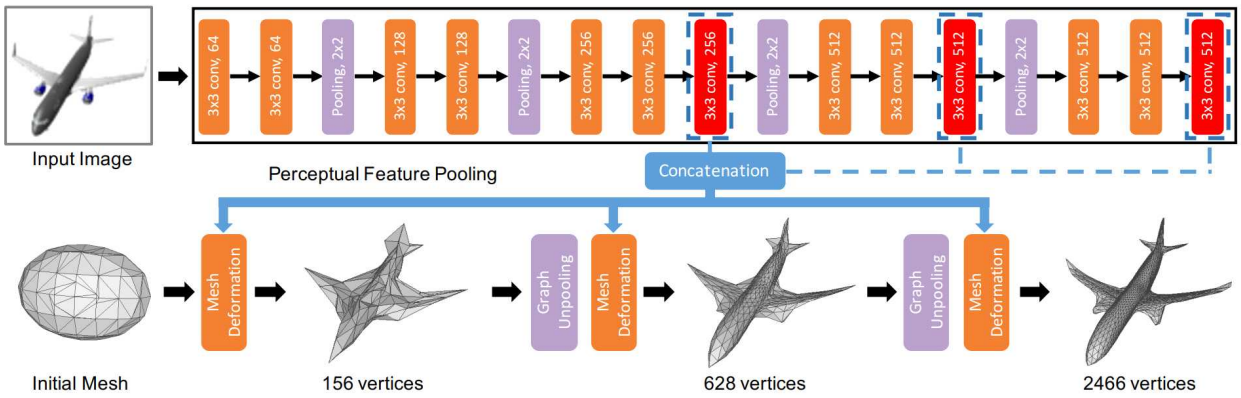


(그림 3) Fan 등이 제안한 PointOutNet의 구조[13]

(그림 3)은 바닐라 버전, 즉 단순한 구조의 버전으로 발표된 PointOutNet의 구조를 보여주고 있다. PointOutNet에는 인코더 단계와 예측기 단계가 있다. 인코더는 이미지 I와 임의 벡터 r의 입력쌍을 임베딩 공간에 매핑한다. 예측기는 N × 3 행렬 M으로 출력하며, 각 행에는 한 점의 좌표가 포함된다. 인코더는 컨볼루션 및 ReLU 레이어로 구성된다. 또한 랜덤 벡터 r이 포함되어 이미지 I의 예측을 교란시킨다. 예측기는 FCN(Fully Connected Layer)를 통해 N 점의 좌표를 생성한다. 이 버전은 간단하지만 실제로는 상당히 잘 동작하는 것으로 평가받고 있다.

5. 폴리곤 메쉬 모델 생성 딥-뉴럴 네트워크 특성

메쉬 표현방법은 고해상도의 영상을 메모리 부하 없이 경량화(light_weight) 하여 표현할 수 있고, 물체의 표면을 부드럽게 세부 정보 손실을 최소화 하여 3차원 형상 복원 수행할 수 있는 방법이다. 앞서 언급한 포인트 표현방식보다 세부정보를 보존하며, 폴리곤 기반으로 변형이 쉽기 때문에 점진적 학습을 통한 사물의 변형절차를 실험하기에 적합한 방법이다. 대부분 메쉬 기반 기법들은 탬플릿 모델을 형상에 맞추어 변형하고 해상도를 업-샘플링(up-sampling)하는 과정을 거친다.

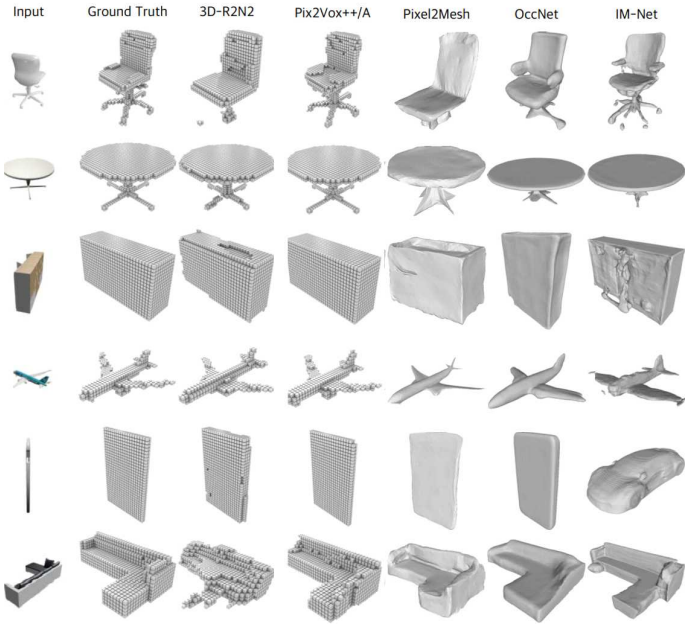


(그림 4) Pixel2Mesh 네트워크 구조[9]

Pixel2Mesh[9]은 한 장의 컬러 이미지로부터 3D 메쉬 모델을 생성하는 대표적인 ene-to-end 딥러닝 프레임워크이다. Pixel2Mesh의 구조는 (그림 4)와 같다. 전체 네트워크는 이미지 특징 네트워크와 계단식 메쉬 변형 네트워크로 구성된다. 이미지 특징 네트워크는 입력 이미지에서 지각 특징을 추출하는 2D CNN이며, 이는 메시 변형 네트워크에 의해 활용되어 타원체 메시를 원하는 3D 모델로 점진적으로 변형한다. 계단식 메쉬 변형 네트워크는 그래프 기반 컨볼루션 네트워크(GCN)로, 두 개의 그래프 풀링 레이어가 교차하는 세 개의 변형 블록을 포함한다. 각 변형 블록은 정점에 부착된 3D 형상 특징이 있는 현재 메쉬 모델을 나타내는 입력 그래프를 가져와 새로운 정점 위치 및 특징을 생성한다. 반면 그래프 풀링 레이어는 꼭짓점의 수를 늘려 세부 사항을 처리할 수 있는 능력을 높이면서도 삼각형 메시 토폴로지를 유지한다. 더 적은 수의 정점에서 시작하여 모델은 점진적으로 변형하고 메쉬 모델에 거친 방식에서 미세한 방식으로 세부 사항을 추가하는 방법을 학습한다. 안정적인 변형을 생성하고 정확한 메쉬를 생성하도록 네트워크를 훈련하기 위해 PointOutNet[13]에서 사용하는 Chamfer Distance loss를 확장한 손실함수와 세 가지 다른 메쉬 특정 손실, 즉, 표면 법선 손실(surface normal loss), 라플라시안 정규화 손실(Laplacian regularization loss) 및 가장자리 길이(Edge length loss) 손실을 사용한다.

6. 3D 모델 생성 결과 비교

(그림 5)는 각각 복셀 표현, 포인트 클라우드 표현, 폴리곤 메쉬로 3D 모델을 생성한 대표적인 딥-뉴럴 네트워크의 모델 생성결과를 비교한 결과를 보여주고 있다. 복셀표현 및 포인트 클라우드로 표현된 3D 모델의 표면을 표현하기 위해서는 마칭 큐브 알고리즘과 같이 표면 메쉬 모델 (surface mesh model)을 생성하기 위한 후처리를 수행하여야 한다. 반면, 메쉬 생성 네트워크의 경우, 2D 이미지 한 장을 입력으로 3D 메쉬모델을 바로 출력할 수 있는 방법이나 메쉬모델의 해상도를 고해상도로 생성하기 위해서는 아직도 메모리와 계산량의 부담이 존재하고 있어, 메쉬모델의 상세도를 높이기 위한 접근방법들이 연구되고 있다[19].



(그림 5) ShapeNet 데이터를 기반한 단일 영상 기반 3D 모델 생성 결과 비교 [6]

7. 결론

본 논문에서는 단일 이미지로부터 3D 모델을 자동 생성하는 기존 딥-뉴럴 네트워크들을 생성되는 3D 모델의 유형에 따라 분류하고, 주요 딥-뉴럴 네트워크의 형태와 특징, 그리고 모델 생성의 성능을

비교하였다. 각 네트워크들은 모델의 상세도를 높이기 위한 다양한 손실함수들과 모델 상세화 계층을 포함하고 있었다. 그러나 아직도, 생성 모델의 해상도를 높이기 위해서는 메모리와 계산량의 부담이 높아져서 생성되는 3D 모델의 해상도에 제약이 있음을 알 수 있었다.

향후 연구로서 3D 모델의 상세한 특성을 표현하기에 유리한 손실함수의 정의, 모델 상세화 계층의 설계 및 구현을 진행하여, 경량의 네트워크이면서 생성 3D 모델의 상세정보는 유지할 수 있는 딥-뉴럴 네트워크를 설계하고자 한다.

참고문헌

[1] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An Information-Rich 3D Model Repository", Tech. Rep. arXiv:1512.03012 [cs.GR], 2015.

[2] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman, "Pix3d: Dataset and methods for single-image 3d shape modeling", in Computer Vision and Pattern Recognition (CVPR), 2018.

[3] A. Dai, C. Ruizhongtai Qi, and M. Niessner, "Shape completition using 3d-encoder-predictor cnns and shape synthesis", In CVPR, pages 5868 - 5877, 2017.

[4] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction", in ECCV, 2016.

[5] Tulsiani, S., Zhou, T., Efros, A. A., & Malik, J. , Multi-view supervision for single-view reconstruction via differentiable ray consistency. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2626-2634), 2017.

[6] Xie, H., Yao, H., Zhang, S., Zhou, S., & Sun, W., "Pix2Vox++: multi-scale context-aware 3D object reconstruction from single and multiple images", International Journal of Computer Vision, 128(12), 2919-2935, 2020.

[7] Yang, S., Xu, M., Xie, H., Perry, S., & Xia, J. , "Single-view 3D object reconstruction from shape priors in memory", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 3152-3161, 2021.

[8] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry, "Atlasnet: A papier-mâché approach to learning 3D surface generation," arXiv preprint arXiv:1802.05384, 2018.

[9] Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., & Jiang, Y. G., "Pixel2Mesh: Generating 3d mesh models from single rgb images", In Proceedings of the European conference on computer vision (ECCV) (pp. 52-67), 2018.

[10] Wang, W., Ceylan, D., Mech, R., & Neumann, U. , "3DN: 3D deformation network", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1038-1046, 2019.

[11] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., & Geiger, A., "Occupancy networks: Learning 3d reconstruction in function space", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4460-4470, 2019.

[12] Chen, Z., & Zhang, H., "Learning implicit fields for generative shape modeling", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5939-5948, 2019.

[13] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3D object reconstruction from a single image," in CVPR, 2017.

[14] Mandikal, P., Navaneet, K. L., Agarwal, M., & Babu, R. V., 3D-LMNet: Latent embedding matching for accurate and diverse 3D point cloud reconstruction from a single image. arXiv preprint arXiv:1807.07796, 2018.

[15] Chen, H., & Zuo, Y., 3D-ARNet: An accurate 3D point cloud reconstruction network from a single-image. Multimedia Tools and Applications, 81(9), 12127-12140, 2022.

[16] Mandikal, P., KL, N., & Venkatesh Babu, R. , 3d-PSRNET: Part segmented 3D point cloud reconstruction from a single image. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018.

[17] Chen, Z., & Zhang, H., Learning implicit fields for generative shape modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5939-5948, 2019.

[18] G. Riegler, A. O. Ulusoy, and A. Geiger. OctNet: Learning deep 3D representations at high resolutions. In CVPR, pages 6620 - 6629. IEEE, 2017

[19] Park, J. J., Florence, P., Straub, J., Newcombe, R., & Lovegrove, S. (2019). DeepSDF: Learning continuous signed distance functions for shape representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 165-174).