한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 27, No. 3, P. 65~74

3차원 형상 복원을 위한 점진적 점유 예측 네트워크

김용규° 김덕수*

한국기술교육대학교 컴퓨터공학부

{caca209, bluekds}@koreatech.ac.kr

Progressive occupancy network for 3D reconstruction

Yonggyu Kim° Duksu Kim*

KOREATECH (Korea University of Technology and Education)

요 약

3차원 형상 복원(3D reconstruction)은 이미지 또는 영상 속 물체를 3차원 형상으로 복원하는 것을 말한다. 본 연구는 물체의 전반적 형상을 넘어 세부적인 모습까지 복원할 수 있는 표현력을 가진 3차원 형상 복원 네트워크인, 점진적 점유 네트워크를 제안한다. 본 연구가 제안하는 네트워크는 이미지 전체의 정보를 담고 있는 특징(feature)을 사용하는 기존 점유 네트워크와 달리, 수용 영역(receptive field)의 크기에 따라 다양한 수준의 이미지 특징을 추출해서 사용한다. 그리고, 다양한 수준의 이미지 특징을 디코더(decoder) 내 디코더 블록(decoder block)들에 순차적으로 반영하여, 형상 복원의 품질이 단계적으로 개선하는 네트워크 구조를 제안한다. 본 연구는 또한, 다양한 수준의 이미지 특징을 적절히 조합하여 사용하는 디코더 블록 구조를 제안한다. 본 연구는 제안하는 네트워크의 성능 검증을 위해 ShapeNet 데이터 세트를 사용하였으며, 기존의 점유 네트워크(ONet) 및 다양한 수준의 이미지 특징을 사용하는 최신 연구(DISN)와 성능 비교하였다. 그 결과, 기존 점유 네트워크 대비 세 가지 검증 지표 모두에서 높은 성능을 달성하였으며, DISN과는 대등한 수준의 성능을 보여주었다. 그리고 복원 형상의 시각적 비교 결과, 본 연구의 점진적 점유 네트워크가 기존 점유 네트워크 대비, 물체의 세부 모습을 잘 복원하는 것을 확인하였다. 또한, DISN이 복원 실패한 물체의 얇은 부분 또는 이미지에서 가려진 부분을 본 연구의 네트워크는 잘 잡아내는 결과를 확인할 수 있었다. 이러한 결과는 본 연구가 제안하는 점진적 점유 네트워크의 유용성을 검증하는 결과다.

Abstract

3D reconstruction means that reconstructing the 3D shape of the object in an image and a video. We proposed a progressive occupancy network architecture that can recover not only the overall shape of the object but also the local details. Unlike the original occupancy network, which uses a feature vector embedding information of the whole image, we extract and utilize the different levels of image features depending on the receptive field size. We also propose a novel network architecture that applies the image features sequentially to the decoder blocks in the decoder and improves the quality of the reconstructed 3D shape progressively. In addition, we design a novel decoder block structure that combines the different levels of image features properly and uses them for updating the input point feature. We trained our progressive occupancy network with ShapeNet. We compare its representation power with two prior methods, including prior occupancy network(ONet) and the recent work(DISN) that used different levels of image features like ours. From the perspective of evaluation metrics, our network shows better performance than ONet for all the metrics, and it achieved a little better or a compatible score with DISN. For visualization results, we found that our method successfully reconstructs the local details that ONet misses. Also, compare with DISN that fails to reconstruct the thin parts or occluded parts of the object, our progressive occupancy network successfully catches the parts. These results validate the usefulness of the proposed network architecture.

키워드: 3차원 복원, 음함수, 딥러닝, 점유 네트워크
Keywords: 3D reconstruction, implicit function, deep learning, Occupancy network

*corresponding author: Duksu Kim/Korea University of Technology and Education(bluekds@koreatech.ac.kr)

Received : 2021.06.11./ Review completed : 1st 2021.06.18. / Accepted : 2021.06.25.
DOI : 10.15701/kcgs.2021.27.3.65
ISSN : 1975-7883(Print)/2383-529X(Online)

1. 서론

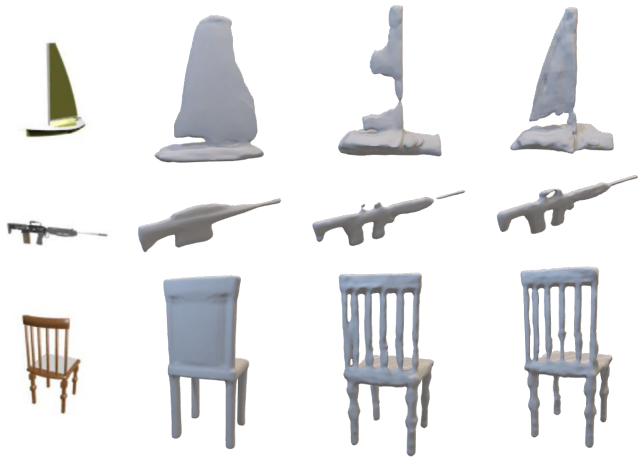
3차원 형상 복원(3D reconstruction)은 이미지 또는 영상 속 물체의 형상을 3차원 모델로 복원하는 기술이다. 전통적인 3차원 형상 복원은 여러 시점에서 획득한 이미지를 사용하는 다중시점 스테레오(multi-view stereo) 알고리즘들 [1]을 기반으로 발전해 왔으며, 단일 이미지로 부터 물체 형상을 복원하는 것은 도전적인 문제였다. 하지만, 딥러닝 기반의 인공지능 기술의 발전에 힘입어 단일 이미지 기반 3차원 형상 복원 기술이 크게 발전하였다 [2, 3, 4, 5].

인공신경망을 이용한 3차원 형상 복원 기술은 네트워크의 출력 형태에 따라 크게 명시적 표현 기법과 음함수 기반의 암묵적 표현 기법으로 분류할 수 있다. 명시적 표현 기법은 복셀(voxel), 포인트(point), 메시(mesh) 등 과 같은 기하 객체를 이용하여 형상을 표현하는 기법으로, 형상 표현이 직관적이고 사용이 쉽다는 장점을 가진다 [6, 3, 2, 7, 8, 9, 10, 4]. 하지만, 명시적 표현 방법을 사용하는 기술들은 제한적인 해상도 또는 낮은 토폴로지(topology) 표현의 한계를 가진다.

물체의 내외부 경계면을 표현하는 음함수(implicit function)를 기반으로 물체의 형상을 복원하는 암묵적 표현을 사용하여, 명시적 표현 기법의 한계인 해상도 문제를 해결할 수 있다. Mescheder 등 [11]은 주어진 좌표(또는 포인트)의 물체 내부(점유) 여부(확률)를 출력하는 점유 네트워크(occupancy network, ONet)를 제안하였다. ONet은 입력 포인트의 점유 여부만 판단하기 때문에, 입력 포인트의 개수에 제한이 없이 원하는 해상도로 물체 형상을 복원 할 수 있다는 장점을 가진다. 그 결과, ONet은 이미지 속 물체 종류에 맞추어 전반적인 형상을 성공적으로 복원하는 결과를 보여주었다. 하지만, 복원된 물체의 세부 모습(예, 의자 등받이의 기둥들)은 잘 표현되지 않는 한계를 가진다.

본 연구는 물체의 세부적 모습까지 복원할 수 있도록 표현력을 높인 점진적 점유 네트워크(progressive occupancy network)를 제안한다. 이미지 전체의 정보를 담고 있는 전역 특징(global feature)을 사용해서 물체의 형상을 추론하는 ONet [11]과 달리, 본 연구는 수용 역영(receptive field)의 크기에 따른 다양한 수준의 이미지 특징을 추출해서 사용함으로서, 물체의 세부적 모습 복원에 대한 네트워크의 표현력을 높인다(3.3장 및 3.4장). 또한, 다양한 수준의 이미지 특징을 단계적으로 사용함으로서, 3차원 형상 복원 품질을 점진적으로 향상시키는 점진적 점유 네트워크 구조를 제안한다(3.2장). 그리고 물체 세부 형상 복원 과정에서 전역적 특징이 깨지는 문제를 해결할 수 있도록 서로 다른 수준의 이미지 특징을 조합해서 사용하는 디코더 네트워크 및 디코더 블록 구조를 제안한다(3.5장).

본 연구는 제안하는 네트워크를 ShapeNet [12] 데이터 세트를 사용하여 학습시켰으며, 그 성능을 두 가지 기존 연구(ONet [11], DISN [13])와 비교 하였다(4장). 그 결과, 3차원 복원 품질을 측정하는 세 가지 지표 모두에서 ONet대비 높은 성능을 달성하였으며, DISN와 대등한 수준의 점수를 얻었다. 또한, 시각적 복원 결과 비교를 통해 ONet 및 DISN이 표현하지 못한 물체의 세부적 형상을 잘 잡아내는 모습을 확인 할 수 있었다(Figure 1과 Figure 5).



(a) Input (b) ONet [11] (c) DISN [13] (d) Ours

Figure 1: 3D reconstruction results using ONet [11], DISN [13], and our progressive occupancy network.

2. 관련 연구

한장의 이미지로 부터 3차원 형상을 복원하는 인공신경망 기술은 출력의 형태에 따라 명시적 표현 기법과 음함수(implicit function) 기반의 암묵적 표현 기술로 구분된다. 명시적 표현 기법은 사용하는 기하 객체(geometric primitive)에 따라 다시 복셀(voxel) 기반 [6, 3], 포인트 클라우드(point cloud) 기반 [2, 7, 8, 9], 그리고 메시(mesh) 기반 [10, 4, 14] 기법으로 구분된다.

복셀은 데이터 구조가 단순하고 사용이 쉽다는 장점을 가지며, 복셀의 격자에 대한 3차원 합성곱 신경망을 이용한 3차원 형상 복원 기술들이 연구되었다 [6, 3]. 하지만 복셀은 해상도 향상에 따른 메모리 부하 크다는 단점을 가지며, 복셀 기반의 인공 신경망을 통해 처리할 수 있는 격자의 해상도가 상대적으로 낮다 (예, 32^3 또는 64^3)는 한계를 가진다. 최근에는 팔진트리(octree)와 같은 적응형 데이터 구조를 사용하는 등의 방법으로 이러한 해상도의 한계를 극복하기 위한 연구들이 진행되었다 [15, 16, 17]. 하지만, 여전히 해상도가 일정 수준으로 제한되며(예, 256^3) 얕은 구조(shallow architecture)에서 작은 크기의 배치를 사용해야 한다는 단점을 가진다.

복셀 기반 3차원 형상 복원 기법의 대안 중 하나로, 물체의 형상을 포인트 클라우드 형태로 출력하는 인공신경망 기술들이 연구되었다 [2, 7, 8, 9]. 포인트는 필요한 부분에만 위치하면 되기 때문에 해상도(포인트의 수)에 따른 메모리 부하가 복셀에 비해 적다는 장점을 가진다. 하지만, 포인트는 위치에 대한 자유도가 높고 연결성(connectivity)을 갖지 않기 때문에 물체의 형상을 정확히 표현하는데 한계를 가진다. 또한, 물체의 표면(surface)을 표현하기 위해서는 마칭큐브(marching cube) [18] 등과 같은 후처

리를 필요로 한다.

메시는 3차원 형상을 표현하기 위해 가장 널리 사용되는 기학 객체 중 하나로, 직접 메시를 생성하는 3차원 형상 복원 인공신경망 기술들도 연구되었다 [10, 4, 14]. Wang 등 [4]은 템플릿 도형(예, 타원체)을 형상에 맞추어 변형(메시를 구성하는 정점의 이동)하고 메시의 해상도를 업-샘플링(up-sampling)하는 과정을 반복하는 Pixel2Mesh 방법을 제안하였다. 그들은 카메라 정보를 이용하여 3차원 공간 상 정점을 이미지 평면으로 투사(projection)하고, 해당 위치에 대한 이미지 특성인 지각적 특성(perceptual feature)을 추출하여 사용하였다. 그 결과, 3차원 형상의 세세한 부분까지 높은 품질로 복원하는 결과를 보여주었다. 하지만, 낮은 토폴로지의 간단한 템플릿 도형을 사용하며 템플릿과 다른 토폴로지를 가지는 형상을 복원할 수 없다는 아쉬움을 가진다. Groueix 등 [14]은 파라메트릭 표면(parametric surface)들을 이용하여 물체의 3차원 형상을 복원하는 AtlasNet을 제안하였으며, 다양한 토폴로지의 물체 형상 을 높은 품질로 복원하였다. 하지만, 폴리곤 패치를 붙인 형태로 메시 사이의 연결성 정보를 표현하지 못하고 서로 중첩이 발생 가능하다. 그 결과, 결과물들의 표면이 매끄럽게 표현되지 못한다는 단점을 가진다.

위에서 살펴본 것과 같이, 기하 객체를 사용한 명시적 표현 방법을 사용하는 기술들은 제한적인 해상도 또는 낮은 토폴로지 표현의 한계를 가진다. 음함수 기반의 형상 표현은 이러한 한계를 극복할 수 있는 대안 중 하나다 [11, 19, 20, 13, 21]. Mescheder 등 [11]은 주어진 3차원 포인트가 물체의 내부(점유)일 확률을 추론하는 점유 네트워크(ONet)를 제안하였다. 그들은 3차원 격자의 각 셀의 점유를 예측하고, 점유 셀을 분할 및 다시 점유예측을 수행하는 것을 반복함으로서 해상도 향상에 따른 메모리 부하 문제를 해결하였다. 또한, 마칭 큐브를 이용하여 최종 메시를 생성함으로써 메시 중첩 및 연결성 정보 부재 문제를 해결하였다. 그 결과, ONet은 입력 이미지 속 물체의 형상을 높은 해상도를 가지는 메시로 복원해 냈다. 하지만, 이미지의 전체의 정보를 담고 있는 특성 벡터(전역 특성)로 잠재 공간(latent space)을 조절함으로서, 물체의 세부 상세를 잘 표현하지 못하는 한계점을 가진다(Figure 1(b)). Saito 등 [19]은 지역적 이미지 특성(local image feature)을 사용하여 옷을 입은 사람의 형상을 세부 상세까지 높은 품질로 복원하였다. Xu 등 [13]은 이미지의 전역적 특성(global feature)과 지역적 특성(local feature)을 동시에 사용하여 부호가 있는 거리 장(signed distance field, SDF)을 예측하는 DISN(Deep Implicit Surface Network)을 제안하였다. DISN은 전역적 특성과 지역적 특성 각각으로 예측한 SDF를 합하여 최종 SDF를 예측하며, 마칭큐브 알고리즘으로 복원한 물체의 형상이 세부 상세까지 잘 잡아내었다. 하지만, 물체의 매우 얇은 부분이 복원 시 손실되는 경우가 종종 있다는 한계를 가진다(Figure 1(c)).

본 연구가 제안하는 네트워크는 ONet 구조 [11]에 기반을 두고 있다. 하지만, 이미지 전체 정보를 담은 전역적 특성뿐만이 아닌 3차원 포인트의 위치에 기반을 둔 지역적 특성을 추출하고 전역적 특성과 함께 사용한다는 점에서 ONet과 차별성을 가진다. 또한 본 연구는 지역적 특성을 사용한다는 측면에서는 Saito 등 [19] 및 DISN [13]과 유사하지만, 전역적 특징에서 시작하여 점점 더 세부적인 지역적 특징을 점진적으로 사용한다는 점에서 차별성을 가진다.

3. 제안하는 방법

3.1 시스템 개요

본 연구는 ONet [11]에서 제시한 3차원 형상 복원 시스템을 기반 시스템으로 사용한다. ONet은 3차원 적응형 격자(예, 팔진트리), 점유 네트워크(occupancy network), 그리고 메시 생성 모듈로 구성된다. 최초의 격자는 낮은 해상도를 가지며, 격자의 각 포인트를 점유 네트워크로 전달한다. 점유 네트워크는 이미지와 3차원 좌표(포인트)를 입력으로 받고, 해당 포인트가 물체 형상의 내부(점유)일 확률을 출력한다. 격자 포인트 중 임계값 이상의 점유 확률을 가지면 포인트 주변의 복셀은 8분할 되며, 분할된 격자의 점들은 다시 점유 네트워크로 전달되어, 점유 여부가 결정된다. 목표 해상도에 도달할 때까지 위 과정이 반복되며, 목표 해상도에 도달한 격자는 메시 생성 모듈로 전달된다. 메시 생성 모듈은 주어진 격자에 마칭큐브(marching cube) 알고리즘 [18]을 적용하여 결과 메시를 생성한다.

본 연구는 ONet의 점유 네트워크 대비, 물체 형상의 세부를 정확히 표현할 수 있는 점진적 점유 예측 네트워크를 제안한다.

3.2 점진적 점유 예측 네트워크 개요

본 연구가 제안하는 점유 예측 네트워크는 입력 이미지 촬영에 사용된 카메라 정보를 알고 있는 경우를 대상으로 한다. 즉, 네트워크의 입력은 이미지, 카메라 정보, 그리고 3차원 좌표(포인트)이며, 출력은 해당 포인트 위치의 점유 여부(확률)다. 이미지로부터의 얻은 정보(관측)를 $x \in \mathcal{X}$, 주어진 포인트를 $p \in \mathbb{R}^3$, 그리고 카메라 정보를 $c \in \mathcal{C}$라고 했을 때, 제안하는 네트워크가 표현하는 함수는 수식 1과 같다.

$$f_\theta : \mathbb{R}^3 \times \mathcal{X} \times \mathcal{C} \rightarrow [0, 1] \tag{1}$$

Figure 2는 본 연구가 제안하는 점진적 점유 네트워크의 전체 구조를 보여준다. 제안하는 네트워크는 인코더, 지역 특징 추출기, 디코더 블록들, 그리고 점유 추론 블록으로 구성되어 있다. 입력 이미지는 224×224 크기로 전처리되어 인코더의 입력으로 들어간다. 인코더(encoder)는 합성곱 신경망(convolution neural network, CNN)을 사용하며, 입력 이미지에서 특징을 추출하는 역할을 한다(3.3장). 지역 특징 추출기(local feature extractor)는 카메라 정보와 포인트 위치 정보를 기반으로 이미지에서 해당 포인트 위치에 대응되는 이미지 특징인 지역적 특징(local feature)을 추출하는 역할을 한다(3.4장). 디코더(decoder) 블록들은 지역 특징들을 반영하여 포인트 특징(point feature)을 갱신한다

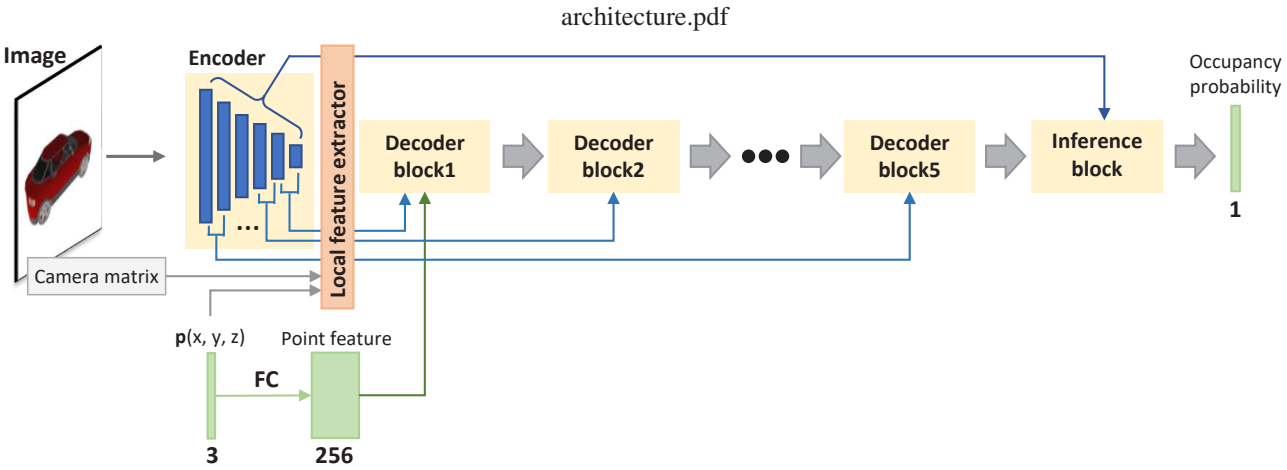


Figure 2: Progressive occupancy network architecture

(3.5장). 디코더 네트워크는 총 다섯 개의 디코더 블록으로 구성되어 있으며, 각 디코더 블록을 서로 다른 수준의 지역적 특징들을 사용하여 포인트 특징을 점진적으로 갱신한다. 마지막 점유 추론 블록은 디코더 블록들을 거쳐 나온 포인트 특징을 입력으로 받아 점유 여부를 출력한다(3.6장).

3.3 인코더

본 연구에서 사용하는 인코더 네트워크는 여섯 개의 합성곱 블록(convolution block)으로 구성된 합성곱 신경망으로, Table 1은 각 합성곱 블록 및 네트워크 전체의 구조를 요약한 것이다. 각 합성곱 층(Conv2D)은 활성 함수로 ReLU를 사용하였으며, Table 1에는 간결성을 표기를 생략하였다. 각 합성곱 블록의 출력(Table 1의 굵은 폰트)은 서로 다른 수용 영역을 가진 특징 맵으로, Conv. block1에서 Conv. block5로 진행할수록 수용 영역이 넓어진다. 수용 영역이 좁을수록 특징 맵의 각 원소는 이미지의 좁은 영역에 대한 지역적 정보를 담게 되며, 수용 영역이 넓어지면 각 원소가 이미지의 넓은 영역에 대한 정보를 담게 된다. 마지막 합성곱 블록(Conv. block6)의 출력은 이미지 전체에 대한 정보를 담고 있으며, 본 연구는 이를 전역적 특징(global feature)로 정의한다. 또한, Conv. block1에서 Conv. block5의 특징맵을 지역적 특징맵(local feature map)으로 정의한다.

Table 1: Encoder network configuration

Type	Layer	Kernel size, stride, padding	Output shape (d,w,h)
Conv. block1	Conv2d	3,1,1	(16,224,224)
	Conv2d	3,1,1	(16,224,224)
	Conv2d	3,2,2	**(32,112,112)**
Conv. block2	Conv2d	3,1,1	(32,112,112)
	Conv2d	3,1,1	(32,112,112)
	Conv2d	3,2,1	(64,56,56)
	Conv2d	3,1,1	(64,56,56)
	Conv2d	3,1,1	**(64,56,56)**
Conv. block3	Conv2d	3,2,1	(128,28,28)
	Conv2d	3,1,1	(128,28,28)
	Conv2d	3,1,1	**(128,28,28)**
Conv. block4	Conv2d	5,2,2	(256,14,14)
	Conv2d	3,1,1	(256,14,14)
	Conv2d	3,1,1	**(256,14,14)**
Conv. block5	Conv2d	5,2,2	(512,7,7)
	Conv2d	3,1,1	(512,7,7)
	Conv2d	3,1,1	(512,7,7)
	Conv2d	3,1,1	**(512,7,7)**
Conv. block6	Conv2d	3,2,2	(256,5,5)
	Conv2d	3,2,1	(256,3,3)
	Conv2d	3,2,1	(256,2,2)
	Linear	(input, output) (1024, 256)	**(256)**

본 연구는 카메라의 정보를 이용하여 지역적 특징맵으로 부터 입력 포인트의 위치(x,y,z 좌표)에 대응하는 특징들을 추출하여 사용한다(3.4장).

3.4 지역적 특징 추출

지역적 특징맵으로 부터 포인트 위치에 대응하는 지역적 특징을 추출하기 위해 본 연구는, Wang 등 [4]의 연구에서 제안한 지각적 특징 풀링(perceptual feature pooling) 방법을 사용한다. Figure 3은 포인트에 대응하는 지역 특징 추출 방법을 보여주며, 그 과정은 다음과 같다. 우선, 이미지 상에서 포인트가 대응되는 위치를 얻기 위해 카메라 정보를 이용하여 3차원 포인트를 2차원 이미지 평면으로 투영한다. 그리고 지역적 특징맵의 크기에 맞추기 위한 축소 연산을 수행하여, 특징맵 상에서 포인트의 대응 위치를 얻는다. 마지막으로 특징맵에서 포인트에 대응하는 특징을 추출한다. 특징맵 상의 포인트 위치는 정수가 아닐 수 있으며, 그 경우 포인트 주변 네 개 특징점의 특징값을 쌍-선형보간(bilinear interpolation)하여 계산한다.

포인트에 대응하는 지역적 특징 추출은 모든 지역적 특징맵에 대해 적용되어 계산되며, 그렇게 추출된 지역적 특징들은 디코더로 전달되어 점유 네트워크 표현력 향상을 위해 사용된다(3.5장).

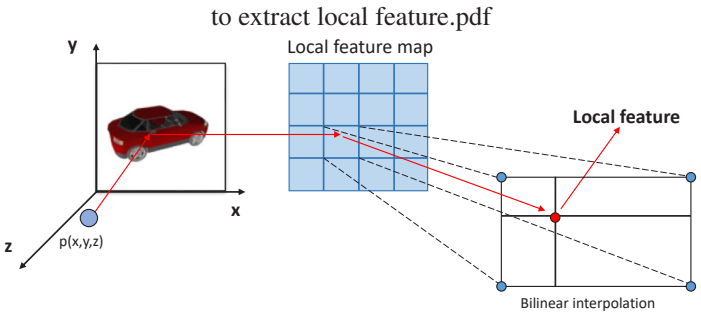


Figure 3: The process of local feature extraction

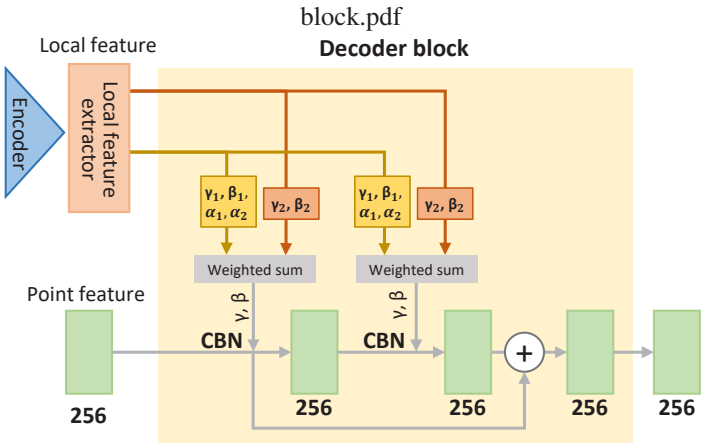


Figure 4: Decoder block architecture

3.5 디코더

본 연구가 제안하는 점진적 점유 예측 네트워크의 디코더는 다섯 개의 디코더 블록으로 구성되어 있으며, Figure 4는 디코더 블록의 구조를 보여준다. 디코더 블록의 입력은 256차원의 포인트 특징과 지역적 특징 추출기를 통해 추출된 지역적 특징들이다. 최초의 포인트 특징은 3차원의 입력 포인트를 완전 연결계층(fully connected layer)을 통과 시켜 256차원으로 확대하여 생성한다(Figure 2). 디코더로 들어온 포인트 특징은 조건부 배치 정규화(conditional batch normalization, CBN) 후, ReLU 활성함수층과 합성곱 층(Conv1d)을 통과한다. 이 과정(CBN → ReLU → Conv1d)는 한차례 더 반복된다. 디코더 블록의 최초 입력 포인트 특징은 잔차(residual)을 통해 전달되어, 위 과정을 거쳐 나온 포인트 특징에 더해져서 디코더 블록의 최종 출력을 생성한다.

ONet [11]는 조건부 배치 정규화를 위한 매개변수 γ와 β를 인코더의 최종 출력(예, Conv. block6의 출력)을 통해 추론하고 사용하였다. 그 결과, 물체의 전체적 형상은 잘 복원해내지만 세부적인 모습까지는 잘 잡아내지 못하는 한계를 보여주었다. 본 연구는 이러한 한계를 극복하기 위해 조건부 배치 정규화 매개변수 추론을 위해 다양한 수준의 지역적 특징을 사용한다. 디코더 블록에는 두 개의 지역적 특징이 입력으로 들어오며, 각 블록으로 들어오는 지역적 특징들은 서로 다르다. 첫 번째 디코더 블록(decoder block1)은 인코더의 가장 후미에 있는 합성곱 블록에서 출력된 특성맵(Conv. block5와 Conv. block6의 출력)에서 추출된 지역적 특징이 입력으로 주어지며, 두 번째 디코더 블록(decoder block2)은 인코더의 네 번째와 다섯 번째 지역적 특징맵(Conv. block4와 Conv. block5의 출력)에서 추출된 지역적 특징을 입력으로 받는다. 동일한 규칙으로 마지막 디코더 블록까지의 지역적 특징 입력이 결정된다(Figure 2). 두 개의 지역적 특징은 조건부 배치 정규화의 매개변수를 추론하는데 사용되며, 그 과정은 다음과 같다. 우선, 각 지역적 특징을 통해 (γ_1, β_1)과 (γ_2, β_2)를 추론한다. 그리고 두 추론 결과를 가중치 합하여 최종 γ와 β를 결정한다. 가중치 합을 위한 가중치 α_1과 α_2는 두 지역적 특징 중 더 넓은 수용영역을 가지는(인코더 뒤쪽의) 지역적 특징을 사용하여 추론하도록 하였다. 매개변수 및 가중치를 추론하는 네트워크는 두 개 층으로 이루어진 합성곱 신경망을 사용하였으며, ReLU활성함수를 사용하였다. 가중치 합을 위한 공식은 다음 수식 2과 3와 같다.

$$\gamma = \alpha_1\gamma_1 + (1-\alpha_1)\gamma_2 \tag{2}$$

$$\beta = \alpha_2\beta_1 + (1-\alpha_2)\beta_2 \tag{3}$$

인코더의 전반부(예, Conv. block1)에서 생성된 지역적 특징맵의 각 원소는 이미지의 국지적 영역에 대한 정보를 담고 있으며, 후반부(예, Conv. block6)로 갈수로 점점 이미지의 전체적 특징을 담게 된다. 따라서, 디코더 블록에 입력으로 들어오는 지역적 특징에 따라 각 디코더 블록이 내포하는 표현력도 물체의 세부적 모습에서부터 물체의 전체적 형상까지 서로 달라진다. 본 연구는 물체의 전체적 형상을 추론하는 디코더 블록을 앞쪽에, 그리고 물체의 세부 형상을 추론하는 블록을 뒤쪽에 배치하는 구조를 선택하여, 네트워크가 물체의 전체적 형상을 먼저 추론하고 점진적으로 세부적인 모습을 추론할 수 있도록 설계하였다(4.4장).

디코더에서 출력된 최종 포인트 특징은 점유 추론 블록에 전달되어, 포인트 위치에 대한 점유 확률 예측을 위해 사용된다.

3.6 점유 추론 블록

점유 추론 블록은 디코더에서 생성한 포인트 특징과 인코더의 모든 블록에서 추출된 지역적 특징(Conv. block6경우, 전역적 특징)을 결합(concatenation)한 종합적 이미지 특징 벡터를 입력으로 받는다. 포인트 특징은 종합적 이미지 특징 벡터를 기반으로 한 조건부 배치 정규화를 거친 후, 완전 연결층으로 전달된다. 마지막으로, 완전 연결층은 점유 확률을 출력한다.

본 연구는 디코더의 출력을 바로 완전 연결층으로 전달하는 경우, 카메라의 시점에 대한 종속성이 강하게 작용한다는 것(예, 가려진 부분의 형상 표현을 못함)을 경험적으로 발견하였다. 그리고 종합적 이미지 특징 벡터의 사용이 이런 현상을 완화해준다는 것을 확인하였다.

3.7 손실함수

본 연구는 제안하는 점진적 점유 예측 네트워크 학습을 위해 손실 함수(loss function)는 이진 교차 엔트로피(binary cross entropy, BCE)를 사용하였다. 이미지 i에 대한 관측과 카메라 정보가 x_i, c_i, 그리고 j 번째 입력 포인트를 p_{ij}라고 할 때, 손실함수는 다음 수식 4와 같다.

$$Loss = \sum_{j=1}^{k} BCE(f_\theta(p_{ij}, x_i, c_i), o_{ij}) \quad (4)$$

4. 구현 및 결과

4.1 데이터 세트 및 실험 환경

본 연구는 제안하는 점진적 점유 네트워크의 성능을 확인하고 기존 연구들과의 비교를 위해, 3차원 형상 복원 연구를 위해 널리 사용되는 ShapeNet [12] 데이터 세트를 사용하였다. 그 중, Choy 등 [6]과 같은 13가지 종류에 대한 데이터 세트를 사용하여 학습 및 테스트를 진행하였다. 전체 데이터 샘플은 43,755개로 학습(training), 검증(validation), 테스트(test) 데이터로 각각 70%, 10%, 20%의 샘플을 사용하였다.

제안하는 모델은 PyTorch를 기반으로 구현하였으며, 최적화 방법(optimizer)으로는 Adam 알고리즘(learning rate = 0.0001, $\beta 1$ = 0.9, $\beta 2$ = 0.999, ϵ= 10^{-8})을 사용하였다. 조건부 배치 정규화의 μ와 σ 운동량(momentum)은 0.1, ϵ는 10^{-5}를 사용하였다. 최종 메시 생성에는 Mescheder 등 [11]이 제안한 다중 해상도 등가면 추출(multiresolution isosurface extraction, MISE) 방법을 사용 하였다. MISE를 위한 최초 해상도는 32^3이며 활성 복셀에 대한 최대 해상도는 256^3으로 ONet [11]과 동일하게 설정하였다.

본 연구는 기존 인공신경망 기반 3차원 복원 네트워크 중, 대표적인 음함수 표현 모델인 ONet [11]과 이미지에서 추출한 전역적 정보와 지역적 특징을 함께 사용하는 특징을 가지는 DISN [13]와 성능을 비교한다. 두 모델 모두 저자가 github에 공개한 코드를 사용하였으며, 제안하는 모델과 동일한 데이터 세트와 컴퓨팅 환경에서 학습을 진행하였다.

4.2 검증지표

본 연구는 3차원 복원 연구에 널리 사용되는 세 가지 지표인, Chamfer distance, IoU, 그리고 법선 일관성(normal consistency) 지표를 사용하여 제안하는 네트워크의 성능을 검증하였다 [11].

Chamfer distance는 포인트 클라우드 사이의 유사도를 측정하는 지표 중 하나로, 가장 가까운 점들 사이의 거리를 기준으로 사용한다. 따라서, 값이 낮을 수록 높은 유사도를 의미한다. 본 연구는 정답 메시와 네트워크를 통해 추론된 메시의 유사도를 분석하기 위해, 각 메시에서 10만 개의 포인트를 임의 샘플링(random sampling)하여 포인트 클라우드들을 만들고, Chamfer distance를 계산하였다.

IoU(Intersection over Union)는 두 메시 사이의 유사도를 측정하는 지표로, 두 메시의 부피(volume)합 대비 두 메시가 겹치는 부분의 부피 비율을 계산하는 방법으로 유사도를 측정한다. 따라서, IoU는 높을 수록 높은 유사도를 의미한다.

법선 일관성(normal consistency)은 두 메시 표면 사이의 법선 유사도를 표현하는 점수로, 두 메시 사이의 고차원 유사도를 측정하는 지표다 [12]. 따라서 법선 일관성은 높을 수록 높은 유사도를 의미한다.

4.3 결과 및 비교

Table 2는 본 논문이 제안하는 점진적 점유 네트워크(Ours)와 두 가지 비교대상 모델인 ONet과 DISN의 테스트 데이터 세트에 대한 Chamfer distance, IoU, normal consistency 점수다. 이미지의 전역적 특징과 지역적 특징을 함께 사용하는 DISN과 Ours가 모든 지표에서 ONet보다 평균적으로 좋은 성능을 보여주는 것을 확인 할 수 있다. Figure 1와 Figure 5는 입력 이미지와 복원된 형상의 예를 보여주고 있으며, ONet에 대비하여 DISN과 Ours가 물체의 세부적 형상을 잘 표현하는 모습을 볼 수 있다. Figure 1에서 총의 손잡이 부분, 의자 등받이의 구멍과 다리의 굴곡 등이 그 대표적인 예들이다. 이러한 결과는 이미지에서 지역적 정보를 추출해서 사용하는 것이 물체의 세부적인 형상을 추론하는데 도움이 된다는 것을 보여주는 결과다.

DISN과 비교하여 Ours는 Chamfer distance와 IoU에서는 조금 좋은 성능을, 그리고 normal consistency에서는 조금 낮은 성능을 보이는 등 세 가지 평가지표에서 대등한 수준의 성능을 보여준다. 하지만, 복원 결과물의 시각적 품질에서는 차이를 확인할 수 있다(Figure 1와 Figure 5). DISN은 형상 복원 과정에 물체의 얇은 부분(예, Figure 1에서 배의 돛, 총의 가늠쇠 및 총구)을 복원하는 데 실패하는 모습을 확인 할 수 있다. 본 연구가 제안하는 점진적 점유 네트워크는 DISN이 놓친 얇은 부분까지 잘 복원해 내는 모습을 확인할 수 있다(Figure 1와 Figure 5). 이는 제안하는 점진적 점유 네트워크의 디코더 내 각 디코더 블록이 내포하는 표현력이 물체의 전역적 특성과 세부적 특성을 모두 아우르고 있으며, 단계별로 포인트 특징을 갱신함에 따라 얻을 수 있는 결과다(4.4 장).

4.4 디코더 블록의 효과

본 연구는 디코더에서 서로 다른 수준의 지역적 특징을 활용하는 디코더 블록의 효과 확인 및 디코더 블록 수에 따라 성능 변화를 분석하기 위해, 각 디코더 블록에서 나온 포인트 특징을 사용하여 3차원 복원을 수행해 보았다. Figure 6는 각 디코더 블록의 출력을 사용하여 측정한 세 가지 지표에 대한 점수를 보여준다. 그래프에서 볼 수 있듯, 디코더 블록을 지나감에 따라 세 가지 지표 모두에서 점진적으로 성능이 향상됨을 확인 할 수 있었다. 본

Figure 5: The input image and the visualization of 3D reconstruction results using ONet [11], DISN [13], and our progressive occupnacy network

Table 2: Result comparison of implicit function based network

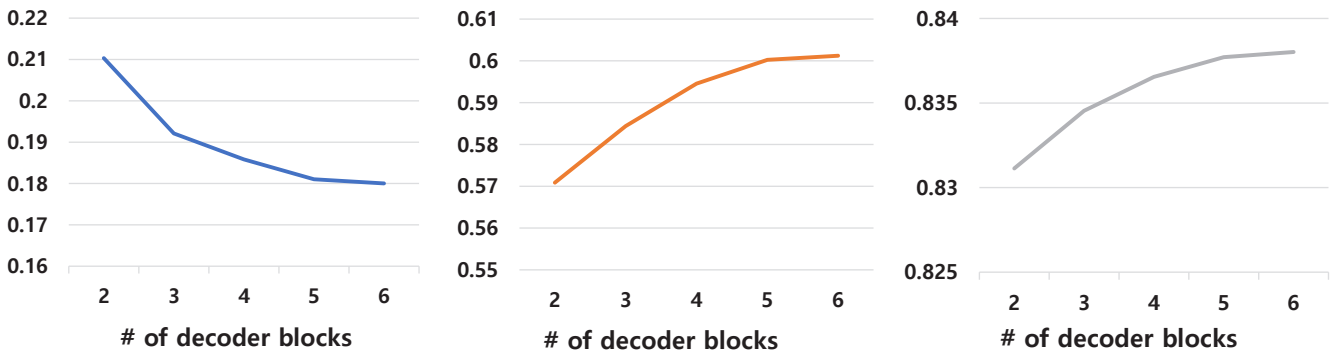
	Chamfer Distance (↓)			IoU (↑)			Normal Consistency (↑)		
category	ONet	Disn	Ours	ONet	Disn	Ours	ONet	Disn	Ours
airplane	0.147	**0.130**	0.134	0.571	0.570	**0.578**	**0.840**	0.829	0.831
bench	**0.155**	0.165	0.171	**0.485**	0.451	0.465	**0.813**	0.797	0.802
cabinet	0.167	0.164	**0.163**	0.733	**0.742**	0.741	0.879	**0.883**	0.880
car	**0.159**	0.173	0.169	**0.737**	0.728	0.734	**0.852**	0.846	0.847
chair	0.228	**0.211**	0.212	0.501	**0.532**	0.529	0.823	**0.826**	0.824
display	0.278	0.222	**0.215**	0.471	**0.555**	0.552	0.854	**0.874**	0.864
lamp	0.479	0.268	**0.261**	0.371	0.445	**0.449**	0.731	**0.757**	0.751
loudspeaker	0.300	0.241	**0.236**	0.647	**0.701**	0.697	0.832	**0.860**	0.859
rifle	0.141	0.111	**0.104**	0.474	0.552	**0.557**	0.766	**0.801**	0.799
sofa,couch,lounge	0.194	0.186	**0.184**	0.680	0.690	**0.693**	0.863	**0.870**	0.869
table	0.189	**0.168**	0.173	0.506	0.538	**0.541**	**0.858**	**0.858**	**0.858**
telephone	0.140	0.141	**0.131**	**0.720**	0.712	0.715	**0.935**	0.930	0.929
vessel	0.218	0.194	**0.191**	0.53	0.558	**0.562**	**0.794**	0.789	0.783
mean	0.215	0.183	**0.180**	0.571	0.598	**0.601**	0.834	**0.840**	0.838

연구는 또한, 인코더 블록의 수와 디코더 블록의 수를 더 늘려보았지만, 일곱 개 이상의 블록에서는 유의미한 성능 향상이 없음을 확인하였다. Figure 7은 배와 트럭 두 가지 샘플에 대해 디코더 블록의 수 증가에 따른 복원 결과를 가시화 한 결과다. 배의 경우 초기 디코더 블록이 배의 전체적 윤곽을 잡아냈으며, 뒤쪽으로 갈수록 세부적인 모양이 점진적으로 개선되는 것을 확인 할 수 있다. 트럭의 경우도 초기에는 배와 비행기의 중간 모습을 표현하다가, 후반으로 갈수록 트럭의 모습으로 세부 형상이 잡혀 가는 것을 확인할 수 있다. 이러한 결과는 본 연구가 제안하는 디코더 내 디코더 블록들이 물체 형상을 전역적 모습에서 지역적 특징 표현으로 점진적으로 개선하고 있음을 확인할 수 있는 결과다.

4.5 삭마 연구(ablation study)

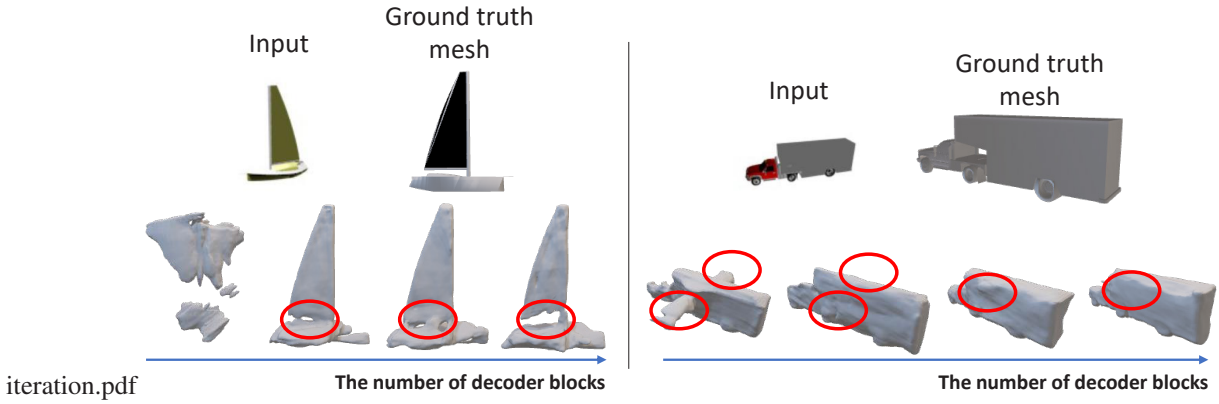
본 연구는 제안하는 네트워크를 구성하는 주요 요소의 효과를 분석하기위해 각 요소를 배제하고 성능을 측정하는 삭마 연구를 진행하였다. 삭마 연구에 사용된 네트워크는 아래와 같다.

- CBN_{concat}: 디코더 블록마다 다른 지역적 특징을 넣는 것 대신, 각 인코더 블록의 출력들을 모두 결합(concatenation) 해서 조건부 배치 정규화 조건으로 사용하는 네트워크다.
- CBN_{single}: 각 디코더 블록에서 하나의 지역적 특징만을 사용해서 조건부 배치 정규화를 수행하는 네트워크로, 전역적 특징(Conv. block6의 출력)은 사용하지 않고 다섯 개 블록만 사용하였다.



(a) Chamfer distance (↓) (b) IoU (↑) (c) Normal consistency (↑)

Figure 6: Scores changes of three evaluation metrics according to the number of decoder blocks



iteration.pdf

Figure 7: Reconstruction visual results that depend on the network's iterations

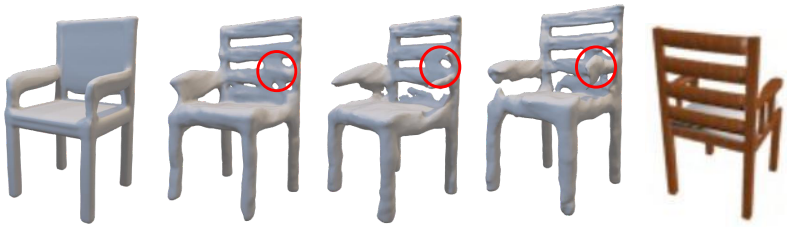
Table 3: Results of for different network architectures

	Chamfer distance	IoU	Normal consistency
ONet	0.215	0.571	0.834
CBN_{concat}	0.200	0.578	0.831
CBN_{single}	0.198	0.582	0.832
Ours	**0.180**	**0.601**	**0.838**

Table 3는 ONet과 Ours, 그리고 삭마 연구를 위해 사용된 두 개 모델의 테스트 세트에 대한 점수를 보여준다. Figure 8은 의자 샘플에 대한 네 가지 모델의 복원 결과를 가시화 한 결과다. CBN_{concat}는 다양한 수준의 지역적 특징과 전역적 특징을 함께 사용함으로서 ONet 대비 Chamfer distance와 IoU를 개선시키는 것을 확인 할 수 있다. CBN_{single} 또한 ONet 대비 높은 성능을 보여주었으며, CBN_{concat}와 비교해서는 IoU에서는 좋은 모습을, 그리고 Chafer distance와 normal consistency에서는 대등한 성능을 보여주었다. CBN_{single}는 CBN_{concat} 보다 지역적 특징에 더 집중을 함으로서 세부모습을 잘 잡아내서 이러한 결과가 나오는 것으로 판단된다. 두 변형 모두 normal consistency에서는 ONet 대비 낮은 성능을 보여주었다. 이는 두 모델이 지역적 형상을 잘 잡아내는 반면, 지역적 특징 표현 과정에서 표면 표현의 매끄러움이 깨는 경우가 있어 normal consitency가 낮게 나오는 것으로 판단된다. 또한, Figure 8에서 볼 수 있듯, 사진에서 가려진 부분에 대해 ONet은 잘 잡아내지만 CBN_{single}와 CBN_{concat}는 복원에 실패하는 모습을 확인 할 수 있다. 이는 지역적 특징에 집중함으로서 발생하는 단점이다. 하지만, 본 연구가 제안하는 점진적 점유 네트워크는 전역적 특징과 지역적 특징을 적절히 결합해서 사용함으로서, 모든 평가 지표에서 높은 성능을 달성함을 확인 할 수 있다.

5. 결론 및 향후 연구

본 논문은 물체의 전체적 형상 및 세부적 특징까지 복원 할 수 있는 표현력을 가진 음함수 표현 3차원 형상 복원 네트워크인 점진적 점유 네트워크 구조를 제안하였다. 제안하는 네트워크는 이미지 전체의 정보를 담고 있는 전역적 특징만을 사용하는 기존의 점유 네트워크(ONet)와 달리, 수용 영역의 크기에 따라 다양한 수준의 이미지 정보를 담고 있는 단계별 지역적 특징맵을 인코더를 통해 생성하고 사용한다. 그리고, 입력 포인트의 위치에 대응하는 지역적 특징을 추출하여 디코더 블록 내 조건부 배치 정규화를 위한 조건으로 사용한다. 디코더 내 블록들은 서로 다른 수준의 지역적 특징을 조건으로 사용함으로서 각각의 디코더 블록이 서로 다른 수준의 형상 표현력을 내포하도록 하였다. 그리고 포인트 특징이 디코더 블록들을 순차적으로 통과하며 특징을 갱신함으



(a) ONet (b) CBN_{concat} (c) CBN_{single} (d) Ours (e) Input

Figure 8: Visualization of reconstruction results using four different occupancy networks

로서 물체의 전역적 형상으로 부터 세부적 형상이 점진적으로 포인트 특징에 반영되도록 하였다. 그 결과, 기존의 점유 네트워크(ONet)보다 다양한 검증 지표에서 좋은 성능을 보였으며, 가시화 결과의 시각적 비교를 통해 ONet이 놓친 물체의 세부적인 형상을 잘 복원해내는 것을 확인 할 수 있었다. 또한, 지역적 정보와 전역적 정보를 동시에 사용하는 기존 연구(DISN)와 검증 지표상 대등한 성능을 달성하였으며, DISN이 복원에 실패한 물체의 얇은 부분이나 이미지에서 가려진 부분을 제안하는 네트워크가 잘 복원해 내는 결과를 확인할 수 있다. 이러한 결과는 본 연구가 제안하는 점진적 복원 네트워크의 유용성을 검증하는 결과다.

본 연구의 점진적 점유 네트워크는 물체 세부 형상은 ONet보다 잘 복원했지만, 전역적 형상의 표현력 및 완만한 표면의 표현은 ONet대비 좋지 못한 모습을 보였다. 향후 연구에서는 세부 형상 성능은 유지하면서 전역적 형상의 표현력을 높일 수 있는 네트워크 구조를 연구하고자 한다. 또한, 가상 환경에서 만들어진 데이터 세트를 넘어 실제 사진에 대한 제안하는 네트워크의 효능성을 탐구하고 실제 환경에서 활용할 수 있는 어플리케이션을 개발하고자 한다.

감사의 글

이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2021R1I1A3048263)

References

[1] A. M. Andrew, “Multiple view geometry in computer vision,” *Kybernetes*, 2001.

[2] H. Fan, H. Su, and L. J. Guibas, “A point set generation network for 3d object reconstruction from a single image,” in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 605–613.

[3] C. Häne, S. Tulsiani, and J. Malik, “Hierarchical surface prediction for 3d object reconstruction,” in *International Conference on 3D Vision (3DV)*, 2017, pp. 412–420.

[4] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, “Pixel2mesh: Generating 3d mesh models from single rgb images,” in *Proc. of the European Conference on Computer Vision (ECCV)*, 2018, pp. 52–67.

[5] Y.-P. Xiao, Y.-K. Lai, F.-L. Zhang, C. Li, and L. Gao, “A survey on deep geometry learning: From a representation perspective,” *Computational Visual Media*, vol. 6, no. 2, pp. 113–133, 2020.

[6] 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 *European conference on computer vision*, 2016, pp. 628–644.

[7] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas, “Learning representations and generative models for 3d point clouds,” in *International conference on machine learning*. PMLR, 2018, pp. 40–49.

[8] G. Yang, X. Huang, Z. Hao, M.-Y. Liu, S. Belongie, and B. Hariharan, “Pointflow: 3d point cloud generation with continuous normalizing flows,” in *Proc. of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4541–4550.

[9] W. Yifan, F. Serena, S. Wu, C. Öztireli, and O. Sorkine-Hornung, “Differentiable surface splatting for point-based geometry processing,” *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–14, 2019.

[10] H. Kato, Y. Ushiku, and T. Harada, “Neural 3d mesh renderer,” in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3907–3916.

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

[12] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, and H. Su, "ShapeNet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[13] Q. Xu, W. Wang, D. Ceylan, R. Mech, and U. Neumann, "DISN: Deep implicit surface network for high-quality single-view 3d reconstruction," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[14] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry, "A papier-mâché approach to learning 3d surface generation," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 216–224.

[15] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs," in *Proc. of the IEEE International Conference on Computer Vision*, 2017, pp. 2088–2096.

[16] G. Riegler, A. Osman Ulusoy, and A. Geiger, "Octnet: Learning deep 3d representations at high resolutions," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3577–3586.

[17] J. Wu, C. Zhang, X. Zhang, Z. Zhang, W. T. Freeman, and J. B. Tenenbaum, "Learning shape priors for single-view 3d completion and reconstruction," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2018, pp. 646–662.

[18] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *ACM siggraph computer graphics*, vol. 21, no. 4, pp. 163–169, 1987.

[19] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proc. of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2304–2314.

[20] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 165–174.

[21] K. Genova, F. Cole, D. Vlasic, A. Sarna, W. T. Freeman, and T. Funkhouser, "Learning shape templates with structured implicit functions," in *Proc. of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7154–7164.

〈저자소개〉

김용규
- 2019년 한국기술교육대학교 컴퓨터공학부 (공학사)
- 2021년 한국기술교육대학원 컴퓨터공학과 (공학석사)
- 관심분야: 3차원 형상 복원, 인공지능
- https://orcid.org/0000-0001-7038-8715

김덕수
- 2008년 성균관대학교 정보통신고학부 (공학사)
- 2014년 KAIST 전산학과 (공학박사)
- 2014년-2018년 한국과학기술정보연구원 선임연구원
- 2018년-현재 한국기술교육대학교 조교수
- 관심분야: 고성능컴퓨팅, 그래픽스/가시화, 인공지능 등
- https://orcid.org/0000-0002-9075-3983