|  |
| --- |
| **RGB 이미지로부터 씬 그래프 기반 상호작용을 위한 3D 공간 재구성** |
|  |
| **Scene Graph based Interactive 3D Scene Reconstruction  from RGB Sequences** |
|  |
| **요 약**  한국정보과학회는 정보과학에 관한 기술을 발전, 보급시키고 회원상호간의 친목을 도모하기 위하여 1973년 3월 3일에 설립되었으며, 정보통신부에 '사단법인 한국정보과학회'로 등록되었다. 학회의 주요 활동은 1) 컴퓨터 기술 및 이론에 관한 새로운 연구결과를 발표하는 기회를 제공하고, 2) 국내의 컴퓨터 관련 기술 개발에 참여 하며, 3) 국제적 학술 교류 및 협력 증진을 도모 하고, 4) 회원 상호간의 친목을 증진시키는 것이다. | | |

**1. 서 론**

최근 증강 현실(AR)과 가상 현실(VR)에 대한 관심이 높아지면서, 공간을 매개로 하는 몰입형 3D 어플리케이션의 수요가 증가하고 있다. 사용자에게 현실감 있는 컨텐츠를 제공하기 위해서는 실제 환경의 기하학적 및 의미론적 정보를 활용하는 것이 중요하다. 장면 그래프(scene graph)는 객체들의 의미론적 및 관계 정보를 결합한 표현 수단으로, 복잡한 3D 정보를 텍스트로 압축하고 적은 리소스로 사용자와 상호작용할 수 있는 직관적인 시각적 표현을 제공한다. 이러한 간결한 구조는 이미지 생성(Image generation), 조작(manipulation), 시각적 질문 답변(visual question answering), 3D 장면 생성(3D scene generation)과 같은 다양한 로봇 공학 및 컴퓨터 비전 작업에 활용되어 광범위한 적용 가능성을 입증하였다.

적은 리소스가 요구되는 AR/VR 환경에서 장면 그래프를 활용하기 위한 선행 연구들은 꾸준히 진행되어 왔다. 연속적인 RGB 및 RGB-D 시퀀스에서 3D 장면 그래프를 생성하는 모델들은 SLAM을 기반으로 점진적으로 3D 공간을 재구성하며 장면 그래프를 예측한다 [1,2]. 이러한 연구는 사용자의 전체 공간을 그래프 수준으로 압축시킴으로써, 공간 데이터를 효율적으로 관리할 수 있도록 하였다. 또 다른 관련 연구로는 장면 그래프를 입력으로 넣어 생성형 모델을 통해 3D 장면을 생성하는 기술들이 제안되었다 [3]. 이들은 Variational Auto-Encoder (VAE) 및 Diffusion을 기반으로 가상 공간을 합성함으로써, 사용자에게 몰입감 있는 경험을 제공하였다. 하지만 이러한 선행 연구들은 개별 모듈 단위로 연구가 진행되어 왔기 때문에, RGB 이미지를 입력으로 받아 그래프를 매개로 3D 가상 공간을 생성해내는 전체 파이프라인에 대한 연구는 부족하였다.

본 논문에서는 사용자가 바라보는 실제 세계를 그래프를 활용하여 3D 재구성 하는 전체 파이프라인을 제안한다. 또한 입력한 프롬프트에 따라 가상 공간을 맞춤형으로 변경할 수 있는 후처리 모듈을 기반으로 사용자와 상호작용할 수 있도록 하였다.

**2. 방법론**

본 연구에서 제안하는 시스템은 RGB 이미지를 입력으로 받아 장면 그래프를 생성한 후, VAE를 활용하여 3D 가상 공간을 단계로 구성되어 있다. 이후 후처리 모듈을 통해 입력 텍스트에 따라 가상 공간에 텍스쳐를 입힐 수 있도록 설계하였다.

**2.1. 3D 장면 그래프 생성**

장면 그래프를 추출하기 위해 본 논문에서는 monoSSG [2]를 활용한다. monoSSG는 두 단계로 구성되어 있으며, 점진적 객체 추정(Incremental Entity Estimation, IEE)와 의미론적 장면 그래프 예측(Semantic Scene Graph Prediction) 과정으로 이루어진다.

점진적 객체 추정은 RGB sequence로부터 객체들 간의 이웃그래프를 추정하는 단계이다. 먼저 ORB-SLAM3 모델을 통해 포인트 클라우드를 추출하며, 이를 라벨 연결과 융합을 통해 각 키프레임마다 각 픽셀의 객체를 추정하여 라벨링한다. 이후 각 키프레임에서의 객체 라벨 일관성을 유지할 수 있도록 연결과 융합 과정을 거쳐 3D 객체들 간에 특정 범위 내 이웃 그래프를 추출한다.

의미론적 장면 그래프 예측 단계에서는 이웃 그래프를 활용하여 장면 그래프를 생성하는 것을 목표로 한다. 이는 message passing 기법을 활용하여 초기 정점 특징과 간선 특징을 업데이트함으로써 3D 의미론적 장면 그래프를 학습한다. 각 키포인트 시점에서 이미지 특징은 정점의 관심영역(Region-of-Interest, ROI)에 대한 이미지 복호화 네트워크에 의해 추출되며, 기하학적 특징은 PointNet [4] 인코더에 의해 추출된다. 이후 GNN에 의해 업데이트되어 최종적으로 전체 공간에 대한 장면 그래프를 구축한다. 이러한 장면 그래프는 물체에 대한 Bounding box (bbox) 좌표와 물체 및 관계의 의미론적 레이블 정보를 가진다.

**2.2. 3D 가상 공간 생성**

구축된 장면 그래프를 기반으로 가상공간을 생성하기 위해 본 연구는 그래프 VAE 기반 생성 모델을 활용한다. 3D 장면 생성을 위해 그래프는 각 정점의 객체 클래스, 형태, 레이아웃 정보를 필요로 한다. 따라서 우리는 장면 그래프를 확장하여, 사전 학습된 DeepSDF [5] 모델을 통해 정점에 형태(shape) 특성을 추가해주었다.

본 모듈에서는 확장된 그래프를 받아 DeepGCN [6]을 기반으로 한 VAE를 통해 3D 장면을 생성한다. 구체적으로 제안한 인코더는 shape 특징과 bbox를 잘 융합하기 위한 Fused layer와, 전체 3D 장면의 일관성을 유지하기 위해 bbox를 중점으로 한 bbox-dominant layer로 구성되어 있다. 물체에 대한 shape 특징과 bbox 특징은 밀접하게 연관되어 있으므로, Fused layer를 통해 관련된 두 특징을 효과적으로 융합할 수 있도록 하였다. 또한 전체 3D 장면을 일관성 있게 구성하는데 있어 bbox의 위치와 방향이 물체 shape보다 더 지배적인 중요도를 가지기 때문에, bbox-dominant layer에서는 fused layer의 output과 bbox 정보를 입력으로 받아 물체들 간 관계를 학습한다. 이후 인코더의 마지막 단계에서 MLP를 통해 잠재 벡터 z의 평균과 표준편차를 구한다. 정규분포로 추출된 z는 디코더에 입력되며, 디코더는 인코더와 동일한 GCN으로 구성된다. 잠재 벡터 z는 bbox 디코더와 shape 디코더를 각각 통과한 후 통합하여 전체 3D 공간을 구성한다. 또한 제안한 시스템은 manipulator를 통해 생성된 3D 장면을 조작할 수 있다. 먼저, 3D 장면 조작을 포함하여 잠재 벡터 z를 수정한 후, 조작한 객체와 관계를 받아 z를 업데이트 한다. 더불어 zero padding과 gaussian noise sampling 등의 기법을 통해 3D 장면의 전체적인 일관성을 유지하며 장면을 조작할 수 있도록 하였다.

**2.3. 3D 장면 그래프 생성**

사용자와 상호작용 가능한 3D 가상공간을 생성하기 위해, 우리는 Scenetex [7] 모델을 이용하여 3D 메쉬를 후처리한다. Scenetex는 depth-to-image diffusion prior를 이용하여 고해상도의 텍스쳐를 생성하는 모듈이다. 이는 implicit texture field를 통해 object 형태를 암호화하며 cross-attention texture decoder를 통해 객체에 일관된 텍스쳐를 입히도록 보장한다. 우리는 이러한 모듈을 3D 메쉬에 적용함으로써, AR/VR 환경에서 원격 집들이, 가상 인테리어와 같은 어플리케이션에 활용할 수 있도록 하였다.

**3. 실험**

**3.1. 데이터셋**

본 연구에서는 3D 장면 생성 평가를 위해 SG-Front [8] 데이터셋을 활용하였다. SG-Front는 실내 공간의 구조와 가구 배치를 포함한 대규모 3D 장면 데이터셋인 3D-Front [9]에 장면 그래프 정보를 레이블링하여 확장한 데이터셋이다. 이는 45,000개의 객체 인스턴스와 세 종류의 장면에서 15가지 관계를 제공한다.

**3.2. 평가 지표**

3D 장면 생성 결과를 평가하기 위해 우리는 베이스라인 모델인 Graph-to-3D과 동일한 평가 지표를 사용하였다. 이는 레이아웃 관계를 기반으로 생성된 3D 장면이 입력 장면 그래프와 비교하여 0.3 IOU 값을 경계 기준으로 일관되게 정렬되는지 평가한다.

**3.3. 결과 분석**

전체적인 3D 장면을 평가하기 위해, 우리는 두 모델 실내, 벽, 바닥재, 인테리어 디자인이(가) 표시된 사진

자동 생성된 설명건물, 벽, 바닥재, 실내이(가) 표시된 사진

자동 생성된 설명모두 shape 특징을 DeepSDF로 생성하여 비교 평가를 수행하였다. 표 1은 3D 장면 생성에 대해 베이스라인 모델인 graph-to-3d과 제안한 모델의 성능을 비교한 결과를 보여준다. 본 시스템은 기존 베이스라인 모델보다 모든 지표에서 동등하거나 능가하는 성능을 보여주었다. 특히 ‘Close by’ 지표와 ‘Symmetrical’ 지표에서 각각 5%, 10%의 성능 향상을 보여주며, 제안한 모듈이 3D 가상 공간 생성에 적합하다는 것을 입증한다. 또한 이는 bbox를 중점으로 그래프 특징을 학습하는 우리의 전략이 유효함을 시사한다.



또한 addition 지표의 경우 Close by가 0.73, Symmetrical 0.55로 기존보다 10%, 8% 높은 성능을 보였으며, 위치 관계를 조작하는 relation change의 경우 특히 Symmetrical 지표에서 16%에 달하는 성능 향상을 보였다. 이는 제안한 인코더를 통해 높은 품질의 잠재 벡터를 추출해 냄으로써, 기존의 모델보다 3D 장면을 더 잘 생성하고 조작할 수 있음을 시사한다.

**4. 결론**

본 연구에서는 RGB 시퀀스를 받아 구축한 장면 그래프를 기반으로 3D 공간을 생성하는 프레임워크를 소개한다. 제안한 통합 모델을 통해 현실과 가상을 효율적으로 연결할 수 있으며, 본 인코더-디코더 아키텍처를 통해 장면 그래프와 정확하게 일치하는 가상 공간을 생성해낸다. 실험 결과, 제안한 모델은 까다로운 장면 그래프 제약 조건에서도 기존 모델보다 뛰어난 생성 및 조작 성능을 발휘한다. 또한 그림 1은 Scenetex 모듈을 통해 다른 텍스트 프롬프트(Bohemian style living room, Korean style living room, Modern style living room with a car poster)를 입력하였을 때의 장면 조작 결과를 보여준다. 이처럼 사용자의 프롬프트를 기반으로 맞춤형 가상 공간을 생성할 수 있도록 함으로써, 상호작용 가능한 AR/VR 환경에 적용할 수 있는 가능성을 확인할 수 있다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

표 1. 씬 생성(Gen.), 추가(Add.), 변경(Chg.)에 따른 정량 성능 비교.

그림 1. 사용자의 프롬프트에 따른 3D 장면 재구성

**5. 참고문헌**

[1] Wu, Shun-Cheng, et al. "Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

[2] Wu, Shun-Cheng, et al. "Incremental 3d semantic scene graph prediction from rgb sequences." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[3] Dhamo, Helisa, et al. "Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[4] Charles, R. Qi, et al. "PointNet: Deep learning on point sets for 3D classification and segmentation." *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2017.

[5] Park, Jeong Joon, et al. "Deepsdf: Learning continuous signed distance functions for shape representation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[6] Li, Guohao, et al. "Deepgcns: Can gcns go as deep as cnns?." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

[7] Chen, Dave Zhenyu, et al. "Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

[8] Zhai, Guangyao, et al. "Commonscenes: Generating commonsense 3d indoor scenes with scene graphs." *Advances in Neural Information Processing Systems* 36. 2024.

[9] Fu, Huan, et al. "3d-front: 3d furnished rooms with layouts and semantics." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.