특집논문 (Special Paper)
방송공학회논문지 제29권 제6호, 2024년 11월 (JBE Vol.29, No.6, November 2024)
https://doi.org/10.5909/JBE.2024.29.6.931
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

ViT 기반 깊이 추정과 MVS 기반 깊이 정보 최적화를 통한 고품질 다시점 깊이 정보 생성 기법

박 준 형[a], 정 종 범[b], 최 재 열[c], 김 영 규[a], 류 은 석[a)‡]

Multi-View Depth Generation Method Using ViT-Based Depth Estimation and MVS-Based Depth Optimization

Jun-Hyeong Park[a], Jong-Beom Jeong[b], Jaeyeol Choi[c], Yeong Gyu Kim[a], and Eun-Seok Ryu[a)‡]

요 약

다시점 실사 영상의 3차원 재구성을 위해서는 다중 시점 간의 일관성을 유지한 고품질의 깊이 맵을 추정하는 것이 필수적이다. 그러나 이는 수많은 기술적 난제를 동반한다. 기존의 다중 시점 스테레오 기법 및 immersive video depth estimation (IVDE)와 같은 다시점 이미지 기반 깊이 추정 방식들은 장면 특성이나 캡처 환경에 따라 정확도가 저하되는 한계를 지니고 있으며, 특히 낮은 텍스처의 데이터에서는 성능 저하가 두드러진다. 본 연구는 vision transformer (ViT) 기반 깊이 추정 모델을 활용해 단일 이미지에서 깊이 정보를 추출하고, 이를 다중 시점 스테레오로 생성된 깊이 정보와 결합하여 구조적 일관성을 유지한 고품질 다시점 깊이 맵 시퀀스를 생성하는 접근법을 제안한다. 제안된 접근법은 단일 이미지에서 추정된 깊이 정보를 조정해 다중 시점에서 일관된 깊이 정보를 생성함으로써, 복잡한 장면에서도 우수한 3차원 재구성 성능을 입증하였다.

Abstract

Achieving high-quality depth maps with multi-view consistency is crucial for 3D reconstruction of real-world scenes. However, this involves numerous technical challenges. Existing multi-view image-based depth estimation methods, such as multi-view stereo (MVS) and immersive video depth estimation (IVDE), have limitations in accuracy depending on scene characteristics or capture environments, with particularly noticeable performance degradation in low-texture data. This study proposes an approach that utilizes a ViT-based depth estimation model to extract depth information from single images and combine it with depth data generated by MVS to produce structurally consistent and high-quality multi-view depth map sequences. The proposed approach meticulously adjusts the depth information inferred from single images to generate consistent depth information across multiple viewpoints, demonstrating superior 3D reconstruction performance even in complex scenes.

Keyword : Virtual reality, Depth map, 6DoF, 3D reconstruction, 3D Gaussian Splatting

Ⅰ. 서 론

가상현실 (VR) 과 증강현실 (AR) 기술의 급속한 발전은 고품질의 몰입형 비디오 애플리케이션에 대한 수요를 증가시키고 있다. 이러한 애플리케이션이 사용자에게 더욱 몰입감 있는 경험을 제공하기 위해서는 6 자유도 (6 degrees of freedom, 6DoF) 환경을 지원해야 한다. 6DoF는 사용자가 가상 공간에서 자유롭게 이동하며 회전하거나 위치를 바꿀 수 있는 기능으로, 몰입형 비디오와 같은 애플리케이션에서 현실감 있는 상호작용을 가능하게 한다. 이를 위해 다양한 시점에서 취득한 비디오 데이터를 기반으로 3차원 공간을 재구성하고, 각 시점에 맞는 뷰를 합성하는 기술이 필수적이다.

Moving Picture Experts Group (MPEG)은 디지털 비디오와 오디오의 압축 표준을 개발하는 국제 표준화 기구로서, 고품질의 멀티미디어 콘텐츠 전송과 저장을 위해 다양한 표준을 제정해왔다. 그중에서도 MPEG Immersive Video (MIV)는 다수 시점에서 촬영된 비디오 데이터를 압축하여 전송하고, 수신 측에서 이를 기반으로 자유시점 뷰를 생성할 수 있도록 하는 기술로, 사용자에게 높은 몰입감과 현실감을 제공하는 6DoF 환경을 구현한다. MIV는 각 시점에 대한 정확한 텍스처 정보, 깊이 정보, 그리고 카메라 포즈를 요구하며, 이러한 데이터를 바탕으로 3차원 공간 내의 모든 시점의 뷰를 생성할 수 있다. 그러나, 고품질의 6DoF 콘텐츠를 생성하기 위해서는 정확하고 일관된 깊이 정보의 제공이 필수적이다. 특히, 실사 기반 다시점 이미지 시퀀스를 이용하여 MIV를 통한 비디오 재구성을 구현할 때, 깊이 정보는 매우 중요한 역할을 한다. 깊이 정보는 3차원 구조의 직접적인 기하학적 신호 (geometry cue)로서, 이를 통해 보다 정밀하고 일관된 3차원 재구성을 실현할 수 있다. 따라서, 깊이 정보의 정확도를 높이고, 이를 효율적으로 추정할 수 있는 방법의 개발이 필요하다.

한편, 3차원 재구성을 위한 또 다른 방식으로 신경망 기반 표현 기법이 주목받고 있다. 특히, Neural radiance fields (NeRF)[1]는 다중 시점 이미지로부터 직접 3차원 장면을 재구성하는 획기적인 기술로, 주어진 카메라 위치에서 광선 (ray) 위 3차원 포인트를 샘플링하여, 각 지점에서의 색상과 밀도를 예측하는 방식으로 동작한다. NeRF는 다층 퍼셉트론 (multi-layer perceptron, MLP)을 활용해 3차원 좌표와 2차원 시점 방향으로부터 해당 지점의 광량 분포와 색상을 예측하며, 이를 통해 포토리얼리스틱한 새로운 시점 합성이 가능함을 입증하였다. 이 모델은 방대한 양의 연산을 요구하는데, 이는 각 픽셀에 대해 수많은 샘플링과 연산이 필요하기 때문이다. 이러한 연산 복잡성으로 인해 NeRF는 실시간 응용에는 적합하지 않으며, 대규모 데이터셋에서의 학습과 추론에 상당한 시간이 소요된다. 결과적으로 NeRF는 고해상도의 3차원 장면을 재구성할 수 있는 능력을 갖추고 있음에도 불구하고, 실시간 처리 속도가 요구되는 응용에서는 한계가 있다.

이러한 문제를 해결하기 위해 최근 연구에서는 3차원 장면을 로컬화하여 명시적인 (explicit) 데이터를 함께 사용하여 모델링하는 새로운 기술들이 등장하고 있다[2-5]. 이는 장면을 소규모의 지역으로 분할하여, 지역별로 독립적으로 최적화함으로써 연산량을 줄이고, 처리 속도를 크게 향상시키는 접근 방식이다. 이 중에서 3D gaussian splatting (3DGS)[6-7]은 다중 시점에서 획득한 광학 정보를 바탕으로 독립적인 스플랫 (splat)을 최적화하는 방식으로 작동하며, 기존 NeRF에 비해 연산 속도와 처리 효율이 크게 개선되었다. 3DGS는 높은 품질의 재구성 결과를 제공하면서도 빠

a) 성균관대학교 실감미디어공학과(Department of Immersive Media Engineering, Sungkyunkwan University)
b) 성균관대학교 컴퓨터교육학과(Department of Computer Science Education, Sungkyunkwan University)
c) 성균관대학교 인공지능융합학과(Department of Applied Artificial Intelligence, Sungkyunkwan University)
‡ Corresponding Author : 류은석(Eun-Seok Ryu)
E-mail: esryu@skku.edu
Tel: +82-2-760-0677
ORCID: http://orcid.org/0000-0003-4894-6105
※이 논문의 연구 결과 중 일부는 한국방송·미디어공학회 2024년 하계학술대회에서 발표한 바 있음.
※This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) support program(IITP-2024-RS-2023-00259497) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).
※This research was supported by Global Standardization and Commercialization of Copyright Technology Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (Project Name: Development of International Standards for CT XR Content Copyright Protection Technologies, Project Number: RS-2024-00439789, Contribution Rate: 50%)
· Manuscript September 5, 2024; Revised October 16, 2024; Accepted October 17, 2024.

른 재구성 속도와 실시간 렌더링을 지원하는 장점이 있다. 그러나 이러한 로컬화 방식은 장면의 글로벌 구조를 고려하지 않기 때문에 특정 장면에서 로컬 최적화에 갇혀 오버피팅 문제가 발생할 수 있다. 예를 들어, 학습에 사용된 단일 시점에 과적합되어 플로팅 아티팩트와 같은 부작용이 나타날 수 있다. 이러한 문제를 해결하기 위해서는 장면 학습 과정에서 깊이 맵과 같은 추가적인 보조 정보의 사용이 필요하다. 깊이 맵은 3차원 장면의 글로벌 정보를 제공함으로써 로컬 최적화에 따른 문제를 보완할 수 있으며, 이를 통해 재구성 품질을 더욱 향상시킬 수 있다. 최근 깊이 맵을 활용하여 성능을 개선한 연구들이 보고되고 있으며, 이는 신경망 기반 3차원 재구성 기술의 발전에 중요한 기여를 하고 있다[8-10]. 따라서, 고품질의 3차원 재구성을 실현하기 위해서는 MIV와 신경망 기반 방식 모두에서 다시점 깊이 맵의 중요성이 강조된다. 다시점 깊이 맵은 3차원 장면의 정확성과 일관성을 확보하는 데 핵심적인 역할을 하며, 이를 효과적으로 생성하고 활용하는 방법이 필수적이다.

3차원 재구성을 위해 여러 가지 다중 시점 광학 정보를 사용한 깊이 맵 생성 기법들이 사용되고 있으며, 그중에서 immersive video depth estimation (IVDE)는 MIV 표준의 decoder-side depth estimation (DSDE) 프로파일을 지원하는 대표적인 기법이다. MIV DSDE 프로파일[11]은 비트스트림에 깊이 맵을 포함하지 않고, 디코더 측에서 깊이 맵을 추정하여 생성하는 방식을 채택하고 있다. 이 프로파일은 텍스처 정보와 메타데이터만을 전송하며, 디코더 측에서 필요한 깊이 정보를 추정하도록 한다. 이를 통해 인코더의 복잡성을 줄이고, 비트레이트 효율성과 렌더링 품질을 동시에 개선할 수 있다. 그림 1은 MIV 표준 DSDE 프로파일의 파이프라인을 나타낸다[12].

Structure from motion (SfM) 및 multi-view stereo (MVS) 기반의 깊이 맵 생성 기법도 널리 사용되고 있으며, 이는 COLMAP 소프트웨어로 구현할 수 있다. SfM은 다중 시점 이미지에서 카메라 포즈를 추정하고, 이를 바탕으로 3차원 공간상의 희소 포인트 클라우드를 생성하는 기법이

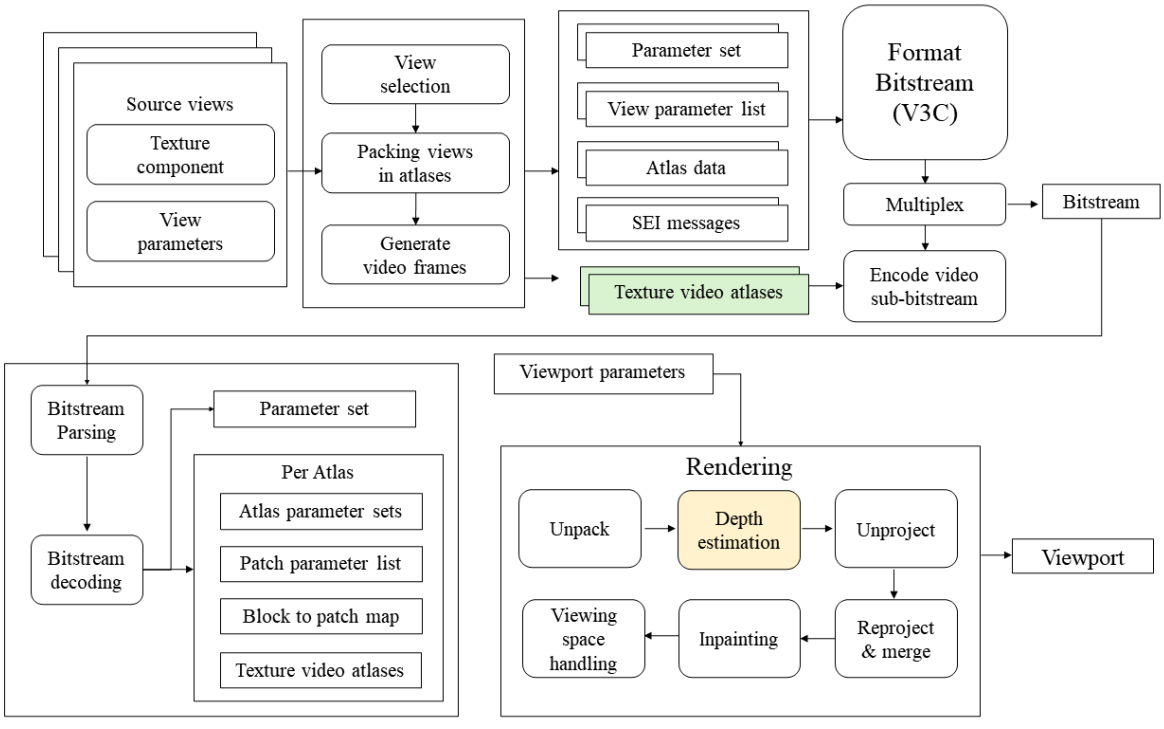


그림 1. MIV DSDE 모드 파이프라인
Fig. 1. MIV DSDE mode pipeline

며, MVS는 SfM에서 추정된 카메라 파라미터와 원본 이미지를 활용하여, 이미지들 간의 픽셀 단위 정합을 통해 밀집된 3차원 재구성을 수행하여 깊이 정보를 생성하는 기법이다[13-14]. 이 과정에서 얻어진 카메라 포즈는 신경망 기반 방식들이 요구하는 필수적인 입력으로 사용되며, 3DGS에서 gaussian splat의 위치 초기값을 지정하기 위해 사용된다.

그러나 이러한 방식들은 여러 한계를 가지고 있다. MVS는 다중 시점의 고품질 이미지가 필요하며, 시점 간의 충분한 중복성 (overlap)과 높은 해상도의 이미지를 요구한다. 그러나 실제 상황에서는 2차원 이미지의 시점 부족이나 저화질의 문제로 인해, 사용 가능한 3차원 특징점을 충분히 검출하지 못할 때가 많다. 이로 인해 생성된 포인트 클라우드가 희소해지면, 결과적으로 깊이 맵의 품질이 저하된다. IVDE 또한 빠르고 효율적인 깊이 맵 생성을 목표로 하지만, 품질 측면에서 한계가 있다. 이러한 문제로 인해 기존의 MVS와 IVDE 방식은 단독으로 사용될 경우 다양한 실제 상황과 데이터에 적용되지 않을 가능성이 있으며, 최종적으로 재구성된 장면의 품질을 떨어뜨리는 결과를 초래할 수 있다.

이러한 한계를 극복하기 위해, 본 논문에서는 다시점 깊이 맵을 생성하는 새로운 방식을 제안한다. Vision transformer (ViT) 기반의 깊이 추정 모델인 dense prediction transformer (DPT) 모델을 사용하여 단일 시점에서 깊이 맵을 생성한 후, 이를 MVS 기반 깊이 정보와 결합하여 구조적 일관성을 유지하는 고품질의 다시점 깊이 맵 시퀀스를 생성하는 방법을 제시한다. DPT 모델은 단일 시점 내에서 고품질의 깊이 정보를 추정할 수 있음에도 불구하고, 단일 시점에서 추정된 깊이 맵의 절대적인 깊이 값이 불명확하며 시점마다 깊이 척도의 스케일과 오프셋이 다르게 나타날 수 있다는 문제점을 내포한다. 다시 말해, 한 시점에서 추정된 깊이 맵이 다른 시점에서 추정된 깊이 맵과 비교할 때 동일한 객체에 대해 다른 깊이 값을 가질 수 있다. 이러한 문제는 다중 시점 간의 깊이 맵을 결합할 때 일관된 3차원 구조를 형성하는 데 어려움을 초래할 수 있다. 이를 해결하기 위해, 본 연구에서는 MVS 기반 깊이 정보를 사용하여 DPT에서 추정된 깊이 정보의 스케일과 오프셋을 조정하는 과정을 포함한다. MVS 기반 깊이 정보는 3차원 구조적 일관성을 가지고 있기에, 이를 기반으로 DPT에서 생성된 깊이 맵의 스케일과 오프셋을 정규화함으로써 시점 간의 깊이 맵 일관성을 확보할 수 있다.

제안된 방법의 검증을 위해, 제안 방법으로 생성된 최종 깊이 맵과 IVDE 기법으로 생성한 깊이 맵을 비교해 제안 기법이 입력 시점에 대한 고품질의 깊이 정보를 제공하는지 확인하였다. 더불어, 제안된 깊이 맵이 실제 3차원 구조성을 얼마나 잘 유지하는지를 검증하기 위해 제안 기법을 통해 최종 출력된 깊이 맵을 활용해 MIV를 통한 3차원 재구성을 진행하고, 새로운 시점의 뷰를 생성하여 최종 비디오 재구성의 품질을 평가하였다. 실험 결과, 제안된 방식으로 생성된 깊이 맵이 기존 기술 대비 더 높은 품질의 깊이 맵을 생성할 수 있으며 생성된 깊이 맵이 3차원 구조와 일관성을 유지함을 확인하였다.

본 논문은 고품질의 다시점 깊이 맵을 생성하는 새로운 방식을 제안하고, 이를 통해 생성된 깊이 맵의 3차원 재구성 기술에 대한 효용성을 검증하였다. 본 논문의 구성은 다음과 같다. 2장에서는 깊이 정보 추정과 뷰 합성 기술에 관한 관련 연구를 검토한다. 3장에서는 제안된 연구 방법론을 상세히 설명하며, 깊이 정보 생성 모델의 구현 과정을 다룬다. 4장에서는 실험 설정과 평가 방법을 설명하고, 제안된 접근법의 성능을 기존 방법들과 비교하여 분석한다. 마지막으로 5장에서는 실험 결과를 요약하고, 본 연구의 기여와 한계점, 그리고 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

1. Structure from Motion 및 Multi-View Stereo

SfM은 여러 시점에서 촬영된 2차원 이미지로부터 3차원 구조를 추정하고, 동시에 각 카메라의 위치와 방향을 계산하는 기술이다. 이 기법은 3차원 재구성에서 중요한 기초적 요소로, 여러 장면에서 카메라 위치뿐만 아니라 장면의 3차원 포인트를 추정하는 데 사용된다. SfM의 기본 원리는 여러 시점에서 촬영된 이미지 사이에서 동일한 특징점들을 비교하여, 이들이 3차원 공간에서 위치하는 곳을 추정하는 것이다. 이를 위해 각 이미지 간의 시차 (parallax)를 이용하여 깊이 정보를 계산하며, 이러한 정보는 동시에 카메라의

위치와 방향을 추정하는데 사용된다.

초기 SfM 연구는 제한된 데이터셋을 처리하는 데 그쳤으나, 최근 기술 발전으로 대규모 이미지 데이터도 처리할 수 있게 되었다. 특히 scale-invariant feature transform (SIFT), speeded-up robust features (SURF)와 같은 고도화된 특징점 추출 알고리즘이 등장하면서, 이미지에서 더욱 정교한 특징점 매칭이 가능해졌으며[15-17], 이로 인해 장면의 정확한 3차원 구조를 보다 신뢰성 있게 추정할 수 있게 되었다. 이러한 알고리즘들은 다양한 환경에서도 특징점을 잘 추출해 내며, 회전이나 스케일 변화에도 영향을 받지 않아, 여러 시점에서 촬영된 이미지 간의 일관된 특징점 매칭을 보장한다. 이를 통해 복잡한 장면의 3차원 재구성이 가능해졌다.

다중 시점 스테레오 (MVS)는 SfM을 통해 얻어진 카메라 파라미터를 바탕으로, 각 픽셀의 깊이 정보를 추정한다. 이 과정에서 여러 시점에서 관찰된 동일한 물체의 위치 변화를 분석하여, 이미지 내에서의 시차를 바탕으로 깊이 맵을 생성한다. MVS의 기본 흐름은 카메라 파라미터 추정, 시차 분석, 깊이 맵 생성의 세 가지 주요 단계로 나뉜다. 먼저, SfM을 통해 각 시점의 위치와 방향값을 포함한 카메라 파라미터가 추정되며, 이를 기반으로 각 시점의 이미지를 서로 정합시킨다. 이후, 인접한 이미지들 간의 픽셀 위치 변화를 시차 분석을 통해 비교하고, 이를 바탕으로 각 픽셀의 깊이 값을 추정하여 최종적으로 3차원 포인트 클라우드를 생성한다.

SfM과 MVS는 3차원 재구성을 위해 상호 보완적인 역할을 한다. SfM은 주로 카메라의 위치와 장면의 대략적인 3차원 구조를 추정하는 데 중점을 두고, MVS는 이를 기반으로 깊이 맵을 생성하여 세밀한 3차원 구조를 복원한다. 이러한 융합 기법은 고해상도의 포인트 클라우드와 깊이 맵을 생성하여, 복잡한 장면에서도 높은 정확도의 3차원 재구성을 가능하게 한다. COLMAP은 SfM과 MVS를 통합적으로 수행할 수 있는 대표적인 상용 소프트웨어로, 이미지 시퀀스에서 카메라 파라미터를 추정하고 다중 시점 간의 정합성을 유지하면서 3차원 포인트 클라우드를 생성할 수 있는 기능을 제공한다. 그림 2는 COLMAP 소프트웨어를 통해 MVS 기반 다시점 깊이 정보를 생성하는 파이프라인을 나타낸다.

MVS 기법은 다중 시점에서 구조적 일관성을 가진 깊이 맵을 생성할 수 있는 효과적인 방법이지만, 텍스처가 부족한 장면이나 이미지의 품질이 저하된 경우 깊이 맵의 정확도가 떨어질 수 있다. 예컨대 평평한 표면이나 특징점이 부족한 장면에서는 이미지 간의 대응점을 찾기 어려워 깊이 정보를 정밀하게 추정하지 못할 가능성이 크다. 또한, MVS와 SfM 과정은 높은 계산 비용을 수반하기 때문에 대규모 데이터셋을 처리하는 데 시간이 많이 소요될 수 있다.

2. 딥러닝 기반 단안 깊이 정보 추정 기법

광학 이미지를 입력으로 받아 해당 이미지의 깊이 정보를 추출하는 신경망 기반 기술은 초기에는 컨볼루션 신경망 (convolutional neural network, CNN) 기반으로 발전해 왔다[18-20]. CNN 기반 깊이 정보 추론 모델은 이미지 내에서 지역적 관계를 학습하고, 이를 통해 픽셀 단위의 상대적인 깊이 정보를 추정해왔다. 이 방식은 일반적으로 이미지의 해상도를 점진적으로 축소하면서 다양한 스케일의 특징을 추출하고, 이를 결합하여 깊이 맵을 생성하는 U-Net 구조를 따르는 방식으로 동작한다. 그러나 CNN은 리셉티브 필드 (receptive field)의 크기가 제한적이기 때문에, 모델의 깊

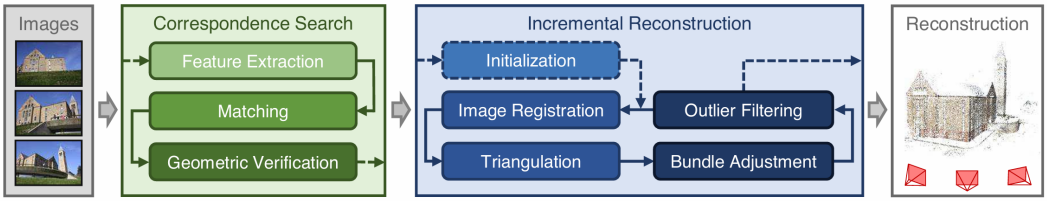


그림 2. Incremental SfM 알고리즘
Fig. 2. Incremental structure-from-motion algorithm

이가 얕을수록 이미지의 전역적인 정보를 처리하는 데 한계가 있었으며, 이러한 구조는 결과적으로 복잡한 장면에서 고해상도 정보를 잃는 문제로 이어졌다.

최근 트랜스포머 아키텍처가 컴퓨터 비전 분야에도 도입되면서 기존의 CNN 기반 모델의 한계를 극복하려는 연구가 활발히 진행되고 있다. 트랜스포머는 원래 자연어 처리에서 도입된 아키텍처로, 셀프 어텐션 (self-attention) 메커니즘을 통해 입력 데이터의 모든 요소 간의 관계를 학습한다. 셀프 어텐션은 각 픽셀 또는 이미지 패치 간의 상호 관계를 고려하여 전역적인 정보를 동시에 학습할 수 있기 때문에, CNN 기반의 지역적인 처리 한계를 극복할 수 있다. DPT는 이러한 트랜스포머 아키텍처를 사용하여 단일 이미지에서 깊이 정보를 추출하는 최신 모델 중 하나이다. DPT는 ViT를 백본 네트워크로 사용하여 입력 이미지를 패치 단위로 분할하고, 각 패치 간의 관계를 학습하는 방식으로 깊이 정보를 추론한다. ViT는 CNN과 달리 모든 단계에서 글로벌 리셉티브 필드를 가지며, 이를 통해 이미지의 전역적인 정보를 처리할 수 있다. 이로 인해 DPT는 이미지의 세부적인 특징과 전역적인 정보를 동시에 학습할 수 있다.

DPT의 핵심 구조는 토큰 (Token) 단위로 이미지를 처리하는 트랜스포머 인코더와, 이를 다시 이미지 형태로 재조립하여 깊이 맵을 생성하는 컨볼루션 디코더로 이루어진다. ViT 백본은 이미지의 패치들을 임베딩하여 각각을 토큰으로 변환하고, 트랜스포머 계층을 통해 각 토큰 간의 관계를 학습한다. 이 과정에서 Multi-Head Self-Attention 메커니즘을 사용하여 이미지 전반에 걸친 전역적인 특징을 학습하고, 모든 처리 단계에서 전역 정보를 고려한다. 트랜스포머는 CNN과 달리 단계적으로 리셉티브 필드를 확장할 필요 없이, 각 단계에서 전역 리셉티브 필드를 유지하며 이미지의 전역적 관계를 학습할 수 있다. DPT는 트랜스포머 인코더를 통해 얻어진 특징 맵을 다양한 해상도에서 결합하여 최종적인 깊이 맵을 생성한다. 이 모델은 깊이 맵을 예측할 때 단순히 픽셀 단위의 정보를 넘어서 이미지 전반에 걸친 전역적 관계를 고려함으로써 더욱 정밀하고 일관된 깊이 정보를 생성할 수 있다. ViT 기반 구조는 모든 단계에서 고해상도의 전역적 특징을 유지하기 때문에 CNN 기반 모델보다 더 뛰어난 성능을 발휘한다[21-23]. 그림 3은 DPT 모델이 이미지를 학습하고 깊이 정보를 추론하는 파이프라인을 나타낸다.

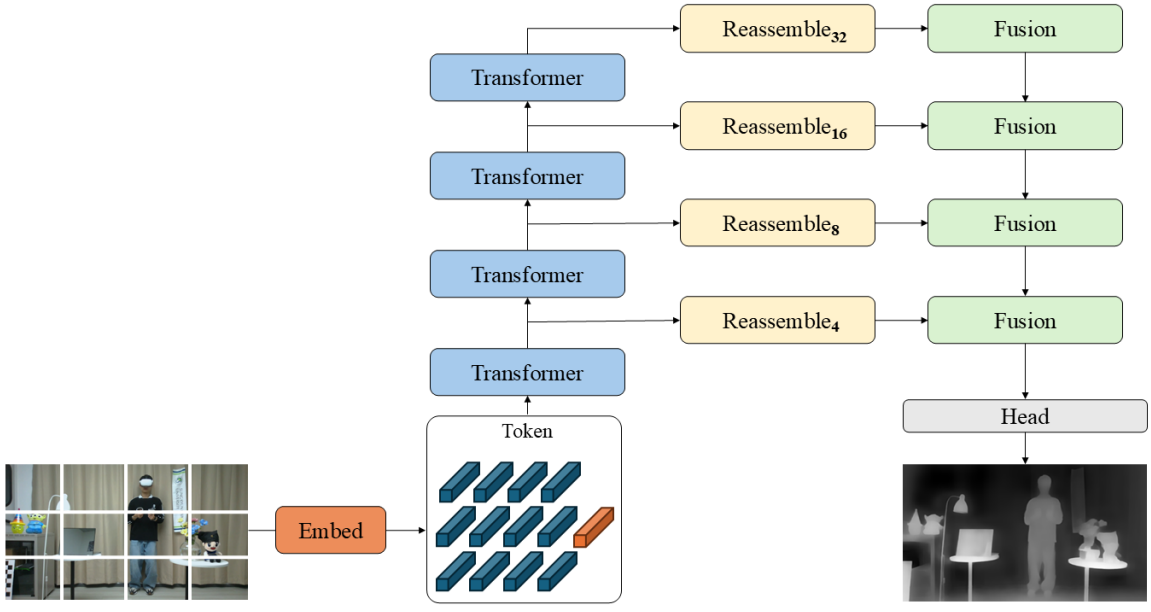


그림 3. Dense prediction transformer 모델 아키텍처
Fig. 3. Dense prediction transformer model architecture

DPT는 단일 시점에서 깊이 정보를 추출하는 데 있어 CNN 기반 모델보다 최대 28% 향상된 성능을 기록하였다. 특히 대규모 데이터셋을 사용한 학습에서 그 성능 차이가 더욱 두드러졌으며, 다양한 깊이 추정 작업에서 새로운 기준을 설정하였다. KITTI와 NYUv2와 같은 대규모 및 소규모 데이터셋에서도 DPT는 기존 모델 대비 높은 정확도를 기록하였다[24].

3. MPEG Immersive Video Standard (MIV)

MIV는 6DoF를 지원하는 몰입형 비디오 데이터를 효율적으로 압축하고 전송하기 위해 visual volumetric video-based coding (V3C) 프레임워크를 기반으로 설계된 차세대 표준이다. V3C는 다중 시점에서 취득한 볼륨 데이터를 효율적으로 압축하여 3차원 비디오 전송을 최적화하는 구조를 제공하며, 이러한 구조는 3차원 장면의 다양한 시점에서 촬영된 데이터를 압축하는 데 필요한 표준을 정의한다.

MIV에서 전송되는 비트스트림은 두 가지 주요 정보로 구성된다. 첫째는 각 카메라 시점에서 촬영된 텍스처 정보로, RGB 데이터를 통해 장면의 시각적 요소를 제공한다. 둘째는 각 픽셀에 대한 3차원 기하학적 깊이 값을 포함하는 깊이 정보로, 물체의 공간적 위치와 형태를 결정한다. 텍스처 정보는 고해상도로 처리될 경우 상당한 데이터를 차지하며, 깊이 정보는 물체의 기하학적 구조를 고정밀도로 표현해야 하므로 이 역시 데이터 양이 매우 크다. 특히, 깊이 정보는 사용자가 시점을 이동할 때 물체가 정확한 위치에서 렌더링 되도록 돕는 중요한 역할을 하며, 이 과정에서 모션 패럴럭스 (motion parallax)가 자연스럽게 구현되기 위해서는 깊이 정보의 정확도가 중요하다. 만약 깊이 정보가 부정확하거나 누락된다면, 시점 이동 시 물체가 왜곡되거나 부유하는 현상이 발생해 몰입감이 저하될 수 있다.

MIV의 DSDE 프로파일은 깊이 정보를 직접 전송하지 않고, 뷰 데이터와 압축된 메타데이터를 사용해 디코더 측에서 깊이 맵을 추정하는 방식이다. 이를 통해 깊이 정보 전송에 필요한 비트레이트를 줄이면서도, 디코딩 단계에서 고품질의 3차원 장면을 재구성할 수 있다. DSDE는 낮은 비트레이트 환경에서도 몰입형 비디오 콘텐츠의 3차원 구조를 정밀하게 유지할 수 있도록 돕는다. 이를 통해 몰입형 비디오 콘텐츠가 대역폭 제약이 있는 환경에서도 높은 품질로 제공될 수 있다.

MIV는 텍스처 정보와 깊이 정보 모두의 압축과 전송을 최적화하기 위한 정교한 기술들을 적용하여 몰입형 비디오 콘텐츠의 품질과 전송 효율성을 모두 확보하고 있다. MIV는 6DoF 비디오 환경에서 고품질의 몰입형 경험을 제공할 수 있으며, 특히 깊이 정보의 정확성과 데이터 전송의 효율성을 높여 사용자에게 자연스러운 3차원 경험을 제공하는 표준으로 자리 잡고 있다[25-27].

III. 다시점 깊이 맵 생성 기법

고품질의 3차원 재구성을 달성하기 위해 깊이 맵은 두 가지 주요 조건을 충족해야 한다. 첫째, 깊이 맵은 밀집된 (dense) 깊이 정보를 제공해야 한다. 둘째, 3차원 구조성을 정확히 반영해야 한다. 이 두 조건을 만족한 깊이 정보는 3차원 재구성의 품질을 향상시키며, 이를 통해 다양한 시점에서 일관된 깊이 정보를 기반으로 현실감 있는 경험을 제공할 수 있다.

먼저, 깊이 맵이 밀집된 정보를 제공해야 한다는 것은 이미지 내 모든 픽셀에 대해 깊이 값을 정확히 추정할 수 있어야 함을 의미한다. 만약 깊이 정보가 희소하면, 장면 전체에 대한 정보가 불충분해져 3차원 재구성 과정에서 빈틈이나 왜곡이 발생할 가능성이 크다. 밀집된 깊이 맵은 장면의 각 물체에 대해 픽셀 단위의 깊이 값을 제공함으로써, 물체의 세부적인 3차원 구조를 정확하게 재구성할 수 있는 기초 데이터를 제공한다. 이는 특히 텍스처가 부족한 영역이나 복잡한 장면에서도 높은 품질의 3차원 재구성을 가능하게 한다.

다음으로, 깊이 맵이 3차원 구조성을 반영해야 한다. 이는 다중 시점에서 관찰한 물체의 위치, 크기, 형태가 일관되게 표현되도록, 깊이 정보가 실제 물리적 구조를 정확히 반영해야 한다는 의미이다. 3차원 구조성이 보장된 깊이 맵은

다중 시점에서 일관된 정보를 제공하여, 물체가 3차원 공간에서 자연스럽게 연결된다. 만약 각 시점에서 생성된 깊이 정보가 물리적 구조를 일관되게 반영하지 못하면, 재구성 과정에서 물체가 왜곡되거나 부유하는 오류가 발생할 수 있다.

이 두 조건을 충족하기 위해, 본 논문은 단일 시점 깊이 맵과 다중 시점 스테레오 기반 깊이 정보를 결합하는 기법을 제안한다. 그림 4는 본 연구에서 제안하는 깊이 정보 생성 기법의 구조를 나타낸다.

1. 단안 깊이 정보 생성

단안 깊이 추정은 단일 이미지로부터 물체 간의 상대적 깊이 정보를 추정하는 기술로, 별도의 3차원 포인트 클라우드나 스테레오 이미지 없이도 깊이 정보를 얻을 수 있다는 이점이 있다. 이러한 기술은 하나의 이미지로부터 장면 내 물체 간의 거리 차이를 효과적으로 학습할 수 있어, 복잡한 장면에서도 유용하게 적용된다. 본 논문에서는 고품질의 깊이 맵을 추정하기 위해 사전 훈련된 DPT 모델을 사용한다.

DPT 모델은 카메라 정보가 없는 단일 이미지에서 고해상도의 밀집된 깊이 정보를 빠른 속도로 추론할 수 있으며, 이를 통해 3차원 재구성 시 세밀한 정보 손실을 최소화하고 복잡한 텍스처나 구조를 정확하게 반영할 수 있다. 또한, DPT는 다양한 이미지 유형에 대해 우수한 성능을 발휘하는데, 실내외 장면, 자연경관, 도시 환경 등 다양한 조건에서도 일관된 성능을 보여준다. 특히, 텍스처가 부족하거나 조명이 복잡한 환경에서도 정확한 깊이 정보를 추출할 수 있어, 기존의 다시점 깊이 정보 추정 기술 대비 다양한 실제 환경에서 깊이 추정이 가능하다.

DPT 모델을 통한 단안 깊이 맵 생성 과정은 크게 세 단계로 구성된다. 첫 번째 단계는 이미지 전처리 단계로, 입력 이미지를 DPT 모델이 처리할 수 있는 크기로 리사이즈하고 정규화하는 작업을 수행한다. 이러한 전처리 과정은 모델이 다양한 크기의 이미지를 일관되게 처리할 수 있도록 하며, 입력 데이터를 보다 효과적으로 학습할 수 있도록 돕는다.

두 번째 단계는 전역적 특징 추출이다. 트랜스포머 아키텍처의 인코더 (encoder)가 이미지의 전역적 특징을 추출하는 역할을 수행한다.

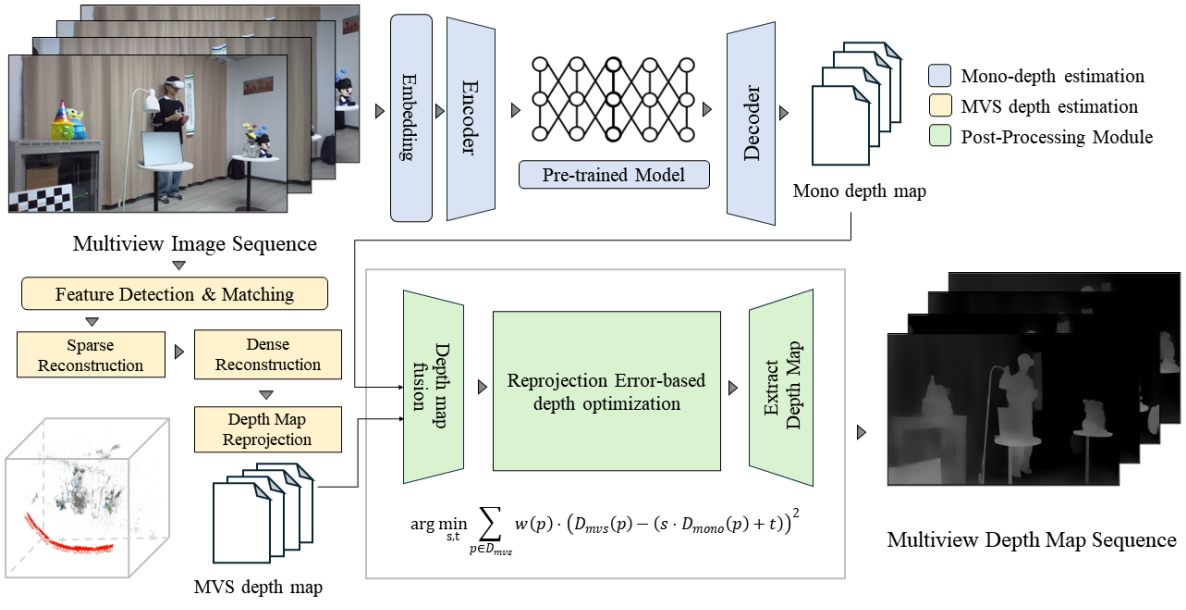


그림 4. 다시점 깊이 맵 생성 기법 파이프라인
Fig. 4. Multi-view depth generation method pipeline

트랜스포머의 self-attention mechanism을 활용하여 이미지 내 모든 픽셀 간의 상호작용을 학습하고, 이미지 전체의 문맥 정보를 고려하여 각 픽셀 간의 관계를 분석한다. 이러한 전역적 특징 학습을 통해 국소적 정보를 처리하는 기존의 CNN 모델들과 달리, 멀리 떨어진 물체 간의 상호작용도 고려한 깊이 추정을 할 수 있다.

세 번째 단계는 깊이 정보 예측 단계이다. 트랜스포머의 디코더 (decoder) 부분이 앞서 추출된 전역적 특징을 바탕으로 각 픽셀의 깊이 정보를 예측한다. 이 과정을 통해 이미지 내 모든 픽셀에 대해 밀집된 깊이 맵이 생성되며, 복잡한 구조와 텍스처를 정확하게 반영한 깊이 맵이 출력된다. 이러한 DPT의 과정은 3차원 재구성에서 세밀한 정보를 잃지 않도록 하며, 사용자 시점 장면을 보다 정교하게 생성하는 데 중요한 역할을 한다.

그러나 DPT 모델을 통해 생성된 깊이 맵은 단일 시점 내 물체 간의 상대적 거리를 잘 표현할 수 있지만, 물체와 카메라 간의 절대적 거리나 실제 크기에 대한 정보는 제공하지 못한다는 한계점을 가진다. 이러한 문제로 인해 여러 시점에서 얻어진 깊이 맵을 결합하여 3차원 재구성을 할 때, 각 시점의 깊이 맵이 서로 다른 스케일로 표현될 가능성이 있다. 이를 스케일 모호성 (scale ambiguity) 문제라고 한다. 이를 해결하지 않고 단안 시점에서 추론한 깊이 맵을 그대로 3차원 재구성에 활용할 경우 물체의 크기나 거리가 시점마다 다르게 표현되어 왜곡된 3차원 구조를 생성하게 될 수 있다. 따라서, 단안 깊이 추정에서 생성된 깊이 맵은 스케일 모호성 문제를 해결하기 위한 추가적인 최적화 과정이 필요하다.

그림 5는 사전 훈련된 DPT 모델을 활용하여 생성한 단안 깊이 맵을 보여준다. 이 예시에서 볼 수 있듯이, 동일한 3차원 공간에 있는 벽면과 사람 간의 거리가 시점별로 다르게 표현되고 있다. 즉, 같은 물체들이 실제로는 동일한 거리감으로 표현되어야 하지만, 각 시점에서 생성된 깊이 맵은 거리감이 일관되지 않게 나타난다. 이는 시점별로 3차원 구조성이 정확하게 정합되지 않음을 의미한다. 이 문제는 여러 시점에서 생성된 깊이 맵을 결합하여 3차원 재구성을 시도할 때, 기하학적 왜곡을 초래할 수 있으며, 물체가 왜곡된

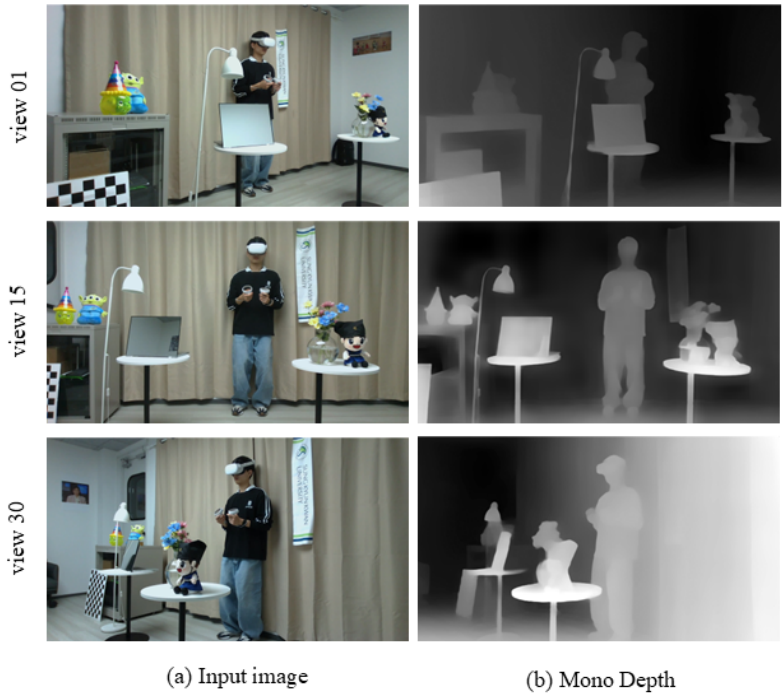


그림 5. 시점별 단안 깊이 맵 추정 결과
Fig. 5. Mono depth generation outupt

형태로 나타나는 결과를 초래할 수 있다. 이를 해결하기 위해서는 단일 시점에서 생성된 깊이 맵에 대해 스케일과 오프셋을 보정하여 시점 간 일관된 3차원 구조를 유지할 필요가 있다.

2. MVS 기반 다시점 깊이 정보 생성

본 연구에서 제안하는 깊이 정보 생성 기법의 두 번째 단계는, MVS 기반 다시점 깊이 정보를 생성하는 것이다. MVS는 다중 시점에서 촬영된 이미지들로부터 각 시점의 시차 정보를 이용하여 각 시점에서 동일하게 관찰된 3차원 포인트 클라우드로 깊이 정보를 추정하는 기술이다. 해당 기술로 얻은 깊이 정보는 비교적 희소한 정보에 기반하므로, 전체 장면의 밀집된 3차원 구조를 정확하게 나타내기에는 한계가 있다. 하지만 각 시점에서 동일하게 관찰된 특징점들이 다중 시점에 걸쳐 일관된 거리 척도를 제공한다는 점에서 앞서 생성한 단일 시점 깊이 맵의 스케일 모호성을 해결하는 최적화 과정에 활용될 수 있다.

3. MVS 기반 최적화를 통한 다시점 깊이 정보 생성

본 연구에서 제안하는 깊이 정보 생성 기법의 최종 단계는 단안 깊이 추정에서 발생하는 스케일 모호성 문제를 해결하기 위한 최적화 과정이다. 앞서 설명한 바와 같이, MVS 기법으로 생성된 깊이 정보를 목표 데이터로 삼아 단일 시점에서 생성된 깊이 맵의 스케일 및 오프셋을 최적화하는 과정을 수행한다. 단안 깊이 추정 모델로 생성된 깊이 맵 D_{mono}을 최적화하는 과정은 다음과 같은 수식으로 표현될 수 있다:

$$D_{mono} = F_{\Theta}(\mathrm{I}) \tag{1}$$

$$D_{optimized} = s \times F_{\Theta}(\mathrm{I}) + \mathrm{t} \tag{2}$$

여기서, D_{mono}는 단일 시점에서 추정된 깊이 값이다. F_{Θ}는 단일 시점 이미지를 입력으로 받아 깊이 맵을 생성하는 네트워크이며, s는 스케일 파라미터, t는 오프셋 파라미터이다. 이 두 파라미터는 단일 이미지로부터 얻어진 상대적 깊이 값을 보정하여 다중 시점 간 일관된 스케일을 유지하는 데 사용된다. MVS로 생성된 포인트 클라우드는 다중 시점에서 공유되는 특징점을 기반으로 물체 간의 실제 물리적 거리 척도를 제공한다. 따라서, 스케일 파라미터 s와 오프셋 파라미터 t는 MVS로 생성된 포인트 클라우드에서 얻은 시점 간 3차원 구조적 정합성이 보장된 깊이 값 $D_{mvs}(p)$와 단일 시점에서 추정된 깊이 값 $D_{mono}(p)$ 간의 차이를 최소화하는 방식으로 최적화된다. 이때 최적화를 위한 손실 함수는 다음과 같이 정의된다:

$$s^*, t^* = \underset{\mathrm{s,t}}{\arg\min} \sum_{p \in D_{mvs}} w(p) \cdot (D_{mvs}(p) - (s \cdot D_{mono}(p) + t))^2 \tag{3}$$

여기서, $D_{mvs}(p)$는 MVS 기반 3차원 포인트 클라우드에서 얻은 깊이 값이고, $D_{mono}(p)$는 단안 깊이 추정 모델에서 얻은 깊이 맵 내부의 각 픽셀의 깊이 값이다. 스케일 파라미터 s와 오프셋 파라미터 t는 두 깊이 값 간의 차이를 최소화하는 방식으로 최적화되며, $w(p)$는 각 픽셀 p의 신뢰도를 나타내는 가중치 함수이다. 본 실험에서 가중치 $w(p)$는 재투영 오류 (reprojection error)를 기반으로 설정된다. 재투영 오류가 작을수록 해당 픽셀의 깊이 정보는 신뢰할 수 있으므로 더 높은 가중치가 부여된다. 최종적으로, 이 최적화 과정을 통해 얻은 파라미터로 단일 시점에서 추정된 깊이 맵을 조정해 일관된 스케일과 구조성을 유지하게 된다. 이를 통해 밀집된 깊이 정보를 제공하며 시점 간의 일관된 척도를 유지하는 두 가지 조건을 모두 만족하는 고품질의 깊이 맵을 생성한다.

Ⅳ. 실험 결과

본 연구에서는 30개의 다시점 카메라 배열을 사용해 실내 환경에서 촬영된 데이터셋을 기반으로 깊이 이미지를 출력하였다. 이 데이터셋은 다양한 텍스처 및 조명 조건을 포함하고 있어, 제안된 깊이 정보 추정 모델의 성능을 종합적으로 평가하는 데 적합하다. 실험 환경은 복잡한 텍스처를 가진 물체와 텍스처가 거의 없는 단조로운 표면을 모두

포함하여, 깊이 추정 과정에서 발생할 수 있는 다양한 문제를 포괄적으로 테스트할 수 있도록 설정되었다.

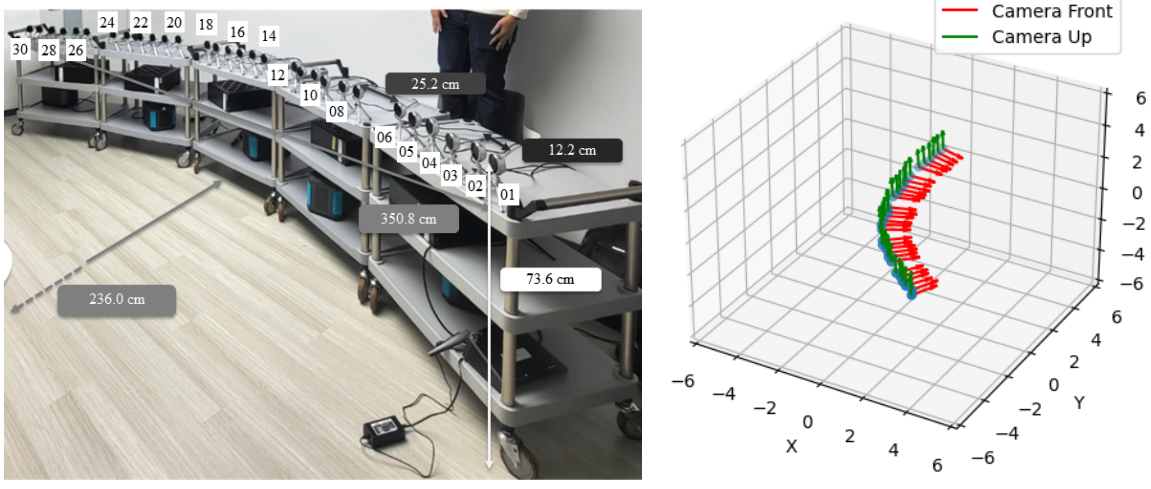
데이터셋 촬영에 사용된 카메라는 Intel RealSense L515 모델로, 이 카메라는 1920x1080 해상도의 24비트 RGB 비디오 파일을 생성한다. 카메라의 초점 거리는 1.88mm이며, 수평 시야각 (FOV)은 69°, 수직 시야각은 42°이다. 총 30대의 카메라가 배치되었으며, 5개의 엣지 디바이스에 각각 6대의 카메라가 연결되었다. 이 중 하나의 엣지 디바이스는 서버 역할을 하며, 서버는 녹화 시작 시 클라이언트들에게 시간 정보를 포함한 시작 신호를 보낸다. 이후 시간 차이를 보정하기 위해 서버의 전역 타임스탬프를 기반으로 후처리가 이루어졌다. 각 카메라의 위치 및 회전 정보는 colmap을 통해 추정되었으며 표 1은 해당 콘텐츠에 대한 기본 설명을 제공하며 그림 6은 데이터셋 촬영 환경 및 카메라 배열을 나타낸다[28-29].

제안 기법을 통해 생성된 깊이 맵의 시각적 품질을 보다 명확히 비교하기 위해, 그림 7과 같이 총 4개의 뷰에서 단안 시점 깊이 맵과 MVS 기반 깊이 맵 그리고 제안 방식을 통해 출력한 깊이 맵을 비교하였다.

표 1. VRroom 데이터셋 개요
Table 1. The description of VRroom dataset

Item	Specification
Number of frames	300
Number of views	30 (v01~v30)
Format	mp4 (H.264 codec)
Resolution	1920 x 1080
FPS	30.0
Total size	373 MB

제안된 방식은 기존 단안 깊이 추정 방식과 마찬가지로, 텍스처가 부족한 영역이나 복잡한 구조를 가진 장면에서도 일관된 고품질의 깊이 정보를 제공하는 성능을 나타내었다. 또한, MVS 기반 깊이 정보를 활용한 최적화 과정을 통해, 기존의 단안 깊이 추정 방식에서 발생하던 시점별 거리 스케일의 불일치 문제가 효과적으로 해결되었다. 기존 방식에서는 각 시점에서 물체 간의 거리가 다르게 표현되어 시점 이동 시 왜곡이 발생했으나, 제안된 방법은 각 시점에서 일관된 스케일을 유지하여 3차원 구조의 정확성이 크게 향상되었다.



(a) (b)

그림 6. VRroom 데이터셋 실험 조건 (a) 데이터셋 취득 환경 (b) 카메라 배열 시각화
Fig. 6. Test condition of VRroom dataset (a) dataset acquisition environment (b) visualization of camera array

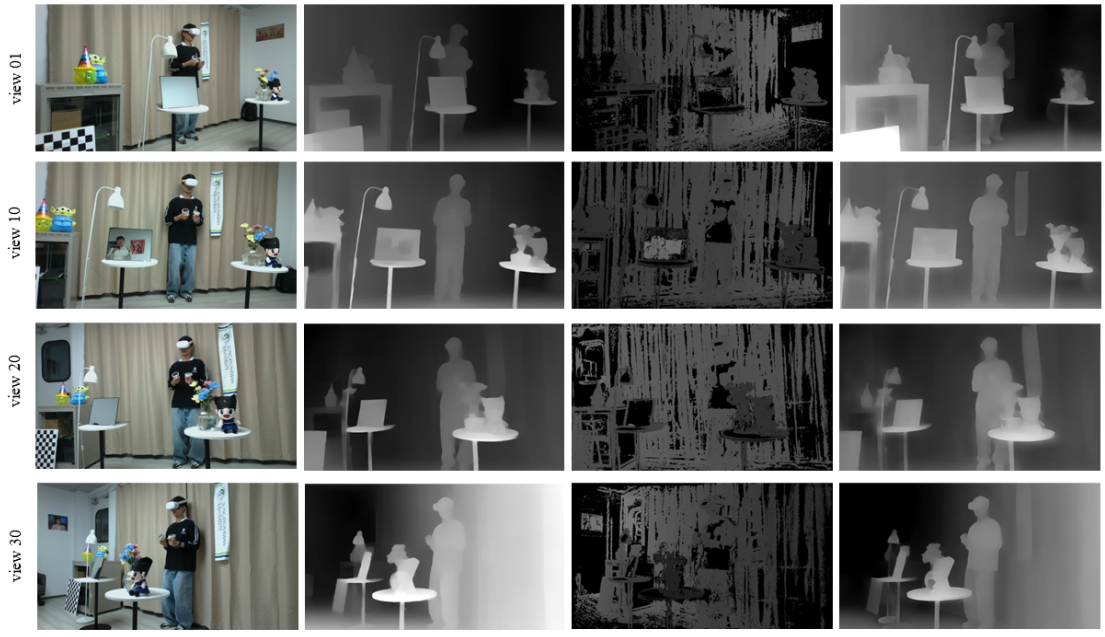


그림 7. 제안 기법으로 생성된 깊이 맵과 단안 깊이 및 MVS 기반 깊이 정보 비교 (a) 원본 이미지 (b) 단안 시점 깊이 맵 (c) 다중 스테레오 기반 깊이 맵 (d) 제안 기법 깊이 맵

Fig. 7. Comparison of depth maps generated by the proposed method, monocular depth, and MVS depth (a) Input image (b) monocular depth map (c) mvs depth map (d) optimized depth map

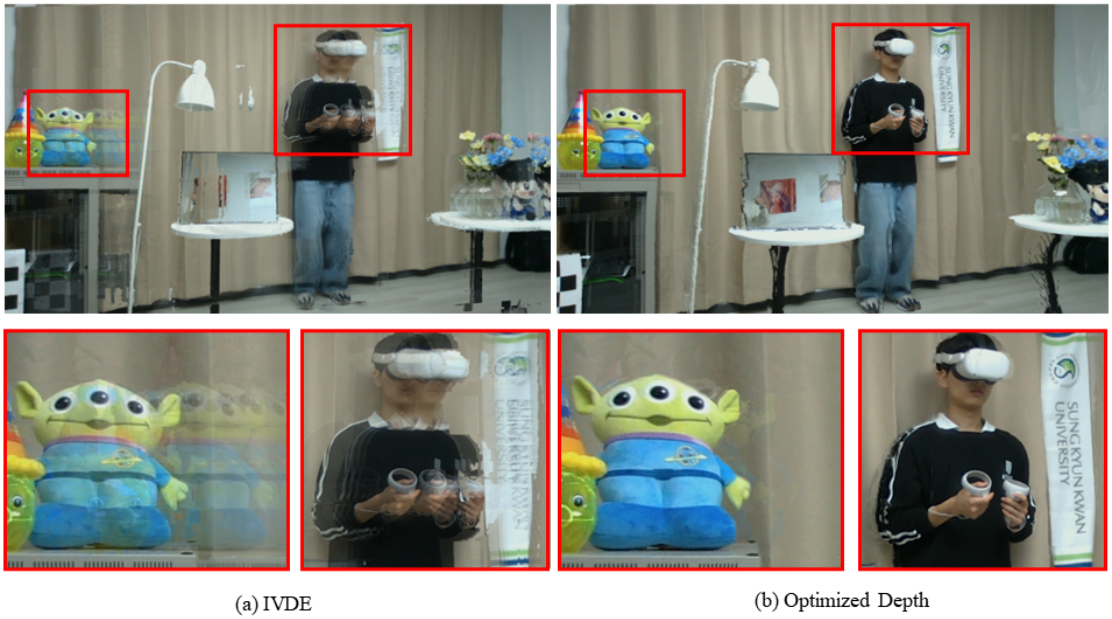


그림 8. IVDE 방식과 제안된 기법을 사용하여 생성된 깊이 맵을 활용해 합성한 중간 시점 이미지 비교

Fig 8. Comparison of novel view synthesis output Using IVDE and the proposed method

추가로, 제안된 기법을 통해 생성된 깊이 맵 시퀀스가 기하학적 연속성을 보장하는지 검증하기 위해, 동일한 실험 조건에서 IVDE와 제안된 기법으로 생성된 깊이 맵을 사용하여 중간 시점을 생성하고 비교하였다. 3차원 재구성을 위해 MIV DSDE 프로파일을 사용했다. 그림 8은 IVDE 방식으로 생성된 깊이 맵을 활용해 합성한 중간 시점 이미지와, 제안된 기법을 통해 생성된 깊이 맵을 기반으로 재구성한 중간 시점 이미지를 비교한 결과를 나타낸다. 실험 결과, 기존 IVDE 방식에서는 중간 시점에서 물체 경계에 발생하는 아티팩트가 다수 관찰된 반면, 제안된 기법에서는 이러한 아티팩트가 현저히 감소하였다. 이를 통해 제안된 기법이 기존 기법 대비 3차원 구조성 정합에서 더 우수한 성능을 발휘함을 검증했다.

V. 결 론

본 연구에서는 ViT 기반의 단안 깊이 예측 모델인 DPT와 다중 시점 스테레오 방식으로부터 얻은 깊이 정보를 결합하여 고품질의 깊이 맵을 생성하는 새로운 방법을 제안하였다. 단안 깊이 추정의 한계점인 스케일 모호성 문제를 해결하기 위해 MVS 기반의 3차원 포인트 클라우드를 활용하여, 각 시점의 깊이 맵을 정교하게 보정하는 최적화 과정을 도입하였다. 이를 통해 각 시점에서 생성된 깊이 맵 간의 일관성을 유지할 수 있었으며, 다중 시점 기반의 정확한 3차원 구조 재구성이 가능해졌다.

실험 결과, 제안된 모델은 기존의 다시점 깊이 정보 추정 기법인 IVDE 방식과 비교하여 정량적, 정성적 평가 모두에서 우수한 성능을 입증하였다. 제안된 모델은 특히 텍스처가 부족하거나 조명이 균일한 영역에서도 깊이 정보의 정확성을 크게 개선하였으며, 처리 속도 또한 크게 향상되어 실시간 응용 가능성까지 확인할 수 있었다.

향후, 본 연구는 몰입형 영상을 활용하는 다양한 분야에서 응용 가능성을 확대할 수 있으며, 다중 시점에서 정확한 깊이 정보 생성이 필요한 분야에서 중요한 기여를 할 수 있을 것이다. 또한, 카메라 포즈 추정이나 물체 추적과 같은 관련 응용 분야에서도 활용될 가능성이 크다. 추후 다양한 환경과 조건에서의 실험을 통해 제안된 방법의 범용성을 검증하고, 이를 실제 응용 시스템에 통합할 수 있는 방법을 모색할 필요가 있다.

참 고 문 헌 (References)

[1] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R., “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1085-1094, June 2020.
doi: https://doi.org/10.1145/350325

[2] Yu, A., Ye, Z., Tancik, M., Kanazawa, A., “PlenOctrees for Real-time Rendering of Neural Radiance Fields”, IEEE International Conference on Computer Vision (ICCV), pp. 5-15, October 2021.
doi: https://doi.org/10.1109/iccv48922.2021.00570

[3] C. Sun, Y. H. Tseng, H. Xu, S. Su, K. Xu, and L. Bao. “Direct Voxel Grid Optimization: Super-Fast Convergence for Radiance Fields Reconstruction”, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 5459-5468 2022.
doi: https://doi.org/10.1109/cvpr52688.2022.00538

[4] Mueller, T., Riegler, G., Nowrouzezahrai, D., & Koltun, V., “Instant Neural Graphics Primitives with a Multiresolution Hash Encoding”, ACM SIGGRAPH Conference Proceedings, pp. 23-33, July 2022.
doi: https://doi.org/10.1145/3528223.3530127

[5] Trevithick, Alex, and Bo Yang. “GRF: Learning a General Radiance Field for 3D Representation and Rendering.” IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2021.
doi: https://doi.org/10.1109/iccv48922.2021.01490.

[6] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuhler, and George Drettakis. “3D Gaussian Splatting for Real-Time Radiance Field Rendering,” ACM Transactions on Graphics, vol. 42, no. 4, pp. 1-14, April 2023.
doi: https://doi.org/10.1145/3592433

[7] Franke, L., Rückert, D., Fink, L., & Stamminger, M. “TRIPS: Trilinear Point Splatting for Real Time Radiance Field Rendering,” Computer Graphics Forum, vol. 43, no. 2, April 2024.
doi: https://doi.org/10.1111/cgf.15012

[8] Deng, Kangle, et al. “Depth-supervised nerf: Fewer views and faster training for free.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
doi: https://doi.org/10.1109/cvpr52688.2022.01254

[9] Roessle, Barbara, et al. “Dense depth priors for neural radiance fields from sparse input views.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
doi: https://doi.org/10.1109/cvpr52688.2022.01255

[10] Turkulainen, Matias, et al. “DN-Splatter: Depth and Normal Priors for Gaussian Splatting and Meshing.” arXiv preprint arXiv:2403.17822 (2024).
doi: https://doi.org/10.48550/arXiv.2403.17822

[11] Dawid Mieloch, Adrian Dziembowski, Jakub Stankowski, Olgierd Stankiewicz, Marek Domanski, Gwangsoon Lee, and Yun Young Jeong. “Immersive Video Depth Estimation,” ISO/IEC JTC 1/SC 29/WG 11 m53407, April 2020.

[12] Adrian Dziembowski, Dawid Mieloch, Jun Young Jeong, and Gwangsoon Lee. “MIV Decoder-Side Depth Estimation Profile,” ISO/IEC JTC 1/SC 29/WG 4 m60667, October 2022.

[13] Johannes L. Schönberger and Jan-Michael Frahm. “Structure-from-Motion Revisited,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4104-4113, June 2016.
doi: https://doi.org/10.1109/CVPR.2016.445

[14] Shimon Ullman. “The Interpretation of Structure from Motion,” Proceedings of the Royal Society of London. Series B. Biological Sciences, vol. 203, no. 1153, pp. 405-426, January 1979.
doi: https://doi.org/10.1098/rspb.1979.0006

[15] David G. Lowe. “Distinctive Image Features from Scale-Invariant Keypoints,” International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, November 2004.
doi: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[16] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. “SURF: Speeded Up Robust Features,” European Conference on Computer Vision (ECCV), pp. 404-417, 2006.
doi: https://doi.org/10.1007/11744023_32

[17] Andrea Vedaldi and Andrew Zisserman. “Efficient Additive Kernels via Explicit Feature Maps,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 3, pp. 480-492, March 2012.
doi: https://doi.org/10.1109/TPAMI.2011.153

[18] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. “Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 10, pp. 2024-2039, October 2016.
doi: https://doi.org/10.1109/TPAMI.2015.2505283

[19] Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. “Joint 2D-3D-Semantic Data for Indoor Scene Understanding,” arXiv preprint arXiv:1702.01105, 2017. Available at https://arxiv.org/abs/1702.01105

[20] Zhao, Chaoqiang, et al. “Monocular depth estimation based on deep learning: An overview.” Science China Technological Sciences,63, 1612-1627, 2020.
doi: https://doi.org/10.1007/s11431-020-1582-8

[21] B. Graham, H. Elsen, J. Touvron, F. Massa, E. Belilovsky, A. Wightman, M. Douze, R. Auli, H. Jégou. “LeViT: a Vision Transformer in ConvNet’s Clothing for Faster Inference,” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 12239-12249, 2021.
doi: https://doi.org/10.1109/ICCV48922.2021.01204

[22] Khan, Salman, et al. “Transformers in vision: A survey.” ACM computing surveys (CSUR) 54.10s, 1-41, 2022.
doi: https://doi.org/10.1145/3505244

[23] Fan, Haoqi, et al. “Multiscale vision transformers.” Proceedings of the IEEE/CVF international conference on computer vision. 2021.
doi: https://doi.org/10.1109/iccv48922.2021.00675

[24] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. “Vision Transformers for Dense Prediction,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12179-12188, October 2021.
doi: https://doi.org/10.1109/iccv48922.2021.01196

[25] J. M. Boyce et al., “MPEG Immersive Video Coding Standard,” Proceedings of the IEEE, vol. 109, no. 9, pp. 1521-1534, September 2021.
doi: https://doi.org/10.1109/JPROC.2021.3062590

[26] Jong-Beom Jeong, Soonbin Lee, and Eun-Seok Ryu. “Rethinking Fatigue-Aware 6DoF Video Streaming: Focusing on MPEG Immersive Video,” International Conference on Information and Communication Technologies (ICT), pp. 1-6, December 2022.
doi: https://doi.org/10.1109/ICOIN53446.2022.9687247

[27] Jong-Beom Jeong, Soonbin Lee, and Eun-Seok Ryu. “DATRA-MIV: Decoder-Adaptive Tiling and Rate Allocation for MPEG Immersive Video,” ACM Transactions on Multimedia Computing, pp. 1-20, January 2024.
doi: https://doi.org/10.1145/3648371

[28] Jaeyeol Choi, Jong-Beom Jeong, Jun-Hyeong Park, Yeongil Ryu, Issac Yang, Jinho Lee, Gun Bang, Eun-Seok Ryu. “[INVR] Color Corrected SKKU_VRroom1D,” MPEG2024/m66418, 144th MPEG meeting of ISO/IEC JTC1/SC29/WG4, January 2024.

[29] Jaeyeol Choi, Yeongil Ryu, Yihyun Choi, Jong-Beom Jeong, Jun-Hyeong Park, Issac Yang, Eun-Seok Ryu. “[INVR] EE2.1-Related: Report with New Natural INVR Video Contents: SKKU_VRroom,” MPEG2023/m64721, 144th MPEG meeting of ISO/IEC JTC1/SC29/WG4, October 2023.

저 자 소 개

박 준 형

- 2018년 3월 ~ 2024년 2월 : 성균관대학교 영상학과 학사
- 2024년 3월 ~ 현재 : 성균관대학교 실감미디어공학과 석사과정
- 2023년 7월 ~ 2023년 8월 : 한국전자통신연구원 학생연구원
- ORCID : https://orcid.org/0009-0000-6524-1559
- 주관심분야 : 실감미디어, 인공지능, 그래픽스, 멀티미디어 통신 및 시스템

정 종 범

- 2018년 8월 : 가천대학교 컴퓨터공학과 학사
- 2018년 9월 ~ 2019년 8월 : 가천대학교 컴퓨터공학과 석사과정
- 2019년 9월 ~ 현재 : 성균관대학교 컴퓨터교육학과 석박통합과정
- 2020년 1월 ~ 2020년 3월 : University of California, Santa Barbara 방문연구원
- 2021년 8월 ~ 2022년 1월 : Purdue University 방문연구원
- 2022년 9월 ~ 2023년 8월 : 성균관대학교 글로벌융합학부 강사
- 2023년 9월 ~ 2024년 8월 : 성균관대학교 실감미디어공학과 강사
- ORCID : https://orcid.org/0000-0002-7356-5753
- 주관심분야 : 멀티미디어 통신 및 시스템, 비디오 압축 표준, MPEG immersive video, video-based dynamic mesh coding

최 재 열

- 2018년 3월 ~ 2024년 2월 : 성균관대학교 컴퓨터교육학과 학사
- 2024년 3월 ~ 현재 : 성균관대학교 인공지능융합학과 석사과정
- 2023년 1월 ~ 2023년 2월 : 한국전자통신연구원 학생연구원
- ORCID : https://orcid.org/0009-0009-2923-1252
- 주관심분야 : 실감미디어, 인공지능, 그래픽스, 멀티미디어 통신 및 시스템

김 영 규

- 2016년 3월 ~ 2024년 8월 : 성균관대학교 중어중문학과 학사
- 2024년 9월 ~ 현재 : 성균관대학교 실감미디어공학과 석사과정
- ORCID : https://orcid.org/0009-0008-5470-3103
- 주관심분야 : 실감미디어, 인공지능, 볼류메트릭 비디오

류 은 석

- 1999년 8월 : 고려대학교 컴퓨터학과 학사
- 2001년 8월 : 고려대학교 컴퓨터학과 석사
- 2008년 2월 : 고려대학교 컴퓨터학과 박사
- 2008년 3월 ~ 2008년 8월 : 고려대학교 연구교수
- 2008년 9월 ~ 2010년 12월 : 조지아공대 박사후과정
- 2011년 1월 ~ 2014년 2월 : InterDigital Labs Staff Engineer
- 2014년 3월 ~ 2015년 2월 : 삼성전자 수석연구원/파트장
- 2015년 3월 ~ 2019년 8월 : 가천대학교 컴퓨터공학과 조교수
- 2019년 9월 ~ 2023년 6월 : 성균관대학교 컴퓨터교육과 조/부교수
- 2023년 9월 ~ 현재 : 성균관대학교 실감미디어공학과 부교수
- ORCID : https://orcid.org/0000-0003-4894-6105
- 주관심분야 : 멀티미디어 통신 및 시스템, 비디오 코딩 및 국제 표준, HMD/VR 응용분야