현실세계를 이해하고 상호작용하는 공간지능 기술 동향

Technology Trends in Spatial Intelligence for Understanding and Interacting with the Real World

김낙우 (N.W. Kim, nwkim@etri.re.kr) 산업콘텐츠지능화연구실 책임연구원
채원석 (W.S. Chae, wschae@etri.re.kr) 산업콘텐츠지능화연구실 책임연구원
김근용 (G.Y. Kim, gykim@etri.re.kr) 산업콘텐츠지능화연구실 책임연구원
유호영 (H.Y. Yoo, yoohy@etri.re.kr) 산업콘텐츠지능화연구실 선임연구원
이상은 (S.E. Lee, sange1104@etri.re.kr) 산업콘텐츠지능화연구실 연구원
김현진 (H.J. Kim, jini@etri.re.kr) 산업콘텐츠지능화연구실 책임연구원/실장

ABSTRACT

Artificial intelligence (AI) is evolving toward spatial intelligence, which enables AI to generate and continuously update three-dimensional (3D) representations of the real world, understand spatial environments, predict future changes, and interact with the physical world. This paper reviews recent advances in spatial intelligence from three perspectives: multimodal spatial generation, real-time spatial reconstruction and dynamic spatial updating, and spatial operational intelligence. We examine advances in 3D/4D spatial generation and dynamic updating using neural radiance fields, 3D Gaussian splatting, and generative AI, along with developments in 3D spatial understanding, world-model-based prediction, physical reasoning, and vision–language–action models. Finally, we discuss the evolution toward integrated 3D spatiotemporal intelligence that combines these capabilities to continuously understand and interact with the real world.

KEYWORDS 공간지능, 동적 공간갱신, 월드 모델, 3DGS, 3D 공간생성

I. 서론

최근 인공지능 기술은 대규모 언어모델(LLM: Large Language Model), 비전-언어모델(VLM: Vision-Language Model), 생성형 인공지능의 발전을 바탕으로 텍스트, 이미지, 영상 등 다양한 형태의 데이터를 이해하고 생성할 수 있는 수준으로 발전하고 있다. 특히 ChatGPT, Gemini, Claude와 같은 언어 기반 인공지능과 Sora, Veo, Seedance 등 영상 생성 모델의 등장은 인공지능의 생성 및 추론 능력을 크게 확장하며 새로운 기술적 전환점을 마련하였다. 그러나 기존 생성형 인공지능의 발전은 주로 디지털 콘텐

* DOI: https://doi.org/10.22648/ETRI.2026.J.410607
* 이 연구는 2026년도 산업통상부 및 한국산업기술기획평가원 연구비 지원에 의한 연구임(과제번호: RS-2026-25544725).

본 저작물은 공공누리 제4유형
출처표시+상업적이용금지+변경금지 조건에 따라 이용할 수 있습니다.
©2026 한국전자통신연구원

츠의 생성과 이해를 중심으로 이루어져 왔으며, 현실세계의 공간적 구조와 변화를 지속적으로 인식하고 이를 기반으로 미래 상태를 예측하며 물리세계와 상호작용을 하는 공간지능(Spatial Intelligence) 기술은 아직 발전 초기 단계에 있다.

공간지능은 현실세계를 단순히 디지털 환경에 재현하는 것을 넘어, 공간의 구조와 의미를 이해하고 변화하는 상태를 지속적으로 반영하며 미래 상황을 예측 · 진단하여 실제 환경과 상호작용을 할 수 있는 차세대 인공지능 기술이다. 자율주행, 로봇, 스마트시티 등 물리적 공간과 직접 상호작용을 하는 응용 분야가 확대되면서, 객체를 개별적으로 인식하는 기존 AI를 넘어 공간 전체의 구조와 객체 간 관계, 시간에 따른 변화 및 물리적 상호작용을 종합적으로 이해하는 기술의 중요성이 증가하고 있다.

최근 World Labs의 Atlas[1], Google DeepMind의 Genie 3[2], NVIDIA의 Cosmos 3[3,4], Meta의 V-JEPA 2.1 등으로 대표되는 월드 모델 연구는 멀티모달 정보를 기반으로 물리세계의 구조와 상태 변화를 학습하고, 세계를 생성 · 표현하는 것에서 나아가 미래 상태의 예측과 추론, 상호작용 및 행동을 통합하는 방향으로 발전하고 있다. 표 1은 주요 월드 모델의 공간지능 특성을 비교한 것으로, 각 모델이 공간 생성 · 재구성, 동적 세계 시뮬레이션, 물리적 추론 및 시공간 표현학습 등 서로 다른 기술적 접근을 취하고 있음을 보여준다. 이러한 연구는 기술적 접근 방식에는 차이가 있으나, AI의 학습 및 추론 대상을 텍스트와 이미지 중심의 디지털 정보에서 실제 3차원 물리세계로 확장한다는 점에서 공간지능 발전의 공통적인 흐름을 나타낸다. 그러나 현재 관련 연구는 멀티모달 데이터로부터 공간을 생성 · 표현하는 기술, 현실 공간의 변화를 지속적으로 관측하여 디지털 공간을 동기화하고 갱신하는 기술, 구축된 공간을 이해하고 미래 상태를 예측하며 행동을 결정하는 기술 등이 서로 다른 연구 영역에서 발전하고 있어, 이들을 유기적으로 연계한 통합적 공간지능 구현에는 여전히 기술적 과제가 존재한다.

이에 본고에서는 공간지능을 생성(Generation), 갱신(Update) 및 운영(Operation)으로 이어지는 연속적인 기술 체계의 관점에서 살펴보고자 한다. 우선 공간 생성 기술은 영상, 텍스트, 3D 데이터 및 공간정보 등 다양한 입력을 활용하여 기하학적 · 의미적 특성과 공간적 관계를 학습하고 새로운 3차원 공간을 생성 · 표현하는 기술을 의미한다. 공간 갱신 기술은 현실 공간을 지속적으로 관측하여 현재 상태

표 1 최신 월드 모델의 공간지능 특성 비교

구분	World Labs Atlas	Google DeepMind Genie 3	NVIDIA Cosmos 3	Meta V-JEPA 2.1
공개일	2026. 9.	2025. 8.	2026. 5.	2026. 3.
핵심 특징	공간 생성·재구성 및 시뮬레이션 통합	상호작용 가능한 동적 세계 생성·시뮬레이션	피지컬 AI의 이해·생성·행동 통합	영상 기반 시공간 표현학습·예측
월드 표현 방식	3차원 공간 맥락 기반 월드 표현	프레임 기반 동적 환경	Omnimodal 통합 월드 표현	잠재 시공간 표현
물리 이해·추론 능력	공간 구조·시간 변화 모델링 중심	생성 과정에서 물리적 상호작용 반영	물리세계 이해·추론 및 행동 연계	시공간 표현·예측 중심
시간적 동역학	강함	매우 강함	강함	강함
모델/코드 공개 여부	비공개, API/서비스	비공개, Project Genie	모델/코드 공개	코드/체크포인트 공개

와 변화를 파악하고, 현실세계와 디지털 공간을 동기화하여 공간정보의 최신성을 유지하는 기술에 해당한다. 마지막으로 공간 운영 기술은 최신화된 공간정보를 기반으로 공간의 구조와 객체 간 관계를 이해하고, 시공간적 변화와 물리적 상호작용을 예측 · 추론하여 의사결정과 행동 계획 및 실제 상호작용으로 연결하는 기술을 말한다.

이러한 세 기술은 독립적인 요소기술이라기보다 공간을 생성 · 표현하고, 현실 변화에 따라 지속적으로 갱신하며, 이를 이해 · 예측하여 판단하고 상호작용을 하는 연속적인 공간지능 구조를 형성한다. 즉, 멀티모달 공간 생성 AI가 공간지능의 기본 표현을 제공하고, 실시간 공간 동기화 및 동적 갱신 기술이 현실세계와 디지털 공간 간의 지속적인 일관성을 확보하며, 공간 운영 인텔리전스가 이를 기반으로 시공간 예측, 물리적 추론, 의사결정 및 행동을 수행함으로써 현실세계와 지속적으로 상호작용을 할 수 있는 통합적 공간지능 체계로 발전할 수 있다.

본고에서는 이러한 관점에서 현실세계를 이해하고 상호작용을 하기 위한 공간지능의 주요 기술 동향을 살펴보고자 하며, 구성은 다음과 같다. 먼저 Ⅱ장에서는 다양한 형태의 데이터를 활용하여 공간을 생성 · 표현하는 멀티모달 공간 생성 AI 기술을 소개한다. 다음으로 Ⅲ장에서는 현실 공간의 변화를 지속적으로 반영하기 위한 실시간 공간 동기화 및 동적 갱신 기술의 주요 연구 동향을 살펴본다. Ⅳ장에서는 생성 · 갱신된 공간을 기반으로 공간의 관계와 상황을 이해하고, 미래 상태의 예측과 물리적 추론을 통해 의사결정 및 행동으로 연결하는 공간 운영 인텔리전스 기술을 분석한다. 마지막으로 Ⅴ장에서는 주요 기술 동향을 종합하고, 생성 · 갱신 · 운영 기술의 연계 및 통합을 중심으로 향후 공간지능 기술의 발전 방향을 제시한다.

Ⅱ. 멀티모달 공간 생성 AI 기술 동향

최근 생성형 AI 기술은 텍스트와 이미지, 영상 중심의 콘텐츠 생성을 넘어 3차원 객체와 장면을 직접 생성하는 방향으로 확장되고 있다. 초기 3D 생성 연구는 주로 텍스트나 단일 영상을 조건으로 개별 객체의 형상과 외관을 생성하는 데 초점을 두었으나, 최근에는 다중시점 영상과 다양한 입력 조건을 활용하여 기하학적 · 공간적 일관성을 갖는 콘텐츠를 생성하는 방향으로 발전하고 있다.

1. 생성형 AI 기반 3D/4D 콘텐츠 생성 동향

3D 콘텐츠 생성은 NeRF[5]와 같은 신경망 기반 3D 표현에 확산모델의 생성 능력을 결합한 Text/Image-to-3D 연구를 중심으로 발전해 왔으며, 최근에는 3차원 가우시안 스플래팅(3DGS: 3D Gaussian Splatting)이 새로운 3D 표현으로 활용되면서 생성 품질과 처리 효율성을 동시에 확보하기 위한 연구가 활발히 진행되고 있다. 대표적으로 DreamGaussian[6]은 3D Gaussian을 생성 표현으로 직접 최적화하는 Generative Gaussian Splatting 방식을 통해 기존 NeRF 기반 생성 방식에서 요구되던 높은 계산 비용과 긴 최적화 시간을 크게 단축하였다. 이와 함께 GaussianDreamer[7]는 2D/3D 확산모델의 사전지식을 결합하여 Gaussian의 초기 형상과 기하학적 일관성을 개선하였고, GSGEN[8]은 3D 생성모델의 사전학습 지식을 활용한 기하구조 생성과 2D 영상 생성모델을 활용한 외형 최적화를 결합하는 방식을 제안하였다. 이러한 연구들은 확산모델의 생성 지식과 3D Gaussian 표현을 결합하여 3D 콘텐츠 생성의 효율성과 형상 안정성, 생성 품질을 향상시키는 방향으로 발전하고 있다.

이후 3D 생성 연구는 서로 다른 시점에서 나타나는 형상 왜곡과 불일치를 개선하기 위해 다시점 정보를 활용하여 기하학적 일관성을 확보하려는 연구가 진행 중이다. Wonder3D[9]는 단일 이미지로부터 다시점 RGB와 기하정보를 생성하여 시점 간 기하학적 일관성을 향상시키는 접근을 제시하였으며, Large Multi-View Gaussian Model[10]은 다시점 이미지를 입력받아 별도의 반복적 최적화 없이 3D Gaussian 표현을 직접 생성함으로써 3D 콘텐츠 생성의 효율성을 높이는 방식을 제안하였다. 이러한 흐름은 다시점 정보를 활용하여 3D 콘텐츠의 기하학적 일관성과 생성 효율성을 함께 향상시키는 방향으로 연구가 발전하고 있음을 나타낸다.

나아가 3D 콘텐츠 생성 연구는 정적인 형상과 외관을 생성하는 것을 넘어 시간에 따른 변화와 움직임을 표현하는 동적 4D 콘텐츠 생성으로 범위가 넓어지고 있다. 특히, Align Your Gaussians[11]는 동적 3D Gaussian 표현과 확산모델을 결합하여 텍스트 조건으로 시간에 따라 변화하는 4D 콘텐츠를 생성함으로써, 정적 3D 콘텐츠 중심의 생성 범위를 시간축을 포함하는 동적 콘텐츠 생성으로 확장하였다.

2. 객체 중심 생성에서 장면·공간 생성으로의 확장

최근 공간 생성 AI의 또 다른 중요한 변화는 생성 대상 자체가 개별 객체에서 복합적인 3D 장면과 공간으로 확대되고 있다는 점이다. 기존 Text/Image-to-3D 연구가 주로 단일 객체의 형상과 외관 생성에 초점을 두었다면, 장면 단위의 공간 생성은 다수 객체와 주변 환경을 함께 구성하고, 객체의 위치 · 배치 및 상호 관계를 공간적으로 일관되게 표현하는 것에 보다 중점을 두고 있다.

이러한 흐름에서 Prometheus[12]는 텍스트로부터 탐색 가능한 3D 장면을 직접 생성함으로써 객체 중심의 Text-to-3D를 장면 수준의 공간 생성으로 확장하였다. 특히 3차원 공간 구조를 고려한 확산모델을 활용하여 장면의 구조와 외형을 함께 생성하고, 생성된 공간을 자유롭게 탐색할 수 있도록 함으로써 대규모 3D 장면 생성의 가능성을 제시하였다. 한편, Structured 3D Latents[13]는 복잡한 3D 형상을 구조적으로 표현할 수 있는 잠재 표현을 도입하여 다양한 3D 자산을 효율적으로 생성 · 편집할 수 있는 확장 가능한 생성 방식을 제시하였다. 개별 3D 자산의 형상과 외관을 생성하는 데서 나아가, 복잡한 구조와 공간 구성을 다룰 수 있는 생성모델로 연구 범위가 확대되고 있음을 선보이고 있다. 이러한 연구들을 통해 공간 생성 AI의 대상은 개별 3D 콘텐츠에서 복잡한 장면과 동적 공간으로 확대되고 있다. 그러나 현실 공간에서는 객체의 이동, 시설물의 추가 · 삭제 및 환경 변화 등이 지속적으로 발생하므로, 생성된 공간을 실제 물리세계와 지속적으로 연결하기 위해서는 초기 공간을 구축하는 것만으로는 충분하지 않고, 이러한 변화를 자동으로 인지하고 디지털 공간에 반영하기 위한 추가적인 기술이 요구된다.

III. 실시간 공간 재구성 및 동적 갱신 기술 동향

공간 생성 AI가 새로운 공간의 디지털 표현을 구축하는 데 초점을 둔다면, 공간 동기화 및 동적 갱신은 현실세계를 지속적으로 관측하고 그 변화를 디지털 공간에 반영하여 현실과의 일관성을 유지하는 것을 목적으로 한다. 기존 3D 재구성 기술은 특정 시점의 현실 공간을 정적으로 복원하는 문제를 주로 다루어 왔으나, 공간지능 환경에서는 객체와 주변 환경의 변화가 지속적으로 발생함에 따라 공간

을 실시간으로 구축하고 시간에 따른 변화를 표현하는 동시에, 장기간에 걸쳐 공간정보의 최신성과 일관성을 유지하는 연구가 활성화되고 있다.

1. Gaussian 기반 3D 공간 표현 및 재구성 동향

현실 공간의 효율적인 디지털 표현을 위한 주요 기반기술로 3DGS[14]가 활용되고 있다. 3DGS는 명시적인 Gaussian Primitive를 이용하여 고품질의 3D 장면 표현과 실시간 렌더링을 가능하게 하였으며, 기존 NeRF 중심의 암시적 공간 표현과 달리 장면을 명시적인 Gaussian 집합으로 표현한다. 이후 관련 연구는 새로운 시점의 영상을 생성하는 수준을 넘어, 실제 공간의 기하학적 구조를 더욱 명확하게 복원하고 활용하기 위한 방향으로 확대되고 있다. 대표적으로 SuGaR[15]는 Gaussian Primitive를 장면의 표면에 정렬하고, 이를 기반으로 다각형 메쉬를 추출하는 방법을 제시하였다. 이를 통해 3DGS의 고품질 렌더링 특성을 유지하면서 장면의 표면과 기하구조를 명시적인 메쉬 형태로 복원할 수 있도록 하였다. 이러한 연구는 Gaussian 기반 공간 표현이 실시간 렌더링을 위한 장면 표현에서 나아가, 현실 공간의 기하학적 구조를 복원하고 활용할 수 있는 공간 표현으로 고도화되고 있음을 보인다.

2. 정적 공간에서 동적 시공간 표현으로서의 확장

정적인 3D 공간 재구성에서 나아가, 새로운 관측 정보를 실시간으로 반영하면서 공간을 지속적으로 구축하는 연구도 활발히 진행되고 있다. 대표적으로 SplaTAM[16]은 3D Gaussian을 RGB-D 기반 카메라 추적과 지도 구축에 적용하고, 이를 추적, 지도 구축 및 렌더링을 위한 공통 공간 표현으로 활용하는 방식을 제시하였다. 이를 통해 3DGS의 활용 범위를 정적인 장면 표현에서 새로운 관측정보를 연속적으로 반영하는 온라인 공간 구축으로 확대하였다. 이와 함께 시간에 따른 공간의 변화와 움직임을 표현하기 위한 연구도 제시되고 있다. 4D Gaussian Splatting[17]은 시간에 따른 Gaussian의 변형을 모델링하여 동적인 3D 장면을 실시간으로 표현하였으며, Spacetime Gaussian[18]은 공간적 특성과 시간에 따른 움직임을 하나의 Gaussian 표현에 통합하여 동적 장면을 효율적으로 표현하는 방법을 제시하였다. 이러한 연구는 정적인 3D 공간 표현에 시간적 변화를 결합함으로써 공간의 구조와 움직임을 함께 나타내는 동적 4D 시공간 표현으로 연구 범위를 확대하였다.

최근에는 동적 공간을 명시적으로 표현하는 연구와 더불어, 관측 영상으로부터 공간 구조와 시간적 변화를 직접 추론하는 학습 기반 4D 재구성 연구도 등장하고 있다. Google DeepMind의 D4RT[19]는 비디오로부터 깊이, 3D 포인트의 움직임 및 카메라 자세를 통합적으로 추론하여 동적 장면의 4D 구조를 재구성하는 방식을 제안하였다. 특히 장면별 반복적 최적화 없이 학습된 모델을 활용하여 공간 구조와 움직임을 효율적으로 추론함으로써 4D 공간 연구가 동적 장면을 표현하는 것을 넘어 관측 영상으로부터 시공간 구조와 변화를 직접 재구성하는 방향으로 확대되고 있음을 선보이고 있다.

3. 지속적 공간 갱신 및 대규모 공간 정합 동향

4D 공간 표현 및 재구성이 특정 관측 구간에서 공간의 구조와 시간적 변화를 표현 · 추론하는 데 초점을 둔다면, 최근에는 기존에 구축된 공간 표현

에 새로운 관측정보를 지속적으로 반영하여 장기간 공간 모델의 최신성과 일관성을 유지하는 Continual Mapping 연구가 활발히 진행되고 있다.

대표적으로 Gaussian Mapping for Evolving Scenes (GaME)[20]는 시간에 따라 지속적으로 변화하는 현실 공간을 대상으로, 새로운 관측정보를 기존 3D Gaussian 표현에 반영하여 공간 모델을 지속적으로 갱신하는 방법을 제시하였다. 특히 새롭게 관측된 공간정보를 추가하는 동시에 현재 상태와 일치하지 않는 과거 정보를 관리함으로써, 공간 전체를 반복적으로 재구축하지 않고도 변화하는 환경의 최신 상태를 유지할 수 있도록 하였다. 이러한 연구는 특정 시점이나 관측 구간의 공간을 재구성하는 방식에서 나아가, 장기간의 환경 변화를 지속적으로 반영하면서 공간 모델의 최신성과 일관성을 유지하는 방향으로 공간 Mapping 연구가 확대되고 있음을 나타낸다.

한편, 산업계에서는 개별 장면을 대상으로 한 공간 구축과 갱신을 대규모 현실 공간에 적용하려는 시도가 이루어지고 있다. Niantic Spatial은 Visual Positioning System[21]을 통해 축적된 실제 장소의 대규모 시각 · 공간정보를 기반으로 Large Geospatial Model을 구축하고, 다양한 장소에서 획득한 공간적 특징을 학습하여 새로운 공간에서도 위치와 공간 구조를 인식할 수 있는 방향을 제시하고 있다. 이는 공간 구축 및 갱신 기술의 적용 범위가 개별 장면에 국한되지 않고, 도시 및 광역 수준의 공간정보 학습과 현실 · 디지털 공간 간 정합으로 이어지고 있음을 의미한다.

실시간 공간 동기화 및 동적 갱신 연구는 3D 공간의 표현과 재구성을 기반으로 실시간 공간 구축, 동적 4D 공간 표현 · 재구성, 변화 탐지 및 지속적 갱신, 대규모 공간 정합을 포괄하는 방향으로 연구 범위를 넓혀가고 있다. 특히 변화하는 현실 공간과 디지털 공간 간의 일관성을 지속적으로 유지하기 위해서는 변화 영역의 효율적인 탐지와 갱신뿐만 아니라, 다양한 센서로부터 획득되는 공간정보를 공통 좌표 및 시간체계에서 정합하고 장기간 운영에서도 공간 표현의 일관성을 유지하는 것이 중요한 과제로 남아 있다.

그러나 현실과 지속적으로 동기화된 공간을 확보하는 것만으로는 공간지능을 구현하기에 충분하지 않으며, 최신화된 공간정보를 기반으로 객체 간 관계와 상황을 이해하고 미래의 공간 상태를 예측하여, 목적에 적합한 의사결정과 행동을 계획할 수 있는 공간 운영 인텔리전스 기술도 요구되고 있다.

IV. 공간 운영 인텔리전스 기술 동향

공간 운영 인텔리전스는 생성 · 갱신된 공간정보를 기반으로 공간의 상태와 객체 간 관계를 이해하고, 미래의 변화를 예측하여 의사결정과 행동으로 연결하는 지능화 기술이다. 기존의 공간 인식 기술이 객체 탐지, 분할 및 위치 추정 등을 통해 현재 공간의 상태를 파악하는 데 주력하였다면, 최근에는 공간의 구조와 관계를 이해하는 수준에서 나아가 시공간적 변화를 예측하고 물리적 상황을 추론하여 실제 행동과 상호작용으로 연결하는 방향으로 연구가 발전하고 있다.

1. 3D 공간 이해 및 공간 추론 연구 동향

초기 멀티모달 모델은 주로 2D 이미지와 텍스트 간의 의미적 관계를 학습하는 데 중점을 두었으나, 최근에는 LLM과 VLM의 지식 및 추론 능력을 실제 3차원 공간으로 확장하려는 연구가 활발히 진행되고 있다. 대표적으로 3D-LLM[22]은 3D Point Cloud와 공간 특징을 LLM과 연계하여 장면에 대한

질의응답과 객체 위치 파악, 경로 탐색 등 다양한 공간 이해 및 추론 기능을 수행함으로써 언어모델의 지식과 추론 능력을 3차원 공간정보와 결합하는 기반을 마련하였다.

이후에는 3D 공간의 객체와 장면을 인식하는 수준에서 나아가 거리와 크기, 방향 및 객체 간 상대적 위치관계 등 공간의 기하학적 관계를 이해하고 추론하려는 연구가 이어지고 있다. SpatialVLM[23]은 대규모 공간 질의응답 데이터를 활용하여 거리, 크기 및 객체 간 위치관계 등에 대한 정량적 공간 추론 능력을 강화하였다. SpatialLLM[24]은 3D 공간정보와 공간 중심의 학습 데이터를 대규모 멀티모달 모델에 결합하여 위치와 방향, 객체 간 관계를 종합적으로 이해하고 추론하는 방식을 제시하였다. 이러한 연구를 통해 멀티모달 AI의 공간 이해 범위는 2D 시각정보와 언어 간의 의미적 이해에서 3D 공간정보의 연계와 기하학적 관계에 대한 추론으로 구체화되고 있다.

2. 공간 예측·추론에서 행동지능으로의 기술 고도화

공간에 대한 이해에서 나아가 최근에는 현재의 관측정보로부터 물리세계의 상태와 변화를 학습하고, 미래 상태를 예측하는 월드 모델 연구가 활발히 진행되고 있다. Meta의 V-JEPA(Video Joint-Embedding Predictive Architecture) 2[25]는 대규모 비디오 데이터로부터 물리세계의 잠재 시공간 표현을 자기지도 방식으로 학습하고, 이를 미래 상태 예측과 로봇의 행동 계획에 활용하는 방법을 제시하였다. V-JEPA 2.1[26]은 영상의 세부적인 공간 특징과 시간에 따른 변화를 더욱 정밀하게 학습할 수 있도록 예측학습 방식을 개선하여, 물리세계에 대한 시공간 이해 능력을 보다 강화하였다. 이러한 연구는 공간지능의 역할이 현재 상태에 대한 인식과 이해를 넘어, 물리세계의 시공간적 변화와 움직임을 학습하고 미래 상태를 예측하는 단계로 고도화되고 있음을 나타낸다.

NVIDIA의 Cosmos-Reason 모델[27,28]은 이러한 시공간 이해를 물리적 상식에 기반한 추론과 결합하여 Physical AI의 의사결정과 행동 계획에 활용하는 데 중점을 두고 있다. 특히 Cosmos-Reason2[28]는 공간과 시간, 물리현상에 대한 이해를 바탕으로 시공간적 상황을 추론하고, 객체에 대한 공간적 식별과 장기적인 맥락 이해를 결합하여 물리환경의 상태를 해석하고 에이전트가 수행해야 할 행동을 판단하도록 설계되었다. V-JEPA 모델이 관측정보로부터 물리세계의 시공간 표현을 학습하고 변화를 예측하는 데 중점을 둔다면, Cosmos-Reason 모델은 물리적 상식과 시공간 추론을 바탕으로 상황을 해석하고 의사결정 및 행동 계획을 지원한다는 점에서 차이가 있다.

예측과 추론을 실제 물리세계의 행동으로 연결하려는 연구도 본격화되고 있다. LEO[29]는 3D 공간의 인식과 객체 위치 파악뿐만 아니라 추론, 계획 및 행동을 하나의 체화형 에이전트에 통합하여 공간정보에 대한 이해를 실제 행동으로 연계하는 접근을 제시하였다. 이후 RT-2[30]와 OpenVLA[31]를 비롯한 Vision-Language-Action(VLA) 모델은 시각적 관측과 자연어 지시를 로봇 행동과 결합함으로써, 멀티모달 환경 이해를 실제 물리적 행동으로 연결하는 범용 로봇 모델의 기반을 마련하였다.

최근 Google DeepMind의 Gemini Robotics[32]는 VLA를 기반으로 멀티모달 환경 인식과 공간 추론, 행동 생성을 통합하는 범용 로봇 지능으로 기술 범위를 확장하고 있다. 주변 환경과 객체 간 공간관계를 파악하고 복합적인 작업을 계획하여 실제 로봇의 조작 및 제어와 연계함으로써 공간 추론과 물리

적 행동 간의 연계성을 강화하였다. 이러한 VLA 연구는 개별 로봇 제어모델의 성능 향상뿐만 아니라, 공간에 대한 이해와 미래 상태의 예측, 물리적 추론을 실제 행동 및 상호작용으로 연결한다는 점에서 공간 운영 인텔리전스의 주요 기술적 발전 방향으로 볼 수 있다.

공간 운영 인텔리전스는 이와 같이 현재 공간의 구조와 객체 간 관계를 이해하는 데서 나아가, 시공간적 변화를 학습하고 미래 상태를 예측하며, 물리적 상황에 대한 추론과 행동 계획을 통해 실제 환경과 상호작용을 하는 지능으로 진화하고 있다. 향후에는 공간 이해, 예측, 추론 및 행동을 개별적으로 수행하기보다 이들 기능을 유기적으로 연계하고, 행동에 따른 공간의 변화를 다시 인식과 의사결정에 반영하는 통합형 공간지능으로 발전할 것으로 예상된다.

3. End-to-End 3D 공간지능 연구 동향

최근 한국전자통신연구원에서는 그림 1과 같이 공간 생성, 실시간 갱신 및 운영 인텔리전스를 통합하여 현실 공간의 상태를 지속적으로 이해하고 미래 변화를 예측하며, 이를 기반으로 자율적인 공간 운영을 지원하는 'End-to-End 3D 시공간 생성 · 운영 초지능 플랫폼' 연구개발을 진행하고 있다. 이를 위해 멀티모달 데이터 기반의 3D 공간 구축과 실시간 갱신 기술을 개발하고, 공간에 대한 인식과 이해, 미래 상태 예측 및 의사결정 기능을 유기적으로 연계하는 통합형 공간지능 플랫폼을 구현하고 있다.

본 연구에서는 현장 전역의 멀티모달 데이터를 정합하여 3D 월드 모델을 구축 · 갱신하고, 이를 공간과 상황 이해 및 미래 상태 예측에 활용하고자 한다. 아울러 현장 전체를 분석 · 판단하는 관제지능과 로봇 · 드론 등의 행동지능을 연계함으로써, 공간 이해와 예측 결과를 실제 의사결정 및 행동으로 연결하는 3D 시공간 운영 플랫폼을 구현하고 있다. 이를 통해 현실 공간의 변화를 지속적으로 반영하면서 상황 인식과 예측, 대응이 연계되는 지능형 공간 운영 체계를 구축하고, 산업 · 도시 · 재난 등 다양한 실환경에서의 적용 가능성을 검증할 예정이다.

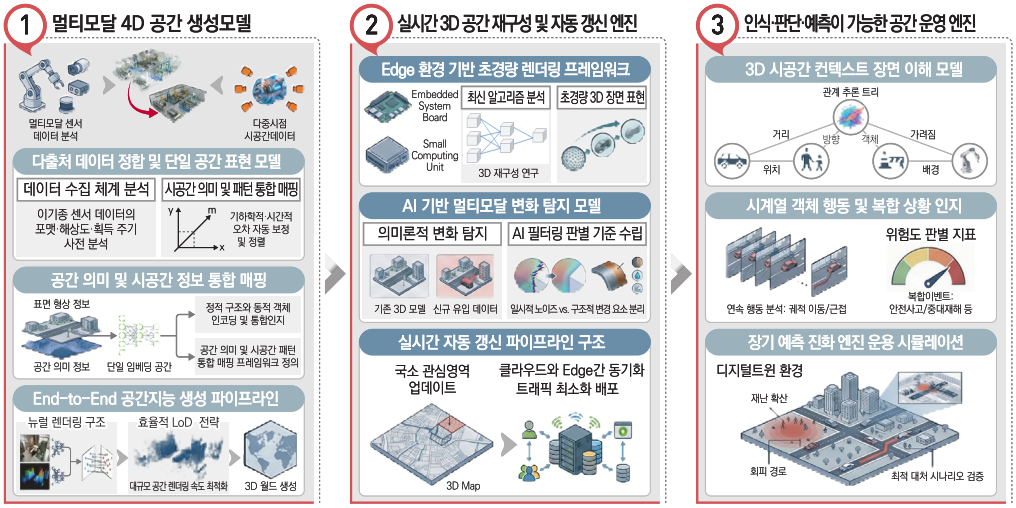

그림 1 End-to-End 3D 시공간 생성·운영 초지능 플랫폼 개념도

V. 결론

본고에서는 공간지능 구현을 위한 핵심 기술을 공간 생성 · 갱신 · 운영의 세 가지 관점으로 구분하여 최근 기술 동향을 살펴보았다. 공간 생성 기술은 텍스트, 이미지, 다중시점 영상 및 3D 데이터 등을 활용하여 개별 객체를 넘어 공간적으로 일관된 3차원 장면을 생성하는 방향으로 발전하고 있으며, 공간 갱신 기술은 3D Gaussian Splatting, 4D 공간 표현, 변화 탐지 및 실시간 Mapping 기술을 기반으로 현실 공간의 변화를 디지털 공간에 지속적으로 반영하는 방향으로 확장되고 있다. 또한, 공간 운영 기술은 공간 이해와 시공간적 표현 학습을 기반으로 미래 상태를 예측하고, 물리적 추론과 행동 계획을 거쳐 실제 물리세계와의 상호작용으로 연결되는 방향으로 발전하고 있다.

최근에는 이러한 공간 생성 · 갱신 · 운영 기술을 개별적으로 발전시키는 것을 넘어, 다양한 기능을 하나의 모델 또는 플랫폼에서 연계하여 통합하려는 월드 모델 연구가 본격화되고 있다. 대표적으로 World Labs의 Atlas, Google DeepMind의 Genie 3, NVIDIA의 Cosmos 3 등은 각각 3D 공간 생성과 표현, 동적 환경 시뮬레이션, 멀티모달 이해 · 생성 및 행동 연계 등을 중심으로 월드 모델의 기술 범위를 확장하고 있다. 특히 Cosmos 3와 같은 최신 모델은 다양한 멀티모달 정보를 통합적으로 처리하고 물리세계에 대한 이해와 예측을 행동으로 연계함으로써, 공간지능이 통합형 지능체계로 발전하고 있음을 보여준다.

향후에는 개별 요소기술의 성능 향상뿐만 아니라 이질적인 공간 표현 간의 정합성, 대규모 환경에서의 실시간 처리, 장기간 운영에 따른 공간정보의 일관성 및 물리적 신뢰성 확보가 중요한 기술적 과제가 될 것으로 예상된다. 또한, 실제 환경에서 자율적으로 동작하기 위해서는 관측과 행동의 결과를 다시 공간 모델과 의사결정 과정에 반영할 수 있는 폐루프 구조와 요소기술 간 유기적인 통합 기술이 요구된다.

궁극적으로 공간지능은 현실 공간을 단순히 디지털 환경에 재현하는 기술을 넘어, 물리세계에 대한 일관된 공간 표현을 토대로 상황을 스스로 판단하고 현실과 지속적으로 상호작용을 하는 지능체계로 발전할 것으로 전망된다. 이러한 기술은 향후 자율주행, 로봇, 스마트시티, 디지털 트윈 및 Physical AI 등 다양한 분야에서 현실세계와 디지털세계를 연결하고 자율적인 지능 서비스를 구현하기 위한 핵심 기반기술로 활용될 것으로 기대된다.

용어해설

공간 생성 멀티모달 데이터를 기반으로 현실 공간의 3D 구조와 의미 정보를 디지털 공간으로 생성·표현하는 기술

공간 갱신 현실 공간의 변화를 지속적으로 관측하여 디지털 공간에 반영하고 최신 상태를 유지하는 기술

공간 운영 생성·갱신된 공간의 상태와 관계를 이해하고 추론·예측하여 의사결정과 행동으로 연계하는 기술

참고문헌

[1] World Labs Team, "Atlas: A World Model for Spatial Intelligence," World Labs Blog, 2026. 9. 1. https://www.worldlabs.ai/blog/atlas

[2] J. Parker-Holder and S. Fruchter, "Genie 3: A New Frontier for World Models," Google DeepMind, 2025. 8. 5.

[3] N. Agarwal et al., "Cosmos World Foundation Model Platform for Physical AI," arXiv preprint, 2025, doi: 10.48550/arXiv.2501.03575.

[4] N. Agarwal et al., "Cosmos 3: Omnimodal World Models for Physical AI," arXiv preprint, 2026, doi: 10.48550/arXiv.2606.02800.

[5] B. Mildenhall et al., "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," in Proc. Eur. Conf. Comput. Vis., 2020, pp. 405-421.

[6] J. Tang et al., "DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation," in Proc. Int. Conf. Learn. Represent., (Vienna, Austria), 2024.

[7] T. Yi et al., "GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), 2024, pp. 6796-6807.

[8] Z. Chen et al., "Text-to-3D using Gaussian Splatting," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), 2024, pp. 21401-21412.

[9] X. Long et al., "Wonder3D: Single Image to 3D using Cross-Domain Diffusion," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), 2024, pp. 9970-9980.

[10] J. Tang et al., "LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation," in Proc. Eur. Conf. Comput. Vis., (Milan, Italy), 2024, pp. 1-18.

[11] H. Ling et al., "Align Your Gaussians: Text-to-4D with Dynamic 3D Gaussians and Composed Diffusion Models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), 2024, pp. 8576-8588.

[12] Y. Yang et al., "Prometheus: 3D-Aware Latent Diffusion Models for Feed-Forward Text-to-3D Scene Generation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Nashville, TN, USA), 2025, pp. 2857-2869.

[13] J. Xiang et al., "Structured 3D Latents for Scalable and Versatile 3D Generation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Nashville, TN, USA), 2025, pp. 21469-21480.

[14] B. Kerbl et al., "3D Gaussian Splatting for Real-Time Radiance Field Rendering," ACM Trans. Graph., vol. 42, no. 4, 2023, pp. 1-14.

[15] A. Gu don and V. Lepetit, "SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), 2024, pp. 5354-5363.

[16] N. Keetha et al., "SplaTAM: Splat, Track & Map 3D Gaussians for Dense RGB-D SLAM," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), 2024, pp. 21357-21366.

[17] G. Wu et al., "4D Gaussian Splatting for Real-Time Dynamic Scene Rendering," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), 2024, pp. 20310-20320.

[18] Z. Li et al., "Spacetime Gaussian Feature Splatting for Real-Time Dynamic View Synthesis," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), 2024, pp. 8508-8520.

[19] C. Zhang et al., "Efficiently Reconstructing Dynamic Scenes One D4RT at a Time," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Denver, CO, USA), 2026, pp. 7382-7392.

[20] V. Yugay et al., "Gaussian Mapping for Evolving Scenes," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Denver, CO, USA), 2026, pp. 18903-18912.

[21] E. Brachmann and V.A. Prisacariu, "Building a Large Geospatial Model to Achieve Spatial Intelligence," Niantic Spatial, 2024. 11. 12.

[22] Y. Hong et al., "3D-LLM: Injecting the 3D World into Large Language Models," Adv. Neural Inf. Process. Syst., vol. 36, 2023, pp. 20482-20494.

[23] B. Chen et al., "SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), 2024, pp. 14455-14465.

[24] W. Ma et al., "SpatialLLM: A Compound 3D-Informed Design towards Spatially-Intelligent Large Multimodal Models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Nashville, TN, USA), 2025, pp. 17249-17260.

[25] M. Assran et al., "V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning," arXiv preprint, 2025, doi: 10.48550/arXiv.2506.09985

[26] L. Mur-Labadia et al., "V-JEPA 2.1: Unlocking Dense Features in Video Self-Supervised Learning," arXiv preprint, 2026, doi: 10.48550/arXiv.2603.14482

[27] A. Azzolini et al., “Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning,” arXiv preprint, 2025, doi: 10.48550/arXiv.2503.15558

[28] NVIDIA, “Cosmos-Reason2,” NVIDIA Cosmos Documentation, 2025-2026. https://docs.nvidia.com/cosmos/latest/reason2/index.html

[29] J. Huang et al., “An Embodied Generalist Agent in 3D World,” in Proc. 41st Int. Conf. Mach. Learn., (Vienna, Austria), 2024, pp. 20413-20451.

[30] B. Zitkovich et al., “RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control,” in Proc. 7th Conf. Robot Learn., (Atlanta, GA, USA), 2023, pp. 2165-2183.

[31] M.J. Kim et al., “OpenVLA: An Open-Source Vision-Language-Action Model,” arXiv preprint, 2024, doi: 10.48550/arXiv.2406.09246

[32] Google DeepMind, “Gemini Robotics 2,” 2026.