생성형 AI 기반의 3D 입체영상 생성 및 지능형 저작 기술 동향

Trends in Generative AI-based Spatial Video Generation and Intelligent Authoring

김지완 (J.W. Kim, jiwankim@etri.re.kr) 콘텐츠융합연구실 선임연구원
박성진 (S.J. Park, seongjin.park@etri.re.kr) 콘텐츠융합연구실 책임연구원
정유구 (Y.G. Jung, u9jung@etri.re.kr) 콘텐츠융합연구실 선임연구원
장호욱 (H.W. Jang, hwjang@etri.re.kr) 콘텐츠융합연구실 책임연구원
김도형 (D.H. Kim, kdh99@etri.re.kr) 콘텐츠융합연구실 책임연구원

ABSTRACT

Recent advances in generative artificial intelligence (AI) have shifted digital content creation from manual post-editing to model-driven generation. Beyond 2D image synthesis, current research is increasingly focusing on the production of high-dimensional 3D stereoscopic video content. Although neural scene representations such as NeRF and 3D Gaussian Splatting enable the generation and rendering of raw 3D assets, intelligent authoring has emerged as a key technology for aligning these outputs with professional intent while preserving multiview spatiotemporal consistency.

This study surveys the end-to-end pipeline for intelligent 3D authoring and organizes it into five stages—that is, (1) raw generation (including retake and re-angle capabilities), (2) preprocessing (object recognition, segmentation, and tracking), (3) authoring (multimodal-guided editing using text, points, and poses), (4) postprocessing (inpainting and neural rendering-based restoration/compositing), and (5) integration into industry-standard tools and plugins. We summarize recent technical trends—such as optimization-free multiview structural propagation and multimodal precision control—which aim to reduce computational overhead and improve practical usability. Based on these developments, we outline future directions for scalable immersive content production enabled by AI-driven authoring tools for high-fidelity 3D stereoscopic experiences.

KEYWORDS 3D Authoring, 3D Editing, 3D Stereoscopic Video, Gaussian Splatting, Generative AI, Intelligent Authoring, Neural Rendering

* DOI: https://doi.org/10.22648/ETRI.2026.J.410201
* 공동 제1저자 김지완, 김도형
* 교신저자 김도형
* This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2022-0-00022, RS-2022-II220022, Development of immersive video spatial computing technology for ultra-realistic metaverse services).

I. 서론

최근 미디어 산업은 단순한 평면 영상 중심의 소비를 넘어, 사용자에게 높은 몰입감을 제공하는 실감형 콘텐츠로 무게중심이 빠르게 이동하고 있다. 이러한 변화 속에서 GAN(Generative Adversarial Networks)과 확산 모델(Diffusion Models)을 필두로 한 생성형 AI 기술의 비약적인 발전은 디지털 콘텐츠 제작 패러다임을 '사후 편집'에서 '원천 생성'으로 근본적으로 전환시키고 있다[1,2]. 특히 초기 생성형 AI가 2차원 이미지의 화질 개선이나 단순 합성에 국한되었던 것과 달리, 최신 기술은 텍스트나 이미지를 가이드로 삼아 고차원적인 3차원 공간 정보와 시간축이 결합된 4D 입체영상을 직접 생성하는 단계에 진입하였다[3,4].

이러한 생성형 AI 기반의 입체영상 생성 기술은 전통적인 촬영 및 그래픽 제작 방식이 가진 물리적 제약을 획기적으로 극복한다. 가상 공간 내에서 피사체의 위치나 조명을 자유롭게 재구성하는 재촬영(Retake) 기능이나, 물리적인 카메라의 이동 없이도 임의의 가상 시점을 정밀하게 생성해내는 시점 변경(Re-angle) 기술은 제작자에게 폭넓은 창작의 자유를 제공하고 있다[5]. 특히 3차원 가우시안 스플래팅(3DGS)이나 뉴럴 래디언스 필드(NeRF: Neural Radiance Field)와 같은 최신 3D 표현 기법의 등장은 실시간 렌더링 품질과 속도를 동시에 끌어올리며, 실감형 콘텐츠 제작의 실용성을 뒷받침하고 있다[6,7].

그러나 생성형 AI가 도출한 초기 결과물은 대규모 데이터의 확률적 분포를 기반으로 하므로, 제작자가 의도한 세밀한 예술적 디테일이나 특정 구조적 변경 사항을 완벽히 반영하기에는 여전히 한계가 존재한다. 이에 따라서 생성된 콘텐츠를 인간의 의도에 맞게 정밀하게 교정하고 최적화하는 지능형 저작(Intelligent Authoring) 기술이 생성형 AI 시대의 필수 보완책이자 핵심 경쟁력으로 대두되고 있다[8,9]. 또한 3D 입체영상은 시점 변화에도 불구하고 다중 시점 간 기하학적 일관성(Geometric Consistency)이 엄격히 유지되어야 하며, 좌우안 영상 간 불일치로 인한 시각적 피로도를 최소화해야 한다. 따라서 생성과 저작의 전 과정에서 고도의 시공간적 연속성 확보가 요구된다.

그간 학계와 산업계에서는 3D 공간상의 구조적 편집을 위해 다양한 시도를 해 왔으나, 대부분의 기존 방식은 시점 간 일관성을 확보하기 위해 장면당 막대한 최적화 연산(Per-Scene Optimization)을 요구해 왔다. 이러한 방식은 처리 시간이 길고 높은 컴퓨팅 자원을 필요로 하여 실시간성이 요구되는 제작 현장이나 모바일 등 자원이 제한적인 환경에 적용하기에는 큰 제약이 있었다[6,7]. 또한 2차원 확산 모델 기반의 편집 도구들은 개별 프레임의 품질은 우수하나, 다중 시점 관찰 시 구조적 표류(Structural Drift) 현상이 발생하는 등 입체영상 저작 관점에서 해결해야 할 난제들이 여전히 산적해 있다[10-12].

이에 본고에서는 생성형 AI를 활용하여 입체영상을 생성하는 원천 기술부터 이를 지능적으로 저작하기 위한 전처리 · 저작 · 후처리 및 통합 도구에 이르는 전 공정의 기술 동향을 체계적으로 살펴보고자 한다. 아울러 자유로운 시점 제어, 다중 시점 일관성, 연산 효율 등 실감형 제작 과정에서 반복적으로 제기되는 핵심 이슈들을 중심으로 관련 연구 흐름을 정리하고, 향후 입체영상 제작 시스템의 발전 방향을 조망하고자 한다[4,9,13].

II. 생성형 AI 기반 입체영상 생성 및 지능형 저작 파이프라인 구성

생성형 AI를 활용한 3D 입체영상 제작은 지능화된 모델을 통해 기초 에셋을 생성하는 과정과, 이를

제작자의 의도에 맞게 정밀하게 가공하는 저작 과정이 유기적으로 결합된 체계를 갖는다[14,15]. 이러한 공정 체계는 단순히 결과물을 도출하는 순서를 나열한 것이 아니라, 입체 콘텐츠 특유의 복잡한 기하학적 구조를 제어하고 시각적 피로도를 최소화하기 위한 기술적 근거를 바탕으로 구성된다. 특히 원천 데이터를 생성하는 단계에서부터 최종 결과물을 완성하는 후처리 단계까지 상호 보완적인 5개의 핵심 요소를 통해 실감형 콘텐츠의 완성도를 확보한다[16,17]. 그림 1은 생성형 AI 기반 입체영상 생성 및 지능형 저작 파이프라인을 5개 핵심 단계로 요약하여 전체 흐름과 단계 간 연계를 나타낸다.

워크플로우의 출발점인 생성형 AI 기반의 입체영상 생성 단계에서는 텍스트나 이미지 등 멀티모달 입력을 활용하여 고차원적인 3차원 공간 정보를 직접 합성한다[14,15,18,19]. 이 단계의 핵심은 생성형 모델의 잠재 공간(Latent Space)을 제어하여 실제 촬영 없이도 장면의 구도를 다시 구성하는 재촬영 효과를 구현하거나, 생성된 3DGS 또는 NeRF 표현체를 기반으로 자유로운 시점 변경 영상을 실시간으로 렌더링하는 데 있다. 생성된 장면을 정교하게 다듬기 위해 이어지는 지능형 전처리 단계에서는 장면 내 객체를 시각적으로 인지하고 배경으로부터 분리하며, 시점 간 객체 이동을 추적함으로써 저작의 대상이 되는 시맨틱 영역을 정의한다[20,21]. 이러한 객체 단위의 해석 기술은 이후 편집 과정에서 사용자가 원하는 부분만을 선택적으로 제어할 수 있는 기술적 토대가 된다[8,10].

실질적인 저작이 수행되는 지능형 저작 기능 단계에서는 텍스트, 인체 포즈, 포인트 드래깅 등의 가이드를 활용하여 생성된 객체의 형태나 질감을 맥락 기반(Context-aware)으로 수정한다[10-12,22]. 이때 다중 시점 일관성을 확보하는 과정에서 연산 비용이 커지는 문제가 지속적으로 지적되어 왔으며, 최근에는 이러한 부담을 줄이기 위한 다양한 접근이 제안되고 있다[9,13]. 저작 완료 이후에는 편집 과정에서 발생할 수 있는 왜곡이나 비가시 영역을 자연스럽게 보간하는 지능형 후처리 기능이 수행된다[16,17,23]. 영역 복원 및 합성을 통해 시점 간 이질감을 제거함으로써 입체영상 제작의 최종적인 시각적 완성도를 검증한다[24].

마지막으로 이러한 생성 및 저작 기술들을 단일 워크플로우 내에서 직관적으로 활용할 수 있도록 제공하는 저작 도구 환경이 구축되어야 한다[25-27]. 저작 도구는 생성형 AI의 강력한 콘텐츠 생성 능력과 지능화된 편집 기능을 유기적으로 연결하고 제작자에게 실시간 피드백을 제공함으로써, 전문가부터 일반 사용자까지 고품질의 실감 콘텐츠를 효율적으로 생산할 수 있는 인터페이스를 보장한다[26,28,29]. 결론적으로 이러한 계층적 공정 체계는 복잡한 입체영상 제작의 진입 장벽을 낮추고 산업 전반의 생산성을 혁신하는 기술적 근거가 된다.

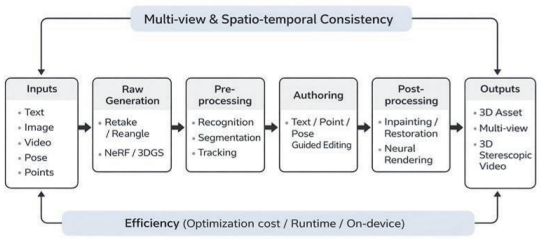

그림 1 생성형 AI 기반 입체영상 생성 및 지능형 저작 파이프라인

III. 생성형 AI 기반 입체영상 생성 및 지능형 저작 기술

1. 생성형 AI 기반의 입체영상 생성

생성형 AI를 활용한 입체영상 생성은 단순한 이미지 합성을 넘어, 3차원 공간 정보와 시간축을 결합한 고차원적인 데이터 생성 단계로 확장되고 있

다[3,4,30]. 초기 생성형 모델이 2차원의 시각적 유사성에 집중했다면, 최신 기술은 잠재 공간(Latent Space) 내에서 3차원 구조를 보다 명시적으로 모델링하고 이를 바탕으로 시점 변화에 견고한 입체영상을 합성하는 데 초점을 맞춘다[14,15]. 특히 최근의 생성형 3D 파이프라인은 전통적인 제작 방식의 제약을 완화하고, 가상공간에서 자유로운 시점 제어를 가능하게 함으로써 실감 콘텐츠 제작의 새로운 지평을 열고 있다[5,30].

1.1 3차원 생성 모델 기술 동향

최근 3차원 생성 기술은 텍스트나 이미지를 가이드로 삼아 고품질의 3D 에셋을 생성 · 복원하는 방향으로 빠르게 발전하고 있다[14,15,18,19]. 초기에는 점구름(Point Cloud)이나 메쉬(Mesh) 형태의 직접 생성이 주를 이루었으나, 현재는 대규모 확산 모델의 사전 지식을 3차원으로 전이하는 SDS (Score Distillation Sampling) 계열 접근이 중요한 흐름으로 자리 잡았다. 대표적으로 DreamFusion[14]과 Magic3D[15]는 텍스트 프롬프트로부터 고해상도의 3차원 기하구조와 텍스처를 생성하는 데 성과를 보였으며, 최근에는 단일 시점 이미지로부터 수초 이내에 정교한 3D 모델을 복원하는 LGM[18]이나 TripoSR[19]과 같은 고속 생성 모델도 제안되고 있다. 이러한 기술은 입체영상 제작 시 필요한 원천 데이터를 방대한 수작업 없이 확보할 수 있도록 제작 효율을 높인다.

1.2 방사 필드 및 가우시안 기반 입체 생성

입체영상의 시각적 충실도와 렌더링 속도를 확보하기 위해 NeRF와 3DGS 기술이 핵심적인 역할을 수행하고 있다. NeRF는 좌표 기반의 신경망 학습을 통해 연속적인 공간 표현을 가능하게 하여 고품질의 뷰 합성(View Synthesis) 성과를 보여주었으며[6], 3DGS는 명시적인 포인트 기반 렌더링 방식을 통해 높은 품질을 유지하면서도 실시간에 가까운 렌더링을 가능하게 하였다[7]. 특히 생성형 AI와 결합된 GaussianDreamer[31]와 같은 연구는 3DGS의 효율적인 표현력을 활용하여 객체와 배경을 신속하게 생성하고, 시점 변화에도 기하학적 일관성이 유지되는 입체 데이터 생성을 지향한다.

1.3 재촬영 및 시점 제어 기술

생성형 AI 기반 입체영상의 주요 차별점은 물리적인 카메라 움직임의 제약 없이도 장면을 재구성할 수 있다는 점이다[5]. 가상공간에서 피사체의 위치나 조명을 생성 단계에서 다시 설정하는 재촬영(Retake) 기능은 제작자가 사후에도 연출 의도를 조정할 수 있는 유연성을 제공한다. 또한 생성형 모델의 시점 추론 능력을 바탕으로, 촬영되지 않은 가상의 경로를 따라 카메라가 이동하는 것처럼 시점을 생성하는 시점 변경 기술은 입체영상의 표현 범위를 확장한다. 최근에는 영상 확산 모델을 3D 구조와 결합하여 카메라 궤적에 따른 시점 변화를 보다 일관되게 생성하려는 SV3D[35]나 VEO[30] 등의 시도가 제시되면서, 고가의 입체 촬영 장비 없이도 시네마틱한 입체영상 제작 가능성을 보여주고 있다.

2. 지능형 저작을 위한 전처리 기술

지능형 저작을 위한 전처리 기술은 영상 분야에서 객체 인지, 객체 분리, 객체 추적으로 구분된다[20,32]. 최근 객체 인지 기술은 단순한 2D 이미지 분석을 넘어, 시간적 맥락 정보와 대규모 파운데이션 모델을 결합하는 방향으로 빠르게 발전하고 있다[21,33]. 객체 인지가 “무엇이 어디에 있는가”를 찾는 기술이라면, 객체 분리는 객체의 정확한 경계와 형태를 픽셀 수준에서 추정하는 더 정교한 기술

이다. 또한 객체 추적은 객체를 인지 · 분리하는 것을 넘어, 시간의 흐름 속에서 동일 객체의 정체성을 지속적으로 유지하도록 하는 기술로 이해할 수 있다[30].

2.1 객체 인지 기술 동향

과거에는 CNN(Convolutional Neural Network) 기반 모델들이 주류를 이루었으나, 최근에는 자연어 처리에서 발전한 트랜스포머(Transformer) 구조가 영상 내 객체 인지 분야의 대표적 접근으로 자리 잡고 있다[33]. ViT(Vision Transformer)[36]는 이미지를 패치 단위로 분할해 처리함으로써 전역 맥락을 효과적으로 반영할 수 있다. DETR은 앵커 박스나 NMS와 같은 복잡한 후처리 없이 객체 인지를 세트 예측(Set Prediction) 문제로 해결하며[37], 이후 RT-DETR 등 실시간 성능을 강화한 변형 모델도 제안되었다[38].

또한 특정 클래스만을 대상으로 학습하던 방식에서 벗어나, 학습되지 않은 객체도 텍스트 프롬프트로 탐지하는 개방형 어휘(Open-Vocabulary) 탐지가 활발히 연구되고 있다[20,39]. CLIP(Contrastive Language-Image Pre-Training)[20], Grounding DINO[39] 등은 텍스트 명령을 이해하고 영상 내에서 해당 객체의 위치를 찾는 기반 기술로 널리 활용된다.

한편 산업 현장이나 엣지 디바이스에서의 활용을 위해 정확도와 속도의 균형을 맞춘 모델들도 지속적으로 발전하고 있다. YOLO 계열은 대표적인 실시간 탐지 모델로, 다양한 변형 모델이 꾸준히 발표되고 있다[40].

2.2 객체 분리 기술 동향

SAM[21]이 객체 분리 분야에 큰 영향을 준 이후, SAM2[41]는 이미지뿐만 아니라 비디오까지 확장된 프롬프트 기반 분할을 제시하며 시간 축에서의 일관성 확보를 주요 과제로 다룬다. 또한, 사용자가 점/박스/텍스트 등의 프롬프트를 입력하면 해당 객체를 픽셀 단위로 분리해내는 인터랙티브 세그멘테이션이 널리 활용되며 사실상 표준적 인터페이스로 자리 잡고 있다.

과거에는 배경(Stuff)과 사물(Thing)을 각각 다루는 분할 기술이 별도로 발전해 왔으나, 최근에는 이를 통합적으로 처리하는 팬옵틱 분리(Panoptic Segmentation)가 주요 흐름으로 자리 잡고 있다. Mask2Former[34], OneFormer[42] 등은 하나의 프레임워크에서 다양한 분할 과제를 통합하는 방향을 제시하며, 이러한 흐름은 의료영상 등 특화 도메인으로도 확장되어 MedSAM[43]과 같은 의료 특화 파운데이션 모델로 발전하고 있다.

2.3 객체 추적 기술 동향

객체 분리에서 촉발된 파운데이션 모델 기반의 흐름은 추적 분야로도 확장되고 있으며, SAM2[41]와 같은 모델은 비디오 환경에서 분할과 추적을 통합적으로 다루는 방향을 중요한 추세로 제시한다. 한편 자율주행 및 로봇을 위한 3D 멀티모달 추적에서는 카메라-라이다 등 이종 센서의 융합과 3D 공간 표현이 핵심 과제로 주목받는다. BEVFusion[44]은 특징(Feature) 단계에서의 융합을 통해 효율적인 성능을 보인 대표적 사례로 널리 인용된다.

3. 생성형 AI 기반 지능형 저작 기술

생성형 AI 기반의 3D 입체영상 저작 기술은 초기 프롬프트에 의존하던 단순 생성 단계를 넘어, 사용자의 구체적인 의도를 반영해 결과물을 수정 · 보정하는 지능형 저작 단계로 고도화되고 있다[8,9]. 지능형 저작 기술은 텍스트, 포인트, 포즈, 스케치 등 멀티모달 가이드를 활용하여 생성된 결과물의 기하학적 구조나 시각적 속성을 정밀하게 제어하는 기

술을 의미한다[10,11,22]. 특히 3차원 입체영상은 시점 변화에 따른 시공간적 일관성 유지가 필수적이므로, 기존 2D 편집 기술을 그대로 적용하기 어렵고 입체 콘텐츠에 특화된 제어 메커니즘이 요구된다.

3.1 가이드 유형별 저작 기술 동향

포인트 드래깅(Point Dragging) 기술은 사용자가 객체의 특정 지점(Handle Point)을 선택한 뒤 원하는 위치(Target Point)로 이동시켜 직관적으로 형상을 변형하는 방식이다. DragGAN[11]과 DragDiffusion[12]으로 촉발된 이 흐름은 최근 3DGS 및 NeRF와 결합하는 방향으로 확장되고 있으며, 3D 공간에서의 드래깅 기반 저작을 다루는 Dragin3D[45]와 같은 연구도 보고되고 있다. 또한, Drag-Your-Gaussian (DYG)[46]은 3DGS 기반에서 드래그 신호를 이용해 기하 편집을 정밀하게 제어하려는 시도를 보여준다.

자연어 프롬프트를 활용한 저작은 사용자가 전문적인 DCC(Digital Content Creation) 도구 없이도 고차원적인 편집을 수행할 수 있도록 한다. 최근에는 확산 기반 시각-언어 모델을 활용하여 객체의 부분적 속성을 이해하고 수정하는 시맨틱 편집 기술이 활발히 연구되고 있다[8,10]. GaussianEditor[9]는 가우시안 스플래팅 표현에서 텍스트 쿼리를 통해 특정 객체를 분리하고, 해당 영역의 텍스처 변경이나 객체 삭제 및 보간 등의 편집을 수행한다. 이는 입체영상 제작 과정에서 배경 요소를 제거하거나 특정 오브젝트의 재질을 변경하는 등 다양한 작업을 더욱 손쉽게 수행할 수 있게 한다.

입체영상 내 캐릭터나 동적 객체를 제어하기 위해 스켈레톤(Skeleton), 깊이 맵, 포즈 정보 등을 가이드로 활용하는 기술도 중요하다. 이러한 구조 가이드는 생성 과정에서 발생할 수 있는 형태적 왜곡을 완화하고, 보다 자연스러운 동작 생성을 지원하는 데 이바지한다. ControlNet[10]의 3D 확장 계열 및 MimicMotion[22] 등의 연구는 참조 비디오나 포즈 시퀀스를 입력으로 받아 동작을 생성 · 제어하는 방향을 제시한다. 또한 최근에는 다중 시점 일관성을 유지하기 위한 학습 기반 메커니즘이 함께 논의되며, 입체영상 재생 시 좌우안 간 불일치로 인한 시각적 피로를 완화하는 데 목적을 두고 있다.

3.2 기하학적 일관성 확보를 위한 3D 저작

3D 장면 편집의 핵심 과제는 여러 각도에서 관찰했을 때 시각적 모순이 없는 기하학적 일관성을 유지하는 것이다. 초기 연구들은 주로 NeRF나 3DGS와 같은 3차원 표현체를 직접 최적화하는 방식을 취해 왔다. 예를 들어 DYG[46]는 스코어 증류 샘플링을 활용해 3DGS의 기하학적 구조를 변형하는 접근을 제시하며, GaussianEditor 및 EditSplat[9,13]은 반복적인 최적화 과정을 통해 외형과 텍스처를 수정함으로써 시점 간 일관성을 확보하고자 한다. 그러나 이러한 최적화 기반 방식은 장면당 연산 시간과 비용이 많이 증가할 수 있어, 실시간 저작 환경에 적용하기에는 제약이 존재한다.

이러한 한계를 완화하기 위해, 최근에는 단안 깊이(Monocular Depth) 정보 등 기하학적 단서를 매개로 활용하여 최적화 의존도를 낮추고 편집 결과를 다중 시점으로 효율적으로 확장하려는 접근도 논의되고 있다. 이러한 흐름은 향후 실시간성 및 제작 효율이 요구되는 입체영상 저작 시스템에서 중요한 연구 방향으로 이어질 것으로 기대된다.

4. 지능형 저작 후처리 기술

생성형 AI를 이용해 3D 입체영상을 제작할 때, 초기 결과물은 구조적 결함이나 노이즈를 포함하는 경우가 많다. 대표적인 결함으로는 좌 · 우 시점 불일치, 가림(Occlusion) 영역에서의 홀(Hole) 발생, 프

레임 간 깜빡임(Flicker), 객체 경계 주변의 깊이 오류, 색 · 노출 불일치 등이 있으며, 이러한 오류들을 제작 현장에서 활용할 수 있는 수준으로 보정 · 개선하는 과정을 지능형 저작 후처리 기술이라 한다. 지능형 저작 후처리 기술은 단순한 영상 보정을 넘어 생성된 데이터를 안정화하고 품질을 향상시키는 방향으로 발전하고 있으며[16,17,23], 실시간 신경 렌더링 후처리, 멀티모달 기반 정밀 제어, 온-디바이스 AI 가속 기술이 핵심 요소로 부각되고 있다.

4.1 실시간 신경 렌더링 후처리 기능

실시간 신경 렌더링 후처리 기술은 화면을 선명하게 만드는 수준을 넘어, 전통적인 그래픽스의 물리 기반 연산과 AI 기반 추론을 결합하여 생성된 3D 영상의 품질을 실시간으로 보정 · 최적화하는 방향으로 고도화되고 있다. 기존 CNN 기반 후처리 방식은 빠르게 움직이는 입체영상에서 잔상이나 깜빡임이 발생할 수 있다는 한계가 지적되어 왔으며, 최근에는 프레임 전역의 맥락과 픽셀 간 상관관계를 함께 분석해 시공간적 안정성을 높이는 Vision Transformer(ViT) 기반 복원 접근이 활발히 연구되고 있다. 대표적으로 SwinIR[16], Video Restoration Transformer(VRT)[17], ACT[47] 등이 보고되었다. 한편 NVIDIA DLSS[24]는 레이트레이싱 과정에서 발생하는 노이즈를 완화하기 위해 학습 기반의 재구성(Reconstruction) 방식을 도입하여, 제한된 광선 샘플로부터 보다 안정적인 렌더링 결과를 얻는 방향을 제시한다. 또한 Pointersect[48], Ray-Distance Volume Rendering[49] 등은 포인트 기반 표현을 활용한 렌더링 및 재구성 관점에서 참고할 만한 연구로 언급될 수 있다.

4.2 멀티모달 기반의 정밀 제어

멀티모달 기반 정밀 제어 기술은 텍스트 프롬프트의 한계를 넘어 이미지, 비디오, 스케치, 오디오, 깊이 맵 등 다양한 입력을 조합하여 사용자의 의도를 더욱 정확하게 반영하는 방향으로 발전하고 있다[10]. 이는 단순히 “무엇을 생성하는가”를 넘어, 특정 부위의 질감이나 객체의 이동 경로, 상호작용 양상을 보다 정교하게 제어하려는 흐름으로 확장되고 있다. 단일 프레임 편집을 비디오로 일관되게 확장하는 접근으로 CoDeF(Content Deformation Field)[23]가 대표적이며, 객체/영역 단위의 모션을 정밀하게 제어하는 연구로 MotionPro[50]가 보고되었다. 또한 후처리 단계에서 멀티모달 LLM이 생성된 입체영상을 점검하고, 일관성이 깨진 구간을 탐지해 추가 보정을 유도하는 피드백 루프 방식도 제안되고 있다[51].

4.3 On-Device AI 가속 기술

온-디바이스 AI 가속 기술은 클라우드 의존 없이 단말기 자체에서 인공지능 연산을 수행함으로써, 입체영상 저작과 같이 연산량이 큰 작업에서 저지연과 개인정보 보호를 동시에 달성하기 위한 핵심 수단으로 주목받고 있다. 애플 M5는 고성능 NPU를 기반으로 온-디바이스에서의 추론 성능을 강화하는 방향이 제시되었으며[52], 메모리와 연산 유닛을 통합한 PIM(Processor-In-Memory) 기술은 데이터 이동에 따른 병목과 전력 소모를 줄이기 위한 대안으로 연구되고 있다[53]. 또한, 모델 경량화 관점에서 지식 증류와 모델 압축 기법은 온-디바이스 배포를 뒷받침하는 핵심 기법으로 널리 인용된다[54,55].

5. 지능형 저작 도구 및 플러그인

생성형 AI 기술의 급격한 발전으로 3차원 콘텐츠 제작 파이프라인은 수작업 모델링 중심의 방식에서 프롬프트 기반 생성 및 지능형 편집 중심의 방식으

로 전환되는 과도기에 있다[25,27]. 이 분야의 지능형 저작 기술은 단순히 텍스트를 3차원 형상으로 변환하는 수준을 넘어, 기존 상용 그래픽 소프트웨어 및 엔진(Unity, Unreal, Blender 등)과 결합하여 텍스처링, 리깅, 애니메이션, 인페인팅 등 제작 전 과정을 자동화 · 효율화하는 제반 기술을 포괄한다.

5.1 상용 소프트웨어 내 지능형 저작 기술

전통적인 그래픽 툴 제조사들은 생성형 AI 모델을 자사 제작 파이프라인에 통합하고 있다. Adobe는 Substance 3D에 Firefly 모델을 연동하여 텍스트 프롬프트만으로 고해상도 재질(Material)을 생성하고 3차원 모델에 적용하는 기능을 제공한다[25]. 또한, After Effects와 Premiere Pro 등에서는 생성형 채우기(Generative Fill) 계열 기능을 통해 2차원 영상 편집 과정에서 객체 제거 · 복원과 같은 인페인팅 기반 보조 기능을 지원한다[26].

캐릭터 애니메이션 분야에서도 수작업의 비중을 줄이기 위한 기술이 도입되고 있다. NVIDIA Omniverse의 Audio2Face-3D[28]는 오디오 파형을 분석하여 3차원 캐릭터의 입 모양과 표정을 자동 생성함으로써 일부 키프레임 작업을 대체할 수 있는 가능성을 제시한다. Autodesk 역시 Maya와 3ds Max에 AI 기반 오토 리깅(Auto-Rigging) 및 모션 생성 기능을 실험적으로 도입하며 저작 보조 도구로써의 활용 범위를 확장하고 있다.

5.2 3차원 생성 및 편집 플러그인

생성형 AI 기반 3차원 생성 · 편집 기술은 상용 엔진에서 활용할 수 있는 플러그인 형태로도 빠르게 구현 · 확산되고 있다. DreamFusion[14]의 SDS(Score Distillation Sampling) 계열 접근을 바탕으로 한 Luma AI, Meshy 등의 서비스는 Unreal Engine 및 Unity 연동 워크플로우를 제공하여, 텍스트나 이미지를 기반으로 생성된 3차원 메쉬를 엔진 에셋으로 활용할 수 있도록 지원한다. 오픈소스 생태계인 Blender에서는 Dream Textures[29]와 같이 뷰포트 환경에서 깊이 정보를 활용해 AI 이미지를 3차원 표면에 투사(Projection)하는 방식이 활용된다. 또한 PhysGaussian[56]은 3DGS 객체에 탄성, 마찰 등 물리 속성을 학습시켜 단순 시각화를 넘어 상호작용 및 물리 시뮬레이션과의 결합 가능성을 제시한다.

국내의 한국전자통신연구원(ETRI)에서는 실사와 CG(Computer Graphics) 합성 및 무안경 입체영상 가시화 등 입체영상 제작에 필요한 핵심 요소 기술을 고도화하려는 연구 개발이 진행되고 있다. 이러한 기술들은 상용 그래픽스 툴 기반 플러그인 형태로 구현되어, 실사 기반 3차원 입체영상에 가상의 CG 모델을 정교하게 증강 · 합성할 수 있는 저작 환경을 제공하는 방향으로 활용될 수 있다. 특히 3차원 공간 정보를 유지한 상태에서 합성이 수행됨으로써 입체감이 보존된 콘텐츠 제작에 이바지할 수 있다. 표 1은 앞서 논의한 전처리 · 저작 · 후처리 단계에

표 1 입체영상 저작 공정의 주요 문제와 대응 기술

주요 문제	대응 기술
다중 시점 기하 불일치	• 3D 표현체(예: 3DGS) 기반 편집 • 다시점 정합·일관성 제약 • 기하 단서(깊이 등) 활용 확장
좌·우안 불일치	• 좌·우 시점 정합 • Disparity·Occlusion 안정화 • 불일치 영역 탐지·보정
가림 영역 및 비가시 결손	• 인페인팅 기반 복원 • 계층적 합성(객체·배경 분리) • 신경 렌더링 보간
시간적 불안정	• 시간 일관성 제약 • 추적·워핑 기반 안정화 • 비디오 복원 모델 적용
깊이/경계 아티팩트	• 정밀 분리·추적 • 깊이 정렬·보정 • 경계 인지형 합성·복원
색·노출·재질 불일치	• 색·노출 정규화 • 참조 기반 스타일 전이 • 멀티모달 가이드 기반 정밀 제어

서 빈번히 발생하는 주요 품질 저하 요인과 대표적 대응 기술을 정리한 것이다.

IV. 결론

생성형 AI 기술의 급격한 성장은 3차원 실감 콘텐츠 제작 방식을 전통적인 수작업 모델링 중심에서 AI 기반의 자동화 · 지능화 파이프라인으로 빠르게 전환시키고 있다. 본고에서 살펴본 바와 같이, 최신 연구와 산업 동향은 단순한 3D 에셋 생성 단계를 넘어 사용자의 의도를 반영해 결과물을 보정 · 정교화하는 '지능형 저작' 단계로 확장되는 흐름을 보여준다. 특히 입체영상은 다중 시점 간 기하학적 일관성과 시공간적 안정성이 품질을 좌우하므로, 생성 · 저작 · 후처리 전 과정에서 이를 안정적으로 확보하기 위한 기술적 진전이 중요한 의미가 있다.

파이프라인 관점에서 보면, 전처리 단계에서는 파운데이션 모델 기반의 객체 인지 · 분리 · 추적 기술이 성숙하며 저작 대상 영역을 보다 정확히 정의할 수 있게 되었다. 저작 단계에서는 텍스트, 포인트, 포즈 등 멀티모달 가이드를 활용한 편집이 고도화되는 한편, 다중 시점 일관성 확보 과정에서 발생하는 연산 부담을 완화하기 위한 효율화 접근이 지속적으로 제안되고 있다. 또한 후처리 단계에서는 신경 렌더링 기반 복원과 품질 향상, 그리고 멀티모달 입력을 통한 정밀 제어가 결합되면서 생성 결과물의 결함을 완화하고 시각적 완성도를 높이는 방향으로 발전하고 있다.

향후 지능형 저작 기술은 시각적 속성의 수정에 그치지 않고, 물리적 타당성과 상호작용을 고려하는 '물리-인지형 저작(Physics-aware Authoring)'으로 확장될 가능성이 크다. 동시에 개인정보 보호와 지연 시간 요구가 높은 실감형 서비스의 특성상, 온-디바이스 AI 가속 및 모델 경량화 기술과의 결합도 중요한 축이 될 것이다. 궁극적으로 이러한 기술들은 상용 DCC 도구 및 게임 엔진과의 연동, 플러그인 형태의 배포, 그리고 VR/AR 및 무안경 입체영상 디스플레이 환경으로의 적용을 통해, 고품질 실감 콘텐츠를 더욱 효율적으로 제작 · 유통할 수 있는 생태계 형성에 이바지할 것으로 기대된다.

참고문헌

[1] I. Goodfellow et al., “Generative adversarial networks,” Commun. ACM, vol. 63, no. 11, 2020, pp. 139-144.

[2] J. Ho et al., “Denoising diffusion probabilistic models,” Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 6840-6851.

[3] J. Ho et al., “Video diffusion models,” Adv. Neural Inf. Process. Syst., vol. 35, 2022, pp. 8633-8646.

[4] Y. Xie et al., “Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency,” arXiv preprint, 2024. doi: 0.48550/arXiv.2407.17470

[5] D.J. Zhang et al., “Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning,” in Proc. Conf. Comput. Vis. Pattern Recognit., (Nashville, TN, USA), June 2025.

[6] B. Mildenhall et al., “Nerf: Representing scenes as neural radiance fields for view synthesis,” Commun. ACM, vol. 65, no. 1, 2021, pp. 99-106.

[7] B. Kerbl et al., “3D Gaussian splatting for real-time radiance field rendering,” ACM Trans. Graph., vol. 42, no. 4, 2023, pp. 1-14.

[8] T. Brooks et al., “Instructpix2pix: Learning to follow image editing instructions,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Vancouver, Canada), June 2023.

[9] Y. Chen et al., “Gaussianeditor: Swift and controllable 3d editing with gaussian splatting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), June 2024.

[10] L. Zhang et al., “Adding conditional control to text-to-image diffusion models,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., (Paris, France), Oct. 2023, pp. 3813-3824.

[11] X. Pan et al., “Drag your gan: Interactive point-based manipulation on the generative image manifold,” in Proc. ACM SIGGRAPH Conf., (Los Angeles, CA, USA), Aug. 2023, pp. 1-11.

[12] Y. Shi et al., “Dragdiffusion: Harnessing diffusion models for interactive point-based image editing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), June 2024.

[13] D.I. Lee et al., “Editsplat: Multi-view fusion and attention-guided optimization for view-consistent 3d scene editing with 3d gaussian splatting,” Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

[14] B. Poole et al., “Dreamfusion: Text-to-3d using 2d diffusion,” arXiv preprint, 2022. doi: 10.48550/arXiv.2209.14988

[15] C.H. Lin et al., “Magic3d: High-resolution text-to-3d content creation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Nashville, TN, USA), June 2023.

[16] J. Liang et al., “Swinir: Image restoration using swin transformer,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, (Montreal, Canada), Oct. 2021, pp. 1833-1844.

[17] J. Liang et al., “Vrt: A video restoration transformer,” IEEE Trans. Image Process., vol. 33, 2024, pp. 2171-2182.

[18] J. Tang et al., “Lgm: Large multi-view gaussian model for high-resolution 3d content creation,” in Proc. Eur. Conf. Comput. Vis., (Milan, Italy), Sep. 2024.

[19] D. Tochilkin et al., “Triposr: Fast 3d object reconstruction from a single image,” arXiv preprint, 2024. doi: 10.48550/arXiv.2403.02151

[20] S. Liu et al., “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” in Proc. Eur. Conf. Comput. Vis., (Milan, Italy), Sep. 2024.

[21] A. Kirillov et al., “Segment anything,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., (Paris, France), Oct. 2023, pp. 3992-4003.

[22] Y. Zhang et al., “Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance,” arXiv preprint, 2024. doi: 10.48550/arXiv.2406.19680

[23] H. Ouyang et al., “Codef: Content deformation fields for temporally consistent video processing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), June 2024.

[24] NVIDIA Corporation, “DLSS 4: Transforming Real-Time Graphics with AI,” https://research.nvidia.com/labs/adli/DLSS4/, Technical Report, 2025.

[25] https://blog.adobe.com/en/publish/2024/03/18/adobe-announces-firefly-powered-features-substance-3d-apps

[26] https://blog.adobe.com/en/publish/2024/10/14/generative-extend-in-premiere-pro

[27] https://www.fab.com/listings/b52460e0-3ace-465e-a378-495a5531e318

[28] https://docs.nvidia.com/ace/audio2face-3d-microservice/latest/text/getting-started/overview.html

[29] https://www.meshy.ai/blog/meshy-unity-texture

[30] https://deepmind.google/models/veo/

[31] 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), June 2024.

[32] R. Girshick, “Fast r-cnn,” in Proc. IEEE Int. Conf. Comput. Vis., (Santiago, Chile), Dec. 2015, pp. 1440-1448.

[33] A. Vaswani et al., “Attention is all you need,” Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[34] B. Cheng et al., “Masked-attention mask transformer for universal image segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (New Orleans, LA, USA), June 2022, pp. 1280-1289.

[35] V. Voleti et al., “Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion,” in Proc. Eur. Conf. Comput. Vis., (Milan, Italy), Sep. 2024.

[36] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv preprint, 2020. doi: 10.48550/arXiv.2010.11929

[37] N. Carion et al., “End-to-end object detection with transformers,” in Proc. Eur. Conf. Comput. Vis., (Glasgow, UK), Aug. 2020, pp. 213-229.

[38] Y. Zhao et al., “Detrs beat yolos on real-time object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), June 2024.

[39] A. Radford et al., “Learning transferable visual models from natural language supervision,” International conference on machine learning July 2021.

[40] A. Wang et al., “Yolov10: Real-time end-to-end object detection,” Adv. Neural Inf. Process. Syst., vol. 37, 2024, pp. 107984-108011.

[41] N. Ravi et al., “Sam 2: Segment anything in images and videos,” arXiv preprint, 2024. doi: 10.48550/arXiv.2408.00714

[42] J. Jain et al., “Oneformer: One transformer to rule universal image segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Vancouver, Canada), June 2023.

[43] J. Ma et al., “Segment anything in medical images,” Nat. Commun., vol. 15, no. 654, 2024.

[44] Z. Liu et al., “Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation,” arXiv preprint, 2022. doi: 10.48550/arXiv.2205.13542

[45] W. Guang et al., “Dragin3D: Image Editing by Dragging in 3D Space,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Nashville, TN, USA), June 2025, pp. 21502-21512.

[46] Y. Qu et al., “Drag your gaussian: Effective drag-based editing with score distillation for 3d gaussian splatting,” in Proc. Spec. Interest Group Comput. Graph. Interact. Techn. Conf. Conf. Papers, (Vancouver, Canada), Aug. 2025, pp. 1-12.

[47] J. Liang et al., “Recurrent video restoration transformer with guided deformable attention,” Adv. Neural Inf. Process. Syst., vol. 35, 2022, pp. 378-393.

[48] J.H.R. Chang et al., “Pointersect: Neural rendering with cloud-ray intersection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Vancouver, Canada), June 2023.

[49] R. Yin et al., “Ray-distance volume rendering for neural scene reconstruction,” in Proc. Eur. Conf. Comput. Vis., (Milan, Italy), Sep. 2024.

[50] Z. Zhang et al., “MotionPro: A Precise Motion Controller for Image-to-Video Generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Nashville, TN, USA), June 2025, pp. 27957-27967.

[51] D.X. Long et al., “VISTA: A Test-Time Self-Improving Video Generation Agent,” arXiv preprint, 2025. doi: 10.48550/arXiv.2510.15831

[52] https://www.apple.com/kr/newsroom/2025/10/apple-unleashes-m5-the-next-big-leap-in-ai-performance-for-apple-silicon/

[53] https://www.sciencedirect.com/science/article/pii/S2773064622000160

[54] G. Hinton et al., “Distilling the knowledge in a neural network,” arXiv preprint, 2015. doi: 10.48550/arXiv.1503.02531

[55] S. Han et al., “Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding,” arXiv preprint, 2015. doi: 10.48550/arXiv.1510.00149

[56] T. Xie et al., “Physgaussian: Physics-integrated 3d gaussians for generative dynamics,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., (Seattle, WA, USA), June 2024.