Journal of Digital Art Engineering & Multimedia Vol.13, No.2, June (2026), pp.171-178
ISSN 2508-9099(p) 2672-1198(e) *http://dx.doi.org/10.29056/jdaem.2026.06.03*

이미지 기반 모델이 볼륨 렌더링을 이해할까?

Do Vision Foundation Models Understand Volume Renderings?

강지윤[1], 안하일[2], 정윤현[3]*

Jiyun Kang[1], Haill An[2], Younhyun Jung[3]*

요 약

직접 볼륨 렌더링은 대규모 자연 이미지 데이터셋을 사전학습시켜 이미지의 형태적 구조와 복잡한 문맥적 구조를 포착하는 능력을 가지고 있는 모델이다. 한편, 직접 볼륨 렌더링 이미지는 3D 볼륨 정보를 2D 공간으로 투영하고 투명도 중첩, 전이 함수의 광학적 속성 매핑 등이 반영되기 때문에 자연 이미지 보다 복합적인 정보를 포함하고 있다. 본 연구는, 자연 이미지를 통해 사전 학습된 VFM 들이 DVR 이미지에 대해서도 의미 있는 특징을 추출할 수 있는지를 분석하고 검증한다. 이를 위해 학습 방식과 입력 모달리티 측면에서 서로 다른 대표적인 시각 기반 모델인 DINO, CLIP, SAM의 이미지 인코더를 분석 대상으로 한다. 해당 모델들을 통해 볼륨 렌더링 이미지의 특징을 시각화함으로써 각 모델의 구조적 표현력과 색상 구분력을 비교 분석한다. 또한, 각 모델의 학습 방식이 특징 시각화 결과에 어떤 영향을 미치는지 살펴봄으로써, 각 모델의 사전 학습 전략에서 발생하는 특성 차이를 분석한다.

핵심어 : 직접 볼륨 렌더링, 시각 기반 모델, 특징 추출

Abstract

Vision Foundation Models (VFMs) are pre-trained on large-scale natural image datasets and possess strong capabilities for capturing both morphological structures and complex contextual patterns in images. In contrast, Direct Volume Rendering (DVR) images contain more intricate information than natural images, as they project 3D volumetric data onto a 2D plane while incorporating transparency accumulation and the optical properties defined by a transfer function. This study investigates whether VFMs pre-trained solely on natural images can also extract meaningful features from DVR images. To this end, we analyze the image encoders of three representative VFMs－DINO, CLIP, and SAM－which differ in their training objectives and input modalities. By visualizing the features extracted from volume-rendered images using these models, we compare their capabilities in terms of structural expressiveness discrimination. Furthermore, by examining

1 School of Computing, Gachon University, Gyeonggi-do, Republic of Korea [Graduate Student]
e-mail: jiyun1215@gachon.ac.kr
2 School of Computing, Gachon University, Gyeonggi-do, Republic of Korea [Graduate Student]
e-mail: xenotic@gachon.ac.kr
3 School of Computing, Gachon University, Gyeonggi-do, Republic of Korea [Professor]
e-mail: younhyun.jung@gachon.ac.kr (Corresponding author)

* 본 과제(결과물)는 2025년도 교육부 및 경기도의 재원으로 경기RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다 (2025-RISE-09-A01). 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2025-00554526).

Received(May 19, 2026), Review Result(1st: June 1, 2026), Accepted(June 13, 2026), Published(June 30, 2026)

how each model's training methodology influences its feature visualization outcomes, we provide an analysis of the characteristic differences that arise from their respective pre-training strategies.

Keyword : Direct Volume Rendering, Vision Foundation Model, Feature Extraction

1. 서론

직접 볼륨 렌더링(Direct Volume Rendering, DVR)은 복잡한 3차원 볼륨 데이터를 2차원 평면에 투영하여 내부 구조의 밀도와 공간적 맥락을 동시에 시각화하는 기술이다 [1]. 이 과정에서 전이함수(Transfer Function, TF)는 볼륨 데이터의 스칼라값을 불투명도와 색상이라는 광학적 속성으로 대응시키는 역할을 수행하며, 이를 통해 사용자는 특정 관심영역(Region of Interest, ROI)을 정의하고 시각적으로 구체화하는 역할을 수행한다. 따라서 DVR 이미지는 단순한 2차원 투영 이미지가 아니라, TF 조작을 통해 사용자의 의도가 반영된 고차원적 시각 정보라고 할 수 있다.

그러나 적절한 TF를 설정하는 과정은 볼륨 데이터의 히스토그램, 내부 구조들의 공간적 관계, 시각적 상호 가림 등을 종합적으로 이해해야 하는 복잡한 작업이며, 시행착오에 의존하는 비직관적이고 반복적인 작업이라는 한계가 존재한다. 이러한 비효율성을 극복하기 위해 자동화된 TF 생성이나 유사 이미지 검색과 같은 기술이 연구되고 있으나, 이를 실현하기 위해서는 먼저 렌더링된 이미지가 포함하고 있는 정보를 의미 있게 해석할 수 있어야 한다. 즉, DVR 이미지를 단순한 픽셀 집합이 아니라 어떤 구조가 강조되었는지, 공간적 정보가 어떻게 표현되었는지와 같은 의미론적 특징을 추출하는 능력이 필수적이다.

최근 인공지능 분야에서는 이러한 특징추출 문제를 해결하기 위한 새로운 접근법으로 대규모 데이터셋을 통해 학습된 기반 모델(Foundation Model, FM)이 주목받고 있다. 자연어 처리 분야(Natural Language Processing, NLP)에서 시작되어 Vision Transformer 모델을 활용하여 컴퓨터 비전 분야로 확장되었으며, 대규모 자연 이미지 데이터셋을 self-supervised 또는 semi-supervised 방식으로 학습한 시각 기반 모델(Vision Foundation Model, VFM)들은 이미지의 형태적 구조와 복잡한 문맥 정보를 정교하게 포착하는 능력을 입증하였다 [2]. 이는 강력한 사전학습을 통해 얻은 일반적인 표현 능력이 DVR 이미지와 같이 시각적으로 난이도가 높은 데이터에서도 유용할 수 있음을 시사한다. 그러나 기존 VFM은 불투명한 표면과 명확한 경계를 가진 자연이미지를 중심으로 학습되었다는 점에서, 투명도 기반의 중첩 표현, TF 변화에 따른 시각적 다양성 등 DVR 이미지 고유의 속성과는 분명한 도메인 격차가 존재한다. 따라서 본 논문에서는 자연이미지 기반으로 학습된 모델이 사용자의 의도가 반영하면서 고유한 특성을 가진 DVR 이미지에서도 여전히 의미 있는 특징을 안정적으로 추출할 수 있는지를 체계적으로 분석, 검증 하고자 한다.

본 연구의 기여는 다음과 같다.

자연 이미지 기반의 VFM 중 볼륨 렌더링 이미지의 구조적, 문맥적 특징을 가장 효과적으로 추출하는 최적의 모델을 탐색한다.

선정된 VFM들을 DVR 이미지에 적용하여 각 모델의 특징 공간이 지닌 특성을 비교 분석한다.

2. 관련 연구

DVR 에서 적절한 TF를 설정하는 것은 볼륨 데이터에 대한 사전 지식을 요하며 사용자의 의도를 시각적으로 구현하기까지 많은 시행착오를 동반한다. 이러한 한계를 극복하기 위해, 최근 연구들은 대규모 데이터로 사전학습된 FM의 의미론적 추론 능력을 활용하여 TF 조작의 복잡성을 낮추고 사용자의 의도와 시각화 결과 간의 의미적인 간극을 줄이려는 시도를 해왔다.

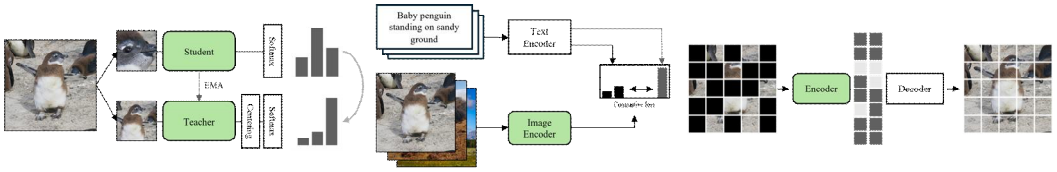
대표적으로 Jeong et al. [3]은 CLIP 모델을 활용하여 볼륨 렌더링 이미지와 이를 설명하는 텍스트 간의 특징 공간을 정렬함으로써, 사용자가 직관적인 텍스트 입력을 통해 원하는 TF 재구성과 DVR 이미지를 생성하고 제어할 가능성을 보여주었다. 또한, Wang et al. [4]은 GPT, Gemini와 같은 Multimodal Large Language Model(MLLM)을 활용하여, 복잡한 조작 없이 대화형 인터페이스를 통해 사용자의 의도를 파악하고 TF를 최적화하는 프레임워크를 제시하였다.

그러나 이러한 선행 연구는 주로 텍스트 기반 제어에 집중되어 있어, VFM의 image encoder가 DVR 이미지의 시각적 특징을 잘 추출할 수 있는지에 대한 분석은 부족한 실정이다. 이에 본 연구는 대규모 자연 이미지로 학습된 VFM의 Image Encoder가 이질적인 도메인인 DVR 이미지의 구조적, 문맥적 특징을 효과적으로 추출할 수 있는지 적합성을 검증하고 분석하는 데 집중한다.

3. Method

3.1 Overview

VFM은 사전학습에 사용되는 입력값의 모달리티에 따라 세 가지로 구분할 수 있다. 단일 이미지, 이미지-텍스트 쌍, 이미지와 이미지 프롬프트(box, point, mask)를 입력으로 하는 FM이 대표적이다. [그림 1]과 같이 해당 세 가지 방식으로 구현된 모델들에 대하여 가장 범용적으로 알려져 있고 backbone 모델로써 활용된 모델인 DINO [5], CLIP [6], SAM [7]을 비교 대상으로 한다.



[그림 1] 세 가지 Vision Foundation Model들의 사전학습 방식 개요 그림이다

[Fig. 1] Overview of pretraining method of 3 Vision Foundation Models

3.2 Model Architecture

선정된 세 가지 모델은 모두 동일한 Vision Transformer 모델을 학습한다. 이 모델을 학습시키는 데에 있어서 서로 다른 입력과 학습 loss를 사용하여 해당 모델을 최적화하는 과정에서의 차이점을 보인다.

DINO는 학습 시에 단일 이미지를 기반으로 동일한 두 ViT 모델을 각각 student, teacher로 설정하여 학습하는 self-distillation 기법을 사용한다. crop 전략을 통해 student에게는 local 정보를 teacher에게는 global 정보를 포함하여 나온 출력값을 확률 분포로 표현하여 정렬하는 방식으로 학습함으로써, 모델이 이미지 내의 “local-to-global” 구조적 관련성을 파악하게 한다.

CLIP은 학습 시에 이미지-텍스트 쌍을 학습 데이터로 ViT 기반의 이미지 인코더와 Transformer 기반의 텍스트 인코더를 활용하여 각각의 모델에서 특징들을 대조학습을 통해 학습한다. 해당 학습 방식을 통해 CLIP은 이미지와 텍스트 간의 연관성을 인식하고 구분할 수 있는 능력을 갖추게 된다.

SAM은 분할에 특화된 FM이다. 해당 모델은 이미지 인코더를 통과한 의미론적 특징을 추출한 뒤에 분할에 도움이 되는 텍스트 프롬프트 또는 시각적 프롬프트 (segmentation box, point, mask)를 입력받아 분할 결과를 출력한다. SAM의 이미지 인코더는 사전학습된 MAE(Masked Auto Encoder) [8]로 사전학습된 모델이며, ViT 모델을 사용하여 의미 정보를 표현한다. MAE는 이미지 패치를 75% 정도 마스킹한 뒤, 남은 패치에서 추출한 특징을 이용해 원본 이미지를 픽셀 단위로 재구성하는 방식으로 학습된다. 이 과정에서 모델은 부족한 정보로 전체 이미지를 복원하려고 시도함으로써, 다양한 구조와 시각적 의미를 학습한다.

3.3 실험 방식

DVR 이미지는 이미지 자체의 구조적 정보와 TF를 통해 조정함으로써 사용자가 보고자 하는 지역적 정보를 모두 포함하고 있다. 해당 복합적인 특징을 모두 확인하기 위해 DINO에서 사용한 Attention Map visualization 전략을 차용한다. Attention Map은 Vision Transformer 모델에서 계산한 각 haed 별 Attention이 이미지의 어떤 부분을 중요하게 보고 있는지를 시각적으로 나타낸다. 이를 통해 DVR 이미지의 특징을 각 모델이 어떻게 추출하는지를 판단한다.

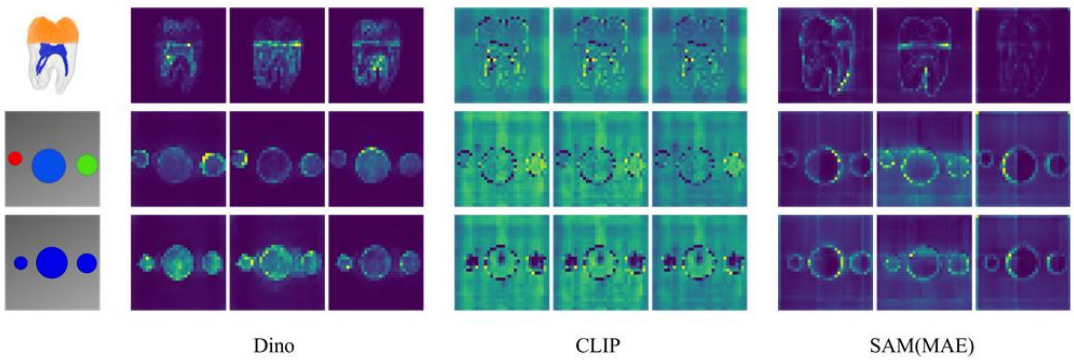
Attention Map visualization은 ViT 모델에 이미지를 넣어 마지막 layer에서 추출된 multi-head Attention 값에 대하여 CLS 토큰을 기준으로 각 패치 간의 중요도를 Attention Map으로써 시각화하는 방식이다. CLS 토큰을 포함한 상태로 사전학습을 진행하는 DINO와 CLIP은 DINO의 전략을 그대로 사용하고, CLS 토큰을 포함하지 않고 학습한 SAM의 이미지 인코더는 패치별 Attention을 평

균값으로 계산하여 대푯값으로 시각화한다.

4. 실험 결과

4.1 Implementation Details

본 실험은 다음 주어진 데이터셋을 사용한다 : Tooth, Three balls. Three balls 이미지는 Voreen 렌더러를 통해 제작되었고 Tooth 이미지는 자체 제작한 VTK 기반 렌더러를 통해 제작되었다. 비교를 위한 DINO, CLIP, SAM 모델은 모두 해당 논문의 github에서 제공하고 있는 모델의 정의와 가중치를 가져와 사용했다. Attention Map을 시각화하는 코드는 DINO 코드에서 제공하고 있는 visualize_Attention.py를 참고하여 각각 CLIP, SAM에 적용했다. [그림 2]는 ViT-B/16 모델이 출력하는 12개의 Attention Map 중 3개를 뽑아 보여준다.



[그림 2] 볼륨 렌더링 이미지에 대한 Attention Map 결과 그림이다
[Fig. 2] Results of Attention Map in Volume Rendering Image

4.2 Tooth 데이터셋에서의 구조적 표현력 분석

치아 데이터셋에서 DINO, CLIP, SAM의 Attention Map을 비교한 결과, 세 모델 모두 치아 윤곽과 주요 구조에 대한 기본적인 전역적 인지는 가능하였으나, 의미론적 관심 영역인 치아의 머리, 뿌리, 신경 세 가지 구조를 분리하는 능력은 모델 간 큰 차이를 보였다. DINO는 각 Attention Head가 서로 다른 구조에 선택적으로 집중하며, 특히 신경과 뿌리처럼 TF에 의해 색상, 불투명도가 명확히 구분된 영역을 독립적 의미 단위로 식별하는 양상을 보였다. 이는 DINO가 Self-distillation기반의 self-supervised 학습을 통해 이미지 패치 간의 구조적, 형태적 일관성을 학습하고, 레이블 없이도

객체 내부의 semantic part를 구성하는 능력을 갖추었기 때문이다. 결과적으로, TF가 강조한 구조적 의미를 DINO는 특징 공간에서 명확한 경계로 재구성함으로 DVR 이미지의 의미론적 ROI 분리를 가장 효과적으로 수행하였다.

반면 CLIP은 전체 구조의 위치 관계나 치아 윤곽과 같은 전역적 맥락은 유지하였으나, 내부의 의미론적 구조는 구분하지 못하고 Attention Map이 전반적으로 확산된 형태를 보였다. 이는 CLIP의 학습 목적이 이미지-텍스트 간의 전역 개념 정렬에 있으며, 이미지 내부의 세밀한 구조적 차이나 TF 기반 색상, 불투명도 차이를 구분하는 것이 주요 목표가 아니기 때문이다 [9]. DVR 이미지에서는 TF가 구조를 구분하는 핵심 신호로 사용되나, CLIP은 이러한 신호를 해석하지 못해 구조 분해 능력이 제한적이었다.

SAM은 모든 Attention Head에서 전경-배경 분리와 외곽 경계에 일관되게 집중하는 경향을 보였다. SAM의 MAE 기반 이미지 인코더는 분할 강건성 확보를 위해 형태 구조와 경계 복원에 최적화되어 있으며, 색상, 강도 변화와 같은 색상 차이를 분할 연관 특성으로 간주하지 않는다. 그 결과, 치아의 윤곽은 가장 명확하게 포착하였으나 내부 ROI 분리는 거의 수행하지 못했다. 이는 SAM이 DVR 이미지에서 구조적 형태는 잘 포착하지만, TF에 의해 드러난 의미론적 구조를 분해할 능력은 부족함을 의미한다.

종합적으로, DVR 이미지의 의미론적 구조는 TF 기반 지역적 대비와 구조적 패턴이 혼합된 형태로 존재하며, 이러한 복합적인 특성을 가장 정교하게 추출한 모델은 DINO임을 보여준다.

4.3 3 balls 데이터셋에서의 색상 구분력 분석

Three balls 데이터셋의 실험 결과, DINO는 색상 차이를 강한 변별 요인으로 활용하며, 세 개의 구가 서로 다른 색을 가질 때 각 공을 독립 객체로 분리하여 Attention head가 일관되게 반응하였다. 이는 DINO가 이미지 패치 간 유사성을 기반으로 특징 공간을 구성하는 self-supervised 학습 구조로 되어 있어, 색상 차이가 패치 유사도 구조를 변화시키는 주요 신호로 작동하기 때문이다. 반면 단일 색상 조건에서는 세 객체 모두 동일한 중요도를 부여하여 색상 변화가 없을 때는 구조적 동일성만을 반영하는 특징 표현을 보였다. 이는 DINO가 DVR 이미지에서 TF 기반 색상 변화 차이를 구조적 구분 신호로 활용할 수 있음을 의미한다.

CLIP은 다중 색상과 단일 색상 조건에서 Attention Map이 유사하게 유지되었으며, 색상 차이에 따른 분할이나 의미론적 분리가 나타나지 않았다. 이는 CLIP의 특징 공간이 이미지-텍스트 정렬을 중심으로 구성되어있어, 색상 대비와 같은 TF 기반 요소가 특징 공간에 중요하게 반영되지 않기 때문이다. DVR 이미지에서 TF 변화는 의미론적 구조 강조 수단이지만, CLIP의 능력은 이러한 TF 기반 변화를 중요 특성 요소로 해석하지 못한다.

SAM도 CLIP과 동일하게 Attention Map의 변화가 거의 관찰되지 않았으며, 모든 경우에서 공의

윤곽만을 강하게 포착하였다. SAM의 분할 중심 특징 공간 표현의 특성상 색상, 불투명도와 같은 요소는 무시되도록 학습된다. 따라서 TF에 의해 강조된 색상 변화는 SAM의 Attention Map에서 무시되며, 이는 DVR 이미지에서 TF 변화 적응성이 낮다는 것을 의미한다.

결과적으로, DVR 이미지의 중요한정보증 하나인 TF 기반 색상 변화가 모델별로 다르게 반영됨을 보여주며, 특히 DINO만이 TF 변화가 의미하는 의도적 의미 차별을 특징 공간에서 안정적으로 분리할 수 있음을 확인하였다.

5. 결론

본 연구는 DVR 이미지에 대한 VFM 들이 추출한 특징에 대한 구조적, 색상 표현력을 분석하고 검증했다. 우리는 세 VFM인 DINO, CLIP, SAM이 출력하는 Attention Map을 활용하여 DVR 이미지의 특징 분석을 위한 기준으로 삼았다. 실험을 통하여 세 VFM 중 DINO가 DVR 이미지의 구조적 정보와 색상 정보 모두에 대하여 우수한 구분력을 보여주는 것을 확인하였다. CLIP은 세부 구조보다도 전역적인 정보를 중요시하는 학습 과정으로 인해 DVR 이미지의 특성을 잘 추출하지 못하고 있음을 나타냈다. 마지막으로, SAM 모델은 경계에 대한 중요도를 높게 판단하였지만, DVR 이미지의 중요 특성 중 하나인 색상에 대해서는 분리하지 못함을 확인하였다.

References

[1] K. Engel, M. Hadwiger, J. M. Kniss, A. E. Lefohn, C. R. Salama, and D. Weiskopf, "Real-time volume graphics," in *ACM SIGGRAPH 2004 Course Notes*, Los Angeles, CA, USA, Aug. 2004, pp. 29-es, doi: 10.1145/1103900.1103929.

[2] R. Bommasani et al., "On the opportunities and risks of foundation models," unpublished.

[3] S. Jeong, J. Li, C. R. Johnson, S. Liu, and M. Berger, "Text-based transfer function design for semantic volume rendering," in *2024 IEEE Visualization and Visual Analytics (VIS)*, Florida, USA, Oct. 13-18, 2024, pp. 196-200, doi: 10.1109/VIS51563.2024.00035.

[4] Y. Wang et al., "IntuiTF: MLLM-guided transfer function optimization for direct volume rendering," unpublished.

[5] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, Oct. 11-17, 2021, pp. 9650-9660, doi: 10.1109/ICCV48922.2021.00954.

[6] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, Virtual, Jul. 18-24, 2021, pp. 8748-8763, doi: 10.48550/arXiv.2103.00020.

[7] A. Kirillov et al., "Segment anything," in *Proceedings of the IEEE/CVF International Conference on*

Computer Vision (ICCV), Paris, France, Oct. 2023, pp. 4015-4026.

[8] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 19-24, 2022, pp. 16000-16009, doi: 10.1109/CVPR52688.2022.01554.

[9] Y. Wang, S. Xu, X. Zhu, and Y. Li, "MSCI: Addressing CLIP's inherent limitations for compositional zero-shot learning," unpublished.