객체-구성적 신경 암시적 표현 기반 신규 시점 객체 영상 렌더링 방법

김원겸 [1], 조경은 [2]
[1] 동국대학교 컴퓨터 · AI 학과 석 · 박사 통합과정
[2] 동국대학교 컴퓨터 · AI 학부 교수
dnjsrua59@gmail.com, cke@dongguk.edu(교신저자)

Novel View Object Image Rendering Method Based on Object-Compositional Neural Implicit Representation

Wongyeom Kim[1], Kyungeun Cho[2]
[1]Department of Computer Science and Artificial Intelligence
[2]Division of Computer Science and Artificial Intelligence

요 약

본 연구는 객체-구성적 신경 암시적 표현에서 객체의 새로운 시점 영상을 렌더링하는 방법을 제안한다. 객체-구성적 신경 암시적 표현 최적화 과정은 오직 2 차원 영상에서 관측된 정보에 의존하고, 객체 형상에 대한 평가가 부재하다. 따라서, 본 연구는 2 차원 객체 영상으로부터 형상을 평가할 수 있도록, 객체-구성적 신경 암시적 표현으로부터 객체 위치를 계산하고 객체 위치를 중심으로 카메라와 영상 평면 위치를 설정하여 새로운 시점의 객체 영상을 렌더링한다.

1. 서론

3 차원 재구성은 2 차원 영상 또는 멀티 모달리티로부터 3 차원 기하학과 외관을 구현하는 기술이다. 장면을 3 차원으로 재구성하는 것은 컴퓨터 비전 분야에서 중요한 연구 주제로, 게임 및 애니메이션 에셋 제작, 증강 현실(AR) 및 가상 현실(VR) 등 다양한 애플리케이션에서 활용된다.

최신 3 차원 장면 재구성 방법론들은 신경 암시적 표현(neural implicit representation)을 기반으로 연구되고 있다. 신경 암시적 표현은 3 차원 좌표로부터 3 차원 기하학 정보와 색상 정보를 계산하는 함수를 신경망으로 모델링한 것이다. 신경 암시적 표현은 3 차원 실측치 없이 일반 카메라로 촬영된 2 차원 동영상으로 최적화할 수 있고, 해상도에 의한 제약이 없어 학습 및 응용에 필요한 컴퓨팅 요구사항도 비교적 낮다. 더 나아가, 신경 암시적 표현에 객체-구성적(object-compositional) 요소를 포함하기 위해, 일부 연구들은 2 차원 인스턴스 맵(instance map)을 최적화 과정에 활용함으로써, 장면을 배경과 객체로 분리하여 재구성할 수 있다. 이는 애플리케이션에서 활용뿐만 아니라, 객체 수준의 이해를 필요로 하는 연구에 기여할 수 있다.

그러나, 신경 암시적 표현 최적화 과정은 오직 2 차원 동영상에서 관측된 정보에 의존하며, 관측되지 않은 정보에 대한 처리가 부재하다. 특히, 여러 객체로 구성된 실내 장면에는 객체가 촬영할 수 없는 위치에 있거나 다른 객체로 인해 폐색되는 공간들이 다수 존재한다. 또한, 관측된 정보에 기반한 단순 재구성에 집중하고 재구성된 객체 형상에 대한 평가가 부재하기 때문에, 잘못된 재구성에 대한 규제를 수행할 수 없다.

따라서, 본 연구에서는 객체의 형상 손실 탐지 및 완전성 평가를 할 수 있도록, 최적화된 객체-구성적 신경 암시적 표현으로부터 새로운 시점에서의 객체 영상을 렌더링하는 방법을 제안하고자 한다.

본 연구의 기여점을 요약하면 다음과 같다.

- 객체-구성적 신경 암시적 표현으로부터 객체 위치를 계산하고, 새로운 시점에서의 객체 영상을 렌더링한다.
- 효율적인 렌더링을 위해, 추정된 객체의 기하학 및 외관 정보로 된 3 차원 그리드를 구성하고, 임의의 포인트에 대한 값을 계산할 때 보간 함수로 활용한다.

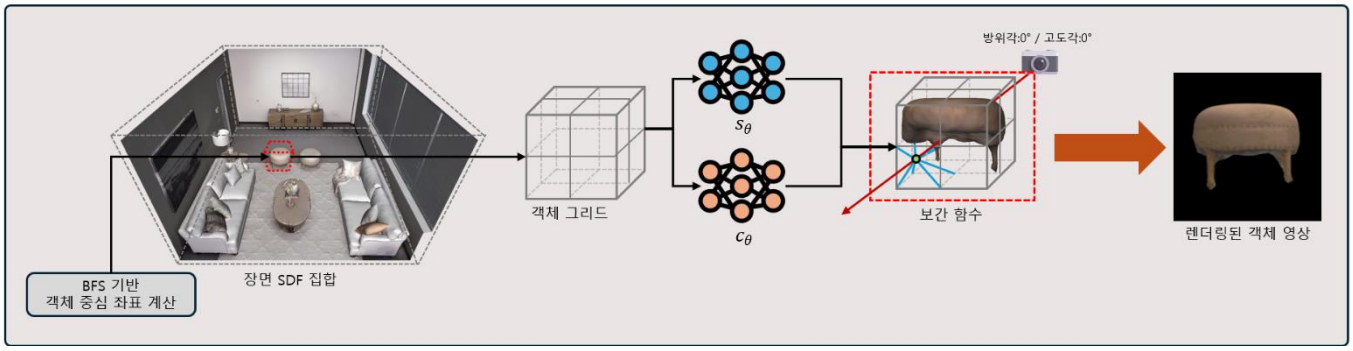

그림 1. 제안하는 연구의 개요

2. 관련 연구

2.1. 신경 암시적 표현

신경 암시적 표현 기법은 뷰 합성(View Synthesis) 또는 3 차원 재구성을 수행하기 위해, 다중 시점 RGB 영상으로 신경망을 최적화하는 방법이다. 해당 기법은 신경망으로 3 차원 좌표에서의 기하학 정보와 색상 정보를 추정한 후, 볼륨 렌더링 기법을 통해 하나의 픽셀로 합성할 수 있다. 신경망 최적화는 합성된 픽셀과 RGB 영상의 픽셀과의 손실을 계산함으로써 달성할 수 있다. 대표적인 출력 형태는 신경 방사 필드(Neural Radiance Field, NeRF)[1, 2]와 부호화된 거리 필드(Signed Distance Field, SDF)[3, 4]가 있다. 특히, SDF 는 표면까지의 거리를 나타내는데, 표면을 부드럽게 표현할 수 있다는 점과 단안 기하학 단서를 추가 활용[5]하여 정확한 표면을 재구성할 수 있다는 점 때문에 장면 재구성에서 주목받고 있다.

하지만, 일반적인 신경 암시적 표현은 배경과 객체를 하나로 간주하여 재구성하고, 이에 따라 객체 수준의 이해를 활용할 수 없다.

2.2. 객체-구성적 신경 암시적 표현

객체-구성적 신경 암시적 표현은 장면 내 구성 요소에 대한 기하학 정보를 개별적으로 추정하도록 설계되었다. 대표적인 연구인 ObjectSDF[6]는 각 요소의 SDF 를 인스턴스 로짓(instance logit)으로 매핑하는 함수를 설계했고, 매핑된 결과와 2 차원 인스턴스 맵 간 손실 함수를 기반으로 신경망을 최적화하는 방법을 제안했다. 장면 재구성으로 확장하기 위한 연구인 RICO[7]와 ObjectSDF++[8]는 요소 간 SDF 가 교차하는 문제를 해결하기 위한 규제 방법을 설계하는 방향으로 연구가 진행되었다.

객체-구성적 신경 암시적 표현의 성과는 장면을 배경과 객체로 분리하여 재구성하기 때문에 애플리케이션에서의 활용성이 증대되고, 객체 수준의 이해를 활용하는 연구가 등장하는 기틀을 마련했다.

그러나, 미관측 정보에 대한 처리 및 재구성된 객체에 대한 평가가 여전히 부재하기 때문에, 장면 내 객체에 인공물이나 구멍이 발생하는 한계점은 남아있다.

3. 제안하는 방법론

3.1. 객체-구성적 신경 암시적 표현 최적화

본 연구에서, 장면에 대한 기하학 정보는 SDF 로 표현된다. 우선, 임의의 3 차원 좌표가 입력되었을 때 k 개의 요소에 대한 SDF 값을 계산하는 함수 $s(\mathbf{x})$를 신경망 s_θ으로 구현한다:

$$s_\theta : \mathbf{x} \in \mathbb{R}^3 \mapsto \left(\hat{\mathbf{s}} \in \mathbb{R}^k, \hat{\mathbf{f}} \in \mathbb{R}^{256}\right) \tag{1}$$

여기서, $\mathbf{x}$는 3 차원 좌표, $\hat{\mathbf{s}}$은 추정된 SDF 값, $\hat{\mathbf{f}}$은 특징 벡터이다. 또한, 색을 추정하는 신경망 c_θ 은 다음과 같다:

$$c_\theta : \left(\mathbf{x} \in \mathbb{R}^3, \mathbf{d} \in \mathbb{S}^2, \hat{\mathbf{n}} \in \mathbb{R}^3, \hat{\mathbf{f}} \in \mathbb{R}^{256}\right) \mapsto \hat{\mathbf{c}} \in \mathbb{R}^3 \tag{2}$$

여기서, $\hat{\mathbf{c}}$은 $\mathbf{x}$와 바라보는 방향 $\mathbf{d}$에 따라 추정된 색상 값, $\hat{\mathbf{n}}$은 $\mathbf{x}$에서의 노멀이다. 신경망을 학습하기 위해, 우선 2 차원 영상의 카메라 매개변수로 계산된 카메라 위치 $\boldsymbol{o}$에서 영상 평면의 각 픽셀을 향해 광선 $\mathbf{r}$ 을 발사한다. 각 광선에서 n 개의 포인트 v_i 을 추출하여 3 차원 좌표 $\mathbf{x}_i = \boldsymbol{o} + v_i\mathbf{d}$를 계산한 후, 각 신경망을 활용해 $\mathbf{x}_i$ 에서의 $\hat{\mathbf{s}}_i$ 와 $\hat{\mathbf{c}}_i$ 을 추정한다. 그리고, 볼륨 렌더링을 사용해 색상 픽셀로 합성한다:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{n} T_i \alpha_i \hat{\mathbf{c}}_i \tag{3}$$

여기서, α_i는 볼륨 렌더링을 위해 $\hat{\mathbf{s}}_i$을 밀도 값으로 변환한 값이고, $T_i = \prod_{i=1}^{j}(1-\alpha_i)$은 투과율이다. 색상 픽셀 대신 깊이 픽셀이나 노멀 픽셀로 합성하기 위해서는 $\hat{\mathbf{c}}_i$를 v_i 또는 $\hat{\mathbf{n}}_i$로 변환하면 된다. 이를 통해 2 차원 실측치를 활용하여 s_θ와 c_θ를 최적화할 수 있다.

3.2. 객체 위치 탐색 및 객체 영상 렌더링

새로운 시점에서 객체 영상을 렌더링하기 위해서는 카메라 매개변수로부터 계산된 것이 아닌, 객체를 중심으로 카메라 및 영상 평면 위치를 계산해야 한다. 최적화된 객체-구성적 신경 암시적 표현에서 각 객체의 SDF 는 규제로 인해 서로 교차하지 않고 분리되게 된다. 또한, 객체의 SDF 에서 객체 표면 내부는 음수, 외부는 양수로 표현된다. SDF 의 특징을 바탕으로, 하나의 축 길이가 L 인 방 전체를 포함하는 3 차원 그리드 G를 정의한 후, L^3개의 모든 포인트에 대해 목표 객체의 SDF 값 $\hat{s}_a$ 을 추정하여 목표 객체의 SDF 집합 $\hat{S}_a$을 구성한다:

$$\hat{S}_a = \{\hat{s}_a(\mathbf{x}_l), \mathbf{x}_l \in G, l = 1, \dots, L^3\} \quad (4)$$

다음으로, 너비 우선 탐색(Breadth First Search, BFS) 알고리즘을 통해 $\hat{S}_a$ 로부터 연속적인 3 차원 음수 집합을 탐색한 후, 가장 큰 집합을 선택하면 3 차원 공간에서의 객체 위치를 추출할 수 있다. 추출된 객체 위치를 중심으로 거리, 방위각, 고도각을 설정하여 카메라와 영상 평면의 위치를 계산하면, 볼륨 렌더링을 통해 새로운 시점의 객체 영상을 시각화할 수 있다.

3.3. 보간 함수 기반 객체 영상 렌더링

객체의 형상을 평가하기 위해서는 영상 내에 객체 형상 전체가 나타나도록 렌더링을 수행해야 한다. 하나의 픽셀을 렌더링하기 위해 광선마다 샘플링되는 포인트의 최소 개수가 n 일 때, 렌더링할 영상의 크기가 m x m 이면 신경망에 입력되는 포인트의 최소 개수는 $m^2 * n$이 된다. $m^2 * n$개의 포인트를 신경망에 입력하는 것은 범용 GPU 의 메모리로는 한계가 있기 때문에, 선명도와 해상도 중 하나를 선택해야 한다.

본 연구에서는 신경망을 통과하는 포인트의 개수는 최소화하면서 렌더링 품질은 유지하기 위해 보간 함수를 활용하는 방법을 제안한다. 객체 위치를 중심으로 객체 전체를 포함하는 3 차원 그리드 G_a를 정의한 후, 색상과 목표 객체의 SDF 값을 추정하여 색상 집합 $\hat{C}$ 과 목표 객체의 SDF 집합 $\hat{S}_a$ 를 구성한다. 그리고, 임의의 3 차원 좌표가 입력되었을 때 $\hat{C}$과 $\hat{S}_a$를 기반으로 입력된 3차원 좌표의 $\hat{\mathbf{c}}$과 $\hat{s}_a$를 계산하는 보간 함수를 정의한다:

$$\hat{s}_a = \text{interp}(\mathbf{x}, \hat{S}_a), \qquad \hat{\mathbf{c}} = \text{interp}(\mathbf{x}, \hat{C}) \quad (5)$$

여기서, interp 는 선형 보간 함수이다. $\hat{C}$ 과 $\hat{S}_a$ 을 구성할 때는 신경망을 활용하지만, 영상 렌더링을 위해 색상과 목표 객체의 SDF 값을 계산할 때는 오직 보간 연산만을 수행한다. 따라서, 그리드의 포인트 개수가 더 적은 경우에는 신경망에 직접 입력하는 것에 비해 메모리 효율성이 좋다. 또한, 영상을 렌더링할 때마다 메모리가 동일하게 소요되는 기존 방식과 달리, 제안하는 방식은 보간 함수를 정의한 후에는 메모리 소요가 적기 때문에 여러 시점의 영상을 동시에 렌더링할 수 있다.

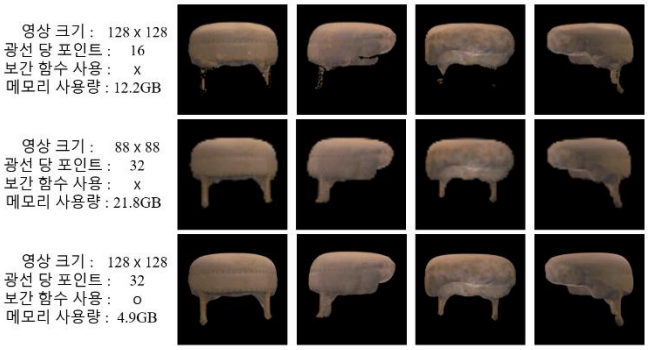

그림 2. 실험 결과

4. 실험 및 결과

4.1. 실험 설정

실험에서 사용된 데이터셋은 합성 장면 데이터셋인 레플리카[9]로, 하나의 장면에 대해 100 개의 RGB 영상과 상응되는 인스턴스 맵을 제공한다. 본 연구에서는 Room 0 을 선정하여 실험을 수행한다.

객체-구성적 신경 암시적 표현을 학습하기 위해, 옵티마이저는 학습률이 5e-4 로 설정된 Adam 을 사용한다. 하나의 RTX 3090 GPU 에서 학습할 수 있다.

4.2. 실험 내용

그림 2 는 Room 0 에서 원형 의자를 방위각 90° 간격으로 회전하며 렌더링한 결과이다. 영상을 렌더링한 후, 384 x 384 리사이즈를 수행한다. 그림 2 의 상단과 중단은 보간 함수를 활용하지 않은 결과로, 컴퓨팅 요구사항을 충족하기 위해 선명도와 해상도에 조정이 필요하다. 상단은 광선 당 포인트 개수를 32 개에서 16 개로 줄여 262,144($128 * 128 * 16$)개의 포인트를 추정한 결과로 다리와 같은 얇은 구조물이 렌더링 되지 않은 문제가 발생한다. 중단은 영상의 크기를 128 x 128 에서 88 x 88 로 줄여 247,808 ($88 * 88 * 32$)개의 포인트를 추정한 결과로, 해상도가 낮아지기 때문에 영상 전체에 블러링이 발생하는 것을 확인할 수 있다.

반면, 보간 함수를 활용한 본 연구의 방법은 하나의 축 길이가 48 인 3 차원 그리드를 사용하여 131,072(48^3)개의 포인트만 신경망에 입력한 후, 그 결과를 보간 함수로 정의한다. 보간 함수를 정의한 후에는, 신경망을 직접 활용하지 않기 때문에 광선 당 포인트 개수나 영상의 크기에 따른 연산량은 큰 영향을 미치지 않는다. 따라서, 그림 2 의 하단과

같이 객체의 형상을 명확하게 렌더링하면서도 사용하는 메모리는 상대적으로 낮은 것을 확인할 수 있다.

5. 결론

본 연구는 최적화된 객체-구성적 신경 암시적 표현으로부터 3 차원 공간상 객체 위치를 계산하고 카메라와 영상 평면 위치를 설정해 객체 영상을 렌더링하는 방법을 제안했다. 범용 GPU 에서 영상의 선명도와 해상도를 갖추기 위해, 3 차원 그리드의 포인트로부터 값을 예측하고 이를 보간 함수를 정의할 때 사용함으로써, 영상 렌더링에는 추가적인 보간 연산만 수행하면 되도록 하였다. 향후 연구에서는 제안한 방법을 바탕으로 객체 영상을 직접적으로 평가하는 방법을 통해 신경망을 최적화할 수 있을 것으로 기대된다.

사사

* 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2022-NR070225).(90%)

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었음(IITP-2025-RS-2023-00254592).(10%)

참고문헌

[1] Ben Mildenhall, et al. “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis.” The European Conference on Computer Vision, 2020.

[2] Thomas Müller, et al, “Instant Neural Graphics Primitives with a Multiresolution Hash Encoding.” ACM Transactions on Graphics, 2022.

[3] Peng Wang, et al. “NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction.” Neural Information Processing Systems, 2021.

[4] Lior Yariv, et al. “Volume Rendering of Neural Implicit Surfaces.” Neural Information Processing Systems, 2021.

[5] Zehao Yu, et al. “MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction.” Neural Information Processing Systems, 2022.

[6] Qianyi Wu, et al. “Object compositional neural implicit surfaces.” The European Conference on Computer Vision, 2022.

[7] Z. Li, et al. “RICO:Regularizing the unobservable for indoor compositional reconstruction,” International Conference on Computer Vision, 2023.

[8] Qianyi Wu, et al. “ObjectSDF++: Improved Object-Compositional Neural Implicit Surfaces.” International Conference on Computer Vision, 2023.

[9] J. Straub, et al. “The replica dataset: A digital replica of indoor spaces.” arXiv preprint arXiv:1906.05797, 2019.