3D Reconstruction의 최근 동향 및 분석

인공지능학과 2021006253 박준우

■ Abstract

3D Reconstruction 문제에 관심이 많아 해당 분야의 최근 모델들을 조사했다. 이 문제는 다양한 각도에서 촬영된 2D 이미지를 기반으로 3D 장면이나 객체를 재구성하는 것으로 AR/VR, 로보틱스, 게임이나 영화 등 다양한 IT 나 엔터테인먼트, 제조업과 같은 분야에서 활용 가능성은 높으나 해결해야 할 문제가 아직 많이 남아있는 분야이다. NeRF, 3D Gaussian Splatting, DUSt3R 모델들은 이 문제를 해결하기 위해 각각 독창적인 접근 방식을 제공한다.

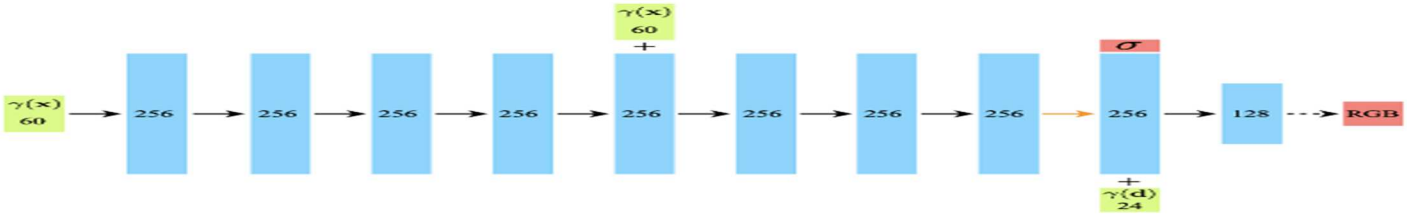
NeRF 는 신경망을 이용해 특정 시점에서 촬영한 이미지들로부터 주어지지 않은 위치와 방향에서 바라본 대상의 모습을 연속적으로 합성해내는 기술이다. 3D Gaussian Splatting 은 NeRF 와 마찬가지로 촬영한 이미지와 카메라 포즈로부터 주어진 방향 이외에서 이미지를 생성하는 뷰 합성 모델이지만 기존 NeRF 방식과 달리 MLP 와 Volume Rendering 를 사용하지 않았으며 3D Gaussian 분포와 Tile Splatting 기법으로 압도적인 렌더링, 학습시간과 퀄리티를 보여준다. DUSt3R 은 NeRF 와 3D Gaussian Splatting 과 달리 카메라 파라미터 없이, 즉 캘리브레이션 되지 않은 이미지 쌍에서 Transformer 모델을 사용하여 3D Point Map 을 예측하고 이를 정렬하여 일관된 3D 재구성을 제공한다.

■ Introduction

3D Reconstruction 기술은 실제의 모습이나 모양을 복원하는 과정이다. 카메라로 사진을 찍을 때, 실제 Real World 의 3D 장면을 2D 평면 상에 사영시켜 깊이 값이 손실되어 3D 형태 정보가 소실되며 컴퓨터가 2D 이미지만을 가지고 3D 형태를 정확하게 복원할 수 없다. 때문에 깊이를 추정하여 3D 장면을 재구성하게 되며 재구성된 3D 장면은 다양한 분야에 활용할 수가 있다.

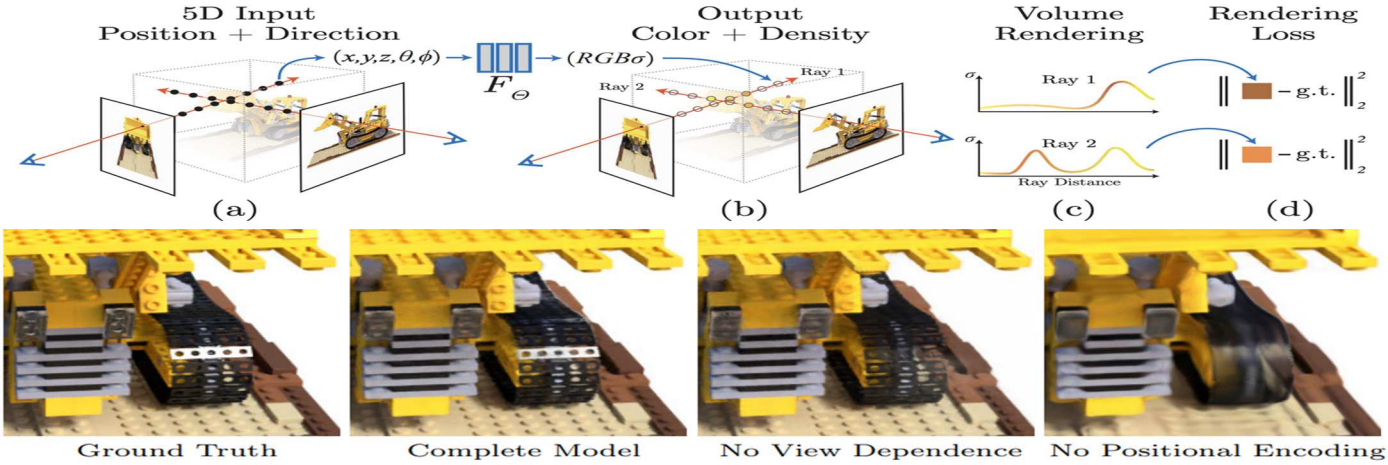
우선 로봇이 공간과 물체를 이해하여 보다 효율적인 로봇 팔의 파지 작업 및 이동이 가능하며, 디지털 트윈에 적용하여 스마트 팩토리나 효과적인 건물 및 인프라 모니터링이 가능하다. 또한 현장감과 몰입감 있는 스포츠나 공연 중계도 가능할 것이며 현재 3D 프린터를 사용하기 위해선 프린팅 할 물체의 정확한 규격을 알기 위해 3D 스캐너를 사용하는데 3D 재구성 기술이 더 발달한다면, 이러한 작업비용이 감소하며 시뮬레이션이나 가상 환경 등으로 교육이나 연구개발에도 많은 긍정적인 영향을 준다고 생각한다.

● *NeRF* 의 접근 방식



- **입력 :** 3D 공간위치 (x, y, z) + 방향 정보 θ,ϕ
- **출력 :** RGB 컬러 정보 + 밀도 값 σ

NeRF 의 구조는 MLP 로 간단하게 구성된다. 입력은 물체의 3D 위치(x, y, z)가 되고, 입력단에서부터 5 번째 레이어까지 위치 정보를 전달한다. 이후, 5 번째 레이어에서 한 번 더 해당 위치 정보가 합쳐지는 일종의 Skip Connection 이 있다. 9 번째 레이어에서는 밀도(σ)가 출력되며, 이 값에 방향 정보(d, θ, ϕ)가 다시 입력으로 들어가서 최종적으로 RGB 값이 출력된다. 밀도 값은 물체의 위치와 관련이 있으므로 물체를 바라보는 각도와 무관하게 위치 정보만으로 예측할 수 있기 때문에 해당 방식 사용한다. 반면 RGB 값은 물체를 바라보는 각도에 따라 달라질 수 있으므로 마지막 레이어에서 방향 정보를 추가하여 RGB 값을 예측한다.



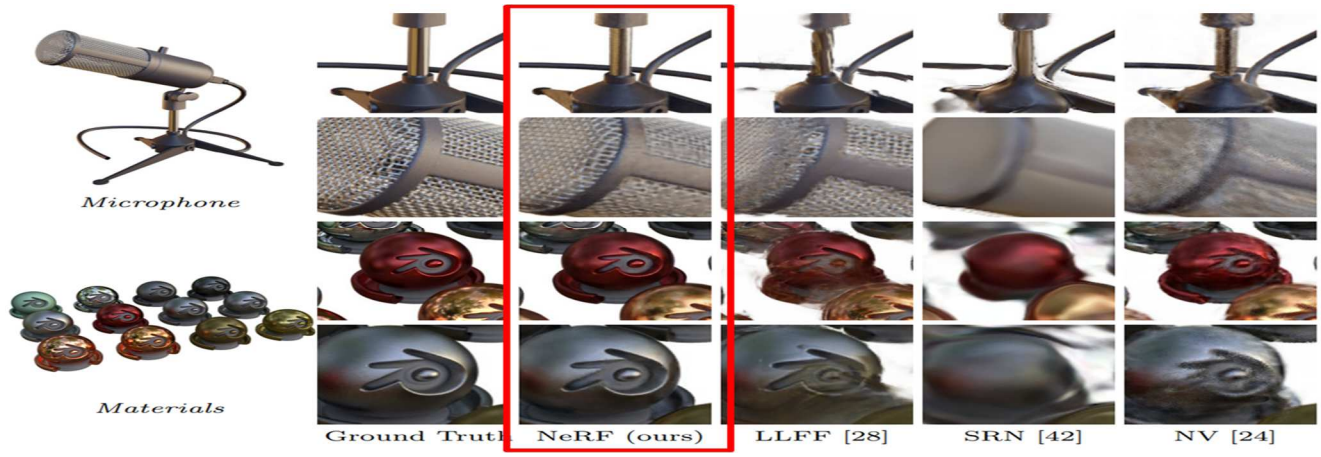
- 생성할 Novel View 의 카메라 중심으로부터 물체로 향하는 Ray 을 발사하고, 이 광선 상에서 여러 점들을 샘플링한다. 그런 다음, MLP 모델을 사용하여 각 좌표에서의 RGB 컬러와 밀도 값을 예측한다. 이때, 물체가 없는 위치에서는 밀도가 낮게 나오고, 물체가 있는 위치에서는 밀도가 높게 나온다. 이렇게 샘플링된 광선 상의 모든 점들의 RGB 와 밀도 값을 다시 원하는 뷰로 투영하고, 연산을 통해 최종 컬러 값을 계산한다.

- NeRF 를 최적화하기 위해서는 두 가지 방법이 사용된다. 첫 번째로, 저차원 정보는 고주파 변화를 표현하기에 부족하기 때문에, 저차원인 위치 정보와 방향 정보를 Positional Encoding 을 통해 고주파 Function 을 활용해서 고차원 공간에 Input 으로 Mapping 하여 고주파 Varation 을 포함하는 데이터가 적절하게 Fitting 되게한다.

- 두 번째로, Hierarchical Sampling 에서 균일 샘플링의 편향성 한계를 극복하기 위해 Stratified Sampling 을 사용한다. 각 격자 구간에서 랜덤하게 포인트를 샘플링하고, Coarse 샘플링 후 Fine 샘플링 과정에서 밀도가 높은 지점에서 추가 샘플링을 수행한다. 이 과정에서, Single Network 대신 Coarse 와 Fine 두 가지 Network 를 동시에 최적화하여, Coarse Network 로 평가한 결과를 바탕으로 추가 샘플링을 수행한다. Coarse Network 는 Ray 를 따라 샘플링 된 모든 색상의 가중치 합으로 재작성된다.

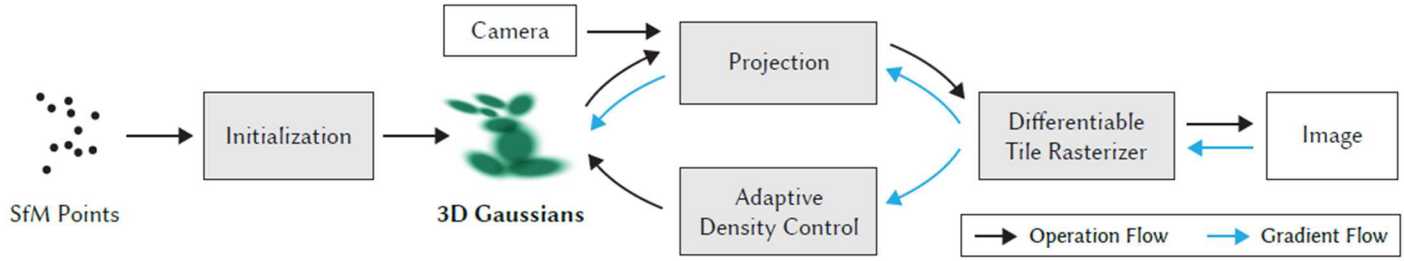
	DeepVoxels			**Generated**			**Real world**		
	Diffuse Synthetic 360° [41]			Realistic Synthetic 360°			Real Forward-Facing [28]		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SRN [42]	33.20	0.963	0.073	22.26	0.846	0.170	22.84	0.668	0.378
NV [24]	29.62	0.929	0.099	26.05	0.893	0.160	-	-	-
LLFF [28]	34.38	0.985	0.048	24.88	0.911	0.114	24.13	0.798	**0.212**
Ours	**40.15**	**0.991**	**0.023**	**31.01**	**0.947**	**0.081**	**26.50**	**0.811**	0.250

- 상기 표는 각 데이터셋에 대하여 이전 연구들과 이미지 유사를 정략적으로 비교한 것으로 NeRF 모델이 이전 모델들에 비해 우수함을 입증한다.



- NeRF 합성 데이터셋에서의 Novel 뷰 합성 결과로 NeRF 모델의 결과물의 우수함을 보여준다.

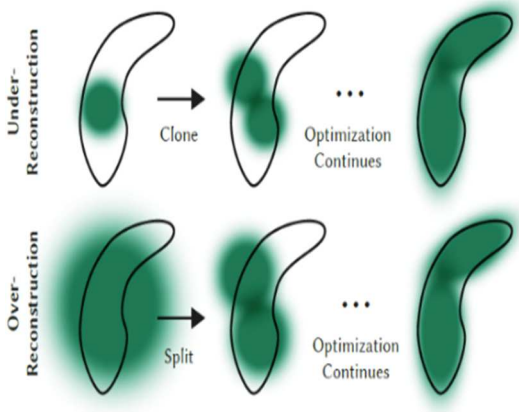
● *3D Gaussian Splatting* 의 접근 방식



- ***Initialization :*** SfM 알고리즘으로 얻은 Camera Pose 와 Point Cloud 정보를 Optimization 되기 위한 평균, 공분산, 투명도, SH 계수 파라미터를 갖는 Gaussian 으로 만든다.
- ***Projection :*** Ground Truth 이미지와 비교하여 Parameter 를 업데이트하기 위해 3D Gaussian 을 Image Plane 으로 사영시켜 2D Gaussian 으로 만든다.
- ***Adaptive Density Control :*** Gradient 를 기반으로 Gaussian 을 제거, 나누기, 복제한다.
- ***Differentiable Tile Rasterizer :*** 미분 가능한형태의 Tile Rasterization 을 통해 2D Gaussian 들을 하나의 Image 로 만든다.

$$\Sigma = RSS^TR^T$$

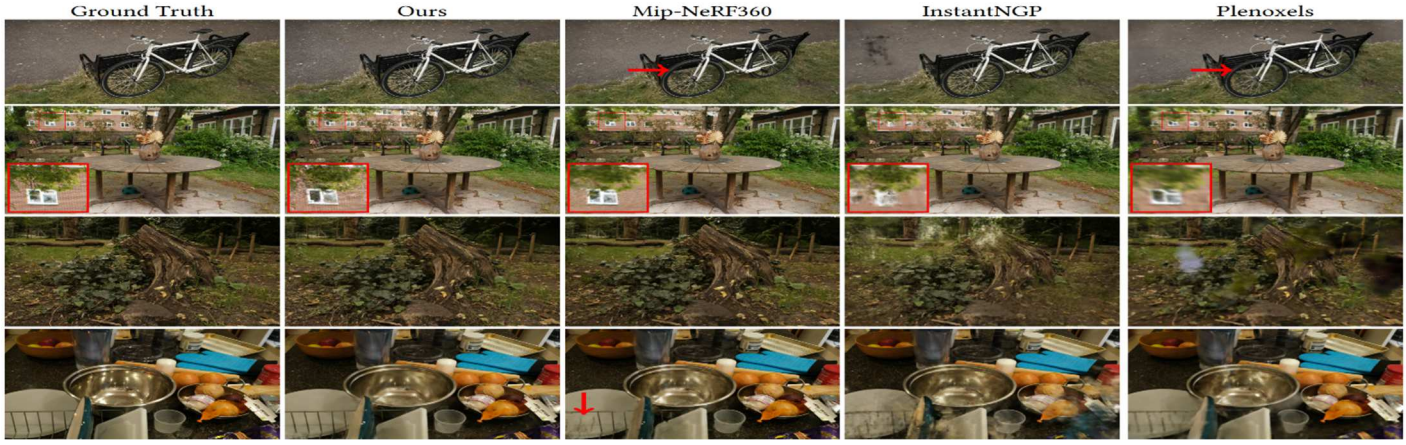
- 이때, Optimization 과정에서 공분산이 음수나 0 값을 갖게 되면 분포의 불안정성, 수치 계산 오류 등이 야기될 수 있다. 이를 해결하기 위해 공분산 행렬이 물리적인 의미를 갖도록 Scaling 과 Rotation 을 독립적으로 Optimization 하여 공분산 행렬이 Positive Definite 하게 표현식을 바꿔주었다.



- 또한, 불필요한 Gaussian 이 많아지거나 필요한 부분의 Gaussian 이 부족하면 재구성의 정밀도가 떨어지고 계산 비용이 증가하게 되는데, 이를 위해 Adaptive Density Control 에서 우선 threshold 보다 작은 α 값을 갖는 투명한 Gaussian 을 제거한다.

- 그런 다음, 특정 Threshold 값에 따라 Under-Reconstruction 된 영역은 Clone 하여 Positional Gradient 방향에 배치하고, Over- Reconstruction 된 영역은 분리하여 초기 Gaussian 의 확률밀도 값에 따라 배치한다.

- 이러한 방법으로 재구성의 정밀도를 향상시킨다.

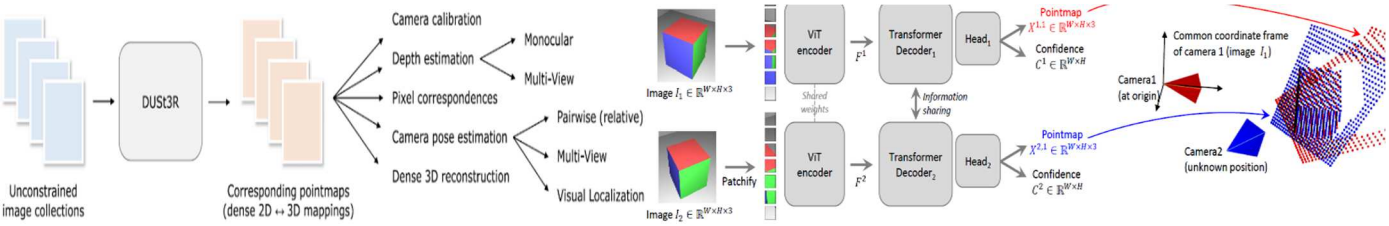


- Mip- NeRF360 데이터셋을 활용한 비교로 3DGS 모델의 결과물이 가장 실물과 비슷함을 보여준다.

Dataset	Mip-NeRF360						Tanks&Temples						Deep Blending					
Method|Metric	SSIM↑	PSNR↑	LPIPS↓	Train	FPS	Mem	SSIM↑	PSNR↑	LPIPS↓	Train	FPS	Mem	SSIM↑	PSNR↑	LPIPS↓	Train	FPS	Mem
Plenoxels	0.626	23.08	0.463	25m49s	6.79	2.1GB	0.719	21.08	0.379	25m5s	13.0	2.3GB	0.795	23.06	0.510	27m49s	11.2	2.7GB
INGP-Base	0.671	25.30	0.371	5m37s	11.7	13MB	0.723	21.72	0.330	5m26s	17.1	13MB	0.797	23.62	0.423	6m31s	3.26	13MB
INGP-Big	0.699	25.59	0.331	7m30s	9.43	48MB	0.745	21.92	0.305	6m59s	14.4	48MB	0.817	24.96	0.390	8m	2.79	48MB
M-NeRF360	0.792†	27.69†	0.237†	48h	0.06	8.6MB	0.759	22.22	0.257	48h	0.14	8.6MB	0.901	29.40	0.245	48h	0.09	8.6MB
Ours-7K	0.770	25.60	0.279	6m25s	160	523MB	0.767	21.20	0.280	6m55s	197	270MB	0.875	27.78	0.317	4m35s	172	386MB
Ours-30K	0.815	27.21	0.214	41m33s	134	734MB	0.841	23.14	0.183	26m54s	154	411MB	0.903	29.41	0.243	36m2s	137	676MB

- 각 세 Dataset 기준으로 기존 SOTA 모델들에 비해 성능이 매우 뛰어나며 FPS 성능이 압도적이다.

● *DUSt3R* 의 접근 방식



- ***ViT, Transformer Decoder :*** 대량의 데이터가 학습될 수 있도록 설계됐으며 동일한 네트워크가 서로 공유되는 Siamese 방법으로 Encoder 를 설계하면서 두 이미지에 대한 Feature 를 비교할 수 있게 한다. Transformer Decoder 에선 입력된 2 개의 이미지 Featrue 가 관계성을 학습할 수 있도록 해준다. 기준 View 에서 나온 Feature 는 Self- Attention 을 하고, 참조 View 에서 나온 Feature 는 Cross- Attention 을 하여, 정렬된 Pointmap 을 만들게 해준다.
- ***Head :*** Decoder Featrue 는 서로 다른 Head Network 를 통해, 각 Pixel 마다 Pointmap 과 신뢰도 값을 출력하게 된다.
- 이러한 구조로 3D Point 의 절대적인 위치 추정없이 상대적인 위치를 추정하게 하여 카메라 파라미터 없이 3D 를 Reconstruction 을 할 수 있게 한다.
- Loss 계산 시 반투명 영역에서는 3D Point 를 정확하게 예측할 수 없어서 이것을 구분하기 위해 신뢰도 값을 사용한다. 확실한 물체의 Point 에 대해서는 신뢰도 값이 높게 나오고, 불확실한 물체에 해당하는 Point 에 대해서는 신뢰도 값이 낮게 나오도록 한다.
- 2 개 이미지에 대한 Pointmap 을 겹쳐두었을 때, 정확히 Point Matching 되지 않는 문제가 발생한다. 이를 해결하기 위해 Nearest Neighbor Search 알고리즘을 통해 2 개 이미지내 Pixel 간의 관계성을 만들게 되며 결과적으로 World 좌표계에서 3D Space 에 배치된 Point 가 가까운 것끼리 관계성을 갖게 된다.
- DUSt3R 을 통해 3D Reconstruction 뿐만 아니라 Camera Calibration, Depth Estimation, Pixel Correspondences, Camera Pose Estimation 등 다양한 Task 에 적용이 가능하다.

Methods	Train	Outdoor						Indoor			
		DDAD[40]		KITTI [35]		BONN [79]		NYUD-v2 [114]		TUM [118]	
		Rel↓	$\delta_{1.25}$ ↑	Rel↓	$\delta_{1.25}$ ↑	Rel↓	$\delta_{1.25}$ ↑	Rel↓	$\delta_{1.25}$ ↑	Rel ↓	$\delta_{1.25}$ ↑
DPT-BEiT[90]	D	10.70	**84.63**	9.45	89.27	-	-	**5.40**	**96.54**	**10.45**	**89.68**
NeWCRFs[173]	D	**9.59**	82.92	**5.43**	**91.54**	-	-	6.22	95.58	14.63	82.95
Monodepth2 [37]	SS	23.91	75.22	11.42	86.90	56.49	35.18	16.19	74.50	31.20	47.42
SC-SfM-Learners [6]	SS	16.92	77.28	11.83	86.61	21.11	71.40	13.79	79.57	22.29	64.30
SC-DepthV3 [120]	SS	**14.20**	**81.27**	11.79	86.39	**12.58**	**88.92**	**12.34**	**84.80**	**16.28**	**79.67**
MonoViT[181]	SS	-	-	**09.92**	**90.01**	-	-	-	-	-	
RobustMIX [91]	T	-	-	18.25	76.95	-	-	11.77	90.45	15.65	**86.59**
SlowTv [116]	T	**12.63**	79.34	(6.84)	(56.17)	-	-	11.59	87.23	15.02	80.86
DUSt3R 224-NoCroCo	T	19.63	70.03	20.10	71.21	14.44	86.00	14.51	81.06	22.14	66.26
DUSt3R 224	T	16.32	77.58	16.97	77.89	11.05	89.95	10.28	88.92	17.61	75.44
DUSt3R 512	T	13.88	81.17	**10.74**	**86.60**	**8.08**	**93.56**	**6.50**	94.09	**14.17**	79.89

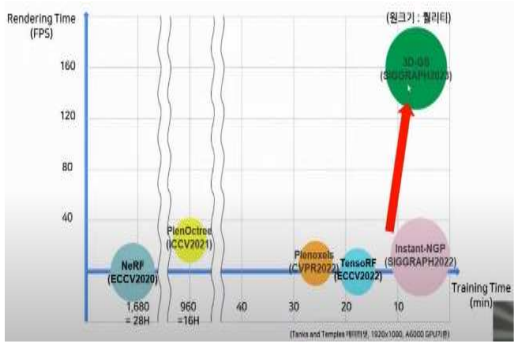
Methods	N Frames	Co3Dv2 [93]			RealEstate10K [185]
		RRA@15	RTA@15	mAA(30)	mAA(30)
COLMAP+SPSG	3	~22	~14	~15	~23
PixSfM	3	~18	~8	~10	~17
Relpose	3	~56	-	-	-
PoseDiffusion	3	~75	~75	~61	- (~77)
DUSt3R 512	3	95.3	**88.3**	**77.5**	**69.5**
COLMAP+SPSG	5	~21	~17	~17	~34
PixSfM	5	~21	~16	~15	~30
Relpose	5	~56	-	-	-
PoseDiffusion	5	~77	~76	~63	- (~78)
DUSt3R 512	5	95.5	86.7	76.5	67.4
COLMAP+SPSG	10	31.6	27.3	25.3	45.2
PixSfM	10	33.7	32.9	30.1	49.4
Relpose	10	57.1	-	-	-
PoseDiffusion	10	80.5	79.8	66.5	48.0 (~80)
DUSt3R 512	10	**96.2**	86.8	76.7	67.7

- 실내와 실외 깊이 추정 Benchmark 에서 대체적으로 우수한 성능을 보이고 있다.
- Co3Dv2, RealEstate10K 데이터셋에서의 Pose Regression 평가 지표에서도 Relative Rotation Accuracy 와 Relative Translation Accuracy, Mean Average Accuracy 부분 모두 COLMAP 등의 기존 SfM 알고리즘과 비교하여 가장 뛰어난 성능을 보여준다.

✓ **접근방식의 유사점**
- 세 모델 모두 3D 장면이나 객체의 재구성을 목표로 한다.
- 다양한 시점에서 촬영된 2D 이미지를 활용하여 3D 구조를 추정한다.
- 3D Gaussian Splatting 과 NeRF 의 경우 COLMAP 과 같은 SfM 알고리즘 과정이 필요하다.

✓ **접근방식의 차이점**
- 3D Gaussian Splatting 의 경우 NeRF 와 DUSt3R 와 달리 네트워크를 사용하지 않는다.
- DUSt3R 는 3D Gaussian Splatting 과 NeRF 와 달리 카메라 캘리브레이션 과정이 없다.
- DUSt3R 의 경우 3D Reconstruction 뿐만 아니라 다양한 3D Vision Task 에도 적용이 가능하다.

● *Discussion*



- NeRF 는 고해상도 이미지를 생성할 수 있어 시각적으로는 매우 정밀하지만, 실시간성이 떨어지며 잘못 학습됐을 때 모델에 치명적인 왜곡이 발생한다. NeRF 의 경우 Stochastic Sampling 으로 노이즈가 발생하며 연산량이 많고, 학습 및 렌더링이 오래 걸리는데, 3D Gaussian Splatting 은 표현 방법을 변경하여 실시간 렌더링 속도 측면에서 매우 뛰어난 성능을 보이며 Gaussian 을 활용하여 물체에 대한 왜곡 현상이 NeRF 에 비해 현저히 줄어들었다. 또한, 3DGS 의 등장으로 기존 NeRF 모델에서는 추출이 어려웠던 투명 재질에 대한 정확도도 개선할 수 있다고 생각한다.
- DUSt3R 의 경우 Camera Extrinsic 뿐만 아니라 Intrinsic 파라미터도 필요하지 않기에, NeRF 나 3DGS 에 비해 다양한 환경에서 활용 가능하며, 3D 재구성 및 다른 Task 에도 적용 가능하여 활용 범위가 넓다.
- 일반적인 퀄리티나 학습시간은 대체적으로 3DGS 가 가장 뛰어나며 실시간성도 압도적이라고 생각하며, 표면부의 세부 퀄리티와 조명/그림자 효과부분은 NeRF 모델이 현재 기준으로는 더 우월하나 개선된다면 이 부분도 3DGS 가 더 좋은 성능을 보일 수 있다고 예상한다. 활용성과 실용성 측면에서는 제약조건은 적으면서 적용 가능한 Task 가 많은 DUSt3R 모델이 최적이라고 생각한다.

■ Review

1. NeRF

- ***Summary*** : NeRF 는 신경망을 활용하여 2D 이미지 세트를 기반으로 3D 장면을 고해상도로 재구성하며 다양한 시점에서의 뷰를 제공하는 방법을 제시한다. NeRF 는 특정 위치와 시점에서의 색상과 방사율을 예측하는 모델로, MLP 를 사용하여 연속적인 공간에서 이 정보를 학습한다. 고해상도의 Photo Realistic 한 이미지를 생성할 수 있으며 입력되지 않은 새로운 시점에서의 뷰를 생성해준다.
- ***Strengths and Weaknesses*** : 기존의 3D 표현 기술과 MLP 구조를 결합한 구조다. 기존의 뷰 합성 기법들은 주로 Mesh 또는 Voxel 기반의 불연속적인 표현을 사용했지만, NeRF 는 연속적인 공간에 대한 표현을 사용하기 때문에 View 이동이 자연스러우며 기존 방식에 비해 저장 공간에 대한 부담이 적다. 컬러값 뿐만 아니라 Volume Density 까지 추정하기 때문에 더욱 사실적인 결과를 뽑아내 준다. 그러나 실시간으로 적용하기엔 부적합하며, 많은 데이터를 요구해서 실용성의 한계가 존재한다.
- ***Questions***: Data 가 되는 객체의 어떤 Feature 가 주로 결과물의 품질에 영향을 미치는지가 궁금하다. 시간이라는 추가적인 차원의 입력을 주어 시간에 따른 동적인 장면에 대해서도 렌더링을 구현할 수 있고, 추가적으로 적은 입력 영상만으로도 3D 렌더링이 가능하다 가정한다면, 이를 CCTV 영상 등에 적용하여 더 사실적인 CCTV 영상을 얻는 것이 가능한 것인지 궁금하다.
- ***Limitations***: Stochastic Sampling으로 노이즈가 발생하며 연산량이 많고 학습과 렌더링이 오래 걸린다. NeRF는 단 하나의 장면을 최적화하기 위해 약 100에서 300K번의 Iteration을 필요로 한다. 많은 입력 영상 또는 이미지가 필요하다. 적은 수의 입력 만으로도 렌더링이 가능하도록 개선이 필요하다. Static이 아닌 Dynamic한 물체 렌더링에 한계가 존재하는데. 현실 세계는 Dynamic의 경우가 더 많기에 시간을 고려한 표현도 가능하도록 개선한다면, 실용성이 더욱 증가할 것이다.

2. 3D Gaussian Splatting

- ***Summary***: NeRF 와 같은 문제를 해결하고자 하며 여러 이미지와 카메라 Pose 가 주어지면 다양한 시점에서 3D Scene 을 렌더링한다. 고해상도에서 Rendering Quality 가 SOTA 인 Mip- NeRF360 보다 뛰어나며, Training Time 에서 SOTA 인 InstantNGP 보다도 시간을 단축시켰다. MLP 를 사용하지 않아 연산량이 비교적 적고, 실시간 렌더링 성능을 보여준다.
- ***Strengths and Weaknesses***: GPU 자원을 효율적으로 활용하여 실시간으로 3D 장면을 렌더링 해준다. MLP 를 사용하지 않으면서 Explicit 하게 표현을 하여 NeRF 와 비교해 연산량이 많이 줄어들었다. NeRF 기반 모델보다 메모리 요구량이 많다.

- ***Questions***: NeRF 의 경우에는 좋은 결과물을 얻기 위한 최적의 입력 영상 촬영 궤적이 존재하는데, 3DGS 의 경우에도 이러한 입력 영상을 위한 최적의 촬영 궤적이 존재하는지 또는 NeRF 궤적과 차이점이 있는지가 궁금하다.
- ***Limitations***: 다른 기존 모델들처럼, 잘 안 보이는 장면에 대해서는 퀄리티 있는 렌더링에 어려움이 존재한다. 메모리 요구량이 많아 고성능 하드웨어가 필요하다. 입력 이미지가 부족한 Scene 에서 큰 타원 형태의 시각적 Artifacts 가 생길 수 있다. Depth 가 갑자기 변화하는 구간에서 생기는 Artifact 에 대해서는 Antialiasing 을 활용하면 개선이 될 것으로 보여진다. Regularization 을 도입한다면 보이지 않은 지역과 Artifact 발생 영역을 잘 처리할 수 있을 것이다.

3. DUSt3R

- ***Summary***: DUSt3R 모델은 여러 이미지로 3D 재구성을 수행하는 새로운 방법론을 제안한다. 카메라 캘리브레이션이나 시점 정보를 사전에 필요로 하지 않고, 임의의 이미지 쌍으로 3D 재구성을 수행하는 패러다임을 제시한다. Transformer 인코더와 디코더를 기반으로 이미지 쌍을 입력 받아 3D Pointmap 을 예측하며 여러 이미지가 제공되면 모든 Pointmap 을 전역 정렬로 공통 참조 프레임으로 표현한다. 이를 통해 깊이 정보와 3D 모델을 직접 제공하며, 3D 재구성뿐만 아니라 다른 3D Vision Task 에 적용할 수 있다. Monocular 과 Multi- View 깊이 추정에서 SOTA 의 성능을 보여준다.
- ***Strengths and Weaknesses***: 기존의 MVS 알고리즘과 다르게 카메라 캘리브레이션이나 시점 정보를 사전에 필요로 하지 않기에 기존의 제약을 극복하고, 더 많은 상황에서 적용 가능할 수 있다. 단일과 다중 뷰 재구성을 통합할 수 있는 방법론의 제시로 독창성을 보여준다. 카메라 포즈 추정과 깊이 추정 등 다양한 3D Vision 작업을 단일 모델로 통합하여 처리할 수 있게 하여 다양한 응용 분야에서 유용함을 보여준다. 실내외 모든 환경에서 안정적으로 동작한다.
- ***Questions***: 비교적 원경에서도 잘 작동하는 것 같은데, 어느 정도의 거리에 있는 물체까지 Pointmap 을 출력할 수 있는지? 모델을 경량화한다면, Visual- SLAM 에 적용하여 실시간 응용을 할 수 있는지? 로보틱스에 적용한다면 매핑이나 경로 설정 등에는 도움이 되겠지만, 사람이나 다른 로봇들이 움직이는 현실 세계에서 주행이나 물체 파지 등에도 적용이 가능한 것인지?
- ***Limitations***: 동적인 물체나 장면에 대한 처리가 어렵다. Transformer 네트워크를 기반하므로 많은 계산 자원을 요구하므로 실시간 응용에는 한계가 존재하여 모델의 경량화가 필요할 것으로 보인다. 또한 가려짐이 심한 상황에서는 정확도가 떨어질 수 있어 Occlusion 에 대한 처리 기법으로 견고성을 추가할 필요가 있다. DUSt3R 모델이 Lidar 등의 다른 센서와 융합한 Sensor Fusion 을 통해 더욱 정확하고 견고한 3D 재구성 및 3D Vision Task 를 가능하게 할 수 있다고 생각한다.

■ Future Work

- 현재 3DGS 의 경우 수많은 Gaussian 파라미터를 저장하느라 GPU 메모리가 20GB 가까이도 사용되는 경우가 있어 권장 GPU 메모리가 24GB 로 매우 큰 편에 속하는데, 해당 수준의 메모리가 탑재된 GPU 는 매우 고가여서 일반인 수준에서는 이를 활용하거나 구현하는 데 큰 제약이 존재한다. 이를 최적화하여 메모리 사용량을 줄일 수 있다면, 보다 적은 자원으로 활용할 수 있게 되어 실용성도 증가하고 많은 사람들이 접근할 수 있어 연구개발 속도도 더 빨라질 것이라 생각된다.
- NeRF 와 3DGS 모두 아직 최소 입력 이미지 개수가 많이 필요한데, 서비스적으로 활용도를 높이려면 Sparse 한 입력 개수와 속도에 대한 개선도 더 필요하다고 생각한다.
- 세 모델 모두 정적인 장면을 재구성할 뿐, 동적인 장면에서는 한계가 있는데 시계열 데이터에 적합한 모델을 사용하여 시간이라는 Dimension 을 추가해 동적인 객체나 장면도 재구성하게 된다면 조금 더 현실세계에서 실생활에 적용할 수 있는 부분이 늘어날 수 있을 것으로 생각한다.
- 실제로 NeRF 모델 구현 시 COLMAP 을 통한 카메라 파라미터 추출 과정에서 많은 시간이 걸리는데, 이를 DUSt3R 모델과 결합한다면 NeRF 와 3DGS 모델의 COLMAP 의존성도 낮추어 제약을 완화시킬 수도 있고 학습 소요시간도 줄일 수 있지 않을까 생각한다.
- 3D Reconstruction 과정에서 실시간 렌더링 능력과 Zero 또는 Few Shot Diffusion 등의 모델을 결합한다면, Geometric 적으로 보지 못한 부분의 사물 모양을 추론하여 보다 효율적인 로봇 팔의 파지 궤적과 조작 제어 학습이 가능할 것으로 예상한다.

■References

[1] Ben M, Pratul P. S. Matthew T., Jonathan T. B., Ravi R and Ren N."NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," ECCV 2020.

[2] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuhler, ¨ and George Drettakis. 3d gaussian splatting for real- time radiance field rendering. ACM Transactions on Graphics 2023.

[3] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, Jerome Revaud, "DUSt3R: Geometric 3D Vision Made Easy," to appear in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.