다 시점 영상 콘텐츠 특성에 따른 딥러닝 기반 깊이 추정 방법론

손호성[1], 신민정[1], 김준수[2], 윤국진[2], 정원식[2], 이현우[2], 강석주[1)†]

{hsson97, hse03197, sjkang}@sogang.ac.kr[1], {joonsookim, kjyun, wscheong, hwlee}@etri.re.kr[2]

Deep learning-based Multi-view Depth Estimation Methodology of Contents' Characteristics

Hosung Son[1], Minjung Shin[1], Joonsoo Kim[2], Kug-jin Yun[2], Won-sik Cheong[2], Hyun-woo Lee[2],

and Suk-ju Kang[1)†]

요 약

최근 다 시점 영상 콘텐츠 기반 3차원 공간(장면) 복원을 위한 다 시점 깊이 추정 딥러닝 네트워크 방법론이 널리 연구되고 있다. 다 시점 영상 콘텐츠는 촬영 구도, 촬영 환경 및 세팅에 따라 다양한 특성을 가지며, 고품질의 3차원 복원을 위해서는 이러한 특성을 이해하고, 적절한 깊이 추정 네트워크 기법들을 적용하는 것이 중요하다. 다 시점 영상 촬영 구도로는 수렴형, 발산형이 존재하며, 촬영 세팅에는 카메라 시점 간 물리적 거리인 baseline이 있다. 본 연구는 이와 같은 다 시점 영상 콘텐츠의 종류와 각 특징에 기반하여 콘텐츠(데이터 셋)의 특성에 따른 적절한 깊이 추정 네트워크 방법론을 다룬다. 실험 결과로부터, 기존의 다 시점 깊이 추정 네트워크를 발산형 또는 large baseline 특성을 가지는 데이터 셋에 곧바로 적용하는데 한계점이 존재함을 확인하였다. 따라서, 각 영상 환경에 적합한 '참조 시점 개수' 및 적절한 '참조 시점 선택 알고리즘'의 필요성을 검증하였다. 결론적으로, 3차원 공간(장면) 복원을 위한 딥러닝 기반 깊이 추정 네트워크 구현 시, 본 연구 결과가 다 시점 영상 콘텐츠 기반 깊이 추정 기법 선택에 있어 가이드라인으로 활용될 수 있음을 확인하였다.

Abstract

Recently, multi-view depth estimation methods using deep learning network for the 3D scene reconstruction have gained lots of attention. Multi-view video contents have various characteristics according to their camera composition, environment, and setting. It is important to understand these characteristics and apply the proper depth estimation methods for high-quality 3D reconstruction tasks. The camera setting represents the physical distance which is called baseline, between each camera viewpoint. Our proposed methods focus on deciding the appropriate depth estimation methodologies according to the characteristics of multi-view video contents. Some limitations were found from the empirical results when the existing multi-view depth estimation methods were applied to a divergent or large baseline dataset. Therefore, we verified the necessity of obtaining the proper number of source views and the application of the source view selection algorithm suitable for each dataset's capturing environment. In conclusion, when implementing a deep learning-based depth estimation network for 3D scene reconstruction, the results of this study can be used as a guideline for finding adaptive depth estimation methods.

1) 서강대학교 전자공학과(Department of Electronic Engineering, Sogang University)

2) 한국전자통신연구원 (Electronics and Telecommunications Research Institute)

† Corresponding Author : 강석주 (Suk-ju Kang)

E-mail: sjkang@sogang.ac.kr

Tel: +82-2-705-8466

※ This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2022-0-00022), Development of immersive video spatial computing technology for ultra-realistic metaverse services), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1004208).

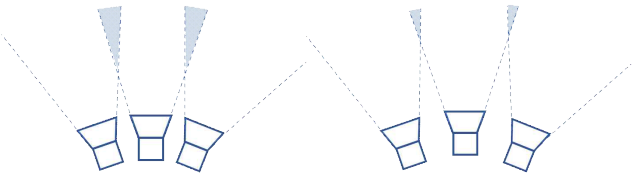


그림 1. 발산형 카메라 기준 baseline에 따른 시점 간 공유 정보 차이
(L: Small baseline, R: Large baseline)

1. 서론

가. 다 시점 영상 콘텐츠

다 시점(multi-view) 영상 콘텐츠는 3차원 장면 혹은 물체의 복원(reconstruction)을 위해 여러 위치 및 방향에서 동시에 촬영하여 얻은 영상 데이터를 말한다. 이는 정지 영상 혹은 한 공간을 여러 방향으로 스캔하듯 촬영한 동영상 데이터일 수 있다.

이러한 다 시점 영상 콘텐츠는 촬영 시점 종류 및 촬영 기준에 따라 각기 다른 특성을 가진다. 촬영 시점에 따른 분류 중 하나로, 수렴형(convergent) 데이터는 주로 정지 객체(static object) 혹은 장면(scene) 주변에 여러 대의 카메라를 수렴형으로 배치하여 취득된다. 대부분 다 시점 기반 데이터들은 이와 같은 카메라 배치 구조 상태에서 취득되었으며, 대표적으로 DTU [1], Tanks&Temples [2] 데이터 셋이 존재한다. 한편, 발산형(divergent) 데이터의 경우, 정지 객체나 장면에 대해 여러 대의 카메라로 영상을 취득하는 것은 수렴형과 동일하지만, 카메라 촬영 시점을 대상에 대하여 발산형으로 배치하여 시점 간 공유되는 정보가 수렴형에 비해 부족하다는 특성이 있다. 이러한 이유로 발산형 카메라 배치로 촬영된 다 시점 영상은 수렴형에 비해 영상 정보가 충분하지 않다. 따라서, 발산형 배치에 대해서는 별도로 고려될 필요가 있다.

한편, 앞서 설명한 수렴형, 발산형 특성의 영상을 모두 포함하면서 RGB-D 단일 카메라로 취득된 데이터도 존재한다. 대표적으로 ScanNet [3]이 있는데, 해당 데이터는 동영상을 기반으로 취득되었으며, 카메라 움직임을 조밀하게 두어 이미지 프레임의 시점 간 포즈 거리가 좁은 특징을 보인다.

이처럼 다 시점 영상 데이터는 그 종류가 다양하며, 그에 따른 특성 또한 다양하게 존재한다. 따라서 본 연구는 앞서 언급한 수렴형, 발산형 데이터 셋의 특성에 따라 여러 깊이 추정 네트워크 구동 결과를 비교, 분석하여 네트워크 성능 개선을 위한 방안을 제시할 예정이다.

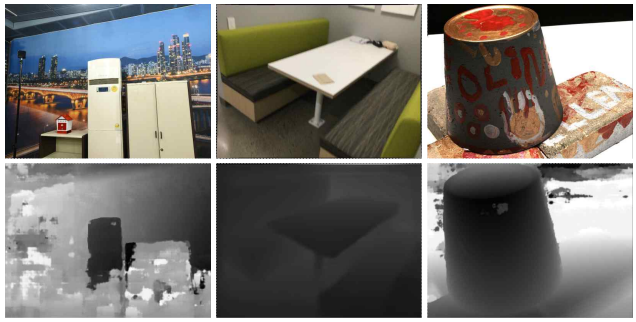
나. 취득 영상의 카메라 간 물리적 거리

Baseline은 영상 취득 과정에서 각 시점 별 카메라의 물리적인 거리를 의미한다. 이때, 그림 1에서와 같이 baseline이 클 때가 작을 때보다 각 시점 간 공유되는 이미지 정보가 적다는 것을 확인할 수 있다. 이와 같은 특성으로 인해, 특히 발산형 데이터의 경우, baseline의 크기가 깊이 추정 네트워크 성능에 큰 영향을 미치게 된다. 이를 해결하기 위해서는 시점 간 공유 정보를 최대한 활용하는 방향으로 깊이 추정 네트워크 방법론 적용이 필요하다. 그러므로 실험에서는 baseline이 큰 발산형 데이터 셋 깊이 추정 결과와 제안 기법을 적용하여 얻은 결과를 비교하여 제안 기법을 적용했을 때 깊이 추정 성능을 개선할 수 있음을 보이고자 한다.

다. 딥러닝 기반 깊이 추정

딥러닝 기반 깊이 추정 네트워크는 참조 시점 이미지를 기준 시점 이미지로 식 (1)과 같이 homography warping [4]을 수행하여 생성한 매칭 비용 볼륨(matching cost volume)[5, 6]에 확률 기반 방식으로 각 픽셀별 깊이 라벨링을 진행하고, edge와 같이 가장 두드러지는 확률에 해당하는 깊이 라벨이 불확실한 영역에 대해서 깊이 추정 회귀(depth estimation regression)[4, 7, 8]의 방식으로 픽셀 당 깊이를 추정한다.

$$H_i(d) = K_i \cdot R_i \cdot (I - \frac{(t_1 - t_i) \cdot n_i^T}{d}) \cdot R_1^T \cdot K_1^{-1} \quad (1)$$



(a) (b) (c)

그림 2. 수렴형-발산형 데이터 셋 구동 결과
(a) 발산형-CasMVS, (b) 수렴형-DVMVS, (c) 수렴형-CasMVS

본 연구에서는 이와 같은 딥러닝 기반 깊이 추정 네트워크를 이용하여 다 시점 깊이 추정 연구를 진행하였다. 우선, 정지 영상 다 시점 깊이 추정 네트워크로써 CasMVS [9]를 이용하였다. [9]의 저자는 종속적 비용 볼륨(cascade cost volume) 생성 네트워크 구조를 제안하였는데, 피라미드 특성맵 네트워크(feature pyramid network) [10]를 사용하여 추출한 다중 해상도 계층적 특성맵을 기반으로 네트워크의 각 단계(stage)에서 깊이 맵을 추정하는 방식을 따른다. 본 연구에서는 이와 같은 네트워크들을 사용하여 정지 영상 수렴, 발산형 데이터 셋에 대한 깊이 추정 실험을 진행하였다. 한편, 동영상 기반의 다 시점 영상 데이터의 깊이 추정을 위해 시간적 일관성(temporal consistency)[11, 12, 13, 14, 15]을 고려하는 DeepVideoMVS [16](이하 DVMVS) 네트워크를 사용하였다. DVMVS [16]는 실시간 구동이 가능하도록 일반적인 3D-CNN 기반의 깊이 정제 네트워크와는 달리 2D U-Net [17] 구조인 인코더(encoder)-디코더(decoder) 구조를 사용하였으며, ConvLSTM [18, 19] 모듈을 인코더와 디코더 사이에 추가하여 해당 네트워크가 이미지 프레임 간 시간적 일관성을 고려할 수 있도록 미세 조정(fine-tuning) 과정을 포함하였다. 이를 통해 동적 움직임이 존재하는 동영상 기반의 다 시점 영상 콘텐츠에 대해서도 우수한 성능 내었다.

2. 제안한 방법론

그림 2는 각 수렴, 발산형 데이터 셋에 대해서 CasMVS [9]로 구동한 결과이다. 실험 결과, 발산형 데이터 셋에 대해서는 깊이 추정 성능이 우수하지 못함을 확인하였고, 이에 따라 깊이 추정 정확도를 개선하기 위해 다음과 같이 추가적인 방안을 고안하였다.

가. 참조 시점 개수

다 시점 깊이 추정 기법에서 참조 시점에서는 영상의 특징(수렴형, 발산형)에 따라 적절한 참조 시점의 개수를 파악하여 깊이 추정 성능을 향상시키는 작업이 요구된다. 참조 시점이란 다 시점 깊이 추정을 위해 필요한 매칭 비용 볼륨(matching cost volume)을 생성하는데 사용되는 기준 시점 이외의 나머지 시점들을 일컫는다. 일반적으로, 다 시점 영상은 시점 간 공유 정보(shared image information)가 존재하고, 해당 정보를 최대한 활용할수록 3차원 영상 복원의 품질을 높일 수 있다.

하지만, 참조 시점의 개수를 일정 기준 이상 확장한다면 오히려 기준 시점과 공유되지 않는 정보를 많이 포함하는 시점들까지 네트워크 학습에 활용될 가능성이 존재한다. 따라서, 다 시점 데이터 셋의 특성(수렴형, 발산형)에 따른 적절한 참조 시점의 개수를 파악하여 깊이 추정 성능을 향상시킬 수 있도록 연구를 진행하였다. 수렴형 배치의 경우 시점 개수를 증가하여도 성능이 곧바로 개선되는 양상이 관찰되지 않으나, 발산형 배치의 경우에는 참조 시점의 개수를 증가시키는 것이 성능 향상에 도움이 되었음을 확인하였다. 그리고 이와 관련해 더욱 상세한 결과는 아래의 실험 결과에서 후술할 예정이다.

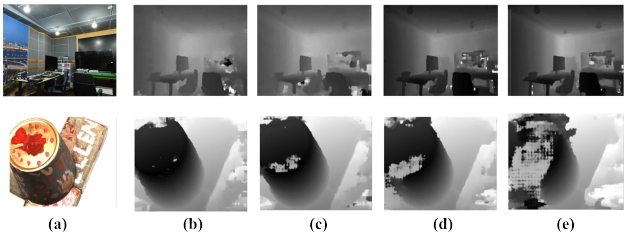


그림 3. (Up) 발산형 데이터 셋 결과 (참조 시점 개수: 3, 5, 8, 24)
(Down) 수렴형 데이터 셋 결과 (참조 시점 개수: 3, 5, 10, 24)

나. 참조 시점 선택 알고리즘

다 시점 깊이 추정 기법은 어떤 참조 시점을 활용하는지에 따라 깊이 추정 성능에 큰 영향을 끼친다. 본 논문에서는 적절한 참조 시점 선택 알고리즘을 이용하여 깊이 추정을 진행할 필요성을 확인하였고, 이를 위해 적절한 참조 시점 선택 알고리즘 활용 방안을 제안하였다.

참조 시점 선택 기법의 종류로는 발산형 데이터 셋에 대한 비용 볼륨(cost volume) 공유 정도에 따른 최인접 시점 선택 기법과 ScanNet [3]과 같은 동영상 데이터에서 주로 사용되는 식 (2)의 카메라 포즈 거리(camera pose distance) [20] 기반 참조 시점 선택 알고리즘 기법이 존재한다.

식 (2)에서, t_{rel}은 프레임 간 상대적 평행이동 벡터(translation vector)를 의미하고, R_{rel}은 상대적 회전 행렬(rotation matrix)를 의미한다.

$$dist[T_{rel}] = \sqrt{\|t_{rel}\|^2 + \frac{2}{3} tr(I - R_{rel})} \quad (2)$$

실험에 사용된 발산형 데이터 셋의 경우, baseline 길이가 큰 특성도 존재하기 때문에 기준 시점에 대하여 주변의 참조 시점과의 공유 정보가 수렴형에 비해 부족하여 시점 간 공유 정보를 최대로 활용하는 것이 필수적이므로 최인접 시점을 참조 시점으로 선택하는 것이 적절하다.

반면, 동영상 기반 데이터 셋의 경우 baseline이 작으므로 프레임 간 공유 정보와 비 공유 정보가 적절히 포함된 참조 시점을 사용할 필요가 존재한다. 최인접 프레임을 사용할 경우 공유 정보의 비중이 지나치게 높아지게 되어 추정 성능의 저하를 발생시킬 수 있기 때문이다. 다시 말해, 최인접 시점 선택이 아닌 카메라 포즈 거리 기반의 참조 시점 선택이 동영상 기반 다 시점 깊이 추정에 적절하다고 할 수 있다. 한편, 깊이 추정의 정성적 결과를 중심으로 하는 내용은 실험 결과에서 후술할 예정이며, 그로부터 데이터 셋 특성에 따른 적절 참조 시점 선택 알고리즘을 확인할 수 있다.

3. 실험

본 실험에서는 다 시점 영상 콘텐츠의 특성 중 baseline 길이와 촬영 시점 종류에 따른 딥러닝 깊이 추정 네트워크의 구동 결과를 토대로 데이터 셋 특성별 네트워크 적합성을 평가하고자 한다.

수렴형(convergent)-발산형(divergent) 데이터 셋 특성 차이에 기반하여 CasMVS [9]와 DVMVS [16]의 성능을 평가하고, 각 상황에서의 문제점을 파악한 후, 제안 기법을 적용해 깊이 추정 성능을 개선하였다.

가. 참조 시점 개수 탐색 실험

그림 3을 보면, 적은 참조 시점 개수로 CasMVS [9]를 구동하였을 때 수렴형보다 발산형 데이터 셋에서 성능이 좋지 못했다. 그러나 그림 3. (a)에서 (e)로의 결과 변화와 같이 참조 시점 개수을 증가시킬 때에는 오히려 발산형에서 우수한 결과를 얻게 되었다. 이를 통해 데이터 셋 특성에 따라 참조 시점 개수를 적절히 설정하면 깊이 추정 성능을 개선할 수 있으며, 동시에 어떠한 데이터 셋의 최적 참조 시점의 개수가 다른 모든 데이터 셋에서도 동일하게 적용될 수는 없음을 확인하였다.

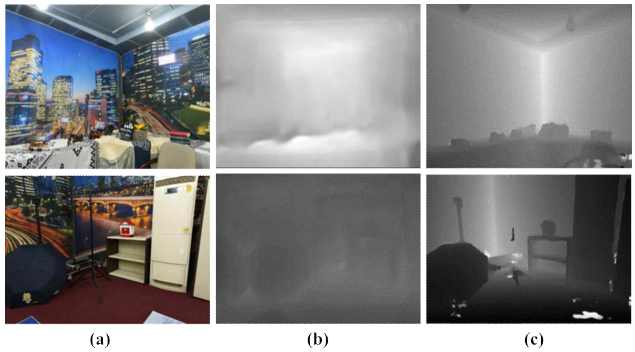


그림 4. 참조 시점 선택 알고리즘 적용 결과
(a) RGB 이미지, (b) DVMVS, (c) CasMVS

결론적으로. 딥러닝 기반의 다 시점 깊이 추정에 있어 목표 도메인(target domain) 특성에 따라 최적 참조 시점의 개수가 다를 수 있고, 해당 사실을 기반으로 데이터 셋의 특성과 맞는 참조 시점 개수를 적절히 설정한다면, 우수한 성능을 얻을 수 있음을 확인하였다.

나. 참조 시점 선택 알고리즘 검토 실험

최적 참조 시점 개수 뿐만 아니라 어떤 참조 시점이 깊이 추정에 사용되는지도 다 시점 깊이 추정 성능에 큰 영향을 미친다. 참조 시점 선택 알고리즘은 카메라 포즈 거리 기반 기법과 최인접시점 선택 알고리즘 등이 존재하며, 목표 도메인(target domain)에 따라 적절한 시점 선택 알고리즘이 다를 수 있다.

실험에서는 baseline이 큰 발산형 데이터 셋을 사용하였기 때문에, 시점 간 공유 정보가 부족하여 적절한 참조 시점 이미지를 선택하는 것이 더욱 중요하다. 실험에서는 최인접 시점 선택 기법과 카메라 포즈 기반 기법 중 특히 baseline이 큰 발산형 데이터 셋에서 어떠한 시점 선택 알고리즘이 적절한지 규명하였다.

그림 4.(b)는 DVMVS [16]는 카메라 포즈 거리 기반의 참조 시점 선택 기법을 사용했을 때, 발산형 데이터 셋에 대한 깊이 추정 결과이다. 그리고 그림 4.(c)는 CasMVS [9]에 대하여 최인접 시점 선택 알고리즘을 이용한 결과이다. 두 실험 결과를 비교했을 때, 실험에서 사용한 baseline이 큰 발산형 데이터셋에 대해서는 최인접 시점 선택 알고리즘을 사용했을 경우가 성능 개선에 효과적임을 확인하였다. 뿐만 아니라 그림 1과 비교하여도, 해당 참조 시점 선택 기법을 적용한 경우에 성능 개선이 월등히 이루어졌음을 확인하였다. 한편, 카메라 포즈 거리 기반 참조 시점 선택 기법의 경우 ScanNet [3]과 같은 동영상 기반 다 시점 데이터 셋에서는 성능이 우수한 것으로 보아 목표 도메인(target domain) 특성에 따른 참조 시점 선택 알고리즘이 적절히 적용된다면 추정 성능을 개선할 수 있지만, 오히려 부적절한 참조 시점 선택 기법을 적용할 경우, 성능 개선에 큰 효과가 발생하지 않음을 확인하였다.

4. 한계

본 연구의 실험에서 CasMVS [9]와 DVMVS [16]의 각 목표 도메인(target domain) 별 깊이 추정 결과를 기반으로 최적 참조 시점 개수와 참조 시점 선택 알고리즘을 중심으로 제안 기법의 필요성을 검증하였다. 그러나 large baseline 특성을 포함한 발산형 데이터에 대한 깊이 추정을 수행할 경우, 단순히 방법론에서 제시한 기법으로 개선을 도모하는 것으로는 깊이 추정에 한계가 존재한다. 그 이유는 각 픽셀에 대하여 겹쳐지는(overlapped) 정보의 차이가 증가함으로 인해 발생하는 깊이 추정 이미지에서의 결함(artifact)을 추가적으로 개선할 필요가 발생하기 때문이다. 따라서, 향후 이와 같은 사항을 고려한 매칭 비용 볼륨 생성 기법에 관한 연구가 수행될 필요가 있다. 그리고 이러한 한계점을 적절히 극복한다면, 더욱 다양한 환경에서 취득된 다 시점 영상 콘텐츠에서도 강건한 깊이 추정이 가능할 것으로 기대한다.

5. 결론

본 연구는 데이터 셋 촬영 특성에 따른 다 시점 깊이 추정 방법론에 대해 다룬다. 일반적으로, 다 시점 깊이 추정 테스크에서는 수렴형 배치일 때 추정 성능이 우수하였지만, 발산형일 때는 그렇지 않음을 확인하였다. 또한, 배치된 카메라 간 기준 거리인 baseline에 따라 추정 결과 차이가 존재했다. 따라서 이러한 다 시점 데이터 셋의 특성들을 고려하여 깊이 추정에 적용할 수 있는 두 가지 기법을 제안하였다. 하나는 최적 참조 시점 개수 설정이고, 다른 하나는 baseline에 따른 적절한 참조 시점 선택 알고리즘의 적용이다. 제안 기법에 기반하여 실험을 진행한 결과, 깊이 추정 결과가 월등히 개선되었다. 이에 따라, 향후 딥러닝 기반 다 시점 깊이 추정 네트워크의 연구 과정에 있어 본 연구의 결과가 깊이 추정 방법에 대한 가이드라인으로 활용될 수 있을 것임을 확인하였다.

6. 참고문헌

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. IJCV, 2016, 120(2):153-168, 2016.

[2] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. TOG, 36(4):78, 2017.

[3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In IEEE *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2423-2443, 2017.

[4] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In ECCV, 2018, pages 767-783, 2018.

[5] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR, 2018*, pages 5410-5417, 2018.

[6] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In CVPR, 2019, pages 3273-3282, 2019.

[7] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV, 2017*, pages 66-75, 2017.

[8] Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P.: End-to-end learning of geometry and context for deep stereo regression. *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, ZuozhuoDai, Feitong Tan, Ping Tan. Cascade Cost Volume for High-Resolution Multi-View Stereo and Steroe Matching. In CVPR 2020, arXiv:1912.06378, 2019.

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR, 2017*, pages 2117-2125, 2017.

[11] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[12] Yiran Zhong, Hongdong Li, and Yuchao Dai. Open-world stereo video matching with deep rnn. In *European Conference on Computer Vision (ECCV)*, September 2018.

[13] Vaishakh Patil, Wouter Van Gansbeke, Dengxin Dai, and Luc Van Gool. Don't Forget The Past: Recurrent Depth Estimation from Monocular Video. In I*EEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* 2020.

[14] Denis Tanæv, Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. Temporally Consistent Depth Estimation in Videos with Recurrent Architectures. In *European Conference on Computer Vision (ECCV)*, 2019.

[15] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1725-1734, 2019.

[16] Arda Düzçeker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, Marg Pollefeys. DeepVideoMVS: Multi-View Stereo on Video with Reccurent Spatio-Temporal Fusion. In CVPR 2021, pages 15324-15333, 2021.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Coputer-Assisted Intervention (MICCAI)*, 2015.

[18] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Waikin Wong, and Wang chun Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *In International Conference on Neural Information Processing Systems*, page 802-810, 2015.

[19] Andrea Palazzi. ConvLSTM_pytorch. https://github.com/ndrplz/convLSTM_pytorch, 2020.

[20] Yuxin Hou, Juho Kannala, and Arno Solin. Multi-View Stereo by temporal Nonparameteric Fusion. *International Conference on Computer Vision (ICCV)*, 2019.