하이브리드 트랜스포머 기반 비지도 단안 카메라 영상 깊이 추정 방법

황승준, 박성준, 김규민, 백중환
한국항공대학교
s-eungju-n@hanmail.net, jdwns1011@naver.com, gyumin93@gmail.com, jhbaek@kau.ac.kr

Unsupervised Monocular Depth Estimation Using Hybrid Transformer

Seung-Jun Hwang, Sung-Jun Park, Gyu-Min Kim, Joong-Hwan Baek
Korea Aerospace Univ.

요 약

본 논문은 비지도 학습 기반 단안 카메라 영상 깊이 추정에 CNN과 비전 트랜스포머 네트워크를 혼합한 하이브리드 네트워크를 적용하는 방법에 대해 기술한다. 최근의 연구는 CNN은 컨볼루션 연산을 통해 지역적 특징을 추출에 적합하며, 비전 트랜스포머는 다중 자기주의 모듈에 기반한 광역적 특징추출에 적합하다고 알려져 있다. 본 연구에서는 일반적인 비지도 깊이 추정 네트워크의 인코더-디코더 구조에서 지역적 특징을 추출하는 단계에서는 CNN을 사용하고, 광역적 특징을 추출하는 단계에서는 비전 트랜스포머를 사용하는 하이브리드 모델 기반 깊이 추정 네트워크를 제안한다. 제안하는 네트워크는 실험을 통해 기존의 네트워크보다 적은 파라미터와 연산량으로 기존 네트워크와 유사한 성능을 보임을 확인하였다.

Ⅰ. 서 론

영상에서 3D 깊이 정보를 예측하는 기술은 자율주행, AR, VR 및 로봇 비전 등 3D 공간 지도 생성이 요구되는 분야에 활발히 적용되고 있다. 기존의 센서 방식의 깊이 추정 방식은 정확도는 높지만, 높은 가격과 낮은 해상도의 문제점이 존재한다. 최근의 카메라 기반의 깊이 추정 연구는 상당히 활발히 진행되었으며, 성능 또한 드론과 같은 상용 제품에 적용될 정도로 높은 수준을 보인다.

카메라 기반의 깊이 추정 연구는 센서로부터 취득된 데이터를 직접 학습하는 지도 학습 모델과 인접한 프레임 두 프레임의 이미지에서 다른 한쪽 이미지를 재구성하는 시점 합성을 통해 깊이를 추정하는 비지도 학습 모델로 나누어진다[1]. 지도 학습 모델이 데이터를 직접 학습하기 때문에 더 높은 성능을 보이고 있으나, 센서로부터 3D 데이터를 획득할 수 없는 경우에는 지도 학습 모델 적용의 어려움이 있다. 최근의 연구는 비지도 학습 모델 방식도 지도학습 방식과 유사한 수준의 성능을 보인다.

또한, 최근의 비전 트랜스포머 모델은 이미지 분류, 분할, 검출 분야에서 최고성능을 달성하였다[2, 3]. 비전 트랜스포머는 패치단위의 임베딩 벡터를 쿼리, 키, 값으로 나누어 연산하는 다중 자기주의 메커니즘과 MLP(Multi Layer Perceptron)로 구성되어 있다. 이를 지도학습에 적용한 연구에서는 CNN 기반의 Resnet과 비전 트랜스포머 모델을 혼합한 하이브리드 모델이 가장 좋은 성능을 보임을 실험으로 증명하였다[4].

본 논문에서는 비지도 깊이 추정 모델에서 효율적인 깊이 네트워크 추정을 위한 하이브리드 트랜스포머 기반 깊이 추정 네트워크를 제안한다. 기존 깊이 추정 네트워크의 인코더-디코더 구조에서 지역적 특징 추출을 위한 CNN기반의 Resnet 모델과 광역적 특징 추출을 위한 트랜스포머의 혼합 모델을 제안한다.

Ⅱ. 본 론

본 논문에서는 최근의 연구에 따라 비지도 학습 기반 깊이 추정을 위해 깊이 및 포즈 네트워크를 동시에 학습한다[5]. 네트워크는 타겟 이미지 I_t와 소스 이미지 I_s로부터 타겟 이미지 시점으로 재구성된 이미지 $\hat{I}_{s\ t}$사이의 재구성 에러 L_p를 최소화하는 뷰합성 과정을 거치면서 학습이 이루어진다. 재구성 이미지는 예측된 타겟 깊이와 포즈를 투영하고 소스이미지로부터 샘플링하여 얻을 수 있다. 이때 예측된 깊이와 포즈는 각각의 네트워크에 의해 추정되어진다. 재구성 이미지 생성을 위한 뷰합성 과정에 대한 식 (1)을 아래에 보인다.

$$\hat{I}_{s\to t} = I_s \ \ proj(D_t, T_{t\to s}, K)\rangle \quad (1)$$

기존의 비지도 모델의 손실 함수는 닮음 유사도 정도인 SSIM(Structural Similarity Index Measure)과 $L1$ 거리를 조합한 이미지 재구성 손실과, 깊이가 이미지 경계 불연속에 따라 달라지도록 제약하는 깊이 유연 손실을 사용한다. 본 연구에서도 기존의 연구를 따라 위의 손실함수를 사용한다.

최근의 연구는 이미지의 뷰합성이 아닌 특징 표현의 뷰합성을 수행하였다. 또한, 재구성 소스 특징과 타겟 특징간의 $L1$거리를 다채널로 누적하는 비용 용량 메커니즘을 적용하여 성능을 향상 시켰다[6]. 본 연구에서는 Resnet을 이용한 기존 모델의 인코더-디코더 구조에서 광역적 특징 추출자에 해당하는 깊은 레이어에 트랜스포머를 적용한다.

CNN기반의 모델은 컨볼루션 연산을 사용하기 때문에 인접 영역에 대한 특징 표현 특성이 있으며, 트랜스포머는 자기주의 메커니즘으로 전체 영역을 표현하는 특성을 갖는다. 따라서, 모든 레이어를 동일한 구조의 연속으로 설계하는 방법이 아닌, 레이어에 따라 다른 모델 블록을 적용하도록 지역적 블록과 광역적 블록을 제안한다.

지역적 블록은 기존 Resnet의 처음 두 레이어를 그대로 사용한다. 광역

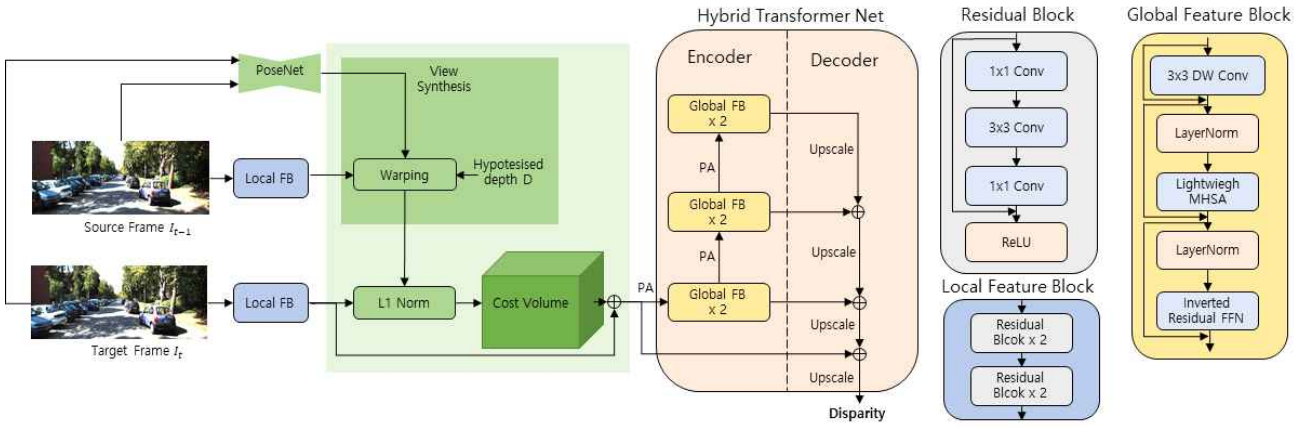

그림 1. 하이브리드 트랜스포머 모델 기반 단안 카메라 깊이 추정 모델

적 블록은 키와 값을 절반으로 줄인 가벼운 다중 자기주의 모듈과 역잔차 피드포워드 네트워크로 구성한다[7]. 광역적 블록은 각 레이어에 2개의 연속된 층으로 구성하였으며, 제안하는 전체 구조도를 그림 1 에 보인다.

Ⅲ. 실험 결과

본 논문에서는 제안하는 모델의 성능 평가를 위해 KITTI 데이터셋 기반 실험평가를 진행한다. 트랜스포머 모델은 다중 자기주의 연산에 따라 컨벌루션보다 더 적은 파라미터와 연산 수로 광역적 특징추출이 가능하다. 기존 Resnet18 기반 인코더-디코더 모델과 하이브리드 인코더-디코더 모델의 파라미터 차이를 표 1에 보인다.

표 1. 인코더-디코더 파라미터

	Params(M)	MACs(G)
Resnet18 [6]	14.277 M	4.461
Hybrid	**11.097 M**	**3.406**

기존 Resnet18 기반과 제안한 하이브리드기반 모델의 정량적 성능 평가를 표 2에 보인다. 평가 지표에서 AbsRel, SqRel, RMSE는 에러 평가지표로 낮을수록 좋으며, $\delta < 1.25$는 임계치 정확도 평가지표로 높을수록 좋은 지표이다. 본 논문에서 제안하는 하이브리드 모델이 기존 모델에 비해 보다 적은 파라미터와 연산량으로 유사한 성능을 나타냄을 알 수 있다.

표 2. 정량적 평가표

	AbsRel	SqRel	RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Resnet18 [6]	**0.098**	0.770	**4.459**	**0.900**	**0.965**	**0.983**
Hybrid	0.099	**0.757**	4.478	0.898	0.964	**0.983**

정성적 평가를 위한 결과 이미지를 그림 2에 보인다. 기존 모델보다 적은 파라미터와 연산량을 필요로 하지만 유사한 깊이 추정 결과를 보임을 확인한다.

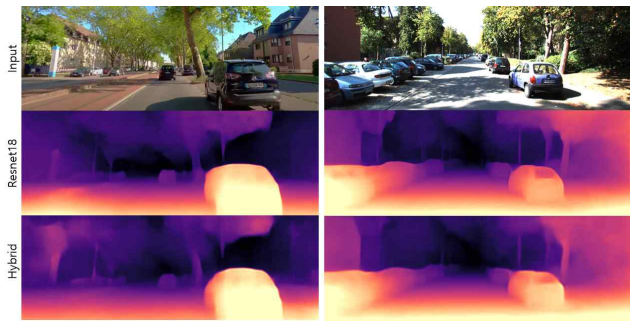

그림 2. 기존 모델과 제안하는 하이브리드 모델의 깊이 추정 결과

Ⅳ. 결 론

본 논문에서는 비지도 학습기반 단안 카메라 영상 깊이 추정에 CNN기반 Resnet18 모델과 비전 트랜스포머 모델을 혼합한 하이브리드 모델을 제안하였다. 제안하는 하이브리드 모델은 기존 모델에 비해 보다 적은 파라미터와 연산량을 필요로 하지만, 기존 모델과 유사한 성능을 확인하였다. 추후 연구로 기존 자세 추정 모델 경량화를 진행할 계획이다. 깊이 추정 네트워크의 자세 추정 모델 또한 Renet18기반의 인코더-디코더 모델을 사용한다. 자세 추정 모델은 두 개의 입력이미지에서 자세를 추정하기 때문에, 두 이미지간의 상관관계를 분석하는 상호 작용 트랜스포머 모델을 사용하여 보다 효율적인 네트워크를 구성할 수 있을 것으로 예상된다.

ACKNOWLEDGMENT

본 연구는 경기도 지역협력 연구센터 사업의 일환으로 수행하였음. [GRRC항공2017-B04, 지능형 인터랙티브 미디어 및 공간 융합 응용 서비스 개발].

참 고 문 헌

[1] Ming, Yue, et al. "Deep Learning for Monocular Depth Estimation: A Review." Neurocomputing (2021).

[2] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." arXiv preprint arXiv:2103.14030 (2021).

[3] Dong, Xiaoyi, et al. "CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows." arXiv preprint arXiv:2107.00652 (2021).

[4] Ranftl, René, Alexey Bochkovskiy, and Vladlen Koltun. "Vision transformers for dense prediction." arXiv preprint arXiv:2103.13413 (2021).

[5] Guizilini, Vitor, et al. "3d packing for self-supervised monocular depth estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[6] Watson, Jamie, et al. "The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[7] Guo, Jianyuan, et al. "CMT: Convolutional Neural Networks Meet Vision Transformers." arXiv preprint arXiv:2107.06263 (2021).