향상된 깊이 추정을 위한 Hierarchical Vision Transformer 기반 모델

김동준, 손경아
아주대학교

dk85440@ajou.ac.kr, kasohn@ajou.ac.kr

Hierarchical Vision Transformer for Enhanced Depth Estimation

Kim Dong Jun, Sohn Kyung Ah
Ajou Univ.

요 약

깊이 추정은 이미지나 영상에서 장면의 구조를 정확하게 파악해야 하는 어려운 문제로, 다양한 응용 분야에서 높은 정확성이 요구된다. 본 논문은 깊이 추정의 정확성을 향상시키기 위해 Hierarchical Vision Transformer (Hiera ViT) 기반의 모델을 제안한다. Hiera ViT 를 도입하여 피처를 계층적으로 추출하고, 깊이 맵의 정밀도와 세밀함을 향상시킨다. 실험 결과, 제안된 모델이 기존 ViT 기반 모델에 비해 RMSE 및 Abs Rel 지표에서 유의미한 성능 향상을 보였다.

Ⅰ. 서 론

깊이 추정은 컴퓨터 비전의 핵심 과제로, 자율주행, 로봇 공학, 증강 현실(AR) 등 다양한 정보통신 응용 분야에서 중요한 역할을 한다 [1, 2]. RGB 이미지만을 기반으로 정확한 깊이 맵을 생성하는 것은 많은 연구자들이 집중해온 과제이며, 최근 딥러닝 기술의 발전으로 성능이 크게 향상되었다. 특히, Vision Transformer(ViT)[3]는 자기 주의(Self-Attention) 메커니즘을 통해 이미지의 글로벌 컨텍스트를 효과적으로 학습할 수 있어 깊이 추정 분야에서 주목받고 있다. 그러나 기존 ViT 는 계층적 피처 추출이 부족하여, 세부적인 공간 정보의 보존에 한계가 있었다. 본 논문에서는 Hierarchical Vision Transformer (Hiera ViT)[4]를 도입하여 다양한 해상도의 피처를 계층적으로 추출하고, 효과적으로 포착하여 깊이 추정의 정확성과 세밀함을 향상시키고자 한다.

Ⅱ. 본론

1. 모델 구조

본 논문에서 제안하는 모델은 Hiera ViT 를 인코더로 사용하여, 다양한 해상도의 피처를 계층적으로 추출하고 이를 효과적으로 결합해 정확한 깊이 맵을 생성한다. 일반적인 ViT 가 전역적인 패치 단위 처리를 통해 이미지 전체의 맥락을 학습한다면, HieraViT 는 계층적 피처 추출 메커니즘을 통해 고해상도와 저해상도 정보를 동시에 포착한다. 이러한 계층적 피처는 깊이 추정에 필요한 공간적·구조적 단서를 더욱 세밀하게 확보할 수 있으며, 결과적으로 작은 물체나 경계 영역의 깊이를 보다 정확하게 추정하게 된다. 이후 계층적으로 추출된 피처 맵을 Multi-Scale Decoder 로 전달하여, 각 스케일에서 중요도가 높은 정보를 효율적으로 통합한다. 이를 통해, 전역적인 컨텍스트와 세밀한 정보가 반영된 깊이 맵이 생성된다. 그러나 해당 깊이 맵 만으로는 복잡한 경계나 작은 물체의 깊이를 완벽히 표현하기 어렵기 때문에 RefineNet[5]을 추가로 도입하여 이러한 맵을 정제한다. 전역적인 깊이 분포와 세부적인 경계·물체 윤곽 정보를 조화롭게 결합함으로써, 경계의 모호성이 발생하거나 세부 정보가 소실되는 문제를 완화한다. 결과적으로, 제안 모델은 전역적 정확도와 세밀함을 동시에 만족시키는 깊이 맵을 산출한다.

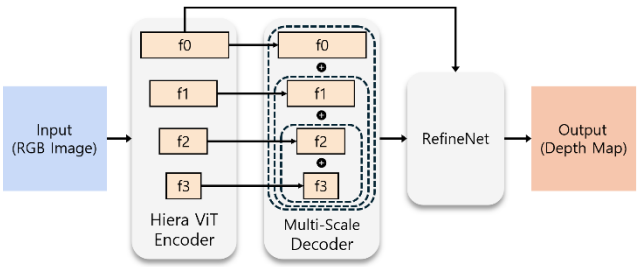

그림 1: 모델 구조

본 연구에서 사용된 손실 함수는 크게 두 범주를 결합하여 사용한다. 먼저, 전역적인 정확도를 보장하기 위해 Scale-Invariant Loss 와 L1 Loss 를 혼합한 형태를 사용하였다. 이는 거리의 절대값보다는 비례 관계를 중요시하는 깊이 추정의 특성을 반영하여, 원근감이 크게 변하는 장면에서도 모델이 안정적으로 학습할 수 있도록 돕는다. 또한, 경계 및 소형 물체의 정밀도를 높이기 위해 Gradient Loss 와 SIMM(Structural Similarity) Loss 를 추가적으로 적용함으로써 경계 영역과 물체 형태가 더욱 정밀하게 표현되도록 유도한다. Gradient Loss 는 깊이 맵의 기울기를 목표와 일치시키면서

경계부의 표현을 개선하고, SIMM Loss 는 인접 픽셀 간의 구조적인 유사도를 향상시켜 깊이 추정 과정에서 발생 가능한 블러 현상이나 세부 정보 손실을 보완한다.

$$L_{total} = \alpha(L_{SI} + L_{L1}) + \beta L_{\nabla} + \gamma L_{SSIM}$$

여기서, L_{SI} 는 Scale-Invariant Loss, L_{L1} 는 L1 Loss, L_{∇} 는 Gradient Loss, L_{SSIM} 는 SIMM Loss 를 의미하며, α, β, γ 는 각각의 손실 함수에 대한 가중치를 조절하는 파라미터이다.

$$L_{SI} = \frac{1}{N}\sum_{i=1}^{N}(\log_{10}(\hat{y}_i) - \log_{10}(y_i))^2$$

$$L_{L1} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

$$L_{\nabla} = \frac{1}{N}\sum_{i=1}^{N}(|\nabla_x\hat{y}_i - \nabla_x y_i| + |\nabla_y\hat{y}_i - \nabla_y y_i|)$$

$$L_{SSIM} = 1 - SSIM(\hat{y}, y)$$

2. 실험 방법

본 논문에서는 실내 환경에서 촬영된 다양한 장면을 포함하는 NYU Depth V2[6]데이터셋으로 모델을 학습하고 평가하였다. 해당 데이터셋은 가구, 소형 물체, 복잡한 배경 등을 담고 있어, 모델이 실내 환경에서 발생할 수 있는 복합적인 물체 배치를 학습하기에 적합하다. 또한, 다양한 거리와 크기를 지닌 물건들을 처리하는 능력을 향상시킬 수 있으며 깊이 추정 모델의 일반화 성능을 평가하는 데에도 유용하다.

성능 평가는 깊이 추정 분야에서 주로 쓰이는 RMSE(Root Mean Squared Error), RMSE log 그리고 Abs Rel(Absolute Relative Error) 지표를 사용한다. RMSE 는 예측 깊이와 실제 간 오차의 전역적 크기를 한눈에 파악하기 유용하다. log10 은 깊이 값의 범위가 매우 넓거나, 특정구간에서 급격한 변동이 있는 장면에서도 예측 품질을 균형 있게 평가할 수 있도록 돕는다. 그리고 Abs Rel 은 실제 깊이에 대한 상대적 오차를 측정함으로써, 모델의 일반화 성능을 보여준다

3. 실험 결과

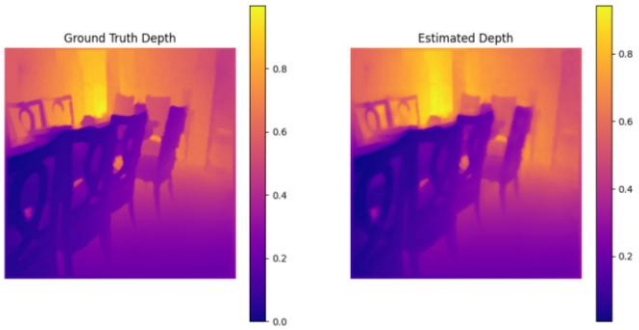

그림 2: 깊이 맵 추정 결과

Models	Abs_Rel ↓	RMSE↓	log10↓
DPT[1]	0.110	0.357	0.045
DepthFormer[2]	0.094	0.331	0.044
Ours	0.051	0.268	0.038

표 1: 깊이 추정 실험 결과

실험 결과, Hiera ViT 기반 모델이 기존의 ViT 기반 모델 대비 모든 지표에서 개선된 성능을 보임을 확인할 수 있다. 이러한 성능 향상은 Hiera ViT 의 효과적인 계층적 피처 추출과 RefineNet 을 통한 효율적인 통합 덕분으로 해석된다. 이는 본 논문에서 제안한 접근 방식이 깊이 추정 모델의 표현력과 세밀함 복원 능력을 향상시키는데 효과적임을 뒷받침한다.

Ⅲ. 결론

본 연구에서는 Hierarchical Vision Transformer (Hiera ViT)를 도입하여 깊이 추정 모델의 정확성과 세밀함을 향상시켰다. Hiera ViT 는 계층적 피처 추출을 통해 다양한 해상도의 정보를 효과적으로 학습할 수 있었으며, Multi-Scale Decoder 와 RefineNet 의 결합을 통해 깊이 맵의 정밀도를 크게 개선하였다.

실험 결과, 제안된 모델은 기존 ViT 기반 모델에 비해 우수한 성능을 보였으며, 특히 세부적인 공간 정보의 보존 측면에서 큰 장점을 나타냈다. 향후 연구에서는 RefineNet 의 구조를 더욱 강화하고, 다양한 데이터셋에서 학습하여 모델의 성능을 더욱 향상시킬 계획이다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업(IITP-2024-RS-2023-00255968) 및 대학 ICT 연구센터사업(IITP-2024-2021-0-02051)의 연구 결과로 수행되었음.

참 고 문 헌

[1] Rene Ranftl, Alexey Bochkovskiy, Vladlen Koltun. (2021). Vision Transformers for Dense Prediction. International Conference on Computer Vision (ICCV).

[2] Z Li, Z Chen, X Liu, J Jiang. (2023). Depthformer: Exploiting Long-Range Correlation and Local Information for Accurate Monocular Depth Estimation. Machine Intelligence Research, Springerkra

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. (ICLR)

[4] Hassani, A., Walton, S., Shah, N., Shi, H., Suh, Y., & Shi, H. (2023). Hiera: A Hierarchical Vision Transformer Without the Bells-and-Whistles. International Conference on Computer Vision (ICCV).

[5] Guosheng Lin, Anton Milan, Chunhua Shen, Ian Reid. (2017). RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. (CVPR).

[6] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. (2012). Indoor segmentation and support inference from rgbd images. (ECCV).