Attention Model 을 이용한 단안 영상 기반 깊이 추정 네트워크

*정근호 **윤상민
국민대학교 컴퓨터공학과 HCI 연구실
*ehwk9200@kookmin.ac.kr, **smyoon@kookmin.ac.kr

Single Image-based Depth Estimation Network using Attention Model

*Geunho, Jung **Sang Min, Yoon
HCI Lab., Computer Science, Kookmin University

요 약

단안 영상에서의 깊이 추정은 주어진 시점에서 촬영된 2 차원 영상으로부터 객체까지의 3 차원 거리 정보를 추정하는 것이다. 최근 딥러닝 기반으로 단안 RGB 영상에서 깊이 정보 추정에 유용한 특징 맵을 추출하고 이를 이용해서 깊이를 추정하는 모델들이 기존 방법들의 성능을 넘어서면서 관련된 연구가 활발히 진행되고 있다. 또한 Attention Model 과 같이 특정 특징 맵의 채널 혹은 공간을 강조하여 전체적인 네트워크의 성능을 개선하는 연구가 소개되었다. 본 논문에서는 깊이 정보 추정을 위해 사용되는 특징 맵을 강조하기 위해서 Attention Model 을 추가한 AutoEncoder 기반의 깊이 추정 네트워크를 제안하고 적용 부분에 따른 네트워크의 깊이 정보 추정 성능을 평가 및 분석한다.

1. 서론

컴퓨터 비전 및 영상처리 분야에서 영상의 깊이 추정은 주어진 시점에서 영상에 위치한 주요 객체까지의 거리 정보를 추정하는 것이다. 깊이 추정은 2 차원 영상에서 3 차원 정보인 깊이 정보를 추정하는 영상 이해부터 3 차원 복원 등 다양한 연구부터 자율주행, 네비게이션과 증강 현실 등 여러가지 산업 분야에도 응용 가능한 주요 연구 분야이다. 영상의 깊이 정보는 센서를 이용해 획득하거나, 두 대 이상의 카메라를 이용해 촬영한 영상들의 시차를 이용해서 획득할 수 있다. 하지만 이러한 정보 없이 단안 영상으로부터 깊이 정보를 획득하는 것은 어려운 문제이다. 컴퓨터 비전 분야에 딥러닝이 접목된 연구가 활발해지면서 NYU V2[1]나 KITTI[2] 등 깊이 정보를 포함하는 데이터셋을 기반으로 단안 영상에서 깊이를 추정하는 연구가 이루어졌다.

DenseDepth[3]는 ImageNet[4]으로 사전 학습된 Enco-der를 이용해서 입력 영상으로부터 특징을 추출하고 Decoder 를 이용해서 깊이 영상을 복원하는 AutoEncoder 기반의 깊이 복원네트워크를 제안했으며 기존의 방법에 비해 좋은 성능을 보여주었다. BTS[5]는 DenseDepth[4]와 달리 Multiscale Local Guidance Layer 를 Decoder 로 사용해서 보다 정확한 깊이 정보를 복원하고 더 좋은 성능을 보여주었다.

SENet[6]은 특징 맵의 채널을 강조해서 네트워크의 성능을 향시키는 Channel Attention 을 제안하였고 SCA-CNN [7]은 Spatial Attention 모델을 제안해서 특징 맵의 공간 정보를 강조했다. 본 논문에서는 기존 AutoEncoder 기반의 네트워크에 Attention Model 을 추가하여 네트워크의 특징 맵을 강조해서 입력 영상으로부터 정확한 깊이 정보를 복원하는 네트워크를 제안하고 성능을 비교 및 분석한다.

2. 깊이 영상 추정 네트워크

본 논문에서는 DenseDepth[3]를 기반으로 깊이 정보를 추정하는 네트워크 설계했으며 제안하는 Attention 을 추가한 네트워크는 다음 그림 1, 그림 2 와 같다.

2.1 제안하는 네트워크

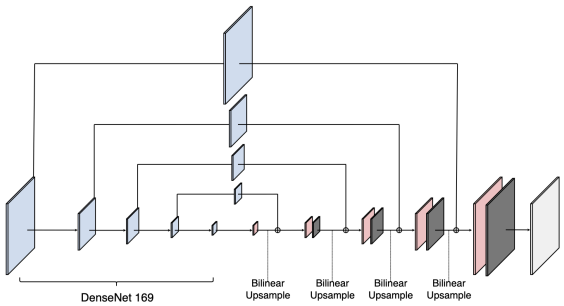

그림 1. Decoder 에 Attention Model 을 추가한 네트워크

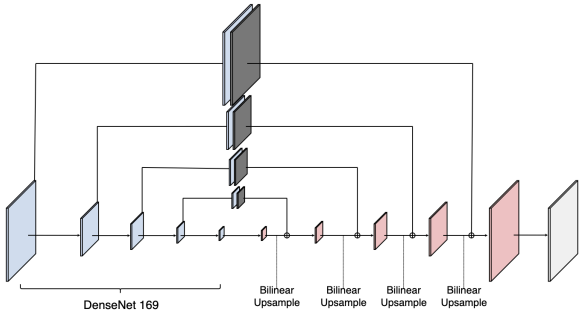

그림 2. Skip-connection 에 Attention Model 을 추가한 네트워크

두개의 네트워크는 ImageNet[4]으로 사전 학습된 DenseNet-169[8]를 Encoder 로 사용하며 입력 RGB 영상으로부터 특징을 추출함과 동시에 Skip Connection 을 통해 특징 맵을 Decoder 로 전달한다. 추출된 특징 맵은 Decoder 에서 Bilinaer 보간법과 여러 개의 Convolution Layer 를 거쳐 최종적으로 입력 영상의 1/2 크기를 갖는 깊이 정보 영상을 복원한다.

본 논문에서는 Encoder 에서 Decoder 로 전달되는 특징 맵을 강조하기 위해서 Skip Connection 에 Attention 을 적용하는 네트워크와 Decoder 의 복원 성능을 높이기 위해 Decoder 에 Attention 을 적용한 네트워크를 제안한다. 또한 Attention 을 적용하지 않은 네트워크를 학습하여 두개의 네트워크의 성능과 비교한다.

2.2 손실 함수

본 논문에서는 DenseDepth[3]에서 사용한 손실 함수를 참고해서 제안하는 네트워크의 손실 함수로 사용하였으며 다음 수식 1 과 같다.

$$L(y,\hat{y}) = \lambda L_{depth}(y,\hat{y}) + L_{grad}(y,\hat{y}) + L_{SSIM}(y,\hat{y}) \quad (1)$$

$L(y,\hat{y})$ 은 네트워크의 전체 손실 함수로 정확한 깊이 정보를 복원하기 위해서 세개의 손실 함수를 사용한다.

$$L_{depth}(y,\hat{y}) = \frac{1}{n}\sum_p^n |y_p - \hat{y}| \quad (2)$$

수식 2 는 Depth 정보에 대한 손실 함수로써 복원된 깊이 정보와 Ground-Truth 깊이 정보의 차이를 L1-Norm 으로 계산한다.

$$L_{grad}(y,\hat{y}) = \frac{1}{n}\sum_p^n |\mathbf{g_x}(y_p,\hat{y}_p)| + |\mathbf{g_y}(y_p,\hat{y}_p)| \quad (3)$$

수식 3 은 $\mathbf{g_x}$와 $\mathbf{g_y}$는 복원된 깊이 영상 $\hat{y}$와 원본 깊이 영상 y의 기울기를 x, y 축에 대해 계산하고 각각 픽셀 사이의 차를 구한 것이다.

$$L_{SSIM}(y,\hat{y}) = \frac{1-SSIM(y,\hat{y})}{2} \quad (4)$$

수식 4 의 $L_{SSIM}(y,\hat{y})$ 은 영상 복원 분야에서 널리 사용되는 Structural Similarity (SSIM)을 이용해 손실 함수로 사용한 것이다. 파라미터 λ는 L_{depth}에 가중치로 DenseDepth[3]를 참고하여 0.1 로 설정하였다.

2.3 Attention 모델

본 논문에서는 기존 AutoEncoder 에 Attention 모델을 적용한 네트워크를 기반으로 특징 맵을 강조는 네트워크를 설계하였다. Attention 모델은 SENet[6]에서 제안한 채널 Attention 모델과 SCA-CNN[8]이 제안한 공간 Attention 모델을 이용하였다. 다양한 차원에서 특징을 강조하기 위해 기존에 제안된 Channel 뿐만 아니라 Width, Height 방향으로도 Channel Attention Model 을 적용했으며, Spatial Attention Model 또한 (W,H), (C, H), (C, W) 세 방향으로 적용하였으며 강조된 3 방향의 특징 맵들을 A. G. Roy [9] 등이 제안한 방법을 참고해서 설계하였다. 그림 3 과 4 는 Channel Attention Model 과 Spatial Attention Model 의 구조를 나타낸 그림이다. Attention 모델들을 다양한 방향 (그림 5)에서 적용하기 위해 출력된 특징 맵을 전치시킨 뒤 여러 방향으로 채널 Attention Model 을 적용하였다.

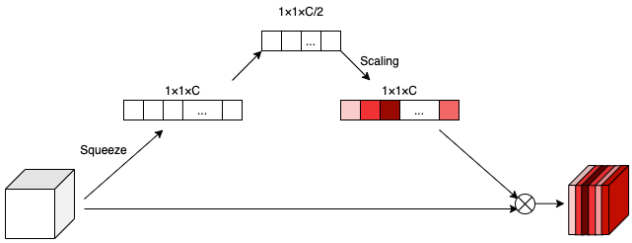

그림 3. Channel Attention Model

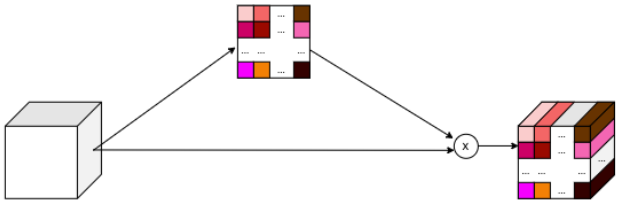

그림 4. Spatial Attention Model

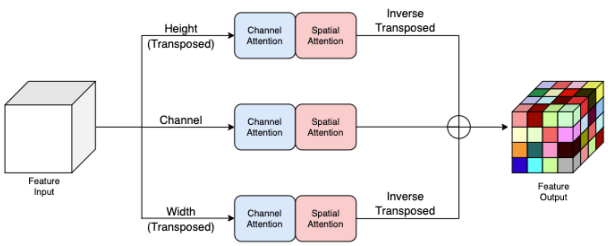

그림 5. 다양한 방향에 적용한 Channel & Spatial Attention Model

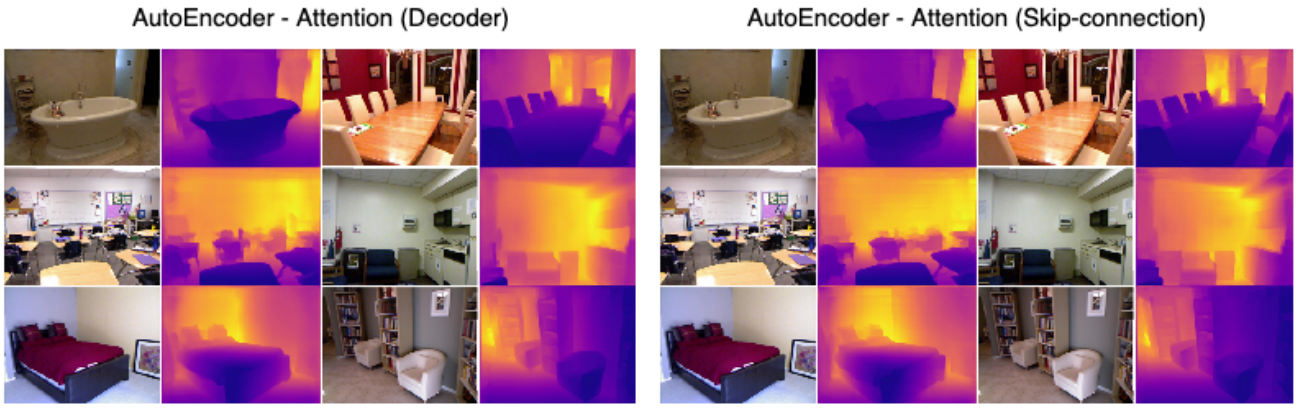

그림 6. 제안하는 네트워크로 깊이 영상을 추정한 결과

3. 실험

3.1 학습 데이터 및 실험 환경

본 논문에서는 NYU V2 RGB-D 데이터셋[1]을 실험 데이터로 사용하였다. NYU 데이터셋[1]은 실내를 촬영한 영상 데이터셋으로 640×480 해상도의 RGB 영상과 320× 240 크기의 깊이 영상 쌍으로 구성되어있다. 학습 데이터는 총 120K 개의 영상으로 구성되어 있으며 50K 개 영상을 학습데이터로 사용하였고 654 개의 테스트 데이터로 사용하였다. 제안하는 네트워크는 Tensorflow 를 기반으로 구현하였으며 NVIDIA TitanXP 로 학습하였다.

3.2 실험 결과 및 분석

본 논문에서는 다양한 방법으로 Attention 모델을 추가한 네트워크들의 성능을 비교하였으며, 평가 방법으로는 깊이 추정 연구 분야에서 널리 사용되는 RMS (Root Mean Square)를 비롯한 여러 평가 방법을 이용하여 성능을 평가 및 비교하였다.

각 실험에 대한 결과는 다음 표 1 과 같다. 제안하는 네트워크가 다른 방법으로 Attention 모델을 적용한 네트워크보다 성능이 좋은 것을 알 수 있으며, 각 방향에 대해 Channel Attention 과 Spatial Attention 을 직렬로 연결하여 특징을 강조하고 세 방향의 강조된 특징들을 더하는 방법이 깊이 정보 복원에 도움이 된다는 것으로 볼 수 있다.

그림 6 은 제안하는 네트워크를 이용해 깊이 영상을 추정한 결과이다. Ours (Middle)은 Attention Model 을 Encoder 와 Decoder 사이 중간에 연결한 네트워크이고, Ours (Branch)는 특징을 추출하는 Encoder 와 특징을 강조하는 Attention Encoder 로 분할하여 나오는 결과를 더해 강조하는 네트워크이다. 전체적으로 비슷한 결과를 보였지만 종합적으로 Decoder 에 Attention 모델을 추가한 Ours (Decoder)의 성능이 가장 좋았다.

표 1. 깊이 영상 추정 실험 결과

	a1	a2	a3	REL	RMS	$\log_{10}$
Ours (Skip)	0.833	0.970	0.992	0.132	0.573	0.056
Ours (Decoder)	**0.842**	**0.971**	**0.993**	**0.129**	0.564	**0.055**
Ours (Middle)	0.839	0.969	0.992	**0.129**	**0.559**	**0.055**
Ours (Branch)	0.833	0.967	0.992	0.132	0.574	0.056

4. 결론 및 향후 연구

본 논문에서는 H, W, C 세 방향으로 Channel Attention Model 과 Spatial Attention Model 을 적용해서 특징 맵을 보다 다양한 관점에서 강조하는 깊이 영상 추정 네트워크를 제안하였으며, 그 결과 단일 방향보다 다양한 방향으로 Attention 을 적용한 네트워크가 경우에 따라 좋은 성능을 보이는 것을 알 수 있었다. 향후 다양한 Attention Model 과 적용 방안을 연구할 계획이다.

감사의 글

이 논문은 2016 년도 정부(교육부)의 재원으로 연구재단 기본 연구 지원 사업의 지원으로 수행된 연구임. (NRF-2016R1D1A1B04932889)

참고문헌

[1] Silberman, Nathan, et al. "Indoor segmentation and support inference from rgbd images." *European conference on computer vision*. Springer, Berlin, Heidelberg, 2012.

[2] Geiger, Andreas, et al. "Vision meets robotics: The kitti dataset." *The International Journal of Robotics Research* 32.11 (2013): 1231-1237.

[3] Alhashim, Ibraheem, and Peter Wonka. "High quality monocular depth estimation via transfer learning." *arXiv preprint arXiv:1812.11941* (2018).

[4] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.

[5] Lee, Jin Han, et al. "From big to small: Multi-scale local planar guidance for monocular depth estimation." *arXiv preprint arXiv:1907.10326* (2019).

[6] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[7] Chen, Long, et al. "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[8] Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[9] Roy, Abhijit Guha, Nassir Navab, and Christian Wachinger. "Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2018.