특집논문 (Special Paper)
방송공학회논문지 제30권 제4호, 2025년 7월 (JBE Vol.30, No.4, July 2025)
https://doi.org/10.5909/JBE.2025.30.4.526
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

오토인코더를 활용한 잠재 확산 모델 기반 단안 깊이 추정

김 현 우[a)], 김 원 준[a)‡]

Monocular Depth Estimation via Autoencoder-guided Latent Diffusion Model

Hyunwoo Kim[a)] and Wonjun Kim[a)‡]

요 약

최근 다양한 도메인에서 대용량 데이터로 사전 학습된 확산 모델(Diffusion Model)을 활용하여 정밀하게 단안 깊이 추정을 수행하는 연구가 활발히 진행되고 있다. 그러나, 이러한 방식은 역방향 확산 과정에서 예측한 잡음(Noise)을 통해 깊이 지도를 간접적으로 생성하기 때문에 깊이 경계 예측에 있어 성능 저하가 발생할 수 있다. 본 논문에서는 정답 깊이 지도(Ground Truth)로 사전 학습한 오토인코더(Autoencoder)를 이용하여 잡음 생성 결과를 명시적으로 가이드 할 수 있는 방법을 제안한다. 기준 데이터셋(Benchmark Dataset)에서의 성능 평가를 통해 제안하는 방법이 학습하지 않은 도메인에서도 단안 깊이 추정 성능을 효과적으로 개선할 수 있음을 보였다.

Abstract

Recent studies leverage diffusion models pre-trained on large multi-domain datasets for accurate monocular depth estimation. However, the indirect nature of this approach, which involves reconstructing depth maps from predicted noise in the reverse diffusion process, could weaken depth-edge precision. In this paper, we propose a method to explicitly guide the predicted noise by employing an autoencoder pre-trained on ground-truth depth maps. Experimental results on the benchmark datasets demonstrate that the proposed method can effectively improve the monocular depth estimation performance in unseen domains.

Keyword : Monocular depth estimation, Latent diffusion model, Autoencoder, Unseen domain

a) 건국대학교 전기전자공학부(Department of Electrical and Electronics Engineering, Konkuk University)
‡ Corresponding Author : 김원준(Wonjun Kim)
E-mail: wonjkim@konkuk.ac.kr
Tel: +82-2-450-3396
ORCID: https://orcid.org/0000-0001-5121-5931
※ 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2023-NR076462)
※ This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-NR076462)
· Manuscript May 21, 2025; Revised July 1, 2025; Accepted July 1, 2025.

Ⅰ. 서 론

자율 주행, 증강 현실 등 다양한 산업 분야에서 실제 환경 이미지의 깊이 지도를 정밀하게 생성하는 기술에 대한 수요가 지속적으로 증가하고 있다. 그러나, 실제 환경 이미지는 학습 데이터와 다른 특성(예를 들어, 장면 구성, 카메라 내부 매개변수 등)을 보이는 경우가 많아 깊이 지도를 생성함에 있어 성능 저하가 발생할 수 있다. 학습되지 않은 도메인에서 깊이 지도 생성 방법의 일반화 성능을 개선하기 위해서 최근에는 다양한 도메인의 데이터로 사전 학습된 확산 모델(Diffusion Model)을 활용한 연구가 활발히 진행되고 있다. 깊이 추정을 위한 신경망은 이미지 내 장면의 구조나 시각적 형태 등을 학습하게 되는데, 확산 모델 또한 이와 유사하게 이미지 생성에 필요한 이미지 내 물체 간의 관계와 전반적인 맥락 등을 학습한다. 따라서 대용량 데이터로 학습된 확산 모델을 이용하면 깊이 추정의 일반화 성능 향상을 기대할 수 있다. 그러나 확산 모델은 입력 데이터의 확률 분포를 학습하여 고품질의 이미지를 생성할 수 있다는 장점이 있지만, 이는 다단계의 복잡한 추론 과정을 요구하여 긴 생성 시간과 높은 계산 비용을 초래하는 한계가 있다.

잠재 확산 모델(Latent Diffusion Model)[1]은 사전 학습된 오토인코더(Autoencoder)가 매핑(Mapping)한 잠재 공간(Latent Space)에서 가우시안(Gaussian) 잡음으로부터 점진적으로 잡음을 제거하는 과정을 통해 고해상도 이미지를 생성한다. 최근에는 잠재 확산 모델의 U자형 신경망 구조(U-Net)에 RGB 이미지를 조건 입력으로 사용하는 방법에 대한 연구가 활발히 진행되고 있다. Zhao[2] 등은 사전 학습된 텍스트 인코더[3]로 추출한 임베딩(Embedding)을 U자형 신경망 구조의 크로스 어텐션(Cross-attention) 모듈에 조건 입력으로 사용하는 잠재 확산 모델 기반의 단안 깊이 추정 모델을 제안하였다. Patni[4] 등은 RGB 이미지로부터 트랜스포머(Transformer)[5]로 추출한 특징을 조건 입력으로 사용하는 구조를 제안하여 단안 깊이 추정 성능을 향상시켰다. Ke[6] 등은 RGB 이미지와 잡음을 주입한 정답 깊이 지도를 연접(Concatenate)하여 잠재 확산 모델의 U자형 신경망 구조에 입력하고, 미세 조정 절차(Fine-tuning Protocol)를 통해 고정밀 단안 깊이 추정을 수행하는 방법을 제안하였다. 그러나, 잠재 확산 모델 기반의 단안 깊이 추정 모델은 잠재 공간에서만 깊이 정보를 학습하기 때문에 픽셀 공간에서의 깊이 정보에 대한 학습이 부족하다는 한계가 있다.

본 논문에서는 잠재 공간과 픽셀 공간에서의 깊이 정보 학습을 통해 깊이 경계 예측 성능을 개선할 수 있는 신경망 구조를 제안한다. 제안하는 방법은 잠재 확산 모델이 예측한 잡음을 깊이 지도로 재구성하고, 정답 깊이 지도로 사전 학습된 오토인코더를 통해 잡음 생성 결과를 픽셀 공간에서 가이드 하여 깊이 추정 성능을 개선하였다. 기준 데이터셋(Benchmark Dataset)에 대한 성능 평가를 통해 기존 방법 대비 제안하는 방법이 확산 모델 기반 단안 깊이 추정에 효과적임을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 제안하는 깊이 지도 재구성 및 오토인코더 결합 구조에 대해 자세히 설명하며, 3장에서는 다양한 실험을 통해 제안하는 방법이 기존 확산 기반 깊이 추정 방법보다 효과적임을 검증한다. 마지막으로 4장에서는 본 논문의 결론을 서술한다.

Ⅱ. 제안하는 방법

제안하는 방법은 오토인코더를 이용해 잠재 공간과 픽셀 공간에서 잠재 확산 모델의 잡음 생성 결과를 가이드하여 깊이 추정 성능을 개선한다. 본 장에서는 먼저 확산 모델을 이용한 단안 깊이 추정 방법에 대해 설명한다. 이어서 제안하는 신경망 구조에 대해 자세히 설명하고, 마지막으로 제안하는 신경망의 훈련에 사용된 손실함수에 관해 설명한다.

1. 확산 모델(Diffusion Model)을 이용한 단안 깊이 추정

단안 깊이 추정 과제를 잡음 제거를 통한 조건부 깊이 지도 생성 과제로 정의하고, 모델이 RGB 이미지를 조건으로 하는 깊이 분포 $D(\boldsymbol{d}|\boldsymbol{x})$를 학습하도록 훈련한다. 여기서 $\boldsymbol{d}$는 깊이 지도 $\boldsymbol{d}\in R^{W\times H}$, $\boldsymbol{x}$는 RGB 이미지 $\boldsymbol{x}\in R^{W\times H\times 3}$를 의미한다.

확산 과정은 모델의 학습을 위해 깊이 지도 $\boldsymbol{d}_0 := \boldsymbol{d}$에 단

계적으로 가우시안(Gaussian) 잡음을 주입하여 잡음이 섞인 깊이 지도 $\boldsymbol{d}_t$를 얻는 과정이다.

$$\boldsymbol{d}_t = \sqrt{\overline{\alpha}_t}\,\boldsymbol{d}_0 + \sqrt{1-\overline{\alpha}_t}\,\boldsymbol{\epsilon}, \tag{1}$$

여기서 $t \in \{1, ..., T\}$는 확산 시간 단계, $\boldsymbol{\epsilon} \sim N(0, I)$는 가우시안 잡음, $\overline{\alpha}_t := \prod_{s=1}^{t} 1-\beta_s$, 그리고 $\{\beta_1, ..., \beta_T\}$는 분산 스케줄을 의미한다. 잡음 제거 과정에서 조건부 잡음 예측 모델 $\boldsymbol{\epsilon}_\theta(.)$는 확산 과정의 확산 시간 단계 t에서 주입된 잡음을 제거하여 $\boldsymbol{d}_t$로부터 $\boldsymbol{d}_{t-1}$을 복원한다.

훈련 과정에서는 확산 과정과 잡음 제거 과정을 통해 매개변수 θ를 갱신한다. 먼저, 확산 시간 단계 t와 가우시안 잡음 $\boldsymbol{\epsilon}$을 샘플링하고, 훈련 데이터 쌍$(\boldsymbol{x}, \boldsymbol{d})$의 깊이 지도 $\boldsymbol{d}$에 잡음을 주입한다. 다음으로 모델은 확산 과정으로 얻은 잡음이 섞인 깊이 지도 $\boldsymbol{d}_t$에서 잡음 $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta(\boldsymbol{d}_t, \boldsymbol{x}, t)$을 예측하고, 목적함수를 최소화한다. 잡음 제거 확산 모델의 목적함수[7]는 수식 (2)와 같다.

$$L = E_{\boldsymbol{d}_0, \boldsymbol{\epsilon} \sim N(0,I), t \sim u(T)} \|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}\|_2^2. \tag{2}$$

추론 과정에서는 가우시안 분포에서 가우시안 잡음 $\boldsymbol{d}_T$을 샘플링하고, 학습된 잡음 제거 모델 $\boldsymbol{\epsilon}_\theta(\boldsymbol{d}_t, \boldsymbol{x}, t)$을 이용해 점진적으로 잡음을 제거하여 깊이 지도 $\boldsymbol{d}_0 := \boldsymbol{d}$를 생성한다.

본 논문에서는 잠재 확산 모델이 RGB 이미지를 조건으로 하는 깊이 분포 $D(\boldsymbol{d}|\boldsymbol{x})$를 학습하도록 훈련한다. 잠재 확산 모델은 잠재 공간에서 확산 과정과 잡음 제거 과정을 통해 RGB 이미지를 조건으로 깊이 지도를 생성하여 효율적으로 고해상도 깊이 지도를 생성할 수 있다. 잠재 확산 모델은 기존 확산 모델과 달리 먼저 RGB 이미지 $\boldsymbol{x}$와 깊이 지도 $\boldsymbol{d}$를 각각 사전 학습된 VAE[8]의 인코더 E를 통해 $\boldsymbol{z}^{\boldsymbol{x}} = \mathrm{E}(\boldsymbol{x})$와 $\boldsymbol{z}^{\boldsymbol{d}} = \mathrm{E}(\boldsymbol{d})$로 압축한다. 이어서 $\boldsymbol{z}^{\boldsymbol{d}}$에 잡음을 주입하는 확산 과정을 진행하고, 잡음 제거 모듈은 $\boldsymbol{z}^{\boldsymbol{d}}_t = \sqrt{\overline{\alpha}_t}\,\boldsymbol{z}^{\boldsymbol{d}} + \sqrt{1-\overline{\alpha}_t}\,\boldsymbol{\epsilon}$에 주입된 잡음 $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}^{\boldsymbol{d}}_t, \boldsymbol{z}^{\boldsymbol{x}}, t)$를 예측한다. 마지막으로 사전 학습된 VAE[8]의 디코더 D는 잡음 제거 모듈이 점진적으로 잡음을 제거하여 생성한 잠재 벡터 $\boldsymbol{z}^{\boldsymbol{d}}_0$를 깊이 지도 $\hat{\boldsymbol{d}} = \mathrm{D}(\boldsymbol{z}^{\boldsymbol{d}}_0)$로 복원한다.

2. 오토인코더를 이용한 잡음 제거 모듈 조정

그림 1은 제안하는 신경망의 훈련 과정과 추론 과정을 보여주며, 그림의 SD는 Stable Diffusion v2를 의미한다. 본 논문에서는 사전 학습된 Stable Diffusion v2[1] 모델을 사용

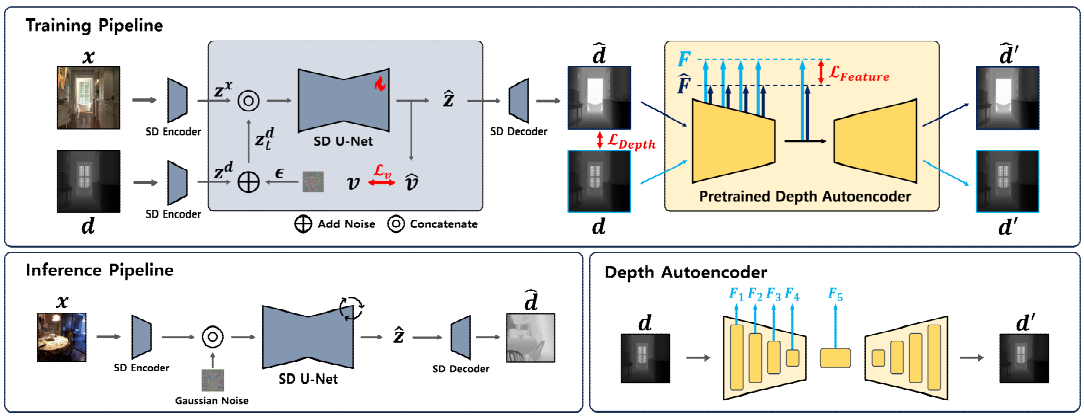


그림 1. 제안하는 방법의 훈련 과정과 추론 과정. 훈련 과정에서는 Stable Diffusion v2[1]의 잡음 제거 모듈을 미세 조정하기 위해 사전 학습된 깊이 오토인코더를 활용한다. 깊이 오토인코더는 깊이 지도를 저차원 잠재 벡터로 압축하고 이를 다시 깊이 지도로 복원하는 과정을 학습하였다.
Fig. 1. Training and Inference Pipeline. In the training pipeline, a pre-trained depth autoencoder is utilized to fine-tune the denoising U-Net of Stable Diffusion v2[1]. The depth autoencoder is trained to compress depth maps into low-dimensional latent vectors and then reconstruct them into depth maps.

하며, 훈련 과정에서 잡음 제거 모듈의 가중치를 미세 조정하여 고정밀 깊이 지도를 생성한다. Stable Diffusion v2[1]는 학습 안정성을 위해 잡음 $\boldsymbol{\epsilon}$가 아닌 속도 $\boldsymbol{v}=\alpha_t\boldsymbol{\epsilon}-\sigma_t\boldsymbol{z}_0$를 예측하도록 학습되었다. 여기서 $\alpha_t=\sqrt{\overline{\alpha_t}}$는 신호 크기 계수, $\sigma_t=\sqrt{1-\overline{\alpha_t}}$는 잡음 크기 계수를 의미한다.

기존 방법은 조건부 깊이 분포 $D(\boldsymbol{d}|\boldsymbol{x})$를 학습하도록 이미지 생성 모델을 미세 조정하여 RGB 이미지로부터 깊이 지도를 생성한다. 이 과정에서 이미지 생성 모델은 기존과 동일한 구조 및 목적함수로 미세 조정되어, 깊이 정보를 이미지 생성 관점에서 처리할 가능성이 있다. 결과적으로, 모델이 입력 데이터로부터 조건부 깊이 분포를 학습함에도 불구하고, 깊이 지도의 경계나 세부 구조와 같은 고유한 특징을 정확히 파악하고 표현하는 능력이 부족할 수 있다는 한계가 있다. 또한, 기존 방법은 잠재 공간에서 잡음 제거 과정을 통해서만 깊이 지도를 생성하기 때문에, 잠재 벡터에 표현되기 어려운 깊이 경계와 같은 아주 미세한 세부 사항에 대한 학습이 제대로 이루어지지 않을 가능성이 있다. 제안하는 방법은 이러한 두 가지 한계를 극복하고 모델이 픽셀 공간에서 깊이 지도의 본질적인 특징을 효과적으로 학습하도록 유도하기 위해 사전 학습된 깊이 오토인코더를 이용한다. 깊이 오토인코더는 깊이 지도의 경계 정보를 효과적으로 보존하면서 저차원 잠재 벡터로 압축하고 다시 깊이 지도로 복원하는 과정을 학습하여, 깊이 오토인코더의 각 인코더에서 추출한 특성 지도는 깊이 지도의 중요한 특징과 구조적 정보를 함축한다. 이러한 특성을 활용하여, 정답 깊이 지도로부터 추출된 특성 지도와 모델이 생성한 깊이 지도로부터 추출된 특성 지도를 비교하면 확산 모델이 픽셀 수준에서 더욱 정밀한 깊이 지도를 생성하기 위해 필요한 깊이 특징을 학습하도록 유도할 수 있다.

훈련 과정에서, Stable Diffusion v2[1]의 잡음 제거 모듈이 예측한 $\hat{\boldsymbol{v}}=\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t^{\boldsymbol{d}},\boldsymbol{z}^{\boldsymbol{x}},t)$를 기반으로 잡음이 제거된 잠재 벡터 $\hat{\boldsymbol{z}}=\alpha_t\boldsymbol{z}_t^{\boldsymbol{d}}-\sigma_t\hat{\boldsymbol{v}}$를 계산하고, VAE[8]의 디코더를 통해 깊이 지도 $\hat{\boldsymbol{d}}$로 재구성하여 깊이 오토인코더에 입력한다. 추론 과정에서 해당 오토인코더는 사용하지 않고 미세 조정한 Stable Diffusion v2[1] 모델만 사용한다.

3. 제안하는 신경망 훈련을 위한 손실함수

본 논문에서는 Stable Diffusion v2[1]에서 사용된 손실함수 $L_{\boldsymbol{v}}$와 GDN[9]에서 제안된 손실함수를 사용하여 제안하는 신경망을 학습한다. 먼저, 평균 제곱 오차를 사용하여 확산 과정에서 계산한 $\boldsymbol{v}$와 잡음 제거 모듈이 예측한 $\hat{\boldsymbol{v}}$ 간의 차이를 수식 (3)과 같이 계산한다.

$$L_{\boldsymbol{v}}=E_{t,\boldsymbol{z}_0,\boldsymbol{\epsilon}}[\|\boldsymbol{v}-\hat{\boldsymbol{v}}\|^2]. \tag{3}$$

다음으로, BerHu[9][10] 손실함수를 사용해 정답 깊이 지도 $\boldsymbol{d}$와 예측 깊이 지도 $\hat{\boldsymbol{d}}$의 차이를 수식 (4)와 같이 계산한다.

$$L_{BerHu}=\begin{cases}|\boldsymbol{d}-\hat{\boldsymbol{d}}|, & \text{if } |\boldsymbol{d}-\hat{\boldsymbol{d}}|<c,\\ \dfrac{|\boldsymbol{d}-\hat{\boldsymbol{d}}|^2+c^2}{2c}, & otherwise,\end{cases} \tag{4}$$

여기서 $c=0.2\times max(|\boldsymbol{d}-\hat{\boldsymbol{d}}|)$이다. L_{BerHu}는 L1 손실과 L2 손실로 구성되며, c 값에 따라 적응적으로 손실을 계산한다.

또한, 경사(Gradient) 일치 손실함수를 사용해 깊이 지도 $\boldsymbol{d}$와 $\hat{\boldsymbol{d}}$의 경사 값 차이를 수식 (5)와 같이 계산한다.

$$L_{GM}=\frac{1}{N}\sum_{i}^{N}|\boldsymbol{d}_{h,i}-\hat{\boldsymbol{d}}_{h,i}|+|\boldsymbol{d}_{v,i}-\hat{\boldsymbol{d}}_{v,i}|, \tag{5}$$

여기서 N은 깊이 지도의 크기, $\boldsymbol{d}_{h,i}$와 $\boldsymbol{d}_{v,i}$는 정답 깊이 지도의 수평과 수직 방향의 i번째 경사 값, $\hat{\boldsymbol{d}}_{h,i}$와 $\hat{\boldsymbol{d}}_{v,i}$는 예측 깊이 지도의 수평과 수직 방향의 i번째 경사 값을 의미한다. 그림 1의 L_{Depth}는 L_{BerHu}와 L_{GM}으로 구성된다.

$$L_{Depth}=L_{BerHu}+L_{GM} \tag{6}$$

마지막으로, 평균 제곱 오차를 사용해 깊이 오토인코더가 추출한 $\boldsymbol{d}$의 특성 지도와 $\hat{\boldsymbol{d}}$의 특성 지도 간의 차이를 계산한다.

$$L_{Feature} = \sum_{j}^{5}\left(\frac{1}{M_j}\sum_{k}^{M_j}|\boldsymbol{F}_{j,k} - \hat{\boldsymbol{F}}_{j,k}|^2\right), \quad (7)$$

여기서 $\boldsymbol{F}_j$와 $\hat{\boldsymbol{F}}_j$는 깊이 오토인코더의 j번째 인코더에서 추출한 $\boldsymbol{d}$와 $\hat{\boldsymbol{d}}$의 특성 지도를 의미하며, 그림 1과 같이 깊이 오토인코더는 5개의 인코더와 4개의 디코더로 구성된다. M_j는 특성 지도 $\boldsymbol{F}_j$와 $\hat{\boldsymbol{F}}_j$의 크기, $\boldsymbol{F}_{j,k}$와 $\hat{\boldsymbol{F}}_{j,k}$는 j번째 특성 지도의 k번째 특성 값을 의미한다. 최종적인 손실 함수는 수식 (8)과 같다.

$$L_{total} = L_v + L_{Depth} + L_{Feature}. \quad (8)$$

Ⅲ. 실험 결과 및 분석

본 논문에서는 제안하는 방법의 학습을 위해 두 개의 합성 데이터셋(Synthetic Dataset)을 사용한다. Hypersim[11]은 461개의 실내 장면으로 구성된 합성 데이터셋이며, 훈련을 위해 365개 장면에 해당하는 53,885개의 이미지와 깊이 지도 데이터 쌍을 사용했다. Virtual KITTI[12]는 도로 주행 장면을 기반으로 다양한 날씨와 카메라 시점에서 합성된 데이터셋이며, 훈련을 위해 4개의 장면에 해당하는 20,148개의 이미지와 깊이 지도 데이터 쌍을 사용했다.

제안하는 방법의 성능 평가를 위해 다섯 개의 벤치마크 데이터셋(Benchmark Dataset)을 사용하였다. NYUv2[13]와 ScanNet[14]은 Kinect 센서를 이용해 수집한 실내 장면으로 구성된 데이터셋이다. 654개의 이미지 쌍으로 구성된 NYUv2[13]의 테스트 데이터셋을 성능 평가에 사용하고, ScanNet[14]의 검증 데이터셋에서 800개의 이미지 쌍을 무작위로 선택하여 성능 평가에 사용했다. KITTI[15]는 LIDAR 센서를 이용해 수집한 도로 주행 영상 데이터셋이며, Eigen[16] 등을 따라 652개 이미지 쌍을 사용했다. ETH3D[17]와 DIODE[18]는 LIDAR 센서를 이용해 수집한 고해상도 데이터셋이다. ETH3D[17]는 전체 454개 이미지 쌍을 성능 평가에 사용하고, DIODE[18]에서는 검증 데이터셋 전체 325개의 실내 장면 이미지 쌍과 446개의 실외 장면 이미지 쌍을 사용한다.

제안하는 방법은 PyTorch[19] 프레임워크를 기반으로 구현되었다. 본 논문에서는 신경망 가중치를 최적화하기 위한 알고리즘으로 Adam[20]을 사용하였고 β_1과 β_2 값은 각각 0.9와 0.999로 설정하였다. 학습 속도(Learning Rate)는 3×10^{-5}으로 설정하고, 배치 크기 32로 18,000 이터레이션(Iteration) 동안 학습을 진행하였다. 학습 데이터는 무작위로 원본 영상을 수평으로 반전하는 데이터 증강 방법을 적용하였다. 훈련에서는 DDPM 잡음 스케줄러[7]를 사용하여 확산 과정을 1,000번의 확산 단계로 설정하고, 추론 시에는 DDIM 스케줄러[21]를 사용하여 확산 모델이 50번의 잡음 제거 과정을 수행한다. 최종 예측 깊이 지도는 서로 다른 가우시안(Gaussian) 잡음으로부터 10번의 추론을 수행한 결과를 종합하여 얻는다. 학습과 성능 평가에는 Intel(R) Xeon(R) E5-1650 v4 @3.60GHz CPU와 NVIDIA RTX 3090 GPU 2대가 이용되었다.

제안하는 방법의 성능 개선 효과를 검증하기 위해, 단안 깊이 추정에서 주로 사용하는 두 가지 평가 지표로 제안하는 방법의 성능을 측정하였다. 구체적으로, 깊이를 정규화하여 예측 깊이와 실제 깊이 간의 오류를 계산하는 절대 상대 오차(Absolute Relative Error)와 예측 깊이가 특정 임계값 안에 속하는지 평가하는 임계값 미만의 정확도(Accuracy Under Threshold)를 활용하여 정량적으로 성능을 비교하였다. 절대 상대 오차는 AbsRel로 표기하고, 임계값 미만의 정확도는 임계값 1.25에 해당하는 δ_1으로 표기하였으며, 실험 결과를 표 1에 나타내었다. 대부분의 기존 방법의 실험 결과는 Marigold[6]에서 제공하는 수치를 사용했으며, Marigold[6] 방법은 제안하는 방법과 동일한 환경에서 재학습하였다.

표 1에 제시된 실험 결과에서 가장 좋은 성능을 굵은 글씨로 표시하고, 두 번째로 좋은 성능은 밑줄로 표시하였다. 절대 상대 오차는 수치가 낮을수록 좋은 성능이며, 임계값 미만의 정확도는 수치가 높을수록 좋은 성능이다. 정량적인 성능 비교 결과, 제안하는 방법이 대부분의 기준 데이터셋에서 절대 상대 오차와 임계값 미만의 정확도에서 기존 방법보다 성능이 향상되어, 제안하는 방법이 정밀한 깊이 지도 생성에 효과적임을 보였다.

표 2에서 베이스라인 모델과 제안하는 방법의 성능 비교

표 1. 기준 데이터셋에서의 정량적 평가 비교
Table 1. Performance comparison on the benchmark datasets

	NYUv2[13]		KITTI[15]		ETH3D[17]		ScanNet[14]		DIODE[18]	
	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1
DiverseDepth[22]	11.7	87.5	19.0	70.4	22.8	69.4	10.9	88.2	37.6	63.1
MiDaS[23]	11.1	88.5	23.6	63.0	18.4	75.2	12.1	84.6	33.2	71.5
LeReS[24]	9.0	91.6	14.9	78.4	17.1	77.7	9.1	91.7	27.1	76.6
Omnidata[25]	7.4	94.5	14.9	83.5	16.6	77.8	7.5	93.6	33.9	74.2
HDN[26]	6.9	94.8	11.5	86.7	12.1	83.3	8.0	93.9	24.6	78.0
DPT[27]	9.8	90.3	10.0	90.1	7.8	94.6	8.2	93.4	**18.2**	75.8
Marigold[6]	5.7	96.0	10.1	90.7	6.4	95.7	6.9	94.4	30.0	77.9
Proposed Method	**5.6**	**96.2**	10.1	**91.1**	**6.3**	**95.9**	**6.3**	**95.5**	29.9	**78.1**

표 2. 베이스라인[6]과 제안하는 방법의 임계값 미만의 정확도 비교
Table 2. Accuracy under threshold comparison of baseline[6] and proposed method

	NYUv2[13]			KITTI[15]			ETH3D[17]			ScanNet[14]			DIODE[18]		
	δ_1	δ_2	δ_3	δ_1	δ_2	δ_3	δ_1	δ_2	δ_3	δ_1	δ_2	δ_3	δ_1	δ_2	δ_3
Baseline[6]	96.0	**99.0**	**99.7**	90.7	**98.5**	99.5	95.7	**98.9**	**99.5**	94.4	98.5	**99.6**	77.9	89.4	93.9
Proposed Method	**96.2**	**99.0**	**99.7**	**91.1**	**98.5**	**99.6**	**95.9**	**98.9**	**99.5**	**95.5**	**98.7**	**99.6**	**78.1**	**89.6**	**94.0**

를 위해 임계값 $1.25(\delta_1)$, $1.25^2(\delta_2)$, $1.25^3(\delta_3)$에 대한 임계값 미만의 정확도를 비교하였다. 5개의 벤치마크 데이터셋에서 임계값의 변화에도 제안하는 방법이 베이스라인 모델 대비 성능 향상을 보였다.

깊이 오토인코더와 깊이 손실함수가 깊이 경계에서의 성능 향상에 효과적임을 입증하기 위한 실험의 결과를 표 3에 나타내었다. 5개의 기준 데이터셋에서 절대 상대 오차(AbsRel)와 임계값 $1.25(\delta_1)$에 대한 임계값 미만의 정확도를 비교하였다. 표 3의 실험 결과를 통해 깊이 손실함수와 깊이 오토인코더를 결합하는 경우에 깊이 오토인코더와 깊이 손실함수가 정밀한 깊이 지도 생성에 효과적임을 보였다.

그림 2와 3은 Marigold[6]와 제안하는 방법의 정성적인 단안 깊이 추정 결과를 보여준다. 그림 2, 3의 실험 결과에서 Marigold[6] 방법으로 생성한 깊이 지도는 객체의 경계를 부드럽게 표현하고, 정답 깊이 지도와 달리 서로 다른 객체가 동일한 깊이로 표현되는 경우가 있다. 그러나 제안하는 방법으로 생성한 깊이 지도는 객체의 경계를 선명하게 표현

표 3. 깊이 오토인코더와 깊이 손실함수의 효과 검증 실험 결과
Table 3. Ablation study of the depth autoencoder and the depth loss function

	NYUv2[13]		KITTI[15]		ETH3D[17]		ScanNet[14]		DIODE[18]	
	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1
Baseline[6]	5.7	96.0	10.1	90.7	6.4	95.7	6.9	94.4	30.0	77.9
Proposed Method (w/ Depth Loss, w/o Depth Autoencoder)	**5.6**	96.1	**10.0**	**91.2**	6.4	**96.0**	6.6	94.9	30.0	**78.1**
Proposed Method (w/ Depth Loss, w/ Depth Autoencoder)	**5.6**	**96.2**	10.1	91.1	**6.3**	95.9	**6.3**	**95.5**	**29.9**	**78.1**

그림 2. NYUv2[13] 데이터셋에 대한 단안 깊이 추정 결과 (a): 입력 이미지 (b): 정답 깊이 지도 (c): 입력 이미지로부터 생성된 Marigold[6]의 깊이 지도 (d): 제안하는 방법으로부터 생성된 깊이 지도

Fig. 2. Results of monocular depth estimation on the NYUv2[13] dataset (a): Input images (b): Ground Truth depth maps (c): Depth maps generated by Marigold[6] from the input image (d): Depth maps generated by the proposed method from the input image

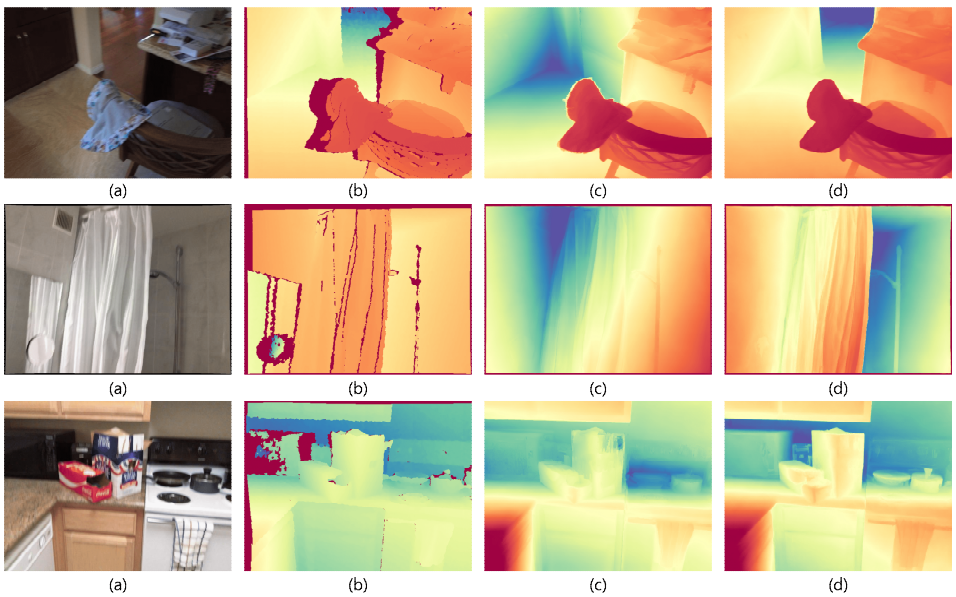


그림 3. ScanNet[14] 데이터셋에 대한 단안 깊이 추정 결과 (a): 입력 이미지 (b): 정답 깊이 지도 (c): 입력 이미지로부터 생성된 Marigold[6]의 깊이 지도 (d): 제안하는 방법으로부터 생성된 깊이 지도

Fig 3. Results of monocular depth estimation on the ScanNet[14] dataset (a): Input images (b): Ground Truth depth maps (c): Depth maps generated by Marigold[6] from the input image (d): Depth maps generated by the proposed method from the input image

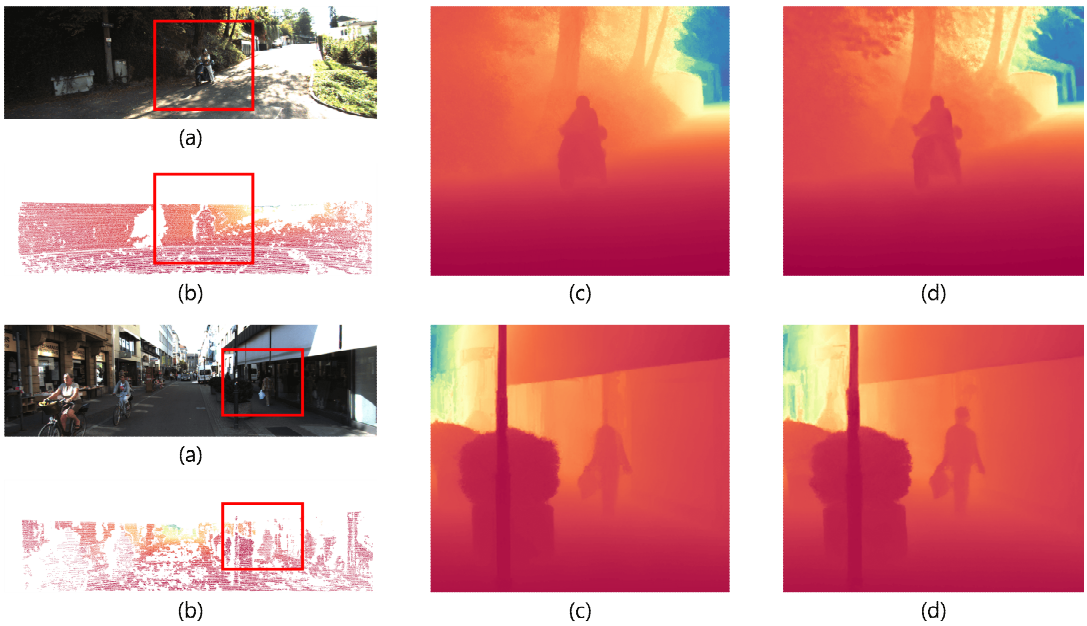


그림 4. KITTI[15] 데이터셋에 대한 단안 깊이 추정 결과 (a): 입력 이미지 (b): 정답 포인트클라우드 (c): 입력 이미지로부터 생성된 Marigold[6]의 깊이 지도 (d): 제안하는 방법으로부터 생성된 깊이 지도
Fig. 4. Results of monocular depth estimation on the KITTI[15] dataset (a): Input images (b): Ground Truth point clouds (c): Depth maps generated by Marigold[6] from the input image (d): Depth maps generated by the proposed method from the input image

하고, 서로 다른 객체의 깊이 차이를 정답 깊이 지도와 유사하게 표현하고 있다. 실험 결과를 통해 제안하는 방법인 깊이 오토인코더를 활용해 픽셀 공간에서 확산 모델의 잡음 생성 결과를 가이드 하는 것이 깊이 경계를 정밀하게 표현하는 데 효과적이었음을 보였다. 그림 4는 Marigold[6]와 제안하는 방법의 깊이 경계에서의 단안 깊이 추정 결과를 보여준다. 제안하는 방법은 기존 방법보다 사람의 형체나 나무의 형태와 같은 미세한 세부사항 표현에서 더 나은 성능을 보였다.

Ⅳ. 결 론

본 논문에서는 잠재 확산 모델 기반의 단안 깊이 추정에서 잡음 예측 결과를 픽셀 공간과 잠재 공간에서 가이드 하는 방법을 제안하였다. 제안하는 방법은 깊이 오토인코더를 활용하여 잡음 제거 모듈의 잡음 예측 결과를 픽셀 공간에서 가이드 한다. 제안하는 방법으로 미세 조정된 잠재 확산 모델은 단일 이미지 입력으로부터 깊이 경계가 정확한 고정밀 깊이 지도를 생성할 수 있다. 다양한 실험 결과를 통해 제안하는 방법이 잠재 확산 모델 기반 단안 깊이 추정 성능을 효과적으로 향상시킬 수 있음을 확인하였다.

참 고 문 헌 (References)

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., Jun. 2022, pp. 10684-10695.
doi: https://doi.org/10.1109/cvpr52688.2022.01042

[2] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, “Unleashing text-to-image diffusion models for visual perception,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2023, pp. 5729-5739.
doi: https://doi.org/10.1109/iccv51070.2023.00527

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., “Learning transferable visual models from natural language supervision,” in Proc. Int. Conf. Mach. Learn., Jul. 2021, pp. 8748 - 8763.
doi: https://doi.org/10.48550/arXiv.2103.00020

[4] S. Patni, A. Agarwal, and C. Arora, “ECoDepth: Effective conditioning of diffusion models for monocular depth estimation,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., Jun. 2024, pp. 28285-28295.
doi: https://doi.org/10.1109/cvpr52733.2024.02672

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Represent., May 2021.
doi: https://doi.org/10.48550/arXiv.2010.11929

[6] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, “Repurposing diffusion-based image generators for monocular depth estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2024, pp. 9492-9502.
doi: https://doi.org/10.1109/CVPR52733.2024.00907

[7] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in Proc. Adv. Neural inf. Process. Syst., Dec. 2020, pp. 6840-6851.
doi: https://doi.org/10.48550/arXiv.2006.11239

[8] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” in Proc. Int. Conf. Learn. Represent., Apr. 2014.
doi: https://doi.org/10.48550/arXiv.1312.6114

[9] M. Song and W. Kim, “Depth estimation from a single image using guided deep network,” IEEE Access, vol. 7, pp.142595-142606, Oct. 2019.
doi: https://doi.org/10.1109/ACCESS.2019.2944937

[10] L. Zwald and S. Lambert-Lacroix, “The BerHu penalty and the grouped effect,” arXiv:1207.6868, Jul. 2012.
doi: https://doi.org/10.48550/arXiv.1207.6868

[11] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind, “Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2021, pp. 10912 – 10922.
doi: https://doi.org/10.1109/ICCV48922.2021.01073

[12] Y. Cabon, N. Murray, and M. Humenberger, “Virtual KITTI 2,” arXiv:2001.10773, Jan. 2020.
doi: https://doi.org/10.48550/arXiv.2001.10773

[13] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from RGBD images,” in Proc. Eur. Conf. Comput. Vis., Oct. 2012, pp. 746 – 760.
doi: https://doi.org/10.1007/978-3-642-33715-4_54

[14] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “ScanNet: Richly-annotated 3D reconstructions of indoor scenes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 2432 – 2443.
doi: https://doi.org/10.1109/cvpr.2017.261

[15] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2012, pp. 3354 – 3361.
doi: https://doi.org/10.1109/CVPR.2012.6248074

[16] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” arXiv:1406.2283, Jun. 2014.
doi: https://doi.org/10.48550/arXiv.1406.2283

[17] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, “A multi-view stereo benchmark with high-resolution images and multi-camera videos,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 3260 – 3269.
doi: https://doi.org/10.1109/CVPR.2017.272

[18] I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F. Z. Dai, A. F. Daniele, M. Mostajabi, S. Basart, M. R. Walter, and G. Shakhnarovich, “DIODE: A Dense Indoor and Outdoor Depth Dataset,” arXiv:1908.00463, Aug. 2019.
doi: https://doi.org/10.48550/arXiv.1908.00463

[19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S.Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “PyTorch: An imperative style, highperformance deep learning library,” in Proc. Conf. Neural Inf. Process. Syst., Dec. 2019, pp. 8024 – 8035.
doi: https://dl.acm.org/doi/10.5555/3454287.3455008

[20] D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization,” in Proc. Int. Conf. Learn. Represent., pp. 1-15, May 2015.
doi: https://doi.org/10.48550/arXiv.1412.6980

[21] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” arXiv:2010.02502, Oct. 2020.
doi: https://doi.org/10.48550/arXiv.2010.02502

[22] W. Yin, X. Wang, C. Shen, Y. Liu, Z. Tian, S. Xu, C. Sun, and D. Renyin, “DiverseDepth: Affine-invariant Depth Prediction Using Diverse Data,” arXiv:2002.00569, Feb. 2020.
doi: https://doi.org/10.48550/arXiv.2002.00569

[23] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 3, pp. 1623-1637, Mar. 2022.
doi: https://doi.org/10.1109/TPAMI.2020.3019967

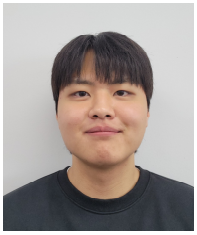
[24] W. Yin, J. Zhang, O. Wang, S. Niklaus, L. Mai, S. Chen, and C. Shen, “Learning to Recover 3D Scene Shape From a Single Image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2021, pp. 4509-4519.
doi: https://doi.org/10.1109/CVPR46437.2021.00451

[25] A. Eftekhar, A. Sax, J. Malik, and A. Zamir, “Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets From 3D Scans,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2021, pp. 10786 – 10796.
doi: https://doi.org/10.1109/ICCV48922.2021.01061

[26] C. Zhang, S. Pan, Z. Hu, P. Huang, J. Qin, and G. Li, “Hierarchical Normalization for Robust Monocular Depth Estimation,” in Proc. Adv. Neural Inf. Process. Syst., Dec. 2022.
doi: https://doi.org/10.48550/arXiv.2210.09670

[27] R. Ranftl, A. Bochkovskiy, and V. Koltun, “Vision Transformers for Dense Prediction,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2021, pp. 12179 – 12188.
doi: https://doi.org/10.1109/ICCV48922.2021.01061

저 자 소 개



김 현 우

- 2020년 3월 ~ 현재 : 건국대학교 전기전자공학부 학사과정
- ORCID : https://orcid.org/0009-0009-3120-7067
- 주관심분야 : 컴퓨터 비전, 영상처리, 심층학습

김 원 준

- 2012년 8월 : 한국과학기술원(KAIST) 박사
- 2012년 9월 ~ 2016년 2월 : 삼성종합기술원 전문 연구원
- 2016년 3월 ~ 2020년 2월 : 건국대학교 전기전자공학부 조교수
- 2020년 3월 ~ 현재 : 건국대학교 전기전자공학부 부교수
- ORCID : https://orcid.org/0000-0001-5121-5931
- 주관심분야 : 컴퓨터 비전, 영상처리, 기계학습