특집논문 (Special Paper)
방송공학회논문지 제29권 제6호, 2024년 11월 (JBE Vol.29, No.6, November 2024)
https://doi.org/10.5909/JBE.2024.29.6.866
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

비정렬 열화상 이미지와 RGB 이미지에 대한 상호 참조 및 융합을 통한 듀얼 모달 깊이 추정

권 병 준[a)], 김 문 철[a)‡]

Dual-Modal Depth Estimation based on Cross Reference and Fusion for Misaligned Thermal and RGB Images

Byeongjun Kwon[a)] and Munchurl Kim[a)‡]

요 약

이미지 깊이 추정 연구는 자율 주행, 가상 현실 등 다양한 분야에 활용도가 높은 연구이다. 이미지 깊이 추정(depth estimation)은 입력 이미지의 각 픽셀이 3차원 공간에 대응되었을 때의 대응된 각 픽셀이 카메라로부터 3차원 공간 내 위치까지의 거리를 추정하는 것을 목표로 한다. 열화상 이미지는 야간과 같이 조명이 부족한 상황에서 어느 정도의 신뢰할 수 있는 영상 정보를 제공하는 반면, RGB 이미지는 낮 동안 일관된 고품질 데이터를 제공하기 때문에, 본 논문에서는 열화상 이미지와 RGB 이미지를 모두 활용하여 이미지 각각의 깊이 정보를 추정하는 효과적인 딥러닝 방법을 제안한다. 또한, 일반적으로 열화상 이미지와 RGB 이미지는 비정렬 상태로 획득되기 때문에 두 이미지를 깊이 정보 추정에 상호보완적으로 사용하기 위해, 본 논문에서는 (i) 비정렬된 열화상 이미지와 RGB 이미지에서의 특징 추출과 이들의 듀얼 모달리티 교차 융합 모듈, (ii) 듀얼 모달리티 입력을 처리하기 위한 공유 인코더-디코더 구조, (iii) 듀얼 모달리티에서의 동시 지도 훈련을 위한 다중 목적 학습 전략을 제시한다. 다양한 실험을 통해 제안 방법의 효과성을 검증하였으며, 단일 모달 입력(열화상 이미지, RGB 이미지)만을 사용하는 경우 대비 7%, 4%의 성능 향상을 확인하였다.

Abstract

Research on depth estimation is highly applicable in various fields, such as autonomous driving and virtual reality. Depth estimation aims to predict the distance from the camera to the position in 3D space corresponding to each pixel when it is mapped from a 2D image. Thermal images provide reliable visual information in low-light conditions, such as at night, while RGB images offer consistent high-quality data during the day. Therefore, this paper proposes an effective deep learning method that utilizes both thermal and RGB images to estimate depth information from each image. Additionally, because thermal and RGB images are generally captured in an misaligned state, this paper introduces (i) feature extraction from misaligned thermal and RGB images and their Cross-fusion module and their dual-modality cross-fusion module, (ii) a shared encoder-decoder structure for processing dual-modality inputs, and (iii) a multi-objective training strategy for simultaneous supervised learning from both modalities. We verify the effectiveness of the proposed method through various experiments, achieving performance improvements of 7% and 4% compared to using single-modal inputs (thermal images or RGB images).

Keywords : Depth Estimation, Dual-modal, Misalignment, Robustness

I. 서 론

3D 공간의 구조를 이해하는 것은 컴퓨터 비전 분야에서 매우 중요하다. 일반적으로 2D 이미지는 3D 공간을 2D 평면에 투영한 결과이므로 깊이 정보를 잃게 된다. 따라서 이미지의 각 픽셀에 해당하는 깊이를 추정하는 것은 3D 구조를 재구성하는 데 필수적이다. 최근 자율 주행 및 증강 현실(AR)과 같은 기술이 발전함에 따라 이러한 과정은 더욱 중요해지고 있다.

LiDAR 센서의 깊이 값을 참조하여 깊이 추정 결과를 비교하며 2D 이미지의 깊이를 추정 학습하는 방식인 지도 학습 기반 단안 깊이 추정 기법들이 활발히 연구되고 있다[3,4,5,6]. 그러나 학습 데이터는 양호한 환경(충분한 조도, 맑은 날씨 등)에서 촬영된 깨끗한 RGB 이미지로 구성되어 있기 때문에, 깨끗한 RGB 학습 데이터를 사용하여 학습된 깊이 추정 모델은 열악한 조건에서 적용될 경우 성능 저하를 초래한다.

열악한 조건에서도 이미지의 깊이 추정 정확도를 유지할 수 있는 강건성(Robustness)을 가지는 모델 학습을 위해, 열화상 이미지를 추가적으로 사용하는 연구가 진행되기 시작하였다[7,8,9]. 하지만 RGB 이미지 대신 열화상 이미지만을 사용하는 경우, 열화상 이미지의 낮은 색 대비 및 적은 텍스처 정보로 인해 성능 저하가 발생한다. 이러한 이유로, 본 논문에서는 깊이 예측 학습의 강건성을 향상시키기 위해 RGB 이미지와 열화상 이미지 각각의 모달리티 장점을 모두 결합한 방법을 사용한다.

그러나 각 모달리티의 카메라는 서로 다른 시야각과 해상도를 가지므로, 동일한 장면을 동시에 촬영하더라도 각 이미지간의 비정렬로 인해 동일한 객체가 각 이미지에서 다른 위치에 나타나게 된다. 따라서, 본 연구에서는 비정렬 RGB 이미지와 열화상 이미지 입력에 대해 두 이미지 각각의 깊이를 예측(dual-modal depth estimation)하는 문제에 초점을 맞추고자 한다.

a) 한국과학기술원 전기전자공학부(School of Electronic and Electrical Engineering, KAIST)
‡ Corresponding Author : 김문철(Munchurl Kim)
E-mail: mkimee@kaist.ac.kr
Tel: +82-42-350-7419
ORCID: https://orcid.org/0000-0003-0146-5419
※본 논문은 저자인 권병준의 카이스트 석사 학위 논문인 “비정렬 열 영상과 자연 영상으로부터의 다중 모달 깊이 추정”[1]을 기반으로 기술되었음.
※본 논문은 한국방송·미디어공학회 2024년 하계학술대회에 제출한 “Multi-modal Depth Estimation from Misaligned Thermal and RGB Images”[2]의 후속 논문임.
※이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. RS-2022-00144444,딥러닝 기반 정적 및 동적 장면의 공간 영상 표현 학습 및 렌더링 연구).
· Manuscript September 4, 2024; Revised October 8, 2024; Accepted October 10, 2024.

II. 관련 연구

1. 단안 깊이 추정 연구

단안 깊이 추정은 하나의 이미지만을 통해서 깊이를 예측해야한다. 스테레오 깊이 추정과 같이 두 이미지의 상관성을 통해 깊이를 예측하지 않기 때문에 이미지에 대한 전반적인 이해가 중요하다. 이를 위해 일반적인 단안 깊이 추정 모델은 인코더와 디코더로 구성된다. 인코더를 통해 특징 맵을 추출한 이후 이미지 전체에 대한 이해를 포함하고 있는 특징 맵을 기반으로 디코더를 통해 깊이를 추정한다. 이후, 예측 깊이와 실제 깊이 사이의 손실 함수를 통해 모델을 학습한다.

단안 깊이 추정은 실용적인 응용(예: 자율 주행 차량 및 스마트폰 기반의 증강 현실)으로 인해 큰 관심을 받고 있다. 그러나 단안 깊이 추정은 단일 이미지에서 얻은 깊이의 스케일 불확정성 문제로 인해 무한히 많은 경우의 수가 존재할 수 있으므로 불안정한 해를 구하는 문제(ill-posed problem)로 간주된다.

Eigen[3]은 손수 설계된 특징(예: 저수준 슈퍼픽셀 또는 윤곽선) 없이 단일 이미지에서 깊이를 예측하기 위해 깊은 컨볼루션 신경망을 훈련하는 개념을 처음 도입하였다. 이 연구는 전역적인 특징을 학습할 수 있는 다중 스케일 컨볼루션 신경망을 제안하여, 세밀한 깊이 맵을 예측하기 위한 개선 네트워크를 사용한다. 또한, 내재된 스케일 불확정성 문제를 해결하기 위해 스케일 불변 손실을 제안하였다.

Fu[4]는 깊이 추정 문제를 이산화된 순서 회귀 문제로 재정의한 첫 번째 연구를 발표하였다. 그들은 회귀 훈련 손실 함수가 훈련 단계에서 느린 수렴과 흐릿한 깊이 추정 결과를 초래하기 때문에 이산화된 깊이 값을 사용하는 다중 클래스 분류 손실을 활용하였다. 또한, 매우 큰 깊이 값이 훈련에 미치는 과도한 영향을 해결하기 위해 간격 증가 이산

화 방법을 제안하였다.

Ranftl[5]는 깊이 추정 및 시멘틱 세그멘테이션과 같은 밀집 예측을 위한 인코더로 비전 트랜스포머를 사용하는 아키텍처를 소개하였다. 트랜스포머 기반의 인코더가 셀프 어텐션 메커니즘을 통해 일관된 특징을 추출하고, 큰 수용 영역을 가지고 있기 때문에 보다 전역적으로 일관된 깊이 맵을 예측할 수 있다. 또한, 트랜스포머 인코더와 컨볼루션 디코더를 결합하여 다중 스케일 특징 맵을 효과적으로 결합하였다.

Yuan[6]는 인접한 깊이 값들 간의 상관 관계에 중점을 두고, 뉴럴 윈도우 완전 연결 조건부 랜덤 필드(CRF) 기반 깊이 디코더를 제안하였다. 계산 복잡성 문제로 인해 이전 연구들은 전체 그래프가 아닌 인접 그래프간의 CRF를 적용하여 완전 연결 CRF의 이점을 포기하였지만 해당 연구는 윈도우 기반 접근 방식을 채택하여 계산 복잡성 문제를 해결하는 동시에 완전 연결 CRF의 이점을 얻을 수 있었다.

깊이 추정 모델의 발전에도 불구하고, 앞서 언급한 모델들은 양호한 환경(충분한 조도, 맑은 날씨 등)에서 얻은 RGB 이미지가 존재해야 한다는 조건에 종속된다. 따라서 저녁이나 비 오는 날과 같은 악조건에서는 성능 저하가 심각하다는 한계가 있다.

2. 열화상 이미지를 사용한 깊이 추정 연구

열화상 카메라는 날씨와 조명 조건의 변화에 적게 영향을 받기 때문에 RGB 카메라에 비해 일관된 이미지 품질을 유지한다. 열화상 이미지의의 강건성은 날씨나 조명 조건에 강건한 깊이 추정 네트워크를 설계하는 데 주목을 끈다.

Kim[7]은 RGB 스테레오 이미지와 열화상 이미지를 사용하는 자가 지도 멀티 모달리티 구조를 제안했다. 스테레오 RGB 이미지와 왼쪽 RGB 이미지에 정렬되어 있는 열화상 이미지를 가지고 있을 때에 열화상 이미지를 통해 추정한 깊이, 오른쪽 RGB 이미지와 카메라의 외부 파라미터를 기반으로 오른쪽 RGB 이미지를 워핑하여 왼쪽 RGB 이미지를 재구성한다. 재구성한 왼쪽 RGB 이미지와 실제 왼쪽 RGB 이미지에 손실함수를 적용하여 열화상 깊이 추정 모델을 학습하였다. 또한 다양한 조건을 포함하는 대규모 멀티스펙트럴 스테레오 데이터셋을 제공한다.

Lu[8]는 RGB와 열화상 도메인 간의 변환을 위한 GAN과 열화상 이미지에서 깊이를 예측하는 컨볼루션 모델을 결합한 비지도 학습 모델을 제안했다. 학습 과정에서 스테레오 RGB 이미지와 단일 열화상 이미지가 모두 사용된다. 왼쪽 RGB 이미지는 GAN을 통해 열화상 이미지로 변환된 후, 실제 열화상 이미지에서 얻은 깊이 추정 결과를 기반으로 이미지를 워핑하여, 워핑된 열화상 이미지와 실제 열화상 이미지 간의 손실함수를 계산하며 학습을 진행한다.

Shin[9]은 다양한 조건(예: 주간, 야간, 비)을 포함하는 대규모 멀티스펙트럴 스테레오(MS2) 데이터셋을 제공하며, 이 데이터셋은 스테레오 RGB, 스테레오 NIR, 스테레오 열화상, 스테레오 LiDAR 데이터를 포함한다. 또한 이들은 단일 모달리티와 스테레오 시나리오 모두에서 효과적으로 작동하는 열화상 깊이 추정 모델을 제안했으며, 단안 및 스테레오 설정에서 학습할 수 있는 지식을 코스트볼륨 방법을 통해 효과적으로 통합하였다.

III. 제안 방법

본 장에서는 비정렬 RGB 및 열화상 이미지로부터 강건하게 깊이 정보를 추정하기 위해, 듀얼 모달리티 교차 융합 깊이 추정 네트워크인 DMCF-SIDE (Dual-Modality Cross-Fusion based Single Image Depth Estimation) Net을 제안한다. 제안한 DMCF-SIDE Net의 주요 특징은 서로 다른 모달리티(RGB 및 열화상 이미지)의 특징을 깊이 추정에 효과적으로 융합하는 것이다. DMCF-SIDE Net은 ‘특징 추출’, ‘특징 융합’, ‘깊이 예측을 위한 디코더’의 세 단계로 구성된다. 열화상 이미지와 RGB 이미지 간의 해상도 불일치 문제를 해결하기 위해, 모든 RGB 이미지를 열화상 이미지와 동일한 해상도를 갖도록 스케일을 조정한다. 그럼에도 불구하고, 서로 다른 모달리티 이미지는 비정렬 상태로 존재하므로, 각 모달리티에 대한 깊이 추정은 자연스럽게 비정렬되고 독립적일 수밖에 없다. 따라서, 본 모델은 각 모달리티에 대해 별도의 깊이 추정을 수행한다.

그림 1은 DMCF-SIDE Net의 전체 개요를 나타낸다. DMCF-SIDE Net은 크게 인코더와 디코더로 구성된 U-Net 구조와 유사한 특징을 가진다. 각 스케일의 인코더 레이어

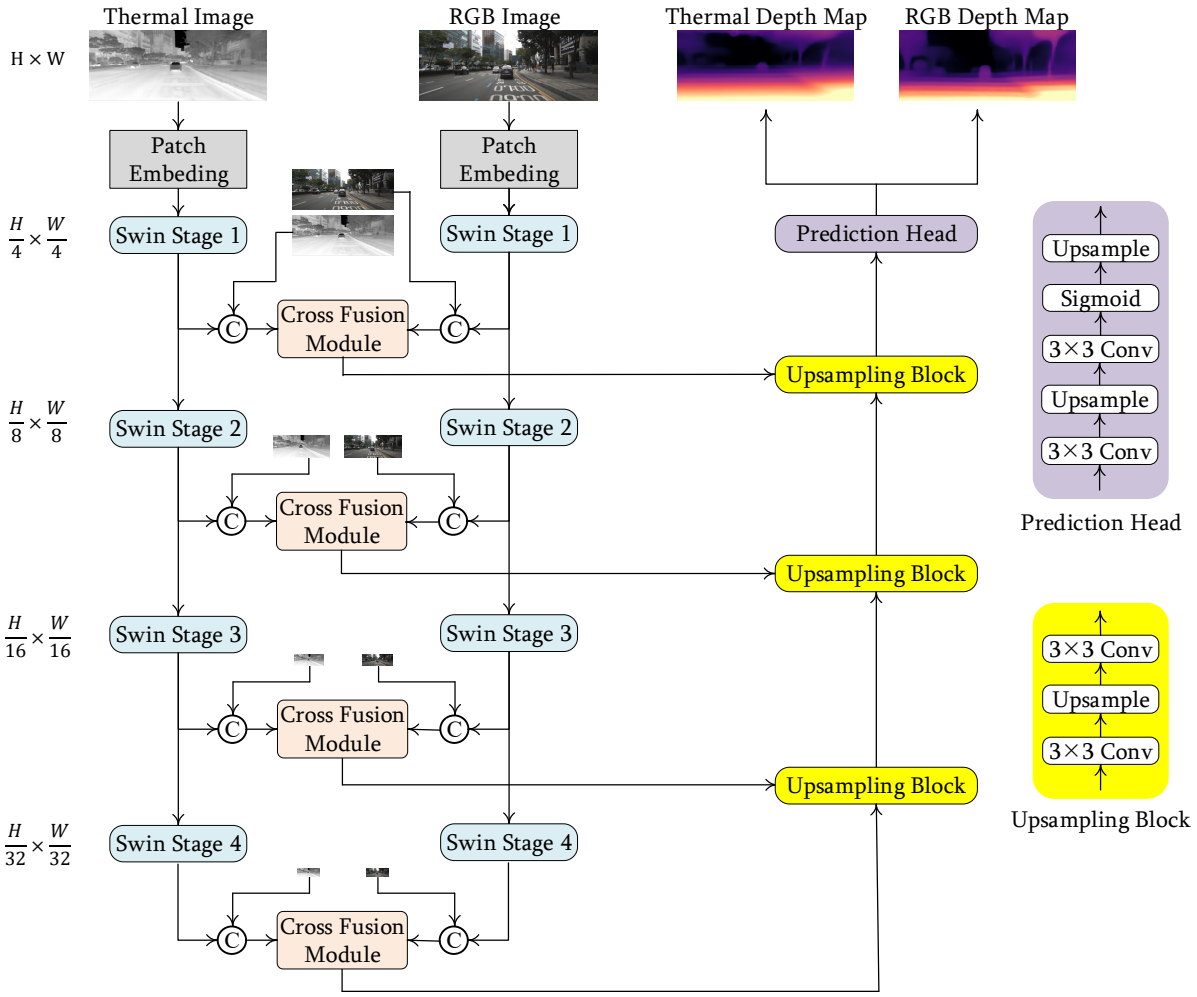

그림 1. 제안한 듀얼 모달리티 교차 융합 깊이 추정 네트워크의 전체 구조
Fig. 1. Overall Framework of Proposed DMCF-SIDE Net

와 해당되는 디코더가 직접 연결되지 않고, 듀얼 모달리티 교차 융합 모듈을 거쳐 연결되어 있다. 열화상 이미지 $\mathbf{I}_{thr}\in\mathrm{R}^{W\times H\times 1}$와 RGB 이미지 $\mathbf{I}_{rgb}\in\mathrm{R}^{W\times H\times 3}$가 Patch Embedding을 거쳐 각각 $I_{thr}^{PE}\in\mathbb{R}^{\frac{W}{4}\times\frac{H}{4}\times 96}$, $I_{rgb}^{PE}\in\mathbb{R}^{\frac{W}{4}\times\frac{H}{4}\times 96}$으로 변환된 이후 첫 번째 Swin Transformer[10] 단계로 입력된다. DMCF-SIDE Net의 최종 출력으로 각 모달리티에 대한 깊이 추정인 $\mathbf{D}_{thr},\mathbf{D}_{rgb}\in\mathrm{R}^{W\times H\times 1}$을 얻는다.

본 실험에서 사용한 훈련데이터셋에는 열화상과 RGB 영상에 대한 GT가 각각 존재하며, 서론에서 설명했듯이 각 모달리티의 카메라는 서로 다른 시야각을 가지므로 두 모달리티의 수퍼비전을 동시에 사용하기 위해 학습시에는 $\mathbf{D}_{thr},\mathbf{D}_{rgb}$를 동시에 추정하지만 추론시에는 원하는 모달리티에 대한 깊이만 추정할 수 있다.

1. 듀얼 모달리티 교차 융합 모듈

이전에서 설명한 바와 같이, 서로 다른 모달리티 이미지 간의 비정렬로 인해 서로 다른 위치에서의 특징들 간의 상관관계가 존재한다. 따라서 서로 다른 위치에 있는 상호보완적 특징을 추출하기 위해 트랜스포머의 어텐션 기반 서

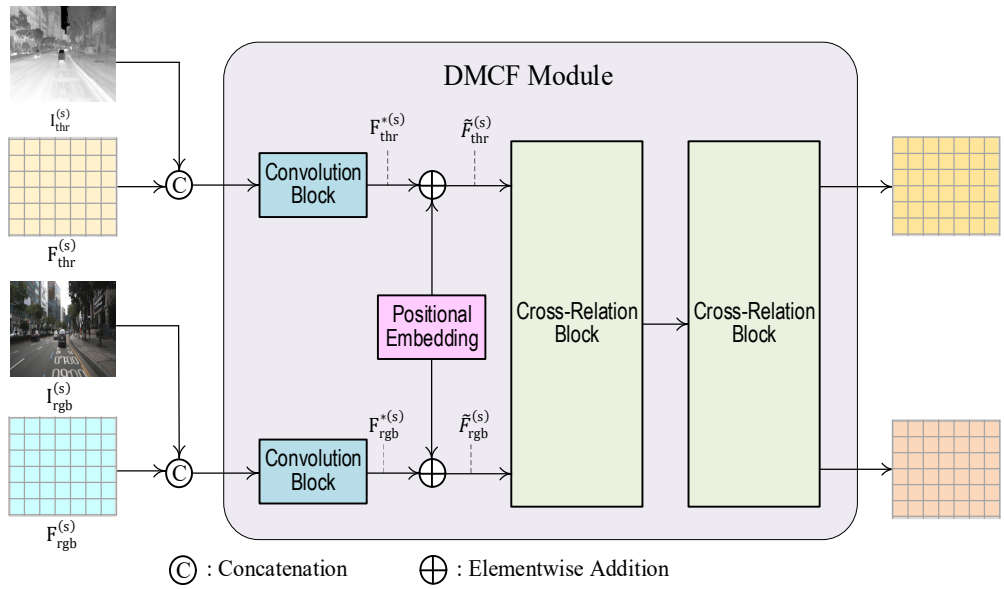

그림 2. 듀얼 모달리티 교차 융합 모듈 구조
Fig. 2. Architecture of a Dual-Modality Cross -Fusion (DMCF) Module

로 다른 모달리티 간의 상호 참조를 통해 상관관계를 효과적으로 융합하고자 한다. 그림 2는 듀얼 모달리티 교차 융합(Dual-Modality Cross Fusion, DMCF) 모듈의 구조를 나타낸다.

각 모달리티의 특징을 향상시키기 위해, GMFlow[11] 구조 내 특징 향상 레이어의 입력 구조를 개선하였다. 기존 특징 향상 레이어는 특징 입력을 직접적으로 사용하는 것에 비해, 제안된 DMCF 모듈은 특징 입력과 다운샘플된 입력 이미지를 concatenation 후에 convolution block을 통과시킨 결과를 입력으로 사용한다. 이렇게 GMFlow의 특징 향상 레이어 입력부 구조를 개선함으로써 깊이 영상의 구조적 정보를 더 효과적으로 추정할 수 있었다. 제안한 DMCF는 각 모달리티의 강점을 상호 보완적으로 활용하는 것을 목표로 한다. 입력 이미지 해상도의 1/4, 1/8, 1/16, 1/32 크기에 해당하는 네 가지 다른 스케일(s = 1/4, 1/8, 1/16, 1/32)에서 특징 $\mathbf{F}^{(s)}_{\text{modal}}$, modal = 'thr' or 'rgb'을 추출한다. 이러한 특징들은 공간 구조를 유지하기 위해 각 모달리티의 다운샘플된 입력 이미지 $\mathbf{I}^{(s)}_{thr} \in \mathrm{R}^{sW \times sH \times 1}$, $\mathbf{I}^{(s)}_{rgb} \in \mathrm{R}^{sW \times sH \times 3}$와 연결된 후, ConvBlock(두 번의 합성곱과 배치 정규화)을 사용하여 채널 크기를 줄인다. 그림 2에서 ConvBlock의 출력은 $\mathbf{F}^{*(s)}_{\text{modal}} = \text{ConvBlock}\left(\text{concat}\left(\mathbf{F}^{(s)}_{\text{modal}}, \mathbf{I}^{(s)}_{\text{modal}}\right)\right)$으로 표현되며, 'concat'은 연결(concatenation)을, $\mathbf{I}^{(s)}_{\text{modal}}$은 평균 풀링을 통해 얻은 다운샘플된 이미지를 나타낸다. $\mathbf{F}^{(s)}_{\text{modal}}$은 스케일 s에서의 Swin Transformer Stage[10] 출력 특징으로 attention weight를 계산할 때에만 상대적 위치 임베딩 정보를 고려하기 때문에 $\mathbf{F}^{(s)}_{\text{modal}}$는 공간적 위치 정보가 부족할 수 있어, 이를 보완하기 위해 [11]에서 설명한 방법을 따라 절대적 위치 임베딩 PE를 사용한다. 이 단계는 $\tilde{\mathbf{F}}^{(s)}_{\text{modal}} = \mathbf{F}^{*(s)}_{\text{modal}} + \text{PE}$로 표현된다.

2. 교차 관계 블록

그림 2의 DMCF 모듈은 2개의 교차 관계(Cross Relation, CR) 블록을 포함하고 있다. 교차 관계 블록은 DMCF 모듈 내에서 듀얼 모달리티에 대한 상호 참조 및 융합 과정을 처리하기 위해 고안되었다. 각 CR 블록의 구조는 그림 3과 같다.

그림 3에서 보듯이 CR 블록은 각 모달리티 내 셀프-어텐션과 듀얼 모달리티 간 크로스-어텐션으로 구성된다. 셀프-어텐션은 각 모달리티 내(Intra-Modality)에서의 고유 정보를

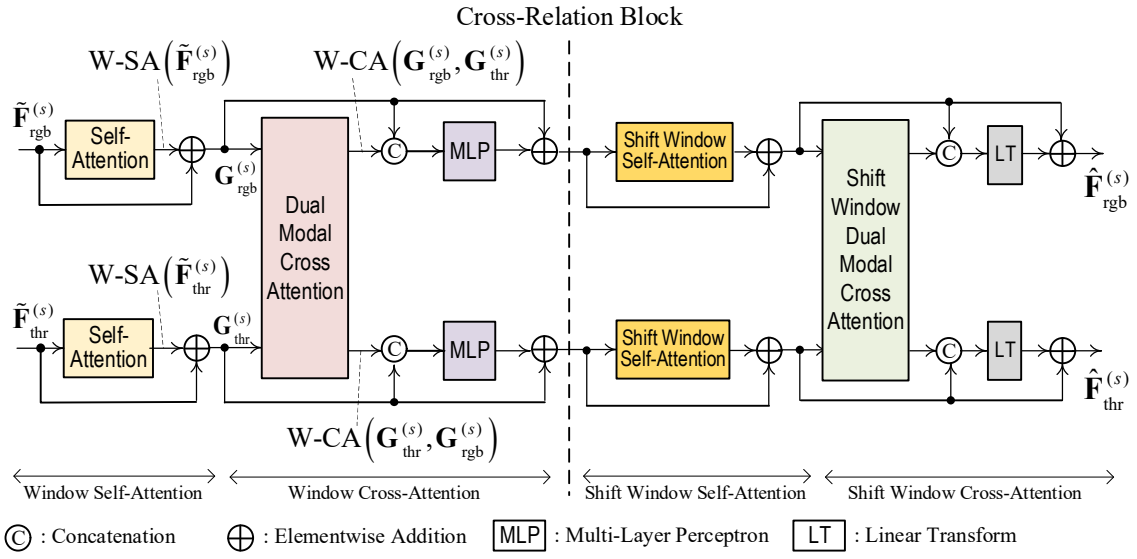

그림 3. 교차 관계 블록 구조
Fig. 3. Architecture of Cross-Relation Block

강화하며, 크로스-어텐션은 두 모달리티 간(Inter-Modality)의 정보 교환을 촉진하여 특징 간 상호 강화 효과를 유도한다. DMCF 모듈을 모든 스케일의 트랜스포머 단계로부터 출력된 특징에 대해 적용하면서도 메모리 사용량을 줄이기 위해 Swin Transformer[10]의 윈도우 셀프-어텐션 방법을 차용하여 어텐션을 계산한다.

어텐션 기반 특징 향상 과정으로서, 스케일 s의 트랜스포머 단계의 출력인 $\tilde{\mathbf{F}}^{(s)}_{\text{modal}}$가 CR 블록의 입력이 된다. 각 모달리티에 대한 윈도우 셀프-어텐션 연산은 다음과 같이 계산된다:

$$\text{W}-\text{SA}\left(\tilde{\boldsymbol{F}}^{(s)}_{\text{modal}}\right)=\text{softmax}\left(\frac{Q_{\text{modal}}\text{K}^{\text{T}}_{\text{modal}}}{\sqrt{d}}\right)\text{V}_{\text{modal}} \quad (1)$$

여기서 $\text{Q,K,V}\in\text{R}^{\text{M}^2\times d}$는 모두 $\tilde{\mathbf{F}}^{(s)}_{\text{modal}}$로부터 계산된다. M과 d는 각각 윈도우 크기와 채널 차원의 개수를 나타낸다. W-SA 연산은 Q, K, V를 동일한 특징으로부터 가져오므로, 윈도우 내에서 셀프-상관된 특징들을 강화한다.

그림 3에서 윈도우 크로스-어텐션 연산은 다음과 같이 계산된다.

$$\text{W-CA}\left(\mathbf{G}^{(s)}_{thr},\mathbf{G}^{(s)}_{rgb}\right)=\text{softmax}\left(\frac{Q_{\text{thr}}K^{T}_{\text{rgb}}}{\sqrt{d}}\right)\cdot V_{\text{rgb}}$$
$$\text{W-CA}\left(\mathbf{G}^{(s)}_{\text{rgb}},\mathbf{G}^{(s)}_{\text{thr}}\right)=\text{softmax}\left(\frac{Q_{\text{rgb}}K^{T}_{\text{thr}}}{\sqrt{d}}\right)\cdot V_{\text{thr}} \quad (2)$$

여기서 $\text{W-CA}\left(\mathbf{G}^{(s)}_{\text{thr}},\mathbf{G}^{(s)}_{\text{rgb}}\right)$와 $\text{W-CA}\left(\mathbf{G}^{(s)}_{\text{rgb}},\mathbf{G}^{(s)}_{\text{thr}}\right)$의 쿼리는 각각 $Q_{\text{thr}},Q_{\text{rgb}}\in\text{R}^{\text{M}^2\times d}$이며 $Q_{\text{thr}}=\mathbf{W}^{\text{thr}}_{Q,s}\cdot\mathbf{G}^{(s)}_{\text{thr}}$와 $Q_{\text{rgb}}=\mathbf{W}^{\text{rgb}}_{Q,s}\cdot\mathbf{G}^{(s)}_{\text{rgb}}$인 선형 변환($\mathbb{R}^{M^2\times C}\rightarrow\mathbb{R}^{M^2\times d}$)으로부터 계산되며, $K_{\text{thr}},K_{\text{rgb}},V_{\text{thr}},V_{\text{rgb}}\in\text{R}^{\text{M}^2\times d}$는 각각 $K^{(s)}_{thr}=W^{thr}_{K,s}G^{(s)}_{thr}$, $K^{(s)}_{rgb}=W^{rgb}_{K,s}G^{(s)}_{rgb}$, $V^{(s)}_{thr}=W^{thr}_{V,s}G^{(s)}_{thr}$, $V^{(s)}_{rgb}=W^{rgb}_{V,s}G^{(s)}_{rgb}$으로부터 계산된다. W-CA는 타겟 특징인 $\mathbf{G}^{(s)}_{\text{thr}}\left(\text{or}\,\mathbf{G}^{(s)}_{\text{rgb}}\right)$로부터 모든 $K^{(s)}_{\text{thr}}\left(\text{or}\,K^{(s)}_{\text{rgb}}\right)$와 $V^{(s)}_{\text{thr}}\left(\text{or}\,V^{(s)}_{\text{rgb}}\right)$를 계산하므로, 이는 CR 블록이 서로 다른 모달리티에서 관련된 특징을 추출할 수 있는 구조를 가지게 된다.

그림 3에서 시프트 윈도우 셀프-어텐션과 시프트 윈도우 크로스-어텐션 연산은 [10]의 방식과 동일하게 윈도우를 x, y 방향으로 $\frac{M}{2}$ 만큼 이동시킨 이후 앞서 설명한 윈도우 셀프-어텐션과 윈도우 크로스-어텐션 연산과 동일한 방식으로 반복된다.

3. 다중 목적 학습 전략 (Multi-objective Training Strategy)

본 논문에서는 열화상 이미지 I_{thr}와 RGB 이미지 I_{rgb}에 대한 깊이 추정을 동시에 수행하는 다중 목적 훈련 전략을 사용한다. 두 모달리티의 결과값인 D_{thr}와 D_{rgb}에 동시에 손실 함수를 적용함으로써, DMCF 모듈이 두 모달리티의 학습 신호 모두로부터 학습이 진행된다. 또한, 특정 모달리티의 실제 깊이만을 사용하여 학습하는 경우에 인코더가 해당 모달리티로부터만 유의미한 특징을 추출하는 것에 편향(bias)될 수 있지만, 다중 목적 훈련을 적용하면 듀얼 모달리티 입력으로부터 의미 있는 특징 추출을 할 수 있다는 장점을 가지고 있다.

제안된 DMCF 모듈은 어텐션 기반 특징 향상 모듈이므로 모든 모달리티에서의 유의미한 특징 추출이 중요하다. 따라서, 본 논문에서는 DMCF 모듈의 상호 보완적 특징을 추출할 수 있는 잠재력을 최대한 활용하기 위해 다중 목표 학습을 적용한다.

4. 멀티 모달리티 학습을 위한 네트워크 구조

DMCF 모듈에서 특징 향상을 위해 사용한 어텐션 계산의 경우 쿼리(Q)와 키(K) 간의 상관관계를 통해 관련성 있는 정보를 강화할 수 있게 한다. 어텐션 계산에서 두 특징 간의 내적을 통해 상관성을 계산하기 때문에 쿼리와 키가 비슷한 특성을 가지고 있어야 유의미한 어텐션 값을 얻을 수 있다. 하지만, 크로스-어텐션을 계산할 때의 쿼리와 키는 서로 다른 모달리티의 특징으로부터 도출되기 때문에 어텐션 결과 값이 유의미한 상관 정보를 가지지 않을 수 있다. 이를 해결하기 위해 본 제안 방법에서는 그림 4에서 보듯이 공유 인코더를 사용하여 특징을 추출하였다. 공유 인코더를 사용하면 학습 과정이 안정화될 뿐만 아니라 성능 향상에도 기여한다. 또한, 공유 인코더를 통해 얻은 특징들을 DMCF 모듈에 입력하여 그 출력으로 얻은 특징들도 서로 유사한 특성을 가지게 되므로 공유 디코더를 사용하여 각 모달리티에서의 최종 깊이 예측을 하였다. 이러한 공유 인코더 및 디코더의 효과적 활용은 학습 과정에서 두 모달리티 정답으로부터의 학습을 가능하게 하여 성능 향상을 이끌어낸다. 인코딩 및 디코딩 단계에서 공유 요소를 포함하는 이 통합된 접근 방식은 더 나은 깊이 예측을 위해 모달리티 간 특징 유사성이 중요함을 강조한다.

5. 손실 함수

단안 깊이 추정은 3D 공간이 2D 공간으로 투영될 때 손실되는 3D 정보를 재구성하는 어려운 작업을 포함하며, 이는 불안정한 해를 구하는 문제(ill-posed problem)로 간주된다. 즉, 동일한 2D 공간으로 투영될 수 있는 무한히 많은 경우가 존재한다. 실제 깊이와 예측된 깊이 간의 스케일 차이에 의해 손실 함수가 영향을 받는 문제를 해결하고자

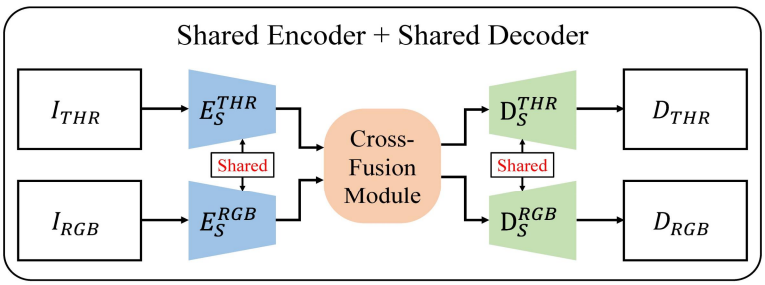

그림 4. 공유 인코더와 공유 디코더를 활용한 네트워크 구조
Fig. 4. Our proposed network architecture utilizing shared encoder and shared decoder

[12]에서 제안된 Scale- Invariant Logarithmic (SILog) 손실 함수를 활용한다. SILog 손실함수는 스케일 차이로 인해 손실 함수 값이 달라지기보다는, 스케일 불변 손실 함수를 통해 값 집합 간 상대적 깊이 값을 비교할 수 있도록 한다. SILog 손실 함수는 다음과 같다.

$$L = \sqrt{\frac{1}{K}\sum_i \Delta d_i^2 - \frac{\lambda}{K^2}(\sum_i \Delta d_i)^2} \quad (3)$$

식에서 Δd_i는 $\log d_i^{'} - \log d_i^{*}$를 나타내고$d_i^{'}$, d_i^{*}은 픽셀 i에서의 실제 깊이 값과 예측 깊이 값을 나타낸다. 또한, K는 유효한 깊이 값, 즉, LiDAR 센서를 통해 실제 깊이 값을 가지고 있는 픽셀 수를 나타낸다. 위 식에서 λ를 1로 설정했을 때 위의 손실함수가 완전히 스케일 불변해질 수 있지만 선행 연구[6,12]에서 λ를 1보다 작게 설정하는 것이 더 좋은 학습 결과를 보여 [6]을 따라 λ를 0.85로 설정하여 학습을 진행하였다.

Ⅳ. 실험 수행 및 결과 분석

1. 구현 세부정보

제안된 DMCF-SIDE 네트워크는 인코더와 디코더로 구성된다. 정성적 및 정량적 비교를 위해 ImageNet 데이터셋으로 사전 학습된 Swin Transformer[10]-Large 모델을 사용한다. 절제 연구(Ablation Study)에서는 학습 속도를 높이기 위해 동일하게 ImageNet으로 사전 학습된 Swin Transformer-Base 모델을 사용한다. 디코더는 Monodepth2 [13]에서 사용된 것과 동일하며, 인코더의 출력에서 스킵 연결을 사용한다. Swin Transformer는 원본 입력 이미지 해상도의 1/4, 1/8, 1/16, 1/32 스케일에서 특징을 출력한다. 이러한 출력의 채널 차원은 Base 모델의 경우 128, 256, 512, 1024이고, Large 모델의 경우 192, 384, 768, 1536이다. DMCF 모듈에서, 컨볼루션 블록은 채널 수를 절반으로 줄이며, 그 결과 Base 모델의 경우 64, 128, 256, 512, Large 모델의 경우 96, 192, 384, 768의 채널 수를 갖는다. 교차 관계 블록에서는, 1/4 및 1/8 스케일의 특징에 대해 4×4 윈도우, 1/16 및 1/32 스케일의 특징에 대해 2×2 윈도우로 특징을 분할한다. 초기 학습률과 최종 학습률은 각각 1×10^{-4}, 1×10^{-5}로 설정한다. Adam 옵티마이저를 사용하며, β_1과 β_2 매개변수는 각각 0.9와 0.999로 설정된다. 학습에 사용된 깊이 값은 0미터에서 80미터까지이며, 평가 시에는 1×10^{-3} 미터에서 80미터까지의 깊이 값을 사용한다.

2. 데이터셋

MS2 데이터셋[9]을 학습 및 테스트에 사용한다. 해당 데이터셋은 다양한 환경 조건(예: 주간, 야간, 비)에서 RGB, 열화상, NIR 스테레오 이미지를 제공하는 실외 자율주행 데이터셋이다. 또한 LiDAR를 통해 얻은 실제 깊이(Ground Truth Depth)와 GPS 센서를 통해 측정된 카메라 위치 정보도 제공된다. 본 연구에서는 자동차 와이퍼로 인해 비가 오는 장면에서 RGB 이미지가 가려지는 것을 배제하기 위해, 학습 및 테스트 모두 맑은 날씨의 주간과 야간에 촬영된 이미지만 사용하였다. 이전 연구에서 학습, 검증, 테스트 목록이 사전에 존재하지 않았기 때문에 새로운 목록을 생성하였다. GPS 센서 데이터를 기반으로 정지된 장면을 데이터셋에서 필터링하여 특정 이미지가 학습에 과도한 영향을 미치는 것을 방지하였다. 데이터 분할은 44,352개의 학습 이미지, 2,240개의 검증 이미지, 그리고 1,000개의 테스트 이미지로 구성된다. 데이터 분할은 주간과 야간 이미지를 동일한 수로 유지한다. 이후의 모든 실험 결과는 이 데이터 분할을 사용하여 학습된 모델을 기반으로 한다.

3. 실험 결과 분석

표 1 및 2는 MS2 테스트셋에서 깊이 추정 네트워크의 성능을 열화상 및 RGB 이미지 각각에 정렬된 깊이에 대해 비교한 내용을 제시한다. 이전의 깊이 추정 연구들의 평가 방식에 따라, 4가지 오류 지표와 3가지 정확도 지표를 사용하여 결과를 평가한다. 오류 지표가 낮거나 정확도 지표가 높을수록 더 나은 결과를 나타낸다. NewCRF1[6]은 RGB 또는 열화상 단일 모달리티 입력만을 사용하여 학습된 모델을 나타낸다. NewCRF2는 공정한 정보 사용을 보장하기 위해 학

표 1. MS2 테스트셋의 열화상 이미지에 정렬된 깊이 추정 결과의 정량적 비교. RGB, THR: 입력 모달리티를 나타냄. Stereo: 모델이 스테레오 입력을 사용했는지 여부. 각 블록에서 최고의 성능은 굵게 표시됨

Table 1. Quantitative Comparison of depth estimation results aligned to thermal images of MS2 test split. RGB, THR: represent modality of input. Stereo: Whether the model utilized stereo inputs. The best performance in each block is highlighted in bold

Methods	RGB	THR	Stereo	TestSet	Error↓				Accuracy↑		
					abs_rel	sq_rel	rms	log_rms	d1	d2	d3
NewCRF[1] [6]		✓		AVG	0.0773	0.306	2.7306	0.1056	0.9346	0.992	0.9991
				Day	0.0707	0.2827	2.6869	0.0998	0.9422	0.9931	0.9991
				Night	0.0838	0.3293	2.7743	0.1115	0.9271	0.9908	0.9992
NewCRF[2] [6]	✓	✓		AVG	0.0763	0.2976	2.687	0.104	0.9357	0.9928	0.9991
				Day	0.0666	0.2575	2.5773	0.0947	0.9476	0.9943	0.9991
				Night	0.0861	0.3376	2.7966	0.1132	0.9239	0.9913	0.9992
Shin *et al.* [9]		✓		AVG	0.0751	0.3013	2.7957	0.1051	0.9362	0.9922	0.9993
				Day	0.0662	0.2652	2.6999	0.0962	0.9465	0.994	0.9995
				Night	0.0841	0.3375	2.8914	0.114	0.9259	0.9905	0.9991
Shin *et al.* [9]		✓	✓	AVG	0.073	0.2801	2.6985	0.102	0.9403	0.9932	0.9994
				Day	0.0644	0.2476	2.6142	0.0934	0.9502	0.9948	0.9996
				Night	0.0815	0.3126	2.7828	0.1106	0.9305	0.9917	0.9993
Ours	✓	✓		AVG	**0.0682**	**0.2406**	**2.4627**	**0.0946**	**0.9497**	**0.9948**	**0.9997**
				Day	**0.0596**	**0.2061**	**2.3275**	**0.0852**	**0.9609**	**0.9959**	**0.9996**
				Night	**0.0769**	**0.2751**	**2.5979**	**0.104**	**0.9385**	**0.9938**	**0.9997**

표 2. MS2 테스트셋의 RGB 영상에 정렬된 깊이 추정 결과의 정량적 비교. RGB, THR: 입력 모달리티를 나타냄. 각 블록에서 최고의 성능은 굵게 표시됨

Table 2. Quantitative Comparison of depth estimation results aligned to RGB images of MS2 test split. RGB, THR: represent modality of input. The best performance in each block is highlighted in bold

Methods	RGB	THR	TestSet	Error↓				Accuracy↑		
				abs_rel	sq_rel	rms	log_rms	d1	d2	d3
NewCRF[1] [6]	✓		AVG	0.085	0.3394	2.9812	0.1137	0.9263	0.9924	0.9993
			Day	0.0758	0.3022	2.8833	0.1042	0.9428	0.9933	0.9993
			Night	0.0941	0.3767	3.0791	0.1232	0.9098	0.9914	**0.9992**
NewCRF[2] [6]	✓	✓	AVG	0.0862	0.3885	3.1013	0.1155	0.9206	0.9898	0.9989
			Day	0.0748	0.33	2.9252	0.1032	0.9403	0.9921	0.9989
			Night	0.0976	0.447	3.2774	0.1278	0.9008	0.9876	0.9989
Ours	✓	✓	AVG	**0.0814**	**0.3299**	**2.886**	**0.1097**	**0.9305**	**0.9929**	**0.9993**
			Day	**0.0727**	**0.2903**	**2.7839**	**0.1008**	**0.9432**	**0.9942**	**0.9994**
			Night	**0.0901**	**0.3696**	**2.9881**	**0.1186**	**0.9178**	**0.9915**	0.9991

습과 테스트 모두에서 열화상과 RGB 이미지를 결합하여 입력으로 사용한다.

3.1 열화상 이미지와 정렬된 깊이 추정 결과:

Shin et al.[9]의 방법은 단안 및 스테레오 설정 모두에서 작동하도록 설계되고 학습되었기 때문에, 제안 방법을 이 두 가지 시나리오 모두와 비교하였다. 표 1에서 볼 수 있듯이, 제안된 네트워크는 모든 지표에서 다른 모든 방법의 성능을 능가한다. 스테레오 입력을 사용하는 Shin et al.[9]의 방법이 두 번째로 좋은 성능을 보였으며, 이에 비해 제안 방법(DMCF-SIDE)은0 'abs_rel' 지표에서 평균 6.6%의 향상을 달성한다. 특히, 낮과 밤 테스트셋에서 각각 7.5%와

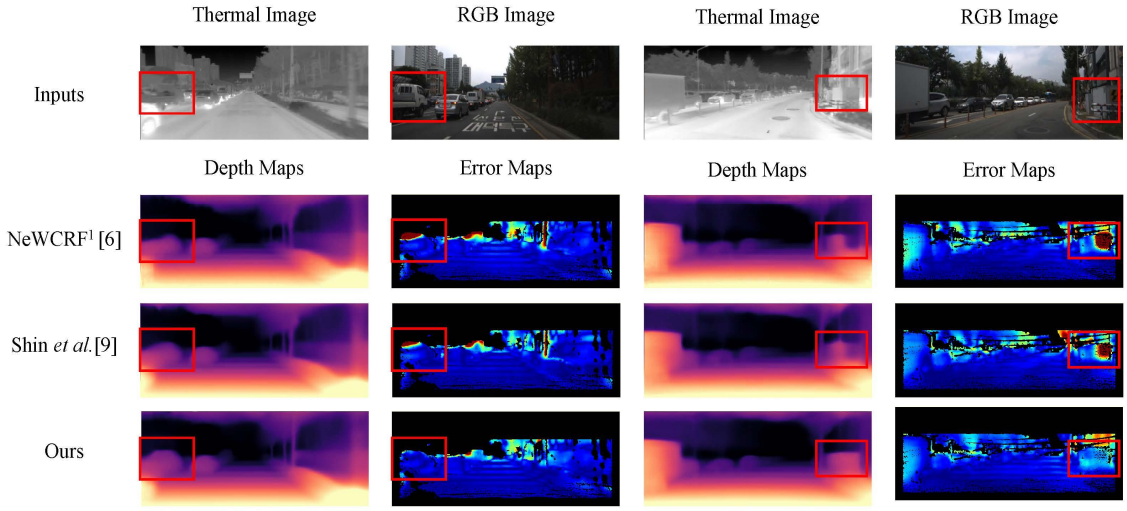

그림 5. MS2 테스트셋의 열화상 이미지에 정렬된 깊이 추정 결과의 정성적 비교. 에러맵에서 파란색은 낮은 값을, 빨간색은 큰 값을 나타냄
Fig. 5. Qualitative comparison of depth estimation results aligned to thermal images of MS2 test split. In error map, blue represents low value and red represents large value

5.6%의 향상을 확인하였다. 더 나은 조명 조건으로 인해 RGB 이미지가 일반적으로 더 많은 유의미한 정보를 포함하는 낮 테스트셋에서 더 큰 성능 향상이 관찰된 것은, 제안 모델이 모달리티 간 상호 보완성을 효과적으로 활용했음을 의미한다.

그림 5는 제안된 모델을 NewCRF1[6] 및 Shin et al.[9](스테레오 설정)와 비교한 정성적 결과를 보여준다. 빨간색 상자로 강조된 영역에서 볼 수 있듯이, 다중 모달리티 입력을 활용한 제안 모델은 객체 구조를 더 잘 보존하고, 객체 내의 일관된 깊이 예측을 보여준다. 이는 제안 모델이 깊이 추정 정확도를 향상시키기 위해 다중 모달리티 정보를 활용하는 것의 효과를 잘 보여준다.

3.2 RGB 이미지와 정렬된 깊이 추정 결과:

RGB 이미지와 정렬된 결과는 열화상 이미지와 유사하게, 제안된 모델이 대부분의 지표에서 최고의 성능을 보인다. 그림 6은 제안 모델을 NewCRF1[6]과 비교한 정성적 결과를 보여준다. 빨간색 상자로 강조된 영역에서 볼 수 있듯이, 열화상 이미지에서 얻은 추가 정보가 네트워크가 더 정확한 깊이를 예측하는 데 도움을 준다. 그림 6에서 RGB

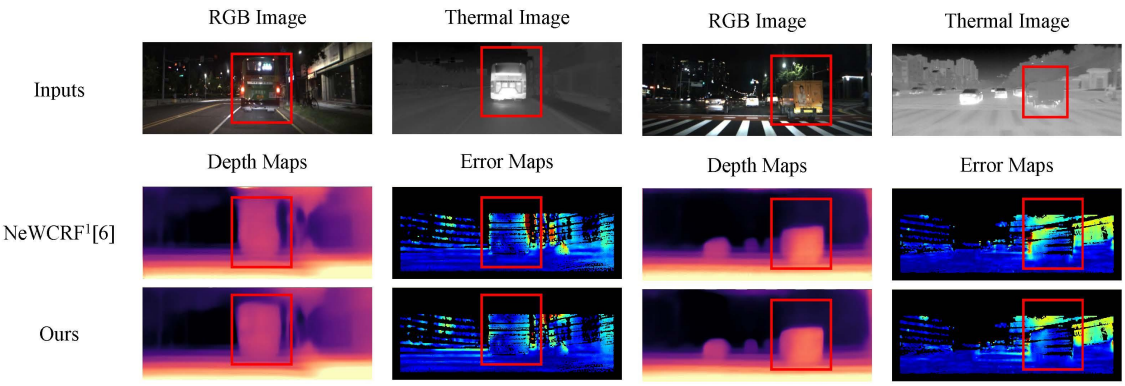

그림 6. MS2 테스트셋의 RGB 영상에 정렬된 깊이 추정 결과의 정성적 비교. 에러맵에서 파란색은 낮은 값을, 빨간색은 큰 값을 나타냄
Fig. 6. Qualitative comparison of depth estimation results aligned to RGB images of MS2 test split. In error map, blue represents low value and red represents large value

입력만을 사용하는 NewCRF1이 객체의 가장자리를 정확하게 감지하는 데 어려움을 겪는 것을 확인할 수 있다.

Ⅴ. 절제 연구 (Ablation Study)

1. 네트워크 구조 절제 연구

이 절제 연구는 공유 인코더와 디코더 구조의 효과를 확인하기 위해 설계되었다. 이를 위해, 제안된 모델 구조의 네 가지 변형과 비교하였다: (i) 두 개의 비공유 인코더와 하나의 공유 디코더(UE+SD), (ii) 두 개의 비공유 인코더와 두 개의 비공유 디코더(UE+UD), (iii) 하나의 공유 인코더와 두 개의 비공유 디코더(SE+UD), (iv) 하나의 공유 인코더와 하나의 공유 디코더(SE+SD, 제안된 모델).

표 3에서 볼 수 있듯이, RGB에 대해 평가할 때, 공유 인코더를 사용하는 경우, (a)를 기준으로 (d)에서 5%의 개선이 있었으며, (b)를 기준으로 (c)에서는 3.7%의 개선이 있었다. 디코더가 공유되거나 비공유되는 조건을 비교했을 때 차이는 1.1% 이내였다. 이러한 결과는 공유 인코더를

표 3. 네트워크 구조 변형에 따른 깊이 추정 결과의 정량적 비교, 'Modal'은 깊이 추정 결과가 어느 모달리티에 정렬되어 있는가를 나타낸 것임, 최고의 성능은 굵게 표시됨
Table 3. Quantitative comparison of depth estimation results between network architecture variants, 'Modal' means where the depth estimation results are aligned to, the best performance in each blockis highlighted in bold

Modal	Methods	Error ↓				Accuracy ↑		
		abs_rel	sq_rel	rms	log_rms	d1	d2	d3
RGB	(a) UE+SD	0.0907	0.4134	3.1864	0.1196	0.9134	0.9902	0.9991
	(b) UE+UD	0.0903	0.4062	3.1109	0.1188	0.9126	0.9897	0.9991
	(c) SE+UD	0.0870	0.3816	3.0533	0.1152	0.9206	0.9907	**0.9992**
	(d) SE+SD	**0.0860**	**0.3536**	**2.9824**	**0.1146**	**0.9248**	**0.9919**	0.9991
THR	(a) UE+SD	0.0710	0.2606	2.5503	0.0976	0.9454	0.9948	**0.9995**
	(b) UE+UD	0.0710	0.2481	2.501	0.0968	0.9487	0.9951	**0.9995**
	(c) SE+UD	0.0699	0.2426	2.4767	0.0955	0.9498	0.9951	0.9994
	(d) SE+SD	**0.0687**	**0.2332**	**2.4481**	**0.0949**	**0.9515**	**0.9953**	**0.9995**

표 4. DMCF 모델에서의 'concatenation' 포함 여부에 따른 깊이 추정 결과의 정량적 비교. 최고의 성능은 굵게 표시됨
Table 4. Quantitative comparison of depth estimation between proposed model and proposed model without 'concatenation' in DMCF Module.The best performance in each block is highlighted in bold

Modal	Concat	TestSet	Error ↓				Accuracy ↑		
			abs_rel	sq_rel	rms	log_rms	d1	d2	d3
RGB		AVG	0.0876	0.3848	3.0708	0.1162	0.9192	0.9902	**0.9991**
		Day	0.0765	0.3162	2.881	0.1049	**0.9372**	**0.9929**	**0.9993**
		Night	0.0987	0.4535	3.2605	0.1276	0.9012	0.9875	**0.9989**
	✓	AVG	**0.0858**	**0.377**	**3.0145**	**0.1138**	**0.921**	**0.9912**	**0.9991**
		Day	**0.0764**	**0.3271**	**2.8549**	**0.1036**	0.9355	0.9927	**0.9993**
		Night	**0.0952**	**0.4269**	**3.1741**	**0.124**	**0.9065**	**0.9897**	**0.9989**
THR		AVG	0.0695	**0.251**	2.4901	0.0962	**0.9486**	0.9941	0.9995
		Day	**0.0594**	**0.2035**	2.3211	0.0856	**0.9622**	**0.9958**	0.9993
		Night	0.0796	0.2985	2.6591	0.1068	0.9349	0.9924	0.9996
	✓	AVG	**0.0693**	0.2529	**2.4669**	**0.0951**	0.9479	**0.9945**	**0.9996**
		Day	0.0599	0.2145	**2.3104**	**0.0852**	0.9603	0.9955	**0.9994**
		Night	**0.0787**	**0.2912**	**2.6234**	**0.1051**	**0.9356**	**0.9935**	**0.9998**

사용하여 다른 모달리티에서 유사한 특징을 추출하는 것이 중요함을 시사한다. 또한, 비공유 인코더와 함께 공유 디코더를 사용하는 경우 성능이 저하되었는데, 이는 서로 다른 특징들이 동일한 디코더를 통해 깊이로 매핑되어야 하기 때문인 것으로 추측하였다. 결과적으로 공유 인코더와 공유 디코더를 함께 사용하는 것이 가장 효과적임을 확인하였다.

2. DMCF 모듈 입력 구조 절제 연구

이 절제 연구는 DMCF 모듈에서의 다운샘플된 입력 이미지와 특징의 concatenation 후에 convolution block을 통과시키는 것의 효과를 확인하기 위해 설계되었다. ‘concat’에 체크가 없는 모델은 제안 모델에서 다운샘플된 입력 이미지의 concatenation만 제외된 것이고 표 4에서 볼 수 있듯이 concatenation을 포함한 제안 모델이 대부분의 성능에서 가장 좋은 것을 확인할 수 있었다.

Ⅵ. 결 론

본 논문에서는 정렬되지 않은 열화상 및 RGB 이미지를 활용한 깊이 추정 연구를 진행하였다. 제안된 교차 융합 모듈은 두 모달리티의 특징을 효율적으로 강화하였고 다중 목적 훈련을 통해 각 모달리티의 특징이 유의미한 특성을 가지도록 유도하였다. 또한, 공유 인코더와 디코더 사용을 통해 특징 간의 상관성을 높여 성능 향상을 이끌어냈다. 그 결과, 제안 방법은 기존의 깊이 추정 방법들보다 더 높은 성능을 달성하였다.

참 고 문 헌 (References)

[1] Byeongjun Kwon, 2024, “Multi-modal Depth Estimation from Misaligned Thermal and RGB Images”, Korea Advanced Institute of Science and Technology (KAIST) Master’s Thesis. https://library.kaist.ac.kr/search/ctlgSearch/posesn/view.do?bibctrlno=1097161&se=t0&ty=B&_csrf=902b96d7-09f5-46a7-9bb2-7fda213b426c

[2] Byeongjun Kwon. “Multi-modal Depth Estimation from Misaligned Thermal and RGB Images”, Korean Institute of Broadcast and Media Engineers Summer Conference, 2024. https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11849202

[3] Eigen, David, and Rob Fergus. “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture.” Proceedings of the IEEE international conference on computer vision. pp. 2650-2658. 2015.
doi: https://doi.org/10.1109/iccv.2015.304

[4] Fu, Huan, et al. “Deep ordinal regression network for monocular depth estimation.” Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2002-2011. 2018.
doi: https://doi.org/10.1109/cvpr.2018.00214

[5] Ranftl, René, Alexey Bochkovskiy, and Vladlen Koltun. “Vision transformers for dense prediction.” Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179-12188. 2021.
doi: https://doi.org/10.1109/iccv48922.2021.01196

[6] Yuan, Weihao, et al. “Neural window fully-connected crfs for monocular depth estimation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3916-3925. 2022.
doi: https://doi.org/10.1109/cvpr52688.2022.00389

[7] Kim, Namil, et al. “Multispectral transfer network: Unsupervised depth estimation for all-day vision.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.
doi: https://doi.org/10.1609/aaai.v32i1.12297

[8] Lu, Yawen, and Guoyu Lu. “An alternative of lidar in nighttime: Unsupervised depth estimation based on single thermal image.” Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3833-3843. 2021.
doi: https://doi.org/10.1109/wacv48630.2021.00388

[9] Shin, Ukcheol, Jinsun Park, and In So Kweon. “Deep depth estimation from thermal image.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1043-1053. 2023.
doi: https://doi.org/10.1109/cvpr52729.2023.00107

[10] Liu, Ze, et al. “Swin transformer: Hierarchical vision transformer using shifted windows.” Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012-10022. 2021.
doi: https://doi.org/10.1109/iccv48922.2021.00986

[11] Xu, Haofei, et al. “Gmflow: Learning optical flow via global matching.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8121-8130. 2022.
doi: https://doi.org/10.1109/cvpr52688.2022.00795

[12] Eigen, David, Christian Puhrsch, and Rob Fergus. “Depth map prediction from a single image using a multi-scale deep network.” Advances in neural information processing systems 27 (2014). https://proceedings.neurips.cc/paper/2014/hash/7bccfde7714a1ebadf06c5f4cea752c1-Abstract.html

[13] Godard, Clément, et al. “Digging into self-supervised monocular depth estimation.” Proceedings of the IEEE/CVF international conference on computer vision. pp. 3828-3838. 2019.
doi: https://doi.org/10.1109/iccv.2019.00393

저 자 소 개

권 병 준

- 2016년 : 연세대학교 전기전자공학부
- 2022년 : 한국과학기술원(KAIST) 전기전자공학부 석사
- 2024년 : 한국과학기술원(KAIST) 전기전자공학부 박사과정
- ORCID : https://orcid.org/0009-0001-5538-2654
- 주관심분야 : 깊이 추정, 이미지신호처리, 딥러닝

김 문 철

- 1996년 8월 : University of Florida Electrical and Computer Engineering 박사
- 2001년 2월 ~ 2009년 3월 : 한국정보통신대학교 공학부 조교수/부교수
- 2009년 3월 ~ 현재 : 한국과학기술원 전기및전자공학부 부교수/정교수
- ORCID : https://orcid.org/0000-0003-0146-5419
- 주관심분야 : 컴퓨터 비전, 딥러닝, 영상 처리, 비디오 코딩