Journal of The Institute of Electronics and Information Engineers Vol.61, NO.3, March 2024 https://doi.org/10.5573/ieie.2024.61.3.68
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2024-61-3-7

단안 깊이 추정을 위한 소실점 위치 정보를 사용하는 향상된 SW-MSA

(An Improved SW-MSA using Vanishing Point Position Information for Monocular Depth Estimation)

조 용 석*, 김 나 라*, 박 호 현**

(Yongseok Jo, Nara Kim, and Ho-Hyun Park©)

요 약

본 논문은 단안 렌즈를 통한 깊이 추정에서 소실점 탐지와 향상된 SW-MSA를 사용한 Swin Transformer 기반의 깊이 추정 모델을 제안한다. 이 모델은 이미지가 입력되면 소실점을 탐색한 후 소실점의 위치에 따른 유형을 파악하여 깊이 추정 모델에 도움이 될 정보를 모델에 전달한다. 소실점 위치 유추는 먼저 이미지에서 캐니 선분 검출기로 외곽선을 추출하여 허프 변환을 통하여 직선 성분만 남기고, 그 직선들을 연장해서 가장 많은 선분의 교점 영역을 소실점으로 설정한다. 소실점의 위치 유형은 3가지로 분류되는데, 유형에 따라 SW-MSA의 셀프 어텐션 방식이 나뉜다. 제안한 모델 성능은 실험 결과를 통하여 기존 단안 깊이 추정 모델과 더 나은 결과를 나타낸다. 본 논문은 소실점이라는 기하학적 특성을 통하여 단안 깊이 추정을 함으로써 훈련 데이터에 의존하지 않고 이미지의 근원적인 특성을 찾아내는 기술을 사용하는 것을 강조한다. 본 논문은 깊이 추정 분야에 중요한 기여를 하고 있으며, 소실점이라는 원근법에서 사용하는 개념을 깊이 추정 분야에서 이용하기 때문에 기술의 잠재력이 크다는 것을 강조한다.

Abstract

This paper proposes a Swin Transformer-based depth estimation model using vanishing point detection and improved SW-MSA in depth estimation through a monocular lens. This model, upon receiving a image, searches for the vanishing point, then identifies the type based on the location of the vanishing point, and conveys information helpful to the depth estimation model. Inference of the vanishing point position first involves extracting the outlines from a image using the Canny edge detector, then retaining only the line components through Hough transformation. These lines are then extended to determine the vanishing point as the area where the most line intersections occur. The types of vanishing point positions are classified into three categories, and the self-attention mechanism of SW-MSA varies according to the type. The performance of the proposed model demonstrates better results than the existing monocular depth estimation models, as shown by experimental result. This paper emphasizes the use of technology that identifies the intrinsic characteristics of an image by estimatioing monocular depth through the geometric feature of vanishing points, thereby not relying on training data. This paper makes a significant contribution to the field of depth estimation, emphasizing the potential of the technology by utilizing the concept of vanishing points from perspective in depth estimation.

Keywords : Deep learning, Depth estimation, Monocular depth estimation, Swin transformer, Vanishing point

*학생회원, **정회원, 중앙대학교 전자전기공학과(Department of Electrical and Electronics Engineering, Chung-Ang University)
© Corresponding Author(E-mail : hohyun@cau.ac.kr)
Received ; December 12, 2023 Revised ; December 14, 2023 Accepted ; December 21, 2023

Copyright © The Institute of Electronics and Information Engineers.

This is an Open-Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-commercial use, distribution and reproduction in any medium, provided the original work is properly cited.

Ⅰ. 서 론

깊이 정보는 이미지의 근본적 정보를 이해하는 데 중요한 구성 요소이며 객체 감지[1], 이미지 분할(image segmentation)[2, 3] 영상의 왜곡 정보 제거[4], 자율 주행 분야[5, 6] 등 여러 분야에 적용할 수 있는 정보이다. 최근 깊이 추정 모델은 딥러닝의 발달로 인하여 성능이 상당히 향상되었다. 딥러닝은 주석 처리(annotation)된 대형 데이터 세트(dataset)만 있다면 높은 성능을 기록할 수 있기 때문이다. 최근에는 자연어 처리(natural language processing)[7]에서 사용하던 트랜스포머 기법(Transformer)을 시각 분야에 적용함으로써[8] 이미지 분할, 깊이 추정에서 비약적인 성능의 향상을 이루었다. 이러한 뛰어난 성능을 가진 딥러닝 모델은 시각(vision) 분야뿐만 아니라 다양한 응용 분야에 활용된다.

그러나 딥러닝은 훈련 데이터(training data)에 따라 성능이 정해지는 문제가 있다. 데이터의 수집 방향에 따라 다른 데이터 세트에 적용할 때는 모델의 성능이 낮아질 수도 있다. 따라서 본 논문은 깊이 추정 시 훈련 데이터에 의존적인 약점을 보완하고, 깊이 추정 정확도를 높이기 위하여 기하학적인 특성인 소실점(vanishing point)을 이용한다. 소실점을 이용한다면 깊이 추정 훈련 시 고정적으로 얻을 수 있는 깊이 관련 정보가 부가적으로 발생하여 이 정보를 추정에 이용할 수 있다. 소실점은 실제 세계에서 다루는 대부분의 이미지에서 발견할 수 있는 요소이기 때문에 실제 세계를 다루는 훈련 데이터 세트에서는 데이터 유형에 상관없이 광범위한 응용이 가능하다.

훈련 데이터에 의존적인 딥러닝 모델의 약점을 보완하고 컴퓨팅 자원의 감소를 위해서 자기 지도 학습(self-supervised learning)의 연구도 활발하다[9]. 자기 지도 학습이란 라벨링(labeling)을 하지 않은 데이터(untagged data)를 기반으로 한 학습이며 신경망 스스로 학습 데이터에 대한 분류를 수행한다. 본 논문은 소실점 탐지를 통한 깊이 추정으로 일반적인 깊이 추정보다 필요 자원의 감소를 이루어내고자 하나, 자기 지도 학습과는 다른 점이 존재한다. 학습을 통한 특성 찾기가 아닌 모델 내부에 인간의 추상적 개념인 소실점의 개념을 탐지할 수 있도록 하는 것이기 때문에 자기 지도 학습과는 다른 작용을 한다고 할 수 있다. 인간이 원근법을 느끼는 원리에 기인하여 이미지 어디서나 찾을 수 있는 특성을 딥러닝 모델이 감지할 수 있도록 하는 것이다.

최근 깊이 추정 분야에서는 Lidar(light detection and ranging) 센서를 활용한 연구도 활발히 이루어졌다[10, 11]. 그러나 Lidar 센서는 얻을 수 있는 정보가 다양하고 정밀하여 활용도나 정확도는 뛰어나지만, 센서가 고비용이기 때문에 스마트폰 같은 개인 장비나 저가의 차량에 보급하기 힘들다는 단점이 있다. 본 연구에서는 이러한 Lidar의 단점을 보완하기 위하여 단안 렌즈를 통한 깊이 추정을 하기로 한 것이다.

본 논문에서 기여하는 점은 다음과 같다. 첫 번째로, 소실점을 활용하여 깊이 추정 모델의 성능을 향상할 수 있다. 두 번째로, 소실점은 모든 이미지 데이터에 존재하기 때문에 데이터 세트에 따라 성능이 큰 편차를 보이지 않는다. 세 번째로 단안 렌즈를 통한 깊이 추정을 함으로써 휴대용 스마트폰을 비롯한 휴대용 개인 장비로 깊이 추정 응용 프로그램을 보급할 수 있도록 한다.

Ⅱ. 관련 연구 및 배경지식

1. 활용 기술

가. 깊이 추정

인간의 눈은 양쪽 눈의 위치 차이에 따라 객체의 깊이를 판단한다. 두 눈은 어느 정도의 거리로 떨어져 있기 때문에, 각각의 눈에서 본 물체의 위치가 약간 다르다. 이 차이로 인해 뇌는 물체의 거리와 깊이를 정확하게 계산할 수 있다. 눈을 렌즈(lens)에 대입한다면 렌즈 사이의 거리를 시차(disparity)라고 말하는데, 시차 정보를 기반으로 렌즈로부터 사물까지의 거리를 추정하는 과정을 다안 깊이 추정(stereo depth estimation)이라고 한다. 렌즈 하나만으로 깊이 추정하는 과정은 단안 깊이 추정(monocular depth estimation)이라고 한다. 렌즈가 하나라면 시차가 존재하지 않게 되는데, 딥러닝 모델로 가상의 시차를 만들어 내어 렌즈로부터 객체까지의 거리를 추출하는 방법으로 단안 깊이 추정의 연구가 본격적으로 이루어졌다[12]. 이후 인코더-디코더(encoder-decoder) 구조를 모델에 추가하여 단안 깊이 추정에서의 성능 향상을 이루었다. 최근에는 자연어 처리 분야에 성공적으로 적용된 어텐션 방식(attention mechanism)을 사용하는 트랜스포머(Transformer)를 단안 깊이 추정 작업에 적용하여 뛰어난 성능을 이룬 모델이 많다. 본 논문에서는 트랜스포머 기반의 단안 깊이 추정 모델을 사용한다.

나. 소실점

소실점(vanishing point)은 물체의 연장선을 그었을 때 선과 선이 만나는 점으로 회화와 건축 분야의 원근법에서 사용하는 개념이다. 유클리드 좌표계의 세계에서 평행선은 만나지 않는다. 하지만 실제 세계에서[13] 평행선이 존재하는 부분을 그린 투시도나 실제 세계를 찍은 사진에서 물체 외곽선의 연장선은 한 점에서 만나게 되는데, 이 점을 소실점이라 한다. 소실점은 가까이 있는 것은 크게 보이고 멀리 있는 것은 작게 보이는 원근 원리에 따라 소실점의 실제 위치는 가장 먼 곳에 존재한다. 일반적으로 사용되는 훈련 데이터 세트는 실제 세계 기반의 이미지이기 때문에 모든 이미지에서 발견할 수 있는 특성이고, 이 때문에 데이터의 종류가 달라지더라도 뛰어난 성능을 보일 수 있다. 소실점은 이미지 데이터의 종류에 따라 한 개 또는 복수의 소실점이 존재한다. 또, 소실점은 이미지 내부에 있을 수도 있지만, 이미지 외부에서 발견될 수도 있다.

다. Swin Transform

기존의 컴퓨터 시각 인공지능 모델(computer vision model)들은 합성곱 신경망(convolution neural network)이 주를 이루었다. 합성곱 신경망에서 객체 탐지 분야는 뛰어난 성능을 달성했지만, 의미 구분 분야(semantic segmentation), 깊이 추정 분야에서는 한계에 봉착하였다. 그 이유는 합성곱 신경망 구조 특성상 주변 픽셀 간의 정보 추출만 가능하기 때문에 멀리 떨어진 픽셀과 관계를 따지기가 어렵다. 따라서 전체적인 문맥을 이해하는 자연어 처리(natural language processing) 기법 중, 어텐션 기법(attention)을 사용한 트랜스포머(Transformer) 모델을 컴퓨터 시각 분야로 적용하는 시도들이 있었다. 트랜스포머 모델을 사용하면 한 픽셀을 단어라고 가정하여 픽셀 간의 관계성까지 고려한 훈련이 가능해진다. 그러나 고해상도의 이미지(image)는 픽셀(pixel) 수가 늘어나기 때문에 모든 조각(patch) 조합에 대해 셀프-어텐션 기법(self-attention)을 사용하기 힘들다. 이러한 문제점을 해결하기 위해 개발된 모델이 Swin Transformer[14]이다. Swin Transformer 이전의 시계열 트랜스포머 모델은 입력 이미지를 작은 조각들로 쪼개며 분석한다면, Swin Transformer는 입력 이미지를 작은 단위의 조각으로 나누고 이를 합쳐 나가며 분석하는 방법을 사용해 고해상도의 이미지 분석에서도 좋은 성능을 낼 수 있었다.

2. 관련 연구 동향

소실점을 추출하거나 소실점 정보를 깊이 추정에 이용하고자 하는 시도들이 있었다. Borji는 소실점을 합성곱 신경망을 통하여 학습할 수 있고, 이를 지역화(localize) 가능함을 보였다[15]. Borji의 소실점 탐지 방식은 훈련 데이터에 소실점 위치를 일일이 주석 처리(annotation)를 하여 유사성 있는 지점을 검출하는 고전적인 방식이었다. 소실점 특성을 활용하여 악천후 같은 저해상도 조건에서 소실점에 의해 차선 및 도로를 효과적으로 감지하는 네트워크의 제안도 있었다[5]. 이 연구에서의 소실점 탐지 방식은 전체 이미지를 사분면으로 나누는 마스크를 적용하여 사분면의 교차점을 소실점으로 추론하는 방식을 사용하였다. Zhou는 소실점을 검출하기 위하여 평면으로 표현된 이미지를 가우시안 구(Gaussian sphere)에 좌표계 이동(mapping)을 한 후 선분(edge)의 교점을 찾아 소실점을 탐지하는 코닉 연산(conic convolution)을 제안하였다[16].

위의 소실점 기반의 연구들은 소실점 특성의 이해 없이 주석 처리된 데이터 세트에 의존한 소실점 탐지이거나 다른 좌표계로 이동 혹은 샘플링(sampling)을 통한 부가 과정이 필요한 탐지이다. 또한, 소실점이 한 개라는 가정에 근거한다. 본 논문에서 제시하는 모델은 2차원 이미지의 평면 좌표계를 활용한 직관적인 소실점 탐지를 통하여 소실점 탐지 과정을 간단히 하였고, 이미지 외부에 있는 소실점도 이미지 분석에 이용하여 효과적인 깊이 추정을 수행할 수 있도록 설계하였다.

Ⅲ. 제안하는 모델

1. 소실점 검출 모듈

소실점을 검출하기 위해서는 먼저 물체의 외곽선을 검출할 필요가 있다. 이는 허프 변환(Hough transform)과 캐니 선분 검출기 알고리즘(Canny edge detector algorithm)을 통하여 검출할 수 있다. 캐니 선분 검출기 알고리즘은 다음과 같은 순서로 진행된다. 먼저 가우시안 필터링(Gaussian filtering)을 통하여 이미지에 존재하는 잡음(noise)을 제거한다. 잡음이 제거된 이미지에 x축, y축으로 소벨 마스크(sobel mask)를 적용하여 필터링 후 기울기(gradient)를 추출하여 선분을 검출한다. 이때, 실제 외곽선보다 많은 선분이 검출되는데, 비최대억제(non-maximum suppression)를 통하여 기울기가 가장 큰 픽셀만을 실제 외곽선으로 추정하여 선택한다.

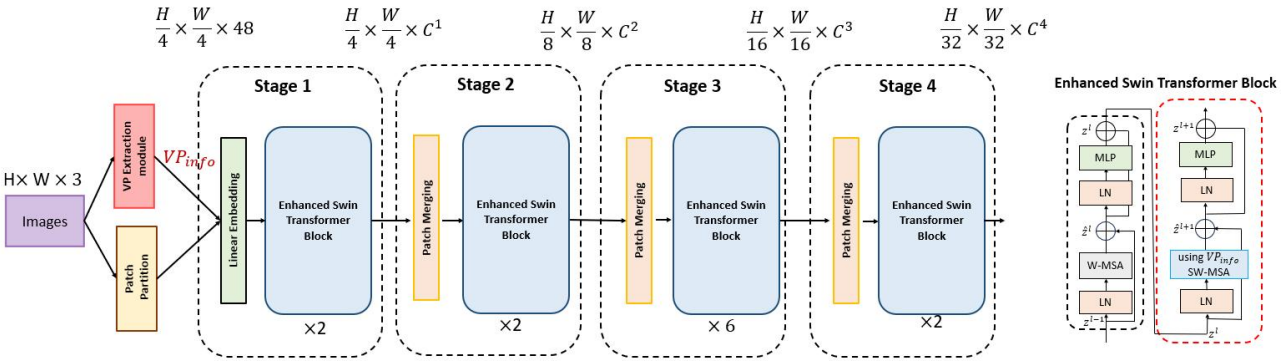

그림 1. 소실점 유추 정보를 사용한 향상된 SW-MSA와 향상된 Swin Transformer Block 구조
Fig. 1. Imporved SW-MSA and improved Swin Transformer Block structure using vanishing point inference information.

허프 변환은 2차원 좌표계의 어떤 점이라도 선 집합의 일부일 수 있다는 가정을 바탕으로 직선을 검출하는 변환법이다. 캐니 선분 검출 알고리즘으로 검출한 선분들에 허프 변환을 적용하여 직선 성분만을 검출한다. 본 논문에서는 이렇게 검출된 직선을 연장하여 직선들의 교점 중 많은 직선이 교차하는 지점을 소실점(영역)으로 유추한다. 이때, 소실점은 이미지 외부에도 있을 수 있다. 이는 이미지 외부에 보강(padding)을 통하여 직선들의 연장선이 이미지 외부에도 존재할 수 있도록 하고, 이미지 외부에서 교점이 발생한다면 해당 교점들의 영역을 소실점으로 인식한다. 인식된 소실점은 다수가 생길 수도 있다.

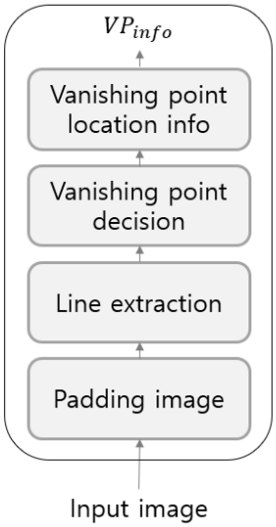

그림 2. 소실점 검출 모듈의 전체 과정
Fig. 2. Process of vanishing point detection module.

검출된 소실점의 위치에 따라 유형을 3가지로 나누고 나누는 기준은 다음과 같다. (1) 소실점의 위치가 이미지 내부에 있을 때, (2) 소실점의 위치가 이미지 외부(보강된 영역)에 1개 있을 때, (3) 소실점의 위치가 2개 이상 발견될 때의 3가지 유형으로 나눈다. (2)의 경우 소실점 위치가 이미지 중심으로부터 좌측에 편향되었을 때와 우측에 편향되었을 때로 나뉜다. 유형에 따라 소실점 위치 정보 VP_{info}를 출력하여 Swin Transformer의 어텐션(attention)을 사용하는 그림 1의 Enhanced Swin Transformer Block으로 전달한다.

2. 향상된 SW-MSA

입력 이미지(input image)를 인코더(encoder)로 다운 샘플링(downsampling) 하기 위하여 먼저 입력 이미지(input image)를 작은 조각 단위(patch)로 쪼갠다. 합성곱(convolution)의 커널(kernel)과 간격(stride)은 서로 같게 한다. NYU-depth V2 데이터 세트[17] 크기인 640×480으로 본 모델에서 조각 크기로 사용하는 4×4 예시를 들어보면, 시퀀스(sequence) 길이는 19,200가 된다. 이후 사전에 정한 합성곱의 채널(channel)에 맞추어 선형 결합(linear embeding)을 한다. 본 논문에서 제시하는 모델의 채널 수는 96이다. 향상된 Swin Transformer block의 입력은 층 정규화(layer normalization)를 거쳐 W-MSA(Window Multi head Self Attention) 모듈로 연결된다. W-MSA 모듈은 시계열 데이터에 대해 트랜스포머를 적용할 때 계산량을 줄이기 위해서 사용한다. 특징 지도(feature map)를 일정한 창으로 나누어 창 안에서의 셀프-어텐션 기법(self-attention)의 출력을 계산한다. W-MSA 모듈 이후에는 다층 퍼셉트론(multi layer perceptron) 모듈이 배치된다. 이 다층 퍼셉트론은 2개의 선형 계층(linear layer)으로 구성되어 있고 GELU(Gaussian Error Linear Uint) 함수로 연결되어 있다. 각 모듈 뒤에는 잔차 연결(residual connection)이 적용된다. W-MSA

모듈과 연결된 다층 퍼셉트론의 출력은 앞선 과정이 반복되는 구조에 연결되는데 W-MSA가 앞서 추출한 소실점 정보를 사용하는 향상된 SW-MSA(Enhanced Shifted Window Multi head Self Attention) 모듈로 대치된다.

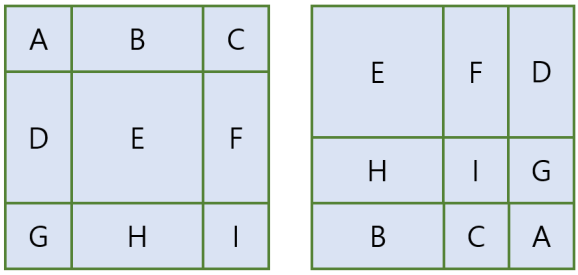

그림 3. SW-MSA에서의 창의 이동 예시
Fig. 3. Example of windows shifting in SW-MSA.

SW-MSA는 W-MSA에서 누락된 특징을 보완하기 위해 사용된다. W-MSA는 나눈 창 안에서만 셀프-어텐션(self-attention)을 계산하기 때문에 계산량은 감소하나 창끼리의 연관성은 알지 못한다. 창끼리의 특징을 추출하기 위해 새로운 모양으로 창을 나누고 창을 이동시킨다. 새로운 모양의 창은 그림 3의 좌측 모양이고 이동된 창은 그림 3의 우측과 같다. 하지만 우측과 같은 상태로 셀프-어텐션 기법을 사용하게 되면 문제가 발생하는데, 창 F-I 영역과 창 D-G 영역은 이동 전 원래 이미지에서 연속된 접합하고 있는 영역이 아니므로 서로 연관이 없다. 창 H-I와 창 B-C의 영역도 마찬가지로 서로 연관이 없다. 따라서 창 A, 창 B, 창 C, 창 D, 창 G에는 마스크(mask) 처리를 하여 어텐션 기법을 사용한다. 마스크 처리는 마스크 처리가 안 될 부분에는 0을, 마스크 처리가 될 부분에는 음수값을 어텐션 매트릭스(attention matrix)에 더하는데 소프트맥스(softmax) 함수 특성상 그 부분은 작은 값이 되어 값이 무시가 된다. 향상된 SW-MSA는 소실점이 이미지에서 가장 깊은 곳이라는 특징과 깊이 추정에서 추정 오류가 높게 발생하는 픽셀이 깊은 픽셀인 것을 근거로 소실점 위치 정보인 VP_{info}의 유형에 따라 다르게 작동한다.

가. 소실점이 이미지 내부에 존재할 때

소실점이 이미지 내부에 존재할 때는 소실점이 1개만 존재하고, 그 소실점이 이미지 중심이거나 중심에 가까운 위치에 존재한다. 즉, 깊이 정보 추출을 위하여 주요 깊게 학습해야 할 픽셀들이 이미지 중앙에 가깝다는 것을 의미한다. 따라서 소실점이 이미지 내부에 존재할 때는 SW-MSA가 한 번 더 반복될 수 있도록 배치한다. 단, SW-MSA 마지막에는 조각 병합(patch merging) 혹은 조각 확장(patch expanding)을 진행하기 위해 다중 배치 텐서(multi batch tensor)를 원래의 조각 단위의 텐서로 정합시키는데 이 과정을 생략한다. 그리고 재차 진행되는 SW-MSA의 창의 이동 방향을 그림 3과 다르게 그림 4와 같이 설정하고, 마스크 적용도 이동된 창에 맞추어 창 C, 창 F, 창 G, 창 H, 창 I에 한다. 중심에 있는 창은 한 번 더 셀프-어텐션 할 수 있도록 하되, 이전 SW-MSA 작용 시 마스크로 무시되었던 창과 셀프-어텐션 값을 계산한다.

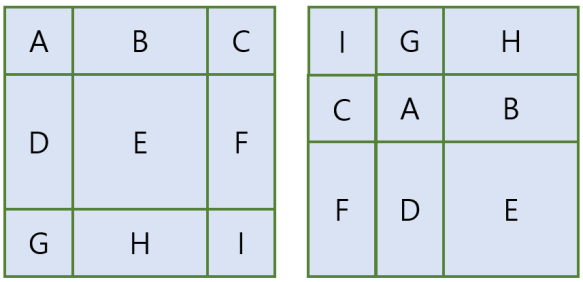

그림 4. 향상된 SW-MSA에서 두 번째 창의 이동
Fig. 4. Shifting of the second window in improved SW-MSA.

나. 소실점이 이미지 외부에 1개 존재할 때

소실점이 소실점 검출 모듈에서 보강된 영역인 이미지 외부에서 1개 발생한 경우 왼쪽이나 오른쪽에 깊이 거리가 치우쳐있는 경우이다. 이 경우 SW-MSA 모듈을 한번 거치지만 소실점의 좌우 편향 정도를 이용하여 창의 이동을 서로 달리 적용한다. 소실점의 좌우 편향도에 따라 마스크 적용을 달리하여 셀프-어텐션을 할 위치를 고려한 것이다. 이미지를 중심으로 소실점이 좌측에 존재하는 경우 그림 4의 우측 그림과 같이 창의 이동을 적용한다. 이미지를 중심으로 소실점이 우측에 존재하는 경우 그림 3의 우측 그림과 같이 창의 이동을 적용한다.

다. 소실점이 2개 이상 존재할 때

데이터 분석 따라 소실점이 2개이거나 3개일 경우 원본 이미지 데이터(raw image data)를 앞선 소실점이 1개만 존재했을 경우들과 유의미하게 분류할 수 있다. 그러나 향상된 SW-MSA로 셀프-어텐션 정도와 위치를 조절하는 본 모델에 적용하기에는 부적합한 분류이

다. 또한 소실점이 4개 이상 검출되었을 경우 실제 세계에서 쓰이는 데이터 특성상 소실점이 4개 이상 검출될 수 없으며, 만일 모듈에서 검출했다면 잘못 추정한 것이다. 따라서 소실점이 2개 이상 검출되었을 경우 그림 3의 SW-MSA 모듈이 작동한다.

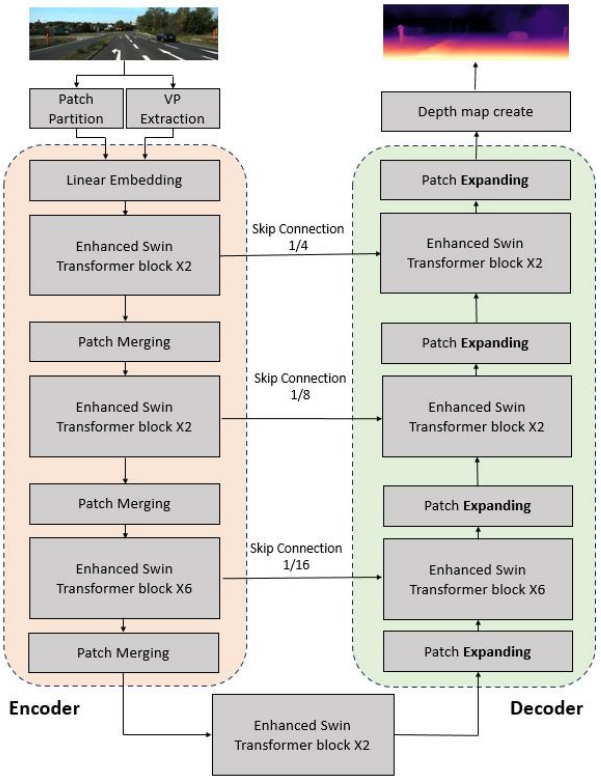

그림 5. 대칭 구조의 향상된 SW-MSA Swin Transformer를 사용한 깊이 추정 모델 구조
Fig. 5. Depth estimation using symmetric Swin Transformer.

3. 인코더-디코더 대칭 구조

인코더-디코더(encoder-decoder) 모델 구조는 깊이 추정 분야에서 많이 사용되고 있는 훈련 구조이다. 본 모델은 인코더와 디코더를 대칭으로 배치하였다. 인코더는 앞서 서술한 Swin Transformer의 동작 방식으로 움직인다. 디코더는 인코더와 대칭적으로 구성하였는데 이는 U-Net[18, 19]의 구조를 참고하였다. 인코더에서는 조각 병합(patch merging)을 통하여 해상도(resolution)과 채널(channel)을 줄이는 다운 샘플링(downsampling)을 하였다면, 디코더에서는 조각 확장(patch expanding)을 통하여 추출된 깊이 특징들을 업 샘플링(upsampling)한다. 생략 연결(skip connection)은 다운 샘플링과 과정에서 발생하는 정보 손실을 막기 위해 다운 샘플링의 얕은 단계에서의 특징과 업 샘플링의 깊은 단계에서의 특징을 연결한다. 디코더를 통해 업샘플링으로 재정립된 특징(feature)은 시그모이드(sigmoid) 함수를 통해 깊이 정보로 나타내고 정답(ground truth)과 비교한다.

4. 손실 함수

단안 이미지 깊이 추정에 필요한 손실 함수는 SI-log(Scale-Invariant logarithmic)[20]를 사용하고, 계산하는 방법은 수식 (1)과 같다.

$$g_i = \log(\hat{d}_i) - \log(d_i)$$
$$L = \alpha\sqrt{\frac{1}{T}\sum_i (g_i)^2 - \frac{\lambda}{T^2}(\sum_i g_i)^2} \quad (1)$$

Bhat[21]의 연구에 따라 $\lambda = 0.85$, $\alpha = 10$ 으로 설정한다. $\hat{d}_i$는 추정된 깊이 지도의 픽셀값 d_i : 정답(Ground truth)의 픽셀값이다.

Ⅳ. 실 험

1. 실험에 사용된 데이터 세트

가. KITTI dataset

KITTI(Karlsruhe Institute of Technology 및 Toyota Technological Institute)[22] 데이터 세트(dataset)는 모바일 로봇 공학 및 자율 주행에 사용되는 가장 인기 있는 데이터 세트 중 하나이다. 고해상도 RGB, 회색 스테레오 카메라 및 3D 레이저 스캐너를 포함한 다양한 센서 양식으로 기록된 몇 시간의 교통 시나리오로 구성된다. 인기가 있음에도 불구하고 데이터 세트 자체에는 의미론적 분할을 위한 기반 정보가 포함되어 있지 않지만, 다양한 연구자들이 필요에 따라 데이터 세트의 일부에 수동으로 주석을 달았다. 본 논문에서는 KITTI 데이터 세트 중 깊이 추정 분야에서 벤치마크(Benchmark) 역할을 하는 KITTI Eigen data[20]를 사용하였다. 23,488개의 데이터 중 20,824개의 데이터를 훈련 데이터(training set)로 사용하여 학습하였고 2,664개의 데이터를 검정 데이터(validation set)로 사용하였다. 697개의 공식적으로 사용되는 테스트 이미지로 모델의 성능을 평가한다.

나. NYU Depth V2

NYU-Depth V2[17] 데이터 세트는 2012년에 발표된 깊이 정보 데이터 세트로써, KITTI 데이터 세트와 더불어 깊이 추정 분야에서 많이 쓰이는 벤치마크 중 하

나이다. NYU Depth V2 데이터 세트는 Microsoft Kinect의 RGB 및 깊이 카메라로 녹화된 다양한 실내 장면의 비디오 연속장면(Sequence)으로 구성된다. 24,231개의 이미지 데이터 중 21,324개의 데이터를 훈련 데이터로 사용하였고 2,907개의 데이터를 검정 데이터로 사용하였다. 654개의 공식적으로 사용되는 테스트 이미지로 모델의 성능을 평가한다.

2. 성능 지표

결과 분석에 쓰일 지표들은 깊이 추정 분야에서 주로 쓰이는 절대 상대 오차(absolute relative error), 제곱 평균 오차(root mean square error), 로그 스케일 제곱 평균 오차(log scale Root Mean Square Error), 임곗값 미만 정확도(accuracy under a threshold)이다. 절대 상대 오차는 수치 계산 시 절댓값을 사용하는 L1 거리 공식(Manhattan distance)을 사용하여 모든 에러를 가중치 없이 평가한다.

$\hat{d}_p$는 특정 픽셀을 깊이 추정한 값을 예측하고, d_p은 특정 픽셀의 깊이 정답 값을 의미하는데, 절대 상대 오차, 제곱 평균 오차, 로그 스케일 제곱 평균 오차를 구하는 방법은 수식 (2)와 같다. 이 값들은 오차값이기 때문에 낮을수록 깊이 추정의 성능이 좋다는 의미이다.

$$\text{절대 상대 오차} = \frac{1}{T}\sum_{p}\frac{|d_p-\hat{d}_p|}{d_p}$$
$$\text{제곱 평균 오차} = \sqrt{\frac{1}{T}\sum_{p}(d_p-\hat{d})^2}$$
$$\text{로그 제곱 평균} = \sqrt{\frac{1}{T}\sum_{p}(\log(d_p)-\log(\hat{d}_p))^2} \quad (2)$$

임곗값 미만 정확도는 Depth Estimation 모델의 Accuracy를 측정하기 위해 도입되었다. 라이다(light detection and ranging) 포인트 클라우드를 이용하여 만든 결핍 깊이 지도(sparse depthmap)와 깊이 추정 모델을 통해 얻은 깊이 지도 예측 간의 정확도를 측정하는 지표이다. $\hat{d}_p$와 d_p를 분수로 비교하는데 이때 큰 수를 분자에 두어 항상 1보다 큰 값을 얻도록 한다. 임곗값(threshold)인 δ를 이용하여 비율이 δ보다 낮으면 성공적인 예측으로 간주하는데, 이때 깊이 추정에서 전통적으로 $\delta = 1.25,\ 1.25^2,\ 1.25^3$ 을 임곗값으로 정의한다. 정확도를 측정하는 지표이므로 오차와 달리 높을수록 높은 성능을 보이는 깊이 추정 모델이다.

3. 모델 학습 및 평가

모델의 학습 환경은 다음과 같다.

- Intel Xeon E5-2620V
- NVIDIA Geforce RTX 2080 11GB Dual
- Ubuntu 18.04.6 LTS
- Pytorch 라이브러리 사용
- AdamW 최적화 알고리즘 사용

표 1. KITTI Dataset 에서의 성능 비교.
Table 1. Performance comparison on KITTI Dataset.

Model	Error			Accuracy		
	Abs Rel	RMSE	RMSE log	$\delta<\rho$	$\delta<\rho^2$	$\delta<\rho^3$
DORN[23]	0.072	2.727	0.120	0.932	0.984	0.994
GCNDepth[24]	0.104	4.494	0.181	0.888	0.965	0.984
DNet[25]	0.113	4.812	0.191	0.877	0.960	0.981
본 논문의 모델	**0.065**	**2.113**	**0.088**	**0.968**	**0.997**	**0.999**

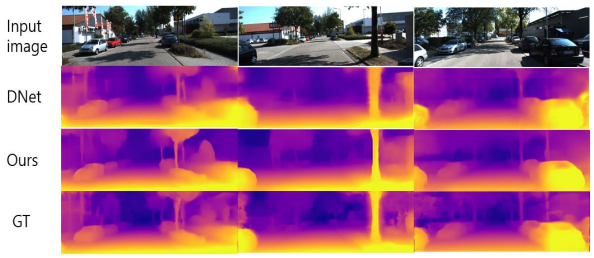

그림 6. KITTI 데이터 세트에 대한 정성적 비교
Fig. 6. Qualitative comparison on the KITTI dataset.

표 1은 KITTI 데이터 세트에서의 다른 모델과의 성능 비교이다. 표 1에서 절대 상대 오차는 AbsRel, 제곱 평균 오차는 RMSE, 로그 제곱 평균은 RMSE log로 표현하였다. 또한, ρ는 1.25이다. 본 모델은 이전의 제안된 모델들보다 상당한 성능 향상을 보인다. Dnet와 비교하면, AbsRel, RMSE, RMSE log는 각각 0.048, 2.699, 그리고 0.103 감소하였다. 정확도 지표 δ, δ^2, δ^3은 각각 0.091, 0.037, 그리고 0.018 상승하였다. 그래프 합성곱 신경망(Graph Convolution Network) 기반의 GCNDepth보다도 더 예측을 잘하는 모델이라고 할 수 있다. 그림 6은 제시하는 모델이 비교 모델인 DNet 모델보다 자세히 깊이 특징을 나타내며 정답(ground truth)에 가깝게 깊이 지도를 생성함을 보인다.

표 2는 NYU-depth V2에서의 성능 비교이다. 표 1에서 절대 상대 오차는 AbsRel, 제곱 평균 오차는

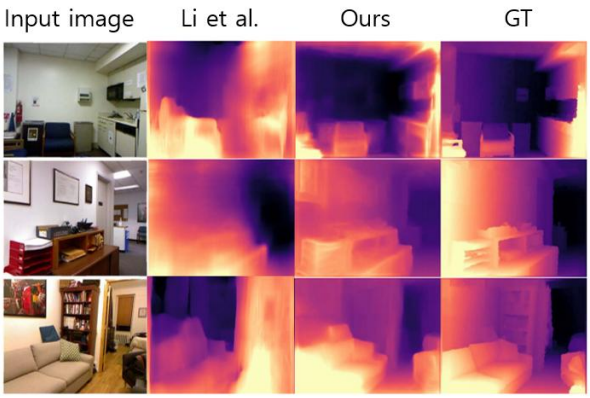

그림 7. NYU-depth V2 데이터 세트에 대한 정성적 비교
Fig. 7. Qualitative comparison on the NYU-depth V2 dataset.

RMSE, 로그 제곱 평균은 RMSE log로 표현하였다. 또한, ρ는 1.25이다. 다른 모델보다 AbsRel, RMSE가 각각 최대 0.101, 0.358 낮고 정확도 지표인 δ, δ^2, δ^3은 각각 0.843, 0.974, 0.994를 기록하면서 비교군에서 가장 좋은 성능을 나타냈다. 그림 7에서처럼 본 모델이 여타 모델보다 정답(ground truth)에 가깝게 깊이 지도를 잘 생성한다.

표 2. NYU Depth V2 Dataset 에서의 성능 비교.
Table 2. Performance comparison on NYU Depth V2 Dataset.

Model	Error		Accuracy		
	AbsRel	RMSE	$\delta < \rho$	$\delta < \rho^2$	$\delta < \rho^3$
eigen et al.[26]	0.158	0.641	0.769	0.950	0.988
Xu et al.[27]	0.163	0.655	0.706	0.925	0.981
Relative Depth[28]	**0.131**	0.538	0.837	0.971	**0.994**
본 논문의 모델	**0.131**	**0.463**	**0.843**	**0.974**	**0.994**

V. 결 론

본 논문은 장면에 추상적으로 존재하는 개념인 소실점을 찾고 소실점의 위치 정보를 이용하여 깊이 추정을 할 수 있는 신경망을 제안하였다. 기존 Swin Transformer에 소실점 탐지 모듈을 추가하고 소실점 위치 정보를 향상된 SW-MSA에 전달한다. 향상된 SW-MSA는 소실점의 위치 정보에 따라 창의 이동(Shifting window) 방식이 달라진다. 이를 통하여 깊이 추정 분야에서 정확도가 떨어지는 깊은 곳을 위주로 셀프-어텐션(self-attention)을 강화하였다. 또한, 향상된 SW-MSA가 추가된 Swin Transformer를 대칭 인코더-디코더 구조로 배열하여 깊이 추정을 할 수 있도록 설계하였다. 그 결과 KITTI 데이터 세트에서와 NYU-depth V2 데이터 세트에서 이전의 깊이 추정 연구들보다 정성적으로나 정량적으로 더 나은 성과를 도출하였다. 본 논문은 데이터 특성에 맞추어 작동 신경망의 유형을 선택하는 모듈을 삽입한다는 점에서 깊이 추정 분야에서의 성과뿐만 아니라 인공지능 신경망 연구 측면에서도 성과를 거두었다. 소실점의 유형을 나누는 방식을 데이터 분석적으로가 아닌 기하학적으로 정의하고 보강(padding)이나 가우시안 구(Gaussian sphere)로의 좌표계 이동(mapping) 없이 소실점을 찾아 손실 함수(loss function)에 사용할 수 있는 정보로 변환하는 연구가 추후의 연구이다. 그리고 소실점 모듈을 깊이 추정 분야뿐만 아니라 SLAM[29]이나 자율 주행 같은 실제 산업에 적용하는 것 또한 추후의 연구 과제이다.

Acknowledgement

이 연구는 2023년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구이고 (P0023718, 2023년 산업 전환형 무기발광 디스플레이 전문인력양성사업), 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2022R1F1A1074775).

REFERENCES

[1] M. Ding, Y. Huo, H. Yi, Z. Wang, J. Shi, Z. Lu and P. Luo, “Learning depth-guided convolutions for monocular 3d object detection,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops,* pp. 1000-1001, June 2020.
[2] Wang, Weiyue and U. Neumann, “Depth-aware cnn for rgb-d segmentation,” in *Proc. of the European Conf. on Computer Vision,* pp. 135-150, Munich, Germany, September 2018.
[3] T. H. Vu, H. Jain, M. Bucher, M. Cord, M and P. Pérez, “Dada: Depth-aware domain adaptation in semantic segmentation,” in *Proc. of the IEEE/CVF International Conf. on*

Computer Vision, pp. 7364-7373, Seoul, Korea, October 2019.

[4] C. H. Pyo and C. H. Lim, "Image dehazing and attenuation Coefficeint Estimation using Depth Estimatino in a Single Image," *Journal of the Institute of Electronics and Information Engineers,* vol. 60, no. 6, pp. 65-72, June 2023.

[5] S. Lee, J. Kim, J. S. Yoon, S. Shin, O. Balio, N. Kim, T. H. Lee, H. S. Hong, S. H. Han and I. S. Kweon, "Vpgnet: Vanishing point guided network for lane and road marking detection and recognition," in *Proc. of the IEEE International Conf. on Computer Vision*, pp. 1947-1955, Venice, Italy, October 2017.

[6] J. H. Ko, "2D Spatial Map based on Lidar and Stereo Camera for Autonomous Driving," *Journal of the Institute of Electronics and Information Engineers*, vol. 58, no. 4, pp. 69-74, April 2021.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. A. Gomez, Ł. Kaiser and I. Polosukhin, "Attention is all you need," Advances in *Neural Information Processing Systems,* vol. 30, California, USA, December 2017.

[8] 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," *arXiv Preprint arXiv:2010.11929,* October 2020.

[9] X. Zhai, A. Oliver, A. Kolesnikov and L. Beye, "S4l: Self-supervised semi-supervised learning," in *Proc. of the IEEE/CVF International Conf. on Computer Vision,* pp. 1476-1485, Seoul, Korea, October 2019.

[10] Y. Wang, W. Chao, D. Garg, B. Hariharan, M. Campbell and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition,* pp. 8445-8453, June 2019.

[11] J. L. GonzalezBelloand and M. Kim, "Forget about the lidar: Self-supervised depth estimators with med probability volumes," Advances in *Neural Information Processing Systems,* vol. 33, pp. 12626-12637, December 2020.

[12] C. Godard, O. Mac Aodha and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 270-279, Hawaii, USA, July 2017.

[13] J. Coughlan and A. L. Yuille, "The manhattan world assumption: Regularities in scene statistics which enable bayesian inference," Advances in *Neural Information Processing Systems,* vol. 13, November 2000.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. of the IEEE/CVF International Conf. on Computer Vision,* pp. 10012-10022, Montreal, Canada, October 2021.

[15] A. Borji, "Vanishing point detection with convolutional neural networks," *arXiv Preprint arXiv:1609.00967*, September 2016.

[16] Y. Zhou, H. Qi, J. Huang and Y. Ma, "Neurvps: Neural vanishing point scanning via conic convolution," Advances in *Neural Information Processing Systems,* vol. 32, Vancouver, Canada, December 2019.

[17] N. Silberman, D. Hoiem, P. Kohli and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proc. of the European Conf. on Computer Vision,* Part V 12, pp. 746-760, Florence, Italy, October 2012.

[18] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conf.,* Part III 18, pp. 234-241, Munich, Germany, October 2015.

[19] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian and M Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Proc. of the European Conf. on Computer Vision*. pp. 205-218, Tel Aviv, Israel, October 2022

[20] D. Eigen, C. Puhrsch and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," Advances in *Neural Information Processing Systems,* vol. 27, Montreal, Canada, December 2014.

[21] S. F. Bhat, I. Alhashim and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proc. of the IEEE/CVF Conf. on*

Computer Vision and Pattern Recognition, pp. 4009-4018, Tennessee, USA, June 2021.
[22] A. Geiger, P. Lenz, C. Stiller and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research,* vol. 32, no. 11, pp. 1231-1237, August 2013.
[23] H. Fu, M. Gong, C. Wang, K. Batmanghelich and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 2002-2011, Salt Lake City, USA, June 2018.
[24] A. Masoumian, H. A. Rashwan, S. Abdulwahab, J. Cristiano, M. S. Asif and D. Puig, "GCNdepth: Self-supervised monocular depth estimation based on graph convolutional network," *Neurocomputing,* vol. 517, pp. 81-92, January 2023.
[25] F. Xue, G. Zhuo, Z. Huang, W. Fu, Z. Wu and M. H. Ang, "Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications," in *Proc. IEEE/RSJ International Conf. on Intelligent Robots and Systems,* pp. 2330-2337, Las Vegas, USA, October 2020.
[26] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. of the IEEE International Conf. on Computer Vision,* pp. 2650-2658, Las Condes, Chile, December 2015.
[27] D. Xu, E. Ricci, W. Ouyang, X. Wang and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 5354-5362, Hawaii, USA, July 2017.
[28] J. Lee and C. Kim, "Monocular depth estimation using relative depth maps," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition,* pp. 9729-9738, Long Beach, USA, June 2019.
[29] B. S. Kim, I. J. Nam, J. K. Ryu, H. C. Moon, "A Comparative Study on GPS Module and SLAM Module for Driving Stability of Waypoint-based Autonomous Delivery Robot". *Journal of the Institute of Electronics and Information Engineers,* vol. 59, no. 12, pp. 65-72, December 2022.

저 자 소 개

조 용 석(학생회원)
2021년 중앙대학교 전자전기공학부 학사 졸업.
2021년~중앙대학교 전자전기공학과 석사 과정.
<주관심분야 : 컴퓨터 비전, 깊이 추정, 비전 트랜스포머, 빅데이터>

김 나 라(학생회원)
2021년 한국교통대학교 전자공학과 학사 졸업.
2022년~중앙대학교 전자전기공학과 석사 과정.
<주관심분야 : 빅데이터, 인공지능, 컴퓨터 비전, 자연어처리>

박 호 현(정회원)
1987년 서울대학교 계산통계학과 학사 졸업.
1995년 KAIST 정보통신공학과 석사 졸업.
2001년 KAIST 전자전산학과 박사 졸업.
1987년~2002년 삼성전자 통신연구소 수석연구원
2003년~현재 중앙대학교 전자전기공학부 교수
<주관심분야 : 빅데이터, 인공지능, 컴퓨터 비전, 정보 보안, 임베디드 시스템>