(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2023-0166464
(43) 공개일자 2023년12월07일

(51) 국제특허분류(Int. Cl.)
G06T 7/50 (2017.01) *G06T 3/40* (2006.01)
G06T 7/11 (2017.01) *G06T 9/00* (2019.01)
(52) CPC특허분류
G06T 7/50 (2017.01)
G06T 3/40 (2013.01)
(21) 출원번호 10-2022-0066476
(22) 출원일자 2022년05월31일
심사청구일자 2022년05월31일

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
최유경
서울특별시 광진구 능동로 209, 세종대학교 대양AI센터 601호(군자동)
한대찬
서울특별시 광진구 광나루로19길 44-1, 201호(군자동)
(뒷면에 계속)
(74) 대리인
송인호, 최관락

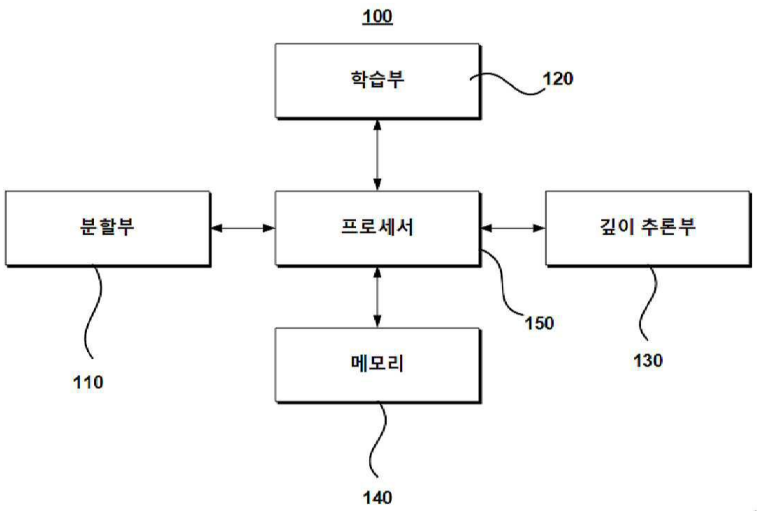
전체 청구항 수 : 총 13 항

(54) 발명의 명칭 **트랜스포머 기반 깊이 추정 장치 및 방법**

(57) 요 약

트랜스포머 기반 깊이 추정 장치 및 방법이 개시된다. 트랜스포머 기반 깊이 추정 장치는 스윈-트랜스포머(Swin Transformer) 인코더와 픽셀-단위 스킵 어텐션(Pixel-Wise Skip-Attention) 기반 디코더 모듈을 통해 인코더 특징맵과 이전 단계의 디코더 특징맵을 융합하여 스윈-트랜스포머 인코더에서 추출된 전역적 정보를 보충하고, 인코더 특징맵을 입력으로 픽셀 레벨 주의 맵을 생성하여 전역 정보만 포함된 디코더 특징맵에 국소적 정보를 향상시킬 수 있으며, 디코더에서 나온 최고해상도의 깊이 영상을 라벨로 활용하여 중간 레벨의 깊이 추정 결과와 비교하는 자가 증류 손실을 이용하여 학습될 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G06T 7/11 (2017.01)
G06T 9/00 (2019.01)
G06T 2207/10028 (2013.01)
G06T 2207/20081 (2013.01)
(72) 발명자
신정민
서울특별시 광진구 능동로28길 22, 203호(능동)
김남일
경기도 용인시 수지구 동천로153번길 7, 1103동 1401호(동천동, 한빛마을 래미안 이스트팰리스 1단지)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711159902
과제번호 2021-0-02067-002
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 차세대인공지능핵심원천기술개발
연구과제명 다목적 비디오 검색을 위한 차세대 인공신경망 기술 개발
기 여 율 1/2
과제수행기관명 세종대학교 산학협력단
연구기간 2022.04.01 ~ 2023.03.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711153517
과제번호 2020M3F6A1109603
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 미래국방혁신기술개발
연구과제명 전장 환경 인식 및 무기체계 성능을 지원하는 AI 카메라 기술 개발
기 여 율 1/2
과제수행기관명 세종대학교 산학협력단
연구기간 2022.02.01 ~ 2023.01.31

명 세 서

청구범위

청구항 1

입력 영상을 겹치는 부분들이 없도록 복수의 패치들로 나누는 분할부; 및

상기 복수의 패치들을 트랜스포머 기반 깊이 추정 모델에 적용하여 깊이 맵을 생성하는 깊이 추정부를 포함하되,

상기 트랜스포머 기반 깊이 추정 모델은,

계층 구조를 가지는 n개의 스윈 트랜스포머 인코더로 구성되며, 상기 n개의 스윈 트랜스포머 인코더를 통해 상기 복수의 패치들이 통과되어 다중 스케일 인코더 특징맵을 생성하는 인코더부; 및

복수의 디코더 모듈을 가지되, 상기 복수의 디코더 모듈 중 적어도 일부를 통해 상기 인코더 특징맵과 이전 디코더 모듈의 디코더 특징맵을 융합하여 다중 스케일의 디코더 특징맵을 생성하고, 상기 다중 스케일의 디코더 특징맵을 이용하여 다중 스케일의 깊이 맵을 생성하는 디코더부를 포함하는 것을 특징으로 하는 트랜스포머 기반 깊이 추정 장치.

청구항 2

제1 항에 있어서,

상기 복수의 디코더 모듈 중 최후단에 위치된 디코더 모듈은 n개의 스윈 트랜스포머 인코더 중 최후단에 위치된 스윈 트랜스포머 인코더에 의해 출력되는 인코더 특징맵($\bar{F}_4^E \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32}}$)을 입력받은 후 잔차 블록을 적용하고 업샘플링한 후 1 x 1 컨볼루션 레이어를 적용하여 디코더 특징맵($F_4^D \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$)을 생성하되,

상기 디코더 특징맵($F_4^D \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$)은 상기 최후단에 위치된 디코더 모듈의 전단에 위치된 디코더 모듈로 입력되는 것을 특징으로 하는 트랜스포머 기반 깊이 추정 장치.

청구항 3

제2 항에 있어서,

상기 복수의 디코더 모듈 중 적어도 일부는 각각,

동일 해상도의 인코더 특징맵과 후단에 위치된 디코더 모듈의 디코더 특징맵을 각각 페어(pair)로 입력받은 후 상기 입력된 인코더 특징맵과 디코더 특징맵을 융합하고, 상기 입력된 인코더 특징맵을 이용하여 생성된 픽셀 레벨 주의 맵을 상기 융합된 특징맵에 합산하여 디코더 특징맵을 생성하는 어텐션 모듈; 및

상기 디코더 특징맵에 2개의 컨볼루션 레이어를 적용하여 깊이 맵을 생성하는 깊이 맵 생성 모듈을 포함하는 것을 특징으로 하는 트랜스포머 기반 깊이 추정 장치.

청구항 4

제3 항에 있어서,

상기 픽셀 레벨 주의 맵은,

상기 인코더 특징맵에 잔차 블록을 적용하고 1 x 1 컨볼루션 레이어를 적용한 후 소프트맥스(softmax)를 적용하여 생성되는 것을 특징으로 하는 트랜스포머 기반 깊이 추정 장치.

청구항 5

제3 항에 있어서,

상기 디코더 특징맵은 상기 인코더 특징맵과 상기 이전 디코더 모듈의 디코더 특징맵을 요소별 합산한 제1 특징맵과 상기 픽셀 레벨 주의 맵을 요소별 곱셈 연산하여 제2 특징맵을 생성한 후 상기 제2 특징맵에 잔차 블록을 적용한 후 상기 제1 특징맵과 합산하여 제3 특징맵을 생성한 후 1 x 1 컨볼루션 레이어를 적용한 후 바이리니어 업샘플링하여 생성되는 것을 특징으로 하는 트랜스포머 기반 깊이 추정 장치.

청구항 6

제1 항에 있어서,

상기 다중 스케일의 깊이 맵 중 최고 해상도의 깊이 맵은 정답 라벨(Pseudo Label)로 활용되되,

상기 최고 해상도 깊이 맵은 광도 손실(Photometric loss)을 적용하고, 상기 정답 라벨을 제외한 깊이 맵은 상기 정답 라벨과의 자가 증류 손실(Self-distillation loss)을 적용하여 상기 트랜스포머 기반 깊이 추정 모델이 학습되는 것을 특징으로 하는 트랜스포머 기반 깊이 추정 장치.

청구항 7

제6 항에 있어서,

상기 자가 증류 손실은 하기 수학식을 이용하여 계산되는 것을 특징으로 하는 트랜스포머 기반 깊이 추정 장치.

$$\mathcal{L}_{sd} = \lambda_{adp} * \frac{1}{3}\sum_{i=1}^{3} |upsample(D_i, 2^i) - D_0|_1 * M$$

$$\lambda_{adp} = \begin{cases} epoch/epoch_{thr}, & epoch < epoch_{thr} \\ 1, & othersize \end{cases}$$

여기서, 로 적응적 가중치를 나타내며, D_i 는 다른 깊이 맵을 나타내고, i는 깊이 맵 인덱스를 나타내며, D_0 는 정답 라벨을 나타내고, M은 마스크를 나타냄.

청구항 8

(a) 입력 영상을 겹치는 부분들이 없도록 복수의 패치들로 나누는 단계;

(b) 상기 복수의 패치들을 계층 구조를 가지는 n개의 스윈 트랜스포머(Swin transformer) 인코더에 적용하여 n개의 다중 스케일 인코더 특징맵을 생성하는 단계; 및

(c) 계층 구조를 가지는 복수의 디코더에 상기 인코더 특징맵과 이전 디코더의 디코더 특징맵을 적용하여 융합

하여 다중 스케일의 디코더 특징맵을 생성하고, 상기 다중 스케일의 디코더 특징맵을 이용하여 다중 스케일의 깊이 맵을 생성하는 단계를 포함하되,

상기 복수의 디코더는 각각, 동일한 해상도의 인코더 특징맵을 입력받으며,

상기 다중 스케일의 깊이 맵 중 두번째 고해상도의 깊이 맵이 깊이 추론에 이용되는 것을 특징으로 하는 트랜스포머 기반 깊이 추정 방법.

청구항 9

제8 항에 있어서,

상기 (c) 단계에서,

상기 복수의 디코더 모듈 중 최후단에 위치된 디코더 모듈은 n개의 스윈 트랜스포머 인코더 중 최후단에 위치된 스윈 트랜스포머 인코더에 의해 출력되는 인코더 특징맵($\bar{F}_4^E \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32}}$)을 입력받은 후 잔차 블록을 적용하고 업샘플링한 후 1 x 1 컨볼루션 레이어를 적용하여 디코더 특징맵($F_4^D \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$)을 생성하되,

상기 디코더 특징맵($F_4^D \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$)은 상기 최후단에 위치된 디코더 모듈의 전단에 위치된 디코더 모듈로 입력되는 것을 특징으로 하는 트랜스포머 기반 깊이 추정 방법.

청구항 10

제8 항에 있어서,

상기 (c) 단계는,

상기 복수의 디코더 각각은,

동일 해상도의 인코더 특징맵과 후단에 위치된 디코더 모듈의 디코더 특징맵을 각각 페어(pair)로 입력받은 후 상기 입력된 인코더 특징맵과 디코더 특징맵을 융합하고, 상기 입력된 인코더 특징맵을 이용하여 생성된 픽셀 레벨 주의 맵을 상기 융합된 특징맵에 합산하여 디코더 특징맵을 생성하는 어텐션 모듈; 및

상기 디코더 특징맵에 2개의 컨볼루션 레이어를 적용하여 깊이 맵을 생성하는 깊이 맵 생성 모듈을 포함하되,

상기 다중 스케일의 디코더 특징맵 중 최고 해상도 디코더 특징맵은 정답 라벨로 설정되어 학습에 이용되는 것을 특징으로 하는 트랜스포머 기반 깊이 추정 방법.

청구항 11

제10 항에 있어서,

상기 픽셀 레벨 주의 맵은, 상기 인코더 특징맵에 잔차 블록을 적용하고 1 x 1 컨볼루션 레이어를 적용한 후 소프트맥스(softmax)를 적용하여 생성되는 것을 특징으로 하는 트랜스포머 기반 깊이 추정 방법.

청구항 12

제10 항에 있어서,

상기 디코더 특징맵은 상기 인코더 특징맵과 상기 이전 디코더 모듈의 디코더 특징맵을 요소별 합산한 제1 특징맵과 상기 픽셀 레벨 주의 맵을 요소별 곱셈 연산하여 제2 특징맵을 생성한 후 상기 제2 특징맵에 잔차 블록을 적용한 후 상기 제1 특징맵과 합산하여 제3 특징맵을 생성한 후 1 x 1 컨볼루션 레이어를 적용한 후 바이리니어

업샘플링하여 생성되는 것을 특징으로 하는 트랜스포머 기반 깊이 추정 방법.

청구항 13

제8 항 내지 제12항 중 어느 하나의 항에 따른 방법을 수행하기 위한 프로그램 코드를 기록한 컴퓨터로 판독 가능한 기록매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 트랜스포머 기반 깊이 추정 장치 및 그 방법에 관한 것이다.

배 경 기 술

[0003] 종래의 경우, 자기지도 학습 기반 단일 영상 깊이 추정을 위해 대부분 컨볼루션 뉴럴 네트워크(CNN, 이하 CNN이라 칭하기로 함)를 깊이 네트워크로 활용하였다. 그러나 CNN의 구조적 한계인 국지적 수용 영역(local receptive field)를 가지고 있어 버스, 기차, 트레일러 등 도로 주행 환경에서 쉽게 마주할 수 있는 대형 물체 또는 도로와 같은 전체적인 영상 속의 문맥 정보를 제대로 포착 및 추론하지 못하는 단점이 있다.

발명의 내용

해결하려는 과제

[0005] 본 발명은 트랜스포머 기반 깊이 추정 장치 및 방법을 제공하기 위한 것이다.

[0006] 이를 통해, 본 발명은 국지적인 영역과 전역적인 영역 모두 정확하고 일관성 있는 깊이 추정이 가능한 트랜스포머 기반 깊이 추정 장치 및 방법을 제공하기 위한 것이다.

과제의 해결 수단

[0008] 본 발명의 일 측면에 따르면 트랜스포머 기반 깊이 추정 장치가 제공된다.

[0009] 본 발명의 일 실시예에 따르면, 입력 영상을 겹치는 부분들이 없도록 복수의 패치들로 나누는 분할부; 및 상기 복수의 패치들을 트랜스포머 기반 깊이 추정 모델에 적용하여 깊이 맵을 생성하는 깊이 추정부를 포함하되, 상기 트랜스포머 기반 깊이 추정 모델은, 계층 구조를 가지는 n개의 스윈 트랜스포머 인코더로 구성되며, 상기 n개의 스윈 트랜스포머 인코더를 통해 상기 복수의 패치들이 통과되어 다중 스케일 인코더 특징맵을 생성하는 인코더부; 및 복수의 디코더 모듈을 가지되, 상기 복수의 디코더 모듈 중 적어도 일부를 통해 상기 인코더 특징맵과 이전 디코더 모듈의 디코더 특징맵을 융합하여 다중 스케일의 디코더 특징맵을 생성하고, 상기 다중 스케일의 디코더 특징맵을 이용하여 다중 스케일의 깊이 맵을 생성하는 디코더부를 포함하는 것을 특징으로 하는 트랜스포머 기반 깊이 추정 장치가 제공될 수 있다.

[0010] 상기 복수의 디코더 모듈 중 최후단에 위치된 디코더 모듈은 n개의 스윈 트랜스포머 인코더 중 최후단에 위치된 스윈 트랜스포머 인코더에 의해 출력되는 인코더 특징맵($\bar{F}_4^E \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32}}$)을 입력받은 후 잔차 블록을 적용하고 업샘플링한 후 1 x 1 컨볼루션 레이어를 적용하여 디코더 특징맵($F_4^D \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$)을 생성하되, 상기 디코더 특징맵($F_4^D \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$)은 상기 최후단에 위치된 디코더 모듈의 전단에 위치된 디코더 모듈

로 입력될 수 있다.

[0011] 상기 복수의 디코더 모듈 중 적어도 일부는 각각, 동일 해상도의 인코더 특징맵과 후단에 위치된 디코더 모듈의 디코더 특징맵을 각각 페어(pair)로 입력받은 후 상기 입력된 인코더 특징맵과 디코더 특징맵을 융합하고, 상기 입력된 인코더 특징맵을 이용하여 생성된 픽셀 레벨 주의 맵을 상기 융합된 특징맵에 합산하여 디코더 특징맵을 생성하는 어텐션 모듈; 및 상기 디코더 특징맵에 2개의 컨볼루션 레이어를 적용하여 깊이 맵을 생성하는 깊이 맵 생성 모듈을 포함할 수 있다.

[0012] 상기 픽셀 레벨 주의 맵은, 상기 인코더 특징맵에 잔차 블록을 적용하고 1 x 1 컨볼루션 레이어를 적용한 후 소프트맥스(softmax)를 적용하여 생성될 수 있다.

[0013] 상기 디코더 특징맵은 상기 인코더 특징맵과 상기 이전 디코더 모듈의 디코더 특징맵을 요소별 합산한 제1 특징맵과 상기 픽셀 레벨 주의 맵을 요소별 곱셈 연산하여 제2 특징맵을 생성한 후 상기 제2 특징맵에 잔차 블록을 적용한 후 상기 제1 특징맵과 합산하여 제3 특징맵을 생성한 후 1 x 1 컨볼루션 레이어를 적용한 후 바이리니어 업샘플링하여 생성될 수 있다.

[0014] 상기 다중 스케일의 깊이 맵 중 최고 해상도의 깊이 맵은 정답 라벨(Pseudo Label)로 활용되되, 상기 최고 해상도 깊이 맵은 광도 손실(Photometric loss)을 적용하고, 상기 정답 라벨을 제외한 깊이 맵은 상기 정답 라벨과의 자가 증류 손실(Self-distillation loss)을 적용하여 상기 트랜스포머 기반 깊이 추정 모델이 학습될 수 있다.

[0015] 상기 자가 증류 손실은 하기 수학식을 이용하여 계산되되,

[0016]
$$\mathcal{L}_{sd} = \lambda_{adp} * \frac{1}{3}\sum_{i=1}^{3} |upsample(D_i, 2^i) - D_0|_1 * M$$

$$\lambda_{adp} = \begin{cases} epoch/epoch_{thr}, & epoch < epoch_{thr} \\ 1, & othersize \end{cases}$$

[0017] 여기서, 로 적응적 가중치를 나타내며, D_i 는 깊이 맵을 나타내고, i는 깊이 맵 인덱스를 나타내며, D_0 는 정답 라벨을 나타내고, M은 마스크를 나타낸다.

[0019] 본 발명의 다른 측면에 따르면, 트랜스포머 기반 깊이 추정 방법이 제공된다.

[0020] 본 발명의 일 실시예에 따르면, (a) 입력 영상을 겹치는 부분들이 없도록 복수의 패치들로 나누는 단계; (b) 상기 복수의 패치들을 계층 구조를 가지는 n개의 스윈 트랜스포머(Swin transformer) 인코더에 적용하여 n개의 다중 스케일 인코더 특징맵을 생성하는 단계; 및 (c) 계층 구조를 가지는 복수의 디코더에 상기 인코더 특징맵과 이전 디코더의 디코더 특징맵을 적용하여 융합하여 다중 스케일의 디코더 특징맵을 생성하고, 상기 다중 스케일의 디코더 특징맵을 이용하여 다중 스케일의 깊이 맵을 생성하는 단계를 포함하되, 상기 복수의 디코더는 각각, 동일한 해상도의 인코더 특징맵을 입력받으며, 상기 다중 스케일의 깊이 맵 중 두번째 고해상도의 깊이 맵이 깊이 추론에 이용되는 것을 특징으로 하는 트랜스포머 기반 깊이 추정 방법이 제공될 수 있다.

[0021] 상기 (c) 단계에서, 상기 복수의 디코더 모듈 중 최후단에 위치된 디코더 모듈은 n개의 스윈 트랜스포머 인코더 중 최후단에 위치된 스윈 트랜스포머 인코더에 의해 출력되는 인코더 특징맵($\bar{F}_4^E \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32}}$)을 입력받은 후 잔차 블록을 적용하고 업샘플링한 후 1 x 1 컨볼루션 레이어를 적용하여 디코더 특징맵

($F_4^D \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$)을 생성하되, 상기 디코더 특징맵($F_4^D \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$)은 상기 최후단에 위치된 디코더 모듈의 전단에 위치된 디코더 모듈로 입력될 수 있다.

발명의 효과

[0023] 본 발명의 일 실시예에 따른 트랜스포머 기반 깊이 추정 장치 및 방법을 제공함으로써, 국지적인 영역과 전역적인 영역 모두 정확하고 일관성 있는 깊이 추정이 가능하다.

도면의 간단한 설명

[0025] 도 1은 본 발명의 일 실시예에 따른 트랜스포머 기반 깊이 추정 장치의 내부 구성을 개략적으로 도시한 블록도.

도 2는 본 발명의 일 실시예에 따른 트랜스포머 기반 깊이 추정 모델의 상세 구조를 도시한 도면.

도 3은 본 발명의 일 실시예에 따른 어텐션 모듈의 상세 구조를 도시한 도면.

도 4는 픽셀 레벨 주의 맵의 적용 유무에 따른 깊이 추정 결과를 비교한 도면.

도 5는 본 발명의 일 실시예에 따른 트랜스포머 기반 깊이 추정 방법을 나타낸 순서도.

도 6은 종래와 본 발명의 일 실시예에 따른 깊이 추정 결과를 비교한 도면.

발명을 실시하기 위한 구체적인 내용

[0026] 본 명세서에서 사용되는 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "구성된다" 또는 "포함한다" 등의 용어는 명세서상에 기재된 여러 구성 요소들, 또는 여러 단계들을 반드시 모두 포함하는 것으로 해석되지 않아야 하며, 그 중 일부 구성 요소들 또는 일부 단계들은 포함되지 않을 수도 있고, 또는 추가적인 구성 요소 또는 단계들을 더 포함할 수 있는 것으로 해석되어야 한다. 또한, 명세서에 기재된 "...부", "모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어 또는 소프트웨어로 구현되거나 하드웨어와 소프트웨어의 결합으로 구현될 수 있다.

[0027] 이하, 첨부된 도면들을 참조하여 본 발명의 실시예를 상세히 설명한다.

[0029] 도 1은 본 발명의 일 실시예에 따른 트랜스포머 기반 깊이 추정 장치의 내부 구성을 개략적으로 도시한 블록도이고, 도 2는 본 발명의 일 실시예에 따른 트랜스포머 기반 깊이 추정 모델의 상세 구조를 도시한 도면이고, 도 3은 본 발명의 일 실시예에 따른 어텐션 모듈의 상세 구조를 도시한 도면이고, 도 4는 픽셀 레벨 주의 맵의 적용 유무에 따른 깊이 추정 결과를 비교한 도면이다.

[0030] 도 1을 참조하면, 본 발명의 일 실시예에 따른 트랜스포머 기반 깊이 추정 장치(100)는 분할부(110), 학습부(120), 깊이 추정부(130), 메모리(140) 및 프로세서(150)를 포함하여 구성된다.

[0031] 분할부(110)는 입력 영상을 복수의 패치로 분할하기 위한 수단이다. 분할부(110)는 입력 영상을 중복 없이 복수의 패치들로 분할할 수 있다.

[0032] 학습부(120)는 트랜스포머 기반 깊이 추정 네트워크 모델을 학습하기 위한 수단이다.

[0033] 이해와 설명의 편의를 도모하기 위해, 트랜스포머 기반 깊이 추정 네트워크 모델에 대해 도 2를 참조하여 보다 상세히 설명하기로 한다.

[0034] 트랜스포머 기판 깊이 추정 네트워크 모델(200)은 인코더부(210)와 디코더부(220)를 포함한다. 여기서, 인코더부(210)는 도 2에서 보여지는 바와 같이, 계층 구조를 가지는 스윈 트랜스포머(Swin Transformer) 기반 인코더로 구성될 수 있다. 예를 들어, 인코더부(210)는 4개의 계층 구조를 가지는 스윈 트랜스포머 기반 인코더로 구성될 수 있다. 이와 같은 경우, 인코더부(210)는 도 2에 도시된 바와 같이, 4개의 스윈 트랜스포머 인코더를 포함할 수 있다.

[0035] 입력 영상은 겹치는 부분이 없도록 패치들로 나뉘며, 처음 생성된 패치는 픽셀을 16개 포함하는 4 x 4 크기의 작은 영상일 수 있다. 학 픽셀당 RGB값 3개를 가지고 있으므로, 처음 생성된 패치 하나는 4 x 4 x 3 크기일 수 있으며, 이렇게 나뉜 패치들은 토큰처럼 고려될 수 있으며, 패치의 수는 $\frac{H}{4} \times \frac{W}{4}$만큼 생성될 수 있다. 각각의 패치(law patch)는 선형 임베딩 레이어(linear Embedding layer)를 통과하며, 선형 임베딩 레이어를 통과하여 4 x 4 x C 크기를 갖게 된다. 즉 선형 임베딩 레이어를 통과하며 채널 차원이 증가될 수 있다. 채널 차원이 증가된 패치들은 스윈 트랜스포머 인코더()을 통과하며 다중 스케일의 특징맵이 생성될 수 있다. 특징맵의 해상도는 각각의 계층(단계)를 통과할 때마다 2배만큼 다운샘플링될 수 있다.

[0036] 이와 같은 인코더부(210)의 특징맵(편의상 인코더 특징맵이라 칭하기로 함)은 디코더부(220)로 입력되어 깊이 추정을 위한 디코더 특징맵을 생성하는데 이용될 수 있다.

[0037] 인코더부(210)를 통과한 후 마지막 인코더 특징맵($\bar{F}_4^E \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32}}$)은 잔차 블록을 이후 업샘플링되어 디코더 특징맵($F_4^D \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$)을 형성하기 위해 다른 1 x 1 컨볼루션이 적용될 수 있다.

[0038] 디코더부(220)는 복수의 디코더 모듈을 가지되, 각각의 디코더 모듈은 인코더 특징맵과 이전 디코더 모듈의 특징맵을 이용하여 깊이 맵을 생성하기 위한 수단이다.

[0039] 각각의 디코더 모듈은 두개의 스트림으로 진행된다. 즉, 각 디코더 모듈은 인코더 특징맵과 이전 디코더 모듈의 특징맵(디코더 특징맵)을 융합하여 스윈 트랜스포머 인코더에서 추출된 전역 정보를 보충하고, 인코더 특징맵을 입력으로 픽셀 레벨 주의 맵을 생성하여 전역적 정보만 포함된 디코더 특징맵에 국소적 정보를 향상시킬 수 있다.

[0040] 도 2를 참조하여 디코더부(220)에 대해 보다 상세히 설명하기로 한다.

[0041] 인코더부(210)에 의해 출력된 마지막 인코더 특징맵($\bar{F}_4^E \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32}}$)은 잔차 블록 이후 업샘플링된 후 1 x 1 컨볼루션이 적용되어 디코더 특징맵($F_4^D \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$)이 생성된다.

[0042] 이후, 각 디코더 모듈에는 인코더 특징맵(F_{n-1}^E)과 이전 디코더 모듈의 디코더 특징맵(F_n^D)이 입력되어 디코더 특징맵(F_{n-1}^D)을 생성할 수 있다. 이러한 디코더 특징맵(F_{n-1}^D)은 다음 디코더 모듈의 입력으로 사용되며, 동시에 깊이 맵 생성을 위해 사용될 수 있다.

[0043] 예를 들어, 인코더부(210)가 제1 스윈 트랜스포머 인코더, 제2 스윈 트랜스포머 인코더, 제3 스윈 트랜스포머 인코더, 제4 스윈 트랜스포머 인코더를 가지며, 디코더부(220)가 제1 디코더 모듈, 제2 디코더 모듈, 제3 디코더 모듈, 제4 디코더 모듈 및 제5 디코더 모듈을 가지는 것을 가정하기로 한다.

[0044] 제4 스윈 트랜스포머 인코더를 통해 출력된 인코더 특징맵은 제5 디코더 모듈에 의해 잔차 블록이 적용된 이후 업샘플링되고, 1 x 1 컨볼루션 레이어가 적용되어 디코더 특징맵이 출력될 수 있다.

[0045] 제4 디코더 모듈에는 제3 스윈 트랜스포머 인코더의 출력값(인코더 특징맵)과 제5 디코더 모듈의 출력값(디코더 특징맵)이 입력되어 융합되고, 인코더 특징맵을 잔차 블록을 적용하고 1 x 1 컨볼루션 레이어를 적용한 후 소프트맥스(Softmax)를 적용하여 픽셀 레벨 주의 맵을 생성한 후 융합된 특징맵과 합산하여 디코더 특징맵이 도출될 수 있다. 마찬가지로, 제3 디코더 모듈에는 제2 스윈 트랜스포머 인코더의 출력값과 제4 디코더 모듈의 출력값이 입력되고, 제2 디코더 모듈에는 제1 스윈 트랜스포머 인코더의 출력값과 제3 디코더 모듈의 출력값이 입력될 수 있다. 마지막으로, 제1 디코더 모듈에는 제1 스윈 트랜스포머 인코더의 출력값을 업 샘플링하여 입력하고, 제2 디코더 모듈의 출력값이 입력될 수 있다.

[0046] 각 디코더 모듈의 상세 기능은 하기의 설명에 의해 보다 명확하게 이해될 것이다.

[0047] 각 디코더 모듈은 어텐션 모듈(222)과 깊이 맵 생성 모듈(224)로 구성될 수 있다.

[0048] 어텐션 모듈(222)은 인코더 특징맵과 이전 디코더 모듈의 디코더 특징맵을 이용하여 디코더 특징맵을 생성하기 위한 수단이다.

[0049] 계층 구조를 가지는 복수의 스윈 트랜스포머 인코더를 포함하는 인코더부(210)는 로컬 상세 정보가 부족하기 때문에 조밀한 깊이 추정에 적합하지 않다. 따라서, 어텐션 모듈(222)이 풍부한 전역적 정보와 국소적 정보(컨텍스트)를 포함하는 디코더 특징맵을 생성할 수 있도록 픽셀 레벨 스킵(pixel-wise skip)을 통한 픽셀 레벨 주의 맵을 생성하고, 이를 활용하여 디코더 특징 맵을 생성할 수 있다.

[0050] 도 3을 참조하여 이에 대해 보다 상세히 설명하기로 한다.

[0051] 어텐션 모듈(222)은 도 3에서 보여지는 바와 같이, 두개의 스트림으로 진행될 수 있다.

[0052] 우선, 어텐션 모듈(222)은 인코더 특징맵(F_{n-1}^{E})과 이전 디코더 특징맵(F_{n}^{D})을 융합하여 스윈 트랜스포머에서 추출된 전역적 정보를 보충한 요소별 합산 특징맵(S_{add})을 생성할 수 있다.

[0053] 또한, 어텐션 모듈(222)은 인코더 특징맵(F_{n-1}^{E})에 잔차 블록을 적용하고 1 x 1 컨볼루션 레이어를 적용한 후 소프트맥스(Softmax)를 적용함으로써 위치 인식 정보를 잘 인코딩하는 것으로 알려진 픽셀 레벨 주의 맵(S_{att})을 생성할 수 있다.

[0054] 어텐션 모듈(222)은 이와 같이 두 개의 스트림으로 진행되어 생성된 요소별 합산 특징맵과 픽셀 레벨 주의 맵에 대해 요소별 곱셈 연산을 적용하여 공간적으로 세부 정보가 향상된 특징맵(S_{scaled})을 생성할 수 있다.

[0055] 이를 수학식으로 나타내면, 수학식 1과 같이 나타낼 수 있다.

수학식 1

[0056]
$$S_{add} = F_{n-1}^{E} + F_{n}^{D}$$
$$S_{att} = (softmax \circ embedding \circ resblock)(F_{n-1}^{E})$$
$$S_{scaled} = S_{add} \odot S_{att}$$

[0057] 여기서, resblock는 각각 ReLU 활성화가 뒤따르는 두개의 3x3 컨볼루션 레이어로 구성되며, embedding은 1 x 1 컨볼루션 레이어로 구성되고, Á는 요소별 곱셈(Hadamard 곱)을 나타낸다.

[0058] 초기 단계에서 리스케일된 특징맵(S_{scaled})이 지나치게 강조되거나 무시되는 것을 완화시키기 위해, 어텐션 모듈(222)운 리스케일된 특징맵(S_{scaled})에 또 다른 잔차 블록(resblock)을 적용한 후 요소별 합산 특징맵(S_{add})을 추가하여 향상된 특징맵($\tilde{S}$)을 생성할 수 있다.

[0059] 디코더 모듈은 향상된 특징맵($\tilde{S}$)에 1 x 1 컨볼루션 레이어를 두번 적용하여 바이리니어 업-샘플링된 디코더 특징맵(F_{n-1}^{D})을 생성할 수 있다.

[0060] 이를 수학식으로 나타내면, 수학식 2와 같다.

수학식 2

$$\tilde{S} = resblock(S_{scaled}) + S_{add}$$
$$F_{n-1}^{D} = (embedding \circ upsample_2)(\tilde{S})$$

[0061]

[0062] 도 4는 디코더 모듈에서 픽셀 레벨 주의 맵 적용 유무에 따른 특징맵을 비교한 결과이다. 410은 입력 영상이며, 420은 픽셀 레벨 주의 맵이 적용되지 않은 특징맵이며, 430은 픽셀 레벨 주의 맵이 적용된 특징맵이다.

[0063] 도 4의 420에서 보여지는 바와 같이, 픽셀 레벨 주의 맵이 적용되지 않은 특징맵은 전역적으로 활성화되며 객체의 경계가 명확하게 표시되지 않는 것을 알 수 있다.

[0064] 그러나, 도 4의 430에서 보여지는 바와 같이, 픽셀 레벨 주의 맵이 적용된 경우, 특징맵의 객체 경계는 배경 픽셀과 구별되는 것을 알 수 있다. 이와 같이, 픽셀 레벨 주의 맵을 적용함으로써 세부 사항이 개선되는 것을 알 수 있다.

[0065] 깊이 맵 생성 모듈(224)는 각 어텐션 모듈(222)에서 출력되는 디코더 특징맵을 이용하여 깊이 맵을 생성하기 위한 수단이다.

[0066] 이에 대해 보다 상세히 설명하기로 한다.

[0067] 깊이 맵 생성 모듈(224)는 2개의 3 x 3 컨볼루션 레이어와 디스패리티 이산 볼륨으로 구성될 수 있다.

[0068] 학습부(120)는 디코더부(220)의 추가적인 디코더 모듈에서 출력된 최고해상도 깊이 맵을 정답 라벨(Pseudo Label)로 활용하여 각 깊이 맵 생성 모듈(224)에서 생성된 깊이 맵 결과와 비교하는 자가 증류 손실(self-distillation loss)을 적용하여 트랜스포머 기반 깊이 추정 네트워크 모델을 학습시킬 수 있다.

[0069] 이에 대해 보다 상세히 설명하기로 한다.

[0070] 복잡성에 크게 기여하는 가장 큰 해상도의 깊이 예측은 일반적으로 가장 좋은 성능을 보여준다. 또한, 가장 큰 해상도의 역 깊이(inverse depth)에서 미세한 상세 정보(detail)을 예측하는 능력을 증류에 의해 끝에서 두번째 예측으로 이전할 수 있다.

[0071] 본 발명의 일 실시예에 따른 자가 증류 손실은 스윈 트랜스포머 인코더 훈련의 불안정성을 완화하기 위해 정답 레이블의 올바른 신호로 중간 특성맵을 훈련하는데 도움이 된다.

[0072] 자가 증류 손실($\mathcal{L}_{sd}$)을 계산하기 위해, 학습부(120)는 정답 레이블(D0)과의 해상도를 일치시키기 위해 정답 레이블(D0)의 해상도를 제외하고 깊이 맵을 업샘플링할 수 있다.

[0073] 이어, 업샘플링된 깊이 맵과 의사 레이블간의 L_1 손실이 계산될 수 있다. 학습 안정성을 위해, 자동 마스크(M)가 사용될 수 있다. 또한, 학습 초기 단계에서 정답 레이블의 부정확성으로 인한 원치 않는 오류 전파를 완화하기 위해 학습 진행에 따라 손실 가중치를 점진적으로 제어하는 적응적 가중치(λ_{adp})가 적용될 수 있다.

[0074] 이를 수학식으로 나타내면, 수학식 3과 같이 나타낼 수 있다.

수학식 3

$$\mathcal{L}_{sd} = \lambda_{adp} * \frac{1}{3}\sum_{i=1}^{3} |upsample(D_i, 2^i) - D_0|_1 * M$$

[0075]

$$\lambda_{adp} = \begin{cases} epoch/epoch_{thr}, & epoch < epoch_{thr} \\ 1, & othersize \end{cases}$$

[0076] 여기서, 로 적응적 가중치를 나타내며, D_i는 다른 깊이 맵을 나타내고, i는 깊이 맵 인덱스를 나타내며, D_0는 정답 라벨을 나타내고, M은 마스크를 나타낸다.

[0077] 다시 정리하면, 최고 해상도 깊이 맵은 광도 손실(Photometric loss)을 적용하고, 최고 해상도 깊이 맵(즉, 정답 라벨)을 제외한 나머지 깊이 맵은 정답 라벨과의 자가 증류 손실(Self-distillation loss)을 적용하여 트랜스포머 기반 깊이 추정 네트워크 모델이 학습될 수 있다.

[0078] 깊이 추정부(130)는 학습된 적용하여 트랜스포머 기반 깊이 추정 네트워크 모델에 입력 영상을 적용하여 깊이 맵을 생성한다. 트랜스포머 기반 깊이 추정 네트워크 모델의 구성 및 동작은 도 2 내지 도 3을 참조하여 설명한 바와 동일하므로 중복되는 설명은 생략하기로 한다.

[0079] 전술한 바와 같이, 다중 스케일 깊이 맵 중 최고 해상도 깊이 맵은 정답 라벨로 활용되며, 두번째 고해상도 깊이 맵이 깊이 추정에 이용될 수 있다.

[0080] 깊이 맵을 생성하는 방법 자체는 전술한 바와 동일하므로 중복되는 설명은 생략하기로 한다.

[0081] 메모리(140)는 본 발명의 일 실시예에 따른 트랜스포머 기반 깊이 추정 방법을 수행하기 위해 필요한 명령어들을 저장한다.

[0082] 프로세서(150)는 본 발명의 일 실시예에 따른 트랜스포머 기반 깊이 추정 장치(100)의 내부 구성 요소들(예를 들어, 분할부(110), 학습부(120), 깊이 추정부(130), 메모리(140) 등)을 제어하기 위한 수단이다.

[0084] 도 5는 본 발명의 일 실시예에 따른 트랜스포머 기반 깊이 추정 방법을 나타낸 순서도이고, 도 6은 종래와 본 발명의 일 실시예에 따른 깊이 추정 결과를 비교한 도면이다.

[0085] 단계 510에서 트랜스포머 기반 깊이 추정 장치(100)는 입력 영상을 중복 없이 복수의 패치들로 분할한다.

[0086] 단계 515에서 트랜스포머 기반 깊이 추정 장치(100)는 복수의 패치들을 트랜스포머 기반 깊이 추정 네트워크 모델에 적용하여 다중 스케일 인코더 특징맵을 생성한다.

[0087] 예를 들어, 트랜스포머 기반 깊이 추정 장치(100)는 분할된 복수의 패치들을 계층 구조를 가지는 n개의 스윈 트랜스포머(Swin transformer) 인코더에 적용하여 n개의 다중 스케일 인코더 특징맵을 생성할 수 있다.

[0088] 단계 520에서 트랜스포머 기반 깊이 추정 장치(100)는 트랜스포머 기반 깊이 추정 네트워크 모델을 통해 인코더 특징맵과 이전 디코더의 디코더 특징맵을 적용하여 융합하여 다중 스케일의 디코더 특징맵을 생성하고, 다중 스케일의 디코더 특징맵을 이용하여 다중 스케일의 깊이 맵을 생성한다.

[0089] 예를 들어, 트랜스포머 기반 깊이 추정 장치(100)는 n개의 스윈 트랜스포머 인코더 중 최후단에 위치된 스윈 트랜스포머 인코더에 의해 출력되는 인코더 특징맵($F_4^E \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}}$)을 최후단에 위치된 디코더 모듈에 입력

하여 잔차 블록을 적용하고 업샘플링한 후 1 x 1 컨볼루션 레이어를 적용하여 디코더 특징맵($F_4^D \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$)을 생성할 수 있다. 이어, 트랜스포머 기반 깊이 추정 장치(100)는 디코더 특징맵($F_4^D \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$)을 전단에 위치된 디코더 모듈로 입력할 수 있다.

[0090] 이어, 각 디코더 모듈은 도 2에 도시된 바와 같이, 동일 해상도의 인코더 특징맵과 이전 디코더 모듈의 특징맵을 입력받은 후 이를 융합하고, 인코더 특징맵에 잔차 블록을 적용하고 1 x 1 컨볼루션 레이어를 적용한 후 소프트맥스(softmax)를 적용하여 픽셀 레벨 주의 맵을 생성한 후 이를 융합된 특징맵에 반영하여 디코더 특징맵을 각각 생성할 수 있다.

[0091] 또한, 트랜스포머 기반 깊이 추정 장치(100)는 디코더 특징맵에 2개의 3 x 3 컨볼루션 레이어를 적용하여 깊이 맵을 생성할 수 있다.

[0092] 또한, 트랜스포머 기반 깊이 추정 장치(100)는 다중 스케일의 깊이 맵 중 최고 해상도의 깊이 맵은 정답 라벨(Pseudo Label)로 활용되되, 최고 해상도 깊이 맵은 광도 손실(Photometric loss)을 적용하고, 상기 정답 라벨을 제외한 깊이 맵은 상기 정답 라벨과의 자가 증류 손실(Self-distillation loss)을 적용하여 트랜스포머 기반 깊이 추정 모델을 학습시킬 수 있다.

[0093] 도 6은 종래와 본 발명의 일 실시예에 따른 깊이 추정 결과를 비교한 도면이다.

[0094] 도 6에서 보여지는 바와 같이, 본 발명의 일 실시예에 따른 트랜스포머 기반 깊이 추정 모델이 로컬 세부 사항을 유지하면서 트랜스포머 기반 인코더에서 전역 컨텍스트를 효과적으로 집계하여 첫번째 행의 오른쪽에 있는 보행자의 모양을 정확하게 도출하는 것을 알 수 있다.

[0095] 또한, 본 발명의 일 실시예에 따른 트랜스포머 기반 깊이 추정 모델이 두번째 및 세번째 행에서 유조선과 열차의 형상을 성공적으로 예측하는 것을 알 수 있다.

[0096] 즉, 본 발명의 일 실시예에 따른 트랜스포머 기반 깊이 추정 모델이 광도 손실과 자가 증류 손실로 훈련됨에 따라 정확하고 조밀한 깊이 맵을 추정하는 것을 알 수 있다.

[0098] 본 발명의 실시 예에 따른 장치 및 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 컴퓨터 판독 가능 매체에 기록되는 프로그램 명령은 본 발명을 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 분야 통상의 기술자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media) 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0099] 상술한 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0100] 이제까지 본 발명에 대하여 그 실시 예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시 예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

부호의 설명

[0102] 100: 트랜스포머 기반 깊이 추정 장치

110: 분할부

120: 학습부

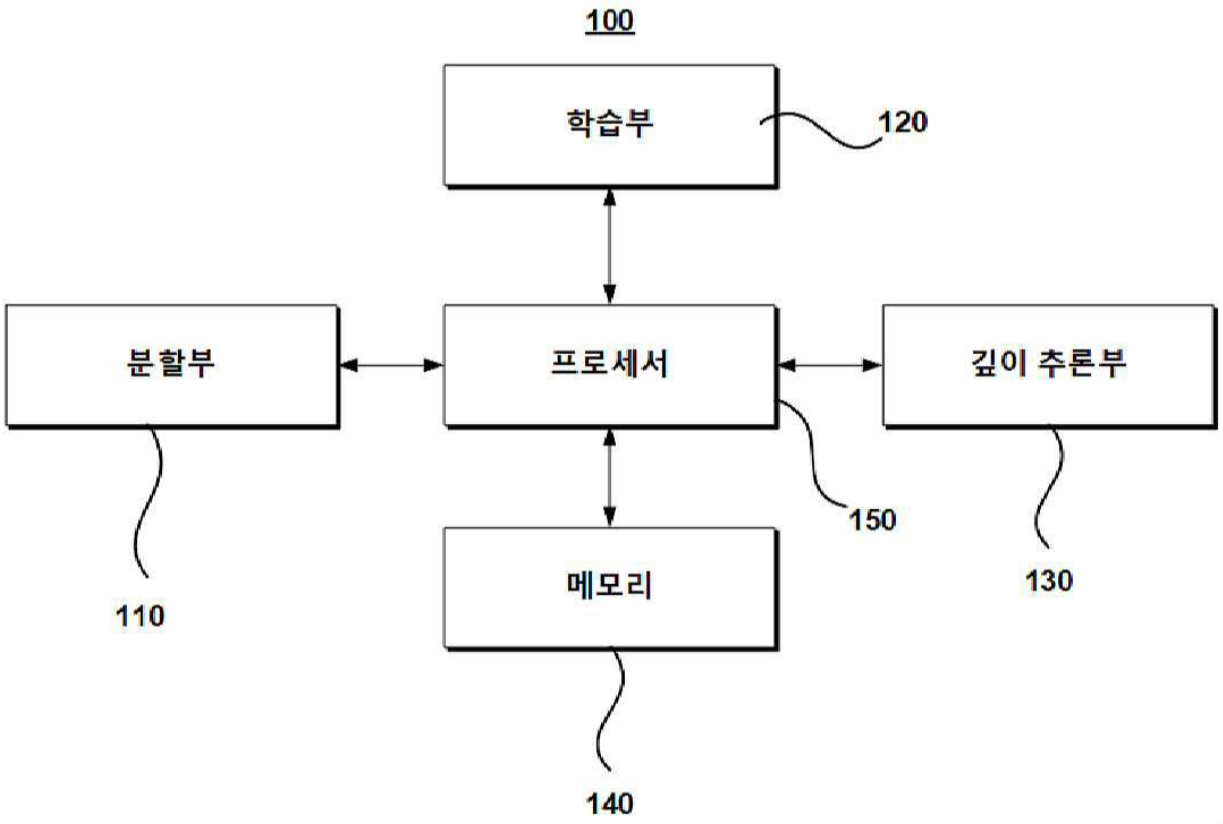
130: 깊이 추정부

140: 메모리

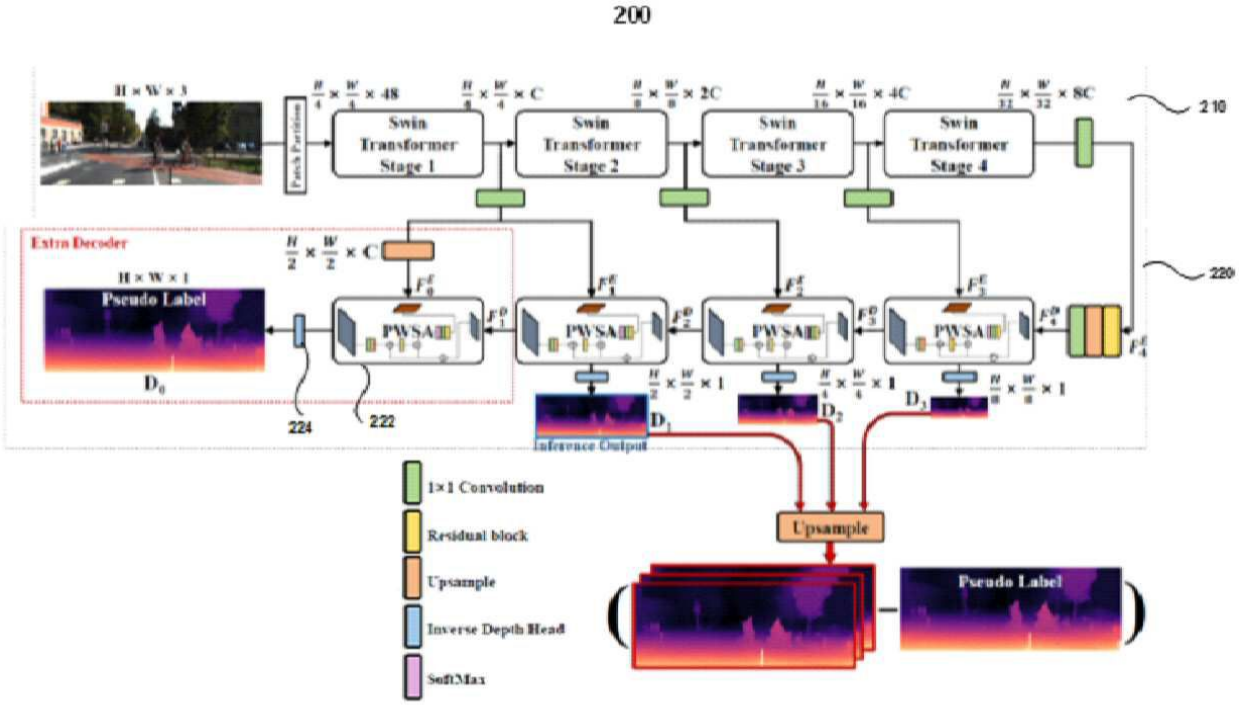
150: 프로세서

도면

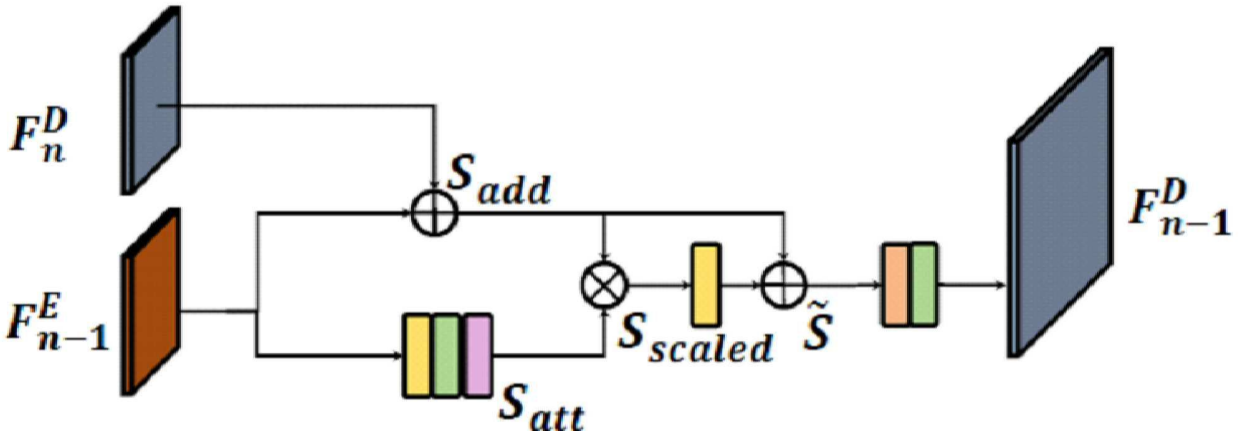
도면1



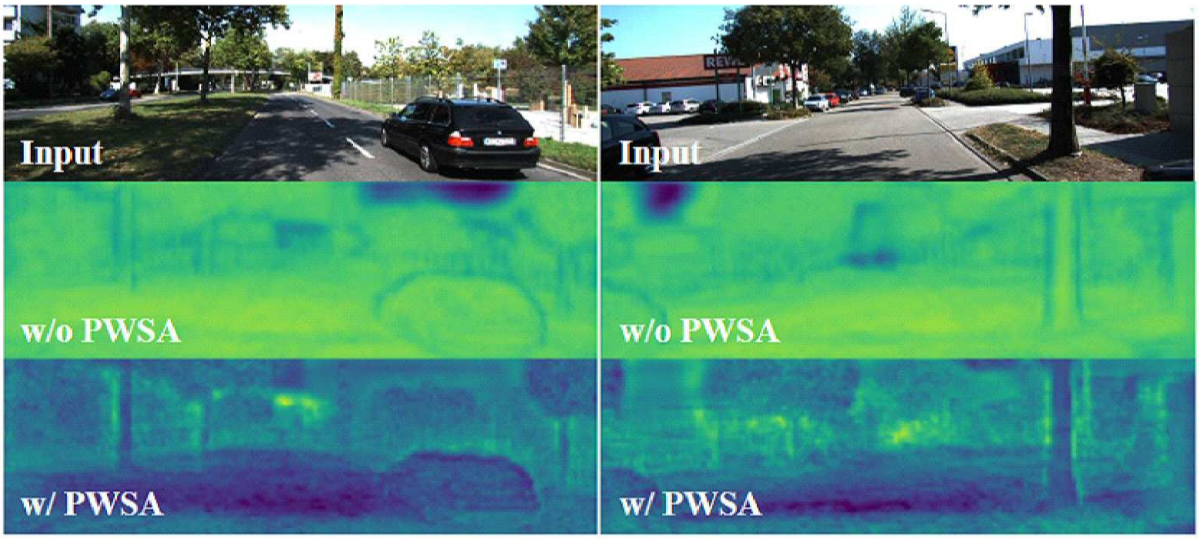
도면2


200
Swin Transformer Stage 1
Swin Transformer Stage 2
Swin Transformer Stage 3
Swin Transformer Stage 4
210
220
Extra Decoder
Pseudo Label
PWSA
224
222
Inference Output
1×1 Convolution
Residual block
Upsample
Inverse Depth Head
SoftMax
Upsample
Pseudo Label

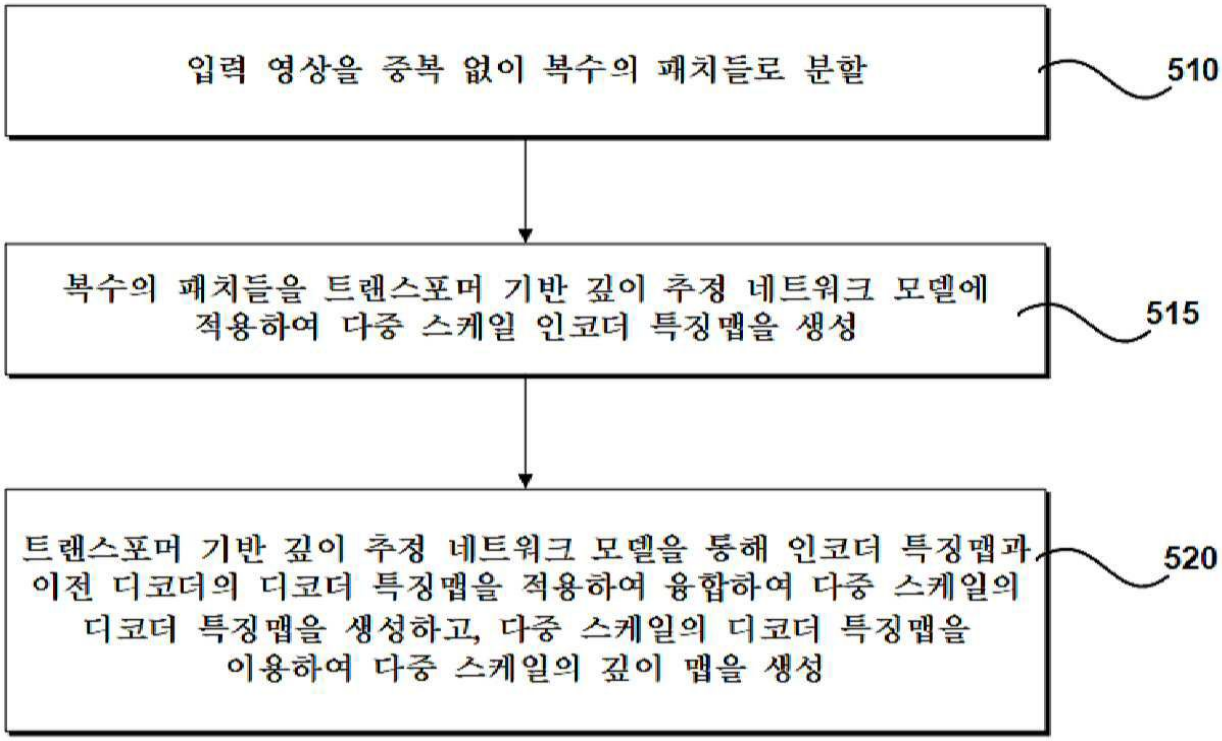
도면3


F_n^D
F_{n-1}^E
S_{add}
S_{scaled}
$\tilde{S}$
S_{att}
F_{n-1}^D

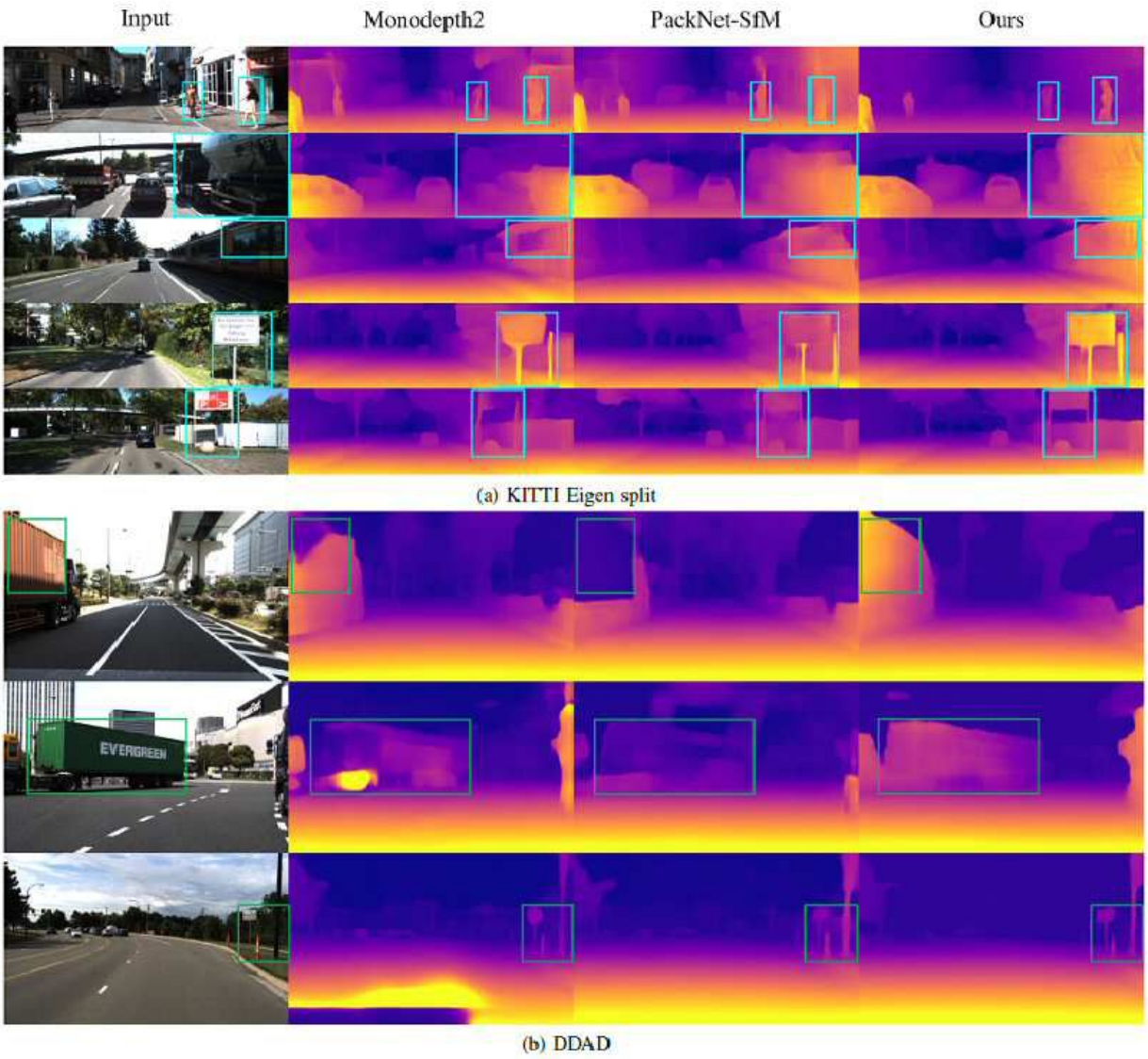
도면4


Input
Input
w/o PWSA
w/o PWSA
w/ PWSA
w/ PWSA

도면5


입력 영상을 중복 없이 복수의 패치들로 분할
510
복수의 패치들을 트랜스포머 기반 깊이 추정 네트워크 모델에 적용하여 다중 스케일 인코더 특징맵을 생성
515
트랜스포머 기반 깊이 추정 네트워크 모델을 통해 인코더 특징맵과 이전 디코더의 디코더 특징맵을 적용하여 융합하여 다중 스케일의 디코더 특징맵을 생성하고, 다중 스케일의 디코더 특징맵을 이용하여 다중 스케일의 깊이 맵을 생성
520

도면6



(a) KITTI Eigen split

(b) DDAD