(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0107889
(43) 공개일자 2024년07월09일

(51) 국제특허분류(Int. Cl.)
G06T 5/00 (2024.01) *G06F 18/214* (2023.01)
G06N 3/0455 (2023.01) *G06N 3/092* (2023.01)
G06T 3/40 (2024.01) *G06T 5/50* (2024.01)
G06T 7/11 (2017.01)
(52) CPC특허분류
G06T 5/77 (2024.01)
G06F 18/214 (2023.01)
(21) 출원번호 10-2022-0190941
(22) 출원일자 2022년12월30일
심사청구일자 없음

(71) 출원인
서강대학교산학협력단
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(72) 발명자
강석주
경기도 고양시 일산서구 후곡로 36, 403동 1003호 (일산3동, 후곡마을4단지아파트)
김지현
서울특별시 마포구 백범로 35 (신수동, 서강대학교)
(74) 대리인
유미특허법인

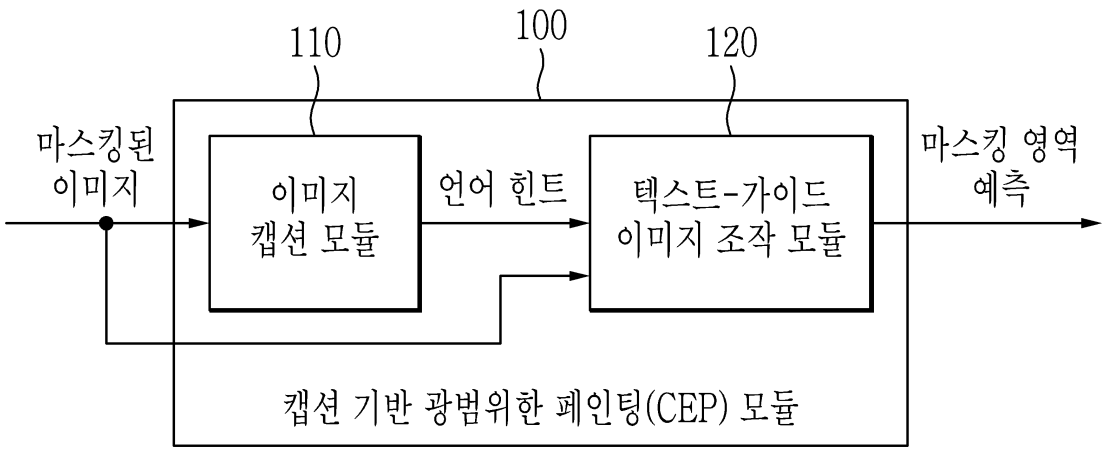
전체 청구항 수 : 총 15 항

(54) 발명의 명칭 **인공신경망 기반의 이미지 캡션 생성을 통한 이미지 확장 방법 및 그 장치**

(57) 요 약

본 개시에 따르면, 캡션 기반 광범위한 페인팅 작업을 수행하는 컴퓨팅 장치로서, 마스킹 처리된 영역을 포함하는 이미지를 입력받아 입력 이미지를 설명하는 상기 마스킹 처리된 영역에 대한 언어 힌트를 생성하는 이미지 캡션 모듈, 그리고 상기 입력 이미지와 상기 언어 힌트를 입력받아 상기 언어 힌트의 가이드에 따라 상기 마스킹 처리된 영역의 이미지를 예측하고, 예측한 이미지로 상기 마스킹 처리된 영역을 채운 확장 이미지를 출력하는 텍스트-가이드 이미지 조작 모듈을 포함할 수 있다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/0455 (2023.01)
G06N 3/092 (2023.01)
G06T 3/40 (2024.01)
G06T 5/50 (2024.01)
G06T 7/11 (2017.01)
G06T 2207/20081 (2013.01)
G06T 2207/20084 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711159124
과제번호 2020M3H4A1A02084899
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 나노미래소재원천기술개발사업
연구과제명 2축 신축성 AMLED용 TFT 및 LED 집적 기술
기 여 율 1/2
과제수행기관명 경희대학교 산학협력단
연구기간 2020.07.01~2024.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711164983
과제번호 2021R1A2C1004208
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초지원사업(중견연구)
연구과제명 극한 실외 환경 변화에 강인한 비디오 처리용 프로그래머블 초고계조 영상 생성 신경망 기술 및 FPGA 기반 하드웨어 가속기 개발
기 여 율 1/2
과제수행기관명 서강대학교 산학협력단
연구기간 2021.03.01~2024.02.29

명 세 서

청구범위

청구항 1

캡션 기반 광범위한 페인팅 작업을 수행하는 컴퓨팅 장치로서,

마스킹 처리된 영역을 포함하는 이미지를 입력받아 입력 이미지를 설명하는 상기 마스킹 처리된 영역에 대한 언어 힌트를 생성하는 이미지 캡션 모듈, 그리고

상기 입력 이미지와 상기 언어 힌트를 입력받아 상기 언어 힌트의 가이드에 따라 상기 마스킹 처리된 영역의 이미지를 예측하고, 예측한 이미지로 상기 마스킹 처리된 영역을 채운 확장 이미지를 출력하는 텍스트-가이드 이미지 조작 모듈

를 포함하는, 컴퓨팅 장치.

청구항 2

제1항에서,

상기 이미지 캡션 모듈은,

입력 이미지로부터 시각적 특징을 추출하는 인코더, 그리고

상기 시각적 특징으로부터 일련의 단어들을 생성하는 디코더

를 포함하는, 컴퓨팅 장치.

청구항 3

제1항에서,

상기 이미지 캡션 모듈은,

입력 이미지로부터 상기 입력 이미지를 설명하는 자연어로 이루어진 텍스트를 출력하도록 사전 학습된 언어 모델을 랜덤하게 마스킹된 이미지들로 구성된 학습 데이터를 이용하여 학습되는, 컴퓨팅 장치.

청구항 4

제3항에서,

상기 이미지 캡션 모듈은,

상기 학습 데이터와 SCST(self-critical sequence training) 방법을 사용하여 학습되는, 컴퓨팅 장치.

청구항 5

제3항에서,

상기 이미지 캡션 모듈은,

교차 엔트로피 손실(cross entropy loss)과 강화 학습(reinforcement learning)을 이용하여 학습되는, 컴퓨팅 장치.

청구항 6

제1항에서,

상기 텍스트-가이드 이미지 조작 모듈은,

텍스트와 이미지가 쌍으로 이루어진 학습 데이터를 이용하여 텍스트에 상응하는 이미지를 출력하도록 학습되는, 컴퓨팅 장치.

청구항 7

제1항에서,

상기 이미지 캡션 모듈과 상기 텍스트-가이드 이미지 조작 모듈은,

인공신경망 모델인 컴퓨팅 장치.

청구항 8

컴퓨팅 장치에 의해 수행되는 이미지 아웃페인팅 방법으로서,

임의의 영역이 마스킹 처리된 이미지를 입력받는 단계,

입력 이미지로부터 일련의 단어들을 생성하도록 학습된 이미지 캡션 모듈을 이용하여, 상기 마스킹된 이미지에서 마스킹 처리되지 않은 이미지에 대한 의미 및 텍스트 정보를 자연 언어로 설명한 언어 힌트를 생성하는 단계,

상기 언어 힌트를 이용하여 상기 임의의 영역의 이미지를 예측하는 단계, 그리고

상기 예측한 이미지를 포함하는 확장 이미지를 출력하는 단계

를 포함하는, 이미지 아웃 페인딩 방법.

청구항 9

제8항에서,

상기 예측하는 단계는,

입력 텍스트에 상응하는 이미지를 생성하도록 학습된 텍스트-가이드 이미지 조작 모듈에 상기 마스킹 처리된 이미지와 상기 언어 힌트를 입력하여, 상기 언어 힌트의 가이드에 따라 상기 마스킹 처리된 임의의 영역의 이미지를 예측하는, 이미지 아웃 페인딩 방법.

청구항 10

제9항에서,

상기 텍스트-가이드 이미지 조작 모듈은,

텍스트와 이미지가 쌍으로 이루어진 학습 데이터를 이용하여 텍스트에 상응하는 이미지를 출력하도록 학습된 인공신경망 모델인, 이미지 아웃 페인딩 방법.

청구항 11

제8항에서,

상기 이미지 캡션 모듈은,

사전 학습된 언어 모델이 랜덤하게 마스킹된 이미지들로 구성된 학습 데이터를 이용하여 학습된 인공신경망 모델이고,

상기 사전 학습된 언어 모델은,

입력 이미지로부터 상기 입력 이미지를 설명하는 자연어로 이루어진 텍스트를 출력하는 모델인, 이미지 아웃 페인팅 방법.

청구항 12

컴퓨팅 장치에 의해 수행되는 광범위한 이미지 블렌딩 방법으로서,

제1 이미지와 제2 이미지가 연결되는 지점을 마스킹 영역으로 설정하는 단계,

상기 마스킹 영역을 기준으로 양측에 상기 제1 이미지와 상기 제2 이미지가 각각 배치된 이미지를 생성하는 단계,

입력 이미지로부터 일련의 단어들을 생성하도록 학습된 이미지 캡션 모듈을 이용하여, 상기 생성한 이미지에서 마스킹 처리되지 않은 이미지에 대한 의미 및 텍스트 정보를 자연 언어로 설명한 언어 힌트를 생성하는 단계,

상기 언어 힌트를 이용하여 상기 마스킹 영역의 이미지를 예측하는 단계, 그리고

상기 제1 이미지, 상기 예측한 이미지, 및 상기 제2 이미지가 순차적으로 배치된 파노라마 이미지를 출력하는 단계

를 포함하는, 광범위한 이미지 블렌딩 방법.

청구항 13

제12항에서,

상기 설정하는 단계 이전에,

이전 단계에서 예측된 출력을 다음 단계의 입력으로 사용하는 이미지 아웃페인팅을 반복하여 상기 제1 이미지를 생성하는 단계, 그리고

상기 제1 이미지와 반대 방향으로 상기 이미지 아웃페인팅을 반복하여 제2 이미지를 생성하는 단계를 포함하고,

상기 이미지 아웃페인팅은,

마스킹 영역에 대한 이미지를 예측하는 작업인, 광범위한 이미지 블렌딩 방법.

청구항 14

제13항에서,

상기 이미지 아웃페인팅은,

상기 마스킹 영역이 포함된 이미지에서 마스킹되지 않은 영역에 대한 의미 및 텍스트 정보를 자연 언어로 설명한 언어 힌트를 사용하여 상기 마스킹 영역의 이미지를 예측하고, 상기 마스킹 영역을 예측한 이미지로 채운 확장 이미지를 생성하는 작업을 포함하는, 광범위한 이미지 블렌딩 방법.

청구항 15

제14항에서,

상기 이미지 아웃페인팅은,

텍스트에 상응하는 이미지를 출력하도록 학습된 인공신경망 모델에 상기 언어 힌트와 상기 마스킹 영역이 포함된 이미지를 입력하여 상기 언어 힌트의 가이드에 따라 상기 마스킹 영역의 이미지를 예측하는 작업을 포함하는, 광범위한 이미지 블렌딩 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공신경망 기반의 이미지 캡션 생성을 통한 이미지 확장 방법 및 그 장치에 관한 것이다.

배 경 기 술

[0002] 이미지 완성은 이미지에서 누락된 영역에 적절한 이미지를 생성하는 연구 분야로서, 대표적으로 이미지 아웃페인팅(Image outpainting)과 광역 이미지 블렌딩(wide-rage image blending)이 있다.

[0003] 이미지 아웃 페인팅은 주어진 이미지의 경계 밖을 생성하여 이미지를 확장하는 연구 분야이다. 이미지 아웃 페인팅은 원본 이미지 왜곡을 최소화하며 각기 다른 디스플레이 종횡비에 맞게 이미지 크기를 조정하는 이미지 리타겟팅(image retargeting)에 활용될 수 있다. 더불어, 이미지 아웃페인팅과 광역 이미지 블렌딩은 파노라마 이미지 생성에도 적용될 수 있다.

[0004] 광역 이미지 블렌딩은 각기 다른 두 이미지 사이를 채워 넣어 하나의 자연스러운 이미지를 생성하는 연구 분야이다.

[0005] 이미지 완성에서 두 과업이 처리하는 누락된 영역의 크기가 상대적으로 크기 때문에 난도가 높은 것으로 평가받고 있다.

[0006] 기존의 일부 딥러닝 기반 이미지 아웃페인팅 알고리즘들은 누락된 영역의 큰 크기를 보완하기 위하여 이미지 형태의 힌트를 생성하였다. 종래의 어떤 방법은 입력 이미지를 반으로 나눈 뒤 양옆을 서로 뒤바꾸어 힌트로 사용하고, 이와 비슷하게 종래의 다른 방법은 입력 이미지를 뒤집어 힌트로 사용한다. 하지만 이러한 방법들은 대칭 이미지에만 적용되며, 비대칭 이미지에 대하여는 부자연스러운 결과물을 생성한다. 이러한 한계를 보완하기 위해, 종래의 또 다른 방법은 이미지에서 적절한 패치를 골라 힌트 이미지를 생성한다. 하지만 이 방법은 다른 방법들과 마찬가지로 이미지 아웃페인팅에 제한되고, 광역 이미지 블렌딩과 같은 다른 다양한 이미지 완성 분야에 대하여는 적용할 수 없다.

[0007] 광역 이미지 블렌딩은 최근에 새롭게 제시된 연구 분야이기 때문에 딥러닝 기반의 모델만 존재하며, 힌트를 사용하지 않는다.

[0008] 종래의 이미지 아웃 페인팅 기술들은 이미지 형태의 힌트를 생성하기 위해 주어진 이미지의 일부를 뒤집는 등의 방법으로 사용하여 누락된 영역에 대한 예측을 수행했다. 하지만 해당 이미지 힌트들은 주어진 이미지 구조에 의존적이기 때문에 다양한 형태의 누락된 영역에 적용하지 못한다는 한계가 있다. 예를 들어, 주어진 이미지를 뒤집어서 이미지와 일정한 간격으로 정렬하여 영상을 수평으로 확장하는 방법은 수직으로 이미지를 확장하거나, 다른 두 이미지가 입력으로 주어지는 광역 이미지 블렌딩에 적용할 수 없다. 이와 같이, 종래의 이미지 아웃 페인팅 기술들은 적용할 수 있는 조건이 까다롭다는 한계점이 존재한다.

발명의 내용

해결하려는 과제

[0009] 본 개시는, 텍스트 형태의 이미지 힌트를 사용하여 이미지를 확장하는 방법 및 그 장치를 제공하는 것이다.

[0010] 본 개시는, 인공신경망 네트워크인 이미지 캡션 모듈과 텍스트-가이드 이미지 조작 모듈을 포함하며, 이미지 캡션 모듈에 의해 마스킹된 영역에 대한 언어 힌트를 생성하고 텍스트-가이드 이미지 조작 모듈에 의해 언어 힌트의 가이드에 따라 마스킹된 영역의 이미지를 예측함으로써, 이미지를 확장하는 방법 및 그 장치를 제공하는 것이다.

[0011] 본 개시는, 마스킹 영역에 대한 언어 힌트를 사용하여 마스킹 영역의 이미지를 예측하는 동작을 통해 이미지 아웃페인팅 작업을 지원하는 이미지 확장 방법 및 그 장치를 제공하는 것이다.

[0012] 본 개시는, 마스킹 영역에 대한 언어 힌트를 사용하여 마스킹 영역의 이미지를 예측하는 이미지 아웃페인팅 작업을 좌우 방향으로 반복하고, 이미지 아웃페인팅 작업을 통해 확장된 이미지를 결합하여 결합된 영역을 마스킹 영역으로 설정한 후 이미지 아웃페인팅 작업을 거쳐 파노라마 이미지를 생성하는 광범위한 이미지 블렌딩 작업을 지원하는 이미지 확장 방법 및 그 장치를 제공하는 것이다.

과제의 해결 수단

[0013] 한 특징에 따르면, 캡션 기반 광범위한 페인팅 작업을 수행하는 컴퓨팅 장치로서, 마스킹 처리된 영역을 포함하는 이미지를 입력받아 입력 이미지를 설명하는 상기 마스킹 처리된 영역에 대한 언어 힌트를 생성하는 이미지 캡션 모듈, 그리고 상기 입력 이미지와 상기 언어 힌트를 입력받아 상기 언어 힌트의 가이드에 따라 상기 마스킹 처리된 영역의 이미지를 예측하고, 예측한 이미지로 상기 마스킹 처리된 영역을 채운 확장 이미지를 출력하는 텍스트-가이드 이미지 조작 모듈을 포함한다.

[0014] 상기 이미지 캡션 모듈은, 입력 이미지로부터 시각적 특징을 추출하는 인코더, 그리고 상기 시각적 특징으로부터 일련의 단어들을 생성하는 디코더를 포함할 수 있다.

[0015] 상기 이미지 캡션 모듈은, 입력 이미지로부터 상기 입력 이미지를 설명하는 자연어로 이루어진 텍스트를 출력하도록 사전 학습된 언어 모델을 랜덤하게 마스킹된 이미지들로 구성된 학습 데이터를 이용하여 학습될 수 있다.

[0016] 상기 이미지 캡션 모듈은, 상기 학습 데이터와 SCST(self-critical sequence training) 방법을 사용하여 학습될 수 있다.

[0017] 상기 이미지 캡션 모듈은, 교차 엔트로피 손실(cross entropy loss)과 강화 학습(reinforcement learning)을

이용하여 학습될 수 있다.

[0018] 상기 텍스트-가이드 이미지 조작 모듈은, 텍스트와 이미지가 쌍으로 이루어진 학습 데이터를 이용하여 텍스트에 상응하는 이미지를 출력하도록 학습될 수 있다.

[0019] 상기 이미지 캡션 모듈과 상기 텍스트-가이드 이미지 조작 모듈은, 인공신경망 모델일 수 있다.

[0020] 다른 특징에 따르면, 컴퓨팅 장치에 의해 수행되는 이미지 아웃페인팅 방법으로서, 임의의 영역이 마스킹 처리된 이미지를 입력받는 단계, 입력 이미지로부터 일련의 단어들을 생성하도록 학습된 이미지 캡션 모듈을 이용하여, 상기 마스킹된 이미지에서 마스킹 처리되지 않은 이미지에 대한 의미 및 텍스트 정보를 자연 언어로 설명한 언어 힌트를 생성하는 단계, 상기 언어 힌트를 이용하여 상기 임의의 영역의 이미지를 예측하는 단계, 그리고 상기 예측한 이미지를 포함하는 확장 이미지를 출력하는 단계를 포함한다.

[0021] 상기 예측하는 단계는, 입력 텍스트에 상응하는 이미지를 생성하도록 학습된 텍스트-가이드 이미지 조작 모듈에 상기 마스킹 처리된 이미지와 상기 언어 힌트를 입력하여, 상기 언어 힌트의 가이드에 따라 상기 마스킹 처리된 임의의 영역의 이미지를 예측할 수 있다.

[0022] 상기 텍스트-가이드 이미지 조작 모듈은, 텍스트와 이미지가 쌍으로 이루어진 학습 데이터를 이용하여 텍스트에 상응하는 이미지를 출력하도록 학습된 인공신경망 모델일 수 있다.

[0023] 상기 이미지 캡션 모듈은, 사전 학습된 언어 모델이 랜덤하게 마스킹된 이미지들로 구성된 학습 데이터를 이용하여 학습된 인공신경망 모델이고, 상기 사전 학습된 언어 모델은, 입력 이미지로부터 상기 입력 이미지를 설명하는 자연어로 이루어진 텍스트를 출력하는 모델일 수 있다.

[0024] 또 다른 특징에 따르면, 컴퓨팅 장치에 의해 수행되는 광범위한 이미지 블렌딩 방법으로서, 제1 이미지와 제2 이미지가 연결되는 지점을 마스킹 영역으로 설정하는 단계, 상기 마스킹 영역을 기준으로 양측에 상기 제1 이미지와 상기 제2 이미지가 각각 배치된 이미지를 생성하는 단계, 입력 이미지로부터 일련의 단어들을 생성하도록 학습된 이미지 캡션 모듈을 이용하여, 상기 생성한 이미지에서 마스킹 처리되지 않은 이미지에 대한 의미 및 텍스트 정보를 자연 언어로 설명한 언어 힌트를 생성하는 단계, 상기 언어 힌트를 이용하여 상기 마스킹 영역의 이미지를 예측하는 단계, 그리고 상기 제1 이미지, 상기 예측한 이미지, 및 상기 제2 이미지가 순차적으로 배치된 파노라마 이미지를 출력하는 단계를 포함한다.

[0025] 상기 설정하는 단계 이전에, 이전 단계에서 예측된 출력을 다음 단계의 입력으로 사용하는 이미지 아웃페인팅을 반복하여 상기 제1 이미지를 생성하는 단계, 그리고 상기 제1 이미지와 반대 방향으로 상기 이미지 아웃페인팅을 반복하여 제2 이미지를 생성하는 단계를 포함하고, 상기 이미지 아웃페인팅은, 마스킹 영역에 대한 이미지를 예측하는 작업일 수 있다.

[0026] 상기 이미지 아웃페인팅은, 상기 마스킹 영역이 포함된 이미지에서 마스킹되지 않은 영역에 대한 의미 및 텍스트 정보를 자연 언어로 설명한 언어 힌트를 사용하여 상기 마스킹 영역의 이미지를 예측하고, 상기 마스킹 영역을 예측한 이미지로 채운 확장 이미지를 생성하는 작업을 포함할 수 있다.

[0027] 상기 이미지 아웃페인팅은, 텍스트에 상응하는 이미지를 출력하도록 학습된 인공신경망 모델에 상기 언어 힌트와 상기 마스킹 영역이 포함된 이미지를 입력하여 상기 언어 힌트의 가이드에 따라 상기 마스킹 영역의 이미지를 예측하는 작업을 포함할 수 있다.

발명의 효과

[0028] 본 개시에 따르면, 종래와 같이 단순히 주어진 이미지를 재배열하거나 비슷한 내용의 이미지 힌트를 생성해 내어 이미지에 이어 붙이는 것이 아니라, 주어진 이미지의 내용을 담은 텍스트 형태의 이미지 힌트, 즉, 언어 힌트를 사용하고 언어 힌트의 가이드에 따라 마스킹된 영역의 이미지를 예측함으로써, 이미지 구조에 의존적이지 않기 때문에, 방향 또는 영상 완성 과업에 대한 제약 없이 모든 영상 완성 상황에 적용이 가능하다.

도면의 간단한 설명

[0029] 도 1은 한 실시예에 따른 캡션 기반 광범위한 페인팅(Captioning-based Extensive Painting, CEP) 모듈의 구성을 나타낸 블록도이다.

도 2는 도 1에서 이미지 캡션 모듈의 구성을 나타낸 블록도이다.

도 3은 한 실시예에 따른 이미지 캡션 모듈의 학습 과정을 설명하는 순서도이다.

도 4는 한 실시예에 따른 CEP 모듈의 동작을 설명하는 순서도이다.

도 5는 한 실시예에 따른 이미지 아웃페인팅 동작을 설명한다.

도 6은 한 실시예에 따른 광범위한 이미지 블렌딩(Wide-range Image Blending) 동작을 설명한다.

도 7은 한 실시예에 따른 광범위한 이미지 블렌딩 절차를 설명한다.

도 8은 본 발명의 실시예와 종래 기술에 따라 수행된 이미지 아웃페인팅 작업 결과를 비교한 도면이다.

도 9는 본 발명의 실시예와 종래 기술에 따라 수행된 광범위한 이미지 블렌딩 작업 결과를 비교한 도면이다.

도 10은 실시예에 따른 컴퓨팅 장치의 하드웨어 구성을 나타낸 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0030] 아래에서는 첨부한 도면을 참고로 하여 본 발명의 실시예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다. 그러나 본 발명은 여러가지 상이한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 그리고 도면에서 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략하였으며, 명세서 전체를 통하여 유사한 부분에 대해서는 유사한 도면 부호를 붙였다.

[0031] 설명에서, 도면 부호 및 이름은 설명의 편의를 위해 붙인 것으로서, 장치들이 반드시 도면 부호나 이름으로 한정되는 것은 아니다.

[0032] 설명에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다.

[0033] 또한, 명세서에 기재된 "…부", "…기", "…모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0034] 본 개시의 장치는 적어도 하나의 프로세서가 명령어들(instructions)을 실행함으로써, 본 개시의 동작을 수행할 수 있도록 구성 및 연결된 컴퓨팅 장치이다. 컴퓨터 프로그램은 프로세서가 본 개시의 동작을 실행하도록 기술된 명령어들(instructions)을 포함하고, 비일시적-컴퓨터 판독가능 저장매체(non-transitory computer readable storage medium)에 저장될 수 있다. 컴퓨터 프로그램은 네트워크를 통해 다운로드되거나, 제품 형태로 판매될 수 있다.

[0035] 본 개시의 인공지능 모델(Artificial Intelligence model, AI model)은 적어도 하나의 태스크(task)를 학습하는 기계학습모델로서, 프로세서에 의해 실행되는 컴퓨터 프로그램으로 구현될 수 있다. 인공지능 모델이 학습하는 태스크란, 기계 학습을 통해 해결하고자 하는 과제 또는 기계 학습을 통해 수행하고자 하는 작업을 지칭할 수 있다. 인공지능 모델은 컴퓨팅 장치에서 실행되는 컴퓨터 프로그램으로 구현될 수 있고, 네트워크를 통해 다운로드되거나, 제품 형태로 판매될 수 있다. 또는 인공지능 모델은 네트워크를 통해 다양한 장치들과 연동할 수 있다.

[0036] 이미지 아웃페인팅(image outpainting)은 주어진 이미지를 원래의 경계 너머로 확장하는 것을 목표로 하는 작업이다.

[0037] 광역 이미지 블렌딩(wide-range image blending)은 서로 다른 두 이미지 사이에 중간 이미지를 생성하여 하나의 파노라마 이미지를 형성하는 것을 목표로 하는 작업이다.

[0039] 도 1은 한 실시예에 따른 캡션 기반 광범위한 페인팅(Captioning-based Extensive Painting, 이하, 'CEP'라 통칭함) 모듈의 구성을 나타낸 블록도이고, 도 2는 도 1에서 이미지 캡션 모듈의 구성을 나타낸 블록도이다.

[0040] 도 1에 따르면, CEP 모듈(100)은 적어도 하나의 프로세서에 의해 동작하는 컴퓨팅 장치로서, 본 개시에서 설명하는 동작을 위한 컴퓨터 프로그램을 탑재하고, 컴퓨터 프로그램은 프로세서에 의해 실행된다.

[0041] CEP 모듈(100)은 이미지 캡션(image captioning) 모듈(110) 및 텍스트-가이드 이미지 조작(text-guided image manipulation) 모듈(120)을 포함할 수 있다.

[0042] 이미지 캡션 모듈(110)과 텍스트-가이드 이미지 조작 모듈(120)은 각각의 인공지능(AI) 네트워크 또는 인공지능 모델일 수 있다.

[0043] 이미지 캡션 모듈(110)은 임의의 일부 영역이 누락된 마스킹된 이미지를 입력받아, 입력 이미지의 의미 및 텍스트 정보를 캡쳐하여 자연어로 이루어진 언어 힌트를 생성한다. 이미지 캡션 모듈(110)은 마스킹된 이미지가 입력되면 마스킹된 이미지에 자막을 넣음(captioning)으로써 언어 힌트를 생성한다.

[0044] 확장 페인팅(extensive painting) 동안 언어 힌트를 생성하는 목적은 입력 이미지에 대해 가능한 자세한 정보를 제공하는 것이다. 따라서, 이미지 캡션 모듈(110)은 이미지의 내용을 자연어로 설명하도록 학습된다.

[0045] 실시예에 따르면, 이미지 캡션 모듈(110)은 OFA(One For ALL), ClipCap와 같은 알고리즘이 사용될 수 있다.

[0046] 이미지 캡션 모듈(110)은 무작위 마스크를 사용하여 대규모 캡션 데이터 세트로 이미지 캡션 모델을 최적화할 수 있다. 이미지 캡션 모듈(110)은 마스킹된 이미지에 적절한 캡션, 즉, 언어 힌트를 생성하기 위해 무작위 마스크를 사용하는 SCST(self-critical sequence training) 방법을 사용하여 최적화될 수 있다.

[0047] 종래의 캡션 모델은 완전한 이미지에서 대해서만 훈련되었기 때문에 마스킹된 이미지를 위한 적절한 캡션을 생성하지 못하는 문제가 있다. 예컨대, 종래의 캡션 모델은 두개의 다른 이미지들의 결합 이미지로서 마스킹된 이미지를 반복적으로 인식하고 "두개의 사진(two pictures of)"과 같은 부적절한 접두사(Prefix)를 생성하여 결국 의미론적으로 어색한 이미지를 생성한다. 그러나, 이미지 캡션 모듈(110)은 랜덤하게 마스킹된 이미지와 그에 매칭되는 캡션을 가진 SCST를 사용하는 최적화 프로세스를 통해 마스크 형태에 관계없이 적절한 캡션을 예측할 수 있다. 따라서, 이미지 캡션 모듈(110)은 종래의 의미론적으로 어색한 이미지를 생성하는 문제를 해결할 수 있다.

[0048] 도 2를 참조하면, 이미지 캡션 모듈(110)은 인코더(111) 및 디코더(112)를 포함하는 구조로 이루어진다.

[0049] 인코더(111)는 입력 이미지로부터 시각적 특징(visual features)을 추출한다. 디코더(112)는 시각적 특징으로부터 일련의 단어들을 생성한다.

[0050] 이미지 캡션 모듈(110)은 교차 엔트로피 손실(cross entropy loss)과 강화 학습(reinforcement learning)을 사용하여 훈련될 수 있다. 이러한 훈련을 통해 이미지 캡션 모듈(110)은 미분할 수 없는 캡션 메트릭을 최적화 목표로 사용할 수 있다.

[0051] 텍스트-가이드 이미지 조작 모듈(120)은 마스킹된 이미지, 그리고 언어 힌트를 입력받아 마스킹된 영역의 이미지를 예측하고 예측을 통해 마스킹된 영역의 이미지를 생성한다. 텍스트-가이드 이미지 조작 모듈(120)은 언어 힌트의 안내에 따라 마스킹되지 않은 영역의 콘텐츠에 맞춘 마스킹된 영역의 시각적 콘텐츠를 생성한다. 즉, 텍스트-가이드 이미지 조작 모듈(120)은 언어 힌트를 사용하여 이미지의 마스킹된 영역을 채울 수 있다.

[0052] I_{GT}를 정답(ground-truth) 이미지라 하고, M을 이진(Binary) 이미지라 하면, 미완성 이미지, 즉, I_{IC}를 수학식 1과 같이 나타낼 수 있다.

[0053] [수학식 1]

[0054] $$I_{IC} = I_{GT} \odot (1 - M)$$

[0055] 여기서, ⊙는 아다마르 곱(Hadamard product) 연산을 나타낸다.

[0056] 이미지 캡션 모듈(110)은 미완성 이미지를 입력으로 사용하여 언어 힌트, 즉, T_{hint}를 생성하며, 이를 수식화하면, 수학식 2와 같이 나타낼 수 있다.

[0057] [수학식 2]

[0058] $$T_{hint} = G_{CAP}(I_{IC})$$

[0059] G_{CAP}는 이미지 캡션 모듈(110)을 의미한다.

[0060] 이미지 캡션 모듈(110)은 수학식 3과 같이 랜덤하게 마스킹된 데이터셋(Dataset), 즉, I_{rand}로 추가 훈련될 수 있다.

[0061] [수학식 3]

[0062] $$I_{rand} = I_{GT} \odot (1 - \tilde{M})$$

[0063] 여기서, $\tilde{M}$은 랜덤 마스크이다.

[0064] 이미지 캡션 모듈(110)은 SCST 접근 방식으로 대용량 데이터셋에 대해 사전 훈련된 언어 모델을 이용하여 최적화된 모델일 수 있다. SCST 접근 방식은 보상이 테스트 시간에 사용되는 메트릭(metric)으로 설정되는 강화 학습(REINFORCE) 알고리즘에 기초할 수 있다.

[0065] 훈련 동안 정책(policy)으로부터 샘플링된 문장(T_{hint}^{S})과 매개변수(Θ)가 주어지면 부정적인 예상 보상(negative expected reward)을 최소화할 수 있으며, 이를 수식화하면, 수학식 4와 같다.

[0066] [수학식 4]

[0067] $$L_R(\theta) = -\mathbf{E}_{T_{hint}^s \sim p_\theta}\left[r(T_{hint}^s)\right]$$

[0068] 여기서, $L_R(\Theta)$은 그래디언트(gradient)를 의미한다. r은 생성된 시퀀스를 정답 시퀀스와 비교하여 평가 메트릭(예, CIDEr)에 의해 계산된다.

[0069] 훈련을 통해 수학식 4의 그래디언트는 수학식 5와 같이 근사화될 수 있다.

[0070] [수학식 5]

[0071] $$\nabla_\theta L_R(\theta) \approx -\left(r(T_{hint}^s) - r(\hat{T}_{hint})\right)\nabla_\theta \log p_\theta(T_{hint}^s)$$

[0072] 여기서, r($\hat{T}_{hint}$)은 모델, 즉, 이미지 캡션 모듈(110)을 탐욕스럽게(greedily) 디코딩하여 얻은 기본 보상(baseline reward)이다.

[0073] 그래디언트는 현재 모델의 보상보다 훈련 중에서 정책에서 캡션 샘플링의 확률을 높이는 경향이 있다. 이러한 과정을 통해 최적화된 이미지 캡션 모듈(110)은 마스크 형태에 관계없이 적절한 캡션을 예측할 수 있다.

[0074] 본 발명의 실시예에서 획득하는 언어 힌트는 종래의 이미지 힌트에 비해 이미지 완성 작업과 관계없이 잘 작동하는 장점이 있다.

[0075] 종래에 이미지 힌트 기반의 방식들, 예컨대, BR(Bidirectional Rearrange), MR(Mirror Arrange), IAH(Image-Adaptive Hint)은 이미지 구조에 대한 의존도가 크다. BR은 입력 이미지의 반대되는 부분을 왼쪽과 오른쪽으로 전환하여 힌트로 활용하고, 두 분할된 이미지 사이의 간격을 채워준다. MR은 미러 플립 입력 영상을 힌트로 사용하여 누락된 영역을 예측한다. 이처럼, BR과 MR은 입력 영상을 왼쪽과 오른쪽을 전환하여 재정렬하거나 가려진 영역 옆에 입력 영상을 미러링하기 때문에 대칭 영상에만 적용할 수 있다. 또한, IAH는 힌트 기반 방법을 비대칭 이미지로 확장한 기술로서, Vision Transformer를 사용하여 이미지 적응 방식으로 이미지 힌트를 생성하며, 출력 힌트는 이미지 형식으로 제한되어 고정 크기 힌트만 생성할 수 있다. 하지만, 본 발명의 이미지 캡쳐 모듈(110)은 다른 모달리티(modality)를 힌트로 사용, 즉, 누락된 영역에 대한 텍스트 형태의 언어 힌트를 사용하기 때문에 제한없이 어떠한 형태의 이미지 완성에도 적용할 수 있다. 종래의 방식은 기본적으로 이미지 구조에 대해 종속적이기 때문에 본 발명에서와 같은 언어 힌트를 생성할 수 없는 구조라는 점에서 본 발명과 차별된다.

[0076] 텍스트-가이드 이미지 조작 모듈(120)은 이미지 캡쳐 모듈(110)에 의해 생성된 언어 힌트를 사용하여 마스킹된 영역에 대한 이미지를 생성한다.

[0077] 실시예에 따르면, 텍스트-가이드 이미지 조작 모듈(120)은 GLIDE, Imagen, DALL-E, textdrive blended diffusion과 같은 대규모(large-scale) 언어(language)-비전(vision) 모델을 사용할 수 있다. 이러한 모델은 대규모 텍스트-이미지 쌍 데이터 셋을 이용하여 훈련되므로, 특정 데이터 세트에서 훈련된 모델과 비교해도 광범위한 도메인에서 제로(zero)-샷(shot) 이미지 생성에 능숙한 장점이 있어 전례없는 수준의 일반화를 달성한 모델이다.

[0078] 텍스트-가이드 이미지 조작 모듈(120)은 마스킹된 이미지(I_{IC})와 언어 힌트(T_{hint})를 입력으로 사용하여 누락된 영역이 채워진 완전한 이미지, 즉, I_{pred}를 생성하며, 이를 수식화하면 수학식 6과 같다.

[0079] [수학식 6]

[0080] $$I_{pred} = G_{IM}(I_{IC}, T_{hint})$$

[0081] G_{IM}는 텍스트-가이드 이미지 조작 모듈(120)을 의미한다.

[0082] 이와 같이, CEP 모듈(100)은 이미지 캡션 모듈(110)과 텍스트-가이드 이미지 조작 모듈(120)은 한쌍으로 작동하여 이미지 아웃페인팅(image outpainting) 작업 또는 광범위한 이미지 블렌딩(wide-range image blending) 작업을 수행할 수 있다.

[0084] 도 3은 한 실시예에 따른 이미지 캡션 모듈의 학습 과정을 설명하는 순서도이다.

[0085] 도 3에 따르면, CEP 모듈(100)은 랜덤하게 마스킹된 이미지들로 구성된 마스킹된 데이터 셋을 훈련 데이터로 입력받는다(S101).

[0086] CEP 모듈(100)은 훈련 동안 정책으로부터 샘플링된 문장, 그리고 훈련 데이터를 이용하여 부정적인 예상 보상을 최소화하도록 매개변수를 포함하는 이미지 캡션 모듈(110)을 학습시킨다(S102).

[0088] 도 4는 한 실시예에 따른 CEP 모듈의 동작을 설명하는 순서도이다.

[0089] 도 4에 따르면, CEP 모듈(100)은 누락된 영역이 있는 마스킹된 이미지 입력받는다(S201).

[0090] CEP 모듈(100)의 이미지 캡션 모듈(110)은 S201에서 입력받은 이미지의 의미 및 텍스트 정보를 자연 언어로 설명한 언어 힌트를 생성한다(S202).

[0091] CEP 모듈(100)의 텍스트-가이드 이미지 조작 모듈(120)은 S201에서 입력받은 마스킹된 이미지와 S202에서 입력받은 언어 힌트를 입력으로 사용하여 마스킹된 영역의 이미지를 예측하고, 예측한 이미지가 포함된 완성된 확장 이미지를 출력할 수 있다(S203).

[0093] 도 5는 한 실시예에 따른 이미지 아웃페인팅 동작을 설명한다.

[0094] 도 5에 따르면, CEP 모듈(100)은 마스킹 영역(빗금친 부분)이 포함된 마스킹된 이미지(P1)를 입력받아 마스킹 영역이 채워진 완성된 이미지(P2)를 출력한다.

[0095] CEP 모듈(100)은 주어진 정보가 이미지 내부 뿐일 때, 이미지 외부를 생성할 수 있다. CEP 모듈(100)은 가로 방향으로 이미지의 단면 또는 양면을 예측할 수 있다.

[0096] CEP 모듈(100)은 랜덤하게 마스킹된 이미지들에 대해 최적화되었으므로, 누락된 영역, 즉, 확장하고자 하는 영역에 해당하는 마스킹한 영역을 입력받아 이미지 아웃페인팅을 수행할 수 있다.

[0097] 이미키 캡션 모듈(110)은 마스킹된 이미지(P1)를 입력받아 언어 힌트를 생성한다. 언어 힌트는 마스킹된 이미지(P1)에서 마스킹되지 않은 영역의 콘텐츠를 설명하는 자연 언어로 된 텍스트로서, 예를 들어, "The skyline of the city on a cloudy day"일 수 있다.

[0098] 텍스트-가이드 이미지 조작 모듈(120)은 마스킹된 이미지(P1)와 언어 힌트를 입력받고, 언어 힌트를 사용하여 마스킹된 영역(빗금친 부분)의 이미지를 예측한다.

[0100] 도 6은 한 실시예에 따른 광범위한 이미지 블렌딩(Wide-range Image Blending) 동작을 설명한다.

[0101] 도 6에 따르면, 광범위한 이미지 블렌딩 동작은 크게 세가지 단계(Stage)로 이루어진다.

[0102] 단계 1은 이전 단계에서 예측된 출력을 다음 단계의 입력으로 사용하여 아웃페인팅을 반복하는 다단계 예측을

수행한다. 구체적으로, 단계 1은 이미지 아웃페인팅 제1 방향의 마스킹 영역(빗금친 부분)을 포함하는 마스킹된 이미지를 입력으로 사용하여 도 1 ~ 도 5에서 설명한 CEP 모듈(100)의 동작을 N번 반복한다. CEP 모듈(100)의 N회 반복 동작을 통해 제1 방향으로 이미지를 확장할 수 있다.

[0103] 단계 2는 외삽된 두 이미지가 연결될 때까지 단계 1과 반대 방향인 제2 방향으로 아웃페인팅을 반복한다. 구체적으로, 단계 2는 제2 방향의 마스킹 영역(빗금친 부분)을 포함하는 마스킹된 이미지를 입력으로 사용하여 도 1 ~ 도 5에서 설명한 CEP 모듈(100)의 동작을 N번 반복한다. CEP 모듈(100)의 N회 반복 동작을 통해 제2 방향으로 이미지를 확장할 수 있다.

[0104] 단계 3은 단계 1과 단계 2를 통해 확장된 각각의 이미지를 연결시키는데, 연결이 끊어진 영역에 마스크를 적용하여 마스킹된 이미지를 CEP 모듈(100)의 입력으로 사용하여 마스킹 영역을 채울 수 있다.

[0105] 단계 1과 단계 2를 통해 확장된 각각의 이미지는 서로 확장한 방향에서 마주하게 되고, 마주한 지점을 연결하는데, 연결된 지점은 완전히 이어지지 못하고 끊어질 수 있다. 이러한 끊어진 지점의 콘텐츠를 생성하기 위해 끊어진 지점을 마스킹 영역(빗금친 부분)으로 생성하고 마스킹 영역의 좌측에 단계 1을 통해 확장된 이미지를 배치하고 마스킹 영역의 우측에 단계 2를 통해 확장된 이미지를 배치한 이미지를 CEP 모듈(100)에 입력한다.

[0106] CEP 모듈(100)은 입력받은 이미지에 대해 언어 힌트를 생성하고, 생성한 언어 힌트의 가이드에 따라 마스킹 영역을 예측하고, 예측한 이미지를 포함하는 완성된 이미지를 생성하게 된다.

[0108] 도 7은 한 실시예에 따른 광범위한 이미지 블렌딩 절차를 설명하며, 도 6의 동작을 순차적으로 설명한 도면이다.

[0109] 도 7에 따르면, CEP 모듈(100)은 제1 방향으로 누락된 영역이 있는 마스킹된 이미지를 입력받아 마스킹된 영역을 예측한 이미지로 채운 제1 이미지를 생성한다(S301). 여기서, 누락된 영역은 확장하고자 하는 영역에 해당한다.

[0110] CEP 모듈(100)은 S301의 예측 횟수가 N번 반복하였는지 판단(S302)하고, N번 반복되지 않았다면, S301에서 예측을 통해 생성된 제1 이미지에서 타겟 방향, 즉, 제1 방향으로 누락 영역을 추가한 마스킹된 이미지를 생성한다(S303).

[0111] CEP 모듈(100)은 S303에서 생성한 마스킹된 이미지를 입력으로 사용하여 S301을 수행한다.

[0112] S302에서 N번 반복으로 판단되면, CEP 모듈(100)은 제1 방향과 반대 방향으로 누락된 영역이 있는 마스킹된 이미지를 입력받아 누락된 영역을 예측하고, 예측한 영역이 포함된 제2 이미지를 생성한다(S304).

[0113] CEP 모듈(100)은 S304의 예측 횟수가 N번 반복하였는지 판단(S305)하고, N번 반복되지 않았다면, S304에서 예측을 통해 생성된 제1 이미지에서 타겟 방향, 즉, 제2 방향으로 누락 영역을 추가한 마스킹된 이미지를 생성한다(S306).

[0114] CEP 모듈(100)은 S306에서 생성한 마스킹된 이미지를 입력으로 사용하여 S304를 수행한다.

[0115] CEP 모듈(100)은 N번 반복으로 판단되면, S301을 통해 생성한 제1 이미지, 그리고 S304을 통해 생성한 제2 이미지를 연결시키고, 연결 부위의 끊어진 지점을 마스킹 처리한 이미지를 입력 이미지로 사용하여 마스킹된 영역의 이미지를 예측하며, 마스킹된 부분을 예측한 이미지로 채운 제3 이미지를 생성한다(S307). 즉, 제3 이미지는 좌우 방향으로 확장된 이미지를 블렌딩한 파노라마 이미지로 생성될 수 있다.

[0117] 도 8은 본 발명의 실시예와 종래 기술에 따라 수행된 이미지 아웃페인팅 작업 결과를 비교한 도면이다.

[0118] 이미지 아웃페인팅은 기본적으로 주어진 이미지를 원래의 경계 너머로 확장하는 작업을 말한다.

[0119] 도 8의 (a)는 이미지 캡션 모듈(110)을 통해 생성한 언어 힌트("A view of the ocean with rocks in the water")를 사용하여 텍스트-가이드 조작 모듈(120)에 의해 누락된 영역의 이미지를 예측하고, 예측한 이미지로 채워진 확장 이미지를 나타낸다. 도 8의 (b)는 종래의 방식으로 이미지 아웃 페인팅 작업이 수행되어 확장된 이미지를 나타낸다. 종래의 방식은 pallette 알고리즘이 사용되었다.

[0120] 도 8의 (c)는 이미지 캡션 모듈(110)을 통해 생성한 언어 힌트("A city street with a tree in front of a

building")를 사용하여 텍스트-가이드 조작 모듈(120)에 의해 누락된 영역의 이미지를 예측하고, 예측한 이미지로 채워진 확장 이미지를 나타낸다. 도 8의 (d)는 종래의 방식으로 이미지 아웃 페인팅 작업이 수행되어 확장된 이미지를 나타낸다.

[0121] 도 8의 (a)와 도 8의 (b), 도 8의 (c)와 도 8의 (d)를 비교하면, 본 발명을 적용한 확장 이미지가 원래의 이미지와 동일하게 연장된 자연스러운 이미지를 연출함을 확인할 수 있다.

[0123] 도 9는 본 발명의 실시예와 종래 기술에 따라 수행된 광범위한 이미지 블렌딩 작업 결과를 비교한 도면이다.

[0124] 광범위한 이미지 블렌딩 작업은 서로 다른 두개의 이미지 사이에 중간 이미지를 생성하여 하나의 파노라마 이미지를 생성하는 작업이다.

[0125] 도 9의 (a)는 이미지 캡션 모듈(110)을 통해 생성한 언어 힌트("An aerial view of a city and a body of water")를 사용하여 텍스트-가이드 조작 모듈(120)에 의해 누락된 영역의 이미지를 예측하고, 예측한 이미지로 채워진 확장 이미지를 나타낸다. 도 9의 (b)는 종래의 방식으로 이미지 블렝딩 작업이 수행되어 확장된 이미지를 나타낸다. 종래의 방식은 BRIDGE 알고리즘이 사용되었다.

[0126] 도 9의 (c)는 이미지 캡션 모듈(110)을 통해 생성한 언어 힌트("A view of the city at night with mountaions in the background")를 사용하여 텍스트-가이드 조작 모듈(120)에 의해 누락된 영역의 이미지를 예측하고, 예측한 이미지로 채워진 확장 이미지를 나타낸다. 도 9의 (d)는 종래의 방식으로 이미지 블렝딩 작업이 수행되어 확장된 이미지를 나타낸다.

[0127] 도 9의 (a)와 도 9의 (b), 도 9의 (c)와 도 9의 (d)를 비교하면, 본 발명을 적용한 파노라마 이미지가 원래의 이미지와 동일하게 연장된 자연스러운 이미지를 연출함을 확인할 수 있다.

[0129] 이상 기재한 바에 따르면, CEP 모듈(100)은 종래와 같이 단순히 주어진 이미지를 재배열하거나 비슷한 내용의 이미지 힌트를 생성해 내어 이미지에 이어 붙이는 것이 아니라, 주어진 이미지의 내용을 담은 텍스트 형태의 이미지를 생성한다. 텍스트 형태의 힌트, 즉, 언어 힌트는 이미지 구조에 의존적이지 않기 때문에, 방향 또는 영상 완성 과업에 대한 제약 없이 모든 영상 완성 상황에 적용이 가능하다.

[0130] 이러한 언어 힌트를 사용하기 위해 CEP 모듈(100)은 이미지 캡션 모듈(110)을 통해 주어진 이미지의 캡션인 언어 힌트를 생성한 뒤, 텍스트-가이드 조작 모듈(120)을 통해 언어 힌트의 가이드에 따라 누락된 영역의 이미지를 예측한다. 종래의 이미지 캡션 모델들은 이미지 일부가 누락되었을 때 잘못된 캡션을 생성할 수 있지만, 본 발명의 이미지 캡션 모듈(110)은 랜덤 마스크로 최적화되었으므로, 누락 영역에 대한 정확한 캡션, 즉, 언어 힌트를 예측할 수 있다.

[0131] 이미지 캡션 모듈(110)과 텍스트-가이드 조작 모듈(120)은 텍스트-이미지 데이터 셋으로 사전 학습되어 뛰어난 성능을 나타냄을 도 8과 도 9를 통해 증명하였다.

[0133] 한편, 도 10은 실시예에 따른 컴퓨팅 장치의 하드웨어 구성을 나타낸 블록도로서, 도 1 ~ 도 9에서 설명한 CEP 모듈(100)은 컴퓨팅 장치로 구현될 수 있다.

[0134] 도 10을 참조하면, 컴퓨팅 장치(200)는 하나 이상의 프로세서(210), 프로세서(210)에 의하여 수행되는 프로그램을 로드하는 메모리(220), 프로그램 및 각종 데이터를 저장하는 스토리지(230), 및 통신 인터페이스(240)를 포함하고, 이들은 버스(250)를 통해 연결된다. 다만, 상술한 구성 요소들은 본 개시에 따른 컴퓨팅 장치(200)를 구현하는데 있어서 필수적인 것은 아니어서, 컴퓨팅 장치(200)는 위에서 열거된 구성요소들 보다 많거나, 또는 적은 구성요소들을 가질 수 있다. 예컨대 컴퓨팅 장치(200)는 출력부 및/또는 입력부(미도시)를 더 포함하거나, 또는 스토리지(230)가 생략될 수도 있다.

[0135] 프로그램은 메모리(220)에 로드될 때 프로세서(210)로 하여금 본 개시의 다양한 실시예에 따른 방법/동작을 수행하게끔 하는 명령어들(instructions)을 포함할 수 있다.

[0136] 프로세서(210)는 도 1 ~ 도 9에서 설명한 이미지 캡션 모듈(110) 및 텍스트-가이드 조작 모듈(120)의 동작을 수행하는 명령어들을 실행함으로써, 본 개시의 다양한 실시예에 따른 방법/동작들을 수행할 수 있다. 프로그램은

기능을 기준으로 묶인 일련의 컴퓨터 판독가능 명령어들로 구성되고, 프로세서에 의해 실행되는 것을 가리킨다.

[0137] 프로세서(210)는 컴퓨팅 장치(200)의 각 구성의 전반적인 동작을 제어한다. 프로세서(210)는 CPU(Central Processing Unit), MPU(Micro Processor Unit), MCU(Micro Controller Unit), GPU(Graphic Processing Unit) 또는 본 개시의 기술 분야에 잘 알려진 임의의 형태의 프로세서 중 적어도 하나를 포함하여 구성될 수 있다. 또한, 프로세서(210)는 본 개시의 다양한 실시예들에 따른 방법/동작을 실행하기 위한 적어도 하나의 애플리케이션 또는 프로그램에 대한 연산을 수행할 수 있다.

[0138] 메모리(220)는 각종 데이터, 명령 및/또는 정보를 저장한다. 메모리(220)는 본 개시의 다양한 실시예들에 따른 방법/동작을 실행하기 위하여 스토리지(230)로부터 하나 이상의 프로그램을 로드할 수 있다. 메모리(220)는 RAM과 같은 휘발성 메모리로 구현될 수 있을 것이나, 본 개시의 기술적 범위는 이에 한정되지 않는다.

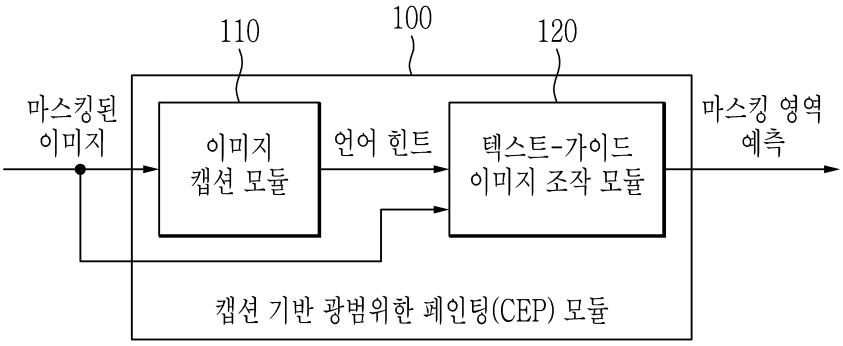
[0139] 스토리지(230)는 프로그램을 비임시적으로 저장할 수 있다. 스토리지(230)는 ROM(Read Only Memory), EPROM(Erasable Programmable ROM), EEPROM(Electrically Erasable Programmable ROM), 플래시 메모리 등과 같은 비휘발성 메모리, 하드 디스크, 착탈형 디스크, 또는 본 개시가 속하는 기술 분야에서 잘 알려진 임의의 형태의 컴퓨터로 읽을 수 있는 기록 매체를 포함하여 구성될 수 있다. 통신 인터페이스(240)는 유/무선 통신 모듈일 수 있다.

[0141] 이상에서 설명한 본 발명의 실시예는 장치 및 방법을 통해서만 구현이 되는 것은 아니며, 본 발명의 실시예의 구성에 대응하는 기능을 실현하는 프로그램 또는 그 프로그램이 기록된 기록 매체를 통해 구현될 수도 있다.

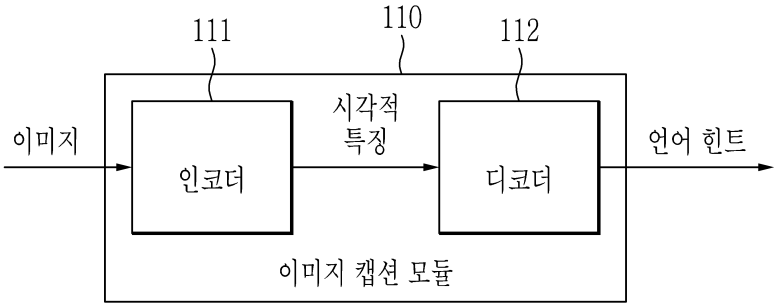
[0142] 이상에서 본 발명의 실시예에 대하여 상세하게 설명하였지만 본 발명의 권리범위는 이에 한정되는 것은 아니고 다음의 청구범위에서 정의하고 있는 본 발명의 기본 개념을 이용한 당업자의 여러 변형 및 개량 형태 또한 본 발명의 권리범위에 속하는 것이다.

도면

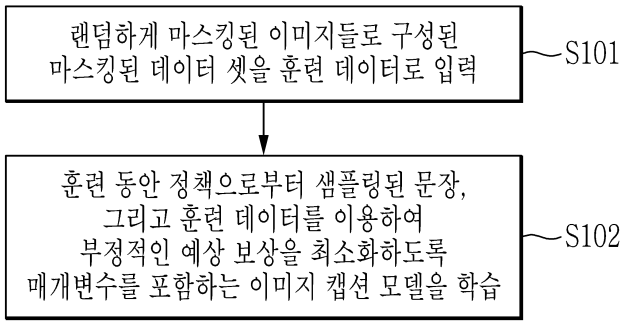
도면1



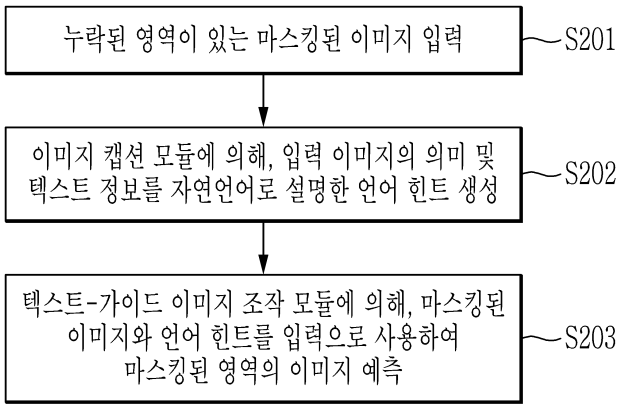
도면2


111
110
112
이미지
인코더
시각적
특징
디코더
언어 힌트
이미지 캡션 모듈

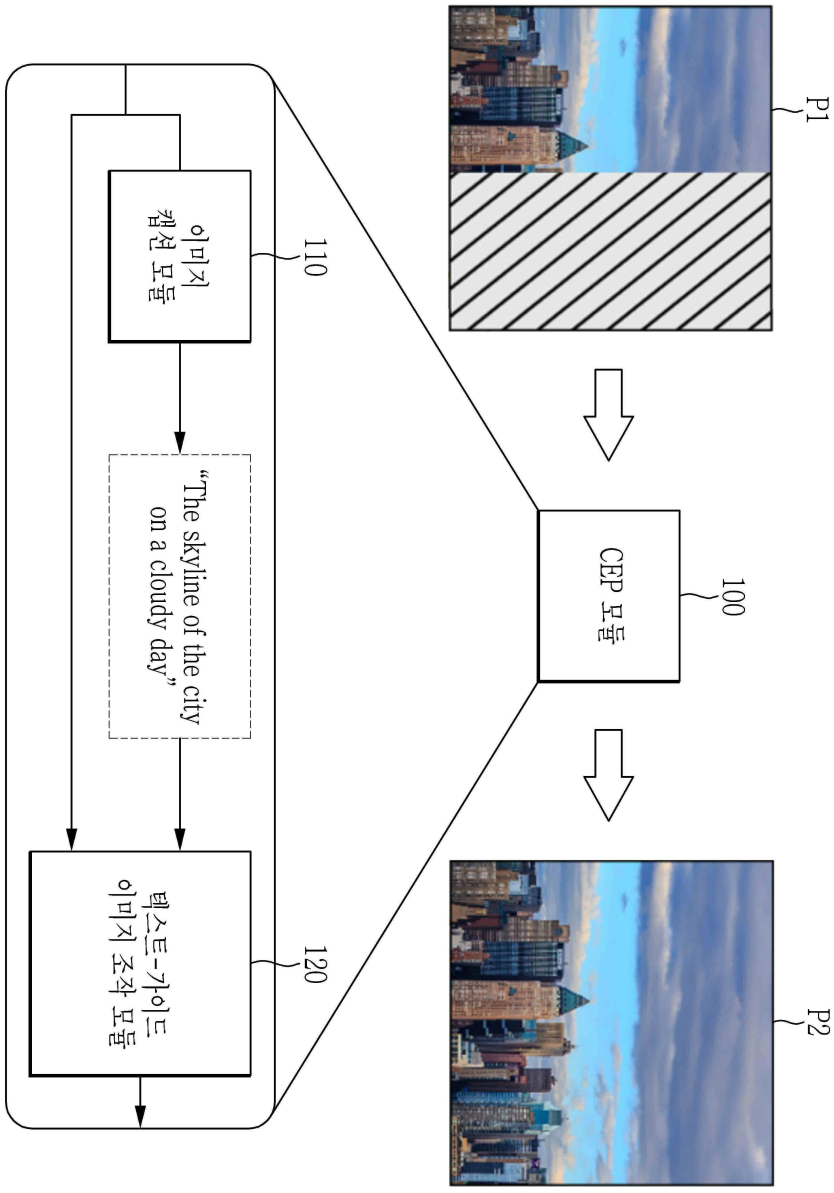
도면3


랜덤하게 마스킹된 이미지들로 구성된
마스킹된 데이터 셋을 훈련 데이터로 입력
S101
훈련 동안 정책으로부터 샘플링된 문장,
그리고 훈련 데이터를 이용하여
부정적인 예상 보상을 최소화하도록
매개변수를 포함하는 이미지 캡션 모델을 학습
S102

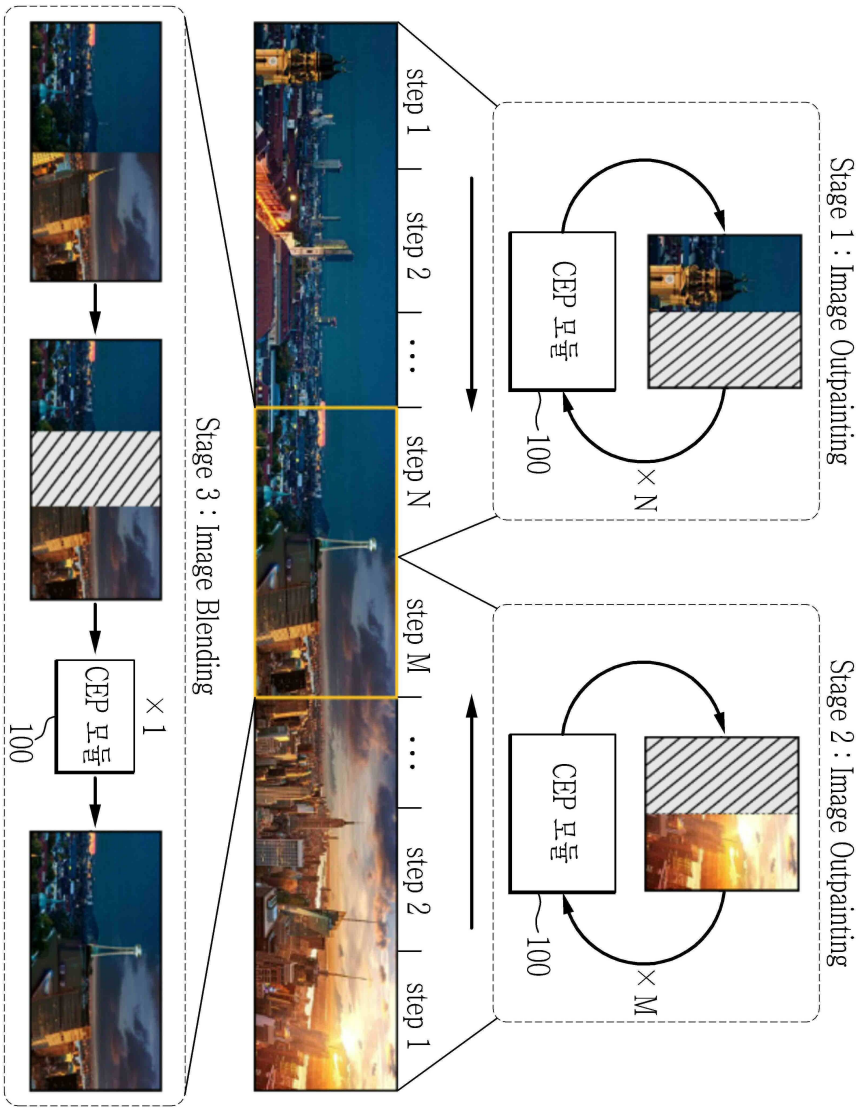
도면4


누락된 영역이 있는 마스킹된 이미지 입력
S201
이미지 캡션 모듈에 의해, 입력 이미지의 의미 및
텍스트 정보를 자연언어로 설명한 언어 힌트 생성
S202
텍스트-가이드 이미지 조작 모듈에 의해, 마스킹된
이미지와 언어 힌트를 입력으로 사용하여
마스킹된 영역의 이미지 예측
S203

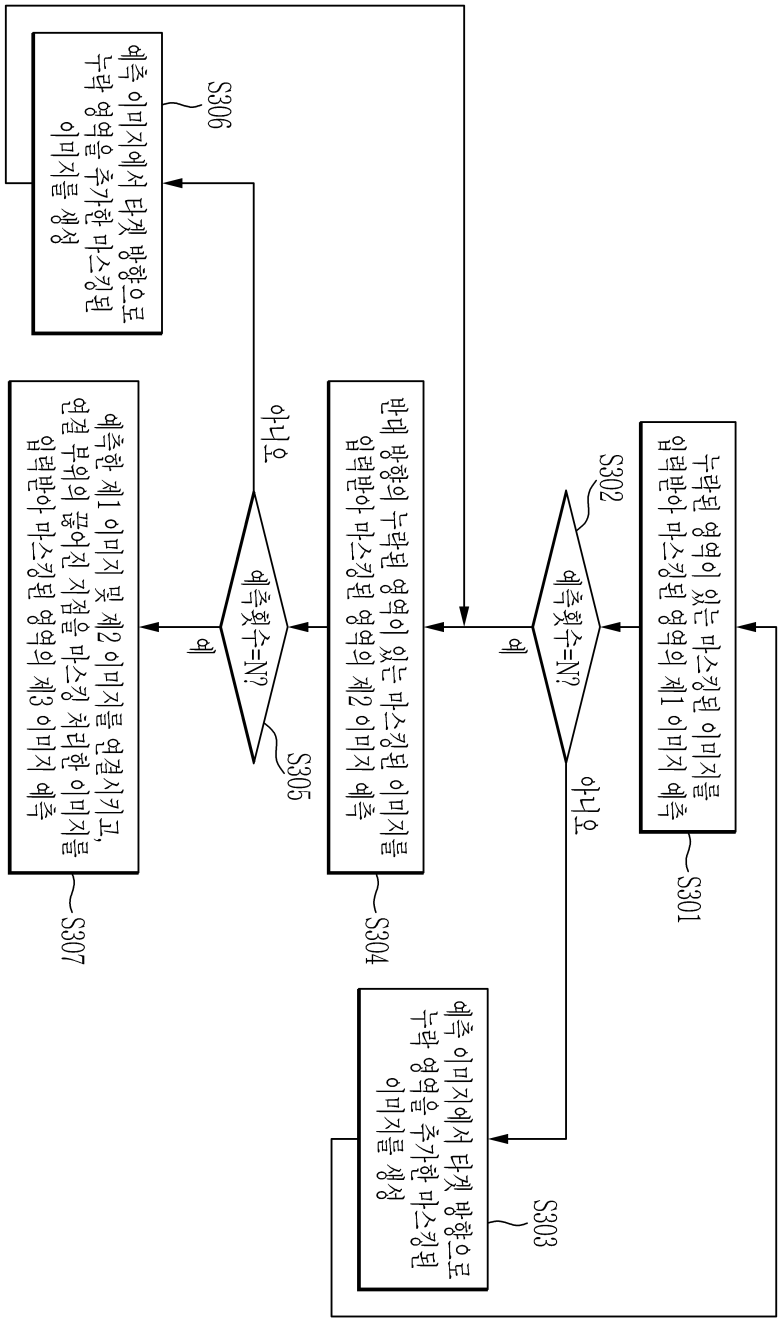
도면5


P1
이미지 캡션 모듈
110
"The skyline of the city on a cloudy day"
CEP 모듈
100
텍스트-가이드 이미지 조작 모듈
120
P2

도면6


Stage 1 : Image Outpainting
CEP 모듈
100
× N
Stage 2 : Image Outpainting
CEP 모듈
100
× M
step 1
step 2
...
step N
step M
...
step 2
step 1
Stage 3 : Image Blending
CEP 모듈
100
× 1

도면7


누락된 영역이 있는 마스킹된 이미지를 입력받아 마스킹된 영역의 제1 이미지 예측
S301
예측횟수=N?
S302
예
아니오
예측 이미지에서 타겟 방향으로 누락 영역을 추가한 마스킹된 이미지를 생성
S303
반대 방향의 누락된 영역이 있는 마스킹된 이미지를 입력받아 마스킹된 영역의 제2 이미지 예측
S304
예측횟수=N?
S305
예
아니오
예측한 제1 이미지 및 제2 이미지를 연결시키고, 연결 부위의 끊어진 지점을 마스킹 처리한 이미지를 입력받아 마스킹된 영역의 제3 이미지 예측
S307
예측 이미지에서 타겟 방향으로 누락 영역을 추가한 마스킹된 이미지를 생성
S306

도면8

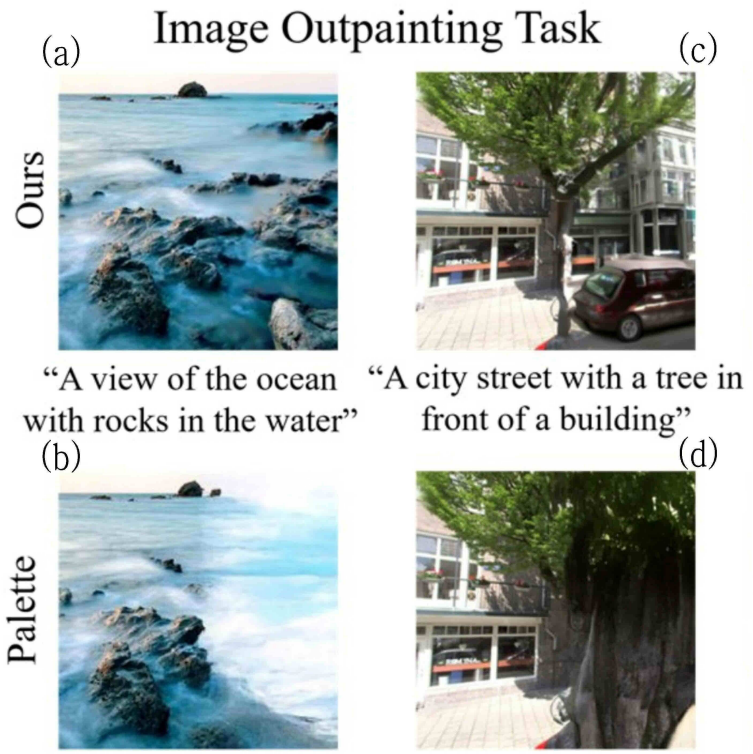
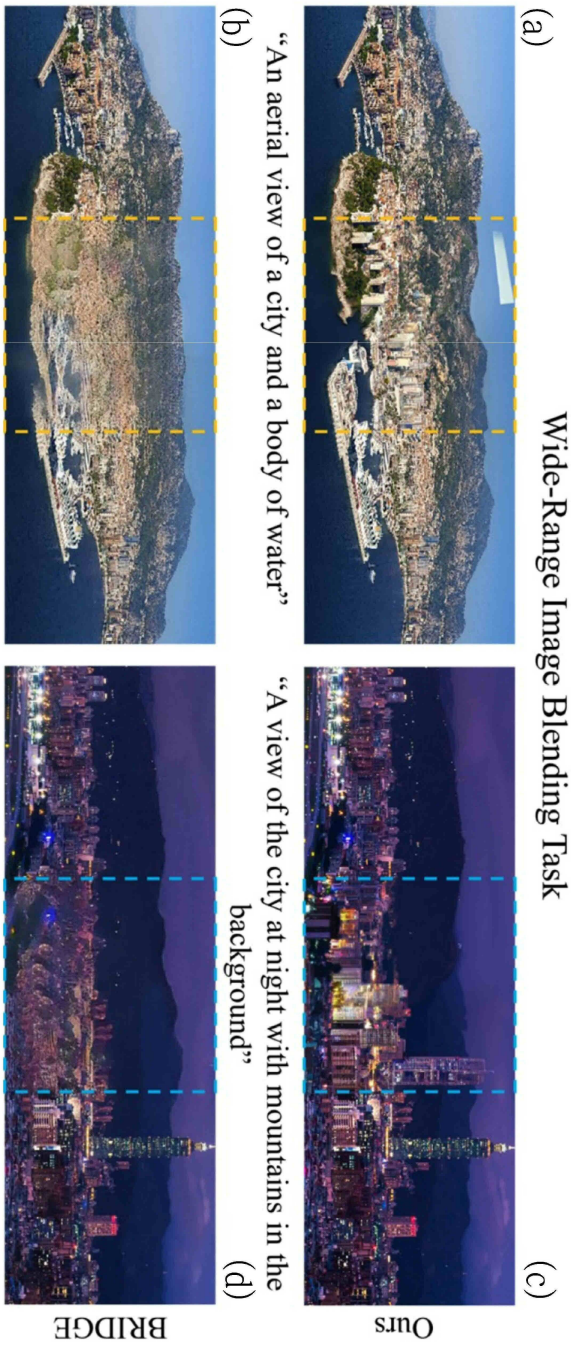


Image Outpainting Task
(a)
(c)
Ours
"A view of the ocean with rocks in the water"
"A city street with a tree in front of a building"
(b)
(d)
Palette

도면9


Wide-Range Image Blending Task
(a)
(b)
(c)
(d)
"An aerial view of a city and a body of water"
"A view of the city at night with mountains in the background"
Ours
BRIDGE

도면10