(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0070795
(43) 공개일자 2024년05월22일

(51) 국제특허분류(Int. Cl.)
G06V 20/52 (2022.01) *G06V 40/16* (2022.01)
H04N 19/30 (2014.01) *H04N 19/60* (2014.01)
(52) CPC특허분류
G06V 20/52 (2023.08)
G06V 40/167 (2022.01)
(21) 출원번호 10-2022-0152138
(22) 출원일자 2022년11월15일
심사청구일자 2022년11월15일

(71) 출원인
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
김용국
경기도 성남시 분당구 동판교로 156, 911동 503호(삼평동, 봇들마을9단지금호어울림아파트)
레 투안 비엣
서울특별시 광진구 능동로 209, 세종대학교 컴퓨터공학과 AI빌딩 722(군자동)
(74) 대리인
송인호, 최관락

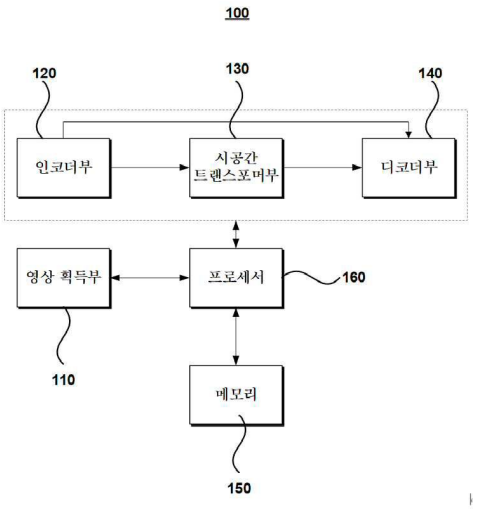
전체 청구항 수 : 총 9 항

(54) 발명의 명칭 **트랜스포머 기반 이상 상황 검지 장치 및 그 방법**

(57) 요 약

트랜스포머 기반 이상 상황 검지 장치 및 그 방법이 개시된다. 트랜스포머 기반 이상 상황 검지 장치는, 피라미드 구조를 가지는 n개의 인코딩 블록을 포함하며, 각각의 인코딩 블록을 통해 피라미드 구조를 가지는 다중 스케일 특징맵을 각각 추출하는 인코더부; n개의 시공간 트랜스포머를 가지되, 각각의 시공간 트랜스포머를 통해 상기 각 인코딩 블록에서 추출된 특징맵에 대한 시간 특징과 공간 특징이 강조된 다중 스케일 시공간 어텐션 맵을 각각 생성하는 시공간 트랜스포머부; 및 역피라미드 구조를 가지는 n개의 디코더를 포함하며, 상기 n개의 디코더를 통해 상기 다중 스케일 시공간 어텐션 맵과 이전 계층 디코더의 출력을 이용하여 예측 프레임을 생성하는 디코더부를 포함하되, 상기 예측 프레임을 이용하여 비정상 이벤트를 검출한다.

대 표 도 - 도1



(52) CPC특허분류
H04N 19/30 (2015.01)
H04N 19/60 (2015.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711152582
과제번호	2019-0-00231-004
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보보호핵심원천기술개발(정진기금)
연구과제명	공공 인프라 안전을 위한 인공지능 기반 영상보안 기술 및 시스템 개발
기 여 율	1/1
과제수행기관명	세종대학교 산학협력단
연구기간	2022.01.01 ~ 2022.12.31

명 세 서

청구범위

청구항 1

피라미드 구조를 가지는 n개의 인코딩 블록을 포함하며, 각각의 인코딩 블록을 통해 피라미드 구조를 가지는 다중 스케일 특징맵을 각각 추출하는 인코더부;

n개의 시공간 트랜스포머를 가지되, 각각의 시공간 트랜스포머를 통해 상기 각 인코딩 블록에서 추출된 특징맵에 대한 시간 특징과 공간 특징이 강조된 다중 스케일 시공간 어텐션 맵을 각각 생성하는 시공간 트랜스포머부; 및

역피라미드 구조를 가지는 n개의 디코더를 포함하며, 상기 n개의 디코더를 통해 상기 다중 스케일 시공간 어텐션 맵과 이전 계층 디코더의 출력을 이용하여 예측 프레임을 생성하는 디코더부를 포함하되,

상기 예측 프레임을 이용하여 비정상 이벤트를 검출하는 것을 특징으로 하는 이상 상황 검지 장치.

청구항 2

제1 항에 있어서,

상기 제n 디코더는 상기 이전 계층의 디코더 출력을 업샘플링한 결과와 상기 제n 시공간 트랜스포머로부터 출력되는 제n 시공간 어텐션 맵을 결합한 후 컨볼루션 레이어를 적용하는 것을 특징으로 하는 이상 상황 검지 장치.

청구항 3

제1 항에 있어서,

상기 제n 트랜스포머는,

상기 제n 인코더로부터 출력되는 상기 제n 특징맵을 입력받은 후 서로 다른 프레임에서 동일한 공간적 위치를 갖는 패치들을 비교하여 시간적 어텐션 맵을 생성하는 시간적 셀프 어텐션(temporal self-attention) 모듈; 및

상기 시간적 셀프 어텐션 모듈 후단에 위치되며, 상기 시간적 셀프 어텐션 모듈의 출력을 이용하여 공간적 어텐션 맵을 생성하는 공간 셀프 어텐션(spatial self-attention) 모듈을 포함하되,

상기 시간적 어텐션 맵과 상기 공간적 어텐션 맵을 더하여 상기 제n 시공간 어텐션 맵이 생성되는 것을 특징으로 하는 이상 상황 검지 장치.

청구항 4

제3 항에 있어서,

상기 공간 셀프 어텐션 모듈은 공간 셀프 어텐션 적용 전에 공간 축소 레이어를 이용하여 공간 축소를 적용하는 것을 특징으로 하는 이상 상황 검지 장치.

청구항 5

제3 항에 있어서,

상기 제n 시공간 어텐션 맵은 피드 포워드 네트워크 모듈을 통해 디코더로 전달되되, LN(Layer Norm) 적용된 후 상기 피드 포워드 네트워크 모듈로 전달되는 것을 특징으로 하는 이상 상황 검지 장치.

청구항 6

(a) 피라미드 구조를 가지는 n개의 인코딩 블록을 포함하며, 각각의 인코딩 블록을 통해 피라미드 구조를 가지는 다중 스케일 특징맵을 각각 추출하는 단계;

(b) n개의 시공간 트랜스포머를 가지되, 각각의 시공간 트랜스포머를 통해 상기 각 인코딩 블록에서 추출된 특징맵에 대한 시간 특징과 공간 특징이 강조된 다중 스케일 시공간 어텐션 맵을 각각 생성하는 단계; 및

(c) 역피라미드 구조를 가지는 n개의 디코더를 포함하며, 상기 n개의 디코더를 통해 상기 다중 스케일 시공간 어텐션 맵과 이전 계층 디코더의 출력을 이용하여 예측 프레임을 생성하는 단계; 및

상기 예측 프레임을 이용하여 스코어를 계산하고, 상기 스코어가 임계치 이상인 경우 비정상 이벤트가 존재하는 것으로 판단하는 단계를 포함하는 이상 상황 검지 방법.

청구항 7

제6 항에 있어서,

상기 제n 디코더는 상기 이전 계층의 디코더 출력을 업샘플링한 결과와 상기 제n 시공간 트랜스포머로부터 출력되는 제n 시공간 어텐션 맵을 결합한 후 컨볼루션 레이어를 적용하는 것을 특징으로 하는 이상 상황 검지 방법.

청구항 8

제6 항에 있어서,

상기 제n 시공간 어텐션 맵을 생성하는 단계는,

상기 제n 인코더로부터 출력되는 상기 제n 특징맵을 입력받은 후 서로 다른 프레임에서 동일한 공간적 위치를 갖는 패치들을 비교하여 시간적 어텐션 맵을 생성하는 단계;

상기 시간적 셀프 어텐션 모듈 후단에 위치되며, 상기 시간적 셀프 어텐션 모듈의 출력을 이용하여 공간적 어텐션 맵을 생성하는 단계; 및

상기 시간적 어텐션 맵과 상기 공간적 어텐션 맵을 더하여 상기 제n 시공간 어텐션 맵을 생성하는 단계를 포함하는 것을 특징으로 하는 이상 상황 검지 방법.

청구항 9

제6 항 내지 제8 항 중 어느 하나의 항에 따른 방법을 수행하기 위한 프로그램 코드를 기록한 컴퓨터로 판독 가능한 기록 매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 트랜스포머 기반 이상 상황 검지 장치 및 그 방법에 관한 것이다.

배 경 기 술

[0003] 감시 카메라는 다양한 장소에 설치되어 상황을 기록하고 예상치 못한 사건을 분석 및 예방하는데 있어 매우 유용한 역할을 한다. 그러나, 사람이 매 순간을 모니터링하거나 수 많은 감시 영상을 분석하여 사건을 찾아내는

것은 비효율적이다. 따라서, 감시 카메라 영상에서 비정상적 이상 상황을 감지하기 위한 다양한 연구가 딥러닝 분야에서 활발하게 진행 중이다.

[0004] 그러나, 종래의 딥러닝 기술 기반 이상 상황 검지 기술은 시간적 특징을 제대로 반영하지 못하는 단점이 있다.

발명의 내용

해결하려는 과제

[0006] 본 발명은 트랜스포머 기반 이상 상황 검지 장치 및 그 방법을 제공하기 위한 것이다.

[0007] 또한, 본 발명은 장거리 종속성을 반영하기 위해 트랜스포머를 인코더로 활용하여 다중 스케일 특징맵 추출이 가능한 트랜스포머 기반 이상 상황 검지 장치 및 그 방법을 제공하기 위한 것이다.

[0008] 또한, 본 발명은 다중 스케일 특징맵의 공간 영역과 시간 영역의 특징 활용이 가능하며, 이를 반영하여 프레임 예측이 가능한 트랜스포머 기반 이상 상황 검지 장치 및 그 방법을 제공하기 위한 것이다.

과제의 해결 수단

[0010] 본 발명의 일 측면에 따르면 트랜스포머 기반 이상 상황 검지 장치가 제공된다.

[0011] 본 발명의 일 실시예에 따르면, 피라미드 구조를 가지는 n개의 인코딩 블록을 포함하며, 각각의 인코딩 블록을 통해 피라미드 구조를 가지는 다중 스케일 특징맵을 각각 추출하는 인코더부; n개의 시공간 트랜스포머를 가지되, 각각의 시공간 트랜스포머를 통해 상기 각 인코딩 블록에서 추출된 특징맵에 대한 시간 특징과 공간 특징이 강조된 다중 스케일 시공간 어텐션 맵을 각각 생성하는 시공간 트랜스포머부; 및 역피라미드 구조를 가지는 n개의 디코더를 포함하며, 상기 n개의 디코더를 통해 상기 다중 스케일 시공간 어텐션 맵과 이전 계층 디코더의 출력을 이용하여 예측 프레임을 생성하는 디코더부를 포함하되, 상기 예측 프레임을 이용하여 비정상 이벤트를 검출하는 것을 특징으로 하는 이상 상황 검지 장치가 제공될 수 있다.

[0012] 상기 제n 디코더는 상기 이전 계층의 디코더 출력을 업샘플링한 결과와 상기 제n 시공간 트랜스포머로부터 출력되는 제n 시공간 어텐션 맵을 결합한 후 컨볼루션 레이어를 적용할 수 있다.

[0013] 상기 제n 트랜스포머는, 상기 제n 인코더로부터 출력되는 상기 제n 특징맵을 입력받은 후 서로 다른 프레임에서 동일한 공간적 위치를 갖는 패치들을 비교하여 시간적 어텐션 맵을 생성하는 시간적 셀프 어텐션(temporal self-attention) 모듈; 및 상기 시간적 셀프 어텐션 모듈 후단에 위치되며, 상기 시간적 셀프 어텐션 모듈의 출력을 이용하여 공간적 어텐션 맵을 생성하는 공간 셀프 어텐션(spatial self-attention) 모듈을 포함하되, 상기 시간적 어텐션 맵과 상기 공간적 어텐션 맵을 더하여 상기 제n 시공간 어텐션 맵이 생성될 수 있다.

[0014] 상기 공간 셀프 어텐션 모듈은 공간 셀프 어텐션 적용 전에 공간 축소 레이어를 이용하여 공간 축소를 적용할 수 있다.

[0015] 상기 제n 시공간 어텐션 맵은 피드 포워드 네트워크 모듈을 통해 디코더로 전달되되, LN(Layer Norm) 적용된 후 상기 피드 포워드 네트워크 모듈로 전달될 수 있다.

[0017] 본 발명의 다른 측면에 따르면, 트랜스포머 기반 이상 상황 검지 방법이 제공된다.

[0018] 본 발명의 일 실시예에 따르면, (a) 피라미드 구조를 가지는 n개의 인코딩 블록을 포함하며, 각각의 인코딩 블록을 통해 피라미드 구조를 가지는 다중 스케일 특징맵을 각각 추출하는 단계; (b) n개의 시공간 트랜스포머를 가지되, 각각의 시공간 트랜스포머를 통해 상기 각 인코딩 블록에서 추출된 특징맵에 대한 시간 특징과 공간 특징이 강조된 다중 스케일 시공간 어텐션 맵을 각각 생성하는 단계; 및 (c) 역피라미드 구조를 가지는 n개의 디코더를 포함하며, 상기 n개의 디코더를 통해 상기 다중 스케일 시공간 어텐션 맵과 이전 계층 디코더의 출력을 이용하여 예측 프레임을 생성하는 단계; 및 상기 예측 프레임을 이용하여 스코어를 계산하고, 상기 스코어가 임계치 이상인 경우 이상 물체가 존재하는 것으로 판단하는 단계를 포함하는 이상 상황 검지 방법이 제공될 수 있다.

[0019] 상기 제n 디코더는 상기 이전 계층의 디코더 출력을 업샘플링한 결과와 상기 제n 시공간 트랜스포머로부터 출력되는 제n 시공간 어텐션 맵을 결합한 후 컨볼루션 레이어를 적용할 수 있다.

[0020] 상기 제n 시공간 어텐션 맵을 생성하는 단계는, 상기 제n 인코더로부터 출력되는 상기 제n 특징맵을 입력받은 후 서로 다른 프레임에서 동일한 공간적 위치를 갖는 패치들을 비교하여 시간적 어텐션 맵을 생성하는 단계; 상기 시간적 셀프 어텐션 모듈 후단에 위치되며, 상기 시간적 셀프 어텐션 모듈의 출력을 이용하여 공간적 어텐션 맵을 생성하는 단계; 및 상기 시간적 어텐션 맵과 상기 공간적 어텐션 맵을 더하여 상기 제n 시공간 어텐션 맵을 생성하는 단계를 포함할 수 있다.

발명의 효과

[0022] 본 발명의 일 실시예에 따른 트랜스포머 기반 이상 상황 검지 장치 및 그 방법을 제공함으로써, 장거리 종속성을 반영하기 위해 트랜스포머를 인코더로 활용하여 다중 스케일 특징맵 추출이 가능한 이점이 있다.

[0023] 또한, 본 발명은 다중 스케일 특징맵의 공간 영역과 시간 영역의 특징 활용이 가능하며, 이를 반영하여 프레임 예측이 가능한 이점이 있다.

도면의 간단한 설명

[0025] 도 1은 본 발명의 일 실시예에 따른 트랜스포머 기반 이상 상황 검지 장치의 구성을 개략적으로 도시한 도면.

도 2는 본 발명의 일 실시예에 따른 트랜스포머 기반 오토인코더 모델을 도시한 도면.

도 3은 본 발명의 일 실시예에 따른 인코딩 블록의 상세 구조를 도시한 도면.

도 4는 본 발명의 일 실시예에 따른 시공간 트랜스포머의 구조를 도시한 도면.

도 5는 본 발명의 일 실시예에 따른 트랜스포머 기반 이상 상황 검지 방법을 나타낸 순서도.

도 6은 종래와 본 발명의 일 실시예에 따른 ROC 커브 결과를 비교한 도면.

도 7은 종래와 본 발명의 일 실시예에 따른 AUC 커브 결과를 비교한 결과를 도시한 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 이상 상황 검지 장치의 구성을 개략적으로 도시한 도면이고, 도 2는 본 발명의 일 실시예에 따른 트랜스포머 기반 오토인코더 모델을 도시한 도면이고, 도 3은 본 발명의 일 실시예에 따른 인코딩 블록의 상세 구조를 도시한 도면이며, 도 4는 본 발명의 일 실시예에 따른 시공간 트랜스포머의 구조를 도시한 도면이다.

[0030] 도 1을 참조하면, 본 발명의 일 실시예에 따른 트랜스포머 기반 이상 상황 검지 장치(100)는 영상 획득부(110), 인코더부(120), 시공간 트랜스포머부(130), 디코더부(140), 메모리(150) 및 프로세서(160)를 포함하여 구성된다.

[0031] 영상 획득부(110)는 감시 카메라 영상을 획득하기 위한 수단이다.

[0032] 인코더부(120)는 영상 획득부(110)를 통해 획득된 감시 카메라 영상에 대한 서로 다른 해상도를 가지는 다중 스

케일 특징맵을 추출하기 위한 수단이다.

[0033] 인코더부(120)는 피라미드 구조를 가지는 n개의 인코딩 블록을 포함하며, n개의 인코딩 블록을 통해 프레임에 대한 다중 스케일 특징맵을 추출할 수 있다. 인코더부(120)의 구조는 도 2에 도시된 바와 같다. 즉, 인코더부(120)는 상향식 계층 구조를 가지는 n개의 인코딩 블록으로 구성된다. 각 인코딩 블록은 감시 카메라 영상 또는 /및 이전 인코딩 블록의 특징맵에 대해 컨볼루션 연산을 수행하여 다른 해상도를 가지는 특징맵을 각각 추출할 수 있다.

[0034] 도 3에는 각 인코딩 블록의 상세 구조가 도시되어 있다. 도 3을 참조하여 각 인코딩 블록의 동작에 대해 간략하게 설명하기로 한다.

[0035] 도 3을 참조하면, 각 인코딩 블록은 패치 임베딩부(310) 및 트랜스포머 블록(320)을 각각 포함한다.

[0036] 패치 임베딩부(310)는 H x W x 3 크기를 가지는 입력 프레임에 S의 스트라이드(stride)를 가지는 컨볼루션을 적용하여 $\frac{H}{S} \times \frac{W}{S} \times C_1$ 크기를 가지는 임베딩 패치들을 생성할 수 있다.

[0037] 트랜스포머 블록(320)은 패치 임베딩부(310)에서 출력되는 임베딩 패치들을 이용하여 특징맵을 추출할 수 있다.

[0038] 예를 들어, 제1 인코딩 블록의 경우, $\frac{H}{S} \times \frac{W}{S} \times C_1$ 크기의 임베딩 패치들을 이용하여 $\frac{H}{4} \times \frac{W}{4} \times C_1$ 크기를 가지는 특징맵(F_1)을 생성할 수 있다.

[0039] 이와 유사하게, 제2 인코딩 블록, 제3 인코딩 블록, 제4 인코딩 블록의 특징맵 F_2, F_3, F_4은 이전 인코딩 블록의 출력을 제공받아 생성될 수 있다.

[0040] 각 트랜스포머 블록(320)은 도 3에 도시된 바와 같이, 공간-축소 어텐션(spatial-reduction attention) 블록(322)과 피드 포워드 네트워크 블록(324)으로 구성된다.

[0041] 임베딩 패치들은 공간-축소 어텐션 블록(322)과 피드 포워드 네트워크 블록(324)에 적용되기 전에, LN(layer nor) 블록을 통해 정규화될 수 있다. 즉, 임베딩 패치들은 LN 블록에 의해 임베딩 패치의 채널들을 각각 정규화할 수 있다.

[0042] LN 블록을 통과한 후 임베딩 패치들은 공간-축소 어텐션 블록(322)과 피드 포워드 네트워크 블록(324)을 통과할 수 있다.

[0043] 각각의 출력은 수학식 1 및 수학식 2와 같이 나타낼 수 있다.

수학식 1

[0044] $$z^{l} = \mathrm{SRA}(LN(z^{l-1})) + z^{l-1}$$

수학식 2

[0045] $$z^{l+1} = \mathrm{FFN}(LN(z^{l})) + z^{l}$$

[0046] 여기서, z^{l} 과 z^{l+1} 은 공간-축소 어텐션 블록의 출력과 피드 포워드 네트워크 블록의 출력을 각각 나타낸다.

[0047] 공간-축소 어텐션 블록(322)은 공간 규모를 줄이기 위해 키K와 값 V에 컨볼루션을 적용하여 어텐션 메커니즘의 계산을 줄일 수 있다. 어텐션 연산은 수학식 3과 같이 계산될 수 있다.

수학식 3

$$\text{Attention}(\mathbf{q},\mathbf{k},\mathbf{v}) = \text{Softmax}(\frac{\mathbf{q}\mathbf{k}^{\mathsf{T}}}{\sqrt{d_{head}}})\mathbf{v}$$

[0048]

[0049] 피드 포워드 네트워크 블록(324)은 두개의 선형 레이어(linear layer)로 구성되며, 깊이(depth)별 컨볼루션과 GELU 활성화 함수가 두개의 선형 레이어 사이에 추가된다.

[0050] 시공간 트랜스포머부(130)는 인코더부(120)의 출력인 특징맵을 입력받아 시간 영역과 공간 영역의 특징이 강조된 시공간 어텐션 맵을 생성하기 위한 수단이다. 이러한, 시공간 트랜스포머부(130)는 복수의 시공간 트랜스포머를 가지며, 각 인코딩 모듈의 출력인 특징맵을 시간 영역과 공간 영역에서 분석하여 다중 스케일 시공간 어텐션 맵을 생성할 수 있다.

[0051] 각 시공간 트랜스포머의 구성은 도 4에 도시된 바와 같다.

[0052] 도 4를 참조하면, 시공간 트랜스포머는 각 인코딩 모듈의 출력인 특징맵을 입력받고, 해당 특징맵에 시간 셀프 어텐션 블록(410)을 적용한 결과에 공간 셀프 어텐션 블록(420)을 적용할 수 있다. 이를 통해, 각 인코딩 모듈에서 출력된 특징맵을 시간 영역과 공간 영역의 특징이 강조된 시공간 어텐션 맵을 생성할 수 있다.

[0053] 시공간 트랜스포머는 모든 S 공간 위치와 T 시간 위치를 사용하여 셀프 어텐션을 계산하기 때문에 $O(T^2S^2)$ 복잡도를 가진다.

[0054] 즉, 이를 수학식으로 나타내면 수학식 4와 같다.

수학식 4

$$z_{s,t}^{l} = \sum_{t'=0}^{T-1}\sum_{s'=0}^{S-1} \text{Softmax}\left(\frac{\mathbf{q}_{s,t}^{l}\cdot\mathbf{k}_{s',t'}^{l}}{\sqrt{D_{head}}}\right)\mathbf{v}_{s',t'}^{l}, \begin{Bmatrix} s=0,\ldots,S-1 \\ t=0,\ldots,T-1 \end{Bmatrix}$$

[0055]

[0056] 여기서, 소프트맥스는 스프트맥스 활성화 함수를 나타낸다.

[0057] 시공간 어텐션의 계산 복잡도를 줄이기 위해, 시간 어텐션과 공간 어텐션을 각각 분리하여 적용하며, 이를 수학식으로 나타내면 수학식 5와 같이 나타낼 수 있다.

수학식 5

$$\tilde{z}_{s,t}^{l} = \sum_{t'=0}^{T-1} \mathrm{Softmax}\left(\frac{\mathbf{q}_{s,t}^{l} \cdot \mathbf{k}_{s,t'}^{l}}{\sqrt{D_{head}}}\right) \mathbf{v}_{s,t'}^{l}, \begin{Bmatrix} s = 0, ..., S-1 \\ t = 0, ..., T-1 \end{Bmatrix}$$

$$z_{s,t}^{l} = \sum_{s'=0}^{S-1} \mathrm{Softmax}\left(\frac{\tilde{\mathbf{q}}_{s,t}^{l} \cdot \tilde{\mathbf{k}}_{s',t}^{l}}{\sqrt{D_{head}}}\right) \tilde{\mathbf{v}}_{s',t}^{l}, \begin{Bmatrix} s = 0, ..., S-1 \\ t = 0, ..., T-1 \end{Bmatrix}$$

[0058]

[0059] 여기서, $\tilde{\mathbf{q}}_{s,t}^{l}, \tilde{\mathbf{k}}_{s',t}^{l}, \tilde{\mathbf{v}}_{s',t}^{l}$ 는 각각 $\tilde{z}_{s,t}^{l}$ 로부터 계산될 수 있다. 이와 같이 분리된 시-공간 어텐션은 계산 복잡도를 $O(T^2S^2)$ 에서 $O(T^2S + TS^2)$ 로 줄일 수 있다.

[0060] 따라서, 본 발명의 일 실시예에 따르면, 시공간 트랜스포머는 제n 인코딩 블록의 출력인 특징맵을 입력받은 후 시간 셀프 어텐션 모듈을 통해 서로 다른 프레임에서 동일한 공간적 위치를 갖는 패치를 비교하여 시간적 어텐션 맵을 생성한다.

[0061] 시간 셀프 어텐션 모듈의 출력인 시간적 어텐션 맵은 공간 어텐션 모듈로 입력되어 시공간 어텐션 맵이 생성될 수 있다. 각 인코딩 블록 이후에 시공간 트랜스포머가 적용되기 때문에 서로 다른 해상도의 특징맵을 사용하여 셀프 어텐션이 계산될 수 있다. 따라서, 본 발명의 일 실시예에서는 공간 어텐션 모듈을 적용하기 전에 공간 축소 레이어를 적용한다. 키 K와 값 V의 공간 스케일은 공간 셀프 어텐션 모듈에 적용되기 전에 공간 축소 레이어에 의해 축소될 수 있다.

[0062] 공간 셀프 어텐션 모듈의 계산 비용은 K와 V의 공간 척도를 줄임으로써 크게 감소될 수 있다.

[0063] 시간적 셀프 어텐션과 공간적 셀프 어텐션은 각각 수학식 6과 같이 계산될 수 있다.

수학식 6

$$\tilde{z}_{s,t}^{l} = \sum_{t'=0}^{T-1} \mathrm{Softmax}\left(\frac{\mathbf{q}_{s,t}^{l} \cdot \mathbf{k}_{s,t'}^{l}}{\sqrt{D_{head}}}\right) \mathbf{v}_{s,t'}^{l}, \begin{Bmatrix} s = 0, ..., S-1 \\ t = 0, ..., T-1 \end{Bmatrix}$$

$$z_{s,t}^{l} = \sum_{s'=0}^{S-1} \mathrm{Softmax}\left(\frac{\tilde{\mathbf{q}}_{s,t}^{l} \cdot \hat{\mathbf{k}}_{s',t}^{l}}{\sqrt{D_{head}}}\right) \hat{\mathbf{v}}_{s',t}^{l}, \begin{Bmatrix} s = 0, ..., S-1 \\ t = 0, ..., T-1 \end{Bmatrix}$$

[0064]

[0065] 여기서, $\tilde{\mathbf{q}}_{s,t}^{l}$ 는 $\tilde{z}_{s,t}^{l}$ 로부터 파생되고, $\hat{\mathbf{k}}_{s',t}^{l}$ 와 $\hat{\mathbf{v}}_{s',t}^{l}$ 는 $\tilde{z}_{s,t}^{l}$ 로부터 계산된다. 그 이후,

공간 축소 레이어에 의해 공간 척도가 축소된다.

[0066] 이후, $\tilde{z}_{s,t}^{l}$ 는 최종 인코딩 결과를 획득하기 위해 피드 포워드 네트워크 블록(430)으로 전달된다.

[0067] 완전 연결 레이어는 입력 벡터의 특징 차원을 증가시키고, GELU 함수는 첫번째 완전 연결 레이어 이후에 활성화 함수로 사용된다. 또한, 도 4에 도시된 바와 같이, 피드 포워드 네트워크 블록(430)에는 로컬 컨텍스트를 활용하기 위해 깊이(depth)별 컨볼루션 블록이 추가된다. 두번째 완전 연결 레이어는 입력 채널의 크기와 일치하도록 채널을 줄일 수 있다.

[0068] 다시, 도 1을 참조하면, 디코더부(140)는 역피라미드 구조를 가지는 복수의 디코딩 블록을 가지며, 다중 스케일 특징맵 또는/및 다중 스케일 시공간 어텐션 맵과 이전 디코딩 블록의 출력을 이용하여 복원 영상을 생성한다.

[0069] 도 2에서 보여지는 바와 같이, 다른 계층의 다중 스케일 특징맵은 특징맵의 공간 해상도를 복원하기 위해 건너뛰기 연결을 통해 디코더로 전달된다. 즉, 이전 단계의 특징맵은 디컨볼루션 레이어를 적용하여 업샘플링된 후 시공간 트랜스포머의 출력인 시공간 어텐션 맵과 결합된다. 결합된 특징맵은 3 x 3 컨볼루션, 배치 정규화 및 RELU 활성화 함수로 구성된 컨볼루션 레이어를 통해 전달될 수 있다.

[0070] 본 발명의 일 실시예에 따른 네트워크 모델에 대한 목적함수에 대해 설명하기로 한다. 예를 들어, 입력 프레임의 시퀀스가 $\{I_1, I_2, \ldots, I_t\}$ 와 같다고 가정하기로 한다.

[0071] 본 발명의 일 실시예에 따른 모델은 실제 프레임 I_{t+1} 에 해당하는 예측 프레임($\hat{I}_{t+1}$)을 예측하는 것을 목표로 한다. 예측 프레임($\hat{I}_{t+1}$)을 실제 프레임(I_{t+1})과 유사하게 하기 위해 세가지 다른 손실 함수가 사용된다.

[0072] RGB 공간에서 모든 픽셀의 유사성을 보장하기 위해 수학식 7과 같이 예측 프레임($\hat{I}_{t+1}$)과 실제 프레임(I_{t+1}) 사이에 l_2 손실이 적용된다.

수학식 7

$$L_{int}(I, \hat{I}) = \left\| I - \hat{I} \right\|_2^2$$

[0073]

[0074] 그러나, l_2 손실을 사용하는 경우, 출력이 흐려진다는 단점이 있다. 따라서, 더 선명한 예측 프레임을 얻기 위해 기울기 손실을 추가적으로 사용한다. 기울기 손실 함수는 예측된 프레임의 두 공간 차원을 따라 절대 기울기와 정답 사이의 기울기를 수학식 8과 같이 계산한다.

수학식 8

$$L_{gra}(I,\hat{I}) = \sum_{i,j} \left\| \left|\hat{I}_{i,j} - \hat{I}_{i-1,j}\right| - \left|I_{i,j} - I_{i-1,j}\right| \right\|_1 + \left\| \left|\hat{I}_{i,j} - \hat{I}_{i,j-1}\right| - \left|I_{i,j} - I_{i,j-1}\right| \right\|_1$$

[0075]

[0076] 또한, MS-SSIM(Multi-Scale Structuralsimilarity)를 이용하여 구조적 차이를 측정한다. MS-SSIM은 다른 해상도에서 영상의 구조적 유사성으로 추정한다. 따라서, 최종 손실은 수학식 9와 같이 강도 손실(L_{int}), 기울기 손실(L_{gra}), 다중 스케일 구조적 손실(L_{mss})로 계산된다.

수학식 9

$$L_{pre}(I,\hat{I}) = \alpha L_{int}(I,\hat{I}) + \beta L_{gra}(I,\hat{I}) + \gamma L_{mss}(I,\hat{I})$$

[0077]

[0078] 여기서, α, β, γ 는 손실 함수의 가중치 균형을 유지하는 계수이다.

[0079] 메모리(150)는 본 발명의 일 실시예에 따른 트랜스포머 기반 이상 상황 검지 방법을 수행하기 위한 프로그램 코드를 저장한다.

[0080] 프로세서(160)는 본 발명의 일 실시예에 따른 이상 상황 검지 장치(100)의 내부 구성 요소들(예를 들어, 영상 획득부(110), 인코더부(120), 시공간 트랜스포머부(130), 디코더부(140), 메모리(150) 등)을 제어하기 위한 수단이다.

[0081] 또한, 프로세서(160)는 예측 프레임을 이용하여 비정상 이벤트를 검출할 수 있다. 이에 대해 보다 상세히 설명하기로 한다.

[0082] 본 발명의 일 실시예에 따른 인코더부(120), 시공간 트랜스포머부(130) 및 디코더부(140)는 정상적인 이벤트로 학습되기 때문에 정상적인 이벤트를 잘 예측한다. 예측 프레임과 실제 프레임 사이의 점수 차이는 비정상 이벤트의 예측 프레임과 실제 프레임 사이의 차이가 크며 그 반대의 경우도 마찬가지다.

[0083] PSNR(피크 신호 대 잡음비)는 이미지 품질을 추정하는데 이용되며, PSNR 값이 클수록 프레임의 품질이 높다. 즉, 예측된 프레임의 PSNR 값이 높으면, 실제 프레임과의 차이가 작다. 결과적으로, 이와 같은 경우, 정상적인 이벤트일 가능성이 높다.

[0084] PSNR은 수학식 10을 이용하여 계산될 수 있다.

수학식 10

$$\mathrm{PSNR}(I,\hat{I}) = 10log_{10}\frac{[max_{\hat{I}}]^2}{\frac{1}{N}\sum_{i=1}^{N}(I_i-\hat{I}_i)^2}$$

[0085]

[0086] 여기서, N은 프레임의 픽셀수를 나타내고, $[max_{\hat{I}}]$ 는 예측 프레임($\hat{I}$)의 최대값을 나타낸다.

[0087] 모든 프레임의 PSNR값을 [0, 1] 범위로 정규화한 다음 수학식 11을 이용하여 프레임의 이상 점수를 계산할 수 있다.

수학식 11

$$S(t) = \frac{\mathrm{PSNR}_t - min(\mathrm{PSNR})}{max(\mathrm{PSNR}) - min(\mathrm{PSNR})}$$

[0088]

[0089] 여기서, $min(\mathrm{PSNR})$ 는 프레임 시퀀스의 최소 PSNR값을 나타내고, $max(\mathrm{PSNR})$ 는 프레임 시퀀스의 최대 PSNR값을 나타낸다.

[0090] 이상 점수 $S(t)$ 는 프레임이 주어진 임계값으로 정상 또는 비정상인지 여부를 추정하는데 사용된다.

[0092] 도 5는 본 발명의 일 실시예에 따른 트랜스포머 기반 이상 상황 검지 방법을 나타낸 순서도이고, 도 6은 종래와 본 발명의 일 실시예에 따른 ROC 커브 결과를 비교한 도면이고, 도 7은 종래와 본 발명의 일 실시예에 따른 AUC 커브 결과를 비교한 결과를 도시한 도면이다.

[0093] 단계 510에서 이상 상황 검지 장치(100)는 피라미드 구조를 가지는 n개의 인코딩 블록을 포함하며, 각각의 인코딩 블록을 통해 피라미드 구조를 가지는 다중 스케일 특징맵을 각각 추출한다.

[0094] 단계 515에서 이상 상황 검지 장치(100)는 n개의 시공간 트랜스포머를 가지되, 각각의 시공간 트랜스포머를 통해 상기 각 인코딩 블록에서 추출된 특징맵에 대한 시간 특징과 공간 특징이 강조된 다중 스케일 시공간 어텐션 맵을 각각 생성한다.

[0095] 단계 520에서 이상 상황 검지 장치(100)는 역피라미드 구조를 가지는 n개의 디코더를 포함하며, 상기 n개의 디코더를 통해 상기 다중 스케일 시공간 어텐션 맵과 이전 계층 디코더의 출력을 이용하여 예측 프레임을 생성한다.

[0096] 단계 525에서 이상 상황 검지 장치(100)는 예측 프레임에 대한 이상 스코어를 계산하고, 이상 스코어를 이용하여 비정상 이벤트를 감지한다.

[0097] 도 6은 종래와 본 발명의 일 실시예에 따른 ROC 커브 결과를 비교한 도면이고, 도 7은 종래와 본 발명의 일 실시예에 따른 AUC 커브 결과를 비교한 결과이다.

[0098] 도 6 및 도 7에서 보여지는 바와 같이, 시공간 트랜스포머를 적용한 결과 성능이 더 좋은 것을 알 수 있다.

[0100] 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.

[0101] 상술한 하드웨어 장치는 본 발명의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0102] 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

[0104] 100: 이상 상황 검지 장치

110: 영상 획득부

120: 인코더부

130: 시공간 트랜스포머부

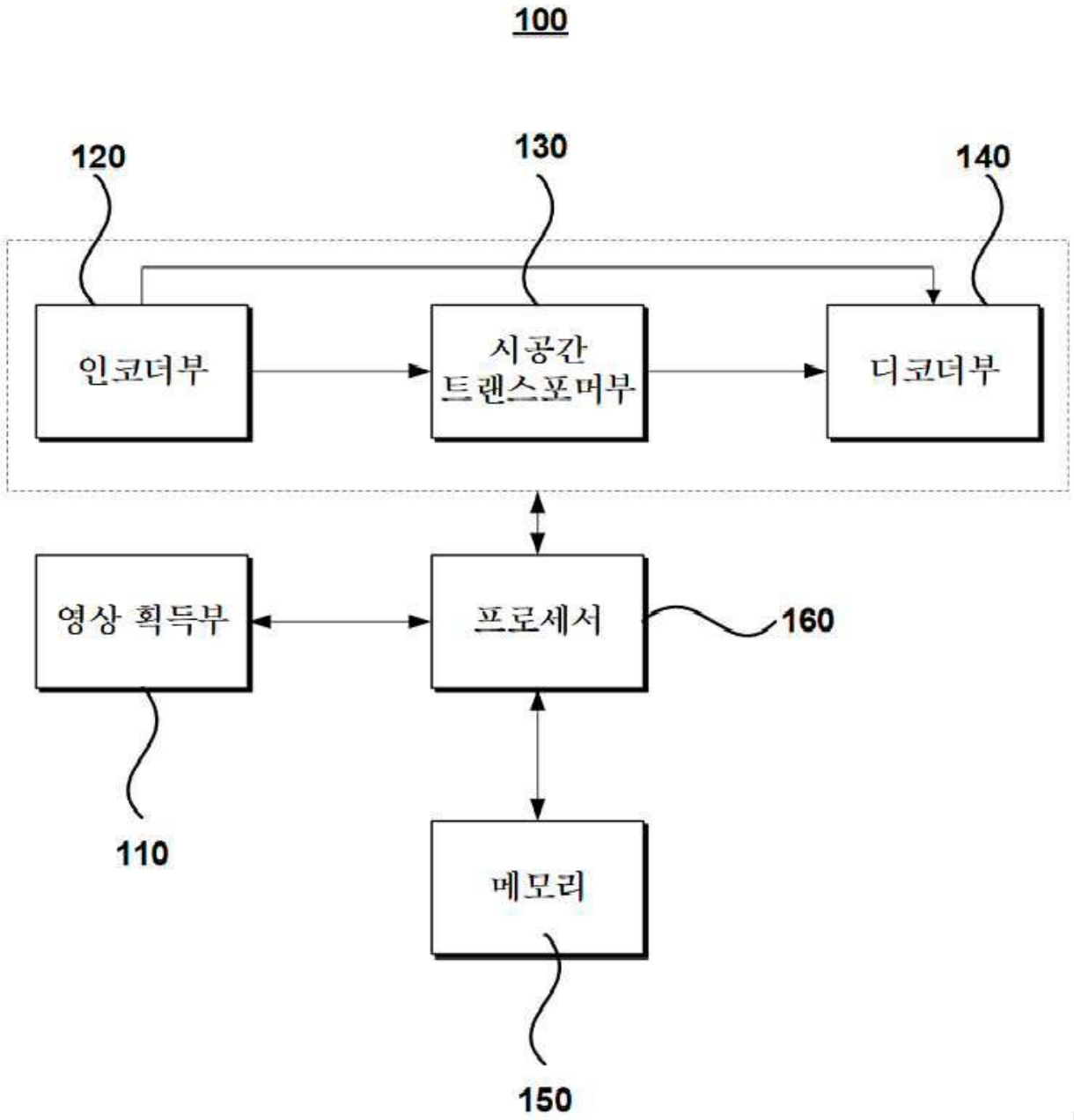
140: 디코더부

150: 메모리

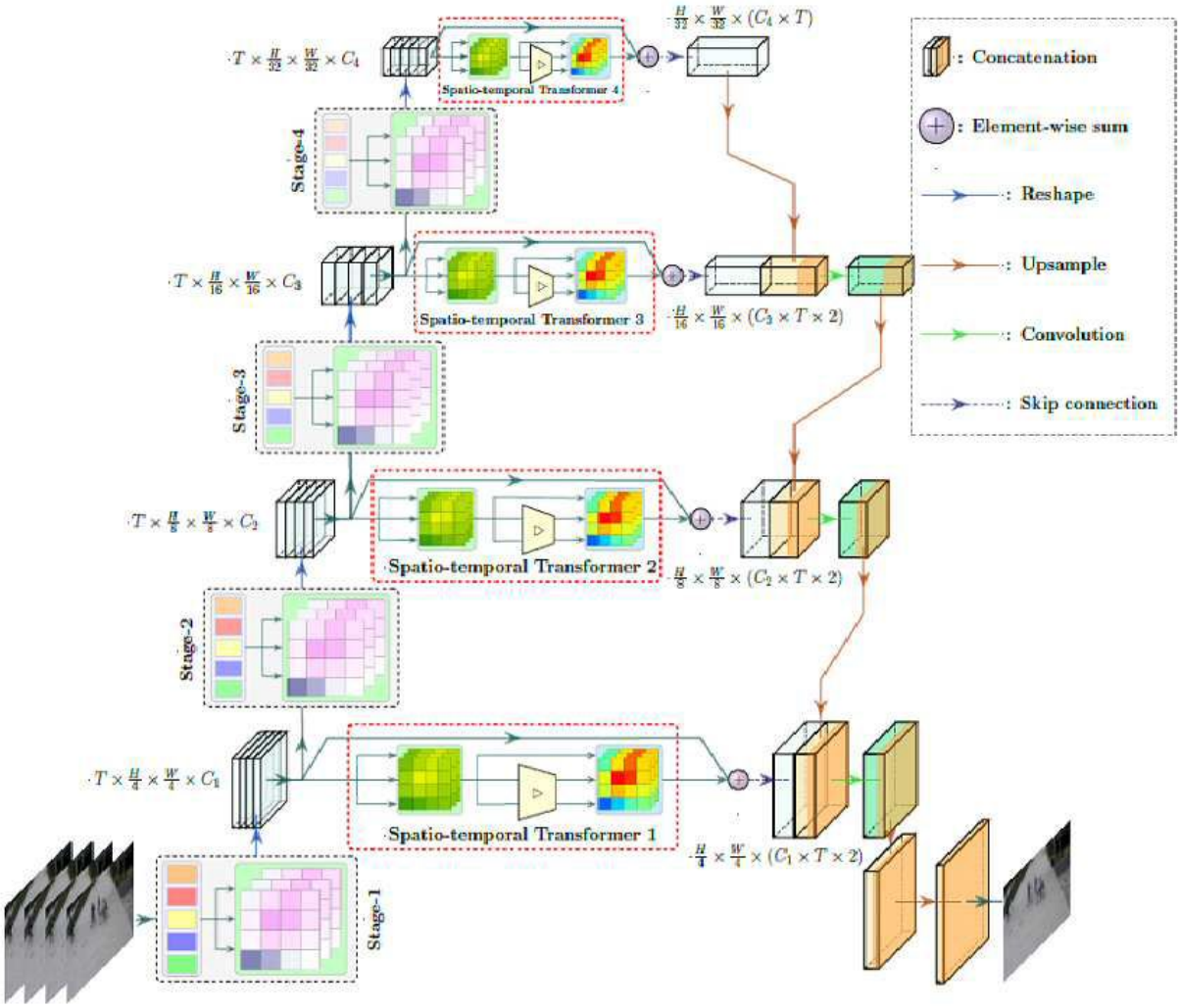
160: 프로세서

도면

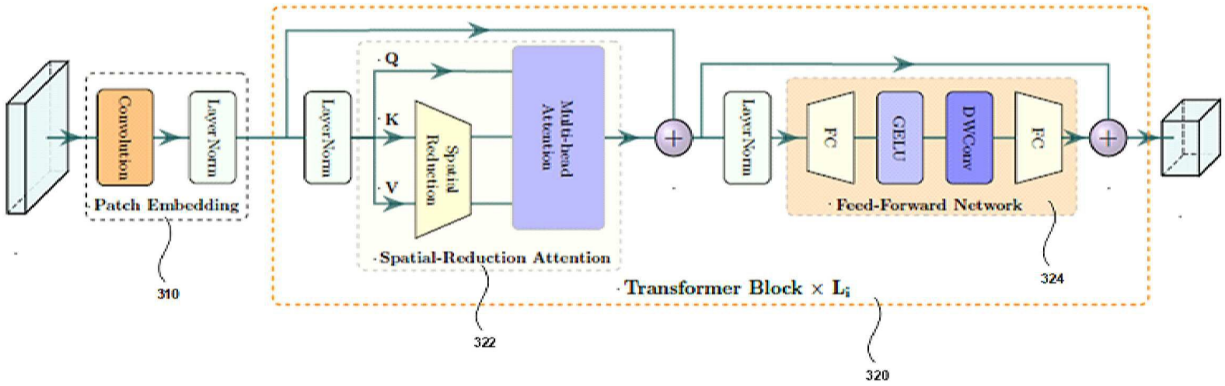
도면1


100
120
130
140
인코더부
시공간
트랜스포머부
디코더부
영상 획득부
프로세서
160
110
메모리
150

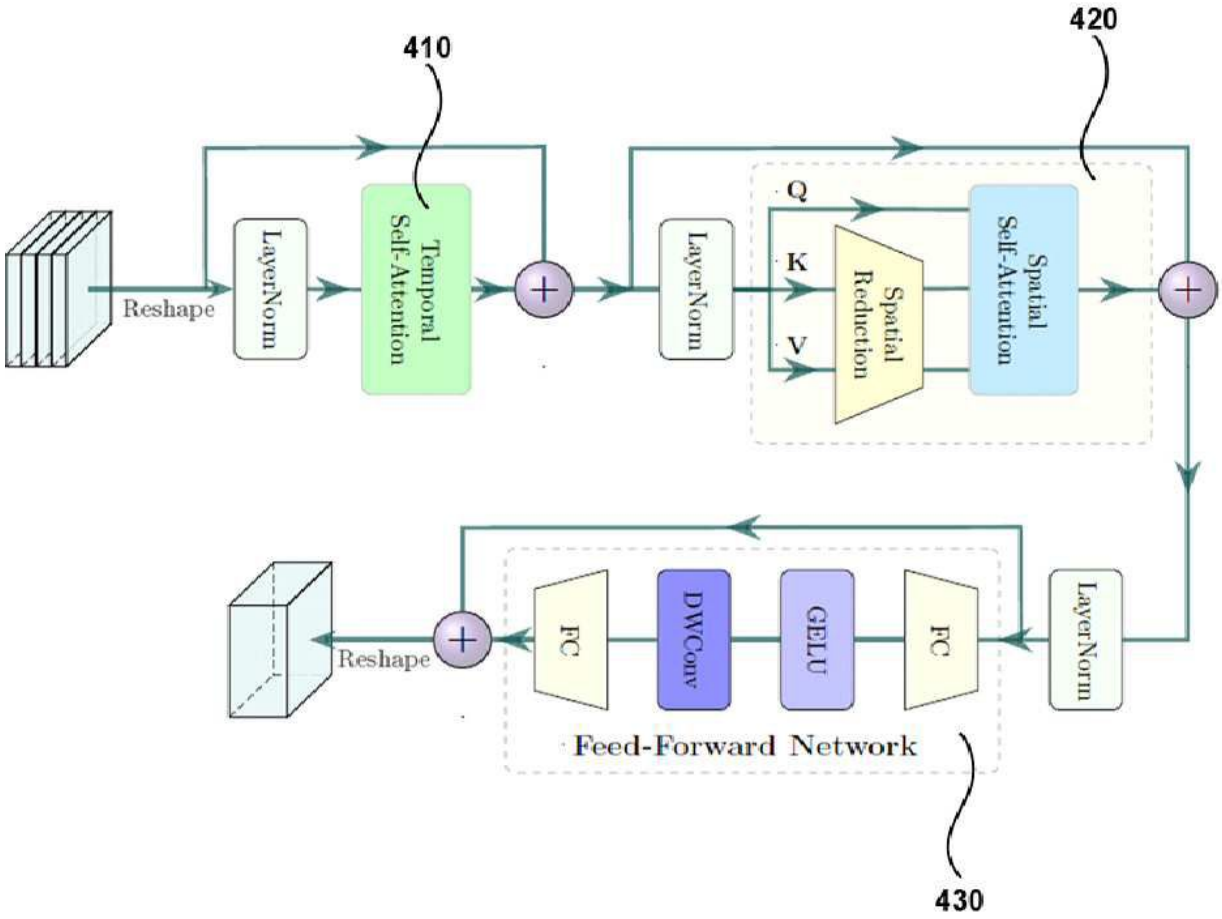
도면2


Stage-4
Stage-3
Stage-2
Stage-1
Spatio-temporal Transformer 4
Spatio-temporal Transformer 3
Spatio-temporal Transformer 2
Spatio-temporal Transformer 1
: Concatenation
: Element-wise sum
: Reshape
: Upsample
: Convolution
: Skip connection

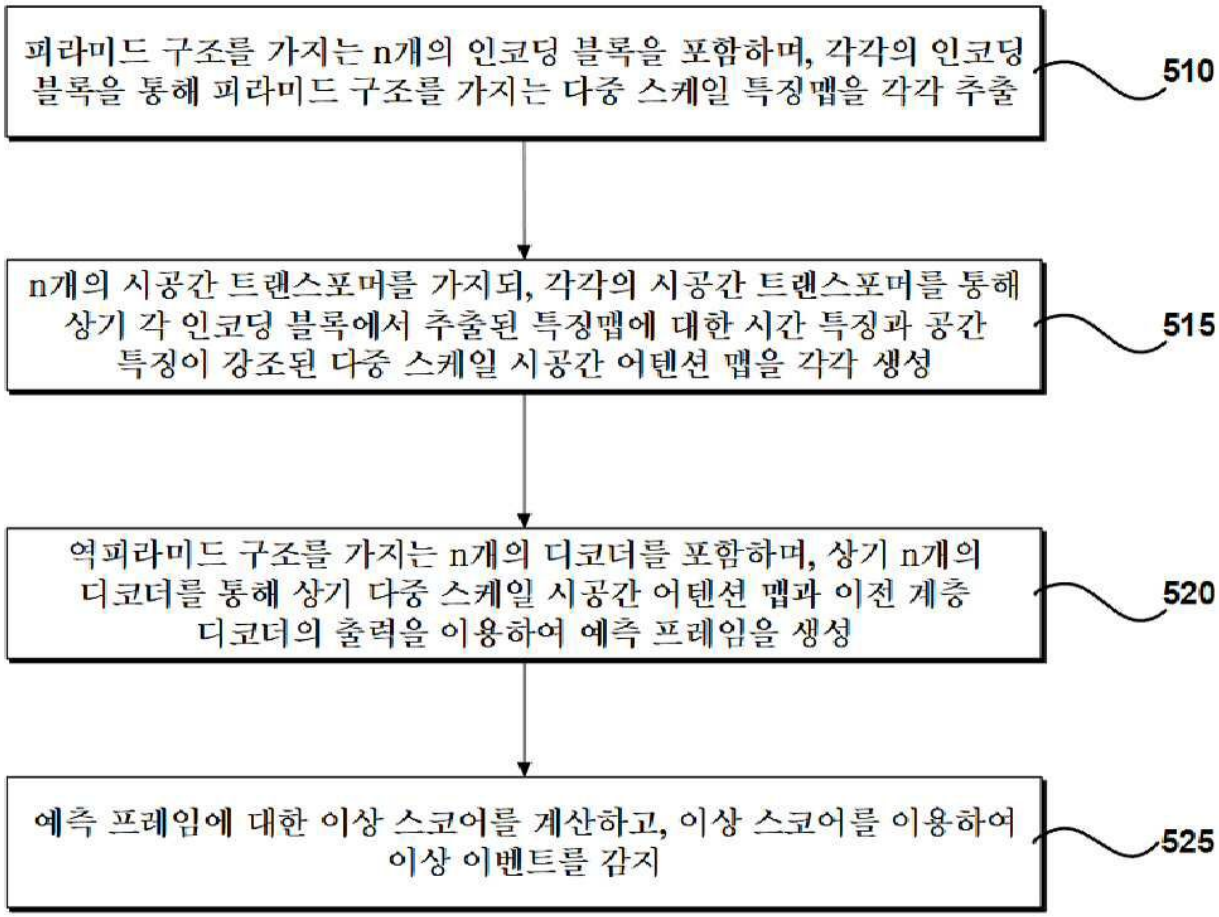
도면3


Convolution
LayerNorm
Patch Embedding
LayerNorm
Q
K
V
Spatial Reduction
Multi-head Attention
Spatial-Reduction Attention
LayerNorm
FC
GELU
DWConv
FC
Feed-Forward Network
Transformer Block × Li
310
322
320
324

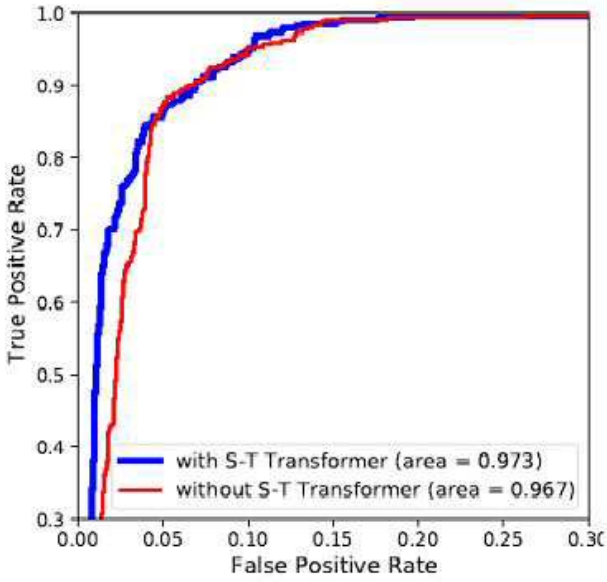
도면4


410
420
Reshape
LayerNorm
Temporal Self-Attention
LayerNorm
Q
K
V
Spatial Reduction
Spatial Self-Attention
LayerNorm
FC
DWConv
GELU
FC
Reshape
Feed-Forward Network
430

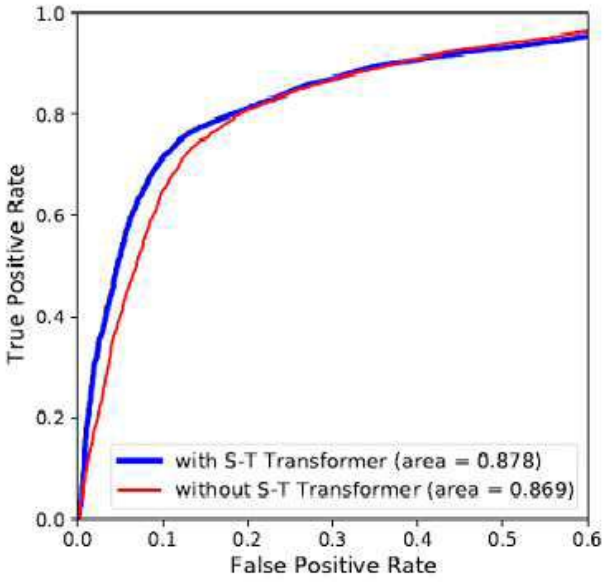
도면5


피라미드 구조를 가지는 n개의 인코딩 블록을 포함하며, 각각의 인코딩 블록을 통해 피라미드 구조를 가지는 다중 스케일 특징맵을 각각 추출
510
n개의 시공간 트랜스포머를 가지되, 각각의 시공간 트랜스포머를 통해 상기 각 인코딩 블록에서 추출된 특징맵에 대한 시간 특징과 공간 특징이 강조된 다중 스케일 시공간 어텐션 맵을 각각 생성
515
역피라미드 구조를 가지는 n개의 디코더를 포함하며, 상기 n개의 디코더를 통해 상기 다중 스케일 시공간 어텐션 맵과 이전 계층 디코더의 출력을 이용하여 예측 프레임을 생성
520
예측 프레임에 대한 이상 스코어를 계산하고, 이상 스코어를 이용하여 이상 이벤트를 감지
525

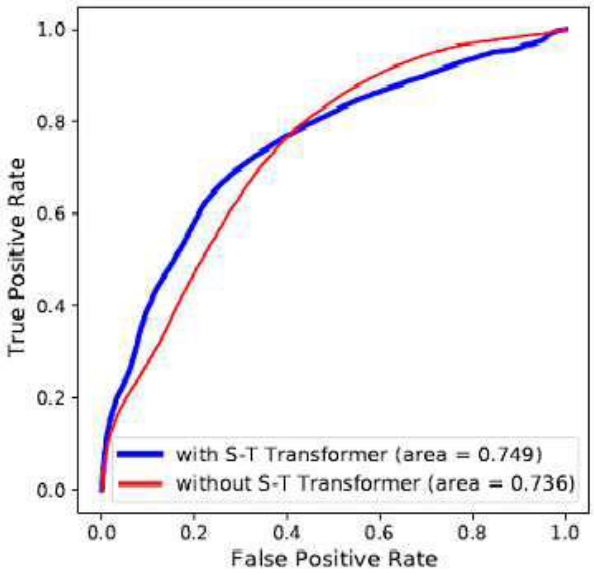
도면6



(a) UCSD Ped2.



(b) CUHK Avenue.



(c) ShanghaiTech.

도면7

Method	UCSD Ped2	CUHK Avenue	ShanghaiTech
Yuan et al. [8] (2021)	96.4	87.0	-
Cai et al. [16] (2021)	96.6	86.6	73.7
Deepak et al. [1] (2021)	83.0	82.0	-
Luo et al. [14] (2021)	-	87.3	74.1
Wu et al. [13] (2021)	92.4	-	-
Luo et al. [3] (2021)	96.2	85.7	73.0
Yang et al. [2] (2021)	93.7	83.2	-
Wang et al. [19] (2021)	96.0	86.3	74.5
Hao et al. [5] (2022)	96.9	86.6	73.8
Chang el al. [15] (2022)	96.7	87.1	73.7
Li et al. [17] (2022)	96.3	87.1	73.6
Guo et al. [12] (2022)	88.1	86.6	-
Liu et al. [4] (2022)	96.6	86.5	-
Wang et al. [20] (2022)	89.6	84.6	70.2
Li et al. [7] (2022)	95.4	86.0	71.4
Borja-Borja [18] (2022)	95.0	83.9	-
Our	**97.3**	**87.8**	**74.9**