< 캡스톤디자인 최종보고서 >

프로젝트 명: DIRL 모델 기반의 Distractor-Immune Remote Sensing Change Captioning

팀원(참여학생): 김희동(소프트웨어융합학과)

요 약

본 프로젝트는 원격 감지(Remote Sensing) 이미지 쌍 사이의 변화를 감지하고 이를 자연어로 설명하는 Remote Sensing Change Captioning(RSICC) 작업을 다룬다. 위성 이미지 분석 시 빈번하게 발생하는 조명 변화나 촬영 시점(Viewpoint) 차이 등의 방해 요소(Distractor)를 효과적으로 극복하기 위해, 본 연구는 방해 요소에 강건한 기존 DIRL 모델을 기반으로 진행되었다. 본 연구의 핵심은 LEVIR-MCI 데이터셋이 제공하는 정교한 세그멘테이션(Segmentation) 마스킹 데이터를 학습에 적극 활용하여 객체의 구체적인 형상이나 픽셀 수준의 위치 정보를 직접적으로 학습하는 것이다. 이를 위해 기존 DIRL 모델에 보조적인 세그멘테이션 헤드(Auxiliary Head)를 결합한 멀티태스크 학습 방법론을 제안한다. 특히, 해당 보조 모듈(Auxiliary Module)의 디코더(Decoder)에는 U-Net 구조를 차용하고 ResNet 의 Skip Connection 기술을 적용하여, 변화 영역의 마스크 복원 성능을 최적화하고 객체 인지 능력을 강화하였다. 실험 결과, 제안된 모델은 기존 DIRL 베이스라인 대비 BLEU-4, METEOR 등의 주요 정량적 지표에서 성능 향상을 기록하였으며, 실제 변화 감지 능력 또한 시각적으로 유효함이 검증되었다.

Github Code: https://github.com/hee-dongdong/DIRL_LEVIR-MCI.git

1.서론

Remote Sensing Change Captioning(RSICC)은 시간 차를 두고 촬영된 두 위성 이미지 사이의 차이점을 분석하여 이를 자연어 문장으로 생성하는 작업이다. 이 작업은 단순한 변화 탐지를 넘어 변화의 의미적 내용을 서술해야 하는 난이도 높은 문제다. 특히 위성 이미지는 조명의

강도 차이, 촬영 시점(Viewpoint)의 변화, 블러 효과, 계절적 변화 등 실제 지형의 변화와는 무관한 다양한 방해 요소(Distractor)들이 존재하여 정확한 캡션 생성을 어렵게 만든다[1].

또한, 기존 캡션 생성 모델들은 변화의 특징을 추상적인 벡터 공간에서만 학습할 뿐, 변화 객체의 구체적인 형상이나 픽셀 수준의 위치 정보를 직접적으로 학습하지 않는 한계가 있다. 이는 모델이 캡션을 생성할 때 "이미지의 어떤 부분이 변했는지"에 대한 명확한 참조 점을 찾지 못하는 '시각적 근거(Visual Grounding)의 부재'로 이어진다.

본 프로젝트의 목표는 방해 요소에 강건한(Distractor-Immune) 특성을 가진 DIRL 모델[2]을 기반으로, 변화 영역에 대한 인지 능력을 더욱 강화하는 것이다. 더불어 변화 영역에 대한 픽셀 단위의 정보를 담고 있는 세그멘테이션 마스크(Ground Truth)를 학습에 활용하여, 모델이 변화가 일어난 위치와 형상을 더 정확히 학습하도록 유도하는 아키텍처를 설계하고 구현한다.

본 연구에서는 DIRL 모델을 베이스라인으로 하여, 세그멘테이션 보조 모듈(Auxiliary Segmentation Module)을 결합한 멀티태스크 학습 모델을 구현하고 LEVIR-MCI 데이터셋을 통해 성능을 검증하였다. 실험 결과, 제안 모델은 기존 베이스라인 대비 BLEU-4, METEOR, ROUGE-L, CIDEr, SPICE metric 에서 유의미한 성능 개선을 보였다.

본 프로젝트의 주요 기여점은 다음과 같이 세 가지로 요약할 수 있다.

첫째, 방해 요소(Distractor)에 강건한 특징 추출 및 시각적 근거(Visual Grounding) 강화를 위한 멀티태스크 학습 프레임워크 제안

본 연구는 조명 차이나 계절 변화와 같은 방해 요소에 강건한(Distractor-Immune) DIRL 구조를 기반으로 하되, 캡션 생성에 필요한 시각적 근거 정보를 보완하기 위해 변화 영역의 픽셀 단위 마스크(Ground Truth)를 활용하는 세그멘테이션 헤드(Auxiliary Head)를 도입하였다. 이를 통해 모델이 실제 변화가 아닌 방해 요소를 효과적으로 배제하고 "어디가 변했는지(Location)"를 명확히 인지하도록 유도함으로써, "무엇이 변했는지(Caption)"를 서술하는 캡션 생성의 인과성과 정확도를 동시에 강화하였다.

둘째, 정교한 마스크 복원을 위한 U-Net 및 Skip Connection 기반 아키텍처 설계

인코딩 과정에서 손실될 수 있는 공간 정보를 보존하고 미세한 변화 영역을 정확히 복원하기 위해, 보조 모듈의 디코더에 U-Net[3] 구조를 차용하고 Skip Connection[4]을 적용하였다. 이

구조는 저수준(Low-level)의 공간 정보를 디코더로 전달하여 건물의 모서리나 얇은 도로와 같은 세밀한 변화 Mask 에 대한 형상을 효과적으로 학습할 수 있게 했다.

셋째, Change Segmentation Mask 추론 보조 과제를 통한 주 과제의 성능 향상(Positive Transfer) 입증
캡션 생성(주 과제)과 세그멘테이션(보조 과제)을 동시에 학습시킬 때, 낮은 가중치(λ_x=0.05)의 보조 손실 함수를 추가하는 것만으로도 주 과제의 성능지표(BLEU-4, CIDEr 등)가 향상됨을 실험적으로 검증하였다. 이는 Change 에 대한 Segmentation Mask Information 을 추가로 주는 것이 자연어 생성이라는 고차원적 작업에 긍정적인 효과를 준다는 것을 입증한다.

2. 관련 연구

2.1 LEVIR-MCI

LEVIR-MCI[1]는 원격 감지 변화 캡션(RSICC) 작업을 위해 구축된 대규모 데이터셋으로, 10,077 쌍의 이미지와 50,385 개의 문장으로 구성되어 있다. 이 데이터셋의 가장 큰 특징은 단순히 변화 전후의 이미지와 캡션만을 제공하는 기존 데이터셋들과 달리, 변화 영역에 대한 정확한 픽셀 단위의 이진 마스크(Binary Segmentation Mask)를 함께 제공한다는 점이다.

본 연구에서는 LEVIR-MCI 가 제공하는 이 세그멘테이션 마스크를 모델 학습의 핵심 요소로 활용하였다. 구체적으로, 제안하는 멀티태스크 학습 프레임워크에서 이 마스크 데이터를 보조 과제(Auxiliary Task)의 정답 레이블(Ground Truth)로 사용하여, 모델이 변화가 발생한 구체적인 위치와 형상을 명시적으로 학습하도록 유도하였다. 이는 모델이 "무엇이(What)" 변했는지를 설명하는 캡션 생성 능력뿐만 아니라, "어디가(Where)" 변했는지를 인지하는 능력을 동시에 기르는 데 결정적인 역할을 한다.

2.2 DIRL (Distractors-Immune Representation Learning)

Tu et al.(2024)이 제안한 DIRL[2]은 위성 이미지 분석 시 빈번하게 발생하는 조명 변화나 계절적 차이와 같은 방해 요소(Distractor)를 제거하는 데 특화된 프레임워크다. DIRL 모델의

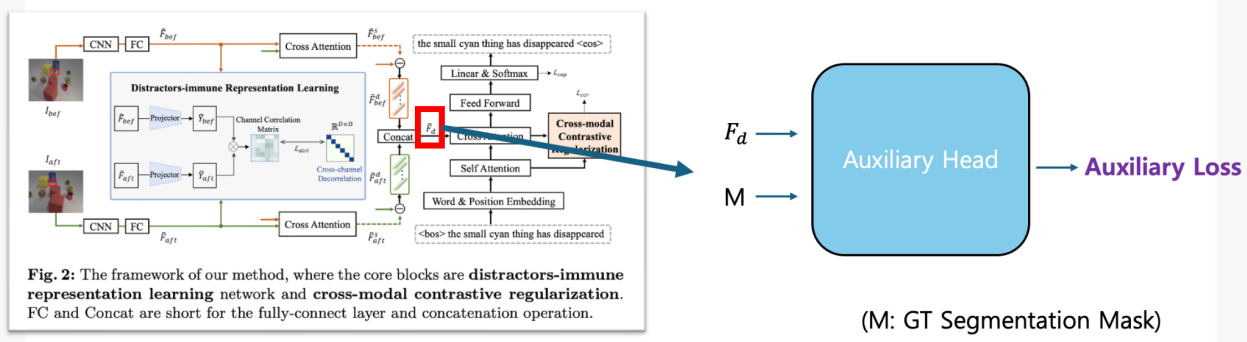
핵심 기전은 두 시점의 이미지(I_{bef} I_{aft})에서 변화하지 않은 안정적 특징(Stable Feature)을 억제하고, 순수한 변화 특징(F_d)만을 추출하는 것이다. 이를 위해 채널 간의 상관관계를 줄이는 L_{DIRL}(Cross-channel Decorrelation)과 Contrastive Learning 을 통해 텍스트가 이미지의 특정 객체를 더욱 discriminative 하게 포착할 수 있도록 하는 L_{CCR}(Cross-modal Contrastive Regularization) 손실 함수를 도입하여 특징 표현(Representation)의 강건성을 확보하였다.

그러나 기존의 DIRL 모델은 변화의 특징을 추상적인 벡터 공간에서 학습하여 캡션을 생성하는 데에는 효과적이나, 변화 객체의 구체적인 형상이나 픽셀 수준의 위치 정보를 직접적으로 학습하지 않는다는 한계가 있다. 즉, 모델이 변화를 인지할 때 "이미지의 어떤 부분이 변했는지"에 대한 명확한 시각적 근거(Visual Grounding)가 부족할 수 있다.

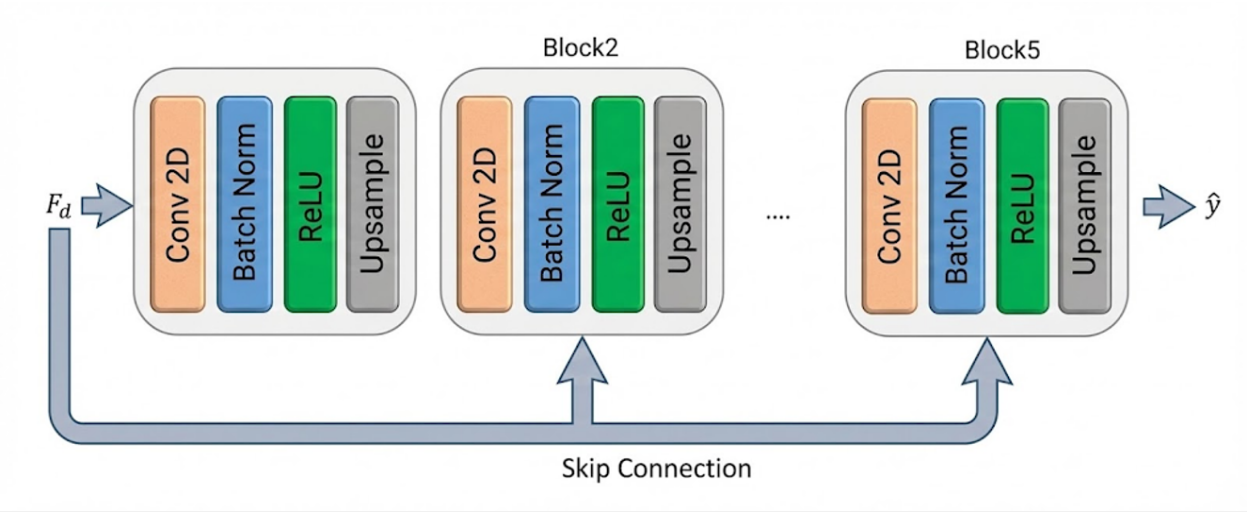
본 연구에서는 이러한 한계를 극복하기 위해 DIRL 구조를 백본(Backbone)으로 채택하되, 이를 확장하여 픽셀 수준의 지도 학습을 수행하는 구조로 개선하였다. 구체적으로, 변화 특징맵(F_d)을 입력받아 변화 영역을 복원하는 보조 세그멘테이션 헤드(Auxiliary Segmentation Head)를 추가하였다. 이를 통해 모델이 방해 요소에 강건한 특징을 추출함과 동시에, 변화 객체의 형상을 정교하게 인지하도록 하여 캡션 생성의 정확도와 신뢰성을 높였다.

3. 제안 방법 (Proposed Method)

[그림 1] GT Segmentation Mask (M)



[Fig 1] Overall Structure (DIRL + Auxiliary Head)



[Fig 2] Details of Auxiliary Head

3.1 전체 아키텍처 (Overall Architecture)

제안하는 모델은 크게 특징 추출기(Feature Extractor), 캡션 생성기(Caption Decoder), 그리고 본 연구에서 새롭게 추가된 보조 세그멘테이션 모듈(Auxiliary Segmentation Module)로 구성된다. 모델은 입력된 이미지 쌍에서 변화 특징 F_d를 추출한 후, 이를 캡션 생성과 마스크 예측이라는 두 가지 경로(Head)로 동시에 전달하여 학습한다.

3.2 아키텍처 개선: Auxiliary Module 설계

기존 DIRL 모델은 변화의 유무를 추상적인 벡터 공간에서만 학습했으나, 본 연구는 모델이 변화의 구체적인 형상을 인지하도록 보조 헤드를 추가하였다. 인코딩 과정에서 줄어든 해상도를 복원하고 정교한 마스크를 생성하기 위해 다음과 같은 기법을 적용했다.

1. U-Net 기반 디코더 (U-Net based Decoder): 세밀한 복원을 위해 기존 Segmentation Task 에서 성능이 입증된 U-Net[2]의 디코더 구조를 차용하였다. 추출된 변화 특징 맵 F_d는 Conv2D → Batch Norm → ReLU → Upsample 로 구성된 블록을 반복적으로 통과하며 원본 이미지 해상도 $H \times W$로 점진적으로 복원된다.
2. Skip Connection 적용: 깊은 신경망 학습 시 발생하는 정보 손실과 기울기 소실 문제를 방지하기 위해 Skip Connection[3] 을 적용하였다. 인코더 단계의 저수준(Low-level) 공간 정보를 디코더로 직접 전달함으로써, 도로의 얇은 선이나 건물의 모서리와 같은 경계 정보를 보존하며 최종 예측 마스크 $\hat{y}$를 생성한다.

3.3 손실 함수 (Loss Function)

모델은 캡션 정확도와 세그멘테이션 정확도를 동시에 최적화하기 위해 멀티태스크 손실 함수를 사용한다.

$$L_{total} = L_{cap} + \lambda_d L_{DIRL} + \lambda_c L_{CCR} + \lambda_x L_{Aux}$$

- L_{cap}: 캡션 생성을 위한 Cross Entropy Loss.
- L_{DIRL}, L_{CCR}: 기존 DIRL 의 Loss.
- L_{Aux} (Auxiliary Loss): 예측된 마스크 $\hat{y}$와 실제 정답 마스크 M 간의 픽셀별 Cross Entropy Loss.
    - 본 실험에서는 λ_x = 0.05 로 설정하였다. 이는 보조 과제가 주 과제인 캡션 생성을 방해하지 않으면서, 변화 영역에 대한 주의(Attention)를 유도하는 가이드 역할을 수행하도록 조절한 값이다.

4. 실험 및 결과 (Experiments)

4.1 데이터셋 및 학습 환경

- 데이터셋: LEVIR-MCI 데이터셋을 사용하였다. 총 10,077 쌍의 이미지와 50,385 개의 문장으로 구성되어 있으며, 무엇보다 변화 영역에 대한 픽셀 단위의 이진 마스크(Ground Truth)를 포함하고 있어 본 연구의 멀티태스크 학습에 적합하다.

- 학습 파라미터: Optimizer 는 Adam, Learning Rate 는 0.0002, Batch Size 는 128 로 설정하였으며, 총 241 Epoch (Max Iteration 13,000) 동안 학습을 진행했다.

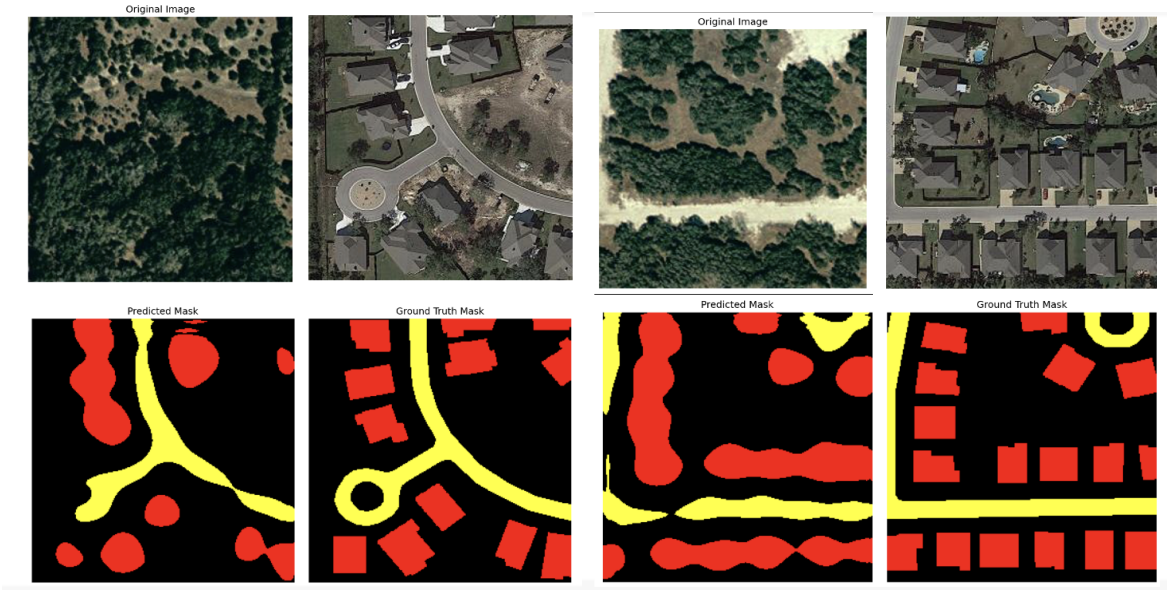
4.2 정량적 평가 (Quantitative Results)

Model	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Original DIRL	57.17	39.13	73.87	133.34	32.18
DIRL + Aux (Ours)	58.39	39.19	73.93	133.39	31.41

제안모델(DIRL + Aux)의 성능을 베이스라인(Original DIRL)과 비교하였다.

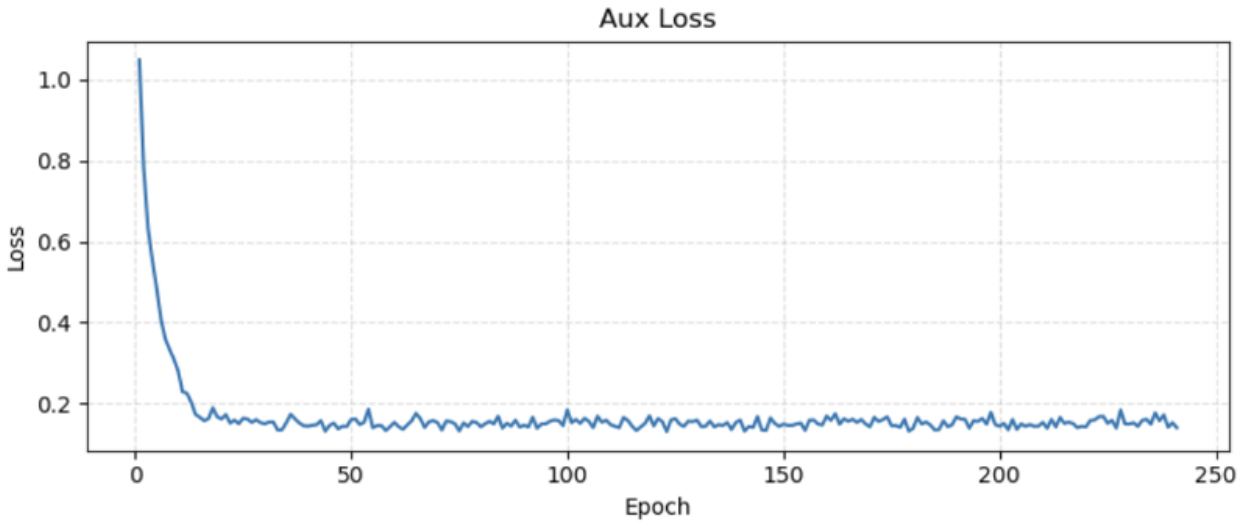
실험 결과, 제안 모델은 베이스라인 대비 BLEU-4(+1.12), METEOR(+0.06), CIDEr(+0.05) 등 문장 생성의 정확도와 유창성을 나타내는 주요 지표에서 성능 향상을 기록했다. 이는 세그멘테이션을 통한 공간 정보 학습이 캡션 품질 개선에 유효함을 입증한다.

4.3 정성적 분석 (Qualitative Analysis)



[그림 2] 위성 사진의 Before-After Image 에 대한 Predicted Change Mask(좌)와 GT Change Mask(우). 노란색은 길에 대한 변화를 의미하며, 빨간색은 건물에 대한 변화를 의미한다.

마스크 예측 시각화: 학습된 모델이 생성한 예측 마스크를 분석한 결과, 복잡한 도심지나 숲 배경 속에서도 건물 신축, 도로 확장 등의 변화 영역을 정확하게 분할(Segmentation)해 내는 것을 확인했다.



학습 안정성: 학습 곡선(Loss Curve) 분석 결과, Aux Loss 가 학습 초기(Epoch 20 이내)에 급격히 감소하며 0.2 이하로 수렴하였다. 이는 모델이 시각적인 변화 위치를 빠르게 파악하고, 이를 바탕으로 캡션 학습을 가속화했음을 시사한다.

5. 고찰 및 한계 (Discussion)

본 연구의 가장 큰 성과는 캡션 생성 모델에 '시각적 근거(Visual Grounding)'를 강제했다는 점이다. 기존 모델은 이미지 특징과 단어 간의 통계적 연관성만을 학습했다면, 보조 모듈이 추가된 모델은 "이 부분이 변했으므로(Location), 이렇게 설명한다(Caption)"는 인과 관계를 강화하여 학습하게 된다. Skip Connection 을 활용한 U-Net 구조는 특징 맵의 공간 해상도를 보존하여 미세한 변화까지 놓치지 않도록 기여했으며, 이는 결과적으로 캡션의 정확도를 높이는 핵심 요인이 되었다.

그러나 절대적인 정량적 성능 지표 측면에서 Chg2Cap 등 현재 해당 분야의 최고 성능(SOTA) 모델들을 상회하지 못했다는 점은 본 연구의 한계로 남는다. 그럼에도 불구하고, 본 연구는 동일한 베이스라인(DIRL) 상에서 추가적인 Segmentation Mask 정보를 제공하는 것만으로도 캡션 생성 성능이 향상됨을 확인하였다. 이는 Segmentation Mask 에 대한픽셀 단위의 지도 학습이 캡션 생성이라는 고차원적 과제에 긍정적으로 기여할 수 있음을 실증적으로 입증했다는 데에 의의가 있다.

6. 결론 (Conclusion)

본 프로젝트에서는 RSICC 작업의 성능 향상을 위해 방해 요소에 강건한 DIRL 모델에 세그멘테이션 보조 모듈을 결합한 개선된 아키텍처를 제안하였다. U-Net 디코더와 Skip Connection 을 활용하여 변화 영역의 위치 정보를 효과적으로 학습시켰으며, 이를 통해 캡션 생성 성능을 베이스라인 대비 유의미하게 향상시켰다. 본 연구는 캡션 생성과 변화 탐지라는 두 가지 과제가 상호 보완적으로 작용할 수 있음을 입증하였으며, 향후 이러한 접근방법을 통해 위성사진의 Change-Captioning Task 에서 더욱 정교한 캡션 생성 모델을 만드는데에 기여할 것으로 기대된다.

---

7. 참고 문헌 (References)

[1] Karaca, Ali Can, et al. "Robust change captioning in remote sensing: Second-cc dataset and mmodalcc framework." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2025).

[2] Tu, Y., et al. "Distractors-Immune Representation Learning with Cross-modal Contrastive Regularization for Change Captioning." European Conference on Computer Vision (ECCV) (2024).

[3] Ronneberger, O., Fischer, P., & Brox, T. "U-Net: Convolutional Networks for Biomedical Image Segmentation." Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2015).

[4] He, K., Zhang, X., Ren, S., & Sun, J. "Deep Residual Learning for Image Recognition." Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (2016).