CLIP 기반 의미론적 지도를 활용한 SAR-RGB 이미지 변환 기법 연구

전지훈, 길태영, 김경태, 김선옥*, 정재훈*
한국항공대학교

wjswlgns1711@kau.kr, gilman2315@kau.kr, kt143056@naver.com, *sunok.kim@kau.ac.kr, *jhjung@kau.ac.kr

A Study on SAR-to-RGB Image Translation Using CLIP-Based Semantic Guidance

Jihun Jeon, Taeyoung Kil, Kyoungtae Kim, Sunok Kim*, Jay Hoon Jung*
Korea Aerospace Univ.

요 약

SAR(Synthetic Aperture Radar, 합성 개구 레이더) 이미지는 구름, 빛, 날씨와 무관하게 안정적으로 촬영이 가능하지만, SAR 이미지의 특성으로 인해 인간이 시각적으로 이해하기 어렵다. 이에 따라 SAR 이미지를 시각적으로 이해하기 쉬운 RGB 이미지로 변환하는 기술은 원격 탐사 및 지리 정보 시스템(GIS) 응용에서 점점 더 중요해지고 있다. 본 논문에서는 기존 Pix2Pix 기반 SAR-to-RGB 변환 구조에 CLIP(Contrastive Language-Image Pretraining) 기반 의미 지도(semantic guidance)를 통합하여, 변환 결과의 의미 정보와 시각적 품질을 동시에 향상시키는 방법을 제안한다. 제안된 모델은 생성된 RGB 이미지와 실제 RGB 이미지 간의 의미 임베딩 유사도를 최대화하는 방향으로 학습하며, 이를 통해 모델이 단순한 픽셀 정합성 이상을 학습한다. 실험 결과, 제안한 기법은 기존 Pix2Pix 대비 SSIM, PSNR, FID 모두에서 향상되었으며, 의미 기반 손실이 SAR-to-RGB 변환에 효과적으로 작용함을 보여준다.

Ⅰ. 서 론

최근 기후 변화, 재해 대응, 국방 감시 등의 목적으로 위성 기반 지구 관측 수요가 증가하고 있으며, 이 중 SAR 이미지는 야간, 악천후 등에서도 안정적인 촬영이 가능하다는 점에서 주목받고 있다. 그러나 SAR 이미지는 레이더 반사 특성을 기반으로 하여, 시각적으로 인간이 인식하기 어려운 텍스처와 노이즈 패턴을 포함하고 있다.[1]

이를 해결하기 위해 SAR 이미지를 RGB 이미지 형태로 변환하는 기술이 제안되고 있으며, Pix2Pix[2], Cycle GAN과 같은 GAN 기반 이미지 변환 모델이 사용되고 있다. 그러나 이러한 모델들은 보통 픽셀 수준의 정합성에 초점을 맞추고 있어, 변환된 이미지가 실제 SAR 이미지의 의미 정보를 제대로 반영하지 못하는 한계가 있다.[3]

본 논문에서는 의미론적 정보 표현에 강점을 가진 CLIP[4]을 활용하여, 생성된 RGB 이미지가 SAR 이미지의 실제 내용과 의미적으로도 일치하도록 학습하는 모델을 제안한다.

Ⅱ. 본 론

2.1 데이터 셋과 전처리

본 연구는 SEN12MS 데이터셋을 기반으로 수행하였다. Sentinel-1의 2채널 SAR 이미지와 Sentinel-2의 13채널 광학 이미지 44,000쌍을 사용하였다. SAR 이미지의 경우, 시각화에 유의미한 실수 부에 해당하는 채널만 사용하였으며, 광학 이미지에서는 RGB에 해당하는 3개의 채널만 사용하였고, 각 이미지는 256×256으로 정규화하였다. 전체 쌍 중 70%는 학습, 20%는 검증에 사용하였다. 나머지 10%를 테스트에 사용하여, 일반화 성능을 평가하였다.

2.2 모델 정의

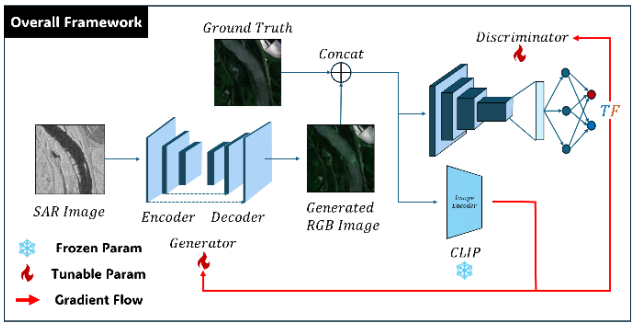

그림 1. CLIP 기반 SAR2RGB 전체 아키텍처

기본 모델은 Pix2Pix 구조를 따른다. 생성기(Generator)는 U-Net 기반의 인코더, 디코더 구조로, SAR 이미지를 입력 받아 대응되는 RGB 이미지를 생성한다. 인코더는 특징 표현을 압축하여 장면의 전역적 구조와 추상적 의미를 학습한다. 디코더는 인코더의 각 단계에서 추출된 특징 맵을 skip connection으로 결합하여 세부적인 구조 정보와 지역적 질감을 보존한다.

판별기(Discriminator)는 Pix2Pix의 PatchGAN 구조를 따르며, PatchGAN은 전체 이미지를 하나의 스칼라 값으로 판별하는 대신, 입력 쌍(SAR, RGB)을 작은 패치 단위로 나누어 각 패치가 실제(real)인지 생성(fake)인지 판별하는 방식이다. 이러한 접근은 이미지의 세부 영역까지 국소적인 일관성을 학습하게 하여, 전체적인 사실감과 함께 미세한 텍스처와 경계 정보까지 보존하는 효과를 낸다.

제안 기법에서는 해당 Pix2Pix 구조 위에 CLIP 기반 의미 지도를 통합한다. CLIP은 대규모 자연 이미지-텍스트 쌍으로 학습된 모델로, 장면의 의미, 물체 유형 등을 표현하는 고차원 임베딩 공간을 제공한다. 이 모델을 기

반으로 생성 결과와 참조 이미지 간 의미 임베딩 유사도를 비교함으로써, 단순한 픽셀 대응을 넘어선 의미적 유사도까지 확보하고자 한다. 생성된 RGB 이미지와 실제 RGB 이미지를 CLIP 이미지 인코더에 각각 통과시켜 임베딩을 추출한다.

최종 손실 함수 구성은 다음과 같다:

$$\mathcal{L}_{GAN} = E_{x,y}[logD(x,y)] + E_x\left[log\left(1 - D(x, G(x))\right)\right] \quad (1)$$

$$\mathcal{L}_1 = E_{x,y}[\| G(x) - y \|_1] \quad (2)$$

$$\mathcal{L}_{CLIP} = 1 - cos\left(f_{CLIP}(G(x)), f_{CLIP}(y)\right) \quad (3)$$

$$\mathcal{L}_{total} = \mathcal{L}_{GAN} + \lambda_1\mathcal{L}_1 + \lambda_{CLIP}\mathcal{L}_{CLIP} \quad (4)$$

본 연구의 최종 학습 손실 함수는 $\mathcal{L}_{GAN}$, $\mathcal{L}_1$, $\mathcal{L}_{CLIP}$의 3가지 항으로 구성된다. $\mathcal{L}_{GAN}$ 는 생성된 이미지가 실제 RGB 이미지와 구분되지 않도록 유도하는 minimax loss로서, 시각적 사실성을 확보하고, $\mathcal{L}_1$는 생성 이미지와 실제 RGB 이미지 사이의 $\mathcal{L}_1 Norm$으로서, 색상과 구조적 정합성을 유지한다.

여기에 본 연구의 핵심인 $\mathcal{L}_{CLIP}$를 추가하여, 생성/실제 이미지에 대해 CLIP 이미지 인코더로 추출한 임베딩 벡터 간 코사인 유사도를 최대화함으로써 생성된 RGB 이미지가 실제 RGB 이미지와 의미적으로 일치하도록 학습한다. 최종적으로, $\mathcal{L}_{total}$은 위 3항의 선형 결합으로 정의되며, 각 항의 기여도는 λ로 조절된다. 여기서 λ는 손실의 가중치를 조절하는 하이퍼파라미터이며, 과도하게 설정될 경우 픽셀 정합성이 저하될 수 있어 실험을 통해 적절한 값을 도출하였다.

2.3 결과 및 비교

표 1. 모델 별 정량적 지표

Method	FID ↓	SSIM ↑	PSNR (dB) ↑
Pix2Pix (Baseline)	131.7202	0.3102	14.95
CLIP-Guided Pix2Pix (Ours)	**113.7680**	**0.3684**	**16.71**

CLIP 기반 모델은 FID[5], SSIM[6], PSNR[7] 지표 모두에서 기존 Pix2Pix에 대비하여 향상되었다. 지표 중 SSIM은 구조적 유사도를, FID는 이미지 분포의 현실성을 반영하는 지표로, 의미 손실이 시각적 일관성을 높이는 데 기여하였음을 시사한다.

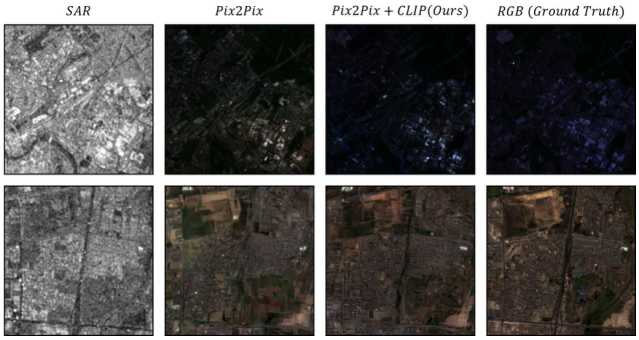

그림 2. 모델 별 정성적 지표

그림 2를 통해 CLIP Loss의 도입은 변환 모델의 시각적 유사도와 의미적 유사도를 동시에 향상시킴을 보여준다. 위와 같이 단순 Pix2Pix 기반 모델은 색감 왜곡과 세부 텍스처의 손실이 발생하는 반면, 제안한 CLIP-Guided Pix2Pix 모델은 실제 RGB 이미지와 보다 유사한 색조를 보이며, 구조적 패턴 재현율이 높음을 확인할 수 있다. 이를 통해 제안된 모델은 단순한 픽셀 별 정합도를 넘어, 시각적 품질과 의미적 유사도를 동시에 향상시키는 효과를 보였다.

CLIP이 위성/원격탐사로 사전 학습되지 않았지만, 도메인이 다르더라도 통용되는 상위 수준의 의미 정보를 제공한다. 이를 통해, 생성 이미지에 제약을 제공해 $\mathcal{L}_1$이 놓치는 형태, 구조, 장면의 일관성을 보완하였다. 그 결과, 제안한 모델은 시각적 품질과 의미적 유사도를 동시에 개선하였음을 알 수 있다.

Ⅲ. 결 론

본 논문은 CLIP 기반 의미론적 지도를 활용하여 SAR-to-RGB 변환 성능을 향상시키는 기법을 제안하였다. 생성기가 구조 및 의미 유사도를 함께 고려하도록 유도함으로써, 정량적 지표(SSIM, PSNR, FID) 뿐 아니라 정성적 지표에서도 개선을 확인하였다. 실험 결과는 제안 기법이 기본적인 GAN 기반 구조를 의미적으로 강화하는 데 효과적임을 보여주며, 향후 다양한 도메인에서의 응용 가능성을 시사한다.

ACKNOWLEDGMENT

본 과제(결과물)은 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 첨단분야 혁신융합대학사업(차세대통신)의 연구 결과입니다.

참 고 문 헌

[1] D. Ao, C. O. Dumitru, G. Schwarz, and M. Datcu, “Dialectical GAN for SAR Image Translation: From Sentinel-1 to TerraSAR-X,” Remote Sensing, vol. 10, no. 10, 1597, Oct. 2018.

[2] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A., "Image-to-Image Translation with Conditional Adversarial Networks," Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1125-1134, Jul. 2017.

[3] 박소연, 곽근호, 황의호, 박노욱, “고해상도 광학영상 복원을 위한 딥러닝 기반 SAR-광학 영상 변환 모델 비교,” Korean Journal of Remote Sensing, vol. 40, no. 6-1, pp. 881-893, Dec. 2024.

[4] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I., "Learning Transferable Visual Models From Natural Language Supervision," Proc. 38th Int. Conf. on Machine Learning (ICML), pp. 8748-8763, Jul. 2021.

[5] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S., "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 6626-6637, Dec. 2017.

[6] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: From error visibility to structural similarity," IEEE Trans. on Image Processing, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[7] Bovik, A. C., "Handbook of Image and Video Processing," Academic Press, pp. 859-862, 2000.