2024년 2월

석사학위 논문

XAI 기반 생성 모델 데이터 분석 및 평가

조선대학교 대학원

항공우주공학과

이 하 늘

XAI 기반 생성 모델 데이터 분석 및 평가

Data Analysis and Evaluation of Generative Models Based on XAI

2024년 2월 23일

조선대학교 대학원

항공우주공학과

이 하 늘

XAI 기반 생성 모델 데이터 분석 및 평가

지도교수 이 현 재

이 논문을 공학석사학위 신청 논문으로 제출함

2023년 10월

조선대학교 대학원

항공우주공학과

이 하 늘

이하늘의 석사학위논문을 인준함

위원장 김태규 (인)

위 원 이현재 (인)

위 원 정성훈 (인)

2023년 12월

조선대학교 대학원

목차

ABSTRACT

Data Analysis and Evaluation of Generative Models Based on XAI

Sky Haneul Lee
Advisor : Prof. Henzeh Leeghim, Ph.D.
Department of Aerospace Engineering,
Graduate School of Chosun University

When performing guided pilot simulations or flight test simulations required for various purposes, the quality of dynamic and synthetic sensor images generated from the sensor models in the given simulation environment is highly important for target recognition, tracking, and behavior for various reconnaissance missions. So this thesis uses artificial intelligence to generate realistic infra-red (IR) images used in flight simulations, and analyzes and evaluates the data through explainable artificial intelligence (XAI).

We construct an IR dataset and a synthetic dataset to generate realistic IR images. Among the Generative Adversarial Network (GAN) models, CycleGAN, which is trained under unpaired dataset, is trained using the constructed dataset. CycleGAN produces high-quality images even though it does not have a correct answer label. However, it is not well-trained on the IR dataset. Therefore, we improved the model's performance based on the structural similarity index measure (SSIM). At this time, we compared the weights of each loss function to find an appropriate value, and analyzed how window sizes of SSIM would affect the synthetic IR image constructed by CycleGAN is analyzed.

Although techniques such as IS and FID have been introduced to evaluate the performance of GAN, it is still difficult to distinguish between synthetic data. Additionally, distinguishing synthetic data generated by artificial intelligence is a big topic because the level of data generation using GAN is improving. Therefore, we introduce XAI for synthetic IR image analysis. Out of various XAI techniques, LRP was used, which detects the model in reverse order through decomposition and relevance propagation, providing a basis for judgment on prediction. Thus, we build a classification network

for IR images and synthetic IR images and then perform LRP analysis. When analyzing LRP, we simulate various transformations of LRP and analyze how LRP draws a heatmap when some transformation is applied to the data, allowing us to distinguish between IR images and synthetic IR images.

I. 서론

현대전에서 핵심 전력으로 꼽히는 정밀타격 유도무기에는 적외선 영상 탐색기가 주로 사용되는데, 이의 성능 검증을 위한 유도 조종 기법 시뮬레이션이나 모의 비행시험 수행 시 탐색기 인식 및 추적, 다양한 정찰 임무를 위한 알고리즘을 평가하기 위해 가상 환경에서의 동적인 합성(sythetic) 적외선 영상이 필요하다. 이러한 가상 영상을 기반으로 하는 유도 조종 시스템은 HILS (hardware-in-the-loop simulation) 기반의 시스템보다 적은 비용으로 시스템 성능 검증이 가능하기 때문에 이에 대한 연구들이 지속되고 있다[1, 2, 3]. 하지만 이러한 가상 영상은 실제 영상과는 차이가 존재하고, 특히 다양한 시간대와 원하는 환경 등에서 적외선 영상 환경을 구축하는 것은 쉽지 않기 때문에 최근 인공지능(Artificial Intelligence, AI) 기법을 이용한 연구들이 활발히 진행되고 있다. 이에 따라 생성 모델 분야에서는 2014년 제시된 GAN (Generative Adversarial Network)[4] 기술을 바탕으로 다양한 영상 획득의 길이 열리게 되었다. 이는 학습 데이터를 바탕으로 생성자와 판별자가 적대적으로 학습하며 실제 데이터와 유사한 데이터를 생성한다. 또한 데이터의 분포를 흉내 내기 때문에 훈련 데이터가 비교적 충분하지 않은 상황에서도 이미지를 생성해 낼 수 있다. GAN은 초반에는 성능이 좋지 않았지만 다양한 구조, 손실 함수의 변형 및 개선 등으로 보다 사실적인 이미지 생성이 가능해졌다.

이미지 생성 분야에서 이미지 생성 외에도 주요하게 꼽히는 점 중 하나는 GAN으로 생성된 이미지를 평가하는 것이다. CNN (Convolution Nerual Network) 등과 다르게 GAN은 결과물에 대한 실제 정답이 없기 때문에 각 GAN들의 아키텍처나 구성을 비교하기가 어렵다. GAN의 연구 초기 단계에서는 평가 지표로 수동적인 방법을 사용하였는데 이는 사람이 직접 각 단계마다 생성된 이미지를 보고 평가하는 것으로, 평가자의 주관적인 견해가 개입되기 때문에 개개인에 따라 지표가 나눠질 수 있다. 따라서 이를 해결하기 위해 IS (Inception Score)[5], FID (Frechet Inception Distance)[6] 등의 기법이 도입되었다. 이는 생성된 데이터의 품질(Fidelity)과 다양성(Diversity)을 기준으로 데이 터셋을 평가하며, 실제 이미지의 확률 분포와 합성 이미지의 확률 분포 차이의 거리를 측정하여 점수를 구하는 방식으로 작동한다. 해당 기법들은 GAN의 성능

평가를 위해 다양하게 적용되고 있지만 데이터셋 전체에 적용되기 때문에 본 연구에서 사용하는 적외선 이미지에 대한 구체적인 평가 기법으로는 적절하지 않다. 따라서 본 논문에서는 이미지의 주파수 스펙트럼을 분석하는 PSD (Power Spectral Density)[7, 8] 기법을 활용해 생성된 이미지 평가에 이용하였다.

정량적인 이미지 평가 지표를 이용해 실제 적외선 이미지와 합성 적외선 이미지의 차이를 식별할 수 있지만 최종적으로는 합성 적외선 이미지를 실제 적외선 이미지와 유사하게 개선하는 작업이 필요하다. 따라서 어떤 부분이 실제 적외선 이미지와의 차이를 야기하는지에 대한 정의가 필요한데, 앞선 평가 지표들은 데이터셋 간 분포만을 계산하거나 이미지의 전체적인 스펙트럼을 나타내기 때문에 이미지의 개선점을 파악하기에는 적절하지 않다. 따라서 이미지 평가와 더불어 보다 구체적인 개선점을 찾기 위해서 설명 가능한 인공지능(eXplainable Artificial Intelligence, XAI) 기법을 도입하였다. GAN을 포함한 대부분의 인공지능은 블랙박스(Blackbox) 구성을 가지고 있기 때문에 인간은 인공지능의 결과물을 보고 이해할 수는 있지만, 인공지능의 의사결정 과정에 대한 이해는 불가능하다. 특히 고품질의 이미지를 생성해내는 GAN의 경우에는 수많은 파라미터들이 내부에서 동작하기 때문에 내부 작동 방식이 불분명하다. 따라서 이와 같은 문제에 해결책을 제공해 주는 XAI 기법이 주목받고 있으며 이를 잘 이용한다면 인공지능 모델의 의사결정 과정에 대해 이해도를 높일 수 있다. 본 연구에서는 XAI의 시각화 기법 중 하나인 LRP (Layer-wise Relevance Propagation)[9]를 이용해 합성 이미지 분석에 이용하였다. 이는 기존 신경망 특징점 시각화 기법들보다 블랙박스를 오인할 가능성이 적으며, 네트워크가 분류한 이미지 결과를 역순으로 탐지하여 분해하기 때문에 이미지 분류에 신경망이 집중한 부분을 보다 명확하게 시각화하여 관찰할 수 있다. 이에 따라 설계된 인공지능 분류 모델에 LRP를 적용하여 실제 이미지와 합성 이미지를 분류하는 데에 네트워크가 이미지의 어떤 부분을 집중하였는지 시각화하였다. 이때 LRP 기법의 다양한 변형을 통해 개선된 히트맵을 제안했으며 그에 따른 분석을 수행하였다.

본 논문은 다음과 같이 구성된다. 먼저 2장에서는 XAI의 개요와 전반적인 개념, 현재 이미지 분석에 사용되는 대표적인 기법들에 대해 소개한다. 3장에서는 알고리즘에 사용된 데이터셋 구축 방법에 대해 소개하고, 4장에서는 LRP를 적용한 네트워크 구성과 알고리즘에 대해 소개한다. 5장에서는 앞서 생성한 데이터와 구축한 네트워크를 통해 시뮬레이션을 수행한다. 마지막으로

6장은 본 논문에 대한 결론으로 마무리한다.

II. 설명 가능한 인공지능

1. 설명 가능한 인공지능

인공지능은 인간의 뉴런을 모방하지만 실제 인간의 의사 결정 방식과는 다르게 주어진 입력 값을 암기하여 의미 있는 정보를 추론하는 정도에 그친다. 물체 인식을 예로 들면, 인간은 물체를 인식할 때 전체적인 영역과 부분적인 영역을 동시에 고려하지만 인공지능은 픽셀 값이 어떠한지만을 고려해 물체를 인식한다. 이는 인공지능이 현실 세계를 이해하고 인식하여 의미 있는 결과를 도출하는 것이 아닌 그저 암기한 것에 불과하며, 인간과 인공지능의 의사 결정 과정에는 큰 차이가 존재한다는 근거가 된다. 하지만 이러한 인공지능의 의사 결정 과정을 인간이 이해할 수 있다면 모델을 어떤 방향으로 개선해야 할지, 어떠한 데이터를 정제하여 입력으로 넣어야 할지 결정해 결과물의 품질을 향상시킬 수 있으며 인공지능이 내린 결정에 대해 신뢰성을 확보할 수 있다. 분류 네트워크를 예로 들면 현재의 네트워크는 거듭되는 발전을 거쳐 ImageNet의 경우 90%이상의 정확도를 달성하였다. 하지만 만약 네트워크가 어떤 특정한 사진에 강아지가 존재하는지 아닌지를 판단할 때 99의 확률로 해당 사진이 강아지라고 판단하였다면, 1의

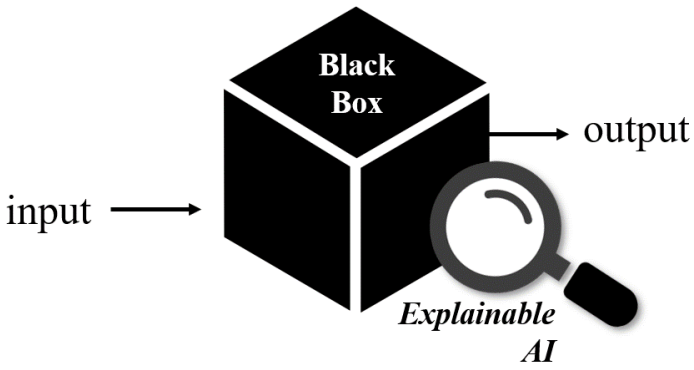


Figure 1 : Black box model of XAI

강아지가 아닐 확률 때문에 이 결과를 완전히 신뢰할 수는 없다. 99의 확률로 객체를 분류하는 모델조차 1의 확률 때문에 모델을 전적으로 신뢰할 수 없는데, 본 연구의 목표인 실제 이미지와 합성 이미지를 분류하는 문제는 보다 더 복잡하고 모델의 결과를 신뢰하기 어려우며, 부합하는 설명이 없다면 인공지능의 판단 기준을 파악하기 어렵다. 따라서 인공지능 모델이 "왜 이것을 합성 이미지로 분류하였는가?"에 대한 설명, 즉 판단 근거를 제공해 줄 수 있다면 모델의 예측을 신뢰할 것인지에 대한 여부를 결정할 수 있을 것이다. 또한 이렇게 XAI가 제공해 주는 설명이 실제 이미지와 합성 이미지의 차이점을 유의미하게 나타낼 수 있다면 분석을 통해 합성 이미지의 개선점을 식별해낼 수 있을 것이다.

본 연구에서 주안점을 두고 있는 것은 시각화(Visualization) 기법으로, 눈에 보이는 직관적인 결과가 인간을 납득시키는 데에 효과적이며 인공지능 기술이 이미지 분류, 생성 등과 같은 영상처리 분야에서 두각을 나타내고 있기 때문이기도 하다. 실제 Occlusion Experiment 연구는 CNN이 사람이 물체를 인식하는 과정과 유사하다는 것을 검증한 바 있는데, 이렇게 설명 가능성이 낮은 딥러닝 모델 중에서도 CNN은 인간의 시신경 구조를 모방해 시각 인지 시스템과 유사하게 동작하기 때문에 CNN 모델의 설명 가능성에 대한 연구들이 많이 진행되었다. 아래 소단원에서는 CNN 네트워크에 적용 가능하며 시각화된 정보를 제공해 주는 모델 시각화 XAI 기법들에 대해 살펴본다.

2. 모델 시각화 설명 기법

모델 시각화(model visualization) 기법은 특성 맵 시각화라고도 하며, 블랙박스 모델을 어떻게 해석하느냐에 따라서 다양한 방법론이 존재한다. 대표적인 특성 맵 시각화 중 하나인 CAM (Class Activation Map)[10]은 CNN 내의 특성 맵(Feature map)을 보고 어떤 픽셀의 활성 함수가 가장 활성화되었는지를 역으로 추정하고 이를 통해 예측 시 가장 중요한 특성 값을 가지고 있던 부분을 히트맵으로 도시하여 시각화한다. 하지만 이는 마지막 컨볼루션 층을 통과해 나온 특성 맵에 대해서만 적용할 수 있기 때문에, 전체적인 신경망의 히트맵은 확인할 수 없다는 단점이 있다. 이러한 한계를 극복하기 위해 가중치를 그래디언트로 대신하여 일반화한 Grad-CAM과 Grad-CAM++와 같은 추가 연구들이 수행되었다[11, 12].

LIME (Local Interpretable Model-agnostic Explanations)[13]은 대표적인 XAI 기법으로, 대리 분석(surrogate analysis) 기법을 이용한다. 본래 분석해야 하는 모델이 f일 때 이를 흉내 내는 모델 g를 만드는 것이 대리 분석의 목표이며 LIME은 이러한 대리 분석을 국소적이면서 구체적으로 구현한 방법으로, 대리 모델이 블랙박스 모델의 예측을 근사하도록 학습된다. LIME은 입력 데이터에 변형을 가했을 때 예측에 어떤 변화가 일어나는지 체크하는데 이때 데이터 변형은 슈퍼 픽셀(super-pixel) 마스킹을 이용하며, 마스킹을 통해 분할된 이미지를 조합해 원본 모델이 대상 분류에 가장 적합하다고 여기는 대표 이미지를 구성한다. Fig.(2)은 슈퍼 픽셀 분할을 통해 구성된 이미지에 대해 모델이 클래스를 예측한 결과이다. 하지만 LIME은 모델의 결정 경계를 확정짓는 방법이 비결정적이며, 데이터 하나에 대해서만 설명하므로 모델에 대한 설명력이 부족하다. 또한 단순히 특성 맵 시각화 방법이기 때문에 입력 값의 기여에 대한 간접적인 해석일 뿐 은닉층 단계에서의 기여에 대해서는 파악할 수 없다는 단점이 있다. 이러한 단점에서 나아가 게임 이론을 바탕으로 하는 섀플리 값(Shapley Value)과 LIME을 결합한 방법론인 SHAP(SHapley Additive exPlanation)[14]이 제안되었다. 섀플리 값은 하나의 특성에 대한 중요도를 알기 위해 여러 특성들의 조합을 구성하고, 해당 특성의 유무에 따른 평균적인 변화를 통해 값을 계산한다. 즉 특정한 변수가 제거되었을 때 예측에 얼마나 영향을 미치는지 살펴보고 그에 대한 답을 섀플리 값으로 표현하는 것이다. SHAP은 LIME보다는 모델에 대한 설명을 효과적으로 제공해 주지만 계산량이 크기 때문에 높은 차원을 가지는 이미지 데이터에는 적용하기 쉽지 않다.

CAM, LIME과 같은 특성 맵 시각화 방법은 모델이 입력 이미지에 어떻게 반응하는지 은

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 2 : Example of LIME algorithm in XAI

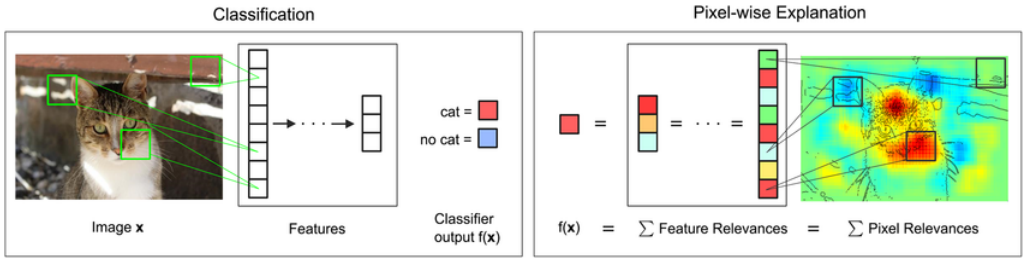


Figure 3 : Example of LRP algorithm in XAI

닉층을 조사하는 방법이다. 하지만 이는 신경망이 깊어질수록 해석력이 떨어지고, 그에 따른 다양한 해석이 존재할 소지가 있다. 또한 전체적인 신경망이 아닌 특정 레이어에 대한 설명을 제공해 주는 것이 대부분이다. 이와 다르게 LRP는 결과를 역추적해서 그에 따른 히트맵을 입력 이미지에 출력한다. 이 히트맵은 블랙박스로 여겨지는 모델이 데이터의 어떤 곳을 주목했는지 시각화해 주기 때문에 앞선 특성 맵 시각화 방법보다 블랙박스를 오인할 가능성이 적다. LRP는 분해(Deconposition)와 타당성 전파(Relevance Propagation) 과정으로 이루어지는데, 순방향(feed-foward)으로 진행하는 보통의 필터 시각화 기법과 다르게 모델을 역순으로 탐지하기 때문에 결과물에 대한 모델의 근거를 보다 합당하게 제공해 줄 수 있다. 분해 과정에서는 입력된 값 하나가 결과 해석에 영향을 얼마나 미치는지에 대한 값을 도출해 내고 이 값을 이용하여 타당성 전파 과정에서는 분해 과정을 마친 은닉층이 예측 결과 출력에 어떤 기여를 하는지 타당성을 계산한다. 타당성 계산으로 모든 은닉층 내 활성화 함수의 기여도를 계산해 이미지 x에서 픽셀별 기여도를 히트맵으로 나타낸다.

본 연구에서는 ImageNet과 같은 일반적인 데이터셋이 아닌 실제 적외선 이미지와 합성 적외선 이미지 대한 설명을 필요로 하므로 네트워크 및 데이터에 대한 보다 구체적인 설명이 필요하다. 따라서 모델의 결정 근거를 보다 합당하게 제공해 주는 LRP가 모델 해석에 적합하다 판단해 해당 기법을 통한 이미지 분석을 수행하였다. LRP 알고리즘에 대한 자세한 내용은 본 논문의 4장에서 다루도록 한다.

III. 데이터셋 구축

1. 학습 데이터셋 구성

해당 장에서는 실제 이미지와 합성 이미지 분류 및 식별을 위한 데이터셋 구축 과정에 대해 소개한다. 기존의 많은 컴퓨터 비전 문제들은 분류 기반 문제로, 어떠한 특성 공간(feature space)이 있을 때 공간 상에서 찾고자 하는 레이블로 구분되는 구분자를 찾는 문제가 대부분을 차지한다. 이에 반해 생성 모델은 정답이 없는 비지도학습(Unsupervised learning)에 속하므로 주어진 학습 데이터의 분포를 학습하고 그 분포를 따르는 유사 데이터를 생성하는 것을 목표로 한다. 이때 유사 데이터란 원본 데이터의 분포는 유사하게 따르되 기존에는 없던 새로운 데이터를 의미하기 때문에, 원하는 결과 데이터를 얻기 위해서는 적절하고 다양한 분포를 가지는 학습 데이터셋을 구축하는 것이 매우 중요하다. 본 연구에서는 실제 적외선 영상과 유사한 실감 적외선 영상을 생성하는 것을 목표로 하기 때문에 생성 모델이 목표로 하는 데이터셋은 실제 적외선 영상, 변환이 수행되는 입력 데이터셋은 합성 영상이 될 것으로 보고 두 가지 클래스에 대한 영상 데이터를 구축하였다.

1) 적외선 영상 데이터셋

다양한 거리 및 시각에서의 적외선 영상을 획득하여 활용하기 위해 비행시험이 필요할 것으로 보고 적외선 카메라가 장착된 헥사로터형 멀티콥터 시스템을 이용하였다. Fig. (4)와 같이 적외선 영상은 12시에 촬영한 각종 구조물 및 자연물이 존재하는 필드와 19시에 촬영한 하천 주변의 두 도메인으로 획득하였으며, 이때 획득한 적외선 이미지로 학습을 진행할 때 모델이 색상 특성을 잘 포착하지 못해 학습이 잘 이루어지지 않는 문제를 방지하기 위해 RGB 이미지를 모두 회색조(Grayscale)로 변환하였다. 사용된 멀티콥터 및 열화상 카메라의 스펙은 Table.(1)과 (2)에 도시하였다. 또한 알고리즘의 편향 방지 및 검증을 위해 FLIR 사에서 무료로 배포하는

FLIR 열화상 데이터셋을 이용하여 추가적인 학습을 진행하였다.

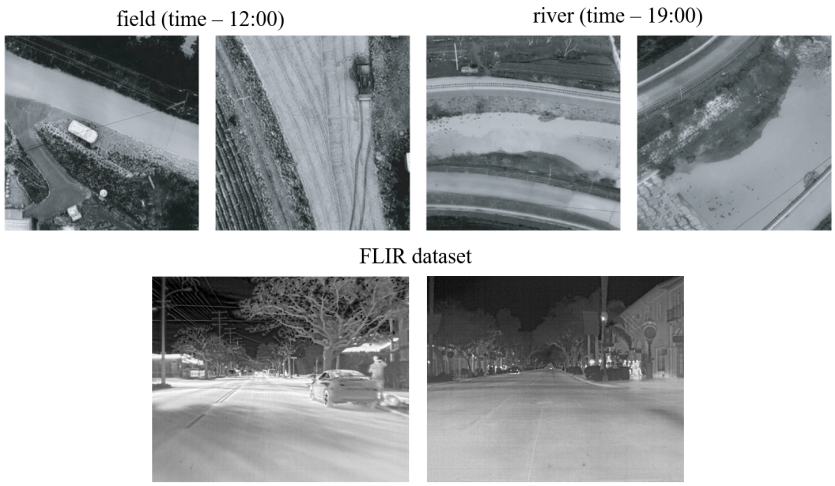


Figure 4 : Real IR images : the left two images were taken in the field at 12:00, the right two images were taken around a small stream at 19:00 and below two images are FLIR images

2) 합성 영상 데이터셋

실제와 유사한 합성 적외선 영상 생성을 위해서는 실제 환경을 정교하게 모사하는 합성 영상이 필요하다. 따라서 본 연구에서는 보다 정교한 실제 환경 모사를 위해 사진 측량(photogrammetry) 및 모델링 소프트웨어인 RealityCapture를 사용하여 멀티콥터를 통해 획득한 영상을 이용한 3D 모델링을 수행하였다. 모델링 시 합성 이미지에 적외선 도메인을 전이시키기 위해 텍스처 매핑(texture mapping)을 거치지 않은 메쉬(mesh) 상태의 모델로부터 이미지를 추출하여 사용하였다. 또한 이미지 데이터가 특성이 흐려져 학습이 잘 이루어지지 않는 것을 방지하기 위해 모든 이미지의 대비(contrast)를 증가시켜 학습에 이용하였다. 학습에 이용하기 위한 합성 데이터 가공 과정은 Fig.(5)에 도시하였다.

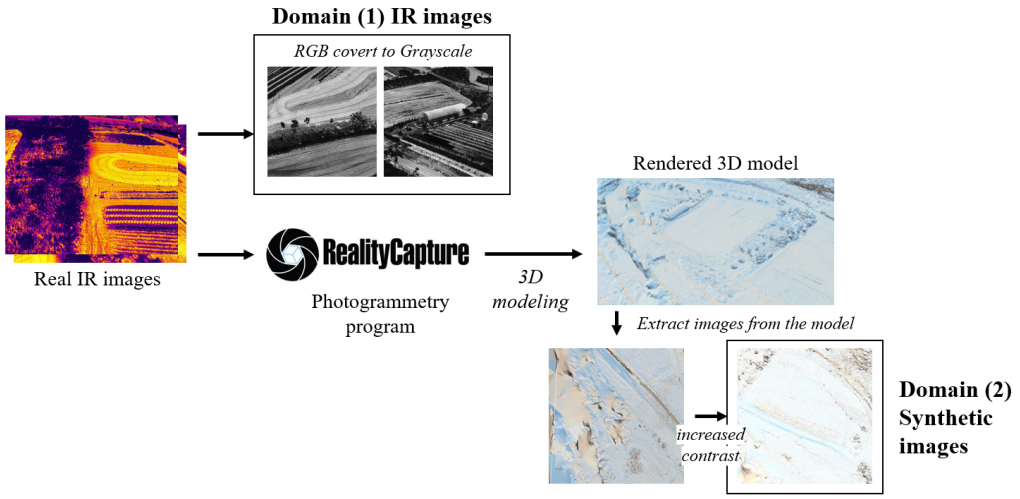


Figure 5 : Flowchart of image domain construction process

Table 1 : UAV specification

Attribute	Value
Description	DJI MATRICE 300 RTK
Weight	6.3 kg (Including two TB60 batteries)
Diagonal length	895 mm

Table 2 : Camera specification

Attribute	Camera Specification
Description	DJI ZENMUSE H20T
Sensor	Vanadium Oxide (VOx) microwave bolometer
Lens	DFOV: 40.6°
	Focal length: 13.5 mm
	Aperture: f/1.0
	Focus: 5 m $\sim \infty$

2. 생성 모델 학습

최근 적외선 합성 영상을 생성함에 있어서 보다 사실적인 구현을 위해 인공지능 기술을 활용하고 있다. 이때 훈련 데이터가 충분하지 않은 상황에서 생성 모델인 GAN 기술을 활용할 수 있는데, GAN이란 생성자(generator)와 판별자(discriminator)가 상호 적대적으로 학습하며 실영상과 유사한 고품질의 영상을 생성해 내는 기법이다. 생성자의 목표는 판별자가 구분하지 못할 정도로 실제 데이터와 유사한 데이터를 생성하는 것이며, 판별자의 목표는 실제 데이터와 가짜 데이터를 잘 구분하는 것이 된다. 잘 구성된 생성 모델에서 학습이 진행된다면 어느 순간부터 생성자는 실제와 거의 유사한 가짜 데이터를 만들 수 있게 되며, 판별자는 결국 참과 거짓을 구분하지 못해 구분 확률이 50%에 수렴하게 된다. 기본적인 GAN의 손실 함수는 다음과 같이 구성된다.

$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[logD(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[log(1-D(G(\mathbf{z})))] \tag{1}$$

또한 GAN은 크게 비조건부 GAN (Unconditional GAN)과 조건부 GAN (conditional GAN, cGAN)으로 나눌 수 있다. 비조건부 GAN은 GAN의 잠재 벡터(latent vector)에 아무런 제약 조건을 걸지 않고 임의의 이미지를 생성하게 된다. 비조건부 GAN에는 학습 과정에서 레이어를 추가하여 점차 해상도를 높여 나가며 높은 해상도의 이미지를 효율적으로 학습시키는 PGGAN[15]과 이미지의 스타일을 부분적으로 제어하는 StyleGAN[16] 등이 연구되었다. StyleGAN의 경우에는 뉴럴 스타일 트랜스퍼(Neural Style Transfer) 기법을 사용하여 입력 이미지의 콘텐츠를 보존하며 참조 이미지의 스타일을 입력 이미지에 적용시키는 기법으로, 하나의 메인 이미지에서 다양한 스타일을 가지는 고해상도 이미지를 생성할 수 있다.

반면에 조건부 GAN이란 생성기와 판별기가 훈련을 하는 동안 레이블이라는 추가적인 정보를 사용해 학습에 방향성을 제시해 주는 기법으로, 이미지 대 이미지 변환 기법들이 cGAN에 속한다. 대표적으로는 Pix2pix[17]와 CycleGAN[18]이 존재하는데, Pix2pix는 일종의 지도학습(Supervised learning)에 속하는 기법으로 레이블 이미지가 존재할 때 그와 유사하게끔 입력 이미지 변환을 수행한다. 하지만 Pix2pix는 입력과 레이블이 쌍을 이루는 쌍체 이미지(paired

image)여야만 한다는 조건이 있는데, 실제 환경에서 쌍체 이미지 데이터셋을 구축하는 것은 쉽지 않기 때문에 비쌍체 이미지(unpaired image)로 학습하는 CycleGAN 관련 연구가 수행되었다.

앞서 언급한 이미지 대 이미지 변환과 스타일 트랜스퍼는 한 이미지를 다른 이미지의 스타일을 가지는 새로운 이미지로 바꾼다는 데에서 비슷해 보이지만 다른 문제로 간주된다. 스타일 트랜스퍼의 경우에는 일반적으로 입력 이미지와 참조 이미지가 존재하며, 생성 모델을 통해 얻어지는 결과는 입력 이미지의 콘텐츠와 참조 이미지의 스타일을 결합한 새로운 이미지이다. 이미지 대 이미지 변환에서는 여러 입력 이미지와 타겟 이미지가 있을 때 입력 도메인에서 타겟 도메인으로의 매핑(mapping) 함수를 찾는 것에 중점을 둔다. 스타일 트랜스퍼는 입력 이미지와 참조 이미지의 가중치를 변경할 수 있고, 입력 이미지에 가장 적합한 스타일을 선택할 수도 있기 때문에 스타일을 변경하고 싶은 구체적이고 적은 양의 데이터가 있는 경우에는 스타일 트랜스퍼를 사용하는 것이 더 효과적이다. 하지만 많은 유사한 사진들 사이에서 어떤 스타일로 변형했는지, 스타일 변환 비중을 얼마나 주었는지가 중요하지 않은 상황에서는 이미지 대 이미지 변환이 더 효과적으로 작동한다. 본 연구에서는 목표로 하는 이미지가 적외선 도메인이기에 이미지의 세부 스타일을 제어하는 것보다는 적외선 도메인을 입히는 것에 더 우선순위를 두고 생성 모델로 이미지 대 이미지 변환에 속하는 CycleGAN을 선정하였다.

1) CycleGAN

Pix2pix가 지도학습에 속하는 반면 CycleGAN은 비지도학습의 일종으로, 레이블에 따른 이미지 매핑이 아닌 타겟 도메인 영역의 이미지 특징을 학습한다. 이때 CycleGAN은 정답이 없는 영역에서 이미지 생성을 수행하므로 모델이 실제 데이터 분포를 전부 다루지 못하고 다양성을 잃는 모드 붕괴(Mode collapse) 문제가 나타날 수 있다. 모드 붕괴가 일어난 CycleGAN 모델은 생성자가 입력 이미지의 특징을 모두 잃어버리고 똑같은 출력을 생성하게 된다. 이를 해결하기 위해 기존 생성자 G에 새로운 생성자 F를 더한 두 개의 생성자와 판별자가 동시에 학습하게 되는데, G 함수는 X 도메인에서 Y 도메인으로 매핑되고 F 함수는 Y 도메인에서 X 도메인으로 매핑되는 생성자이며 D_x와 D_y는 각 생성자에 대한 판별자가 된다. 이러한 순환 구조를 통해 CycleGAN은 입력 이미지를 타겟 도메인으로 매핑할 때 다시 원래의 입력으로 돌아올 수 있게끔

이미지를 변형하여 모드 붕괴를 막게 된다. 이러한 구조를 순환 일관성(Cycle Consistency) 구조라 하며, 입력 이미지 x에서 G와 F를 거쳐 다시 돌아온 $\hat{x}$ 간의 차이를 순환 일관성 손실(Cycle Consistency loss)이라 정의한다. 순환 일관성 손실은 x와 $\hat{x}$ 간 L1 손실을 사용하며, CycleGAN의 적대적 손실과 순환 일관성 손실은 아래와 같다. 또한 CycleGAN의 순환 구조를 Fig.(6)에 도시하였다.

$$L_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[log D_Y(y)] + \mathbb{E}_{x \sim p_{data}(x)}[log(1 - D_Y(G(x)))] \quad (2)$$

$$L_{cyc}(G, F) = \mathbb{E}_{x \sim p_{data}(x)}[\|F(G(X)) - x\|_1] + \mathbb{E}_{y \sim p_{data}(y)}[\|G(F(y)) - y\|_1] \quad (3)$$

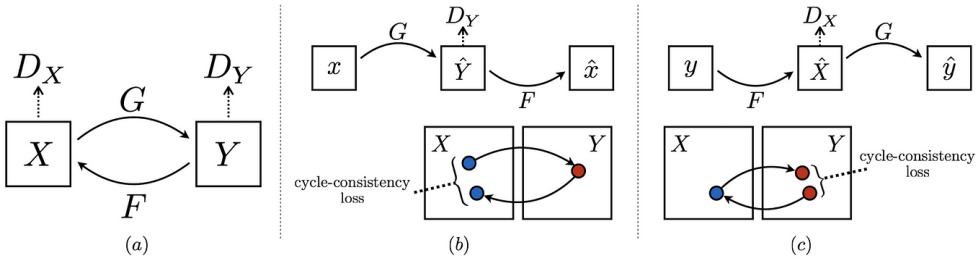


Figure 6 : Consistency structure of CycleGAN

(1) 생성자 및 판별자 구조

CycleGAN의 생성자는 기본적으로 U-Net[19] 구조를 차용한다. Fig.(7)와 같이 신경망이 U자 구조를 가지기에 U-Net이라는 이름이 붙었으며 인코더-디코더(encoder-decoder) 기반 모델에 속하는데, 이는 차원 축소를 수행하는 인코딩(encoding) 단계와 저차원으로 인코딩된 정보를 통해 고차원의 이미지를 복원하는 디코딩(decoding) 단계로 구성된다. 하지만 이때 차원 축소를 수행하며 영상에 대한 고차원적 정보 손실이 일어나고, 디코딩 단계에서도 저차원의 인코딩된 정보를 이용하기 때문에 정보 손실을 막기가 어려워진다. 따라서 U-Net은 저차원뿐만 아니라 고차원의 정보까지 이용하여 디코딩 시의 정보 손실을 최소화한다. 이를 위해서 인코딩 단계에서 얻은 레이어의 특징을 디코딩 단계에 합치는 방법을 사용하는데, 이를 스킵 연결(skip connection) 이라 한다. 인코딩 단계에서 하나의 박스는 3x3 컨볼루션(convolution), 배치 정규화(batch

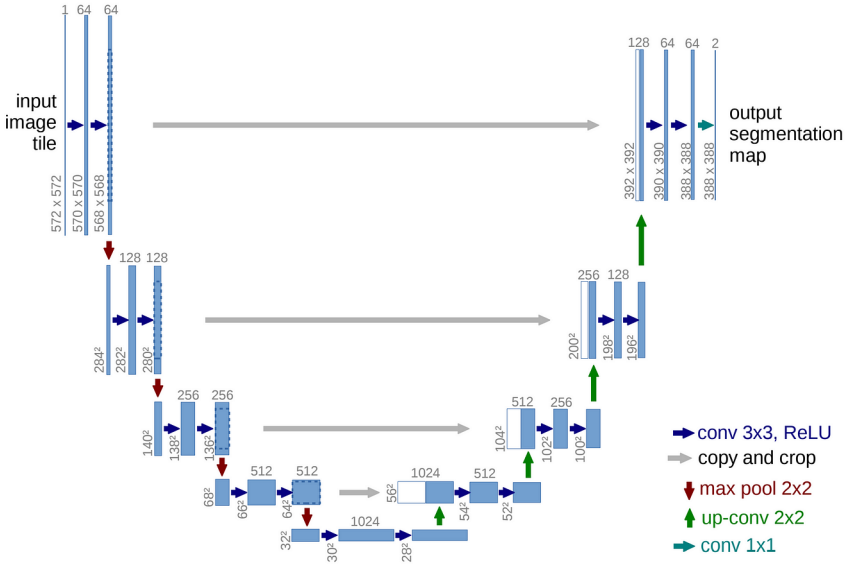


Figure 7 : U-Net structure

normalization), 렐루(ReLU) 활성화 함수로 구성된다. 이러한 구조를 하나의 블록으로 묶어서 사용하고, 2x2 max pooling으로 다운샘플링(downsampling)된 또 하나의 출력을 이용한다. 디코딩 단계에서는 스킵 연결을 통해 인코더로부터 복사된 특성과 전치 컨볼루션(transposed convolution)을 결합하여 저차원 정보뿐만 아니라 고차원 정보도 이용할 수 있게 한다. 이때 U-Net만을 이용하면 인코더에서 디코더로 값이 나올 때 디테일한 부분을 잃는 문제가 있어 고해상도 이미지 처리를 보다 잘 수행하기 위해 잔차 연결(Residaul connection)[20]을 추가하였다. 잔차 연결은 네트워크가 깊게 설계되면 발생하는 기울기 소실/증폭(Gradient vanishing/explosion) 문제를 해결하기 위해 도입된 개념이다. 잔차 연결의 개념은 Fig.(8)와 같으며, 스킵 연결의 개념과 동일하게 기존의 단일 연결 신경망에서 나아가 입력 값을 출력 값에 더해 주는 구조를 가진다. 이렇게 입력으로 들어간 x를 더해 주는 것만으로 정보 손실이 적어질뿐더러 각 층들은 x를 제외한 나머지 부분인 $F(x)$만을 학습하면 되므로 학습량이 상대적으로 줄어드는 효과 또한 발생한다.

기존 GAN의 판별자는 생성자가 만든 입력 이미지의 전부를 보고 진위 여부를 판단하므로 생성자는 판별자를 속이기 위해 데이터의 일부 특징을 과장하려는 경향을 보인다. 이는 사람이 보는 이미지 퀄리티 여부와 관계 없이 판별자를 속이기 위한 데이터 생성을 하게 되므로, 결과

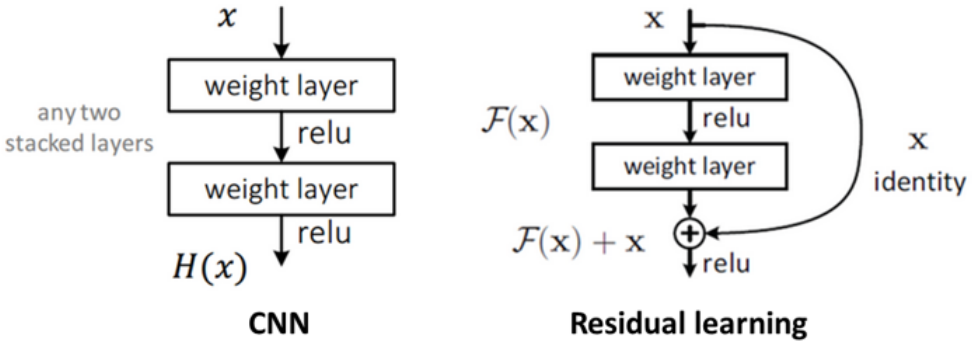


Figure 8 : Concept of residual connection

이미지에 블러가 끼거나 원하지 않는 결과가 도출되는 경우가 잦다. 따라서 전체 이미지에 대한 저주파(Low frequency) 성분을 L1 정규화를 통해 파악한 후, 고주파(High frequency) 성분 파악에 강한 PatchGAN과 결합하는 식으로 판별자를 구성한다. CycleGAN의 판별자도 이러한 PatchGAN 구조를 가지는데, 전체 영역이 아니라 특정 크기의 패치(patch) 단위로 생성자가 생성한 이미지의 진위 여부를 판단한다. 패치의 크기는 전체 이미지 크기에서 특정 픽셀과 다른 픽셀들 간의 연관성이 있는 적절한 범위를 포함해야 하는데, CycleGAN은 70×70 패치를 사용해 이미지를 판별한다. 이는 이미지의 부분별로 연산을 수행하므로 전체 파라미터 개수가 훨씬 적어지며, 전체 이미지 크기에 영향을 받지 않으므로 구조적으로 더 유연하다고 할 수 있다.

(2) 손실 함수

CycleGAN의 최종 손실 함수는 앞서 언급한 적대적 손실과 순환 일관성 손실의 합으로 아래와 같이 구성된다.

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + L_{cyc}(G, F) \tag{4}$$

이때 순환 일관성 손실(L_{cyc})은 아래와 같은 평균 제곱 오차(Mean Square Error, MSE)에 따라 계산된다.

$$\text{MSE} = \frac{1}{mn} \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} [I(i,j) - K(i,j)]^2 \tag{5}$$

이는 실제 값과 예측 값의 픽셀 값 차이에 의존하며 픽셀 단위로 오차가 역전파되기 때문에 네트워크 수렴 시간이 길고, 평균 제곱 오차의 특성에 따라 이미지가 지나치게 평균화되기 때문에 고해상도 텍스터의 디테일을 잡아내지 못해 이미지 품질이 떨어지게 된다. Fig.(9)에서 볼 수 있듯이 실제 학습 결과 일반적인 이미지 세트에서 작동하는 CycleGAN은 좋은 성능과 결과물을 보여 주지만, 적외선 합성 이미지 구성은 학습이 원활하게 이루어지지 않는 점을 포착하였다. 이는 일반적인 CycleGAN의 학습 방식으로는 고품질 적외선 이미지 생성이 어렵다는 점을 사사한다. 따라서 본 연구에서는 CycleGAN의 손실 함수를 일부 수정하여 이미지 생성 품질을 높이는 방법을 고안하였다.

Figure 9 : A single example of learning general dataset (Monet2Photo) and synthetic infra-red dataset Compared to the VIS data, synthetic IR data were not learned well

2) 손실 함수 재구성

(1) 구조적 유사도 지수 측정

구조적 유사도 지수 측정(Structural Similarity Index Measure, SSIM)[21]은 주어진 두 이미지의 유사도(similarity)를 측정하는 측도로 주로 사용된다. 이는 단순한 수치적 오차가 아니라 인간의 시각 시스템을 고안하여 설계되었는데, 밝기(luminance), 대비(contrast), 구조(structure)로 이루어진 3요소를 곱하여 최종 값이 산출된다. SSIM의 최종 결과 값은 0에서 1 사이이며 0에 가까울수록 두 이미지 간의 거리가 먼, 즉 상이한 이미지로 판단되고 1에 가까울수록 유사한 이미지로 판단된다. SSIM은 일반적으로 11×11이나 8×8의 슬라이딩 가우시안 윈도우(sliding Gaussian window)를 이용하여 픽셀 단위가 아닌 영역 단위로 이미지 비교를 수행하게 된다.

밝기 값은 이미지의 픽셀 값이 클수록 이미지가 밝아짐을 의미한다. 이미지의 평균을 이용

해서 두 이미지의 밝기를 비교하는 식은 다음과 같이 도출된다.

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{6}$$

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{7}$$

Eq.(6)을 이용하여 x, y 각 이미지의 평균 값을 구하고 이를 Eq.(7)에 대입하여 두 이미지가 같으면 $l(x,y)$값은 1이 된다. C_1은 분모가 0이 되는 것을 방지하는 상수로 $C_1 = (K_1L)^2$이며 $K_1 = 0.01$, L은 8비트의 픽셀 값을 사용하므로 225를 사용해 $C_1 = (0.01 \times 255)^2 = 6.5025$이 된다.

대비 값은 이미지 내에서 빛의 밝기가 바뀌는 정도를 나타낸 양으로, 픽셀 간 차이를 통해 정량화 가능하므로 픽셀 값의 표준 편차를 이용한다. 이미지의 표준 편차를 이용해서 두 이미지의 대비를 비교하는 식은 다음과 같이 도출된다.

$$\sigma_x = \frac{1}{N-1}\sum_{i=1}^{N}((x_i - \mu_x)^2)^{\frac{1}{2}} \tag{8}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{9}$$

밝기와 동일하게 Eq.(8)을 이용하여 Eq.(9)에 대입하여 두 이미지 간 대비 성분을 비교한다. 표본의 표준 편차를 계산하기 때문에 모표준 편차가 될 수 있도록 분모에 $N-1$을 사용하며, C_2의 경우 $C_2 = (K_2L)^2$이며 $K_1 = 0.03$이므로 $C_2 = (0.03 \times 255)^2 = 58.5225$이 된다.

구조는 픽셀 값 간의 구조적인 차이점을 나타내며 경계선 정보를 포함한다. 두 이미지의 구조적 유사성을 확인하는 것은 두 이미지의 상관 관계(correlation)을 이용하는 것으로, Eq.(6)과 Eq.(8)를 이용하여 픽셀 값을 $(X - \mu_x/\sigma_x)$로 재정의한다면 상관 관계는 다음과 같이 계산할 수 있다.

$$\mathrm{corr}(X,Y) = \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x\sigma_y} = E\left[\frac{(x-\mu_x)}{\sigma_x}\frac{(y-\mu_y)}{\sigma_y}\right] \tag{10}$$

따라서 두 이미지의 상관 관계를 구하는 것은 재정의된 픽셀 값의 곱의 평균을 구하는 것과 같기 때문에, 구조를 비교하는 식은 다음과 같이 도출된다.

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y) \tag{11}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \tag{12}$$

C_3은 최종 수식 도출의 편의를 위해 $C_2/2$로 사용한다. 최종적으로 밝기, 구조, 대비 값을 모두 반영한 SSIM은 다음과 같이 정의된다.

$$\mathrm{SSIM}(x,y) = l(x,y)^{\alpha} \cdot c(x,y)^{\beta} \cdot s(x,y)^{\gamma} \tag{13}$$

이때 $\alpha = \beta = \gamma = 1$이라면 식을 다음과 같이 정리할 수 있다.

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{14}$$

(2) 재구성된 손실 함수

SSIM은 두 이미지 사이의 유사도를 측정하기 때문에 이를 손실 함수로 활용한다면 두 이미지 사이의 거리를 줄이는, 즉 유사하게 만드는 작업을 수행할 수 있다. 또한 기존 평균 제곱 오차와는 달리 영역 단위로 이미지를 비교하기 때문에 보다 지역적인 특성을 보존할 수 있고, 대칭(symmetry) 성질을 만족하기 때문에 이미지 순서에 관계없이 양방향으로 적용할 수 있어 순환 일관성 구조를 가지는 CycleGAN의 손실 함수에 적용이 가능하다. 하지만 SSIM은 1에 가까울수록 두 이미지의 유사도가 커지기 때문에 손실 함수로 사용하기 위해 SSIM 값을 다음과 같이 재정의하였다.

$$L'_{SSIM}(G,F) = 1 - L_{SSIM}(G,F) \tag{15}$$

이때 SSIM의 윈도우 크기는 앞서 언급했듯이 조절이 가능한데, 윈도우 크기에 따라 이미지 품질에 차이가 있을 것이라 판단해 멀티 윈도우(Multi-window) 개념을 적용하였다.

$$L_{MSSIM} = \sum_{i} w_i L'_{SSIM_p} \tag{16}$$

이는 $p \times p$ 윈도우와 해당 윈도우에 가중치 w_i를 곱해서 계산되며, 본 연구에서는 표준 SSIM인 11×11과 22×22 윈도우를 이용하여 가중치 별 이미지 비교를 진행하였다. 손실 함수의 가중치는 $\lambda = 10$, $\alpha = 0.1$로 정의되며[18], 네트워크의 총 손실 함수는 다음과 같다. 또한 합성 이미지 생성을 위한 전체 알고리즘을 Fig.(10)에 도시하였다.

$$\begin{aligned} L(G, F, D_X, D_Y) &= L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) \\ &+\lambda L_{cyc}(G, F) + \alpha L_{identity}(G, F) + \beta L_{MSSIM}(G, F) \end{aligned} \tag{17}$$

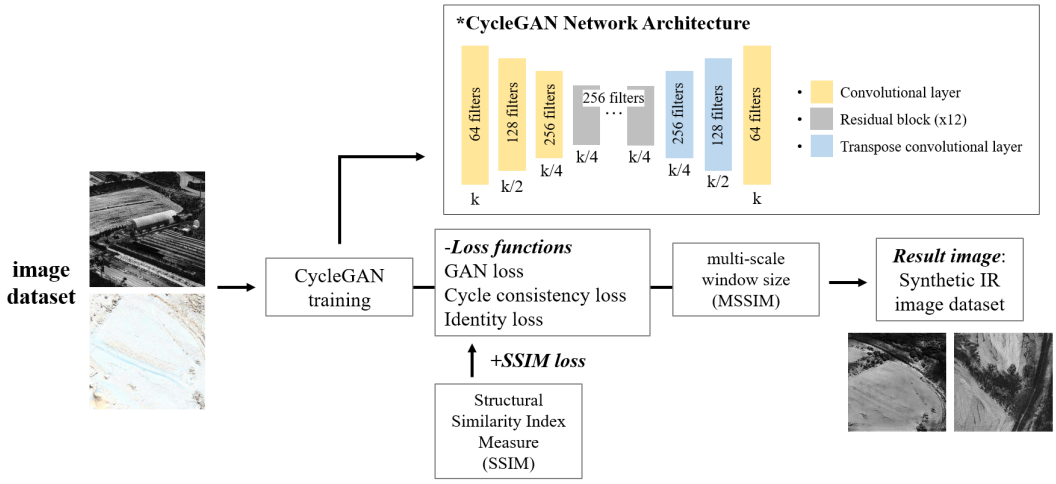


Figure 10 : Flowchart to illustrate the methodological steps of the study

3. 시뮬레이션 결과

1) cyc-MSSIM 가중치 비교

앞서 생성한 적외선 이미지와 합성 이미지를 데이터셋으로 하여 수정된 손실 함수를 가지는 CycleGAN으로 생성 모델 학습을 진행하였다. 학습에 사용된 총 이미지 수는 실제 적외선 이미지 1,908장과 합성 이미지 931장으로 학습 환경은 NVIDIA GeForce RTX 3090 2개로 구성하였다. Epochs는 200, Batch Size는 이미지 품질을 고려해 1로 설정하였다. 200 Epochs까지 학습 시간은 약 9시간이 소요되었다.

L_{MSSIM}의 윈도우 크기가 표준 SSIM인 11×11을 따를 때, L_{cyc}과 L_{MSSIM} 간의 적외선 이미지 생성에 가장 유리한 가중치를 찾기 위하여 $\lambda = 10$일 때 β 값을 다르게 하며 학습을 진행하였다. 이미지 세트에 대한 $\lambda - \beta$ 간 가중치 값은 Table.(3)에 도시하였으며 학습 이미지 결과를 Fig.(11)에 도시하였다. 이때 가장 적절한 가중치를 선별하기 위해 CaseA부터 CaseD까지 각 데이터셋과 실제 적외선 데이터셋과의 FID 점수를 계산하였다. FID 점수는 다음과 같이 정의된다.

$$\text{FID score} = ||m - m_w||_2^2 + \mathrm{T_r}(C + C_w - 2(CC_w)^{\frac{1}{2}}) \tag{18}$$

각 경우에 대한 FID 점수는 Table.(4)에 도시하였다. FID 점수는 실제 이미지와 생성된 이미지에 대해 두 집단 간 벡터 사이의 거리를 계산하며 값이 낮을수록 두 이미지 세트가 유사하다는 지표가 되므로, 점수가 낮을수록 성능이 좋은 모델로 평가된다. 계산 결과 4개의 케이스 중 가장 낮은 점수를 갖는 CaseB의 가중치(1.0)를 채택하여 멀티 윈도우 적용 시 비교를 진행하였다.

Table 3 : Weight parameters in different cases of L_{cyc} between L_{MSSIM}

Weight paramater	CaseA	CaseB	CaseC	CaseD
L_{cyc} (λ)	10	10	10	10
L_{MSSIM} (β)	0.1	1.0	10	100

Table 4 : FID socre in different cases of L_{cyc} between L_{MSSIM}

Weight paramater	FID score
CaseA	242.863
CaseB	**201.806**
CaseC	202.380
CaseD	221.065

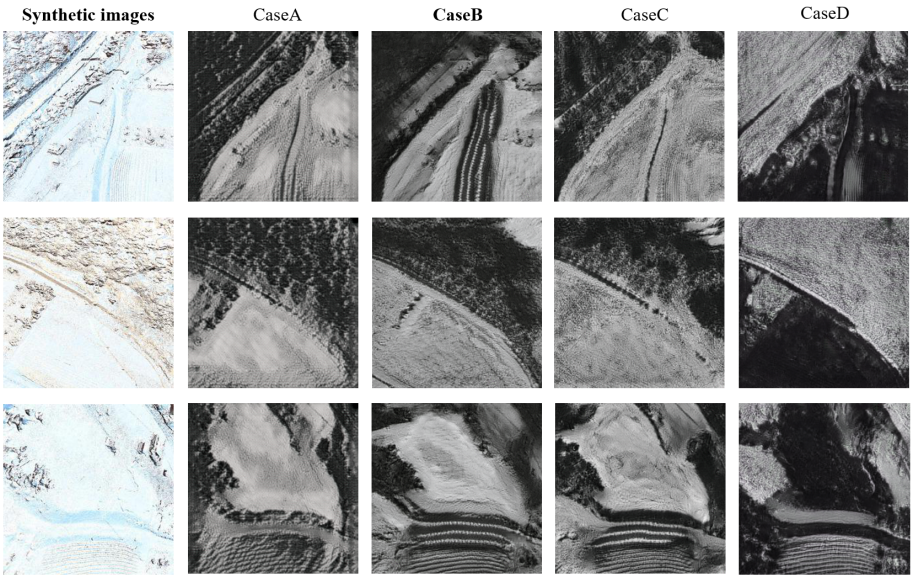


Figure 11 : Constructed image comparison due to the L1 loss(λ) between SSIM loss(β)

2) MSSIM 윈도우 가중치 비교

이번 단원에서는 $\lambda - \beta$ 간 가중치 값이 정해져 있을 때 L_{MSSIM}의 윈도우 크기에 따른 가중치를 다시 부여하여 학습 결과를 비교한다. 위와 동일한 CycleGAN으로 학습을 진행하였으며, 11×11 크기를 가지는 윈도우의 가중치 값을 w_1, 22×22 크기를 가지는 윈도우의 가중치 값을 w_2라 정의하고 Table.(5)와 같이 가중치를 부여하였다. Case1부터 Case5까지 윈도우 가중치 크기에 따른 학습 이미지 결과는 Fig.(12)와 같다. 학습 결과 기존 CycleGAN보다는 L_{MSSIM}이 손실 함수에 추가된 네트워크가, 표준 SSIM보다는 윈도우를 혼합하여 사용한 Case2, 3, 4가 향상된 이미지 생성 결과를 보였으며 Case3의 경우에는 입력으로 사용된 합성 이미지의 특성을 거의 흐리지 않고 적외선 도메인 변환을 수행하는 것을 확인하였다. 이때 위와 동일하게 실제 적외선 데이터셋과의 FID 점수를 계산한 결과, 육안으로도 가장 잘 생성되었다고 판단되는 Case3의 FID 점수가 가장 낮은 것을 확인하였다. 또한 알고리즘의 검증을 위해 FLIR 데이터셋에 대해서도 동일하게 검증을 진행하였다. 학습 이미지 결과는 Fig.(13)와 같으며 이도 Fig.(12)과 동일하게 기존 CycleGAN보다 L_{MSSIM}을 적용한 모델이 더 나은 성능을 보이는 것을 확인하였다. 각 경우에 따른 FID 점수는 Table.(6)에 도시하였다.

Table 5 : Weight parameters in different cases of window size

Case	Case1	Case2	Case3	Case4	Case5
$w_1(11 \times 11)$	1.0	0.6	0.5	0.4	0.0
$w_2(22 \times 22)$	0.0	0.4	0.5	0.6	1.0

Table 6 : FID socre in different cases of window size

Case/dataset	**Custom dataset(real/sythetic IR)**	**FLIR dataset**
L1 loss	208,748	232.017
Case1(CaseB)	201.806	**228.008**
Case3	**198.506**	232.487

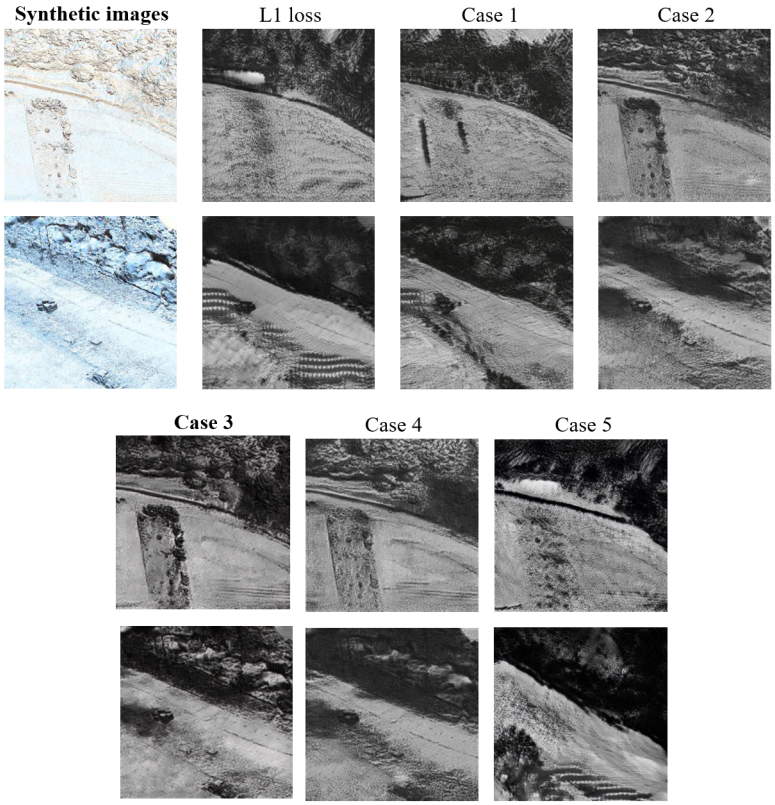


Figure 12 : Constructed IR images comparison due to multi-window weighting parameters

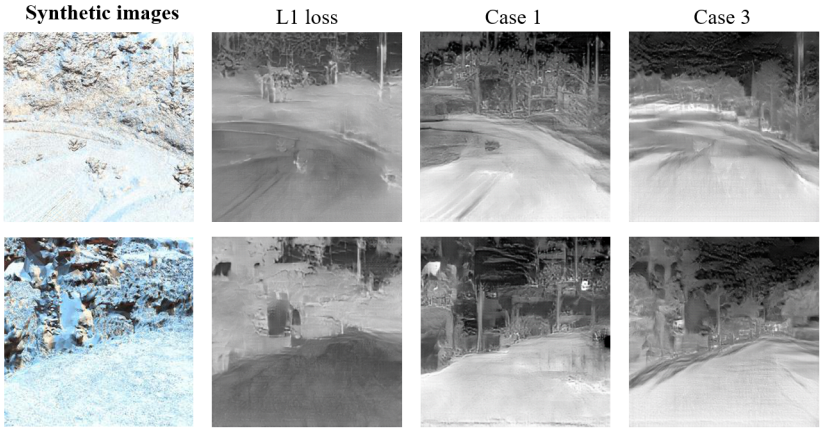


Figure 13 : Constructed FLIR images comparison due to multi-window weighting parameters

IV. 데이터 평가 및 분석 기법

본 장에서는 3장에서 생성한 각종 데이터들에 대한 평가 및 분석법에 대해 다룬다. 이미지 생성뿐만 아니라 이미지가 얼마나 잘 생성되었는지 평가하는 것 또한 중요한 쟁점이며, 인공지능의 발전으로 실제 데이터와 정교하게 생성된 합성 데이터를 구별하는 것은 어려운 문제이다. 이와 관련하여 사람 얼굴 이미지의 경우 인간의 눈으로는 실제 이미지와 합성 이미지를 구별하는 것이 사실상 불가능하다는 연구 결과가 소개된 바 있고, 구글에서는 인공지능 생성 이미지의 악용을 피하기 위해 자사 인공지능 모델로 생성된 이미지에 마크업(markup)을 삽입해 인공지능으로 생성된 이미지라는 것을 표기한다. 사실상 신뢰성 있는 합성 이미지 식별 방법이 없는 것이다. 특히 본 연구에서는 특정 오브젝트를 검출하는 것이 목표가 아니기 때문에 데이터의 명확한 평가 기준을 정립하기가 더 어렵다.

이러한 합성 이미지의 특이점을 조금이나마 식별하기 위해 이미지의 주파수 성분을 분석하는 방법론이 존재한다. 생성 모델에 의해 만들어진 합성 이미지는 주파수 도메인에서 실제 이미지와 유의미한 차이점을 나타낸다는 것인데, PSD 분석을 통해 이러한 이미지의 주파수 성분을 확인할 수 있다. 이는 푸리에 변환(Fourier transform)을 기반으로 이미지가 가지는 주파수를 빈도(frequency)에 따라 시각화한 것인데, 실제 이미지와 합성 이미지의 PSD를 확인해보면 스펙트럼 분포에서 상이한 특징이 발견된다는 것을 확인할 수 있다. PSD 기법을 이용하여 3장에서 생성한 데이터셋을 분석한 결과는 Fig.(14)와 같다. 자연 이미지의 스펙트럼 그래프는 Fig.(14)의 (a)와 같이 선형성을 보이는데, 합성 영상 (b)와 기존 CycleGAN의 결과 이미지 (c)는 저주파 영역의 주파수가 떨어지는 것을 확인할 수 있다. 스펙트럼이 실제 이미지인 (a)를 따르면 실제 이미지와 유사하다고 할 때, L_{MSSIM}을 손실 함수로 추가한 모델의 결과물인 (d)∼(h)가 상대적으로 저주파 영역을 많이 회복하였으며 일부 구간에서는 선형성이 추가된 것을 확인할 수 있었다. 이는 개선된 알고리즘의 합성 이미지가 기존 대비 실제 이미지와 유사하다는 것을 의미한다. 하지만 이는 이미지의 전체적인 스펙트럼 분포를 판단하므로 기존 모델 대비 L_{MSSIM}의 기능은 확인할 수 있으나, 이미지의 어떠한 부분 때문에 합성 이미지의 주파수 스펙트럼이 실제

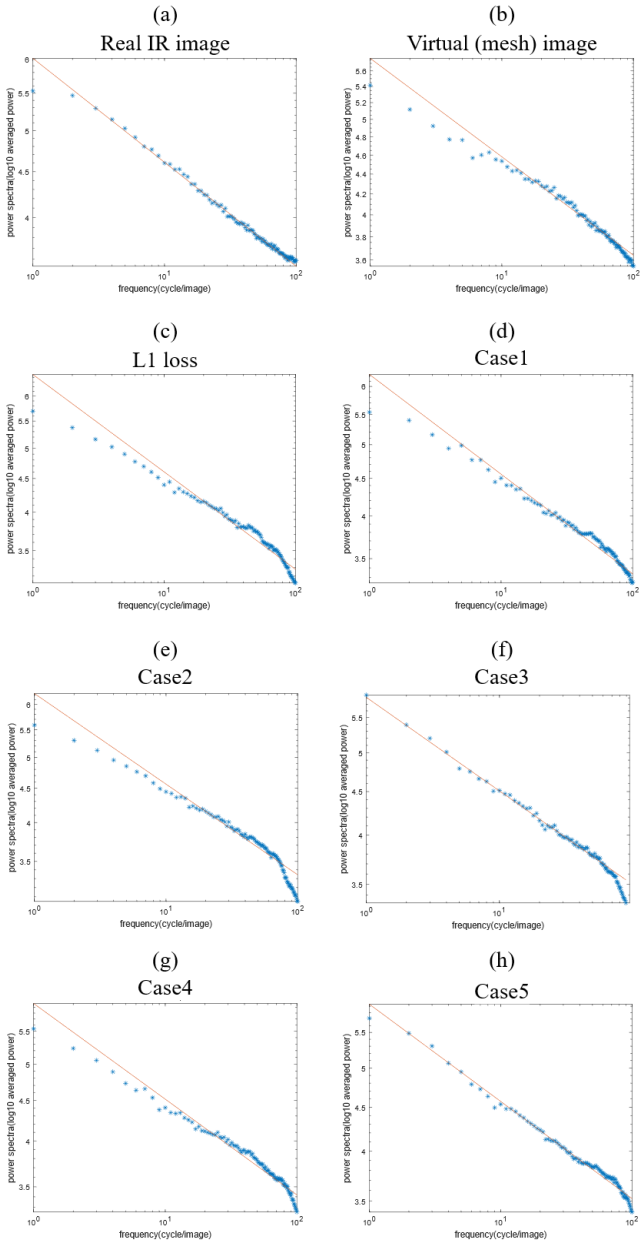


Figure 14 : Power spectrum analysis for the constructed image by CycleGAN proposed and a natural IR image

이미지와 유사해졌는지에 대한 설명으로는 불충분하다. 따라서 본 연구에서는 이미지 주파수 분석에서 나아가 XAI 기법을 활용하여 실제 이미지와 합성 이미지를 유의미하게 식별하고자 하였다.

1. LRP 기반 분석

1) 픽셀 단위 분해

LRP는 분해와 타당성 전파를 이용해 블랙박스 모델을 해부하며, 모델의 출력 값에서 시작해 입력 방향으로 관련성 점수(relevance score)를 계산해 나가며 비중을 분배한다. 임의의 d차원의 입력 값 $x = (x_1, x_2, ..., x_d)$에 대해 모델이 $f(x)$라는 값을 도출했을 때, x의 각 차원에 대한 관련성 점수는 다음과 같다.

$$f(x) = \sum_{d=1}^{V} R_d \tag{19}$$

첫 번째 레이어가 입력 이미지이고 마지막 레이어가 분류기가 출력하는 예측 값일 때, l번째 레이어는 $z = (z_d^{(l)})_{d=1}^{V(l)}$ 차원을 가지는 벡터로 모델링된다. LRP에서는 레이어 $l+1$에서 각 차원 $z_d^{(l+1)}$ 에 대한 벡터의 관련성 점수 $R_d^{(l+1)}$이 있다고 보며, 출력단에서 입력단으로 레이어별 관련성 점수를 계산한다.

$$f(x) = \cdots = \sum_{d\in l+1}^{d} R_d^{(l+1)} = \sum_{d\in l}^{d} R_d^{(l)} = \cdots = \sum_{d}^{d} R_d^{(1)} \tag{20}$$

Fig.(15)와 같은 분류기가 존재한다고 가정하자. 최상위 레이어가 7로 인덱스된 하나의 출력 뉴런일 때, 각 뉴런 i에 대해서 관련성 점수 R_i를 계산하려고 한다면 관련성 점수는 다음을 유지해야 한다.

$$\begin{aligned} R_7^{(3)} &= R_4^{(2)} + R_5^{(2)} + R_6^{(2)} \\ R_4^{(2)} + R_5^{(2)} + R_6^{(2)} &= R_1^{(1)} + R_2^{(1)} + R_3^{(1)} \end{aligned} \tag{21}$$

이때 각 레이어의 연결을 따라 전파될 수 있는 i와 j 사이의 메시지 $R_{i\leftarrow j}^{(l,l+1)}$에 대해 관련성 점수를 표현할 수 있다. 하지만 모델을 역순으로 탐지할 경우 Fig.(15)의 우측에 도시된 것과

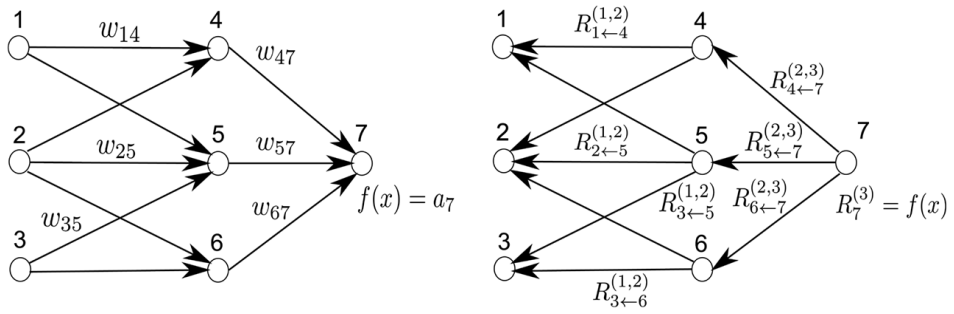


Figure 15 : A neural network-shaped classifier during prediction time

같이 메시지가 출력단에서 입력단으로 흐를 때 마지막 뉴런 7을 제외한 모든 뉴런의 관련성은 다음과 같이 정의한다.

$$R_i^{(l)} = \sum_{k\ :\ i \text{ is input for neuron } k} R_{i\leftarrow k}^{(l,l+1)} \tag{22}$$

뉴런 7은 수신받는 메시지가 없는 대신, 관련성을 $R_7^{(3)} = f(x)$로 정의한다. 이때 Eq.(22)에 따라 Eq.(21)을 다음과 같이 나타낼 수 있다.

$$\begin{aligned} R_7^{(3)} &= R_{4\leftarrow 7}^{(2,3)} + R_{5\leftarrow 7}^{(2,3)} + R_{6\leftarrow 7}^{(2,3)} \\ R_4^{(2)} &= R_{1\leftarrow 4}^{(1,2)} + R_{2\leftarrow 4}^{(1,2)} \\ R_5^{(2)} &= R_{1\leftarrow 5}^{(1,2)} + R_{2\leftarrow 5}^{(1,2)} + R_{3\leftarrow 5}^{(1,2)} \\ R_6^{(2)} &= R_{2\leftarrow 6}^{(1,2)} + R_{3\leftarrow 6}^{(1,2)} \end{aligned} \tag{23}$$

따라서 Eq.(23)에 따라 Eq.(22)을 재정의하면 다음과 같다.

$$R_i^{(l+1)} = \sum_{k\ :\ i \text{ is input for neuron } k} R_{i\leftarrow k}^{(l,l+1)} \tag{24}$$

2) 관련성 전파

특정 픽셀 x가 출력에 얼마나 영향을 주는지는 x의 값이 변화했을 때 $f(x)$ 값의 변화량을 통해 예측이 가능하다. 따라서 $f(x)$에 대해 입력 x_1, x_2의 기여도는 편미분을 통해 나타낼 수

있다.

$$\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \tag{25}$$

따라서 출력값을 기여도로 분해하기 위해 테일러 전개(Taylor series)를 이용한다. 임의의 매끄러운 함수 $f(x)$ 및 실수 a에 대한 $f(x)$의 테일러 전개는 다음과 같다.

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \tag{26}$$

이때 입력이 2개이고 출력이 1개인 신경망에 테일러 전개를 적용하기 위해 다변수 함수의 경우를 고려한다.

$$P(\mathbf{x}) = f(\mathbf{a}) + \frac{\partial f}{\partial x_1}(\mathbf{a})v_1 + \frac{\partial f}{\partial x_2}(\mathbf{a}^2)v_2 + \frac{1}{2}\frac{\partial^2 f}{\partial x_1^2}(\mathbf{a})v_1^2 + \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{a})v_1 v_2 + \frac{1}{2}\frac{\partial^2 f}{\partial x_2^2}(\mathbf{a})v_2^2 + \cdots \tag{27}$$

따라서 d차원의 입력에 대해 1차 다항식으로 근사한 테일러 전개는 다음과 같이 정의할 수 있다. 이는 x_p가 변했을 때 $f(x)$가 얼마나 변했는지, 즉 타당성 점수와 같은 개념이 된다.

$$f(\mathbf{x}) = f(\mathbf{a}) + \sum_{p=1}^{d} \frac{\partial \mathbf{f}}{\partial x_p} f(x)|_{\mathbf{x}=\mathbf{a}}(\mathbf{x}-\mathbf{a}) + \epsilon \tag{28}$$

위 식을 타당성 점수 $f(x)$를 구하는 식과 일치시키려면 $f(a) = \epsilon = 0$을 만족해야 한다. $f(a)$은 테일러 전개의 특성을 통해 $f(a) = 0$인 a를 찾고 그 지점에서 함수 근사화를 통해 0으로 만들 수 있고, ϵ은 신경망에서 출력되기 전에 렐루 활성 함수를 사용한다 가정한다.

$$f(x) = \max(0, \sum_{i=1}^{2} w_i x_i + b) = \begin{cases} 0: \text{ when} \sum_{i=1}^{2} w_i x_i + b) \leq 0 \\ \sum_{i=1}^{2} w_i x_i + b) \ : \text{ when} \sum_{i=1}^{2} w_i x_i + b) > 0 \end{cases} \tag{29}$$

위 식에서 렐루 함수의 특성 상 음수 값은 모두 0이기에 다음과 같이 양수인 경우만 살펴보도록 한다.

$$f(x) = \sum_{i=1}^{2} w_i x_i + b = f(a) + \sum_{i=1}^{d} \frac{\partial f}{\partial x_i}|_{x_i=a_i}(x_i - a_i) + \epsilon \tag{30}$$

현재 가정한 신경망은 $f(x) = w_1x_1 + w_2x_2 + b$이므로 $\partial f(x)/\partial x_1 = w_1, \partial f(x)/\partial x_2 = w_2$이고, 2차 이상 편미분 계수는 모두 0이므로 $\epsilon = 0$을 만족한다. 따라서 Eq.(29)은 최종적으로 다음과 같이 정리된다.

$$f(\mathbf{x}) = f(\mathbf{a}) + \sum_{p=1}^{d} \frac{\partial \mathbf{f}}{\partial x_p} f(x)|_{\mathbf{x}=\mathbf{a}}(\mathbf{x} - \mathbf{a}) \tag{31}$$

2. 설명 향상을 위한 개선된 LRP

1) 관련성 필터

출력 층 $l+1$에서 시작한 관련성 전파는 입력 층 l에 도달할 때까지 아래 규칙에 따라 네트워크의 각 뉴런에 관련성 점수를 할당한다. 뉴런 사이의 관련성은 다음에 따라 계산된다.

$$R_i^{(l)} = \sum_j \frac{a_{i'} w_{i'j}}{\sum_{i'} a_{i'} w_{i'j}} R_j^{(l+1)} \tag{32}$$

이때 양의 가중치만 사용하는 z^+ 규칙[22]을 사용하였다.

$$R_i^{(l)} = \sum_j \frac{a_{i'} w_{i'j}^+}{\sum_{i'} a_{i'} w_{i'j}^+} R_j^{(l+1)} \tag{33}$$

이때 각종 노이즈들이 작은 관련성 값과 연관될 가능성이 높기 때문에, z^+ 규칙을 통해 관련성 점수를 전달할 때 상위 k의 값만 통과하는 필터를 사용한다[23]. 이렇게 노이즈 값을 필터링하여 상위 관련성 점수로만 구성된 선명한 히트맵을 그릴 수 있다.

2) 관련성 점수 및 이미지 정규화

모델의 설명은 인간의 시각 시스템에 부합할수록 좋으며, 그에 따라 하나의 이미지에서 같은 클래스의 물체가 여러 번 발생하는 경우 일부분만 구별하거나, 객체의 전체 영역을 커버하지 못하고 특정 부분에만 히트맵을 그리지 않아야 한다. 하지만 관련성 필터를 사용하는 경우 일정 퍼센트의 상위 값에만 집중하기 때문에 인간의 시각에 부합하는 설명을 제공하기 어렵다는 단점이 있다. 따라서 관련성 필터를 통해 필터링된 점수를 값 정규화(normalize)를 통해 히트맵이 보다 넓은 영역을 커버할 수 있게 하였다. 정규화는 보편적으로 입력 샘플들 내에서 어떤 특성이

가장 중요한지 강조해 주는 효과가 있다. 예를 들어 이미지 분류에서 특정 클래스에 대한 중요한 이미지 영역을 찾는 데에 사용될 수 있기에 모델의 설명력을 높일 수 있다. 본 연구에서는 이에 따라 R값에 따라 행 정규화(row nomalization)를 수행하였다.

$$R_{ij} = \frac{R_{ij}}{\sum_j R_{ij} + \varepsilon} \tag{34}$$

또한 이미지 정규화(image normalization) 기법을 이용하였는데, 이는 전체 데이터셋의 분포를 고르게 해 주기 때문에 특정 값에 히트맵이 집중되는 기존 LRP의 결과를 개선할 수 있을 것이라 판단하여 적용하였다. 이는 입력 이미지 데이터셋의 RGB 채널에 대한 평균과 표준편차를 이용해 해당 값을 가지는 정규 분포가 되도록 데이터셋을 정규화시킨다. 학습에 이용한 각 데이터셋의 평균과 표준편차는 Table.(7)와 같다.

Table 7 : Mean and standard deviation of each dataset

Dataset	**(mean, std)**
Patterns-transition	(0.512,), (0.390,)
ImageNet(public)	(0.485, 0.456, 0.406), (0.229, 0.224, 0.225)
Real-synthetic	(0.452,), (0.227,)

3. 네트워크 구성

본 장에서는 3장에서 생성한 합성 이미지와 실제 이미지 간 차이점을 식별하고자 한다. 이를 위해 LRP 알고리즘을 이용하였고, 네트워크의 분류 결과를 시각화해 주는 LRP 알고리즘에 따라 합성 이미지와 실제 이미지를 적절하게 분류할 수 있는 분류 네트워크를 설계하였다. 네트워크는 VGGNet[24] 기반으로 설계되었는데, VGGNet은 합성곱 계층과 풀링 계층으로 구성된 딥러닝 네트워크로 해당 연구에서 적용된 VGG-16 모델은 16개의 층과 3×3 컨볼루션 필터로 구성되어 있다. 이는 16개라는 깊은 신경망 층임에도 불구하고 기존의 컨볼루션 필터를 1-layer 7×7 필터에서 3-layer 3×3 필터로 변경하며 파라미터가 감소할뿐더러 비선형 함수가 세 번 적용되게 되고, 모델의 비선형성이 증가함에 따라 모델이 특징점을 식별하는 데에 더 용이해진다.

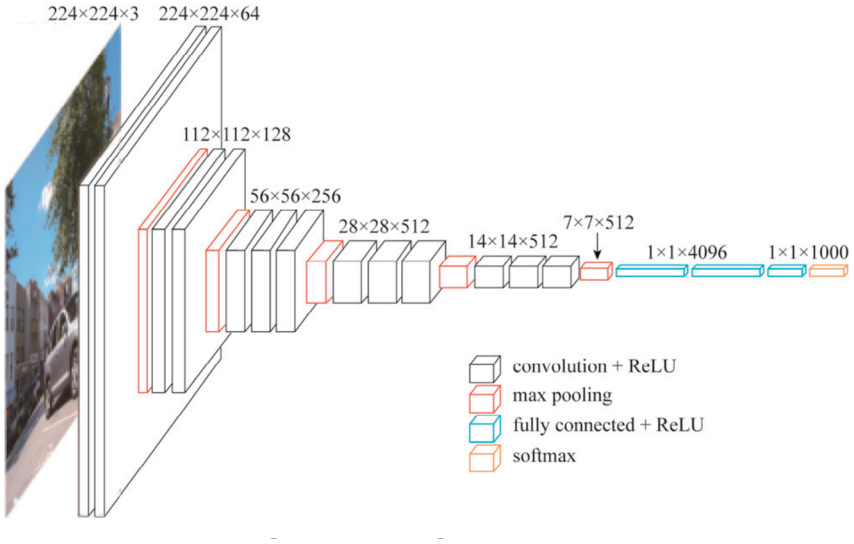


Figure 16 : VGG16 network structure

모델을 구축한 후 465장의 실제 적외선 이미지와 461장의 합성 적외선 이미지로 학습을 진행하였으며 CycleGAN의 학습 결과로 256×256을 가지는 모든 이미지는 학습의 편의를 위해 225×225로 변경한 후 224×224로 랜덤하게 크롭(Crop)하였다. Epochs는 100, Batch Size는 16으로 설정하였고, 학습 환경은 위와 동일하게 NVIDIA GeForce RTX 3090 2개로 구성하였다. 따라서 본 논문에서 최종적으로 제안하는 알고리즘을 Fig.(17)에 도시하였다.

학습이 완료된 분류 모델의 성능을 혼동 행렬(Confusion Matrix)을 통해 분석하였다. 혼동 행렬이란 학습된 모델의 분류 예측 오차가 얼마인지, 어떠한 유형의 예측 오류가 발생하고 있는지 나타내 주는 성능 지표이다. 이는 이진 분류와 다중 분류 모두 적용 가능하며, 이진 분류에서의 혼동 행렬은 Fig.(18)와 같이 구성된다. Fig.(18)에 나타난 혼동 행렬에서의 각 요소들을 통해 모델의 성능을 평가할 수 있다. 정확도(Accuaracy), 재현율(Recall), 정밀도(Precision) 3가지의 지표로 모델을 평가할 수 있으며 다음과 같이 계산된다.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{35}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{36}$$

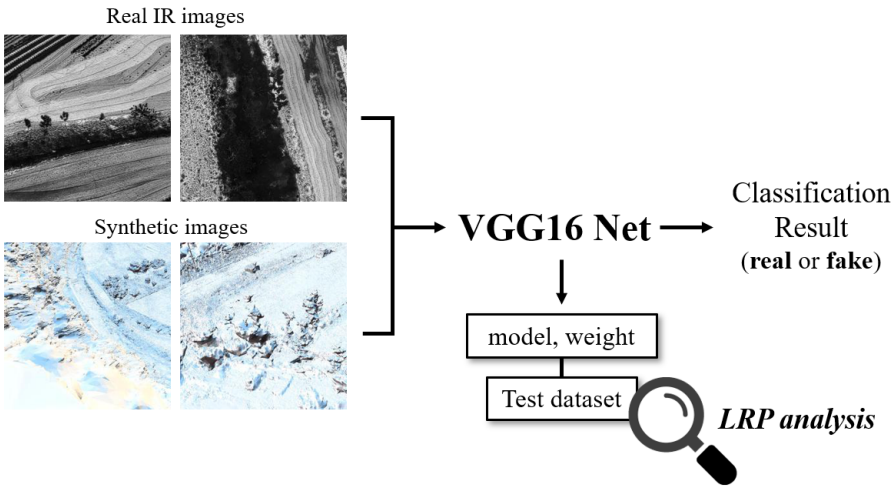


Figure 17 : Flowchart of process by which LRP is applied

$$\text{Precision} = \frac{TP}{TP + FP} \tag{37}$$

정확도란 모델이 얼마나 분류를 잘하는지 나타내는 지표로, 전체 값 중 예측에 성공한 값을 사용하기 때문에 모델이 극단적으로 잘못 학습된 경우에도 높은 값을 가지게 돼 정확도만으로는 모델을 평가하기 어렵다. 이에 따라 재현율과 정밀도를 함께 사용한다. 재현율은 실제로 긍정인 것 중(TP+FN) 긍정으로 예측한 비율(TP), 정밀도는 긍정으로 예측한 비율(TP+FP) 중 실제 긍정인 비율(TP)으로, 재현율과 정밀도는 상충 관계이기 때문에 둘 모두 올리는 것은 어렵다. 따라서 재현율과 정밀도의 조화 평균을 이용하여 모델의 성능을 측정한다. 이는 F1-score라 하며

Actual Values

Predicted Values		Positive	Negative
	Positive	**TP**(True Positive)	**FP**(False Positive)
	Negative	**FN**(False Negative)	**TN**(True Negative)

Figure 18 : Confusion matrix

다음과 같다.

$$\mathrm{F1-score} = 2 \times \frac{1}{\frac{1}{\mathrm{Precision}} + \frac{1}{\mathrm{Recall}}} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{38}$$

최종적으로 모델의 성능을 평가하기 위해 정확도와 F1-score를 고려하였으며, 모델의 혼동 행렬은 Fig.(19)와 같다. F1-score를 계산할 때 실제 이미지인 경우와 합성 이미지인 경우를 고려하여 두 클래스의 점수를 각각 계산해 평균으로 계산하였다. 계산 결과 정확도는 0.81, F1-score는 $0.765 + 0.84/2 = 0.8$의 값을 기록한 것으로 보아 모델이 분류 문제를 잘 해결하는 방향으로 학습하였다고 판단할 수 있다.

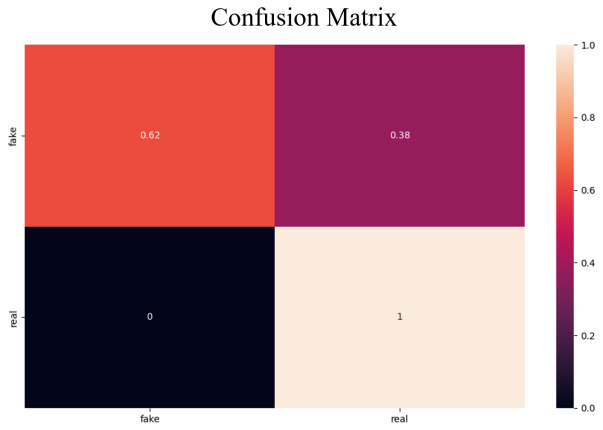


Figure 19 : Confusion matrix for proposed network

V. 생성 모델 데이터 분석

마지막 장인 본 장에서는 앞서 소개한 데이터셋 및 모델의 LRP 분석에 대해 다룬다. 3장에서 합성 적외선 데이터를 생성하였지만 적외선 데이터와 같은 고차원의 데이터셋은 인간의 눈으로 보더라도 구별할 수 없거나 설명이 어려운 부분이 존재한다. 따라서 고차원 데이터에 대한 이해를 높이기 위해 저차원의 데이터부터 LRP의 히트맵이 나타내는 특징점을 분석하고자 하였고, 이에 따라 4장에서 제안한 데이터셋에 추가로 패턴(patterns)-전환(transitions) 데이터셋과 ImageNet 데이터셋을 이용하였으며, 패턴-전환 데이터는 저차원의 데이터, ImageNet 데이터는 일반적인, 즉 중간 차원의 데이터, 적외선 데이터는 고차원의 데이터라 보고 각 데이터셋에 대한 분석을 순차적으로 진행하였다.

이때 LRP 히트맵 개선을 위해 여러 관점에서 이미지 분석을 수행하였는데, 일반 LRP와 관련성 필터를 적용한 LRP, 이미지 정규화를 수행한 뒤의 LRP, 관련성 필터에 관련성 점수 정규화(이하 R-정규화)를 수행한 LRP 결과를 비교해 각 알고리즘별로 히트맵을 어떻게 도시하는지 비교하였다. 또한 LRP가 이미지 내의 구체적인 특징이나 경계선 등에 반응하는 경향으로 미루어 볼 때 입력 데이터에 일부 변형을 가했을 때 LRP 결과가 달라질 것이라고 판단하여 생성 모델을 통해 생성한 변형 데이터셋에 대해 추가적인 분석을 진행하였다. 데이터셋에 대한 모든 분석은 4-3장에서 제안한 네트워크에 대해 수행하였다.

1. 시뮬레이션 결과

1) 패턴-전환 데이터셋

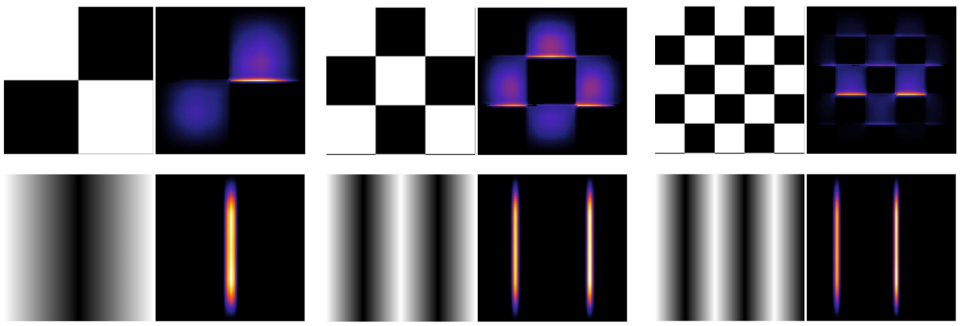
패턴-전환 데이터셋은 패턴 100장, 전환 100장의 데이터로 이루어져 있으며 분류 네트워크는 동일하게 VGG-16을 적용하였다. 해당 데이터에 대해 LRP 분석을 수행한 결과는 Fig.(20)와 같다. 기본적인 LRP 및 관련성 필터를 사용했을 때의 결과 (A)는 패턴 이미지의 경우 이미지의

흑과 백이 전환되는 경계선 일부분에만 히트맵이 집중된 것을 확인할 수 있고, 전환 이미지의 경우에도 이미지가 전환되는 전체적인 영역이 아닌 가운데 부분에만 히트맵이 집중되는 경향을 보였다. 이렇듯 이미지의 단적인 부분만을 기준으로 판단하는 것은 인간의 보편적인 판단 기준과 부합하지 않는데, 인간은 이미지의 단적인 부분이 아닌 전체적인 부분을 보고 판단하기 때문이다. 따라서 LRP 기법의 변형을 통해 해당 데이터셋을 분석한 결과는 (B), (C)와 같다. 이미지 정규화를 수행한 (B)의 경우에는 (A)의 결과보다 히트맵이 보다 이미지의 전체 영역에 생성되는 것을 확인할 수 있었다. 또한 이미지 정규화에 R 점수 정규화를 수행한 (C)의 경우, 히트맵이 보다 이미지 전체에 나타나는 것을 확인할 수 있었다. 해당 결과로부터 기존 LRP보다 변형을 수행한 결과가 더 인간 시각의 분별 시스템과 유사하게 분류 히트맵을 나타냈다는 결론을 도출할 수 있다.

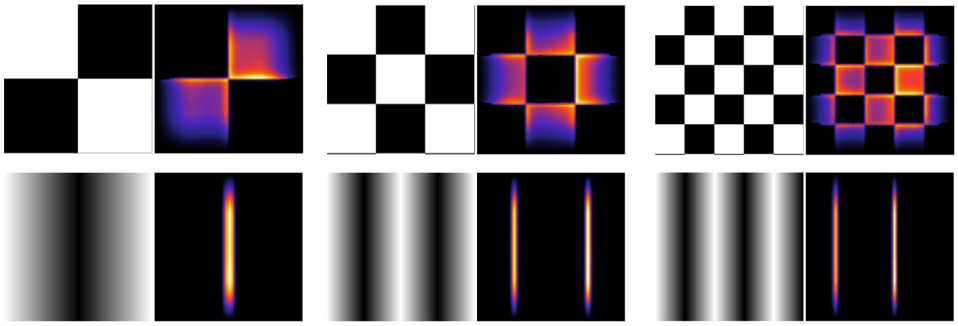
2) ImageNet 데이터셋

패턴-전환 데이터셋에 대한 결과를 토대로 ImageNet 데이터셋에 대한 학습을 진행한 후 LRP 분석을 수행한 결과는 Fig.(21)와 같다. (a)와 같은 입력 이미지가 있을 때 LRP z+ 규칙을 적용한 (b)를 살펴보면 피사체 외의 배경 성분 등에도 히트맵이 생성되는 것을 확인할 수 있다. 이러한 문제점을 해결하기 위해 상위 5%의 점수에만 히트맵을 그리는 관련성 필터를 적용한 결과는 (c)와 같다. (b)에 비해 불필요한 배경 성분이 상당 부분 제거된 것을 확인할 수 있는데, 4행, 5행과 같이 이미지 내에 피사체가 여러 개 있는 경우에는 하나의 피사체에만 집중하여 히트맵이 생성되는 것을 확인하였다. 패턴-전환 데이터셋의 분석 결과와 마찬가지로 인간은 이미지의 단적인 부분을 보고 판단을 수행하지 않기 때문에 이러한 점을 해결하기 위해 이미지 정규화 및 관련성 점수 정규화를 수행한 결과는 (d), (e)와 같다. 제안된 기법의 경우에는 기존 관련성 필터 적용 시보다 이미지의 전체적으로 히트맵이 도시되는 것을 확인할 수 있었고, 특히 피사체가 여러 개 존재할 경우 하나의 피사체에만 집중하는 것이 아니라 주변의 다른 피사체까지 집중하는 히트맵을 그리며 보다 인간 시각의 분별 시스템에 부합하는 결과를 도출하였다.

(A) LRP z+ rule



(B) Image dataset normalization + LRP z+ rule



(C) Image dataset normalization+ LRP z+ rule + R normalization

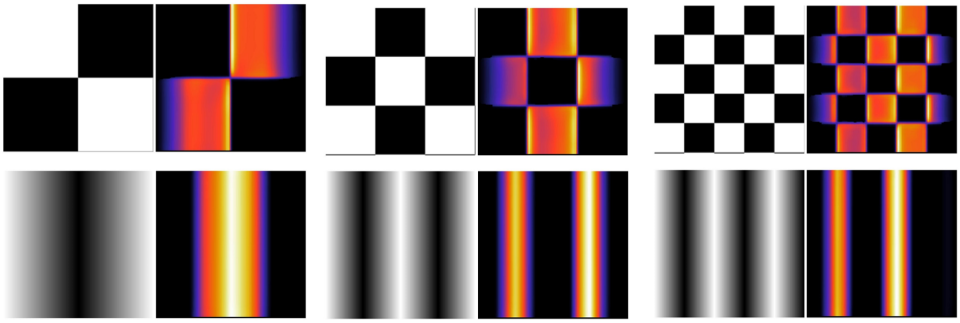


Figure 20 : Confusion matrix for proposed network

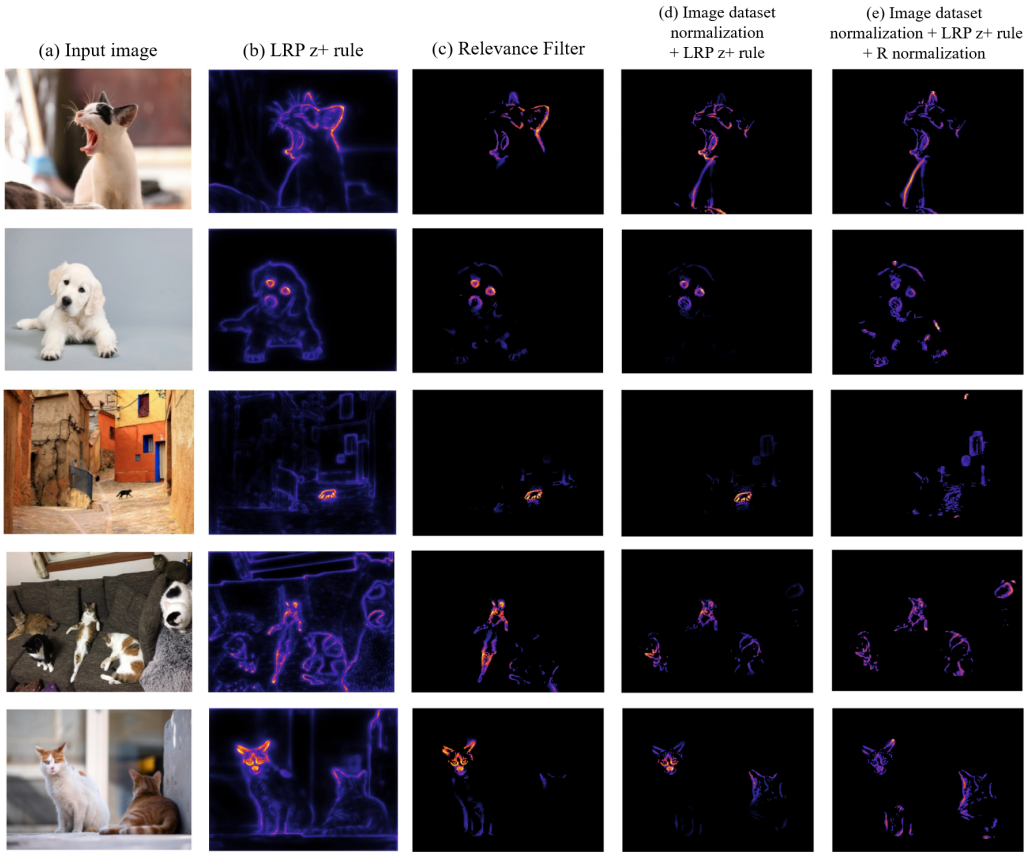


Figure 21 : Confusion matrix for proposed network

3) 합성-실제 적외선 데이터셋

(1) LRP 변형에 따른 결과 분석

앞선 저차원 데이터에 대한 분석 결과를 토대로 합성-실제 적외선 데이터셋에 대한 분석을 진행하였다. LRP 수행 시 양의 가중치만 사용하는 z^+ 규칙을 사용한 LRP를 기준으로 각 테스트 이미지에 대해 관련성 필터를 적용한 결과와 관련성 필터에 R-정규화를 수행한 결과 이미지는 Fig.(22)와 같다. 테스트 데이터셋은 PSD 분석 결과에 따라 다양하게 구축하였으며 1행은 실제 이미지를 모델이 실제로 분류한 경우, 2행은 합성 이미지를 실제로 분류한 경우, 3행은 합성 이미지를 합성으로 분류한 결과이다. R-정규화를 수행한 이미지 기준으로 붉은색과 푸른색 히트맵이 표현된 부분이 모델이 이미지 분류를 수행할 때 집중한 부분이다. 모델이 실제 이미지라고 판단한 1행과 2행 히트맵을 살펴보면, 실제 이미지와 같이 구체적인 특성이나 구조물 등이 이미지에 나타난 경우 그 부분에 대해 집중하는 모습을 보였으며, 그에 따라 히트맵이 상당히 구체화되어서 나타난 것을 확인할 수 있었다. 반면에 합성 이미지로 판단된 합성 이미지의 경우에는 히트맵이 상대적으로 가우시안 블러 처리된 것처럼 흐려지는 경향을 보였고 이미지의 경계선 부분에 상대적으로 집중하는 모습을 보였다.

(2) 데이터에 따른 결과 분석

기존 LRP의 분석 결과를 살펴보면 이미지의 경계선 부분에 집중하는 경향을 보이는 것을 확인할 수 있다. 이미지에서 경계란 값이 급격하게 변화하는 부분으로, 객체 검출 등에 있어서도 이러한 경계선 정보를 활용하기도 한다. 따라서 입력 이미지에 변화를 가했을 때 네트워크 분석 결과 및 LRP 히트맵을 확인하기 위하여 입력으로 사용된 테스트 데이터셋에 일부 변형을 가하는 작업을 수행하였다. 변형 데이터는 3장에서 제안한 생성 모델을 사용하되 이미지 생성 시 적용되는 손실 함수 L_{MSSIM}의 값을 일부 수정하였다. 앞서 언급하였듯이 SSIM은 밝기, 대비, 구조를 비교하여 이미지 품질을 파악한다. 이때 밝기 값을 제외한 대비와 구조 성분은 생성되는 이미지 품질에 영향을 미칠 수 있기 때문에[25] 대비와 구조 값에 관여하는 Eq.(8)과 Eq.(11)을 다음과 같이 재정의하였다.

$$\sigma_m = w_s \sigma \tag{39}$$

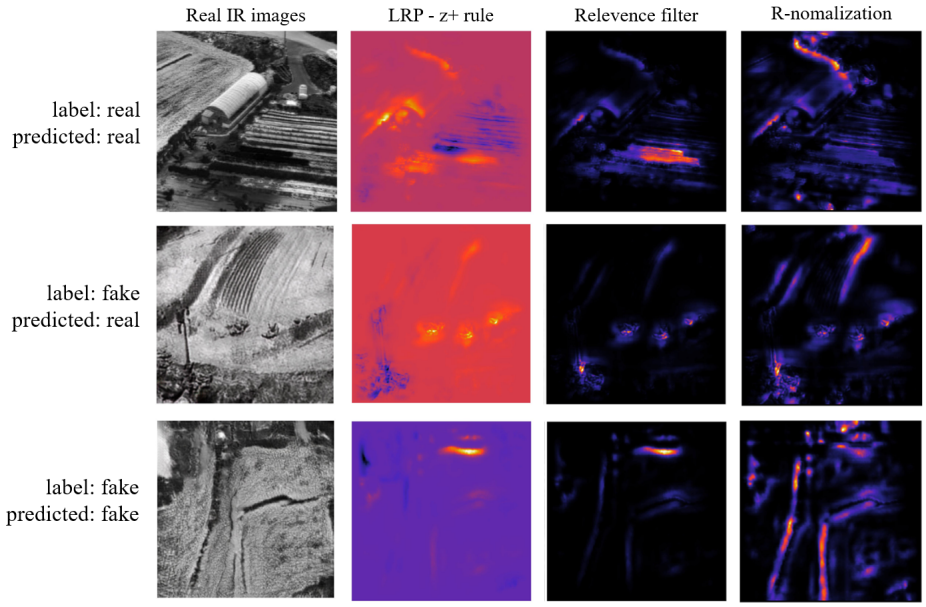


Figure 22 : Constructed image comparision due to LRP transforms

$$\sigma_c = w_c \sigma_{xy} \tag{40}$$

Eq.(39)에 따라 대비 성분이 증가된 이미지가, Eq.(40)에 따라 구조 성분이 증가된 이미지가 생성되며 생성된 이미지와 네트워크 분류 결과는 Fig.(23)에 나타내었다. 2행은 파라미터 변형을 가하지 않은 Case3의 네트워크 생성 이미지이고 3은 w_s에 20만큼의 가중치를 부여해 대비 성분을 증가시킨 이미지, 4행은 w_c에 20만큼 가중치를 부여해 구조 성분을 증가시킨 이미지이다. 세 가지의 케이스 모두 이미지 변환을 잘 수행하였는데 2행의 이미지는 네트워크가 모두 실제 이미지라고 분류한 반면에 3행의 이미지는 모두 합성 이미지로 분류하였으며, 4행은 반만 실제 이미지로 분류하였다. 이렇게 같은 합성 이미지로부터 생성된 이미지임에도 불구하고 이미지 특성에 따라 분류 결과가 상이하게 나오는 것을 확인해, 이를 바탕으로 해당 이미지들에 대한 LRP 분석을 진행하였다.

위 이미지들에 대해 LRP 분석을 수행한 결과는 Fig.(24)와 같다. LRP 분석은 제안된 R-정규화 기법을 적용하여 수행하였으며 전부 실제 이미지로 판단된 Fig.(24)의 1열의 경우에는 Fig.(22)의 결과와 동일하게 히트맵이 구체적인 특성 및 구조물에 집중하는 것을 확인할 수 있

다. 반면에 전부 합성 이미지로 판단된 2열의 경우에는 유사한 이미지임에도 불구하고 히트맵이 전체적으로 선명하지 않고 지형 지물보다는 경계선에 집중하는 경향을 보였다. 일부만 실제 이미지로 판단된 3열의 경우에는 1열만큼은 아니지만 전체적인 히트맵이 구체화된 것을 확인할 수 있다. 이를 통해 합성 이미지의 LRP를 분석했을 때 히트맵 정보를 이용해 해당 이미지가 실제 이미지와 유사하게 잘 생성된 것인지 검증이 가능할 것으로 보인다.

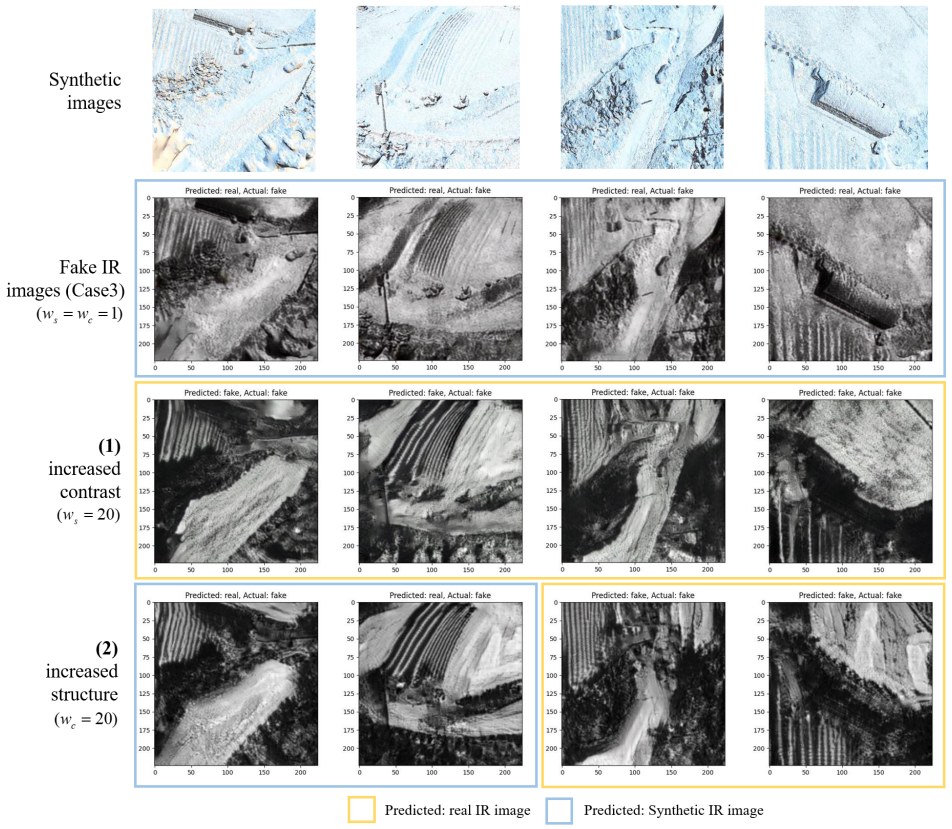


Figure 23 : Network classification results for generated images

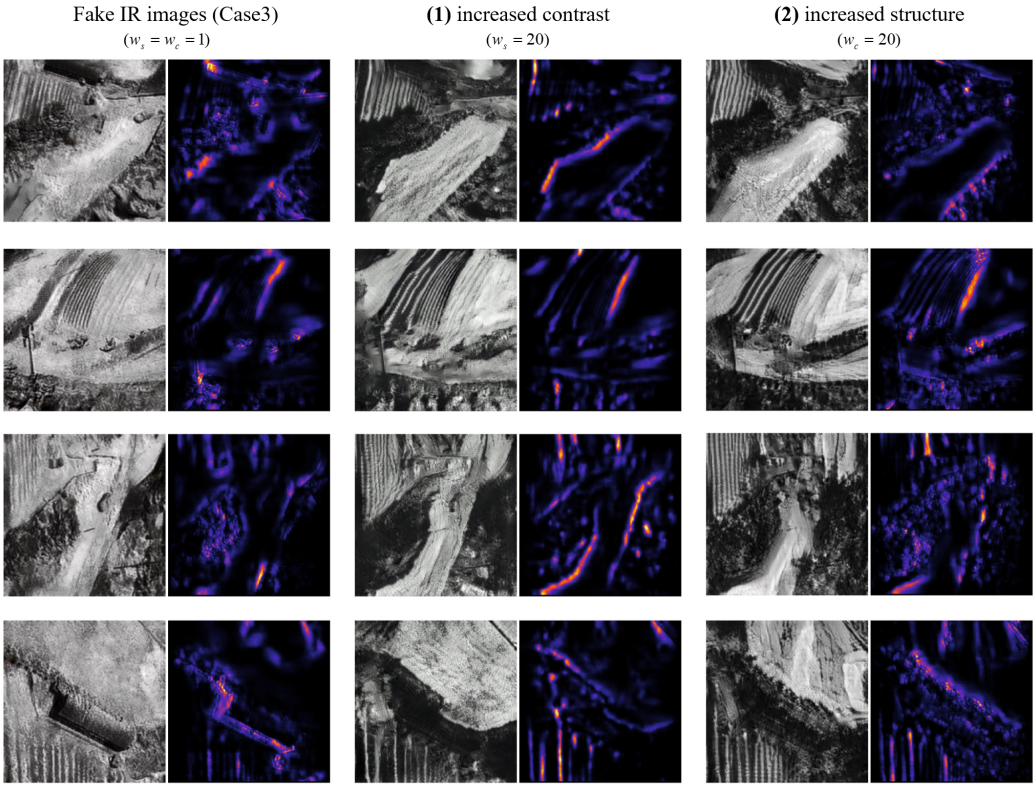


Figure 24 : LRP comparison of constructed images according to changes in SSIM weight parameters

VI. 결론

적외선 이미지는 모의 비행 시험, 탐색기 알고리즘 검증 등에 다양하게 사용되고 있으며 실제 적외선 이미지를 구축하는 데에는 비용이 많이 들고 환경 구축이 어렵기 때문에 합성 영상을 주로 이용한다. 하지만 잘 만들어진 합성 이미지도 인간의 눈으로 구별 가능한 이질감이 존재하는데, 인간은 다양한 정보를 통해 이미지를 인식하기 때문에 영상에서 느껴지는 이질감이 어떤 것 때문이라는 것을 정의 내리기 어렵다. 따라서 본 논문에서는 적외선 합성 데이터셋을 보다 잘 평가하기 위해 생성 모델을 이용하여 합성 이미지를 생성하는 알고리즘을 구축하고, XAI를 활용해 해당 데이터셋을 평가하는 기법에 대해 제안하였으며 시뮬레이션을 통해 이를 검증하였다.

먼저 짝지어진 데이터셋 구축이 어려운 환경에서 적외선 데이터셋과 합성 데이터셋을 구축하여 이를 생성 모델의 일종인 CycleGAN 알고리즘을 이용해 학습하였다. 이때 학습 시 적외선 영상의 특성 상 도메인이 단일화되어 학습이 원활하게 이루어지지 않는 점을 고려해 CycleGAN의 손실 함수에 L_{MSSIM}을 추가함으로써 결과 이미지 향상을 보였고, 이를 다양한 가중치 상에서 비교하였다.

또한 생성한 데이터셋을 평가하기 위하여 기존의 주파수 도메인 해석뿐만 아니라 XAI 방법론을 이용하였다. XAI는 블랙박스 구성인 인공지능의 판단 결과를 인간이 이해할 수 있게끔 표현해 주는 기법으로, XAI 방법론 중 필터 시각화 기법의 하나인 LRP를 이용하였다. LRP는 분류 모델의 결과를 시각화하여 설명해 주기 때문에 이를 적용하기 위해 실제 데이터셋과 합성 데이터셋을 분류하는 분류 모델을 VGGNet 기반으로 구축하였고, LRP 적용 시 관련성 필터와 관련성 점수 정규화를 적용하여 모델의 판단 근거를 시각화하였다. 또한 고차원 데이터에 대한 이해도를 높이기 위하여 저차원의 패턴-전환 데이터셋과 일반적인 조도, 각도, 배경을 포함하는 ImageNet 데이터셋을 이용하여 LRP 결과에 대해 분석하였다.

이때 기존 LRP와 관련성 필터의 경우 히트맵이 이미지 전체에 고르게 퍼지지 않는 경향을 보였는데, 정규화 기법들을 통해 이미지에 전체적으로 히트맵이 도시되게 하며 인간 시각 시스템

에 보다 부합할 수 있는 설명을 이끌어냈다. 또한 적외선 데이터의 경우 이미지 생성에 관여하는 SSIM의 특성을 이용해 입력 이미지의 구조, 대비를 일부 변형시켜 LRP를 적용한 결과, 모델이 실제 이미지라고 판단한 이미지는 히트맵이 고해상도로 생성되는 반면에 합성 이미지라고 판단한 이미지는 상대적으로 저해상도의 히트맵을 나타냈으며 구체적인 지형 지물보다는 경계선에 집중하는 모습을 보이며 두 이미지 간 차이를 식별하였다. 이를 통해 어떠한 합성 이미지가 존재할 때 LRP 분석 결과를 통해 실제 이미지와 유사하게 생성된 것인지 검증이 가능할 것으로 보이며, 본 논문에서 도시한 이미지의 다양한 변형을 통해 합성 이미지의 명확한 개선점 또한 파악할 수 있을 것으로 기대된다.

참고문헌

[1] R. Zhang, C. Mu, M. Xu, L. Xu, Q. Shi, and J. Wang, "Synthetic ir image refinement using adversarial learning with bidirectional mappings," *IEEE Access*, vol. 7, pp. 153 734–153 750, 2019.

[2] V. Kniaz, V. Gorbatsevich, and V. Mizginov, "Thermalnet: a deep convolutional network for synthetic thermal image generation," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 41–45, 2017.

[3] V. Mizginov, "Synthetic thermal background and object texture generation using geometric information and gan," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 149–154, 2019.

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[5] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[7] v. A. Van der Schaaf and J. v. van Hateren, "Modelling the power spectra of natural images: statistics and information," *Vision research*, vol. 36, no. 17, pp. 2759–2770, 1996.

[8] D. Pamplona, J. Triesch, and C. Rothkopf, "Power spectra of the natural input to the visual system," *Vision research*, vol. 83, pp. 66–75, 2013.

[9] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[12] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.

[14] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017.

[15] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[16] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on computational imaging*, vol. 3, no. 1, pp. 47–57, 2016.

[22] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern recognition*, vol. 65, pp. 211–222, 2017.

[23] K. Fischer, "Relevance propagation with pytorch," https://kaifishr.github.io/2021/12/15/relevance-propagation-pytorch.html, 2021.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[25] S. H. Lee and H. Leeghim, "Synthetic infra-red image evaluation methods by structural similarity index measures," *Electronics*, vol. 11, no. 20, p. 3360, 2022.