박사학위논문

색상·주파수 특성 변화에 따른 ResNet 모델의 신뢰도 인과 분석

2026년

한 성 대 학 교 대 학 원

스마트융합컨설팅학과

스마트융합컨설팅전공

권 태 윤

박사학위논문
지도교수 노광현

색상·주파수 특성 변화에 따른 ResNet 모델의 신뢰도 인과 분석

Causal Analysis of ResNet Model Confidence Under Color and Frequency Feature Interventions

2025년 12월 일

한 성 대 학 교 대 학 원

스마트융합컨설팅학과

스마트융합컨설팅전공

권 태 윤

박사학위논문
지도교수 노광현

색상·주파수 특성 변화에 따른 ResNet 모델의 신뢰도 인과 분석

Causal Analysis of ResNet Model Confidence Under Color and Frequency Feature Interventions

위 논문을 컨설팅학 박사학위 논문으로 제출함

2025년 12월 일

한 성 대 학 교 대 학 원

스마트융합컨설팅학과

스마트융합컨설팅전공

권 태 윤

권태윤의 컨설팅학 박사학위 논문을 인준함

2025년 12월 일

심사위원장 김 승 천 (인)

심 사 위 원 노 광 현 (인)

심 사 위 원 이 후 진 (인)

심 사 위 원 남 현 우 (인)

심 사 위 원 임 황 용 (인)

국 문 초 록

색상·주파수 특성 변화에 따른 ResNet 모델의 신뢰도 인과 분석

한 성 대 학 교 대 학 원
스 마 트 융 합 컨 설 팅 학 과
스 마 트 융 합 컨 설 팅 전 공
권 태 윤

딥러닝 기반 이미지 분류 모델, 특히 ResNet 계열의 합성곱 신경망은 다양한 응용 분야에서 높은 성능을 보이고 있지만, 입력 이미지를 어떤 시각적 단서에 근거하여 특정 클래스를 선택하는지 그 내부 의사결정 구조가 명확하게 드러나지 않는다는 한계를 지닌다. 실제 환경에서 발생하는 조명 변화는 밝기 분포를 뒤틀어 명암 대비를 왜곡시키고, 색상 변화는 색조와 채도 정보가 원래 형태를 잃게 만들어 색채 구조를 흔들며, 질감 손실은 미세한 패턴을 사라지게 하여 세부 구조의 인식을 어렵게 만든다. 이러한 비정형 입력은 모델이 특징을 추출하는 전체 과정에 영향을 미쳐 내부 표현을 불안정하게 만들고, 결과적으로 예측 신뢰도가 급격히 감소하는 주요 원인이 된다. 이러한 취약성은 대규모 정제 데이터셋을 중심으로 학습된 모델이 실제 환경에서 나타나는 다양한 밝기 변화, 색채 강도 변화, 윤곽의 대비 변화와 같은 물리적 변동을 충분히 반영하지 못하는 구조적 한계에서 비롯된다. 기존의 설명가능 인공지능 기법은 주로 사후적 시각화를 통해 모델이 어디를 보았는가를

보여주는 수준에 머물러 있으며, 입력의 물리적 변화가 모델의 판단 과정과 결과에 어떤 인과적 영향을 미치는지까지는 규명하지 못한다.

본 연구는 이러한 한계를 극복하기 위해 색상·주파수 개입 기반 인과 분석 프레임워크(Color–Frequency Intervention-based Causal Analysis Framework)를 제안한다. 이 프레임워크는 LAB/HSV 색상 채널과 웨이블릿 서브밴드에 대해 능동적 개입을 수행하고, 변화된 입력이 예측 확률 변화(ΔP), Score-CAM 구조, 주파수 반응에 어떤 인과적 영향을 미치는지 실시간으로 측정한다. 이를 통해 관찰 중심 XAI를 넘어서, 입력–반응–결과의 인과 경로를 실험적으로 규명하는 능동형 분석 체계를 구현한다.

색상 개입 실험에서는 저조도 환경에서 L 채널이 0–20 범위로 압축되고 A/B 색채 정보가 0 근처로 수렴하는 이중 압축(double compression) 현상이 확인되었다. 이로 인해 다양한 장면이 모두 무특징·저조도 영역으로 왜곡되며 모델의 기준 신뢰도(P_{base})가 크게 감소하였다. 본 프레임워크를 통해 L 채널을 확장하고 A/B 색차를 복원한 결과, 윤곽·색채 단서가 회복되었고 예측 신뢰도는 평균 15–30% 증가하였다. 밝은 조도에서도 HSV의 V 채널 과포화와 S 채널 감소로 인해 구조적 정보가 소실되었으나, 조도·채도 개입을 통해 ImageNet 학습 분포에 가까운 특징이 복원되면서 신뢰도 저하가 크게 완화되었다. 주파수 개입 실험에서는 ResNet50이 초기 합성곱 계층이 웨이블릿 서브밴드와 구조적으로 대응함이 확인되었고, LH·HL 대역은 윤곽·경계, HH 대역은 질감을 반영하였다. 특정 서브밴드 증폭 시 형태·질감 기반 특징이 재활성화되며 신뢰도가 평균 0.10~0.25 증가하였다.

본 연구는 이러한 예측 확률 변화를 기반으로 색상 인과 그래프와 주파수 인과 그래프를 제공하여, 입력 조작이 모델의 신뢰도·활성화 구조·주파수 선택성에 미치는 영향을 정량적으로 설명한다. 이는 단순 시각화 수준을 넘어, 모델 편향을 진단–개입–정량화–교정하는 능동적 인과 디버깅 도구로 확장될 수 있다. 향후 개입 자동화·최적화를 통해 이미지 보정, OOD 방어, 데이터 중심 모델 개선으로의 발전이 가능하며, 의료 영상, 포렌식, 이상 탐지, 딥페이크, 산업 검사 등 다양한 비전 분야에서 편향 교정과 신뢰도 향상에 기여할 수 있다. 또한 본 프레임워크는 생성형 AI 모델에도 적용 가능하여, 생성과정

의 편향·비정상 패턴·조작 가능성을 규명하는 인과 기반 XAI 방식으로 확장될 잠재력을 가진다.

【주제어】 색상·주파수 개입 기반 인과 분석, 설명 가능 인공지능, 색상 개입, 주파수 개입, 색상 인과 그래프, 주파수 인과 그래프

목 차

표 목 차

그 림 목 차

약 어

CAM	Class Activation Mapping
CIELAB	Commission Internationale de l'Éclairage L*a*b*
CNN	Convolutional Neural Network
Diffusion	Diffusion Model
Grad-CAM	Gradient-weight Class Activation Mapping
Grad-CAM++	Gradient-weight Class Activation Mapping Plus Plus
GAN	Generative Adversarial Network
HSV	Hue Saturation Value
HDM-WaveNet	High-Dimensional Multi-scale WaveNet
LAB	Luminance-A Channel-B Channel Color Space
MRA	Multi-Resolution Analysis
ResNet	Residual Neural Network
Score-CAM	Score-Weighted Class Activation Mapping
Swin	Swin Transformer
VGGNet	Visual Geometry Group Network
ViT	Vision Transformer
VLM	Vision-Language Model
XAI	Explainable Artificial Intelligence

제 1 장 서 론

제 1 절 연구 배경

1) 딥러닝 모델의 발전과 한계

지난 10여 년간 합성곱 신경망(Convolutional Neural Networks, CNNs)을 중심으로 한 딥러닝 모델은 컴퓨터 비전 분야의 혁신을 이끌어 왔다. ImageNet[1]과 같은 대규모 데이터셋을 기반으로 학습된 CNN은 이미지 분류, 자율주행, 의료 영상 분석 등 다양한 응용 분야에서 인간 수준을 상회하는 인식 능력을 보여주었다. 그러나 수백만 개의 파라미터와 심층 구조로 이루어진 CNN은 내부 의사결정 과정이 명확히 드러나지 않는 블랙박스 특성을 지니고 있으며, 이러한 해석 불가능성은 실제 응용에서 모델의 신뢰성(reliability)과 해석 가능성(interpretability)을 확보하는 데 큰 장애 요소로 작용하고 있다.

2) 설명가능 인공지능의 등장과 발전

이러한 문제를 해결하기 위한 방안으로 설명 가능 인공지능이 등장하였다. 특히 Grad-CAM, Score-CAM 등 Class Activation Mapping 계열 기법[2]은 모델이 입력 이미지의 어떤 영역을 분류 근거로 활용했는지를 시각적으로 표현함으로써 블랙박스 모델을 부분적으로 해석할 수 있는 수단을 제공하였다.

한편, 이미지 복원(Image Restoration)과 초해상도(Super-resolution) 연구에서는 주파수 도메인 분석이 중요한 역할을 해 왔다. 웨이블릿(Wavelet) 변환을 비롯한 주파수 기반 접근은 저주파(전역 구조)와 고주파(세부 텍스처)를

1) Deng, J., Dong, W., Socher, R., Li, L.-J., Li. (2009). ImageNet: A Large-Scale Hierarchical Image Database.

2) Selvaraju et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization.

분리하여 물리적 신호 구조를 보존하는 데 강점을 보이며, 최근에는 이러한 주파수 분석이 딥러닝 모델의 내부 인식 구조를 해석하는 단서로 활용될 가능성이 제시되고 있다.[3)]

3) 도메인 격차와 조명 편향 문제

그럼에도 불구하고, 현 딥러닝 모델은 여전히 학습 데이터와 실제 입력 데이터의 분포가 불일치할 때 성능이 급격히 저하되는 분포 외 문제[4)5)]에 취약하다. 이는 곧 도메인 격차(Domain Gap)로 이어지며, 모델이 학습 환경에 과적합 되어 실제 환경에서 발생하는 다양한 물리적 변동성을 충분히 인식하지 못하는 편향으로 나타난다. 특히 조명 변화는 실제 서비스 환경에서 모델의 성능을 가장 크게 저하시킬 수 있는 대표적 요인이다. 영상이 지나치게 밝아지면 밝기 값이 상한 구간에 몰리면서 질감과 윤곽 정보가 소실되고, 반대로 조명이 부족한 경우에는 명암 분포가 하한 구간으로 수축하여 구조적 단서 표현이 크게 저하된다. 이러한 조도 왜곡은 모델이 의존하는 시각적 근거를 훼손하여 내부 표현을 불안정하게 만들고, 결과적으로 예측 신뢰도의 급격한 붕괴로 이어진다.

4) 기존 XAI 기법의 한계

기존 XAI 기법들은 모델이 어디(Where)를 주목했는지를 시각화하는 데는 성공했지만, 왜(Why) 그 영역을 주목했는지, 혹은 왜 실패했는지에 대한 인과적 설명(Causal Explanation)을 제공하지 못한다. 대부분의 기존 XAI는 수동적 관찰(Passive Observation) 수준에 머물러, 모델의 오류 원인을 규명하기보다는 단순히 사후적으로 결과를 보여주는 역할에 그친다. 예를 들어, 어두

3) Schwalbe, G., & Finzel, B. (2021). A Comprehensive Taxonomy for Explainable AI.

4) Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture.

5) Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations (ICLR 2019).

운 이미지에서 모델이 특징 추출에 실패하여 잘못된 예측을 한다고 할 때, 기존 XAI는 활성화 맵의 약화된 반응을 보여줄 수는 있다. 그러나 그 실패가 LAB 색상 공간의 L(Lightness) 값이 0 부근에 압축되어 정보가 소실되었기 때문인지, 혹은 색채 채널(A/B) 정보까지 함께 붕괴되었기 때문인지와 같은 근본적인 인과 구조(causal mechanism)까지는 밝혀내지 못한다.

또한 기존 XAI 기법들은 대체로 정적인 입력 이미지를 대상으로 분석을 수행한다. 사용자가 밝기, 색상, 채도, 질감, 주파수 등 물리적으로 해석 가능한 입력 변수를 조정하면서 모델의 반응 변화를 능동적으로 실험할 수 있는 기능은 거의 제공되지 않는다.

5) 기존 연구의 한계점

이로 인해 다음과 같은 세 가지 주요 한계가 존재한다. 첫째, 입력 요인과 모델 신뢰도 변화 간의 정량적 관계가 명확히 규명되지 않았다. 색상, 조명, 질감, 주파수 등의 미세한 변화가 예측 확률에 미치는 영향과 신뢰도 하락을 야기하는 주요 채널을 정량적으로 측정하고 비교할 수 있는 체계적 방법론이 부재한 실정이다. 둘째, 기존 XAI는 편향 원인 규명에서 모델 견고성(Robustness) 향상으로의 연결이 미흡하다. 현재 XAI 기법들은 모델의 주목 영역(attention)을 시각화하는 데 그칠 뿐, 모델 안정성 향상을 위한 구체적 조정 방안을 제시하는 능동적 디버깅(Causal Debugging) 관점은 제공하지 않는다. 셋째, 도메인 격차와 조명 편향에 대한 구조적 분석 체계가 미흡하다. 밝기, 색온도, 질감 왜곡 등 도메인 변형 요인이 색 공간 또는 주파수 분해 대역의 어떤 채널을 통해 신뢰도 저하로 전파되는지, 그리고 해당 채널에 대한 개입을 통해 신뢰도를 회복할 수 있는지에 관한 체계적 분석이 부재하다.

이러한 한계로 인해 기존 XAI는 모델의 내재적 편향 및 도메인 격차를 정량적으로 진단하고 교정하는 데 한계가 있다. 따라서 입력 속성에 대한 능동적 개입(Intervention)과 이에 따른 모델 반응의 인과적 측정을 통합하는 새로운 분석 프레임워크의 개발이 요구된다.

제 2 절 연구의 목적 및 범위

1) 연구의 목적

본 연구의 목적은 기존 설명 가능 인공지능(XAI)이 모델의 시각적 주목 영역을 사후적으로 제시하는 수준에 머물렀던 한계를 극복하고, 입력의 색상 및 주파수 성분 변화가 ResNet50 모델의 예측 신뢰도에 미치는 인과적 영향을 규명하는 데 있다. 이를 위해 본 연구는 ResNet50을 핵심 분석 대상 모델로 선정하였다. ResNet50은 ImageNet 벤치마크에서 입증된 표준 심층 신경망으로서의 대표성을 가질 뿐만 아니라, CNN 고유의 공간 위상 보존 특성을 계층 전반에 걸쳐 유지하는 아키텍처이다. 이는 입력 이미지의 2차원 격자 구조를 파괴하지 않으므로, 본 연구의 핵심 방법론인 웨이블릿 변환(Wavelet Transform)과 Score-CAM을 별도의 데이터 재배치(Reshape)나 왜곡 없이 구조적 특징을 적용할 수 있는 수학적 정합성을 제공한다. 따라서 ResNet50은 색상 및 주파수 개입에 따른 모델의 민감도를 분석하기 위한 실용성과 구조적 적합성을 동시에 충족하는 최적의 모델로 판단된다. 본 연구가 채택한 ResNet50과 최신 ViT 모델 간의 구조적 특성 및 방법론적 적합성 비교는 [표 1-1]과 같다.

[표 1-1] 주파수 인과 분석을 위한 ResNet50과 ViT의 구조적 적합성 비교

비교 기준	ResNet50 (제안 모델)	Vision Transformer (비교군)
구조적 적합성	최적 (공간 정보가 보존되어 활성화 맵 추출 및 Upsampling이 수학적으로 자연스러움)	복잡 (공간 재구성이 필요하며, 이 과정에서 인위적인 정보 왜곡 가능성 존재)
공간 위상 보존	보존됨(데이터가 $H \times W$ 격자(Grid) 형태 유지)	파괴됨 (1D 시퀀스 ($N \times D$)로 변환되어 공간성 희석
주파수 계층 구조	명확함 (Layer 깊이에 따라 고주파(Edge) →저주파 (Global) 로 정보가 일치함.)	모호함 (초기 Layer부터 전역 정보를 참조하므로 주파수별 역할 분담이 명확하지 않음.)
Wavelet 적용성	즉시 적용 (별도 전처리 없이 즉각적인 2D 주파수 분해 가능)	변환 필요 (Reshape($1D \rightarrow 2D$) 및 토큰 제거 등) 추가 공정 필수
Score-CAM 호환성	최적 (공간 정보가 주파수 특징과 일치하여 개입을 통한 시각적 패턴 변화 규명 용이)	복잡 (공간 맵 복원 시 핵심 CLS 토큰 배제로 인한 구조적 정보 손실 위험)
선정 결과	분석 방법론의 타당성 검증을 위한 구조적 최적 모델	추후 확장 연구 대상

위의 선정 근거를 바탕으로, 본 연구는 입력 이미지의 LAB/HSV 색상 채널과 웨이블릿 기반 주파수 대역을 체계적으로 조작하는 인과적 개입 실험을 수행하였다. 구체적으로는 각 개입이 모델의 예측 확률 및 신뢰도 변화량에 미치는 영향을 정량적으로 측정하고, Score-CAM 활성화 맵의 위상 변화를 통해 이를 정성적으로 분석하였다. 이러한 실험적 접근을 통해 ResNet50이 명도, 색차, 채도와 같은 색상 정보와 구조, 질감 같은 주파수 정보 중 어떤 요인에 의존하여 분류 결정을 수행하는지를 규명하였다.

2) 연구 범위

본 연구는 ImageNet으로 사전 학습된 ResNet50 모델을 분석 대상으로 선정하고, 각 분석 대상 클래스에 대한 초기 예측 확률을 기준선으로 설정한다. 이러한

설정을 통해 입력 데이터의 색상 및 주파수 성분이 모델의 예측 판단에 미치는 인과적 영향을 분석하며, 특히 조명 편향과 질감 손실이 예측 신뢰도 변화에 미치는 영향을 실험적으로 규명한다. 구체적으로는 LAB/HSV 색상 채널과 웨이블릿(Wavelet) 주파수 대역에 대한 체계적 개입(Intervention)을 수행하고, 이에 따른 예측 신뢰도 변화(ΔP)와 Score-CAM 활성화 맵의 변동을 정량화한다. 이를 통해 입력 특성과 모델 신뢰도 간의 인과 구조를 규명하는 것을 최종 목표로 한다.

본 연구 범위에는 새로운 CNN 아키텍처 제안, 성능(SOTA) 최적화, 모델 재학습 및 하이퍼파라미터 튜닝은 본 논문의 범위에 포함되지 않는다. 본 연구는 모델 구조를 변경하기보다, 입력 이미지의 물리적 속성을 조작하여 모델의 민감도(sensitivity)와 편향 구조(bias structure)를 규명하는 데 초점을 둔다. 또한, 본 연구에서 규명한 인과 구조를 바탕으로 한 자동 정규화 모듈 설계는 중요한 응용 방향이지만, 이는 본 논문의 직접적인 실험 범위를 넘어서는 과제로서 향후 연구(Future Work)로 제안한다.

[표 1-2] 연구 범위 요약

구분	포함되는 범위	포함되지 않는 범위	연구 수행 목적
대상 모델	사전 학습된 ResNet50	새로운 CNN 아키텍처 설계	기존 모델의 인과적 반응 분석
분석 항목	LAB/HSV 색상 개입, Wavelet 주파수 개입	모델 재학습, 하이퍼파라미터 최적화	입력 조작에 따른 신뢰도 변화 규명
설명 방식	Score-CAM 기반 정성·정량 인과 분석	단순 히트맵 시각화 중심 XAI	기존 XAI의 한계를 넘어 선 능동 실험
연구 목표	색상·주파수 인과 그래프 규명	SOTA 정확도 달성 제외	모델 민감성·편향의 원인 해석
향후 과제	인과 분석 기반 자동화 알고리즘 제안 가능성	자동 정상화 모듈 구현(본 논문 범위 아님)	후속 연구의 이론적 기반 제공

[표 1-3] 본 연구의 개입 요소(Intervention Factors)와 분석 방식

개입 요소	조작 변수	기대 반응	분석 목표
색상 채널	L, A, B, H, S, V	밝기·채도·색 대비 변화 → Score-CAM 집중 범위 변화	조명·색상 편향의 인과구조 규명
주파수	LL, LH, HL, HH	전역 구조·윤곽선·질감 성분 변동 → 예측 확률 변화	모델의 주파수 선택성 해석
색상-주파수 교차 분석	색상 조정 + 주파수 조정의 조합 실험	색상 인과 그래프와 주파수 인과 그래프 정량 변화 관측	색상-주파수 간 상호작용 이 모델의 반응 구조에 미치는 영향 정량화

제 3 절 본 연구의 기여

본 연구는 색상·주파수 개입을 기반으로 ResNet50의 신뢰도 변화를 인과적으로 분석할 수 있는 인과 분석 프레임워크를 제안하였으며, 그 핵심 기여는 다음 세 가지로 요약된다.

1) 색상 개입에 따른 모델 신뢰도 인과 분석 규명

조명 변화가 밝기·색채 구조를 왜곡하여 모델 신뢰도를 저하시킨다는 점을 색상 개입 실험을 통해 규명하였으며, 밝기(L) 및 색채(A/B, H/S/V) 조정이 신뢰도 회복과 직접적으로 연결된다는 인과적 메커니즘을 제시하였다.

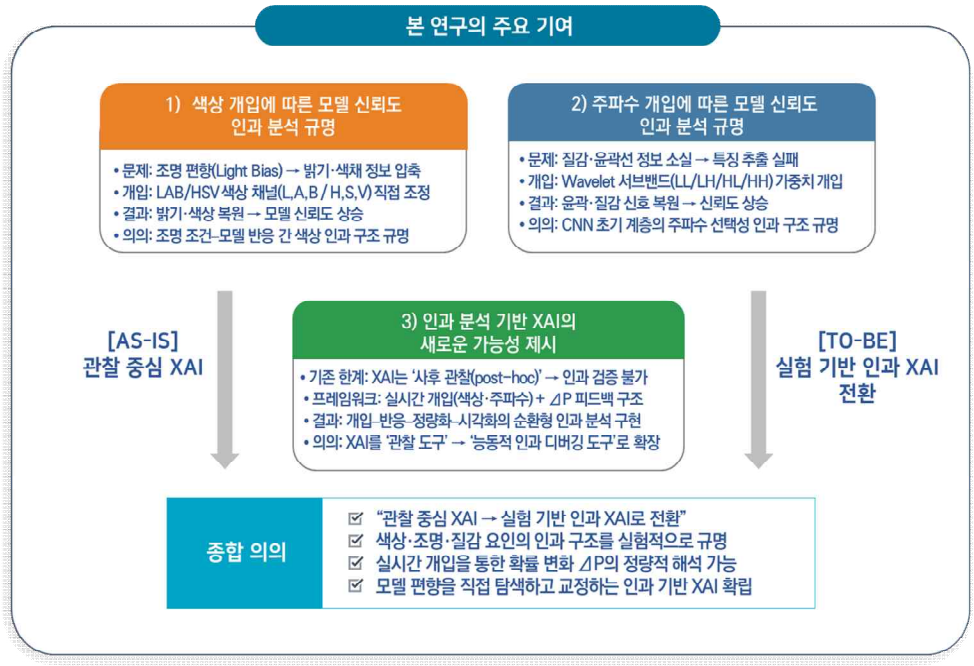
2) 주파수 개입이 모델 신뢰도에 미치는 인과 구조 규명

웨이블릿 서브밴드에 대한 개입 실험을 통해 ResNet50이 특정 주파수 성분에 의존하여 윤곽선·질감 정보를 인식함을 확인하였으며, 주파수 선택성이 신뢰도 변화에 미치는 영향을 정량적으로 제시하였다.

3) 인과 분석 기반 XAI의 새로운 가능성 제시

색상·주파수 개입을 즉각적으로 반영하고 예측 확률 변화 및 Score-CAM 응답을 실시간으로 관찰할 수 있는 인과 기반 XAI 프레임워크를 제안함으로써, 기

존 사후적 시각화 중심 XAI를 능동적 진단·교정이 가능한 구조로 확장하였다.



[그림 1-1] 본 연구의 주요 기여

마지막으로, 본 논문은 다음과 같은 구조로 구성된다. 2장에서는 이론적 배경 및 선행 연구를 정리하고, ResNet50, CAM 계열 기법, 색공간 이론, 웨이블릿 기반 주파수 분석, 그리고 저자의 선행 연구를 고찰한다. 3장에서는 본 논문에서 새롭게 제안하는 색상·주파수 개입 기반 인과 분석 프레임워크의 전체 아키텍처와 색상·주파수 개입 기법의 기술적 세부 사항을 상세히 서술한다. 4장에서는 색상·주파수 개입 기반 대화형 분석 도구를 이용하여 실제 개입을 수행하고, ResNet50의 반응을 정량적으로 측정함으로써 제안 프레임워크의 인과적 유효성을 실험적으로 검증한다. 마지막으로 5장에서는 본 연구의 핵심 성과를 정리하고, 인과 기반 XAI의 확장 가능성 및 향후 연구 방향을 제시하며 논문의 결론을 맺는다.

제 2 장 이론적 배경 및 선행 연구

본 장에서는 색상 및 주파수 조작을 기반으로 한 ResNet50 모델 반응 분석 체계를 구성하는 핵심 기술 요소와 이론적 기반을 체계적으로 고찰한다. 이러한 분석 체계는 색상 공간 변환, 웨이블릿 기반 주파수 분해, Score-CAM 기반 활성화 해석, 그리고 ResNet50의 계층적 특징 표현 구조에 근거하여 설계되었다.

딥러닝 기반 이미지 분류 모델은 심층적 특징 학습을 통해 높은 성능을 보여 왔으나, 내부 판단 구조와 입력 변화 사이의 인과적 관계를 직접 규명하기 어렵다는 한계를 지닌다. 이러한 한계를 해소하기 위해서는 모델의 특징 추출 방식, 내부 표현 시각화 기법, 그리고 입력의 물리적 특성(색상·주파수)이 모델 반응에 미치는 구조적 영향에 대한 다층적 이론 기반이 요구된다.

이에 따라 본 장은 다음과 같이 다섯 개 절로 구성된다. 첫째, 제1절에서는 ResNet50의 구조와 Residual Learning 메커니즘을 고찰하여 색상·주파수 개입의 인과 효과를 해석하기 위한 기반을 마련한다. 둘째, 제2절에서는 Class Activation Mapping의 발전 흐름(CAM→Grad-CAM→Grad-CAM++→Score-CAM)을 정리하고 각 기법의 알고리즘적 특징과 한계를 분석한다. 셋째, 제3절에서는 색 공간 이론을 고찰하여 밝기·색채 분리 방식, 조명 변화로 인한 정보 소실 메커니즘, 그리고 본 연구의 색상 개입 실험과의 연관성을 설명한다. 넷째, 제4절에서는 웨이블릿 변환의 다중 해상도 주파수 분석 원리를 다루며 CNN 초기 계층의 필터 반응과 주파수 대역 구조 간의 상관성을 제시한다. 마지막으로 제5절에서는 저자의 선행 연구를 검토하여 본 연구가 기존 웨이블릿-CAM 융합 방식에서 어떻게 확장되었는지, 그리고 인과적 관점을 도입하는 차별성이 무엇인지 분석한다.

특히 Grad-CAM++와 웨이블릿 기반 멀티스케일 활성화 맵을 결합한 선행 연구의 접근은 본 연구의 색상·주파수 개입 기반 인과 분석 프레임워크의 기술적 출발점이 된다. 이와 같이 제2장은 ResNet50의 구조적 이해 → CAM 계열의 설명 방식 → 색상·주파수 기반 특징 분해 이론 → 선행 연구 검토로 이어지는 논리적 흐름을 통해, 제3장과 제4장에서 제안하는 색상·주파

수 개입 기반 대화형 분석 도구를 이해하기 위한 이론적 기반을 제공한다.

제 1 절 ResNet50 모델 구조와 특징 추출 메커니즘

본 절에서는 본 연구의 대화형 인과 분석 프레임워크가 기반으로 활용하는 사전 학습(Pretrained)모델 ResNet50의 구조적 특징, 특징 추출 메커니즘, 그리고 ImageNet 기반 학습 배경을 심층적으로 고찰한다. ResNet50은 딥러닝 기반 컴퓨터 비전 모델 중 가장 널리 활용되는 백본(backbone) 구조이며, 특히 본 연구에서 관찰된 색상·조도·질감 개입(Intervention)에 따른 반응 특성 ΔP 변화, Score-CAM 변화, 주파수 민감도 변화를 이해하는 데 필수적인 기초 이론을 제공한다.

1) ResNet50의 개념 및 등장 배경

CNN은 이미지 분류·물체 인식·의료 영상 분석 등 광범위한 시각 인지 작업에서 뛰어난 성능을 보이며 딥러닝 분야의 핵심 모델로 자리 잡았다. 그러나 네트워크의 깊이가 증가함에 따라 학습이 불안정해지고, 기울기 소실(Vanishing Gradient) 문제가 발생하여 성능이 오히려 저하되는 Degradation Problem이 보고되었다. He et al.(2015)[6]은 이러한 문제를 해결하기 위해 Residual Learning(잔차 학습) 개념을 도입하였다. 이는 입력을 다음 층의 출력에 직접 더해주는 Skip Connection을 활용하여 네트워크가 변화량(Residual)만 학습하도록 하여 깊이가 50층 이상임에도 안정적인 학습을 가능하게 하였다. 특히 이러한 구조적 안정성은 이후 ResNet 계열 모델이 다양한 XAI 연구 Grad-CAM, Score-CAM, 주파수 기반 해석 기법 등 에서 표준 분석 모델로 자리 잡는 기반이 되었다.

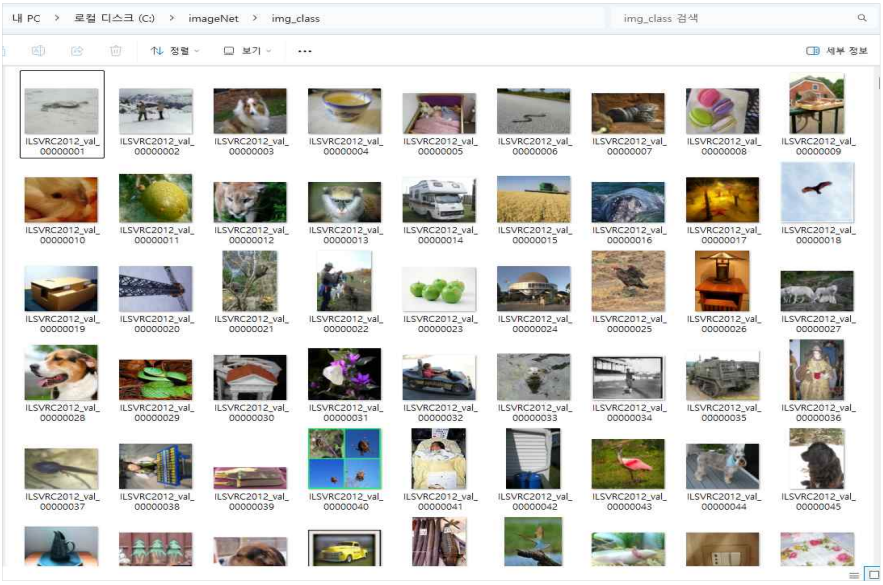
2) ImageNet기반 사전학습

6) He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), 770–778.

ResNet50의 강점은 단순히 깊은 신경망이라는 데 있지 않다. ResNet50은 ImageNet ILSVRC(ImageNet Large Scale Visual Recognition Challenge) 데이터셋으로 사전 학습되었으며, 이 데이터셋은 다음과 같은 특징을 가진다.

[표 2-1] ImageNet 데이터셋 구성

항목	내용
이미지 수	약 1.2M 장(120만 장)
클래스 수	1000개 Object Classes
이미지 크기	다양(고해상도 포함)
이미지 특징	다양한 조명, 질감, 배경, 촬영 환경 포함
학습 효과	저·중·고수준 시각 특징을 폭넓게 학습



[그림 2-1] ImageNet 데이터셋의 예시 이미지

3) ResNet50의 특징 추출 메커니즘

ResNet50은 입력 이미지(224×224×3 RGB)에 대해 계층적으로 심화되는 특징(hierarchical visual features)을 추출하도록 설계된 대표적 심층 합성곱 신경망이다. 네트워크는 초기 저수준 시각 단서에서부터 고수준 의미적 정

보에 이르기까지 단계적으로 시각 표현을 정교화하며, 이러한 특징 추출 과정은 본 연구의 색상·주파수 개입 실험이 모델 신뢰도 변화(ΔP)에 어떠한 방식으로 영향을 미치는지 설명하는 핵심 기반이 된다. 본 절에서는 ResNet50이 학습하는 특징[7]을 저수준–중간 수준–고수준의 세 단계로 구분하여 기술한다.

가) 저수준 특징(Low-level Features; Layer 1~10)

ResNet50 구조에서 Conv1 계층과 초기 잔차 블록은 입력 이미지의 기초적 시각 정보를 계층적으로 추출하는 역할을 담당한다. 이들 초기 계층이 감지하는 정보[8]는 크게 엣지 및 윤곽선, 선형 패턴, 방향성 특징으로 분류될 수 있으며, 이는 모두 저수준 시각 특징에 해당한다. 이러한 신경망 기반 특징 추출 메커니즘을 신호 처리의 관점에서 분석하면, 웨이블릿 변환의 고주파 분해 구조와 현저한 상응성을 발견할 수 있다. 즉, 신경망 필터가 수평, 수직, 대각선 방향의 세부 정보를 선택적으로 강조하는 방식은 웨이블릿의 LH, HL, HH 서브밴드가 각각 해당 방향의 고주파 성분을 분리하는 원리와 기능적으로 유사하다고 해석할 수 있다. 즉, ResNet50의 초기 합성곱 필터들은 사실상 웨이블릿 기반 에지 검출 함수와 유사한 주파수 응답 구조[9]를 가진다. 따라서 본 연구에서 수행한 LH·HL·HH 주파수 개입(Boost/Reduce)이모델의 예측 확률 변화(ΔP)에 매우 민감하게 반응한 이유는, ResNet50의 초기 레이어가 고주파수 대역 정보를 핵심 입력 신호로 사용하고 있기 때문이다.

나) 중간 수준 특징(Mid-level Features; Layer 10~35)

네트워크가 깊어질수록 모델은 단순한 엣지나 색상 변화를 넘어 더욱 구

7) Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. Proceedings of the European Conference on Computer Vision (ECCV 2014), 818–833.

8) Olah et al. (2017). Feature Visualization.

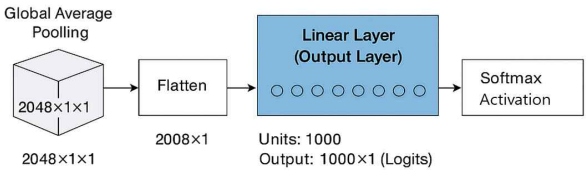
9) Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation.

조적이고 반복적인 시각 패턴을 추출하게 된다. 중간 수준 특징의 대표적 요소는 질감 및 반복 패턴, 재질 특성, 사물의 중간 형태이며, 이는 목재 표면의 결, 물결 패턴, 옷의 섬유 조직, 나뭇잎과 잔디의 반복 배열과 같이 저수준 신호가 결합하여 형성되는 복합적 구조를 의미한다. 본 연구에서 색상 채널(A/B, H/S/V) 개입이 미세 패턴 인식 능력에 영향을 미친 것은 중간 수준 계층이 질감과 색차의 안정성에 크게 의존하기 때문이다. 특히 어두운 조명과 과다 노출 상황에서 색채 채널 정보가 압축되면 이러한 중간 수준 특징이 손실되어 모델 신뢰도가 하락하게 된다.

나) 고수준 특징(High-level Features; Layer 35~50)

ResNet50의 마지막 잔차 블록 및 전역 평균 풀링 직전 계층은 객체 중심의 시각 표현을 학습하는 단계로, 여기에서는 물체의 구체적 형태, 객체의 의미적 부분, 물체 간 구분을 위한 고차원 특징과 같은 의미적 특징이 형성된다. 이 단계에서 추출된 특징은 최종 완전 연결 계층의 1000개 클래스 로짓 입력이 되어, 네트워크가 입력 이미지를 ImageNet의 1000개 객체 범주 중 하나로 분류할 수 있도록 한다. 고수준 특징은 저수준 및 중간 수준 특징의 누적적 조합 위에서 형성되므로, 저수준 색상 정보 또는 고주파 질감 정보가 개입을 통해 보정되면 그 효과가 상위 계층으로 전파되어 고수준 의미적 패턴 역시 회복되며 모델 신뢰도가 상승하게 된다. 이러한 계층적 인과 전파 메커니즘은 본 연구에서 관찰된 색상 인과 그래프 및 주파수 인과 그래프의 신뢰도 상승 현상을 이론적으로 설명해 준다.

4) ResNet50의 마지막 출력 구조



[그림 2-2] ResNet50 최종 출력 구조도

[그림 2-2] ResNet50의 최종 출력 단계(End-to-End Output Pipeline)를 개념적으로 나타낸 구조도이다. 본 그림은 ResNet50이 입력 이미지로부터 최종적으로 1000개 클래스에 대한 예측 확률을 생성하는 전 과정을 직관적으로 시각화한 것이다. 먼저, 네트워크의 마지막 합성곱 계층을 통과한 2048개의 특징 맵들은 Global Average Pooling을 통해 각 채널이 대표하는 공간적 정보를 요약한 2048차원 특징 벡터(2048×1×1)로 축약된다. 이 벡터는 Flatten 과정을 통해 1×2048 형태의 1차원 벡터로 변환된 후, 마지막 Fully Connected Layer에 입력되어 1000개의 logit(선형 활성값)을 산출한다. 이 logit들은 이미지가 각 클래스에 속할 가능성을 나타내는 정규화되지 않은 점수이며, Softmax 함수를 통해 1000개 클래스에 대한 확률 분포로 변환되어 최종 예측 결과를 형성한다. 즉, ResNet50은 GAP → Flatten → Linear (1000 logits) → Softmax의 단계를 거쳐 입력 이미지를 ImageNet 1000개 객체 범주 중 하나로 분류하게 된다.

제 2 절 Class Activation Mapping 계열의 발전적 계보

1) CAM (Class Activation Mapping)

가) 개념 및 등장 배경

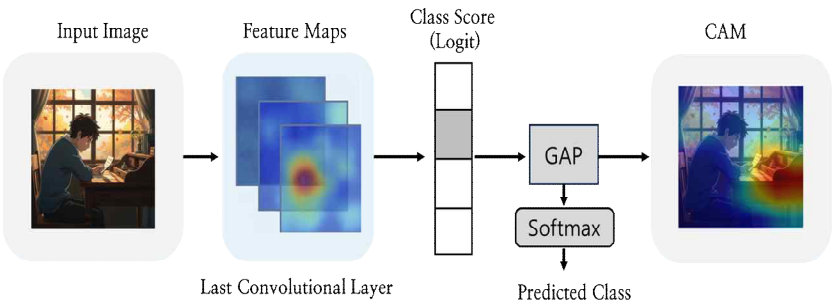
Class Activation Mapping은 신경망이 특정 클래스를 결정하는 과정에서 활용한 이미지의 핵심 영역을 활성화 맵으로 시각화하는 기법이다. 이는 딥러닝 모델의 의사결정 근거를 직관적으로 파악할 수 있도록 하는 초기 설명가능 인공지능(XAI) 접근법이기도 하다. 당시의 합성곱 신경망은 높은 분류 정확도를 달성했음에도 불구하고, 모델이 어떤 시각적 근거(visual evidence)를 기반으로 결론을 도출하는지 확인하기 어려운 블랙박스(black-box) 문제를 갖고 있었다. 이러한 문제의식 속에서, CNN의 마지막 합성곱 계층(last convolutional layer)에서 얻어지는 피처맵(feature map)이 클래스별 판별 정

보(discriminative information)를 포함하고 있다는 점이 주목받기 시작했다.

CAM은 이 피처맵을 활용하여 클래스별 가중합(class-specific weighted sum)을 계산함으로써, 네트워크가 최종 예측을 수행할 때 어떤 영상 영역을 판별 근거로 사용했는지를 열지도 형태로 재구성한다. 다시 말해, CAM은 모델이 무엇을 예측했는가를 넘어 모델이 어디를 보고 판단했는가를 명확히 드러낸 최초의 구조적 설명 기법으로서 중요한 학문적 의의를 지닌다.[10)]

나) 작동 원리

CAM의 기본 구조는 Global Average Pooling을 활용한 CNN 아키텍처에서 정의된다. 모델의 마지막 합성곱 계층에서 얻어진 각 채널의 피처맵을 $A_k(x,y)$라 하면, GAP 레이어를 통해 해당 피처맵의 공간적 평균을 취해 다음과 같은 형태로 표현된다.



[그림 2-3] CAM 기반 활성화 맵 생성과정

CNN은 입력 이미지를 여러 컨볼루션 레이어를 거치며 다양한 특징(예: 질감, 형태, 색상 등)을 학습하고, 그 결과를 피쳐 맵으로 생성한다. CAM은 이 중 가장 고차원적인 의미 정보를 담고 있는 마지막 컨볼루션 레이어에서 생성된k개의 피쳐 맵A_k(각각 $H \times W$ 크기)를 활용한다. 각 피쳐 맵은 특정 시각적 패턴이 이미지의 어느 위치에서 활성화되는지를 나타낸다.

추출된 각 피쳐 맵 A_k는 GAP 레이어를 거쳐 하나의 스칼라 값 F_k로 압축된다. 이는 각 피쳐 맵이 가진 공간적인(spatial) 정보를 하나의 대표값으로 요약하

10) Zhou et al.(2016). Learning deep features for discriminative localization.

는 과정이다.

$$F_k = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} A_k(i,j)$$

- $A_k(i,j)$: k번째 피쳐 맵의 (i,j) 위치에 있는 픽셀(활성화) 값
- F_k: k번째 피쳐 맵의 모든 활성화 값들의 평균으로, 해당 피쳐 맵이 감지한 특징이 이미지 전체에 걸쳐 얼마나 '존재'하는지를 나타내는 단일 값 GAP를 통해 얻은 k개의 스칼라 값 $F_1, F_2, ..., F_k$는 최종 분류를 위한 FC 레이어의 입력으로 사용된다. 특정 클래스 c에 대한 최종 점수(logit) S_c는 각 F_k에 해당 클래스에 대한 가중치 w_k^c를 곱한 후 모두 더하여 계산된다.

$$S_c = \sum_k w_k^c F_k$$

- w_k^c: 매우 중요한 가중치로, k번째 피쳐 맵(F_k로 요약된)이 클래스 c를 예측하는 데 얼마나 중요한지를 나타내는 값 클래스 c에 대한 CAM (M_c)은 마지막 단계에서 사용된 중요도 가중치 w_k^c를 원래의 공간 정보를 보존하고 있는 피쳐 맵 A_k에 다시 곱하여 선형 결합(weighted sum) 함으로써 생성된다.

$$M_c(i,j) = \sum_k w_k^c A_k(i,j)$$

여기서,

- w_k^c: 클래스 c를 예측할 때, k번째 feature map이 갖는 중요도(가중치)
- $A_k(i,j)$: k번째 피쳐 맵이 (i,j) 위치의 활성화 값 클래스 c에 중요한 피쳐 맵(가중치 w_k^c가 큰 것)이 강하게 반응하는 공간 위치를 강조하여 합친 것이 CAM이다. 이렇게 얻어진 M_c는 클래스 c의 예측에 결정적인 역할을 한 이미지 내 영역을 나타내는 히트맵이 되고, 일반적으로 원본 이미지 크기로 업샘플링 하여 시각화

다) 장점 및 한계점

장점은 모델의 판단 근거를 직관적으로 시각화하여 해석 가능성을 제공하고,

별도의 역전파(backpropagation) 계산 없이 단순한 순전파(forward pass) 연산만으로 생성이 가능하다는 점이다.

그러나 CAM은 다음과 같은 명확한 한계점을 갖는다.

(1)모델 구조의 제약: 가장 큰 단점으로, GAP 계층을 사용하는 특정 네트워크 아키텍처에서만 적용이 가능하다. VGGNet과 같이 끝단에 여러 개의 완전 연결 계층이 있는 모델이나 Vision Transformer와 같은 최신 구조에는 직접 적용할 수 없어 범용성이 떨어진다. 이를 적용하기 위해서는 모델 구조 변경 및 재학습이 요구되며, 이 과정에서 성능 저하가 발생할 수 있다.

(2)표현의 한계: GAP 연산은 특징 맵의 공간적 정보를 평균화하므로, 이미지 내에 여러 객체가 존재하거나 미세한 특징을 기반으로 판단하는 경우, 활성화 영역을 정확하게 구분하지 못하고 전반적인 영역을 모호하게 표현하는 경향이 있다.

2) 그래디언트 가중 클래스 활성화 맵 Grad-CAM

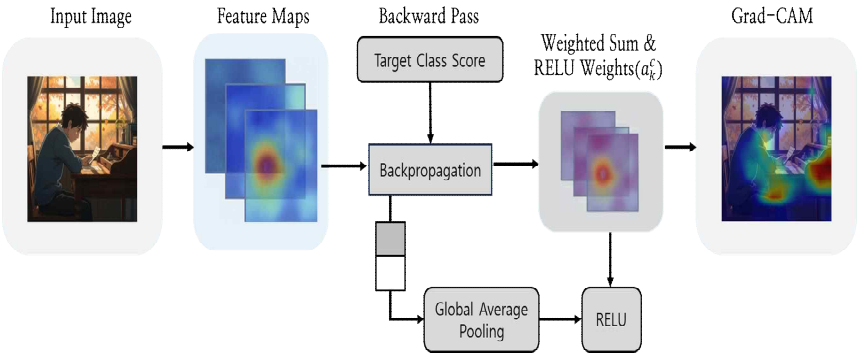
가) 개념 및 등장 배경

CAM은 CNN 모델의 해석 가능성을 제공하는 중요한 방법론이지만, 마지막 컨볼루션 계층 이후에 GAP 계층이 필수적이라는 강력한 구조적 제약을 가진다. 이러한 한계로 인해 VGGNet, ResNet의 일부 변형 모델, Vision Transformer와 같은 다양한 아키텍처에 적용이 불가능하였다.

이러한 문제를 해결하기 위해 Grad-CAM이 제안되었다. Grad-CAM은 모델의 구조를 변경하거나 재학습할 필요 없이, 그래디언트(gradient) 정보를 활용하여 CAM과 유사한 시각화를 생성하는 보다 일반화된 접근법이다. 이 기법은 특정 클래스 예측에 대한 마지막 컨볼루션 계층의 출력 그래디언트를 역전파 여 특징 맵의 중요도를 계산함으로써, 사실상 모든 CNN 기반 아키텍처에 적용 가능한 범용성을 확보하였다.

나) 작동 원리

Grad-CAM의 핵심 아이디어는 FC 계층의 가중치(w_k^c)를 사용하는 대신, 특정 클래스에 대한 예측 점수에 큰 영향을 미치는 활성화 맵을 그래디언트를 통해 식별하는 것이다.



[그림 2-4] Grad-CAM 기반 활성화 맵 생성과정

(1) 순전파 및 예측: 입력 이미지를 CNN 모델에 통과시켜 특정 클래스 c에 대한 예측 점수(소프트맥스 이전의 로짓) y^c를 계산한다.

(2) 그래디언트 계산: 예측 점수y^c를 모델의 마지막 컨볼루션 계층에서 출력된 k번째 특징 맵A_k의 각 픽셀$A_k(i,j)$에 대해 미분하여 그래디언트$\frac{\partial y^c}{\partial A_k(i,j)}$를 구한다. 이 그래디언트는 각 픽셀의 변화가 최종 점수 y^c에 얼마나 큰 영향을 미치는지를 나타낸다.

(3) 특징 맵 중요도 가중치 계산: 계산된 그래디언트 맵 전체에 대해 전역 평균 풀링을 적용하여 k번째 특징 맵에 대한 중요도 가중치 a_k^c를 계산한다.

$$a_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_k(i,j)}$$

여기서 a_k^c는 k번째 특징 맵이 클래스 c의 예측에 미치는 전반적인 중요도를 의미하며, CAM의 가중치 w_k^c와 유사한 역할을 수행한다.

(4) Grad-CAM 생성: CAM과 마찬가지로, 계산된 가중치 a_k^c와 특징 맵 A_k를 선형 결합하여 최종 히트맵을 생성한다. 다만, 그래디언트에는 음수 값이 포함될 수 있으므로, 최종적으로 ReLU(Rectified Linear Unit) 활성화 함수를 적용하

여 양의 영향을 미치는 특징들만 시각화에 사용한다.

$$L^{c}_{Grad-CAM} = ReLU(\sum_{k} \alpha^{c}_{k} A_{k})$$

이 과정을 통해 생성된 히트맵은 모델이 클래스 c를 예측하기 위해 주목한 이미지 내의 영역을 나타낸다.

[표 2-2] CAM과 Grad-CAM의 가중치 계산 방식 비교

구분	CAM	Grad-CAM
적용 가능 구조	GAP(Global Average Pooling) 구조를 포함한 CNN에 한정	GAP 유무와 관계없이 다양한 CNN 구조에 적용 가능
가중치 계산식	$S_c = \sum_k w^c_k F_k$	$\alpha^c_k = \frac{1}{Z}\sum_i \sum_i \frac{\partial y^c}{\partial A^k_{ij}}$
핵심 개념	GAP-FC 구조의 학습된 가중치를 특성 맵 기여도로 활용	클래스별 그래디언트를 통해 채널의 기여도 산출
출력 맵 계산식	$M_c(i,j) = \sum_k w^c_k A_k(i,j)$	$L^{c}_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A_k)$
해석 범위	클래스별 국소적 활성 패턴 (Local activation area)	클래스별 국소적 활성 패턴 (Local activation pattern)

다) 장점 및 의의

(1) 범용성

Grad-CAM은 기존의 클래스 활성화 맵 방식과 달리 글로벌 평균 풀링 계층의 존재를 전제로 하지 않는다. 따라서 ResNet, VGGNet과 같은 전통적인 CNN 아키텍처뿐 아니라, 이미지 캡셔닝(image captioning)이나 시각적 질의응답(Visual Question Answering, VQA)과 같이 CNN 모듈을 포함한 복합 신경망 모델에도 직접 적용 가능하다. 이는 Grad-CAM이 다양한 응용 분야에서 모델 해석 가능성을 확장시킨다는 점에서 학문적·실용적 의의가 있다.

(2) 고해상도 시각화

기존 CAM 대비 더 세밀하고 객체의 특정 부위에 집중된 시각화 결과를 산출하는 경향이 있다. 이는 모델의 내적 판단 근거를 고품질의 시각적 단서로 제시함으로써, 인간 연구자가 인지하기 용이한 형태로 해석 가능성을 높여준다. 따라서 Grad-CAM은 단순히 모델이 특정 객체의 존재를 인식했는지를 확인하는 수준을 넘어, 모델이 입력 이미지 내에서 어떠한 국소적 영역

(local region)과 구조적 특징(structural feature)에 주목하여 판별을 수행하는지를 정량적 및 정성적으로 규명할 수 있는 설명 가능성 도구(Explainability Tool)로서 학문적 의의를 지닌다. 이러한 특성은 모델의 판단 근거를 시각적 형태로 검증할 수 있게 함으로써, 딥러닝 기반 영상 분석의 해석 가능성을 실질적으로 확장시킨다.

(3) 잘못된 예측의 분석 및 디버깅

Grad-CAM은 모델이 잘못된 예측을 내린 경우에도 판단 근거로 활용된 이미지 영역을 직관적으로 확인할 수 있게 한다. 이는 모델이 과적합되었는지, 특정 배경 특성에 의존했는지, 혹은 관심 대상 외의 패턴을 학습했는지를 평가하는 데 유용하다. 다시 말해, Grad-CAM은 모델의 취약점 진단 및 개선 과정(디버깅)에 기여할 수 있는 실질적 도구로 기능한다.

라) 한계점

(1) 그래디언트 포화 문제

모델이 특정 클래스에 대해 지나치게 확신하는 경우, 역전파 과정에서 그래디언트 값이 0에 가까워질 수 있다. 이로 인해 시각화된 활성화 맵의 해상도가 떨어지고, 결과적으로 모델의 주의 영역을 충분히 반영하지 못할 가능성이 있다.

(2) 불완전한 객체 표현

Grad-CAM은 입력 영상 내에서 분류 결과에 기여도가 높은 부분을 강조하는 경향이 있다. 따라서 대상 객체의 일부 두드러진 영역만 시각화되는 경우가 많아, 객체의 전체적인 형태나 구조를 충분히 포착하지 못할 수 있다. 이는 모델 해석의 완전성 측면에서 제한 요소로 작용할 수 있다.

3) Grad-CAM의 일반화된 확장 모델 Grad-CAM++

가) 개념 및 등장 배경

딥러닝 모델은 높은 예측 성능에도 불구하고, 내부의 의사결정 과정을 이

해하기 어렵다는 이른바 블랙박스 문제를 지닌다. 이러한 한계는 모델의 신뢰성과 설명 가능성 확보를 어렵게 만들기 때문에, 최근에는 모델의 예측 근거를 시각적으로 제시하려는 다양한 연구가 활발히 진행되고 있다. Grad-CAM은 대표적인 시각화 기법으로, 모델의 구조적 변형 없이 특정 클래스에 대한 예측에 기여한 이미지 영역을 시각적으로 강조할 수 있다는 장점이 있다. Grad-CAM은 합성곱 신경망(CNN)의 최종 합성곱 계층에서 클래스 점수에 대한 그래디언트를 활용하여 클래스 활성화 맵을 생성하며, 다양한 응용 분야에서 널리 활용되고 있다.

그러나 Grad-CAM에는 다음과 같은 근본적인 한계가 존재한다.
첫째, 하나의 이미지 안에 동일한 클래스에 속하는 객체가 여러 개 존재하는 경우, Grad-CAM은 전체 객체를 고르게 강조하지 못하고 일부 영역만 선택적으로 강조하는 경향이 있다. 둘째, 객체 전체의 형태보다는 모델이 분류에 있어 가장 판별력이 크다고 판단한(discriminative) 국소 영역에 집중하는 경향이 강하다. 이러한 한계는 Grad-CAM이 특징 맵의 모든 위치에서 계산된 그래디언트를 단순 평균하여 중요도를 산출하기 때문이다. 즉, 각 픽셀별 기여도가 정교하게 반영되지 못하고, 위치 정보가 소실되는 문제가 발생한다.

이 문제를 해결하기 위해 기존 Grad-CAM을 수학적으로 일반화하여 픽셀 수준의 중요도를 반영할 수 있는 Grad-CAM++ 기법을 제안하였다.

나) Grad-CAM++의 기본 개념

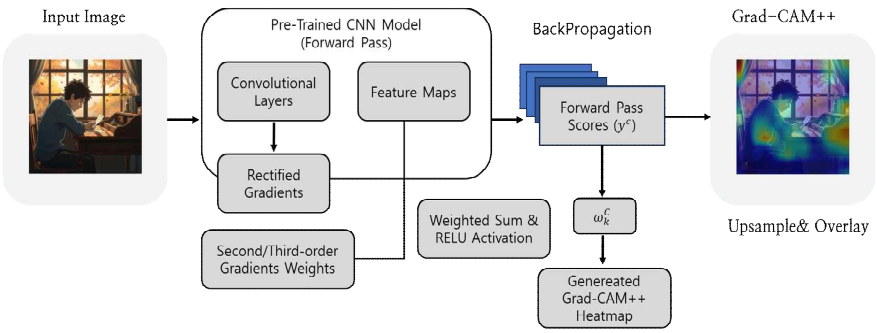
Grad-CAM++[11]는 기존 Grad-CAM의 가중치 계산 방식을 픽셀 수준으로 세분화하여, 보다 정밀한 클래스 활성화 맵을 생성할 수 있도록 고안되었다. 기존 Grad-CAM이 채널별 그래디언트를 단순 평균하여 가중치를 산출한 것과 달리, Grad-CAM++는 픽셀별 중요도를 반영하는 pixel-wise weighted gradients 방식을 도입하였다.

이를 통해 Grad-CAM++는 다음과 같은 장점을 갖는다.

(1) 단일 객체의 경우, 객체의 전체 윤곽을 더욱 정확하게 강조할 수 있다.

11) Chattopadhyay et al.(2018). Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks.

(2) 동일 클래스에 속하는 복수 객체가 존재하는 경우에도 각각의 객체를 효과적으로 구분할 수 있다. 이러한 접근 방식은 기존 Grad-CAM의 계산 과정을 수학적으로 일반화한 결과로, 단순한 평균 연산 대신 위치 가중치 기반의 합산을 통해 세밀한 시각화를 가능하게 한다.



[그림 2-5] Grad-CAM++ 기반 활성화 맵 생성과정

Grad-CAM++의 클래스c에 대한 히트맵은 다음과 같이 정의된다.

$$L^c_{Grad-CAM++} = ReLU(\sum_k a_k^c A_k)$$

Grad-CAM++의 핵심 아이디어는 특징 맵(A_k)의 각 활성화 값($A_k(i,j)$)이 최종 여기서 $A_k(i,j)$는 마지막 합성곱 계층의 k번째 특징 맵의 (i,j)위치에서의 활성화 값이며, ω_k^c는 해당 채널의 중요도를 나타내는 가중치이다.

Grad-CAM과의 가장 큰 차이점은 이ω_k^c의 정의 방식에 있다. 기존 Grad-CAM에서는 ω_k^c가 해당 채널의 그래디언트의 평균으로 계산되는 반면, Grad-CAM++에서는 각 픽셀의 중요도를 반영하여 아래와 같이 정의한다.

$$\omega_k^c = \sum_i \sum_j a_{ij}^{kc}(i,j) \cdot ReLU\left(\frac{\partial y^c}{\partial A_k(i,j)}\right)$$

여기서 a_{ij}^{kc}는 픽셀(i,j)에서의 중요도를 나타내는 가중치이며, 이는 2차 및 3차 미분을 이용해 다음과 같이 계산된다.

$$a_{ij}^{kc} = \frac{\frac{\partial^2 y^c}{\partial A_k(i,j)^2}}{2\frac{\partial^2 y^c}{\partial A_k(i,j)^2} + \sum_a \sum_b A_k(a,b)\frac{\partial^3 y^c}{\partial A_k(i,j)^3}}$$

[표 2-3] Grad-CAM과 Grad-CAM++ 가중치 계산방식 차이

구분	Grad-CAM	Grad-CAM++
적용 가능 구조	일반 CNN 및 GAP 구조 유무와 관계없이 일반적인 CNN에 적용 가능	Grad-CAM과 동일 (모든 CNN 구조에 적용 가능)
가중치 계산식	$\alpha_k^c = \frac{1}{Z}\sum_i\sum_i \frac{\partial y^c}{\partial A_{ij}^k}$	$\alpha_k^c = \frac{1}{Z}\sum_i\sum_i \frac{\partial y^c}{\partial A_{ij}^k}$ /(2차,3차편미분)
핵심 개념	1차 그래디언트(First-order gradient)를 통한 가중치 계산	2차 및 3차 편미분(High-order derivative)을 통한 픽셀별 중요도 반영
계산 방식	단일 그래디언트 기반 역전파	고차 미분 기반 역전파 (2차, 3차 포함)
출력 맵 계산식	$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A_k)$	$L_{Grad-CAM++}^c = ReLU(\sum_k \alpha_k^c A_k)$
해석 범위	국소적 클래스 주목 영역 (Local class attention)	세부 영역별 정밀한 클래스 주목 패턴(Fine-grained attention)

다) Grad-CAM++의 시각적 효과 및 응용

Grad-CAM++는 시각화 품질 측면에서 기존 Grad-CAM보다 뛰어난 성능을 보인다. 단일 객체의 경우 Grad-CAM++는 객체의 외곽선뿐 아니라 내부 영역까지 고르게 강조할 수 있으며, 다중 객체가 존재하는 이미지에서도 각각의 객체를 명확하게 구분해 낼 수 있다. 특히 픽셀 수준의 세밀한 시각화는 모델의 오류 진단과 설명 가능 인공지능(XAI) 분야에서 중요한 역할을 한다.

라) Grad-CAM++의 한계점

Grad-CAM++은 기존 Grad-CAM의 단점을 보완하기 위해 제안된 기법으로, 클래스별 활성화 맵을 산출할 때 다중 위치 픽셀의 기여도를 세분화하여 시각적 정확도를 높인다. 특히, 1차 및 2차 편미분 항을 활용함으로써 동일한 피처맵 내에서도 픽셀 단위의 중요도 차이(importance difference)를 정량화할 수 있다. 그러나 이러한 고차 미분 기반 접근(Second-order Gradient Dependency)은 본질적으로 다음과 같은 한계를 내포한다. 첫째, 그래디언트의 수치적 불안정성으로 인해, 소음(Noise)이나 클래스 불균형(Class Imbalance)이 존재하는 상황에서는 활성화 맵이 왜곡되기 쉽다. 둘째, 모델의

예측 신뢰도보다는 미분의 기울기 분포에 과도하게 민감하여, 실제 예측 근거와 무관한 계산적 인공물(Computational Artifact)을 생성할 수 있다. 셋째, 다중 클래스(Class) 또는 다중 객체(Multi-object) 상황에서의 비선형 상호작용을 충분히 반영하지 못하며, 이는 활성 영역이 중첩되거나 희석되는 문제로 이어진다.

결과적으로 Grad-CAM++은 정확도(precision) 측면에서는 향상되었으나, 안정성(stability)과 설명 일관성(interpretability consistency) 측면에서는 여전히 구조적 제약을 지닌다.

4) 그래디언트 독립적인 시각적 설명 모델 Score-CAM

가) 개념 및 등장 배경

Score-CAM은 그래디언트(gradient)에 대한 의존성을 제거한 최초의 CAM 기반 기법으로, Grad-CAM의 한계를 보완하기 위해 제안되었다.
초기의 CAM은 마지막 합성곱 레이어의 활성화 맵과 Global Average Pooling 가중치를 결합하여 시각화를 수행했지만, GAP 레이어가 포함된 특정 아키텍처에만 적용 가능한 구조적 제약이 있었다. 이후 Grad-CAM은 예측 클래스에 대한 그래디언트를 역전파하여 각 활성화 맵의 중요도를 계산함으로써 이러한 제약을 완화했으나, 그래디언트 소실(vanishing gradient)이나 불안정성으로 인해 히트맵의 신뢰성이 저하되는 문제가 있었다. Score-CAM은 이러한 문제를 해결하기 위해 각 활성화 맵을 입력 이미지에 마스크로 적용한 후, 해당 입력의 순방향 전파(forward pass) 결과로 얻은 클래스 점수(class score)를 중요도로 활용한다.

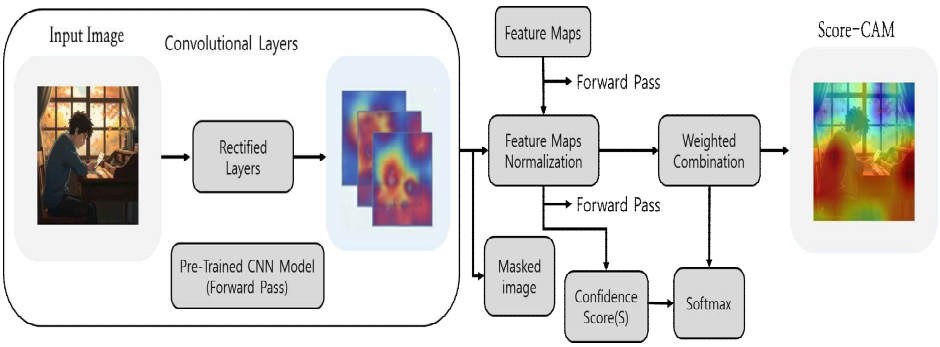
이러한 그래디언트-프리(gradient-free) 접근은 모델의 판단 근거를 보다 충실하게 반영하며, 시각적 일관성과 설명력을 동시에 향상시킨다.

나) Score-CAM의 기본 개념

Score-CAM(Score-weighted Class Activation Mapping)은 CNN모델의 예측 결과를 시각적으로 설명하는 XAI(Explai nable AI) 기법 중 하나이다.

기존의 CAM 기반 방법론, 특히 Grad-CAM이 특정 클래스에 대한 예측의 Gradient를 활용하여 중요 영역을 시각화하는 것과 달리, Score-CAM은 Gradient에 대한 의존성을 완전히 제거한 것이 가장 큰 특징이다. 이 방법론은 각 채널의 활성화 맵(Activation Map)이 모델의 예측에 얼마나 큰 신뢰도 증가(Increase of Confidence)를 가져오는지를 측정하여 가중치를 산출한다.

즉, 특정 활성화 맵을 마스크(Mask)로 사용하여 원본 이미지에 적용했을 때, 모델이 해당 클래스에 대한 예측 점수(Score)를 얼마나 높게 부여하는지를 정량적으로 평가한다. 이 점수를 가중치로 사용하여 활성화 맵을 선형 결합함으로써, 모델이 이미지의 어느 부분을 근거로 판단을 내렸는지 직관적인 히트맵(Heatmap) 형태로 시각화한다. Gradient를 사용하지 않으므로 Gradient 소실(Vanishing Gradients)이나 포화(Saturation) 문제로부터 자유로우며, 이는 결과적으로 더 깨끗하고 객체에 집중된 고품질의 시각화 결과를 생성하는 장점으로 이어진다.[12)]



[그림 2-6] Score-CAM 기반 활성화 맵 생성과정

Score-CAM은 Grad-CAM 계열의 가장 근본적인 제약인 그래디언트 의존성(Gradient Dependency)을 제거하기 위해 제안되었다. 이 기법은 특정 클래스 c에 대한 입력 이미지 I의 활성화 맵 A_k를 이용하여, 해당 맵을 입력에

12) Wang et al.(2020). Score-Weighted Visual Explanations for Convolutional Neural Networks.

주입하고 모델의 출력 스코어 변화를 측정함으로써 가중치를 계산한다.

Score-CAM의 핵심 아이디어는 다음과 같다.

$$\omega_k^c = f_c(I \odot Up(A_k)) - f_c(I_0)$$

여기서 $f_c(\cdot)$는 클래스 c에 대한 모델의 출력 점수(score function), Up(·)은 업샘플링(up-sampling) 연산, I_0는 기준 입력(보통 제로 이미지)이다. 즉, k번째 특성 맵 A_k를 업샘플링하여 원본 이미지 I에 마스크(mask)로 적용한 이미지를 모델에 다시 순방향으로 입력했을 때, 클래스 점수(f_c)가 (베이스라인 $f_c(I_0)$대비) 얼마나 증가하는지를 측정한다. 이 점수 증가량 자체가 A_k의 실제 기여도(contribution)가 되는 그래디언트-프리(gradient-free) 방식이다.

이후 활성화 맵은 다음과 같이 결합된다.

$$L_{Score-CAM}^c = ReLU(\sum_k \omega_k^c A_k)$$

이 식은 Grad-CAM에서의 그래디언트 기반 가중치 계산식과 구조적으로 유사하지만, 미분 연산 없이 순전파(forward pass)만으로 가중치를 산출한다는 점에서 본질적인 차이를 가진다. 이로써 Score-CAM은 그래디언트의 불안정성 문제를 해소하고, 시각화 결과의 일관성과 재현성을 확보한다.

[표 2-4] Grad-CAM++과 Score-CAM의 가중치 계산 방식 비교

구분	Grad-CAM++	Score-CAM
적용 가능 구조	고차 미분이 가능한 CNN 구조	모든 CNN 구조에 적용 가능 (Gradient-free)
가중치 계산식	$\alpha_k^c = \frac{1}{Z}\sum_i\sum_i \frac{\partial y^c}{\partial A_{ij}^k}$ / (2차,3차편미분)	$\omega_k^c = f_c(I \odot Up(A_k)) - f_c(I_0)$
핵심 개념	2차 및 3차 편미분(High-order derivative)을 통한 픽셀별 중요도 반영	그래디언트 없이 클래스 점수(score)의 변화량으로 기여도 산출
계산 방식	고차 미분 기반 역전파 (2차, 3차 포함)	낮음 (Score 기반 안정화)
출력 맵 계산식	$L_{Grad-CAM++}^c = ReLU(\sum_k \alpha_k^c A_k)$	$L_{Score-CAM}^c = ReLU(\sum_k \omega_k^c A_k)$
해석 범위	세부 영역별 정밀한 클래스 주목 패턴(Fine-grained attention)	높음 (각 채널별 forward 반복)

라) Score-CAM의 시각적 효과 및 응용

Score-CAM은 그래디언트에 의존하지 않는 그래디언트-프리(gradient-free) 접근 방식을 채택함으로써 기존의 그래디언트 기반 설명 기법과 뚜렷이 구별되는 시각적 특성을 제공한다. 우선, Score-CAM이 생성하는 활성화 맵은 시각적 선명도(visual cleanliness) 측면에서 높은 품질을 보인다. 이는 Grad-CAM 계열에서 종종 발생하는 그래디언트 노이즈(gradient noise)나 그래디언트 포화(saturation)와 같은 문제로부터 상대적으로 자유롭기 때문이다. 이러한 특성은 모델의 예측 근거와 무관한 계산적 인공물(computational artifact)을 제거하여, 보다 깨끗하고 안정적인 시각적 설명을 가능하게 한다.

또한 Score-CAM은 설명 신뢰도(faithfulness) 측면에서 중요한 장점을 갖는다. 기존 기법들이 그래디언트 값을 가중치의 근사치로 활용한 것과 달리, Score-CAM은 각 특성 맵이 클래스 점수에 기여하는 정도를 순방향 전파(forward propagation)를 통해 직접 측정한다. 이로 인해 모델이 실제로 의존한 시각적 증거를 더욱 충실하게 반영하며, 객체의 일부만 강조되는 경향을 보이는 Grad-CAM 계열과 달리 객체 전체 영역을 보다 일관되게 강조하는 결과를 제공한다. Score-CAM은 이러한 구조적 장점을 바탕으로 다양한 응용 가능성을 가진다. 예를 들어, 높은 신뢰도는 정밀한 모델 디버깅(precise model debugging)에 활용될 수 있으며, 잘못된 예측 상황에서 모델의 오류 원인을 제공하는 영상 영역을 기존 기법보다 명확하게 식별하는 데 도움을 준다. 또한 Score-CAM의 시각적 안정성을 기반으로, 활성을 평활화(smoothing)하거나 여러 계층의 Score-CAM 맵을 통합하는 다층 융합(multi-layer fusion) 접근을 적용하여 설명의 견고성(robustness)과 정밀도를 더욱 향상시키는 방향으로 확장될 수 있다. 마지막으로, Score-CAM의 순방향 기여도 계산 방식은 비-그래디언트(non-gradient) 기반 설명 기법의 이론적 토대를 제시했다는 점에서 학문적 의미를 갖는다. 그래디언트 없이도 시각적 설명이 가능함을 입증하는 XAI 기법이 등장하는 기반을 마련하였다.

제 3 절 특징 분리(Feature Separation)를 위한 색공간 이론

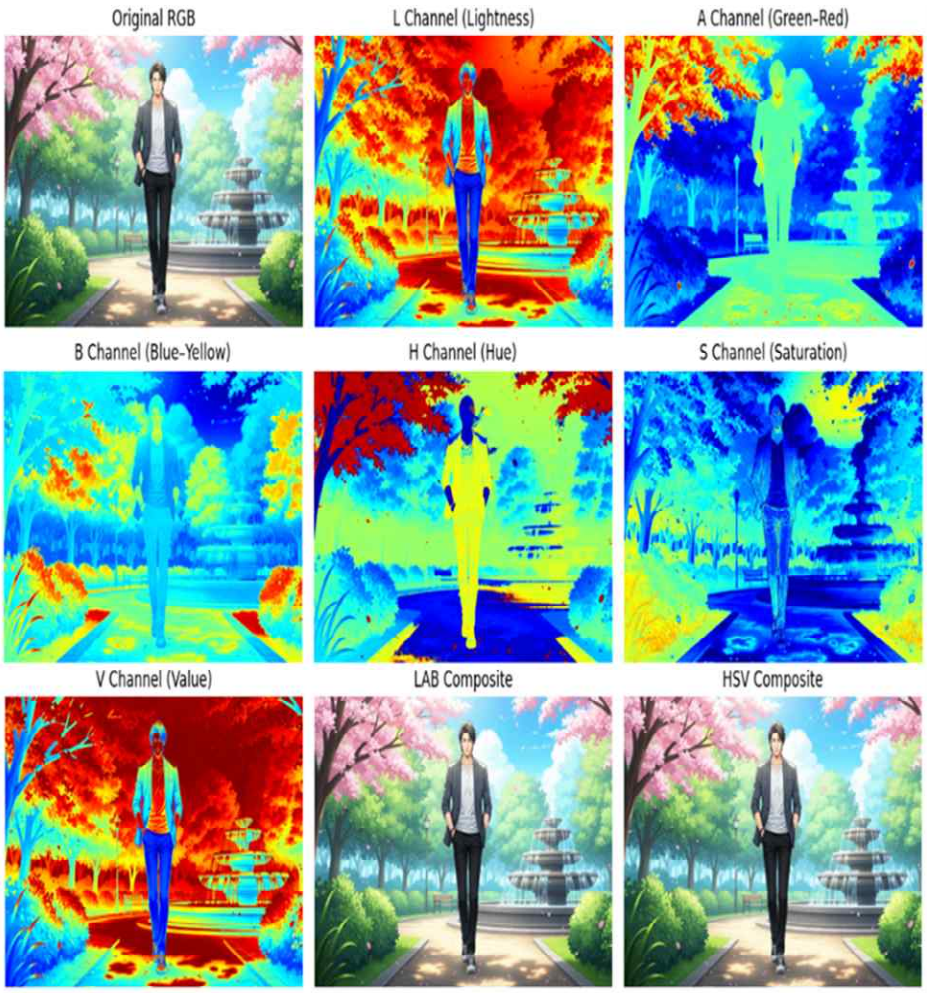
제1절에서 논의한 바와 같이, Class Activation Mapping 계열의 연구들은 모델이 이미지의 어느 공간적 위치에 주목하는지 시각화하는 데 크게 기여했다. 그러나 이러한 활성화 맵(Heatmap)은 모델이 왜 그곳을 보았는가?라는 질문에 대해 색상 때문인지, 형태 때문인지, 혹은 질감(Texture) 때문인지 명확하게 분리하여 설명하지 못하는 한계를 지닌다.

이러한 모호성의 근본 원인 중 하나는 대부분의 CNN이 입력으로 사용하는 BGR (또는 RGB) 색공간의 본질적인 특징 뒤얽힘(Feature Entanglement) 문제에 있다. BGR은 디스플레이와 같은 기계 장치를 위한 색상 표현 방식일 뿐, 인간의 시지각 시스템이 세상을 인식하는 방식과 거리가 멀다[13]. BGR에서 밝기와 색상 정보는 세 채널(B,G,R)에 비선형적으로 혼재되어 있다. 예를 들어, (100,100,100)의 회색 픽셀을 (120,100,100)의 옅은 분홍색으로 변경할 때, 이 변화는 색상(분홍빛)의 변화일 뿐만 아니라 밝기(명암)의 변화까지 동시에 유발한다[14]. 따라서 BGR 기반의 CAM 분석은 모델이 밝기에 반응한 것인지 색상에 반응한 것인지 구분할 수 없게 만든다. 이는 모델의 행동 원리를 인과적으로 규명하고자 할 때 심각한 교란 변수(Confounding Variable)로 작용한다.

본 연구의 핵심 목표인 인과 관계 증명을 달성하기 위해서는, 모델의 반응을 유발하는 핵심 요인을 분리하여 통제하는 개입(Intervention) 실험이 필수적이다. 이를 위해 본 절에서는 BGR의 한계를 극복하고, 인간의 시지각 방식과 유사하게 명암과 색상정보를 분리할 수 있는 지각적 색공간(Perceptual Color Space)인 CIE LAB와 HSV를 이론적 배경으로 고찰한다.

13) Gonzalez, R. C., & Woods, R. E. (2018). Digital Image Processing, 4th ed. Pearson.

14) van de Sande et al.(2011). Color Features for Visual Place Recognition.



[그림 2-7] 이미지의 LAB-HSV 색채널 분리 결과

1) CIE LAB: 인과관계 실험을 위한 핵심 이론

본 연구가 웨이블릿 개입의 핵심 대상으로 삼는 CIE LAB (CIELAB) 색공간은 1976년 국제조명위원회(CIE)에 의해 정의된, 인간의 색상 인지 모델을 수학적으로 근사한 대표적인 색공간이다. LAB의 가장 큰 특징은 색상 정보를 인간의 시각이 인지하는 방식과 유사하게 세 가지 독립적인 채널로 분리한다는 점이다.

가) L 채널 : 구조·형태·텍스처 정보에 집중된 인과개입

L 채널은 인간의 시각 시스템이 가장 민감하게 반응하는 명암(Lightness) 또는 밝기(Brightness) 정보만을 독립적으로 담고 있다. L값은 0(완전한 흑색)에서 100(완전한 백색)까지의 연속 범위를 가지며, 색상(a*, b*) 정보와 완전히 분리되어 있다. L 채널은 이미지의 전역 구조(Structure), 형태(Shape), 경계선(Edge), 질감(Texture) 정보를 대부분 포함하고 있기 때문에 흑백 사진과 유사한 형태의 시각적 정보를 제공한다. 본 연구에서 수행하는 인과개입(Intervention) 실험은 바로 이 L 채널을 중심으로 이루어진다. 저자는 L 채널에 대해 웨이블릿 변환을 수행하고, 그 중에서도 고주파 성분에 선택적으로 개입한다. 이는 색상 정보(a, b)는 완전히 고정한 상태에서 오직 명암 기반 텍스처만을 정밀하게 조작하는 실험 환경을 제공한다.

만약 L 채널의 고주파 성분 변화만으로 모델의 타깃 클래스 신뢰도$P(c)$가 달라진다면, 이는 모델이 색상(Chrominance)이 아니라 텍스처/형태 기반 구조적 정보(Luminance Structure)에 인과적으로 반응한다는 강력한 증거가 된다.

나) A 채널(G－R 색차): 색조 편향의 인과적 검증

A 채널(a*)은 CIE LAB 색공간에서 초록－빨강(Green－Red) 축의 색차(Chromatic Difference)만을 독립적으로 표현한다. A 값이 낮을수록 녹색 기운을, 높을수록 붉은 기운을 나타내며, 명암(L)이나 휘도 구조는 전혀 포함하지 않는다. 따라서 A 채널은 이미지의 색온도(Color Temperature)나 적색 계열 특징에 대한 모델의 민감도를 분석하는 데 최적화된 개입 도구이다. 예를 들어, 피부(dermatology), 과일(fruit), 혈류(blood vessel) 등 적색 성분이 특징인 데이터셋에서 모델이 구조적 정보보다 색조에 과도하게 의존하고 있는지를 파악할 수 있다. 본 연구에서 A 채널을 조작하는 것은, 색상만 변화시키고 구조(L)는 고정한 상태에서 모델의 신뢰도 변화를 측정하는 인과적 실험이다. 만약 A 채널의 조작만으로 $P(c)$가 크게 변한다면, 모델은 형태나 텍스

처가 아닌 색조 편향(Color Bias)에 의해 결정을 내리고 있음을 의미한다. 반대로 변화가 없다면 모델은 적색 계열 변화에 둔감하며 구조적 특징(L기반)에 주로 의존하고 있음을 의미한다.

다) B 채널(B-Y 색차): 조명 기반 인식 편향의 인과 검증

B 채널(b*)은 파랑-노랑(Blue-Yellow) 축의 색차 정보를 표현하며, 조명(illumination)과 환경 색온도(color tone)에 민감하게 반응한다. 노란빛이 강한 실내 조명, 푸른빛이 포함된 자연광 등 촬영 환경의 색조 변화가 B 채널에 직접 반영된다. 따라서 B 채널 개입은 모델이 조명 기반 색편향(illumination-induced bias)을 갖고 있는지 여부를 인과적으로 측정할 수 있는 핵심 실험이다. 본 연구에서 B 채널만 조작하고 L·A 채널을 고정한 상태에서 모델의 신뢰도 변화를 관찰하면, 모델이 형태가 아니라 조명·색온도 변화에 의해 오 판정(Misclassification)되는지를 분리해 분석할 수 있다. 만약 B 채널 변화만으로도 $P(c)$가 크게 흔들린다면, 이는 모델이 대상의 실제 구조보다 조명 조건에 의존하는 취약한 인식 구조로 되어 있음을 의미한다.

2) HSV: 인과관계 실험을 위한 핵심 이론

HSV(Hue-Saturation-Value) 색 공간은 인간의 직관적 색채 인식 방식을 모델링하기 위해 설계된 색 공간으로, 이미지의 색조(H), 채도(S), 명도(V)를 서로 독립된 세 개의 성분으로 분리하여 표현한다. HSV는 기존의 RGB나 BGR과 달리 색상 자체(H), 색의 강도(S), 그리고 밝기(V)를 분리된 차원으로 나타내기 때문에, 인간이 시각적으로 색을 인식하고 구분하는 방식과 더 근접한 표현 구조를 제공한다. 이 색 공간은 색상 변화가 지각적으로 어떻게 인식되는지를 잘 반영하기 때문에, 본 연구와 같이 색채 요인과 구조적 요인을 분리하여 인과적 개입(Intervention)을 수행해야 하는 실험에 매우 적합하다. 특히 HSV는 RGB보다 더 명확하게 색조(Hue)만을 분리하여 조작할 수 있고, 채도(Saturation)와 밝기(Value)를 독립된 변수로 제어할 수 있다는 장점이 있

다. 따라서 HSV 색 공간을 이용하면 H(색조), S(채도), V(명도)를 각각 독립적인 개입 변수로 설정하여, 모델이 색 자체에 반응하는지(H), 색의 강도나 대비에 반응하는지(S), 혹은 밝기 변화에 반응하는지(V)를 인과적으로 검증할 수 있다. 이러한 특성은 본 연구에서 수행하는 색채 기반 개입 실험(Color Intervention)에서 중요한 이론적 기반을 형성한다.

가) H 채널 (Hue): 순수 색조(Hue) 기반 인식 의존성 평가의 핵심

H 채널은 HSV 색 공간에서 색조(Hue)만을 표현하며, 빨강·노랑·초록·파랑 등 색의 고유한 각도(0° ~360°)를 나타낸다. 밝기(Value)나 채도(Saturation)의 영향 없이 '순수 색 자체'를 나타내기 때문에, H 채널은 색조 변화가 모델 인식에 미치는 인과적 영향을 직접 측정할 수 있는 가장 강력한 좌표계이다. H 채널을 조작하는 실험은 다음을 명확하게 구분해 낸다. 만약 H 채널 조작만으로 모델의 $P(c)$가 크게 변한다면, 모델이 색 자체에 과도하게 의존하고 있는 것이며, 이는 색조 기반 편향(Hue-driven bias)의 존재를 인과적으로 입증한다. 반대로 구조·명암 기반으로 판단하는 모델이라면 H 조작은 거의 영향을 미치지 않는다.

나) S 채널 (Saturation): 색 강도(Color Strength) 의존도의 인과적 측정

S 채널은 HSV 공간에서 색의 강도(Saturation) 또는 순도(Purity)를 나타낸다. S값이 낮으면 빛바랜 회색 조에 가까워지고, 값이 높으면 선명하고 강렬한 색이 된다. S 채널을 조작하는 인과 실험은 색의 강함/약함이 모델의 판단에 얼마나 영향을 주는가, 구조(L)와 색조(H)가 동일해도 채도 변화만으로 모델이 오판 정하는지, 모델이 색 대비(Chromatic Contrast)에 민감한가, 텍스처에 민감한가를 저채도가 상황(S↓)에서 모델의 $P(c)$가 급락한다면, 이는 모델이 색의 강도에 의존하고 있음을 보여준다. 반대로 고채도(S↑) 상황에서 반응이 과도하게 증가하면 색 강도 편향(Saturation bias)을 가진 것으로 해석할 수 있다.

다) V 채널 (Value): 색상 영향이 혼합된 명도 기반 인과적 분해

V 채널은 HSV 색 공간에서 밝기(Value)를 나타내지만, LAB의 L과 달리 색상(H, S) 정보가 일부 섞인 밝기다. 즉, V는 '색 영향이 포함된 밝기'라는 특수성을 가진다. 순수한 구조 기반 밝기(L) 변화가 중요했는지, 아니면 색상 영향이 포함된 V 밝기가 모델 인식에 더 크게 작용했는지를 실험하여 동일한 구조(L)를 유지한 채 V값만 증가시키면 이미지는 밝아지면서 동시에 색이 더 강하게 보이는 효과가 나타난다. 모델이 이 변화에 반응한다면 이는 색상 혼합 밝기(Value-driven bias)에 따라 인식하는 경향을 보여준다. 만약 모델의 $P(c)$가 V 변동에는 민감하지만, L 변동에는 둔감하다면, 모델은 구조 기반 인식이 약하고 색-밝기 복합 자극에 의존하는 모델임을 의미한다.

제 4 절 웨이블릿 변환의 원리 및 주파수 분석

본 절에서는 본 연구의 핵심 분석 축을 이루는 웨이블릿 변환(Wavelet Transform)의 수학적 원리와 주파수 해석적 특성을 다룬다.

1) 개념 및 등장 배경

웨이블릿 변환(Wavelet Transform)은 신호나 이미지를 공간 정보(Spatial Information)와 주파수 정보(Frequency Information)를 동시에 분석할 수 있는 수학적 도구로, 푸리에 변환(Fourier Transform)의 한계를 보완하기 위해 제안되었다. 푸리에 변환은 신호 전체의 주파수 성분(frequency component)을 분석할 수 있으나, 시간적·공간적 위치 정보가 소실되어 어느 영역에서 어떤 주파수가 발생하는가를 구체적으로 알 수 없다. 반면, 웨이블릿 변환은 짧은 국소 파형(Localized Waveform)을 이동 및 축소·확대하면서 신호를 분석하기 때문에, 어디서, 어떤 주파수가 발생했는가를 동시에 관찰할 수 있다. 이러한 특성으로 인해 웨이블릿 변환은 이미지 처리, 신호 분석, XAI(Explainable AI) 등 다양한 영역에

서 다층적(multilevel)이고 지역적(localized) 정보를 분석하는 강력한 도구로 사용되어 왔다. 즉, 푸리에 변환이 전체적인 평균 스펙트럼 구조를 보여주는 반면, 웨이블릿 변환은 이미지 내부의 국소적 주파수 변동(Local Frequency Variation)을 시각적·정량적으로 포착할 수 있다. 본 연구에서는 이러한 웨이블릿 변환의 특성을 활용하여, Score-CAM 기반 활성화 맵에 웨이블릿 변환을 적용함으로써 모델의 시각적 판단이 주파수 대역별로 어떻게 반응하는지를 정량적으로 규명하였다.

2) 웨이블릿 변환의 해석적 원리

웨이블릿 변환(Wavelet Transform)은 신호 $f(t) \in L^2(R)$ 상에서 시간(또는 공간)과 주파수(스케일) 정보를 동시에 분석할 수 있는 강력한 해석적 도구로, 비정상 신호(non-stationary signal)의 국소적 특성을 정밀하게 포착하는 데 적합하다. 이는 푸리에 변환(Fourier Transform)이 가진 시간 영역 내 비국소성(non-locality)의 한계를 보완하며, 시간-주파수 해석(time-frequency analysis) 분야에서 핵심적인 위치를 차지한다. 웨이블릿 변환은 기본적으로 모 웨이블릿(mother wavelet)이라 불리는 한 개의 기준 함수$\psi(t)$를 다양한 스케일(scale)과 위치(translation)로 변환하여 전체 신호 공간을 구성하는 방식으로 정의된다. 이러한 접근은 신호의 전체적인 주파수 구성뿐 아니라, 시간(또는 공간) 상의 지역적 변화(local variation)를 동시에 기술할 수 있게 한다.

가) 연속 웨이블릿 변환 [15]

연속 웨이블릿 변환은 신호$f(t)$를 시간-스케일 평면(time-scale plane)으로 사상(mapping)하는 적분 변환으로 정의된다. 모 웨이블릿$\psi(t)$가 허용 조건(admissibility condition)을 만족할 때, 연속 웨이블릿 계수는 다음과 같이 표현된다.

여기서, 웨이블릿 기저 $W_{\psi}(a,b)$는 다음과 같이 정의된다.

15) 김용대, 『디지털 신호처리』, 한티미디어, 2019, pp. 241-247. 강영민, 『웨이블릿 변환과 응용』, 교보문고, 2017, pp. 25-33.

$$\Psi_{a,b} = \frac{1}{\sqrt{|a|}}\Psi(\frac{t-b}{a})$$

- a∈(0,+∞):스케일(scale) 파라미터로, 웨이블릿의 신축을 제어한다. a가 클수록 긴 시간 구간의 저주파 성분(저해상도)을, a가 작을수록 짧은 구간의 고주파 성분(고해상도)을 분석한다.
- b∈R: 이동(translation) 파라미터로, 시간 축 상에서 웨이블릿의 위치를 조정한다.
- $\frac{1}{\sqrt{|a|}}$: L^2-norm 보존을 위한 에너지 정규화 항이다.
- *: 켤레 복소수(Complex Conjugate)를 의미한다.

CWT는 하이젠베르크 불확정성 원리(Heisenberg's Uncertainty Principle)에 기반하여, 저주파 영역에서는 주파수 해상도가 높고 시간 해상도가 낮으며, 반대로 고주파 영역에서는 시간 해상도가 높고 주파수 해상도가 낮은 다중 해상도(multiresolution) 특성을 갖는다. 이러한 특성은 신호 내 급격한 변동이나 경계와 같은 국소 구조를 효과적으로 검출할 수 있게 한다.

나) 이산 웨이블릿 변환 16)

연속 웨이블릿 변환은 모든 a,b에 대해 연속적으로 계산되므로, 계산량이 매우 크고 높은 중복성(redundancy)을 가지는 단점이 있다. 이를 해결하기 위해 스케일과 이동 파라미터를 이산화(discretization)하여 구성한 것이 이산 웨이블릿 변환이다. DWT는 일반적으로 이진 다이아딕(dyadic) 샘플링을 기반으로 하며, 다음과 같은 파라미터 격자를 사용한다.

$$a = 2^j, \quad b = k \cdot 2^j, \quad j,k \in Z$$

이러한 구조는 Mallat(1989)이 제안한 다중 해상도 분석 이론에 의해 정식화되었다. MRA는 신호 공간$L^2(R)$을 해상도 수준(resolution level) j에 따라 계층적으로 분할된 직교 부분 공간들의 합으로 표현한다.

$$L^2(R) = \bigoplus_{j \in Z} W_j, \quad where\ V_{j-1} = V_j \oplus W_j$$

여기서,

16) 김용대, 『디지털 신호처리』, 한티미디어, 2019, pp. 247-253.
강영민, 『웨이블릿 변환과 응용』, 교보문고, 2017, pp. 35-44.

- V_j: 스케일 2^j에서의 근사 공간(approximation space)으로, 스케일링 함수(scaling function)$\phi(t)$에 의해 생성된다.
- W_j: V_{j-1}에 대한의 V_j직교 여공 간으로, 웨이블릿 함수 $\psi(t)$에 의해 생성되는 세부 공간이다. 즉, V_j는 저주파(거친) 성분을, W_j는 고주파(세부) 성분을 표현한다. 이를 반복적으로 적용하면 신호를 다층적 해상도로 분해할 수 있다.

다) 2차원 이산 웨이블릿 변환 및 이미지 분석[17][18][19]

2차원 이산 웨이블릿 변환(2D Discrete Wavelet Transform, 2D-DWT)은 1차원 웨이블릿 변환의 개념을 공간 좌표 (x,y)영역으로 확장하여, 이미지를 다중 해상도의 주파수 성분으로 분해하는 기법이다. 이는 이미지 내의 형태적 구조(form structure)와 세부 질감(texture detail)을 서로 다른 스케일(scale)에서 동시에 분석할 수 있도록 하여, 공간적 국소성과 주파수 정보를 결합한 효율적 표현을 가능하게 한다. 2D-DWT는 1차원 스케일링 함수 $\phi(x)$와 웨이블릿 함수 $\psi(x)$를 두 축으로, 독립적으로 적용하는 분리 가능한 기저(separable basis)를 이용한다. 이를 통해, 2차원 이미지 $f(x,y) \in L^2(R^2)$는 다음과 같이 근사 성분과 세부 성분의 조합으로 표현된다.

- 스케일링 함수 (Scaling Function)

$$\phi(x,y)=\phi(x)\phi(y)$$

이는 이미지의 저주파(approximation) 성분을 표현하며, 전역적인 형태(structure)와 명암 분포(intensity distribution)를 보존한다.

- 웨이블릿 함수 (Wavelet Function)

$\psi^{(LH)}(x,y) = \phi(x)\psi(y)$ (수평 방향 세부 성분, Horizontal detail)

$\psi^{(HL)}(x,y) = \phi(x)\psi(y)$ (수직 방향 세부 성분, Vertical detail)

17) Antonini et al.(1992). Image Coding Using Wavelet Transform.

18) Li, C., & Orchard, M. T. (2001). New Edge-Directed Interpolation.

19) Unser, M. (1995). Texture Classification and Segmentation Using Wavelet Frames. IEEE Transactions on Image Processing, 4(11), 1549-1560.

$\psi^{(HH)}(x,y) = \phi(x)\psi(y)$ (대각 방향 세부 성분, Diagonal detail)

이들은 각각 이미지 내 수평, 수직, 대각 방향의 고주파 정보를 추출하며, 에지(edge), 패턴(pattern), 질감(texture)과 같은 세부 요소를 명확히 분리한다.

- 근사 계수 (Approximation Coefficients)

$$C_{m,n} = < f, \phi_{m,n} >$$

이미지의 저주파 대역(LL)에 해당하며, 전체적인 형태 및 구조를 유지한다.

- 세부 계수 (Approximation Coefficients)

$$d_{m,n}^{(k)} = < f, \Psi_{m,n}^{(k)} >, k \in LH, HL, HH$$

고주파 대역의 세부 정보를 포함하며, 국소적 변화(local variation)나 경계(edge) 정보를 반영한다. 이러한 분해 과정을 통해, 2D-DWT는 원본 이미지를 하나의 저주파 근사 성분(LL)과 세 방향의 고주파 세부 성분으로 구성된 4개의 주파수 대역으로 구분한다. 각 대역은 서로 직교(orthogonal)하여 정보의 중복을 최소화하며, 다음과 같은 물리적·시각적 해석을 제공한다.

[표 2-5] 웨이블릿 분해 대역별 정보 특성 및 시각적 의미

대역	주요 정보	시각적 의미
LL	저주파 근사 성분	전체 형태, 윤곽, 명암 구조
LH	수평 고주파 성분	수평 방향의 경계선, 텍스처 변화
HL	수직 고주파 성분	수직 방향의 경계선, 구조 변화
HH	대각 고주파 성분	미세 질감, 잡음, 세부 패턴

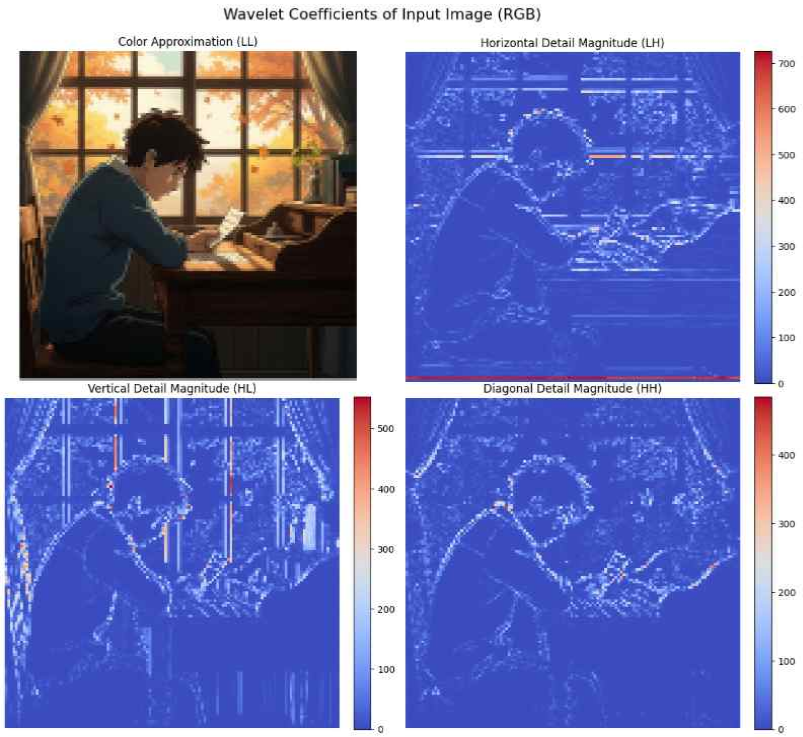
2D-DWT는 신호를 직교 기저(orthonormal basis)에 투영하여 정보 손실 없이 원본 이미지를 다중 주파수 대역으로 분해하는 효율적이고 비 중복적 표현 방식을 제공한다. 이미지 분석 관점에서 LL 성분은 전역적 형태와 명암의 안정적 표현에 유리하며, LH/HL/HH 성분은 세부 구조나 질감 검출에 효과적이다.

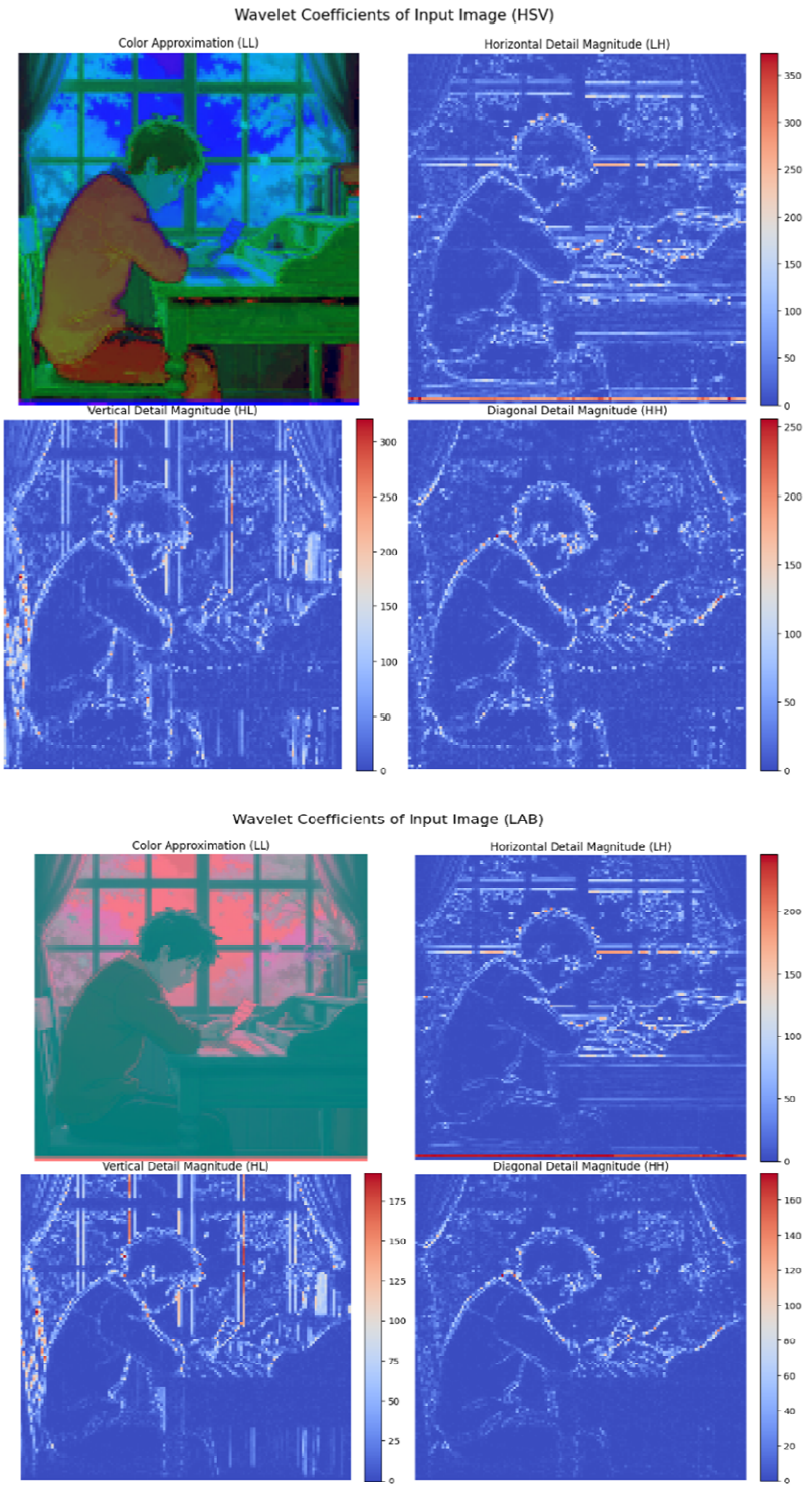
웨이블릿 기반 분해는 공간적 위치와 주파수 정보를 동시에 보존하는 다중 해상도 분석을 가능하게 한다. 이는 공간 필터링으로는 얻기 어려운 정밀한 구조적 표현을 제공한다. 특히 고주파 대역은 각각 수평·수직·대각 방향성

을 반영하므로 방향성 기반 특징 검출에서 뛰어난 성능을 보이며, CNN의 초기 합성곱 필터와 구조적 대응성을 갖는다.

이러한 다층적 정보 분해 특성은 주파수 선택성이나 구조적 민감도 분석과 같은 XAI 기반 인과 실험에 직접 활용될 수 있다는 점에서 학술적·응용적 가치가 크다. 따라서 2D-DWT는 이미지 복원, 압축, 특징 추출, 그리고 설명 가능 인공지능 기반 시각적 해석에 이르기까지 폭넓게 활용될 수 있는 수학적 기반을 제공한다.

웨이블릿 계수의 에너지 분포는 이미지의 구조적 복잡도를 정량적으로 평가하는 데 활용될 수 있어, 엔트로피 기반 정보 측정이나 품질 평가 지표와의 결합도 용이하다. 또한 웨이블릿 대역 간 상관관계를 분석하면 저주파 및 고주파 성분이 모델의 예측 과정에서 보이는 상호작용을 정량적으로 파악할 수 있어 인과적 해석 능력을 확장한다.





[그림 2-8] 색상 공간별 웨이블릿 대역(LL, LH, HL, HH) 비교

[그림 2-8]은 동일한 입력 이미지를 RGB, LAB, HSV의 세 가지 색상 공간으로 변환한 뒤, 각 색상 공간에서 2차원 이산 웨이블릿 변환을 수행하여 얻은 네 개의 서브밴드를 시각적으로 비교한 것이다. RGB 공간에서는 색상 강도(intensity)와 대비(contrast)에 따른 구조적 차이가 두드러지며, LAB 공간에서는 휘도(L)와 색채 성분(a*, b*)의 분리로 인해 명암 대비와 색차가 보다 정교하게 표현된다. 반면, HSV 공간에서는 색상과 채도가 분리되어, 주파수 대역별로 색조 및 밝기 변화가 명확히 드러난다. 특히, 각 색상 공간에 대해 모든 채널(R,G,B/L,a*,b*/H,S,V)의 서브밴드 계수를 통합(summation)함으로써 산출된 총에너지(E_{total})는, 해당 색상 공간이 전체적으로 어느 정도의 주파수 응답(frequecy response)을 포함하고 있는지를 정량적으로 나타낸다.

즉, RGB, LAB, HSV 각각에서 감지된 모든 에지(edge) 및 텍스처(texture) 정보의 누적 강도(ΣE)는 색상 공간의 정의 방식에 따라 주파수 대역별로 상이한 분포 특성을 보인다. 이러한 비교는 단순한 색상 변환의 차이를 넘어, 색상 공간의 구성 원리가 웨이블릿 대역의 주파수 반응 특성에 실질적·정량적 영향을 미친다는 사실을 보여준다. 따라서 본 분석은 색상 공간 선택이 Score-CAM 및 주파수 기반 주의 맵(frequency-based attention map)의 해석적 결과에 직접적인 영향을 미칠 수 있음을 시사한다.

3) Score-CAM 활성화 맵의 주파수 선택성 분석

웨이블릿 변환은 단순한 신호 분해 기법을 넘어, 심층 신경망(Deep Neural Network)의 내부 인식 메커니즘을 주파수 영역(frequency domain)에서 해석할 수 있는 분석적 창구로 확장될 수 있다. 웨이블릿은 신호를 주파수 성분으로 분해함과 동시에, 해당 주파수가 공간적으로 어느 위치에서 발생하는지를 나타내는 위치 정보(localization)를 함께 제공한다.

이러한 특성은 단순한 푸리에 변환 기반 분석과 달리, CNN의 활성화 구조를 다층 해상도(multi-resolution) 관점에서 정량적으로 이해할 수 있게 한다.

본 절의 분석은 저자가 수행한 선행 연구의 실험 결과를 기반으로 하며,

이를 본 논문에서 확장·적용할 수 있음을 입증한다. Score-CAM이 산출하는 활성화 맵 $A^{(c)}(x,y)$은 입력 이미지의 공간적 주의 분포(spatial attention distribution)를 반영한다. 이 맵은 CNN의 마지막 합성곱 레이어 활성화 맵들의 가중합으로 생성되는데, 이 활성화 맵을 만드는 합성곱 필터(convolutional filter)들은 학습 과정을 통해 본질적으로 특정 주파수 패턴(예: 저주파의 형태, 고주파의 질감) 을 감지하는 탐지기(detector)로 기능하게 된다. 따라서 $A^{(c)}(x,y)$는 단순한 공간적 분포를 넘어, 모델이 어떤 주파수 정보에 주로 반응했는지를 나타내는 주파수 응답 특성(frequency response characteristics)을 암묵적으로 내포(embed)하게 된다.

가) 2D-DWT를 통한 주파수 대역 분해

이에 $A^{(c)}(x,y)$에 2차원 이산 웨이블릿 변환을 적용하여, 활성화 맵을 네 개의 주파수 서브밴드로 분해하였다.

$$A^{(c)}(x,y) \xrightarrow{2D-DWT} \{LL^{(c)}, LH^{(c)}, HL^{(c)}, HH^{(c)}\}$$

각 서브밴드는 서로 다른 주파수 영역과 방향성을 가지며, 이를 통해 Score -CAM의 주파수 선택적 반응(frequency-selective response)을 정량적으로 측정할 수 있다. 각 서브밴드의 에너지(Energy)는 해당 주파수 대역에서의 반응 강도와 집중도를 나타내며, 이는 각 밴드의 Frobenius Norm제곱으로 정의된다.

$$E_{band}^{(c)} = \sum_{x,y} \left| Band\, A^{(c)}(x,y) \right|^2, Band \in \{LL, LH, HL, HH\}$$

이때$E_{band}^{(c)}$ 값이 높을수록, 해당 클래스 c인식 과정에서 모델이 그 주파수 대역의 정보를 더 강하게 활용하고 있음을 의미한다.

- $E_{LL}^{(c)}$이 높게 나타나면, 모델이 전역적 형태(global form)나 저주파 중심의 구조적 특징에 의존하고 있음을 의미한다.
- 반대로 $E_{LH}^{(c)}, E_{HL}^{(c)}, E_{HH}^{(c)}$가 상대적으로 높을 경우, 모델은 세부 구조(local detail)즉, 경계(edge), 질감(texture), 패턴 변화(detail variation)에 더 민감하게 반응하고 있음을 시사한다.

나) 계수 가중 및 역변환을 통한 재구성

웨이블릿 서브밴드 $\{LL, LH, HL, HH\}$는 단순 합산으로 원본 맵을 복원할 수 없으며, 이는 반드시 차원 이산 웨이블릿 역변환(2D Inverse Discrete Wavelet Transform, IDWT)을 통해 재구성되어야 한다.

선행 연구에서 각 서브밴드 계수에 가중치ω_{band}를 부여하여 특정 대역의 강도를 조절한 뒤, 다음 식과 같이 새로운 활성화 맵 $\hat{A}^{(c)}$를 재구성하였다.

$$LL'^{(c)} = \omega_{LL} \cdot LL^{(c)},$$

$$LH'^{(c)} = \omega_{LH} \cdot LH^{(c)},$$

$$HL'^{(c)} = \omega_{HL} \cdot HL^{(c)},$$

$$HH'^{(c)} = \omega_{HH} \cdot HH^{(c)}$$

$$\hat{A}^{(c)} = 2D-IDWT(\{LL'^{(c)}, LH'^{(c)}, HL'^{(c)}, HH'^{(c)}\})$$

여기서 ω_{band}는 각 주파수 대역의 상대적 중요도를 조절하는 파라미터이며, 이렇게 재구성된 $\hat{A}^{(c)}$는 특정 주파수 영역을 강조하거나 억제한 주파수 조절형 활성화 맵(Frequency-modulated Activation Map)으로 활용된다.

다) 지배적 주파수 대역 (Dominant Frequency Band, DFB)

모델이 c를 인식하는 과정에서 가장 강하게 반응하는 대역은 다음과 같이 정의된다.

$$\Delta f_c = \arg_{band \in \{LL, LH, HL, HH\}} \max (E_{band}^{(c)})$$

여기서 Δf_c는 클래스c에 대한 모델의 지배적 주파수 대역을 의미하며, 이는 Score-CAM 맵의 주파수 응답 중심(frequency response center)으로 해석된다. 이 지표를 통해 모델이 전역적 특징(저주파) 또는 국소적 세부 구조(고주파) 중 어느 쪽에 더 의존하는지를 정량적으로 비교할 수 있다.

라) 해석적 의의

이와 같은 절차는 Score-CAM의 시각적 설명(visual explanation)을 단순한 공간적 주목(spatial attention)의 차원을 넘어 주파수 선택적 반응(frequency- selective response)으로 확장한다. 즉, 모델이 클래스별로 서로

다른 주파수 대역을 선택적으로 활성화하는 경향은, 딥러닝 신경망이 단순히 공간적 패턴을 학습하는 것이 아니라 주파수적 인지 반응(frequency-aware cognition)을 수행하고 있음을 시사한다. 결과적으로, 웨이블릿 기반 Score-CAM 주파수 해석은 XAI 관점에서 딥러닝 모델의 인식 메커니즘을 다층 해상도 및 주파수 선택성의 틀에서 설명할 수 있는 새로운 분석 패러다임을 제시한다.

제 5 절 웨이블릿-Grad-CAM++ 기반 멀티스케일 활성화 복원 연구

본 절에서는 저자의 선행 연구[20]를 분석하여, 본 연구가 기존 연구를 어떻게 확장하고 새로운 인과적 관점을 도입하는지 비교·정리한다.

1) 선행 연구의 개념 및 등장 배경

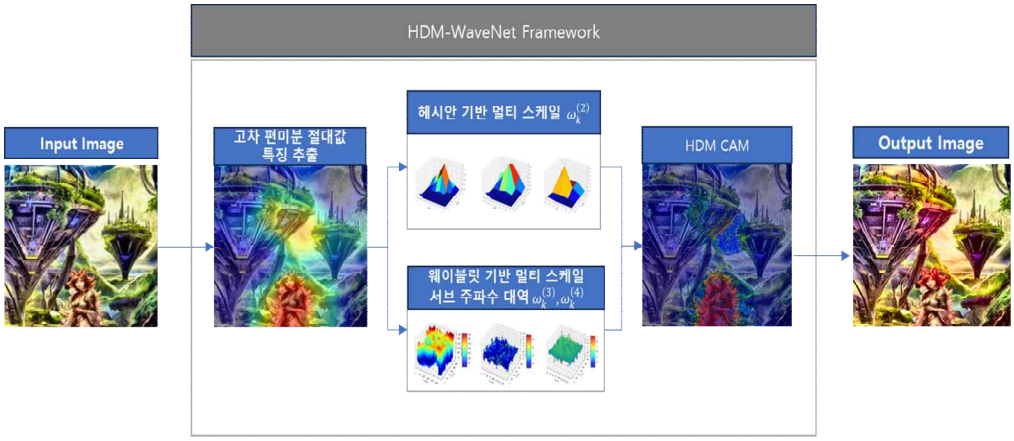
헤시안 기반 Grad-CAM++와 웨이블릿 융합 멀티스케일 활성화 맵으로 고정밀 이미지 복원 연구는 딥러닝 기반 이미지 복원 성능 향상을 목표로 하였다. 기존의 Grad-CAM++와 같은 활성화 맵(CAM) 기법은 단일 스케일(Single-scale)에 의존하여, 이미지의 복잡한 텍스처(고주파 성분)와 전역적인 구조(저주파 성분)를 동시에 포착하는 데 한계가 있었다. 이러한 한계를 극복하기 위해, HDM-WaveNet을 제안하였다. 이 모델의 핵심 목표는 성능 향상을 통하여 XAI 기법을 고도화하여 설명력을 높이는 동시에, 이를 이미지 복원 품질 향상에 직접 적용하는 것이었다.

2) 선행 연구의 해석적 원리: 멀티스케일 융합

HDM-WaveNet의 핵심 원리는 고차 편미분과 웨이블릿 변환을 융합하여 멀티스케일 활성화 맵을 생성하는 데 있다. 헤시안 기반 2차 편미분은 이미지 곡률(Curvature) 정보를 정량화하여 세밀한 텍스처를 탐지하며, 웨이블릿 기

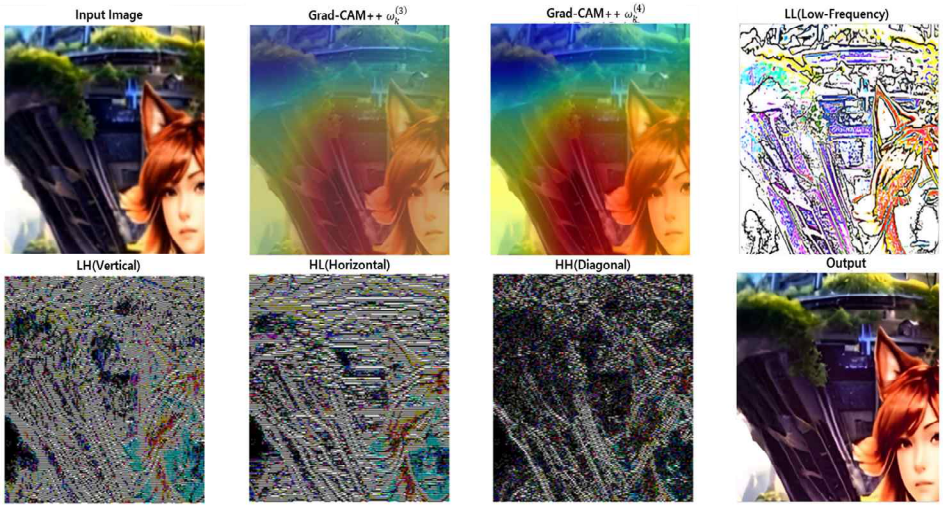
20) 권태윤, 노광현. (2025). 헤시안 기반 Grad-CAM++와 웨이블릿 융합 멀티스케일 고정밀 이미지 복원. 한국정보기술학회논문지, 23(7), 39-48.

반 고차 편미분은 특징 맵을 저주파 및 고주파 서브밴드로 분해하여 스케일 간 정보 손실 없이 활성화 맵을 재구성한다. 이 과정을 통해 단일 스케일 분석의 한계를 극복하고 멀티스케일 기반의 정밀 활성화 맵을 생성하여 IoU, PSNR, SSIM과 같은 성능지표를 향상시켰다.



[그림 2-9] HDM-WaveNet의 프레임워크 구조

[그림 2-9]는 HDM-WaveNet이 입력 이미지로부터 고차 편미분 기반 멀티 스케일 특징을 추출하고, 웨이블릿 서브밴드를 통합하여 HDM-CAM을 생성한 뒤, 최종 복원 이미지를 출력하는 전체 프레임워크를 시각적으로 보여준다.



[그림 2-10] HDM-WaveNet 기반 각 주파수 대역 반영 시각화 비교

[그림 2-10]은 HDM-WaveNet을 적용하였을 때 각 주파수 대역이 활성화 맵에 반영되는 방식과 시각적 차이를 비교한 예시이다. 이를 통해 선행 연구가 멀티스케일 주파수 정보를 실제 활성화 맵 생성 과정에 효과적으로 통합했음을 확인할 수 있다.

3) HDM-WaveNet 해석적 원리

HDM-WaveNet은 ResNet 기반 CNN의 공간적 표현과 Wavelet 변환의 다중 해상도 주파수 표현을 하나의 통합 구조 안에서 결합함으로써, 기존 CNN이 가진 정보 손실 문제를 보완하고 모델의 해석 가능성을 강화하기 위해 제안된 하이브리드 모델이다. 이 모델의 기본 개념은 (1) 고차원 CNN 특징(High-Dimensional Spatial Feature)의 추출, (2) 웨이블릿 변환을 통한 다중 스케일 주파수 분해, (3) 공간-주파수 도메인의 특징 융합(Multi-domain Feature Fusion), (4) Hessian 기반 Grad-CAM++을 통한 해석 가능한 활성화 맵 생성, (5) 다중 스케일 활성화 맵을 활용한 고정밀 이미지 복원의 다섯 단계로 구성된다.

첫째, HDM-WaveNet은 ResNet50의 계층 구조를 그대로 사용하여 다양한 깊이에서 고차원 CNN 특징을 추출한다. 초기 레이어에서 감지되는 에지(edge), 텍스처(texture)와 같은 저수준 특징뿐 아니라, 고층 레이어에서 추출되는 구조적·의미론적 특징까지 포함하는 다층적 고차원 표현을 확보하는 것이 핵심이다. 이러한 특징은 CNN의 공간적 시각 단서를 기반으로 하며, 모델의 인식 구조를 구성하는 핵심적인 특징 공간을 형성한다. 둘째, HDM-WaveNet은 CNN에서 추출된 각 레이어의 특징 맵을 이산 웨이블릿 변환(Discrete Wavelet Transform, DWT) 을 통해 LL,LH,HL,HH의 네 개 서브밴드로 분해한다. LL은 전체적인 밝기·형태 정보를 포함하는 저주파 성분이고, LH/HL/HH는 수평·수직·대각 방향의 고주파 텍스처 정보를 각각 포함한다. 이 과정은 CNN의 단일 스케일 특징이 놓칠 수 있는 다중 해상도(multi-resolution) 및 방향성(orientation) 기반 정보를 보완하며, 특히 복원·재구성 과정에서 중요한 디테일을 보존하는 기제를 제공한다. 셋째, 이렇게

얻어진 CNN 특징과 웨이블릿 서브밴드를 하나의 공통 표현 공간으로 융합(feature fusion) 한다.

$$F_{HDM} = [F_{ResNet} \| LL \| LH \| HL \| HH]$$

의 형태로 통합 표현을 구성한다. 이를 통해 HDM-WaveNet은 공간적 의미(semantic)·구조적 패턴(structure)·다중 주파수 텍스처를 동시에 포함하는 고밀도·고해상도 특징 표현을 확보한다. 넷째, HDM-WaveNet은 모델의 활성화 과정을 단순히 시각화하는 것에서 나아가 Hessian 기반 Grad-CAM++을 적용하여 공간-주파수 도메인의 융합 특징이 최종 모델 출력에 어떠한 기여를 하는지 설명한다. Hessian(절댓값 2차/3차 미분)을 활용한 Grad-CAM++은 기존 Grad-CAM보다 작은 객체나 중첩된 패턴에도 민감하게 반응하며, 각 웨이블릿 서브밴드가 가지는 주파수 중요도를 정밀하게 가중치로 반영할 수 있다. 그 결과, 다중 스케일 활성화 맵을 생성하여 CNN이 주목하는 영역이 어떤 주파수 대역과 연결되는지를 시각적으로 해석할 수 있다.

마지막으로, HDM-WaveNet은 이러한 다중 스케일 활성화 맵을 기반으로 이미지 복원을 수행한다. LL 대역을 활용하여 구조적 일관성을 유지하며, LH/HL/HH 대역을 활용하여 텍스처 세부 성과 샤프니스(sharpness)를 보존한다. 동시에 Grad-CAM++ 기반 가중치는 복원 과정에서 중요한 공간적 위치와 주파수 성분을 선택적으로 강조하여 기존 ResNet 기반 복원 모델보다 높은 PSNR, SSIM, 엔트로피 및 손실을 최소화할 수 있는 성능을 달성한다. 즉, HDM-WaveNet은 단순한 복원 모델이 아니라, 다중 스케일·다중 도메인 정보를 통합하여 데이터의 특성에 따라 스케일별 기여도를 적응적으로 조정하여 수행하는 해석 가능한 CNN 아키텍처로 정의할 수 있다.

4) 선행 연구의 성과와 기술적 한계

선행 연구(HDM-WaveNet)는 XAI 기법이 단순한 사후(Post-hoc) 분석 도구를 넘어, 이미지 복원이라는 특정 태스크의 성능을 향상시키는 엔지니어링 도구로 활용될 수 있음을 입증했다는 성과가 있다. 웨이블릿을 통해 주파수 대역별 정보를 융합(Fusion)하는 접근법이 유효함을 보였다. 하지만, 이 연

구는 기술적·상관적 분석에 머무른다는 한계점을 지닌다. 모델이 왜 그렇게 예측했는가? 라는 근본적 이해보다는, 어떻게 하면 더 나은 복원 결과를 얻을 수 있는가? 라는 응용(Application)에 초점이 맞춰져 있다. HDM-WaveNet은 모델의 예측에 기여하는 특징을 기술(Description)하는 설명 맵을 생성할 뿐, 특정 주파수 성분의 존재 여부가 모델의 예측 신뢰도에 인과적으로 어떤 영향을 미치는지, 그리고 얼마나 미치는지 정량화하지 못한다.

5) 본 연구로의 확장

본 연구는 선행 연구의 주파수 분석이라는 공통 분모 위에서, 기술(Description)에서 인과(Causality)로 응용(Application)에서 이해로 연구 관점을 전환 및 확장한다. 선행 연구가 주파수 정보를 융합(Fusion)하여 성능을 높이는 데 집중했다면, 본 연구는 반대로 색상 및 주파수 정보를 분리(Separation)하고 개입하여 모델의 반응을 측정한다.

HDM-WaveNet이 어떤 주파수가 활성화되는지 보여주는(Show) 데 그쳤다면, 본 연구는 특정 주파수 대역을 의도적으로 제거/변경했을 때, 모델의 예측 신뢰도 변화(ΔP)와 Score-CAM 활성화 변화를 측정하는 관점의 전환을 한다. 이미지 복원 성능 향상이 아닌, ResNet50이라는 특정 모델이 입력의 색상(LAB/HSV)과 주파수 성분에 얼마나, 그리고 어떻게 인과적으로 의존하는지 그 구조를 규명하는 것을 목표로 한다.

결론적으로, 본 연구는 선행 연구에서 활용한 웨이블릿 주파수 분석을 인과 추론의 핵심 도구로 재정의하여, 모델의 블랙박스 내부 작동 원리를 더 깊이 있게 규명한다.

제 3 장 색상·주파수 개입 기반 인과 분석 프레임워크

제 1 절 제안 프레임워크의 개념적 구조

1) 제안의 배경 및 필요성

CAM 계열의 설명 가능 인공지능(XAI) 방법론은 주로 모델의 공간적 주의를 관찰하고 시각화하는 데 초점을 두어 왔다. 이러한 접근은 입력 이미지에서 모델이 어디를 주목하는지를 이해하는 데는 효과적이지만, 해당 영역이 실제 예측의 원인으로 작용했는지를 규명하는 데에는 근본적인 한계가 있다.

기존 연구들은 활성화 맵의 강도, 형태, 엔트로피, 색상 분포 등을 사후적으로 정량 분석하여 모델 반응을 해석하는 방식을 채택해 왔으나, 이는 관찰 기반의 해석에 머물러 인과적 설명을 제공하는 데에는 제한적이다. 그러나 이러한 접근은 모델의 반응을 단순한 상관관계(correlation) 수준에서 해석하는 데 그쳤으며, 활성화된 영역을 왜 주목하는지, 즉 해당 시각적 특징이 모델의 예측에 인과적으로 기여는 지를 실험적으로 입증하지는 못하였다. 이에 본 연구는 관찰 중심 분석에서 개입 중심 검증(Intervention-based Validation)으로의 패러다임 전환을 제안한다. 즉, 모델의 활성화 반응을 단순히 시각화하는 것을 넘어, 특정 요인을 능동적으로 조작(intervention)함으로써 그 변화가 모델의 예측에 미치는 인과적 효과(causal effect)를 정량적으로 검증하는 접근이다. 이를 위해 본 장에서는 데이터, 색 공간, 주파수의 세 가지 요인을 상호 독립적인 분석 단위로 분리하고, 각 요인에 대해 웨이블릿 기반 개입(Wavelet-based Intervention)을 수행할 수 있는 색 공간-주파수-인식 반응 간의 인과적 매핑(causal mapping)을 구성하는 분석 프레임워크를 설계한다.

이 프레임워크는 단순히 모델을 관찰하는 데 그치지 않고, 모델이 의존하는 주파수 성분과 색채 요인에 직접 개입하여 그 결과를 정량적·시각적·인과적으로 규명한다는 점에서, 기존 XAI의 사후적·관찰적 한계를 극복하는 실험

적 방법론을 제공한다.

2) 제안 프레임워크의 구조

본 연구의 색상·주파수 개입 기반 인과 분석 프레임워크는 모델의 내부 반응을 단순히 관찰하는 수준을 넘어, 색상·주파수 성분을 단계적으로 조작하여 모델의 예측 변화를 인과적으로 해석하는 구조로 설계되었다. 프레임워크는 다음 세 가지 개념적 축을 중심으로 구성된다.

가) 색 공간 기반 시각 요인 분리

본 연구의 첫 번째 단계는 입력 이미지를 인지적 색채 공간으로 변환하여 명암, 색차, 색상, 채도와 같은 지각적 요인을 독립적으로 분리하는 것이다. 이 과정을 통해 모델이 인식하는 시각적 요인을 물리적·지각적 단위로 세분화할 수 있으며, 각 색 공간 채널은 이후 웨이블릿 변환을 통해 저주파(전역 구조 정보)와 고주파(세부 질감·에지) 성분으로 나뉜다. 이때 각 채널의 주파수별 에너지 분포를 바탕으로 정보량 맵(Attention Map)이 생성되며, 이는 모델이 반응할 가능성이 높은 시각적 단서(cue)를 공간적으로 어디에 집중하는지를 직관적으로 보여준다. 또한, 이 Attention Map을 원본 이미지 위에 중첩(overlay)하여 Applied Attention Map을 구성함으로써 색 공간–주파수–공간 위치 간의 연관성을 시각적으로 확인할 수 있는 기저 신호 분석[21] (baseline Frequency analysis) 단계가 완성된다.

나) 개입을 통한 인과 추론

두 번째 단계는 주파수 개입을 통해 모델의 예측 반응을 능동적으로 조작하고, 그 결과의 변화를 인과적으로 측정하는 것이다. 이 과정에서 연구자는 특정 주파수 대역을 선택하여 그 에너지나 진폭에 대한 가중치를 조정함으로

21) 기저선 주파수 분석 (Baseline Frequency Analysis): 신호의 기준선을 흔드는 저주파 노이즈(기저선 변동)의 주파수 성분을 분석하여 신호를 보정하는 것.

써 모델의 입력을 부분적으로 변형시킨다. 이때 모델은 동일한 객체를 입력받더라도, LAB 또는 HSV 색 공간으로 변환된 채널을 웨이블릿 변환, 역변환 과정을 거쳐 주파수 개입된 형태로 재구성한 이미지를 입력으로 사용하기 때문에, 주파수 성분의 조정 정도에 따라 예측 확률값과 활성화 맵의 공간적 분포가 서로 다르게 나타난다. 이는 단순한 반응의 차이가 아닌, 모델이 특정 주파수 요인에 인과적으로 반응하는지를 규명하는 실험적 근거로 해석된다. 이러한 접근은 XAI 연구의 기존 패러다임인 관찰 중심 분석[22]을 넘어 개입 중심 검증[23]으로 확장하는 시도라 할 수 있다.

다) 정량적 인과 정립 (Quantitative Causal Validation)

마지막 단계에서는, 앞서 수행된 주파수 개입의 강도와 모델의 예측 변화량 간의 관계를 정량적으로 분석함으로써 모델의 내재 된 주파수 선택성[24] 과 색채 기반 반응(color-based response)을 규명한다. 이를 통해, 모델이 어떤 주파수 대역에 더 민감하게 반응하는지, 또 특정 색 공간 요인이 결과 예측에 얼마나 큰 영향을 미치는지를 정량적 수치로 평가할 수 있다.

이 정량화 단계는 단순한 시각화 수준의 해석을 넘어, 모델의 인식 편향(bias)과 신뢰성을 인과적 측면(causal perspective)에서 규명할 수 있도록 한다. 따라서 본 연구의 프레임워크는 모델의 색공간과 주파수 반응 구조를 실험적으로 분석하고 인과적으로 해석할 수 있는 새로운 정량형 XAI 접근법을 제시한다.

3) 프레임워크의 전체 절차

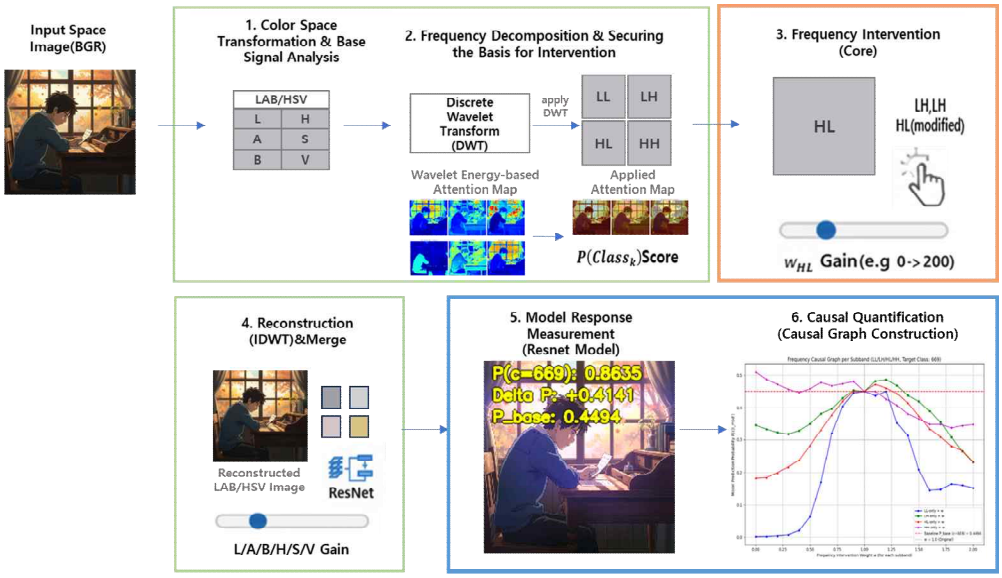
---

22) 관찰 중심 분석(Observation-based Analysis): 모델이나 시스템의 내부 반응을 단순히 관찰(observe)하고 추론(infer)함으로써 현상의 상관관계(correlation)를 해석하는 분석 접근방법.

23) 개입 중심 검증(Intervention-based Validation):입력 요인을 능동적으로 조작하여 모델의 출력 변화를 관찰함으로써 입력 요인과 예측 결과 간의 인과 관계를 정량적으로 검증하는 분석 방법

24) 주파수 선택성(Frequency Selectivity): 모델이 입력 이미지의 다양한 주파수 성분(LL, LH, HL, HH 중 특정 대역)에 대해 선택적으로 강하게 반응하거나 민감하게 반응하는 성질을 의미한다.

본 연구에서 제시하는 색상·주파수 개입 기반 인과 분석 프레임워크는 입력 이미지의 색 공간 분리부터 주파수 조작, 재구성, 모델 반응 측정, 인과적 정량화에 이르기까지의 전체 흐름을 분석 절차 관점에서 단계적으로 정리한 것이다. [그림 3-1]은 이러한 처리 과정을 일련의 실험 단계로 도식화한 것으로, 각 단계가 어떤 분석적 기능을 수행하는지를 총괄적으로 나타낸다.



[그림 3-1] 색상·주파수 개입 기반 인과 분석 프레임워크 흐름도

본 프레임워크는 입력 이미지의 색 공간 변환 및 기저 신호 분석, 주파수 분해 및 개입 기반 확보, 주파수 개입, 역변환, 모델 반응 측정, 인과 정량화의 흐름으로 구성된다. 각 단계는 모델의 반응을 색 공간·주파수에서 분리·조작함으로써 모델의 주파수 선택성과 인과적 반응 구조를 정량적으로 규명한다.

[표 3-1] 색상·주파수 개입 기반 인과 분석 프레임워크의 구성 요약

구분	처리단계	주요 내용
입력	Input Image (BGR/RGB)	원본 이미지는 실험의 기준 레퍼런스로 사용된다. 이후 단계에서 색동간 변환, 주파수 분해, 개입 실험의 입력으로 활용된다.
① 색 공간 변환 및 기저 신호 분석	Color-Space Transformation & Baseline Signal Analysis	입력을 LAB/HSV로 분해해 채널별 특성을 분리하고, 웨이블릿 에너지 Attention Map으로 정보량과 분포를 분석하여 개입할 핵심 신호 대역을 선정한다.
② 주파수 분해 및 개입 기반 확보	Frequency Decomposition & Securing the Basis for Intervention	각 색 공간 채널에 대해 이산 웨이블릿 변환(DWT) 을 적용하여 LL, LH, HL, HH의 4개 주파수 대역으로 분해한다. 각 대역의 에너지 벡터는 형태·질감·명 암 정보를 반영하며, 주파수 개입을 위한 기반 파라미터 ω_i의 기준값으로 정의된다.
③ 주파수 개입	Frequency Intervention	특정 대역(예: HL)의 웨이블릿 계수를 조작(증폭 또는 감쇠)함으로써 모델 입력의 주파수 특성을 능동적으로 변형한다. 이는 인과 실험의 핵심 단계로, 모델의 반응을 유발하는 주파수 성분을 직접 실험적으로 제어한다.
④ 역변환	Reconstruction (IDW)& Merge	개입된 웨이블릿 계수를 역변환(IDWT)하여 수정 이미지$I'_{\mathrm{mod}}(\omega)$ 를 재구성한다. 이 이미지는 모델이 개입된 주파수 정보를 반영한 새로운 입력으로 사용된다.
⑤ 모델 반응 측정	Model Response Measurement	재구성된 $I'_{\mathrm{mod}}(\omega)$이미지를 ResNet 모델 f_θ에 타깃 클래스 c에 대한 출력 확률을 산출한다.
⑥ 인과 정량화 (인과 그래프 도출)	Causal Quantification (ΔP-Based Causal Graph)	개입 강도ω와 예측 변화량 $\Delta P(c;\omega)$의 함수 관계를 통해 모델의 주파수 선택성과 색채 반응을 인과 그래프 $G_{causal}:\omega\mapsto P(c\|I'_{\mathrm{mod}}(\omega))$로 나타낸다.

제 2 절 인과 분석 절차 및 단계별 처리 과정

본 절에서는 제1절에서 제시한 색상·주파수 개입 기반 인과 분석 프레임워크를 실제 실험 과정에 적용하기 위한 구체적 분석 절차를 단계별로 기술한다. 앞 절에서 전체 흐름을 개념적으로 제시한 데에 이어, 본 절에서는 각 단계가 실제 실험 데이터에 대해 어떻게 수행되었는지를 색 공간 변환·주파수 분해·Score-CAM 적용·확률 변화 측정이라는 분석 중심 관점에서 구조적으로 설명한다.

1) 색 공간 변환 및 기저 신호 분석

본 단계는 입력 이미지의 시각적 요인(visual factors)을 인지적[25]으로 분리하여, 후속 주파수 개입 실험의 기반 신호(baseline signal)를 확보하기 위한 사전 분석 과정이다.

입력 이미지

$$I_{RGB} = \{I_R, I_G, I_B\}$$

는 먼저 인간의 시지각 특성을 반영하는 인지적 색채 공간(perceptual color space)으로 변환된다. 본 연구에서는 CIELAB(L*, a*, b*)와 HSV(Hue, Saturation, Value) 두 가지 색 공간을 동시에 적용하였다.

색 공간 변환 과정은 다음과 같이 정의된다.

$$I_{LAB} = C_{LAB}(I_{RGB}) = \{I_L, I_a, I_b\},$$
$$I_{HSV} = C_{HSV}(I_{RGB}) = \{I_H, I_S, I_V\}$$

이 변환을 통해 입력 신호는 명암(Luminance: L/V), 색차(Chromaticity: a, b), 색상(Hue: H) 및 채도(Saturation: S)로 분리된다.

이는 모델의 활성화 반응이 단순한 RGB 강도 변화가 아니라, 지각적 요인(perceptual factors)에 기반함을 정량적으로 분석하기 위한 기초 단계이다.

---

25) 인지적: 이미지를 인간의 시지각 체계에 따라 명암·색상·채도 등의 인지 단위로 분해한다. 라는 뜻

이후 각 채널(*L*, *A*, *B*, *H*, *S*, *V*)에 대해 이산 웨이블릿 변환(DWT, Discrete Wavelet Transform)을 적용하여 해당 채널의 저주파(LL) 및 고주파(LH, HL, HH) 성분을 분해한다.

$$DWT(I_x) = \{LL_x, LH_x, HL_x, HH_x\}$$

분해된 대역의 에너지 분포는 해당 채널이 지닌 정보 복잡도(Information Complexity)를 반영하며, 이를 색상 히트맵으로 시각화한 것이 정보량 맵(Attention Map)이다. Attention Map은 각 채널별로 다음과 같은 시각적 의미를 가진다.

[표 3-2] 색 공간별 채널 특성 및 고에너지 영역의 시각적 의미

색 공간	채널	의미	고에너지 영역의 특징
LAB	L	명암(밝기 구조)	물체의 윤곽·형태 구조
LAB	a	녹–적 축 색차	배경과 피사체의 색 대비
LAB	b	청–황 축 색차	하늘·모래 등 색온도 변화
HSV	H	색상	주요 물체의 고유색 영역
HSV	S	채도	질감 및 포화도 강조 영역
HSV	V	명도	조명·그림자 영향 구역

각 채널의 Attention Map은 웨이블릿 에너지

$$E_x = |LH_x| + |HL_x| + |HH_x$$

를 기반으로 계산하며, 이를 정규화하여 [0, 255]의 히트맵으로 표현한다.

이 히트맵은 시각적으로 정보가 집중된 부분을 드러낸다.

또한, 적용 맵(Applied Attention Map)은 이 에너지 맵을 원본 이미지 위에 오버레이(overlay) 하여 어떤 영역이 주로 활성화되는지를 직관적으로 보여준다. 예를 들어, LAB-B 채널의 Applied Attention Map에서 붉은색 히트맵이 창밖의 단풍 영역에 집중되어 있다면, 이는 해당 영역이 청–황(B) 축의 색상 정보를 가장 풍부하게 포함함을 의미한다. 이미지에 그림자가 포함된 경우, 그림자 내부의 인물은 Applied Attention Map에서 상대적으로 낮은 활성도를 보인다. 이는 그림자가 색 공간 내 명암 대비(Luminance contrast)를 약화시켜, 모델이 해당 영역을 시각적 주목 대상으로 인식하지 못하기 때문이다. 이 단계의 결과는 후속 주파수 개입 실험에서 어떤 채널과 대역이 모델 반응에 결정적으로 작용하는가를 탐색하기 위한 기저 신호로 활용된다.

Input

- M: 사전 학습 CNN 모델 (ResNet50)
- I_{orig}: 원본 입력 이미지 (RGB/BGR)
- T_{class}: Target Class
- *(optional)* $wavelet, level$: DWT 설정 (예: `db2`, $level = 1$)

Output

- $I_{\text{LAB}} = \{I_L, I_a, I_b\}$, $I_{\text{HSV}} = \{I_H, I_S, I_V\}$
- P_{base}: 기준 예측 확률
- *(optional)* A_X: 채널별 기저 에너지/어텐션 맵 $(X \in \{L, a, b, H, S, V\})$

Procedure

1. **색재 공간 변환**

$$I_{\text{LAB}} \leftarrow C_{\text{LAB}}(I_{\text{orig}}) = \{I_L, I_a, I_b\}$$
$$I_{\text{HSV}} \leftarrow C_{\text{HSV}}(I_{\text{orig}}) = \{I_H, I_S, I_V\}$$

2. **기준선(Baseline) 예측**

$$P_{\text{base}} \leftarrow M.\text{Predict}(I_{\text{orig}}, T_{\text{class}})$$

3. **(선택) 채널별 기저 신호 분석**

각 $X \in \{L, a, b, H, S, V\}$에 대해

$$\{LL_X, LH_X, HL_X, HH_X\} \leftarrow DWT(I_X; wavelet, level)$$
$$A_X \leftarrow \text{Energy}(LL_X, LH_X, HL_X, HH_X) \quad (\text{또는 Texture/Grad/Var})$$

필요 시 A_X를 I_{orig}에 오버레이하여 기저 분포를 시각 점검.

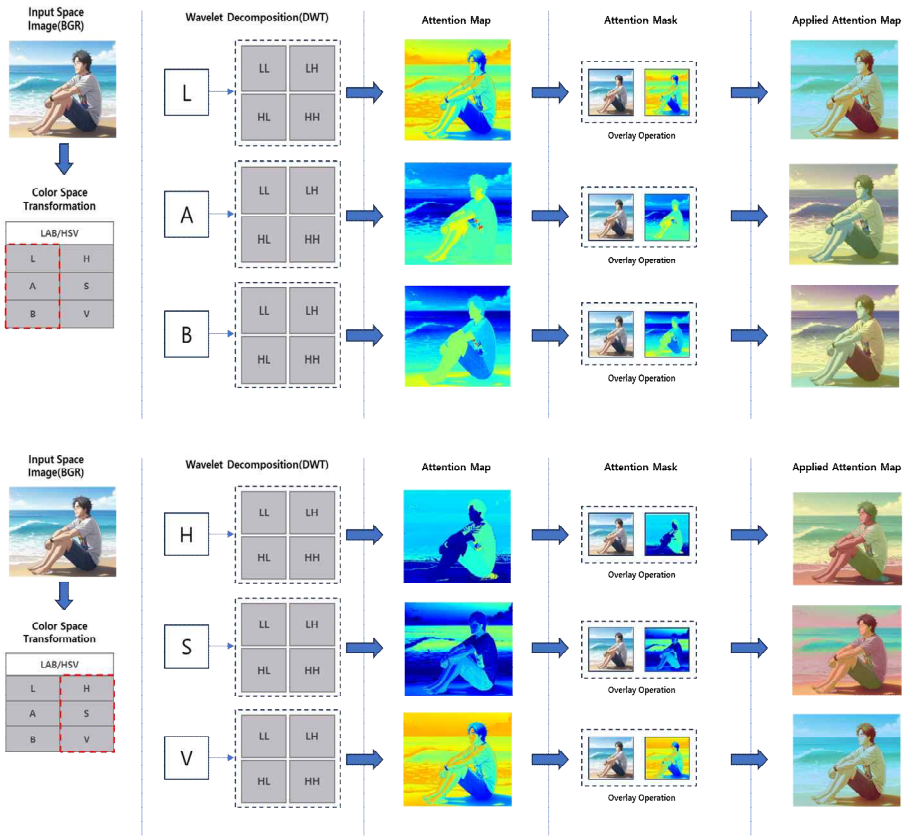
Return

$\{I_{\text{LAB}}, I_{\text{HSV}}, P_{\text{base}}, (A_X)\}$

[그림 3-2] 색 공간 기반 웨이블릿 정보량 맵 생성 기능 요약

[그림 3-2]은 입력 이미지의 색 공간 변환으로부터 각 채널의 주파수 정보를 추출하고 이를 바탕으로 정보량맵(Information Map)과 적용맵(Applied Attention Map)을 생성하는 절차를 정의한다. 우선 입력 이미지 I_{RGB}는 LAB 및 HSV 색 공간으로 변환되어 명암, 색차, 색상, 채도 요인으로 분리된다. 이후 각 채널 $I_x = \{I_L, I_a, I_b, I_H, I_S, I_V\}$에 대해 2차원 이산웨이블릿변환(2D-DWT)을 수행하여 저주파와 고주파 성분을 추출한다. 세 고주파 대역(LH,HL,HH)의 절댓값 합을 통해 해당 채널의 에너지 분포E_x를 계산하고, 이를 정규화하여 색상 히트맵으로 시각화함으로써 Attention Map을 생성한

다. 마지막으로 이 Attention Map을 원본 이미지 위에 중첩(overlay)하여 Applied Attention Map을 얻는다. 본 절차를 통해 각 색 공간 채널이 담고 있는 주파수 에너지의 공간적 분포, 즉 어디에 정보가 집중되어 있는가를 직관적으로 파악할 수 있으며, 이는 이후 단계(주파수 개입 및 인과 검증)의 기초 분석 신호로 활용된다.



[그림 3-3] 색 공간 기반 웨이블릿 정보량 맵 생성 구조도

[그림 3-3]은 입력 이미지(I_RGB)가 LAB 및 HSV 색 공간으로 변환된 후, 각 채널에 대해 웨이블릿 분해가 수행되어 Attention Map 및 Applied Attention Map이 생성되는 과정을 보여준다. 각 색 공간의 채널별 웨이블릿 에너지 분포를 시각화함으로써, 모델이 잠재적으로 반응할 가능성이 높은 정보 영역을 사전 탐색한다.

2) Score-CAM 적용 및 주파수 개입 설계

분석된 색 공간 및 주파수 기저 신호를 바탕으로, 모델이 입력 이미지의 색상·명암·질감 요소에 어떻게 반응하는지를 Score-CAM을 통해 시각화하고, 이에 웨이블릿 기반 주파수 개입을 적용하여 모델의 주파수 선택성과 색채 민감도를 인과적으로 측정하는 절차를 설계한다. 이 과정은 단순한 시각화가 아니라, 입력 이미지의 주파수 성분을 실험적으로 조작하여 모델의 예측 확률 변화($\Delta P(c,\omega)$)를 정량적으로 측정하고, 그 결과를 인과적 설명 가능성(causal explainability)의 근거로 확립하기 위함이다. 입력으로 사용되는 복원 이미지 I_{Recon}은전 절에서 수행된 IDWT(Inverse Discrete Wavelet Transform)에 의해 HSV 및 LAB 색 공간의 채널별로 재구성된 이미지이다. 이 복원 이미지를 CNN 모델f_θ의 입력으로 사용하여 타깃 클래스c에 대한 Score-CAM 맵을 생성한다.

$$A^{(c)}(x,y) = ScoreCAM(f_\theta, I_{Recon}, class = c)$$

Score-CAM은 그래디언트를 사용하지 않고, 출력 점수(score)의 변화량을 이용하여 각 특징 맵의 중요도를 계산한다. 그 결과$A^{(c)}(x,y)$는 모델이 특정 클래스 인식 시 공간적으로 주목하는 영역을 나타낸다.

이 활성화 맵은 다시 웨이블릿 변환(DWT)을 통해 저주파(LL) 및 고주파(LH, HL, HH) 대역으로 분해된다.

$$A^{(c)}(x,y) \xrightarrow{DWT} \{LL, LH, HL, HH\}$$

각 주파수 대역에 대해 가중치 벡터$\omega = (\omega_{LL}, \omega_{LH}, \omega_{HL}, \omega_{HH})$를 정의하고, 개입(intervention)을 수행한다.

$$\omega_i \in [0.5, 2.0]$$

- $\omega_i = 1.0$: 개입 없음 (Baseline)
- $\omega_i < 1.0$: 감쇠 (Attenuation)
- $\omega_i > 1.0$: 증폭 (Amplification)

이 과정을 통해 모델 입력의 주파수 성분을 조작하고, 각 대역의 증폭·감쇠가 모델의 신뢰도 변화에 미치는 영향을 실험적으로 측정한다. 조정된 주파

수 계수는 IDWT를 통해 복원되어 Frequency-weighted Score-CAM Map으로 구성된다.

$$A_f^{(c)}(x,y) = IDWT(\omega_{LL} \cdot LL, \omega_{LH} \cdot LH, \omega_{HL} \cdot HL, \omega_{HH} \cdot HH)$$

이때 모델의 신뢰도 변화량은 다음과 같이 계산된다.

$$\Delta P(c,\omega) = P(c|A_f^{(c)}) - P(c|I_{Recon})$$

이 값이 양수이면 특정 주파수 대역의 증폭이 모델의 인식 강화를 의미하며, 음수이면 해당 대역의 감쇠가 인식을 저해했음을 뜻한다. 즉, $\Delta P(c,\omega)$는 모델이 각 주파수 성분에 대해 가지는 반응 강도의 정량 지표로 활용된다.

Input

- I_{LAB} 또는 I_{HSV}
- X_{target}: 개입 대상 채널 (예: L 또는 V)
- $W_{\text{set}} = \{w_k\}$: 주파수 개입 벡터 집합
  $w_k = (w_{LL}, w_{LH}, w_{HL}, w_{HH})$ 또는 단일 대역 스칼라 w_b
- $wavelet, level$: DWT 설정
- T_{class}, M *(선택: 즉시 응답 측정 시)*

Output

- 개입 이미지 집합 $\{I'_{\text{mod}}(w_k)\}$
- *(선택)* $\{P_{\text{freq}}[k]\}$: 주파수 개입에 대한 모델 반응
- *(선택)* $\Delta P_{\text{freq}}[k] = P_{\text{freq}}[k] - P_{\text{base}}$

Procedure

1. **대상 채널 분해 (DWT)**

$$\{LL_X, LH_X, HL_X, HH_X\} \leftarrow DWT(I_{X_{\text{target}}}; wavelet, level)$$

2. **개입 설계 및 적용**
   각 $w_k \in W_{\text{set}}$에 대해

$$LL'_X = w_{LL} \cdot LL_X, \quad LH'_X = w_{LH} \cdot LH_X,$$
$$HL'_X = w_{HL} \cdot HL_X, \quad HH'_X = w_{HH} \cdot HH_X$$

   (단일 대역 실험이면 해당 w_b만 변동, 나머지는 1.0 고정)

3. **IDWT 복원 및 채널 병합**

$$I'_{X_{\text{target}}}(w_k) \leftarrow IDWT(LL'_X, LH'_X, HL'_X, HH'_X)$$

$$I'_{\text{mod}}(w_k) \leftarrow \begin{cases} LAB2RGB(\{I'_L, I_a, I_b\}) & (X_{\text{target}} = L) \\ HSV2RGB(\{I_H, I_S, I'_V\}) & (X_{\text{target}} = V) \\ \text{해당 색공간에 맞게 병합} & (\text{그 외 채널}) \end{cases}$$

4. **(선택) 즉시 모델 반응 측정**

$$P_{\text{freq}}[k] \leftarrow M.\text{Predict}(I'_{\text{mod}}(w_k), T_{\text{class}}), \quad \Delta P_{\text{freq}}[k] \leftarrow P_{\text{freq}}[k] - P_{\text{base}}$$

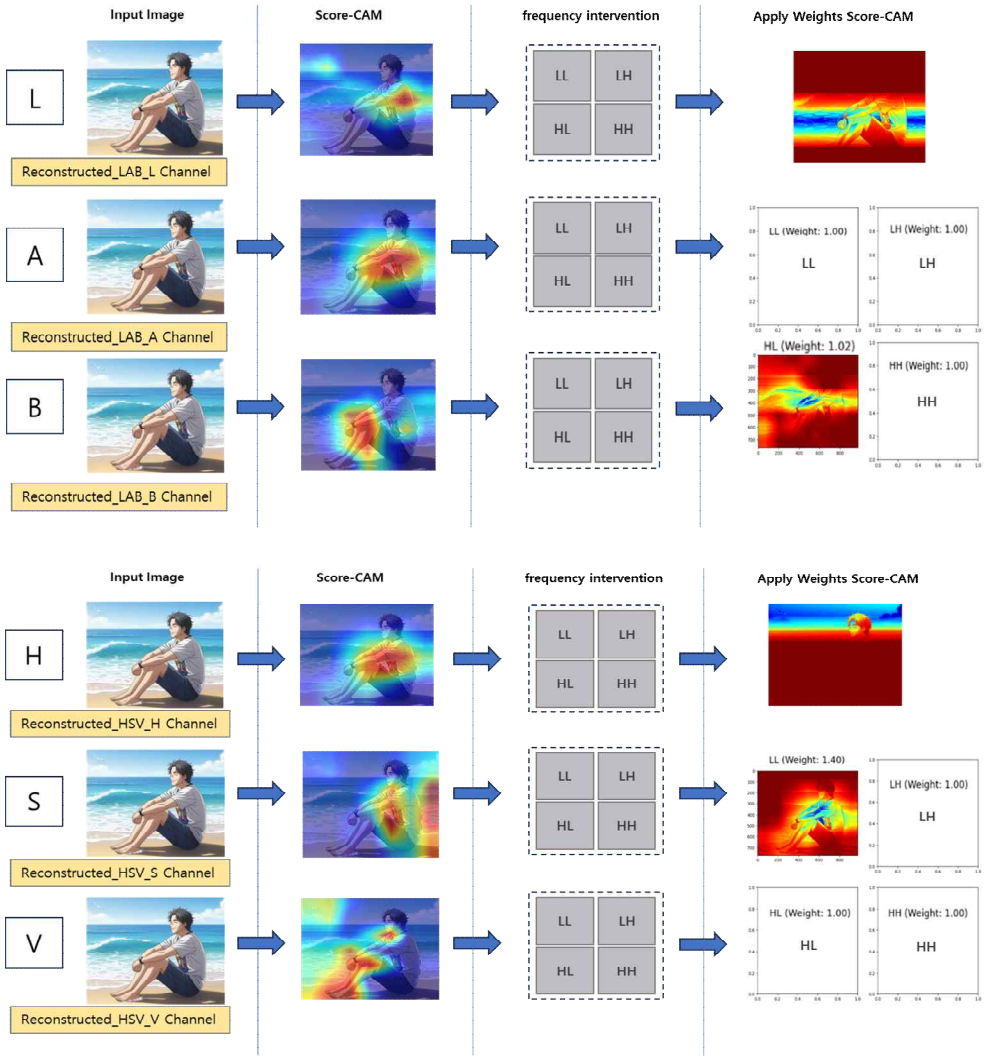
Return

$\{I'_{\text{mod}}(w_k)\}$ *(+ 필요 시* $\{P_{\text{freq}}[k], \Delta P_{\text{freq}}[k]\}$*)*

[그림 3-4] Score-CAM 기반 주파수 개입 절차 기능 요약

[그림 3-4]는 제안된 색상·주파수 개입 기반 인과 분석 프레임워크에서 모델의 시각적 반응을 주파수 단위로 조작하기 위한 핵심 절차를 나타낸다. 이 절차의 입력은 웨이블릿 변환과 역변환을 통해 복원된 색 공간별 채널 이미지로, LAB Channel image, HSV Channel image를 각각 실험의 기반 입력으로 사용된다. 이들 채널 이미지는 모델이 인지적으로 해석할 수 있는 명암(L), 색차(a, b), 색상(H), 채도(S), 명도(V) 등 지각 요인(perceptual factor)을 반영한 형태로 복원된 이미지이며, 이를 통해 모델의 색 공간 기반 반응을 주파수 수준에서 독립적으로 분석할 수 있다. 먼저, 입력된 재구성 이미지(I_Recon)에 대해 Score-CAM을 계산하여 타깃 클래스 c에 대한 활성화 맵 A(c)을 생성한다. 이 활성화 맵을 이산 웨이블릿 변환를 적용하여 저주파와 고주파 대역으로 분해하면, 영상의 전역 구조와 세부 텍스처 정보를 분리하여 모델의 반응 기반을 주파수별로 분해할 수 있다. 이후 각 대역의 에너지 강도를 조절하기 위해 가중치 벡터 $\omega = (\omega_{LL}, \omega_{LH}, \omega_{HL}, \omega_{HH})$를 정의한다.

각 성분ω_i는 [0,2.0] 범위 내에서 증폭 또는 감쇠되며, 특정 주파수 성분이 모델의 인식 반응에 미치는 인과적 영향을 실험적으로 검증할 수 있다. 가중치가 적용된 웨이블릿 계수는 역변환을 통해 새로운 주파수-가중 활성화 맵 $A_f^{(c)}$로 재구성된다. 이 맵은 모델이 실제로 어떤 주파수 성분에 더 큰 주의(attention)를 두는지를 시각적으로 드러낸다. 마지막으로, 이렇게 재구성된 I'_{mod} 맵은 모델의 주목 영역(Attention)이 어떤 주파수 성분으로 구성되어 있는지를 시각적으로 분석하는 근거가 된다. 이 시각적 분석은 제3절에서 수행할 입력 이미지(I) 자체의 주파수 및 색상 채널을 직접 개입하여 예측 확률을 정량적으로 측정하는 본질적인 인과 관계 실험의 해석을 보조한다. 따라서 본 알고리즘은 모델의 주파수 선택성을 시각적으로 진단하고, 제3절의 정량적 인과 분석을 위한 해석적 기반을 제공하는 시각적 분석 모듈(visual analysis module)로서 기능을 제공한다.



[그림 3-5] Score-CAM 기반 주파수 개입 절차 구조도

[그림 3-5]는 [그림 3-4]에서 제시한 Score-CAM 기반 주파수 개입 절차(Frequency-weighted Score-CAM Procedure)의 전체 처리 흐름을 도식화한 것이다. 입력 단계에서는 웨이블릿 변환과 역변환을 거쳐 복원된 LAB Channel 이미지와 HSV Channel 이미지가 모델의 입력으로 사용된다. 각 채널 이미지는 Score-CAM을 적용하여 타깃 클래스c에 대한 활성화 맵$A(c)$을 산출하며, 이 활성화 맵은 이산 웨이블릿 변환를 통해 저주파 및 고주파 성분으로 분해된다. 각 대역은 영상의 구조적 형태, 경계, 질감 등 상이한 시각적

요인을 반영하며, 모델이 주목하는 주파수별 반응의 근거를 분리하여 분석할 수 있다. 이후 특정 주파수 대역에 대한 가중치 $\omega_i \in [0.5, 2.0]$를 설정함으로써 해당 주파수 성분을 증폭 또는 감쇠시키는 주파수 개입을 수행한다. 조정된 주파수 계수는 역 웨이블릿 변환을 통해 새로운 주파수-가중 활성화 맵 $A_f^{(c)}$으로 재구성되며, 이는 개입 강도에 따라 변화하는 모델의 반응 특성을 시각적으로 나타낸다. 재구성된 $A_f^{(c)}$을 모델 f_θ다시 입력하여 예측 확률의 변화량 $\Delta P(c, \omega)$을 계산함으로써 개입 강도 w와 모델의 출력 확률 $P(c|I'_{\text{mod}}(\omega))$ 간의 인과 관계를 정량적으로 도출한다. 이 관계는 우측의 인과 그래프(Causal Graph)로 시각화되어, 모델이 특정 주파수 성분에 얼마나 민감하게 반응하는지를 명확히 보여준다. 따라서 본 그림은 Score-CAM의 공간적 주의 시각화를 주파수 영역 분석으로 확장하여, 색 공간 – 주파수 – 활성화 반응 간의 인과적 연계 구조를 실험적으로 규명하는 본 연구 프레임워크의 핵심 절차를 직관적으로 제시한 것이다.

3) 모델 반응 측정 및 인과 정량화

제안한 색상·주파수 개입 기반 인과 분석 프레임워크의 마지막 단계로서, 모델의 반응을 정량적으로 측정하고, 색채 및 주파수 개입에 따른 인과적 관계를 규명하였다. 이는 단순히 모델이 특정 주파수 대역이나 색채 채널(L,a,b)에 민감하다는 사실을 관찰하는 수준을 넘어, 두 요인이 상호작용하며 모델의 예측 확률에 어떠한 영향을 미치는지를 인과적으로 검증하는 절차이다. 이를 위해 (1) 색채 개입을 통한 모델의 예측 확률 변화를 측정하고, (2) 주파수 개입을 통한 반응 변화를 함께 실험하며, (3) 양자의 상관 및 인과적 상호조절 관계를 분석한다. 이 과정을 통해 모델이 색채 조건(color context)에 따라 주파수 선택성을 달리 보이는지, 즉, 색채가 주파수 반응의 조절 변수(modulatory variable)로 작용하는지를 정량적으로 검증하였다. 이러한 인과적 정량화는 기준선 확률(P_{base}), 개입 후 확률 (P_{mod}),그리고 이 둘의 차이인 영향력(Impact, ΔP)의 세 가지 핵심 지표를 통해 수행된다.

가) 기준선 측정

인과적 영향력을 측정하기 위해서는, 먼저 어떠한 개입도 이루어지지 않은 상태의 기준 확률(baseline probability)이 필요하다. 본 연구에서 기준 확률 P_{base}는 원본 입력 이미지I_{orig}를 ImageNet 사전 학습된 ResNet50 모델 M에 입력하여, 해당 모델이 분석 대상 클래스T_{class}에 대해 산출한 초기 확신도(initial confidence) 로 정의하였다. 이 값은 모델이 현재 입력 이미지를 해당 클래스라고 인식하는 확률을 의미하며, 이후 수행되는 모든 색채·주파수 개입 실험의 비교 기준으로 사용된다. 즉 P_{base}는 인과적 개입 전후의 변화를 정량화하기 위한 출발점(ground truth)으로 기능한다.

나) 개입 후 반응 측정

P_{mod} 는 사용자가 LAB, HSV, 또는 주파수 슬라이더를 조작하여 생성한 개입된 이미지I_{mod}를 동일한 모델 및 Target Class에 재입력하여 얻은 개입 후 확률값이다. 이 값은 사용자의 슬라이더 조작에 따라 이 값은 색채 또는 주파수 요인에 대한 개입 강도 조정에 따라 모델이 보이는 확률적 반응 변화를 즉각적으로 반영한다. 이를 통해 각 개입 요인이 모델의 예측 신뢰도에 미치는 즉각적 영향(Immediate Effect)을 시각적으로 비교·분석할 수 있으며, 이러한 반응 곡선은 이후 인과적 영향력ΔP계산의 실증적 근거로 활용된다.

다) 인과적 영향력 정량화

본 연구의 핵심 지표인ΔP(델타 P)는 기준선 확률과 개입 후 확률의 차이로 정의된다.

$$\Delta P = P_{\text{mod}} - P_{\text{base}}$$

ΔP는 사용자의 개입(원인)이 모델의 확신도(결과)에 미친 순수한 인과적 영향력(Impact) 을 정량적으로 나타낸다.

- $\Delta P > 0$: 개입이 모델의 기존 예측을 강화함을 의미
- $\Delta P < 0$: 개입이 모델의 기존 예측을 약화 또는 반박함을 의미

본 연구는 이ΔP를 기반으로 색채–주파수 상호 조절(Causal Interaction)

을 분석하였다. 즉, 주파수만 개입한 경우 I_{freq}에서의 ΔP_F와, 색채가 먼저 개입된 상태가 먼저 개입된 상태 I_c에 동일한 주파수 개입을 추가한 경우 $I_c + I_{freq}$의 $\Delta P_{F/C}$를 비교함으로써, 색채가 주파수 반응의 조절 요인으로 작용하는지를 검증하였다. 이로써 모델의 반응은 단일 요인(색채 또는 주파수)의 민감도에 국한되지 않고, 양자의 상호작용(interaction)에 의해 조건부인과 구조(Conditional Causal Structure)를 형성함이 확인되었다.

Input

- M: 사전 학습 CNN 모델 (ResNet50)
- I_{orig}: 원본 입력 이미지
- T_{class}: 분석 대상 Target Class
- w_c: 색채 개입 파라미터 (예: S-shift = 0 → Grayscale)
- w_f: 주파수 개입 파라미터 (예: $w_{LL} = 0.5$)

Output

- ΔP_F: 주파수 개입 F의 단독 인과 영향력 ($\text{Ref} \rightarrow F$)
- $\Delta P_{F|C}$: 색채 조건 C 하에서의 주파수 개입 F 영향력 ($C \rightarrow C + F$)
- $Interaction$: 상호조절 효과 크기 ($\Delta P_{F|C} - \Delta P_F$)

Procedure

1. **기준선 (Reference) 측정**

$$P_{\text{ref}} \leftarrow M.\text{Predict}(I_{\text{orig}}, T_{\text{class}})$$

2. **주파수 개입 F의 단독 효과**

$$I_F \leftarrow \text{Modify}_{\text{Frequency}}(I_{\text{orig}}, w_f), \quad P_F \leftarrow M.\text{Predict}(I_F, T_{\text{class}}), \quad \Delta P_F \leftarrow P_F - P_{\text{ref}}$$

3. **색채 개입 C 적용 (문맥 생성)**

$$I_C \leftarrow \text{Modify}_{\text{Color}}(I_{\text{orig}}, w_c), \quad P_C \leftarrow M.\text{Predict}(I_C, T_{\text{class}})$$

4. **색채–주파수 결합 개입 $C + F$**

$$I_{C+F} \leftarrow \text{Modify}_{\text{Frequency}}(I_C, w_f), \quad P_{C+F} \leftarrow M.\text{Predict}(I_{C+F}, T_{\text{class}})$$

5. **인과적 상호조절 정량화**

$$\Delta P_{F|C} \leftarrow P_{C+F} - P_C, \quad Interaction \leftarrow \Delta P_{F|C} - \Delta P_F$$

Return

$\{\Delta P_F, \Delta P_{F|C}, Interaction\}$

[그림 3-6] 인과적 상호 조절 정량화 기능 요약

[그림 3-6]은 색채 개입(Color Intervention)과 주파수 개입(Frequency Intervention)이 모델의 예측 확률에 미치는 조건부인과 영향력(Conditional Causal Effect) 을 정량적으로 평가하기 위해 설계되었다. 즉, 모델이 단일 요인(주파수)뿐 아니라, 다른 요인(색채) 의 존재 여부에 따라 동일한 주파수 자

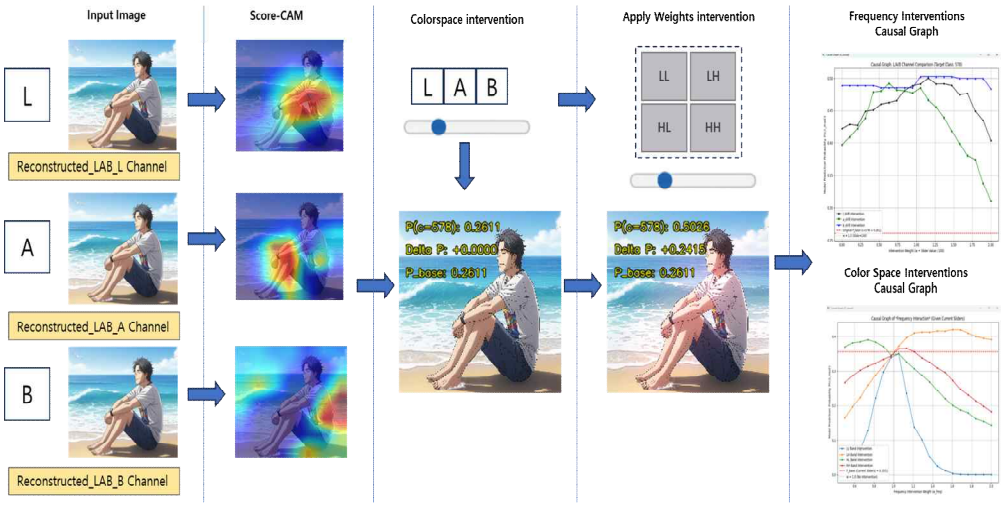
극에 대해 얼마나 상이한 반응을 보이는지를 분석함으로써, 모델 내부의 상호 조절 관계를 규명하는 데 목적이 있다. 우선, 개입이 없는 상태에서 모델의 참조 확률 P_{base}을 측정하고, 이를 인과적 비교의 기준(reference point)으로 설정한다. 이후 주파수 개입만을 수행하여 모델의 확률 변화를 계산하고 이를 단독 주파수 효과(ΔP_F)로 정의한다.

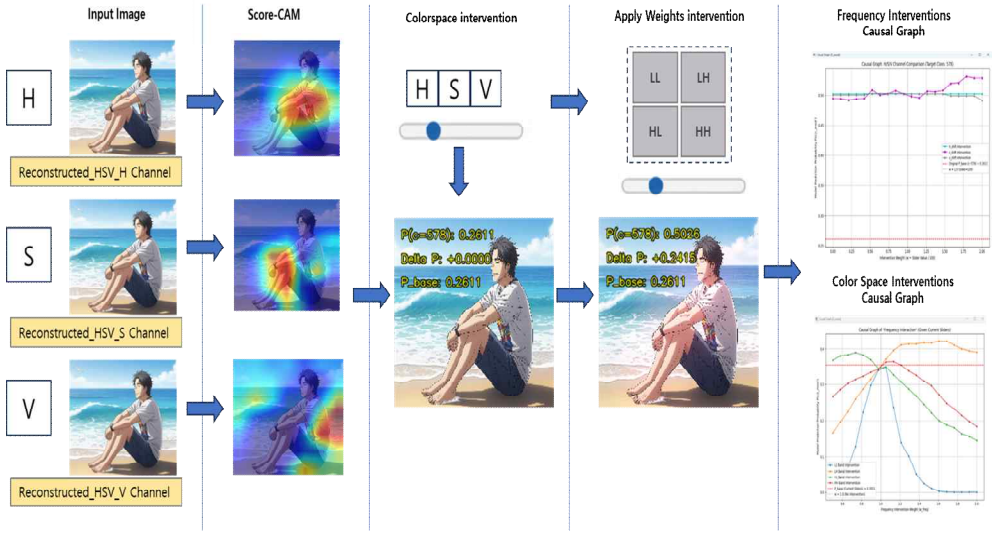
다음으로, 동일한 이미지에 색채 개입을 선행 적용하여 색채 조건이 조성된 상태에서 주파수 개입을 추가함으로써 색채 조건 하에서의 주파수 반응 ($\Delta P_{F/C}$)을 측정한다.

이 두 값의 차이, 즉

$$\text{Interaction} = \Delta P_{F/C} - \Delta P_F$$

은 색채 요인이 주파수 반응에 미치는 조절 효과(Modulatory Effect) 의 크기를 의미한다. 만약 Interaction > 0이면, 색채 조건이 주파수 개입의 효과를 증폭(enhance)시키는 방향으로 작용함을 의미하고, Interaction < 0이면, 색채 조건이 주파수 개입의 효과를 감쇠(suppress)시키는 조절 요인으로 작용함을 의미한다. 따라서 본 알고리즘은 단순한 주파수 민감도 분석을 넘어, 색채와 주파수 요인이 결합될 때 나타나는 조건부 반응 구조(Conditional Response Structure)를 인과적으로 해석하기 위한 정량적 절차이다. 이 결과를 통해 모델의 예측 메커니즘이 단일 요인에 의한 직접적 반응이 아닌, 다중 요인 간 상호작용에 기반한 복합적 인식 구조임을 실험적으로 개입을 통해 입증한다.





[그림 3-7] 색상·주파수 개입 기반 모델 반응 및 인과 그래프 도식화

[그림 3-7]은 [그림 3-6]에서 제시한 인과적 상호 조절 정량화 절차의 시각적 개요를 나타낸 것이다. 왼쪽에서부터 원본 입력 이미지가 LAB 색채 공간으로 분리되어 L,A,B 각 채널의 재구성 이미지가 생성되고, 각 채널별로 Score-CAM을 적용하여 모델이 주목하는 영역(활성화 맵)을 확인한다. 이후 색채 개입(Color-space Intervention) 단계에서는 사용자가 선택한 채널(L, A, B)에 개입 강도를 조정하여 해당 채널의 색채를 변화시키며, 주파수 개입 단계에서는 웨이블릿 서브밴드에 대한 가중치를 조정하여 주파수 성분을 조작한다. 이 두 개입의 조합 결과로 모델의 예측 확률 $P(C)$ 및 변화량 ΔP가 산출되며, 이를 기반으로 색채 개입 인과 그래프(Color-space Causal Graph)와 주파수 개입 인과 그래프가 도출된다. 이러한 시각적 절차는 모델이 색채 문맥(color context)에 따라 주파수 반응(frequency response)을 어떻게 조절(modulate)하는지를 직관적으로 보여준다. 추가적으로, HSV 색 공간에서 V 채널의 값을 감소시켜 색상을 어두운 계열로 변화시킬 경우, 모델의 예측 확률 $P_{\mathrm{mod}}(P(C))$이 점진적으로 감소하는 경향을 보였다.

이는 모델이 입력 이미지의 밝기 정보(luminance information) 에 강하게 의존하고 있으며, 명도 수준이 낮아질수록 타깃 클래스의 시각적 특징(형태,

윤곽, 질감) 을 안정적으로 인식하지 못함을 시사한다. 다시 말해, Value 채널의 감소는 주파수 응답뿐 아니라 모델의 활성화 맵(Score-CAM)에서도 관심 영역의 집중도를 약화하는 방향으로 작용하며, 결과적으로 모델의 확률적 신뢰도($P(C)$)역시 함께 저하되는 것이다. 이러한 현상은 색채 개입(Color Intervention) 중에서도 밝기(Value) 요인이 모델 반응의 주요 조절 변수로 작용함을 보여준다.

제 4 장 색상·주파수 개입 기반 인과 관계 분석 실험 및 결과

제 1 절 인과 관계 분석을 위한 실험 환경

1) 실험 환경 개요

본 연구에서 제안한 색상·주파수 개입 기반 인과 분석 기법을 실험적으로 검증하기 위해, ResNet 50 모델을 대상으로 입력 조작과 모델 반응을 정량적으로 측정할 수 있는 분석 환경을 구성하였다. 이 분석 환경은 사용자가 밝기·색채(LAB/HSV)와 웨이블릿 기반 주파수 성분을 단계적으로 조정할 수 있도록 하며, 각 개입이 모델의 예측 확률과 Score-CAM 활성화 구조에 어떠한 변화를 유발하는지를 즉시 관찰할 수 있도록 설계되었다. 본 장에서는 해당 분석 환경의 구성 요소와 실험 절차를 객관적으로 설명하고, 이를 통해 제안한 인과 구조 분석 방법이 실제로 어떻게 적용되는지를 정량적 실험 결과와 함께 제시한다. 특히 색상 및 주파수 조작에 따른 신뢰도 변화, 활성화 맵 재구성 패턴, 주파수 반응 민감도 등을 체계적으로 비교함으로써, 앞 장에서 제시한 인과적 메커니즘을 실험적으로 검증한다.

2) 분석 도구 구성 및 역할

본 연구에서 수행한 색상·주파수 개입 실험은 특정 입력 속성의 조정이 ResNet 50 모델의 예측 신뢰도 및 활성화 구조에 미치는 영향을 정량적으로 파악하기 위함이다. 이를 위해 실험 환경은 다음의 네 가지 분석 도구로 구성되며, 각 도구는 개입-반응-정량화 과정을 체계적으로 수행하기 위한 최소한의 기능을 제공한다. (1) 입력 속성 개입 도구는 입력 이미지를 LAB 및 HSV 색 공간으로 변환하고, 각 색상 채널의 값을 연속적으로 조정할 수 있는 기능을 제공한다. 또한 2D-DWT를 이용해 서브밴드 계수를 추출하고, 각

대역의 가중치를 독립적으로 변화시키는 주파수 개입 기능을 포함한다. 이 도구는 입력 조작을 수행하는 핵심 모듈로서, 색상·주파수 변화가 모델의 반응에 미치는 인과적 영향을 실험적으로 측정할 수 있는 기반을 제공한다. (2) 모델 반응 측정 도구는 개입된 입력 이미지는 ResNet 50 모델에 재입력되어 예측 확률, 예측 확률 변화량(ΔP), Score-CAM 기반 활성화 맵 변화가 산출된다. 이 측정 도구는 개입을 통해 조작된 입력이 모델의 내부 특징 표현 및 최종 출력에 미치는 영향을 정량적 지표와 시각적 분석을 통해 동시에 측정하는 역할을 수행한다. (3) 시각적 모니터링 도구는 색상 조정 후의 이미지, 주파수 대역 조정 결과, Score-CAM 활성화 맵, 예측 확률 변화 등을 병렬적으로 제시하여 입력 변화 → 모델 반응 변화를 직관적으로 이해할 수 있도록 지원한다. 이 도구는 실험 결과의 실시간 시각적 피드백을 제공하는 보조 기능으로, 개입-반응 구조를 명확히 인지할 수 있도록 도와준다. (4) 인과 그래프 생성 도구는 색상 또는 주파수 성분을 일정 범위로 연속적으로 변화시키면서 예측 확률 변화량을 기록하고, 이를 기반으로 인과 그래프를 생성한다. 이 도구는 본 연구에서 사용된 실험 이미지에 대해 색상·주파수 조정 범위에서 관찰된 신뢰도 변화의 패턴과 그 인과적 영향을 시각적으로 확인할 수 있도록 하며, 수행한 개입 실험 결과를 일관된 형식으로 요약하여 제시하는 역할을 수행한다.

3) 실험 데이터 및 예측 대상 설정

본 연구의 실험은 ImageNet 데이터셋으로 사전 학습된 ResNet 50 모델을 대상으로 수행하였다. 실험에는 ImageNet Validation Set 중 밝기·색채 변화 및 주파수 성분 변형에 따른 모델 반응 차이를 명확히 관찰할 수 있도록 시각적 속성 변화에 민감하게 반응하는 클래스를 선정하여 사용하였다.

4) 분석 도구 아키텍처 개요

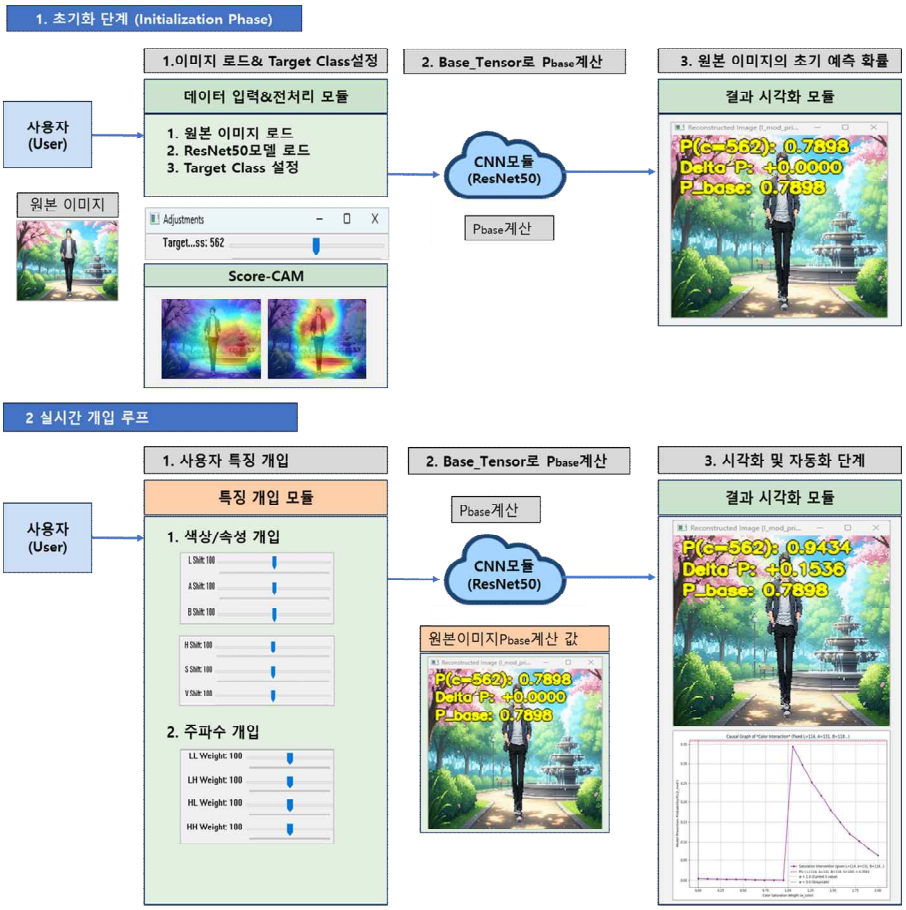
가) 분석 도구의 구성 요소

본 연구에서 수행한 색상·주파수 개입 실험을 효율적으로 실행하고, 개입에 따른 모델 반응 변화를 관찰하기 위해 보조 분석 도구를 구성하였다. 이는 실험을 위한 편의적 인터페이스이며, 4장에서 제시하는 인과 분석 결과를 도출하기 위한 지원 역할을 수행한다. 각 구성 요소는 다음과 같다.

[표 4-1] 전체 분석 도구 구성 기능 요약

모듈명	주요 기능	설명
(1) 입력 조작 모듈	색상·주파수 개입 수행	- LAB/HSV 색상 채널 조정 (L, A, B, H, S, V) - 웨이블릿 서브밴드 가중치 조정 - 밝기·채도·색차 및 저주파·고주파 성분의 단계적 조정 가능 - 실험자가 설정한 개입량에 따라 수정된 입력 이미지 생성
(2) 모델 반응 측정 모듈	모델 신뢰도 변화 및 Score-CAM 반응 산출	- 개입된 입력을 ResNet 50 모델에 재입력하여 예측 확률 변화 (ΔP) 계산 - Score-CAM 반응 변화를 산출하고 개입 전·후 차이를 비교 - 개입-반응 관계를 시각적·수치적으로 확인할 수 있도록 출력
(3) 시각화 및 인과 그래프 생성 모듈	실험 결과 시각화 및 인과 패턴 분석	- 개입 전·후 Score-CAM 맵을 시각적으로 비교 - 색상 및 주파수 조정에 따른 점진적 변화 경향 시각화 - 색상 인과 그래프 생성 - 주파수 인과 그래프 생성 - 각 그래프는 ΔP 변화를 기반으로 개입-반응의 인과 패턴을 요약하여 분석에 활용

나) 색상·주파수 개입 실험의 처리 흐름 및 분석 절차



[그림 4-1] 색상·주파수 개입 기반 인과 분석 실험의 전체 처리 흐름

[그림 4-1]은 본 연구에서 수행한 색상·주파수 개입 실험의 전체 절차를 단계적으로 요약한 것이다. 실험은 먼저 기준 신뢰도를 확보하는 과정으로 시작된다. 원본 이미지와 타깃 클래스를 지정하고, 모델이 개입되지 않은 입력에 대해 산출하는 기준 신뢰도(P_{base})와 초기 Score-CAM 반응을 확보한다.

이는 이후 개입 실험에서 비교 기준점(reference point)으로 사용된다. 두 번째 단계에서는 밝기·색채 또는 주파수 성분에 대한 조정값을 변화시키며, 해당 변화가 모델의 예측 확률과 Score-CAM 활성화 구조에 미치는 영향을 연속적으로 측정한다. 각 조정값에 따른 ΔP 및 활성화 변화는 즉시 반영되어 시각적으로 확인할 수 있으며, 이를 통해 입력 속성의 변화가 모델 반응으로 어떻게 전파되는지를 인과적 관점에서 추적할 수 있다. 이와 같은 단계적

실험 절차를 통해 색상·주파수 성분이 모델 신뢰도 변화에 미치는 영향을 체계적으로 비교할 수 있으며, 본 장 후반부에서 제시하는 색상 인과 그래프와 주파수 인과 그래프 생성의 기초 자료로 활용된다.

제 2 절 분석 도구 및 구성 요소

본 연구에서 수행한 색상·주파수 개입 실험은 Python 기반의 분석 환경에서 진행되었다. 본 연구에서는 ImageNet으로 미리 학습된 ResNet 50이며, 모든 개입 실험은 동일한 모델 파라미터를 유지한 상태에서 수행하였다. 색상 변환, 웨이블릿 기반 주파수 분해 처리는 표준 영상 처리 라이브러리를 활용하여 구현하였으며, 모델의 반응 측정에는 PyTorch 프레임워크를 사용하였다.

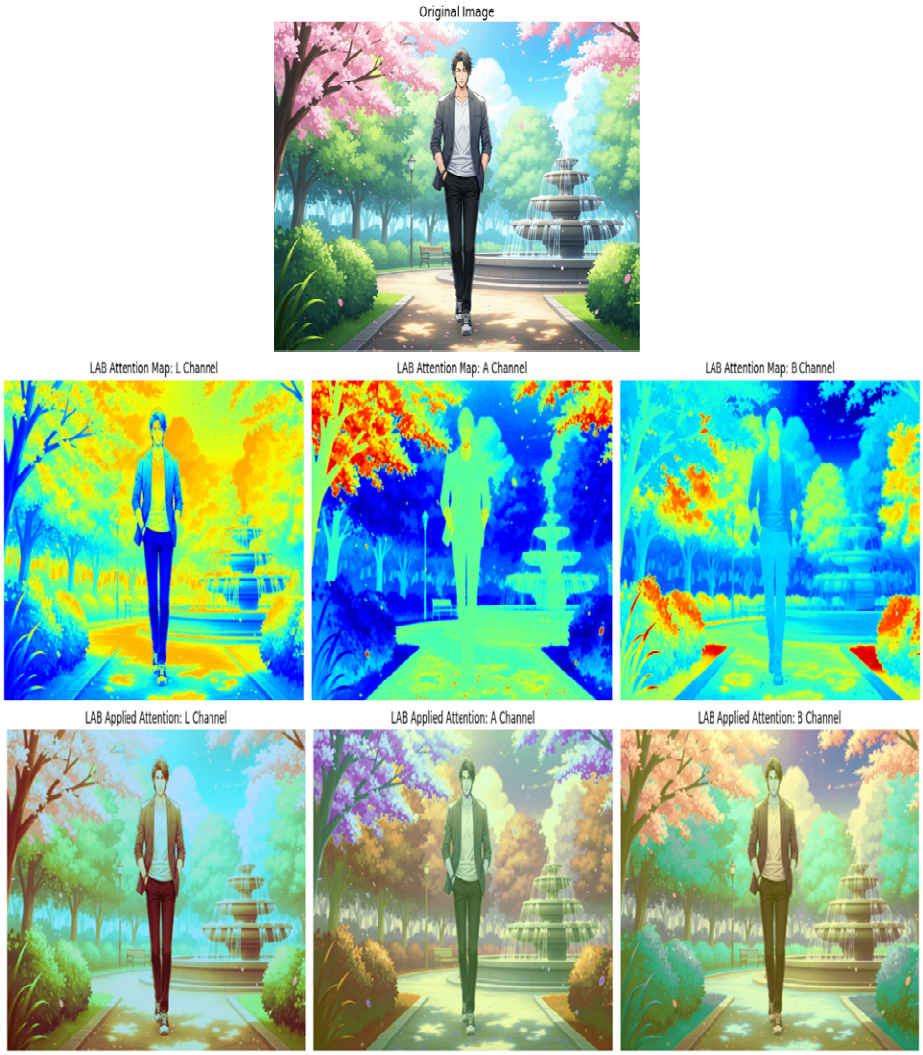
실험 환경은 GPU 가속이 가능한 워크스테이션에서 구성되었으며, 모든 실험은 동일한 하드웨어·소프트웨어 조건에서 반복 수행하여 분석 결과의 일관성을 확보하였다. 데이터 전처리, 입력 개입, Score-CAM 계산, 주파수 분해 및 시각화 과정은 모듈화된 분석 도구로 구성되어, 각 개입 수준에 따른 예측 확률 변화(ΔP)와 활성화 패턴을 단계적으로 비교할 수 있도록 하였다.

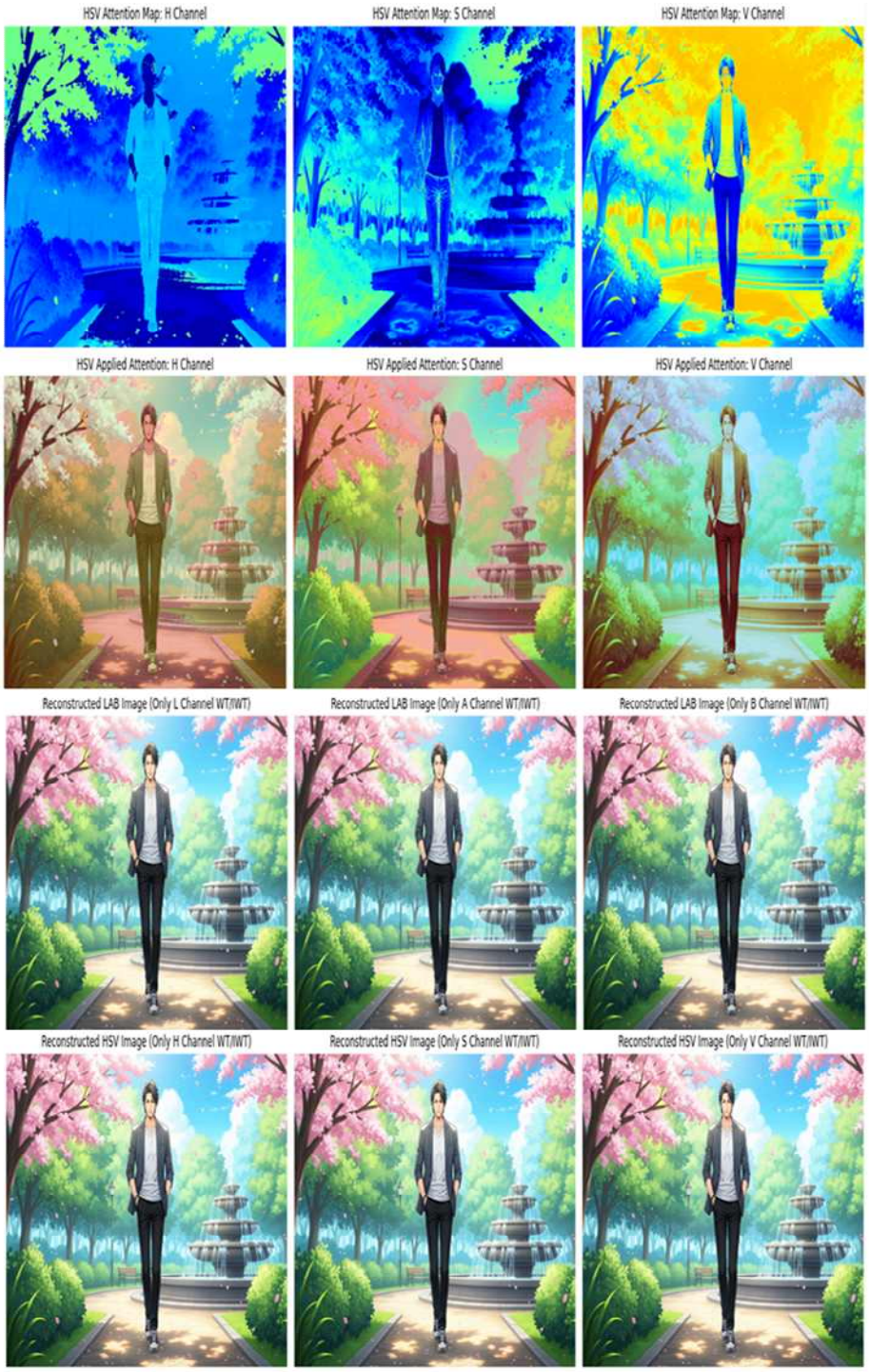
1) 사전 이미지 준비 및 색상 특성 분석

가) 입력 이미지 확보 및 밝기 기반 특성 분석

본 연구는 색상 및 주파수 개입 실험에 앞서 입력 이미지의 밝기 수준 변화가 색상 및 주파수 구성 요소에 미치는 영향을 사전에 분석하였다. 이는 밝기 변화가 모델 내부 표현에 미치는 영향을 명확히 파악하고 이후 개입 실험의 베이스라인 입력을 설정하기 위한 준비 단계이다. 먼저 밝기 과다, 밝기 부족, 정상 밝기의 세 가지 조건에서 이미지를 준비하고, 각 이미지를 RGB에서 LAB 및 HSV 색 공간으로 변환하였다. 색 공간별 채널이 모델 초기 활성화 구조에 미치는 영향을 확인하기 위해 채널별 어텐션 맵을 생성하였으며, 이를 통해 명도, 채도, 색차, 색조 성분이 밝기 조건에 따라 어떻게 차별적으

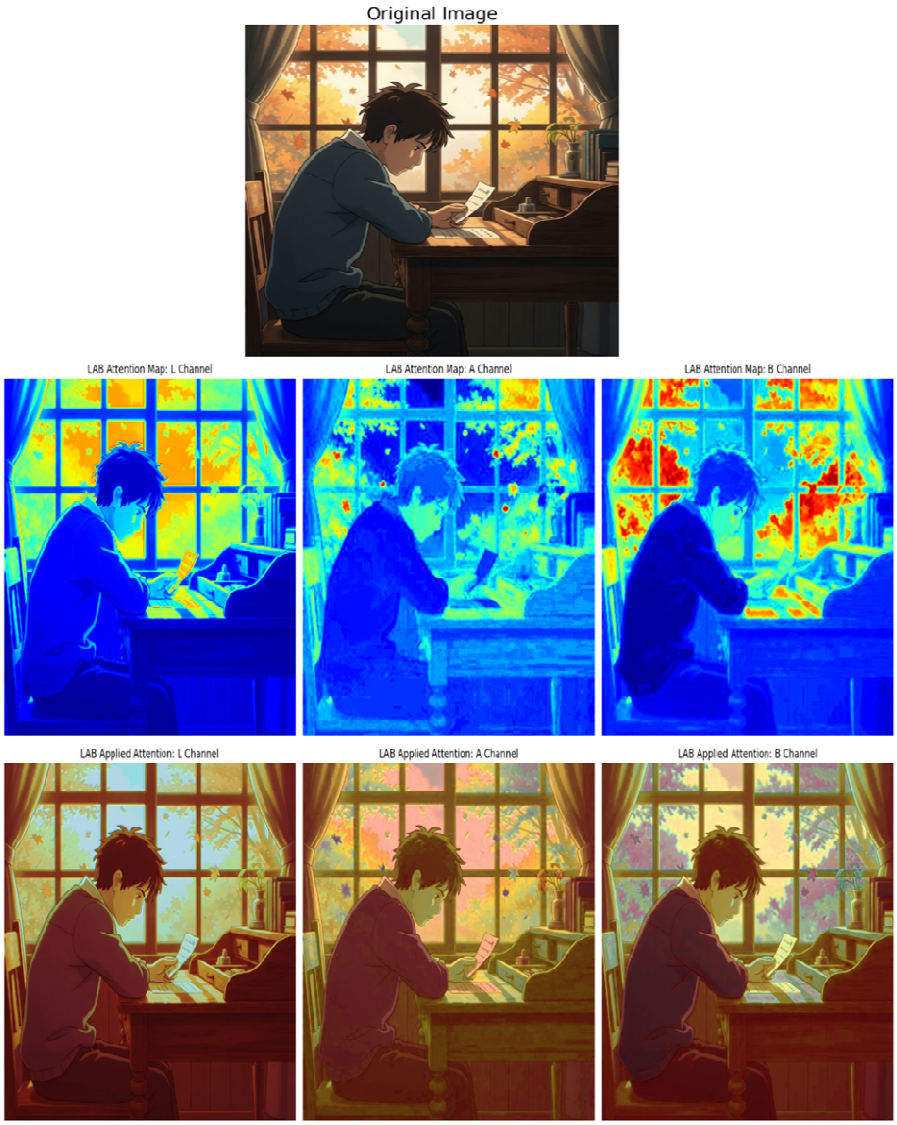
로 강조되는지 정성적으로 비교하였다. 생성된 어텐션 맵을 원본 이미지에 적용하여 적용된 어텐션 맵을 구성한 후, 이를 채널 단위로 분리하고 역웨이블릿 변환을 이용하여 각 색상 및 구조 성분이 반영된 재구성 LAB/HSV 이미지 세트를 복원하였다. 이렇게 확보된 채널별 재구성 이미지는 이후 색상 및 주파수 개입 실험에서 밝기 조건별 베이스라인 입력으로 활용되었으며, 이를 통해 밝기 변화에 따른 색상 민감도와 주파수 선택성을 인과적으로 비교·분석할 수 있는 기반을 마련하였다.

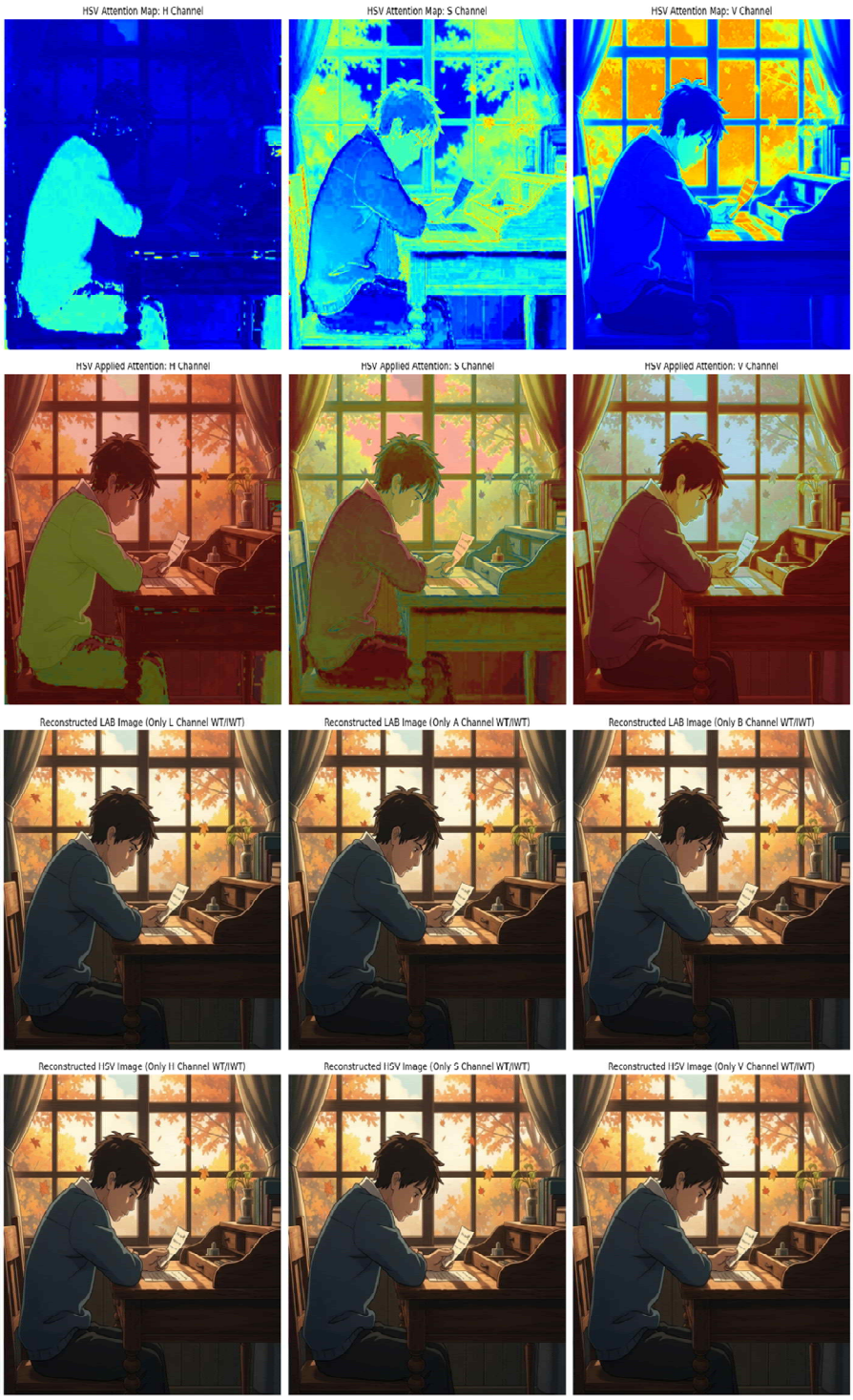




[그림 4-2] 밝은 조명 조건 이미지의 색상·채널 반응 및 채널별 결과

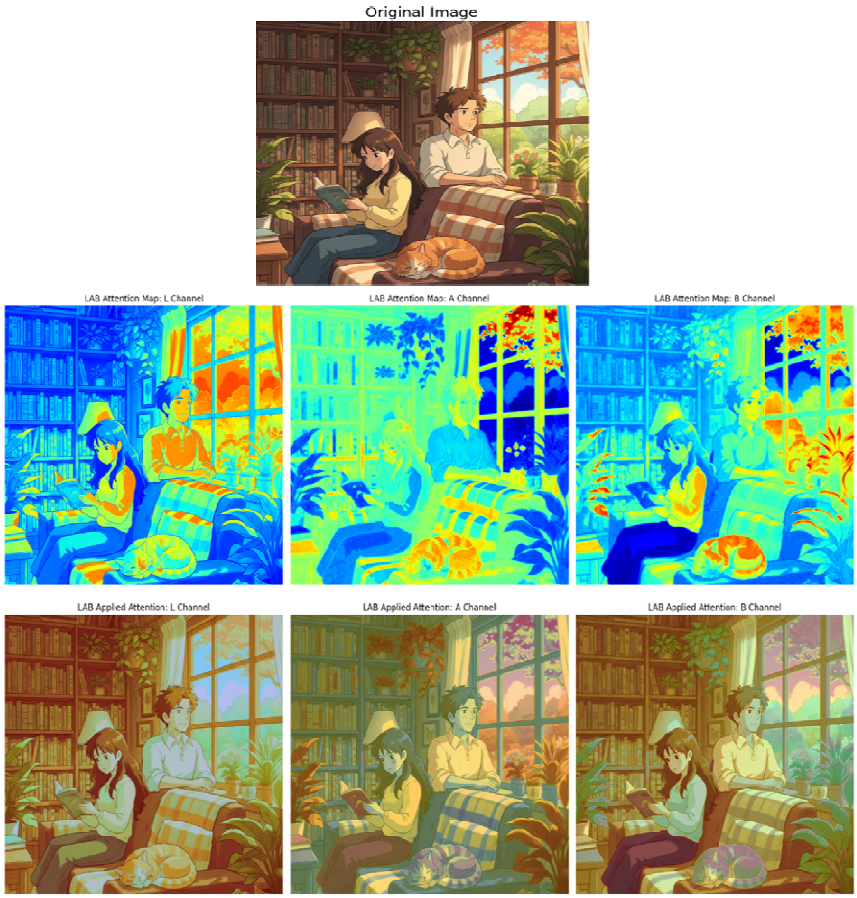
[그림 4-2]는 밝은 조명 조건에서 입력 이미지의 색상 채널 반응을 분석한 전체 과정을 보여준다. 먼저 LAB 및 HSV 채널별 어텐션 맵을 통해 밝은 환경에서 모델이 명도, 색차, 채도, 색조와 같은 색상 성분에 어떻게 반응하는지를 시각적으로 확인하였다. 이후 어텐션 맵을 원본 이미지에 적용하고 각 채널을 분리하여 재구성한 결과를 비교함으로써, 밝은 조명으로 인해 어떤 색상 및 구조 정보가 강조되거나 소실되는지를 직관적으로 관찰할 수 있다. 이 과정은 이후 개입 실험에서 사용되는 조명 조건별 베이스라인 입력 분석을 위한 사전 준비 단계이다.

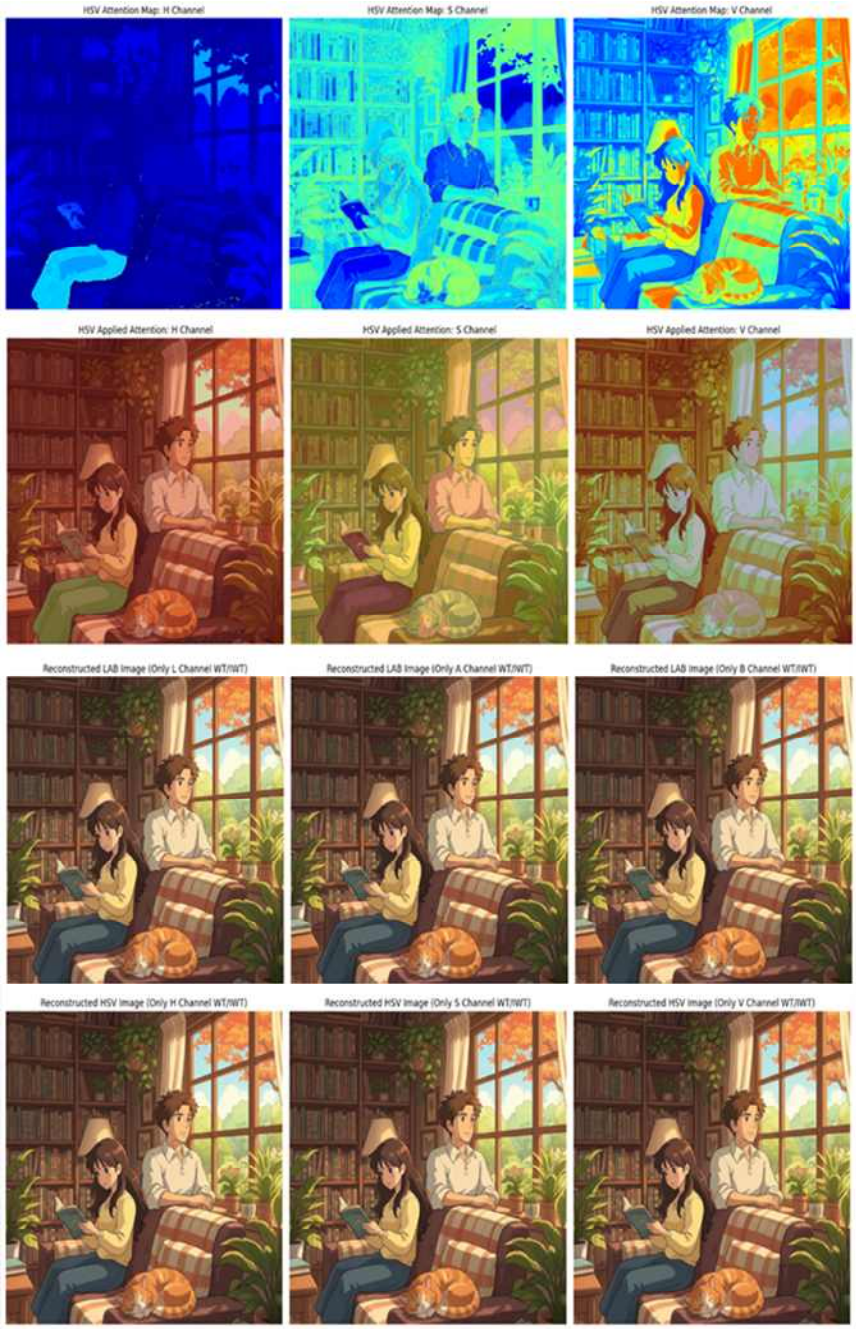




[그림 4-3] 어두운 조명 조건 이미지의 색상·채널 반응 및 채널별 결과

[그림 4-3]은 어두운 조명 조건에서 LAB 및 HSV 색 공간으로 분해된 입력 이미지의 채널별 어텐션 맵, 적용된 어텐션 맵, 역웨이블릿 변환 기반 재구성 이미지를 비교한 결과이다. 어두운 환경에서는 LAB의 L 채널과 HSV의 V 채널이 저값 영역에 집중되어 명도 정보가 충분히 표현되지 못하며, A/B 및 S 채널에서도 색채 및 채도 정보가 약화되어 장면의 시각적 구분도가 감소하였다. 채널별 어텐션을 적용한 결과, 명도 및 색채 정보의 손실이 이미지 전체의 윤곽, 색조, 재질 표현을 저하시켜 모델이 활용할 수 있는 시각적 단서를 제한하는 것으로 확인되었다. 역웨이블릿 변환으로 재구성된 이미지는 조도 저하가 명도, 색채, 구조 성분을 순차적으로 손상시키는 과정을 채널별로 분리하여 보여주며, 어두운 조명에서 모델 내부 표현이 약화되는 구조적 원인을 규명하는 기초 자료로 활용된다.





[그림 4-4] 정상 조명 조건 이미지의 색상·채널 반응 및 채널별 결과

[그림 4-4]는 정상 조명 조건에서 LAB 및 HSV 색 공간으로 분해된 입력 이미지의 채널별 어텐션 맵, 적용된 어텐션 맵, 역웨이블릿 변환 기반 재

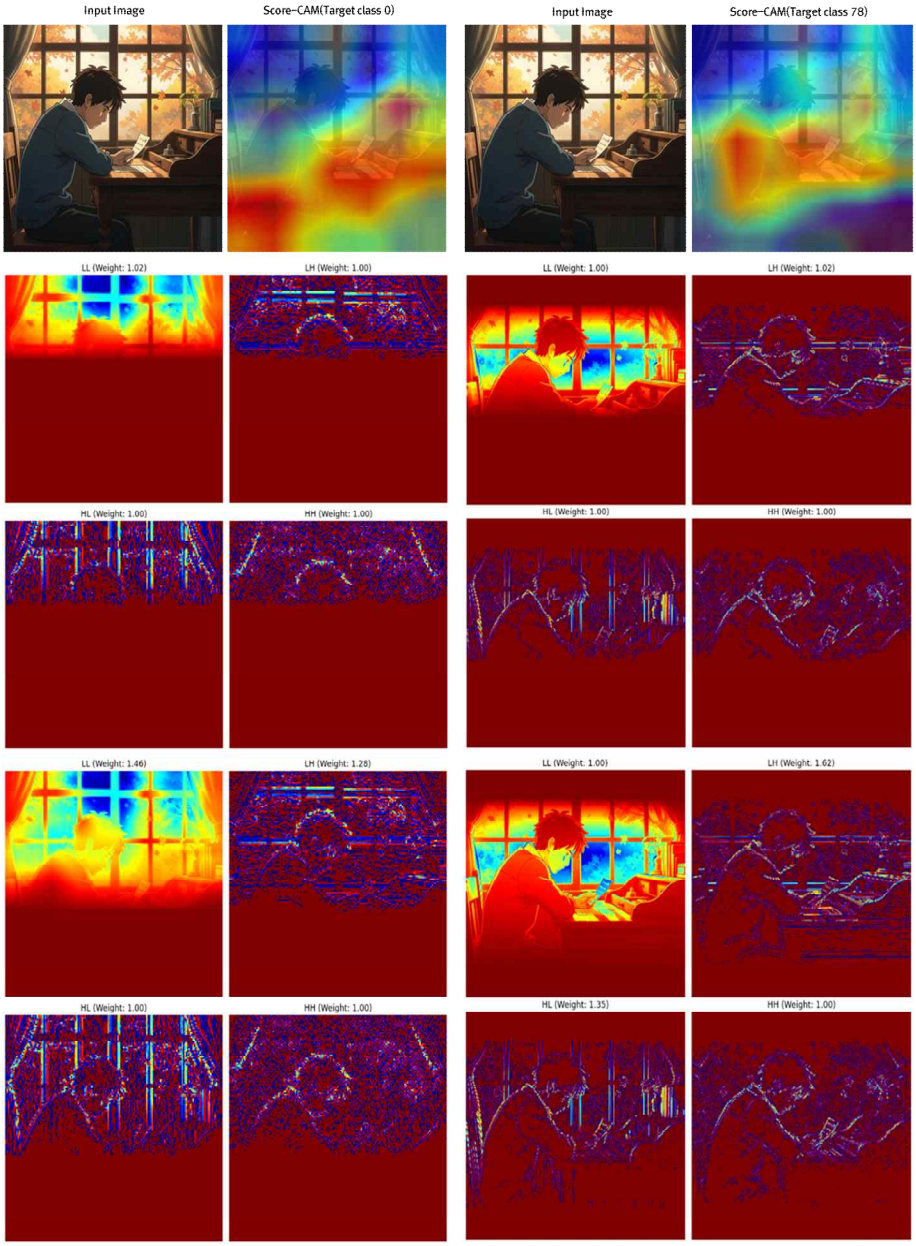
구성 이미지를 비교한 결과이다. 정상 조명에서는 LAB의 L 채널과 HSV의 V 채널에서 명도 값이 안정적으로 분포하며 극단적인 밝기 압축이나 저값 집중 현상이 나타나지 않았다. A/B 및 H/S 채널에서도 색채 및 채도 정보가 균형 있게 유지되어 장면의 색상 구조가 왜곡 없이 표현되었다. 채널별 어텐션을 원본에 적용한 결과, 주요 객체의 윤곽, 색조, 재질 정보가 명확하게 보존되어 모델이 추출하는 시각적 단서가 정상적으로 유지되는 것으로 확인되었으며, 이는 정상 조명이 모델의 내부 표현을 가장 안정적으로 유도함을 의미한다. 역 웨이블릿 변환으로 재구성된 이미지는 각 채널의 명도, 색채, 질감 정보가 왜곡 없이 복원됨을 보여주며, 이는 이후 극단적 조명 조건과 대비하여 조명 변화가 모델의 신뢰도 및 내부 표현에 미치는 영향을 분석하는 기준 입력으로 활용된다.

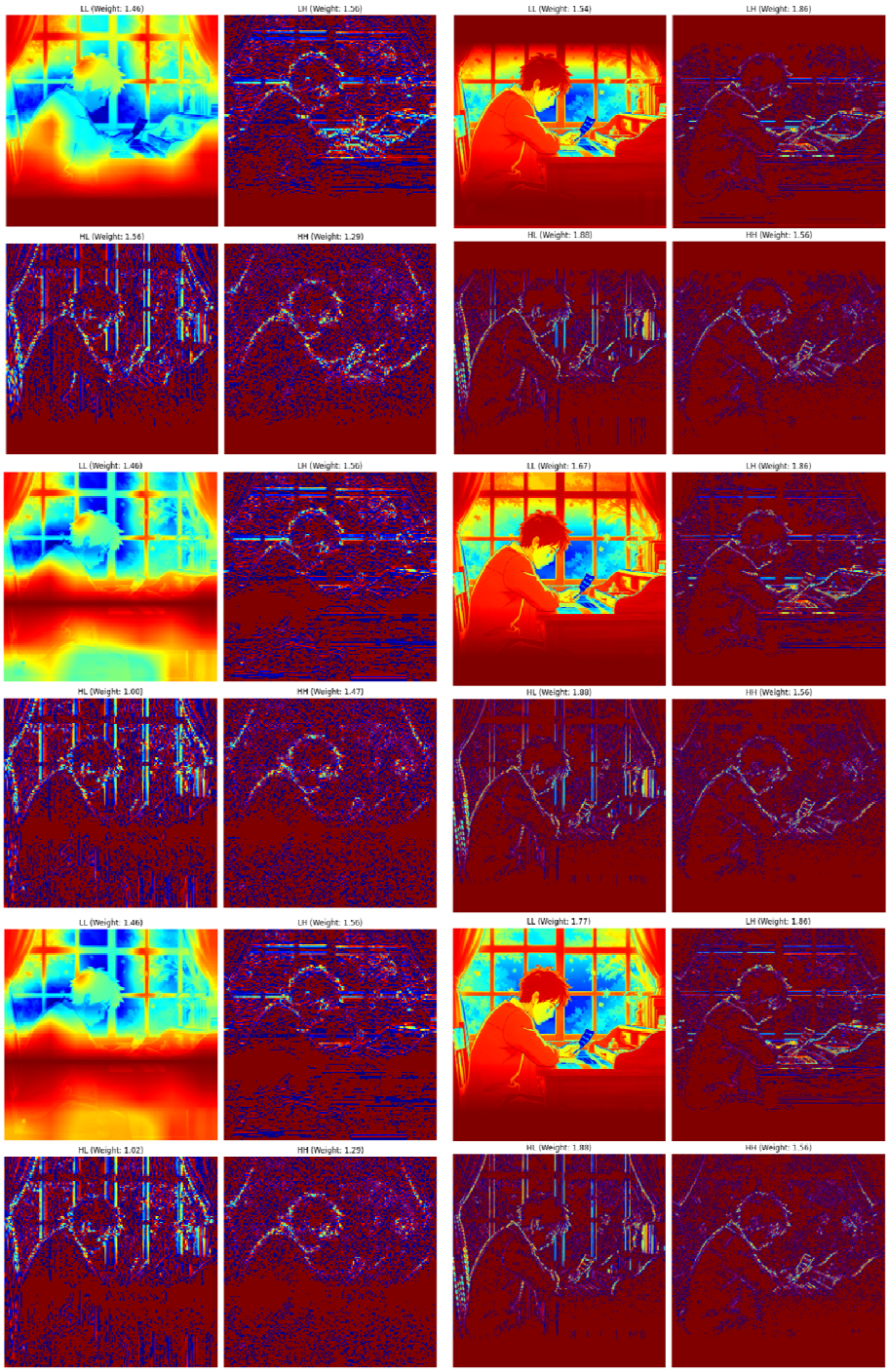
나) Score-CAM 기반 주파수 반응 특성 분석

본 사전 실험은 Score-CAM이 시각화하는 공간적 활성화 구조와 ResNet 50이 실제로 활용하는 주파수 성분 간의 대응 관계를 규명하기 위해 수행되었다. 웨이블릿 기반으로 주파수 대역을 단계적으로 개입한 결과, 조명 조건과 무관하게 동일한 구조적 패턴이 관찰되었다. LL 대역 개입은 Score-CAM 활성화를 거의 변화시키지 않았으며, 이는 모델이 전역적 저주파 정보보다 윤곽·질감과 같은 구조 단서를 포함한 중·고주파 정보에 더욱 의존함을 시사한다. 반면 LH 및 HL 대역 개입에서는 수평·수직 윤곽 중심의 국소적 활성화가 강하게 증가하였고, HH 대역 개입에서는 클래스 판별에 요구되는 미세 질감에 대한 활성화가 일관되게 강화되었다. 이러한 일관된 패턴은 ResNet 50이 조명 조건이 달라져도 고주파(LH/HL/HH) 성분에 선택적으로 반응하는 고유한 주파수 선택성을 가지고 있음을 보여준다.

또한 타깃 클래스의 종류가 달라도 동일한 LH/HL/HH 대역에서 활성화가 강화되는 현상이 반복적으로 관찰되었으며, 이는 예측 확률 변화량의 방향과 크기가 결국 해당 주파수 성분에 의해 결정됨을 사전적으로 검증하는 결과이다. 따라서 본 실험은 이후 수행되는 색상 및 주파수 개입 기반 인과 분석이 왜 신뢰도 변화의 본질적 원인을 주파수 대역에서 찾을 수 있는가를 설

명하는 핵심 이론적 근거를 제공한다.

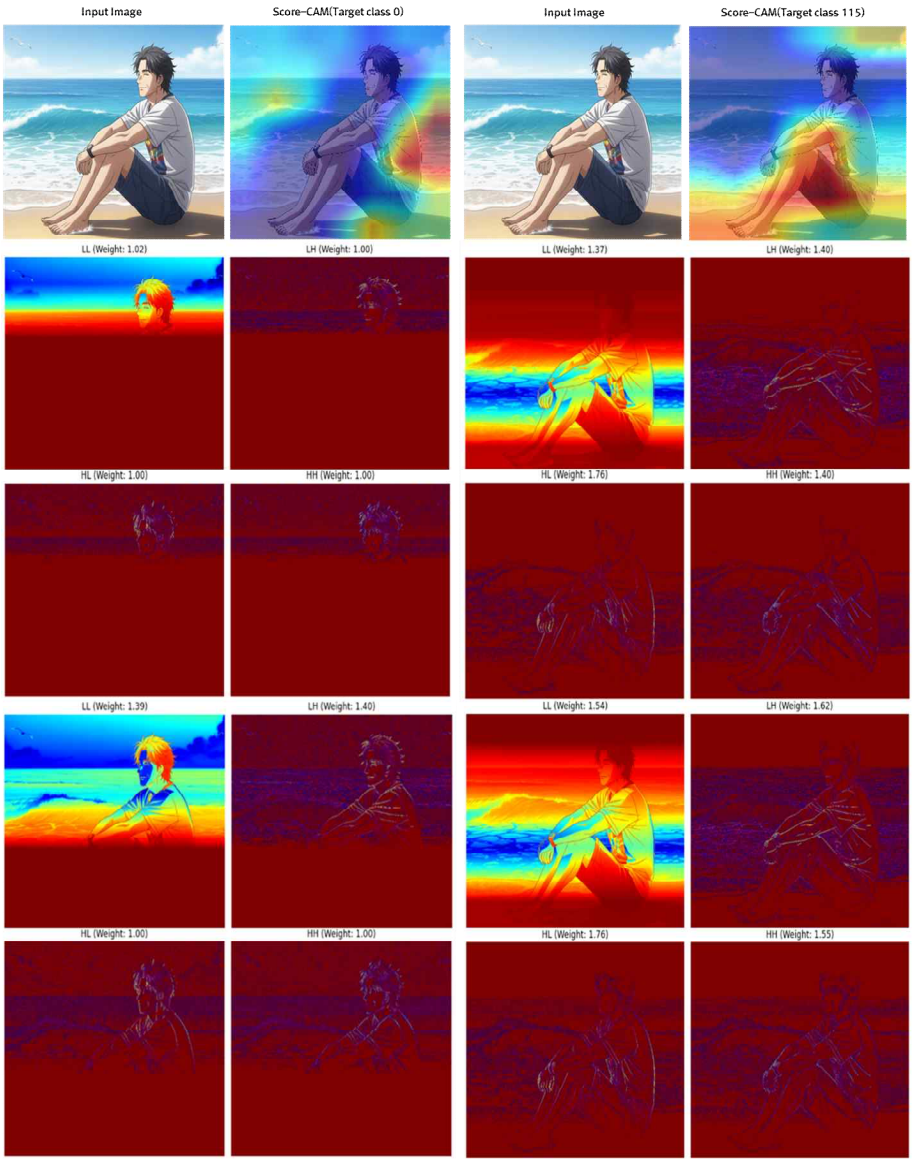


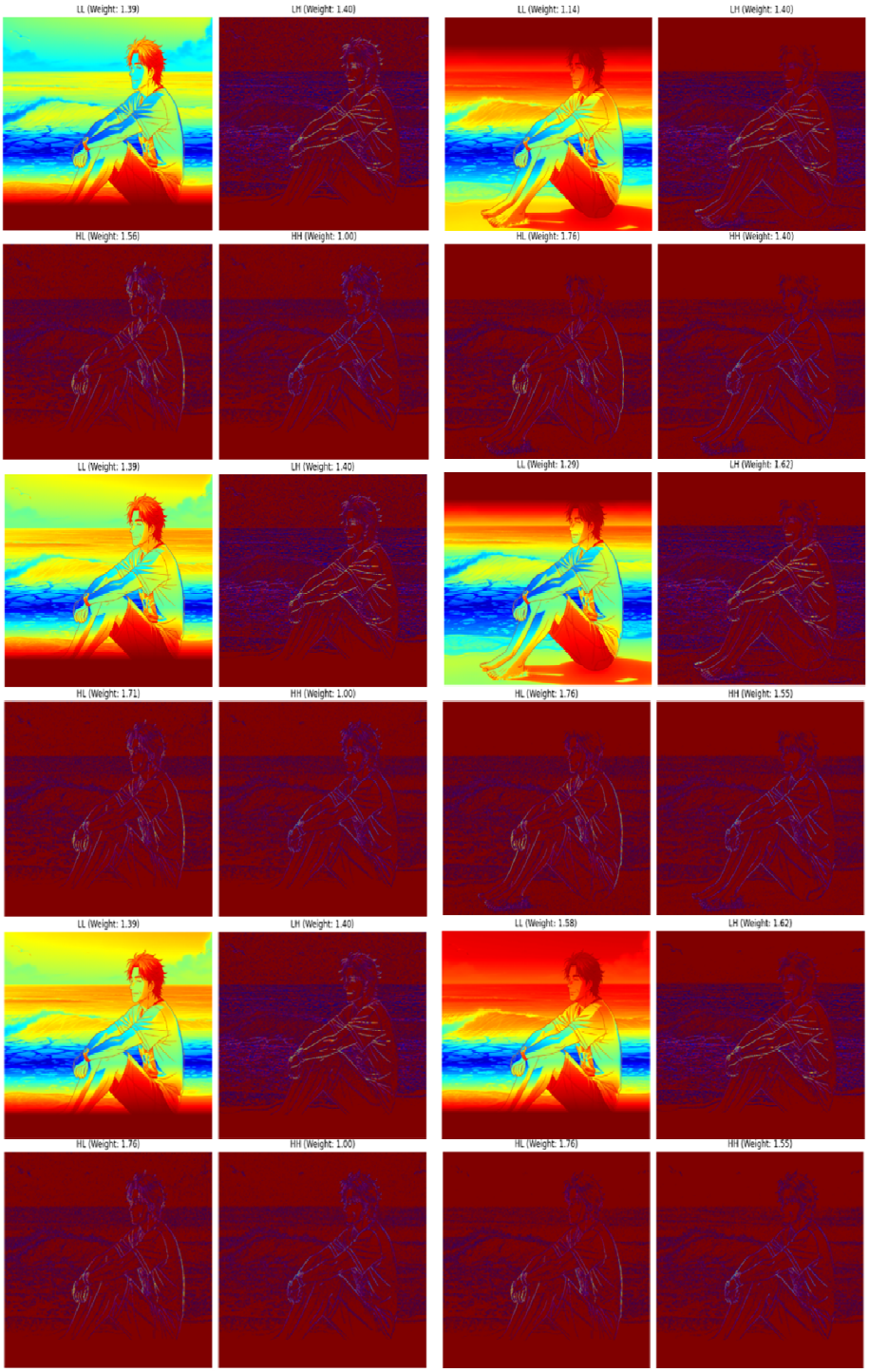


[그림 4-5] 어두운 조명 조건 주파수 개입 Score-CAM 시각적 변화

[그림 4-5]는 어두운 조명(Under-exposure) 하에서 입력 이미지의 웨이

블릿 대역을 개별적으로 조정했을 때 Score-CAM이 시각화하는 활성화 구조가 어떻게 변하는지를 비교한 결과이다. LL 대역 개입은 활성화 구조에 거의 영향을 미치지 않지만, LH·HL 대역 개입에서는 수평·수직 윤곽 중심의 국소적 활성화가 두드러지게 증가하였다. HH 대역 개입에서는 미세 질감에 대한 반응이 상대적으로 강화되었다. 이러한 패턴은 저조도 환경에서도 ResNet 50이 고주파 성분에 선택적으로 의존함을 보여준다.





[그림 4-6] 밝은 조명 조건 주파수 개입 Score-CAM 시각적 변화

[그림 4-6]은 밝은 조명하에서 웨이블릿 주파수 대역을 개입했을 때 나타나는 Score-CAM 활성화 변화를 비교한 것이다. LL 대역 개입은 활성화 반응에 거의 영향을 주지 않았으며, LH·HL 대역 개입에서는 윤곽(edge) 중심 활성화가 강하게 나타났다. HH 대역 개입에서는 클래스 판별에 기여하는 고주파 질감 정보에 대한 활성화가 일관되게 강화되었다. 이는 조도 증가 상황에서도 모델이 중·고주파 기반 구조 단서에 주로 의존하는 고유한 주파수 선택성을 지님을 시사한다. 본 실험은 조명 조건(과다 조명·과소 조명)과 무관하게 ResNet 50의 Score-CAM 활성화 구조가 특정 주파수 대역(LH·HL·HH)에 안정적으로 대응한다는 점을 실증적으로 확인하였다. LL 대역은 활성화 변화가 거의 없어 모델이 전역 저주파 정보에는 크게 의존하지 않음을 보여주고, 반대로 LH·HL·HH 대역은 조명·색상 변화와 상관없이 일관된 구조적 활성화를 유도하였다. 특히 윤곽(edge) 중심의 반응(LH·HL)과 미세 질감(HH)에 대한 반응 강화는 모델의 클래스 판별 과정이 중·고주파 기반 구조 단서를 핵심적으로 활용함을 나타낸다.

이러한 결과는 Score-CAM이 강조하는 공간적 Attention 패턴이 실질적으로 중·고주파 성분에서 기원하며, 색상 밝기 변화의 영향보다 주파수 구조에 더 안정적으로 대응함을 시사한다.

따라서 본 사전 실험은 이후 장에서 수행되는 색상·주파수 개입 기반 인과 구조 분석의 타당성을 뒷받침하는 핵심 근거로 기능하며, ResNet 50의 인식 과정이 주파수 선택성을 중심으로 이루어진다는 구조적 특징을 명확히 확인한 결과라 할 수 있다.

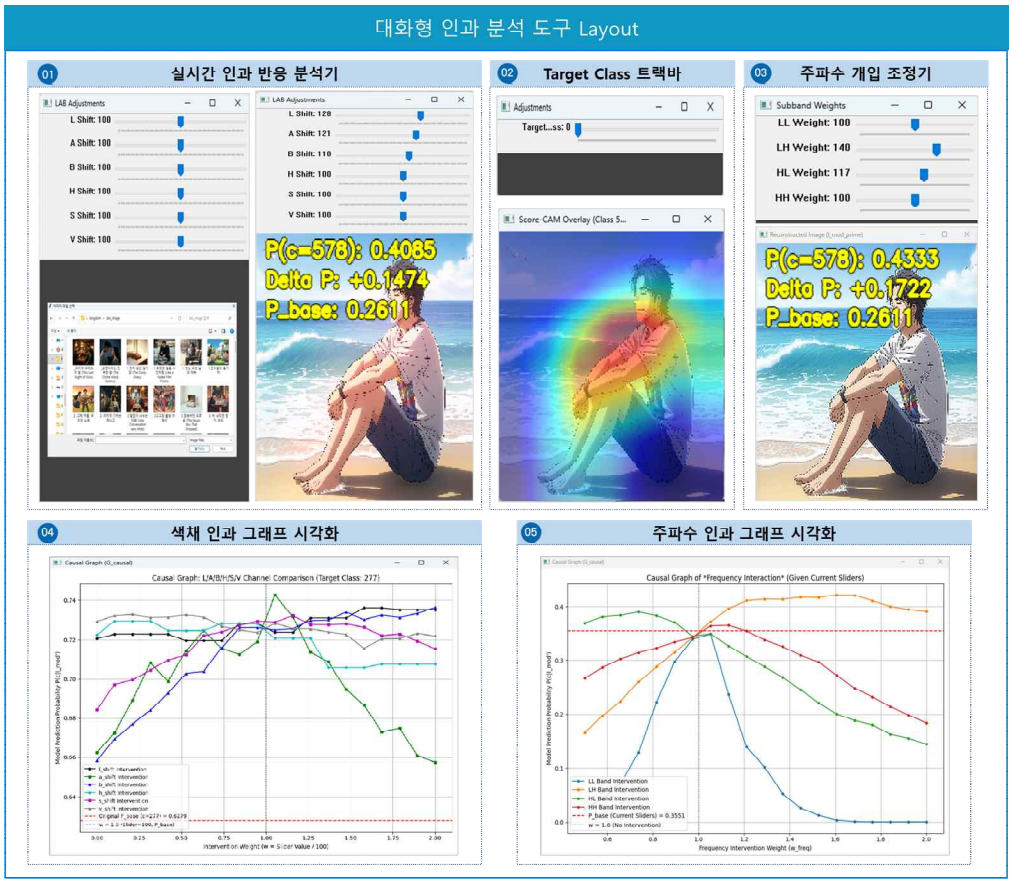
2) 색상·주파수 개입 기반 인과 분석 실험

가) 실험 수행을 위한 대화형 분석 도구의 역할

본 연구는 색상 및 주파수 개입에 따른 모델 반응 변화를 정량·정성적으로 분석하기 위해, 실험자가 입력 파라미터를 단계적으로 조정하며 모델의 반응을 즉시 확인할 수 있는 대화형 분석 도구(Interactive Analysis Tool)를 활용하였다. 이 도구는 다양한 개입 조건을 신속하게 실험하고, 각 개입에 따른

예측 확률 변화와 Score-CAM 활성화 구조를 실시간으로 관찰할 수 있도록 구성된 실험 환경(Experimental Environment)으로 제공된다. 이를 통해 색상·주파수 조작에 따른 모델의 반응 패턴을 체계적으로 비교·분석할 수 있다.

나) 대화형 인과 분석 도구 Layout 구성 및 역할



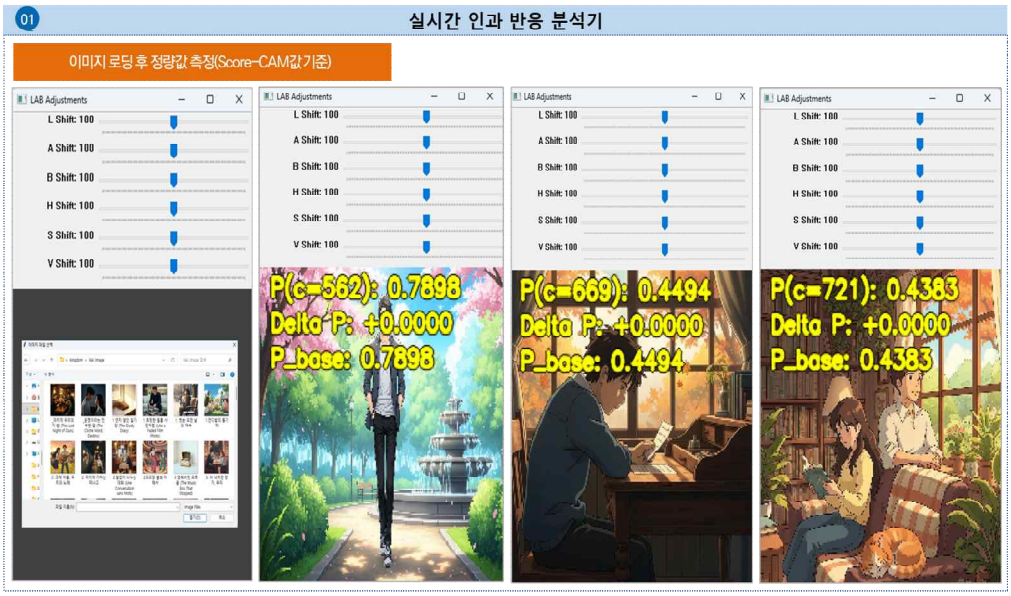
[그림 4-7] 색상·주파수 개입 기반 대화형 분석 도구 구성

[그림 4-7]는 본 연구에서 색상·주파수 개입 실험을 수행하기 위해 활용된 대화형 분석 도구의 전체 구성 요소(UI Layout)를 나타낸다. 이 도구는 (1) 실시간 개입-반응 관찰 영역, (2) Target Class 설정용 트랙바, (3) 주파수 대역 개입 조정기, (4) 색상 개입에 따른 인과 그래프 시각화 영역, (5) 주파수 개입에 따른 인과 그래프 시각화 영역으로 구성된다. 각 구성 요소는 상호 연동되어 개입 → 모델 반응 측정 → 시각화 → 비교 분석의 단계를 통합

적으로 수행함으로써 색상·주파수 조작이 ResNet 50의 신뢰도 변화 및 활성화 구조에 어떻게 영향을 미치는지를 실험적으로 탐색할 수 있도록 지원한다.

(1) 실시간 인과 반응 분석기

이 모듈은 실험자가 이미지를 불러오는 즉시 해당 이미지의 기준 확률을 계산하고, 이후 주파수 파라미터를 조정하면 변형된 입력을 모델에 즉시 재입력하여 예측 확률과 예측 확률 변화량을 실시간으로 제시한다. 이를 통해 특정 주파수 대역 개입이 모델의 신뢰도에 미치는 영향을 즉각적으로 확인할 수 있으며, 주파수 기반 인과 구조를 탐색하는 본 연구의 핵심 분석 기능을 수행한다.



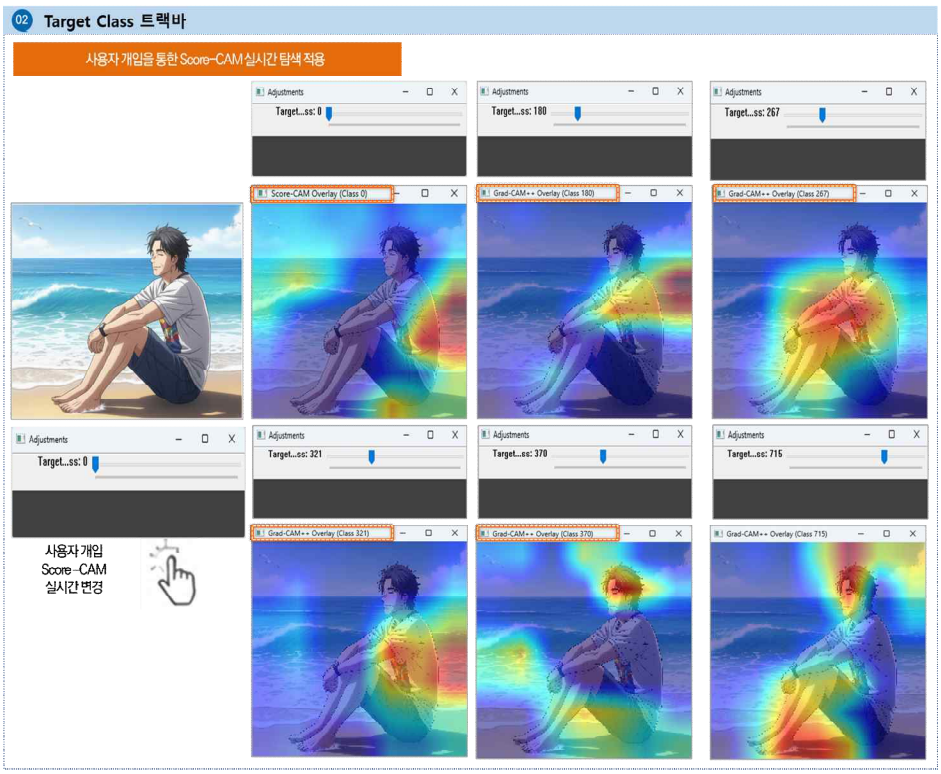
[그림 4-8] 이미지 확률 측정 및 개입에 따른 모델 확률 변화 정량화

[그림 4-8]은 실험자가 색상 파라미터를 조정했을 때 ResNet 50 모델의 기준 확률, 개입 후 예측 확률, 그리고 예측 확률 변화량을 실시간으로 측정·시각화하는 인터페이스를 보여준다. 이 모듈은 특정 색상 채널의 증감이 모델 신뢰도에 어떤 변화를 유발하는지를 즉각적으로 관찰할 수 있도록 지원하며, 색상 개입이 모델의 판단 구조에 미치는 인과적 영향을 탐색하는 실험의 핵심 구성 요소로 사용된

다.

(2) Target Class 트랙바

본 모듈은 실험자가 분석 대상 클래스(Target Class)를 직접 지정·변경할 수 있도록 구성된 제어 요소로, 트랙바 조작을 통해 설정된 클래스값이 즉시 모델에 전달되며 해당 클래스의 Score-CAM 활성화 맵이 실시간으로 갱신된다. 이를 통해 동일한 입력 이미지에 대하여 클래스 번호 변화가 모델의 주의 집중 영역(Class-dependent Attention Focus)을 어떻게 이동시키는지를 직관적으로 관찰할 수 있다. 본 트랙바는 모델의 클래스 의존적 주의 메커니즘(class-dependent attention mechanism)을 실험적으로 확인하고 비교하기 위한 핵심 제어 모듈로 기능한다.



[그림 4-9] Target Class 트랙바 기반 실시간 Score-CAM 반응 시각화

[그림 4-9]는 사용자가 Target Class 트랙바를 조정할 때, Score-CAM

활성화 맵이 실시간으로 업데이트되는 과정을 시각화한 것이다. 각 프레임은 트랙바를 통해 선택된 클래스 값의 변화에 따라 모델이 주목하는 영역(Attention Focus)이 어떻게 이동하는지를 보여준다. 예를 들어, Target Class 클래스 번호가 변경됨에 따라 모델의 주의 영역이 인물의 얼굴, 팔, 배경 등으로 이동하는 패턴을 확인할 수 있다. 이는 모델의 클래스 의존적 주의 집중(Class-dependent Attention) 특성을 실험적으로 검증하는 시각적 근거를 제공하는 인터페이스 화면이다.

(3) 주파수 개입 조정기

본 모듈은 웨이블릿 변환을 통해 분리된 주파수 대역의 가중치를 실시간으로 조정하여 모델 입력을 재구성하는 인터페이스이다. 사용자가 슬라이더를 통해 각 대역의 강도를 조절하면, 재구성된 입력은 Score-CAM을 통해 모델의 시각적 반응 변화를 즉시 시각화하고 예측 확률의 차이(ΔP)를 실시간으로 계산한다. 이를 통해 사용자는 특정 주파수 대역이 모델의 시각적 주의와 신뢰도에 미치는 직접적 영향을 실험적으로 탐색할 수 있다.

[그림 4-10] 주파수 개입 조정기를 통한 Score-CAM 반응 변화 시각화

[그림 4-10]은 사용자가 주파수 개입 조정기의 슬라이더를 통해 웨이블릿 대역의 가중치를 실시간으로 조정할 때, 모델의 시각적 반응과 예측 확률이 어떻게 변화하는지를 보여주는 인터페이스 화면이다. 상단의 슬라이더는 네 개의 주파수 대역에 대한 개입 강도(weight)를 나타내며, 사용자가 특정 대역을 증가시키면 입력 이미지의 전역 구조 정보가 강화되어 Score-CAM 활성화 맵이 더욱 넓고 균질한 형태로 확장된다. 반대로 고주파 대역을 강화하면 모델의 주의 영역이 세부 텍스처나 에지 부분으로 집중되는 패턴을 나타낸다.

하단의 실시간 출력 창에서는 개입 후 모델의 예측 확률P_{mod}과 기준 확률P_{base} 그리고 변화량$\Delta P = P_{\mathrm{mod}} - P_{\mathrm{base}}$가 시각적으로 표기된다. 이를 통해 사용자는 주파수 대역별 조작이 모델의 확률적 신뢰도(probabilistic confidence)와 시각적 주의 집중(attention focus)에 미치는 인과적 영향을 직관적으로 관찰할 수 있다. 본 인터페이스는 단순한 시각화 도구를 넘어, 모델의 주파수 선택성 및 공간 주파수적 인식 편향을 실험적으로 검증하기 위한 핵심 모듈로 만들어졌다.

(4) 색채 인과 그래프 시각화

색상 관련 변수의 변화 범위에 따른 모델 예측 확률 변화를, 단축키를 자동으로 수집·시각화하는 기능을 제공한다. 그래프의 가로축은 색상 개입 강도 세로축은 확률 변화량을 나타내며, 색상 요소별 인과적 영향력을 정량적으로 비교·평가할 수 있다. 본 모듈은 색채 개입 → 모델 반응의 함수 관계를 수치적·시각적으로 표현한다.

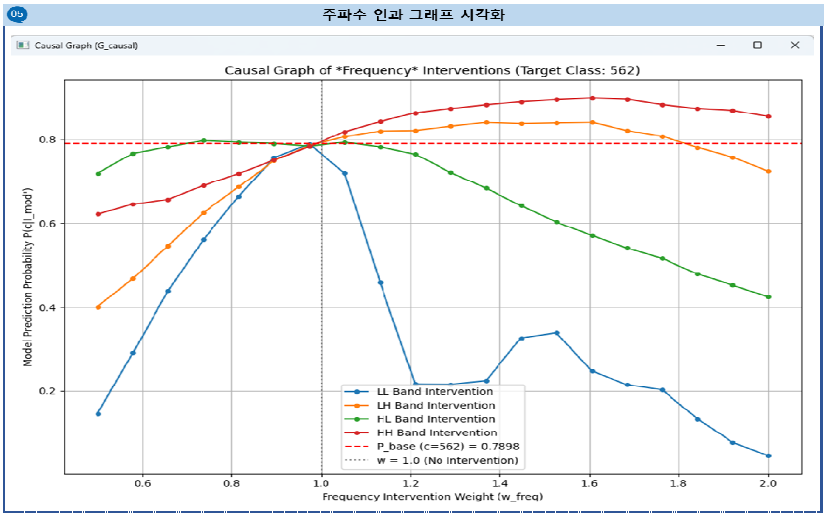
[그림 4-11] 색상 개입 강도에 따른 모델 예측 확률 변화 시각화

[그림 4-11]은 색채 인과 그래프 시각화 모듈의 실행 결과를 보여준다. 본 그래프는 색상 변수의 개입 강도를 단계적으로 변화시키며, 각 단계에서 측정된 모델 예측 확률 변화를 자동으로 수집·시각화한 결과이다. 가로축은 색상 개입 강도(Intervention Weight)를, 세로축은 모델의 예측 확률을 나타낸다. 각 곡선은 개별 색상 채널의 변화에 따른 확률 반응의 추이를 의미하며, 특정 채널의 변화가 모델 신뢰도에 미치는 인과적 영향을 정량적으로 비교할 수 있다. 본 도구의 시각화 결과는 색상 공간별 특성을 다 각도로 분석할 수 있도록, LAB과 HSV 색상 공간을 동시에 비교할 수 있는 통합 시각화 모듈과 각 색상 공간의 인과 반응을 개별적으로 관찰할 수 있는 분리형 시각화

모듈로 구성되었다.

(5) 주파수 인과 그래프 시각화

주파수 대역별 개입 강도 변화에 따른 모델 예측 확률의 변화를 자동으로 수집하여 2D 그래프로 시각화하는 모듈이다. 각 곡선은 서브 밴드별 추이를 나타내며, 특정 대역의 영향 패턴을 인과적으로 분석할 수 있다. 이를 통해 모델의 주파수 민감도(frequency sensitivity)와 색상 요소 간의 상호작용을 정량적으로 해석할 수 있다.



[그림 4-12] 주파수 대역 개입에 따른 모델 예측 확률 변화 시각화

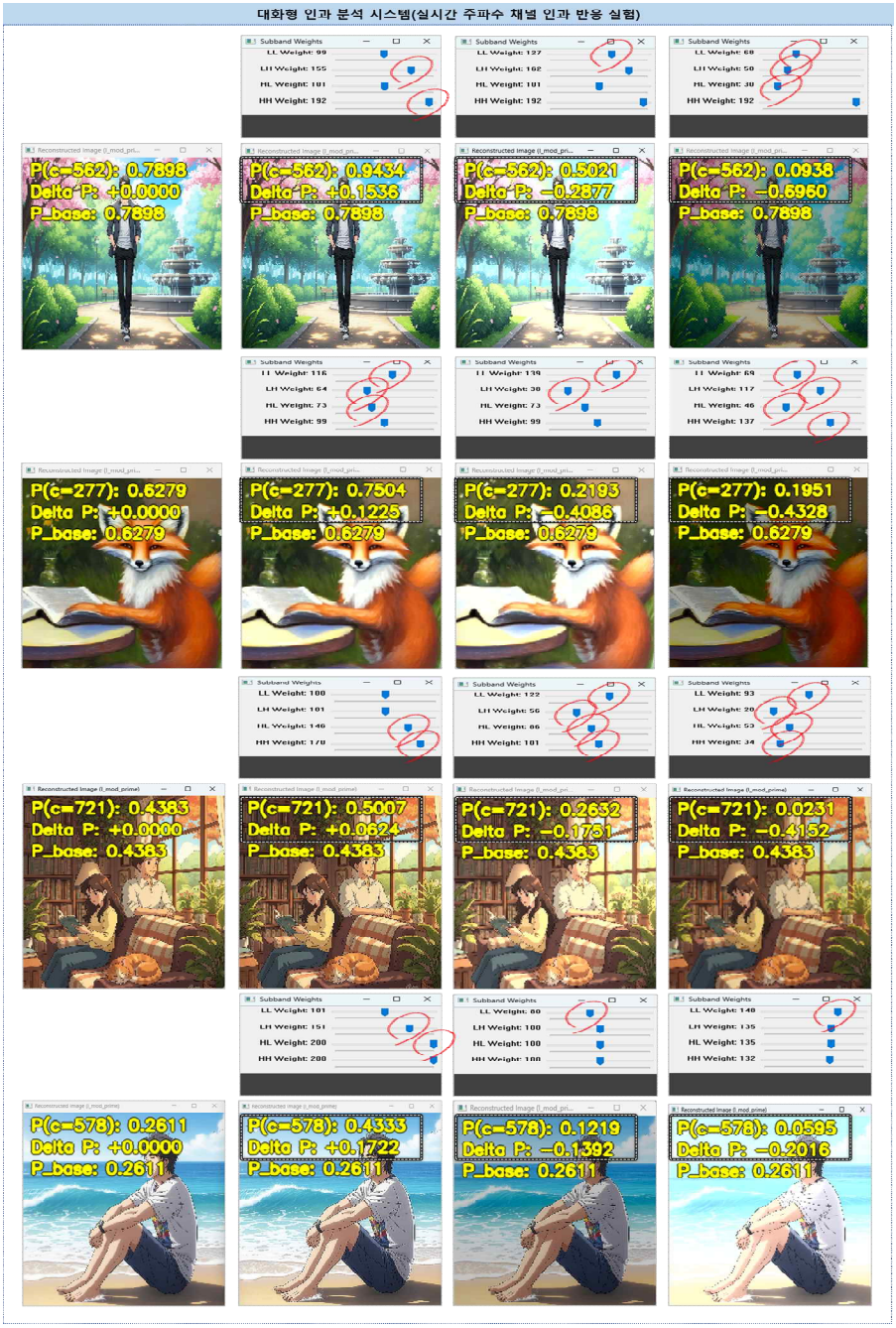
[그림 4-12]는 주파수 인과 그래프 시각화 모듈의 결과를 보여준다. 본 그래프는 웨이블릿 변환으로 분리된 주파수 대역에 대해 각 대역의 개입 강도를 단계적으로 변화시키며 모델의 예측 확률 변화를 자동으로 수집·시각화한 결과이다. 가로축은 주파수 개입 강도를, 세로축은 모델의 예측 확률을 나타낸다. 각 곡선은 개별 서브밴드의 인과적 반응 패턴을 의미하며, 대역별 신호 증폭 또는 감쇠가 모델의 확률 변화에 어떤 영향을 미치는지를 정량적으로 비교할 수 있다.

제 3 절 모델 반응 측정 및 인과 관계 정량화

1) 실시간 개입 기반 시각 반응 실험

본 절에서는 제4장 제2절에서 구현된 대화형 인과 분석 도구를 사용하여 색상 및 주파수 개입이 모델 예측에 미치는 영향을 실험적으로 측정하고 그 인과적 구조를 규명한다. 실험은 실시간 개입 반응 관찰, 색상·주파수 정규화 효과 분석, 그리고 신뢰도 변화의 정량 평가의 세 단계로 구성된다.

[그림 4-13] LAB·HSV 색상 개입에 따른 모델 확률 변화 시각화



[그림 4-14] 주파수 개입에 따른 모델 확률 변화 시각화

가) LAB 및 HSV 색상 정규화의 인과적 해석

대화형 인과 분석 도구의 색상 공간 개입을 통해 모델의 색채 민감도(Color Sensitivity)와 도메인 편향(Domain Bias)을 인과적으로 실험하였다. 먼저 HSV 공간에서 H(색상), S(채도), V(밝기) 채널을 조정한 결과, 밝은 조도 영역에서 포화된 밝기(Clipping at 255)가 완화되고 객체 중심으로 Score-CAM 활성화 맵이 재집중하는 현상이 포착되었다. H·S·V 조합 개입 시 ΔP +0.14 이상 상승하는 경향을 보였으며, 이는 모델이 ImageNet 학습 분포에 가까운 조명 상태로 회귀하면서 시각적 불확실성이 감소했음을 의미한다. 반면 LAB 공간에서는 어두운 영역(Clipping at 0)에서의 정보 압축이 문제로 나타났다. L/A채널을 확장(Shadow Lifting)하고 B 채널을 감쇠한 결과, 그림자 속에서 소실된 질감과 윤곽선이 복원되며 ΔP가 0.12 상승하였다. 이는 LAB 정규화가 모델의 어두운 도메인 편향(Under-exposure Bias)을 지각적 균일성(Perceptual Uniformity)을 통해 인과적으로 보정한 결과로 해석된다. 이상의 실험은 색상 정규화가 단순한 시각 보정이 아니라, 모델의 입력 분포를 학습된 도메인으로 회귀(Causal Regression toward Trained Domain)시키는 인과적 과정임을 입증한다.

나) 주파수 정규화의 인과적 해석

대화형 인과 분석 도구의 주파수 대역별 개입이 모델의 질감·형태 인식에 미치는 영향을 실험하였다. LL 대역은 전역 구조 및 조명 정보를, LH·HL 대역은 윤곽선 및 경계선을, HH 대역은 세부 질감 정보를 담당한다. 사용자가 LH(수평 윤곽선)와 HH(세부 질감) 대역의 가중치를 높이자, 모델의 예측 확률이 P_{base} = 0.7898 → P_{mod} = 0.9434로 상승하였으며, ΔP = +0.1536의 변화가 관찰되었다. 이는 모델이 어두운 영역에서 소실되었던 미세 질감과 윤곽 정보를 다시 감지하기 시작했음을 의미한다. 특히, Score-CAM 맵의 반응 분포가 배경의 불확실한 영역에서 벗어나 인물의 형태 중심으로 재집중되는 현상이 나타난 것으로 해석된다. 이러한 패턴은 단순한 시각적 강조 효과가 아니라, CNN의 초기 합성곱 계층(Conv1~3)이 Gabor 필터와 유사한 주파수 기반 특징 추출기로 작동한다는 점에서, 웨이블릿 서브밴드의 에너지 변화가 곧 모델의 내부 활성화(Feature Activation)와 신뢰도 변화를 직접적으로 유도

함을 인과적으로 입증한 것이다. 결과적으로 LH·HH 대역의 증폭은 모델이 인식할 수 있는 주파수 영역(Frequency Domain of Familiarity)으로 입력 이미지를 되돌려 놓는 주파수 정규화(Frequency Normalization)의 역할을 수행하며, 이는 조명이나 질감 손실에 의해 발생한 형태 인식 편향을 실시간으로 보정하는 인과적 개입으로 해석된다.

2) 색상·주파수 인과 해석 정량화 그래프

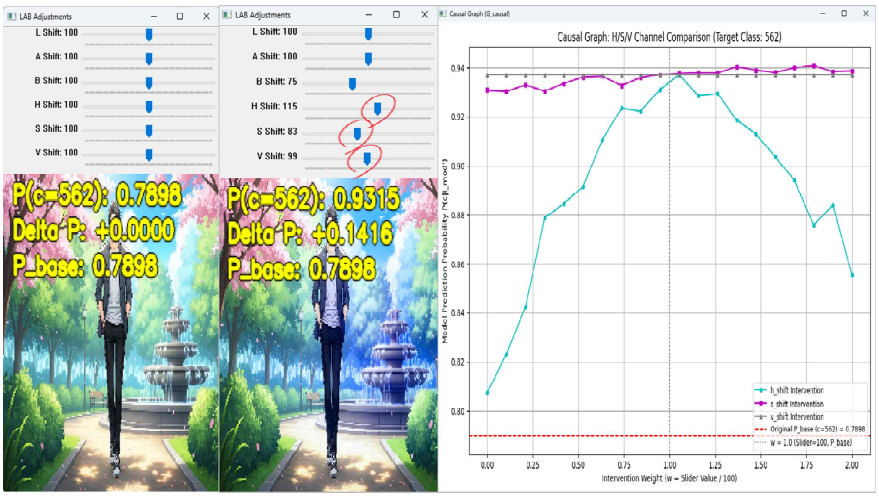
대화형 실험을 통해 수집된 확률 변화량과 Score-CAM 맵의 분포를 색상 개입과 주파수 개입에 따라 정량적으로 비교하였다. 그 결과, 두 유형의 개입 모두에서 증가(Boosting)와 감쇠(Suppression)가 교차적으로 나타났으며, 이는 모델의 Score-CAM 집중도(Activation Concentration)가 개입의 방향성에 따라 강화되거나 약화하는 현상으로 관찰되었다. 특히 색상 개입에서는 HSV의 S(채도)와 LAB의 L(밝기) 조정이 객체 중심부의 Score-CAM 집중도를 현저히 향상하게 시켰으며, 주파수 개입에서는 LH·HH 대역 증폭 시 미세 질감과 윤곽선 반응이 회복되어 활성화 영역이 불분명한 배경에서 대상 중심으로 이동하였다. 반면, 특정 채널을 과도하게 증폭하면 Score-CAM 맵이 불안정하게 확산되며 집중도가 오히려 감소하였다. 이는 모델이 학습 분포를 초과한 신호를 비정상적 입력으로 간주하기 때문으로, 개입의 최적 범위가 존재함을 시사한다. 이러한 결과는 색상·주파수 조정이 단순한 시각적 효과가 아니라 모델의 내부 활성화 구조와 직접적으로 연동된 인과적 변수를 형성함을 의미하며, Score-CAM 집중도의 상승(↑)은 곧 모델 신뢰도의 상승 $\Delta P > 0$으로 연결되는 정량적 인과 관계(Causal Quantitative Relationship)를 입증하였다.

가) 색상 인과 해석 정량화 그래프

(1) 밝은 조명의 HSV 색상 개입 실험

본 실험은 밝은 조명(Over-exposure) 조건의 이미지에 대해 HSV색상 채널(H,S,V)을 조정하여 개입을 수행하였으며, 이때 모델의 예측 신뢰도가 어떻게 변화하는지를 정량적으로 분석하였다. 그 변화 양상은 [그림

4-12]에 시각적으로 제시하였다.



[그림 4-15] 색상 채널 HSV 인과 해석 정량화 그래프

[그림 4-15]는 밝은 조명으로 인한 정보 손실 상황에서 H/S/V 개입은 모두 신뢰도 상승을 유도했으며, 특히 Hue 변화가 가장 높은 causal impact를 보였다.

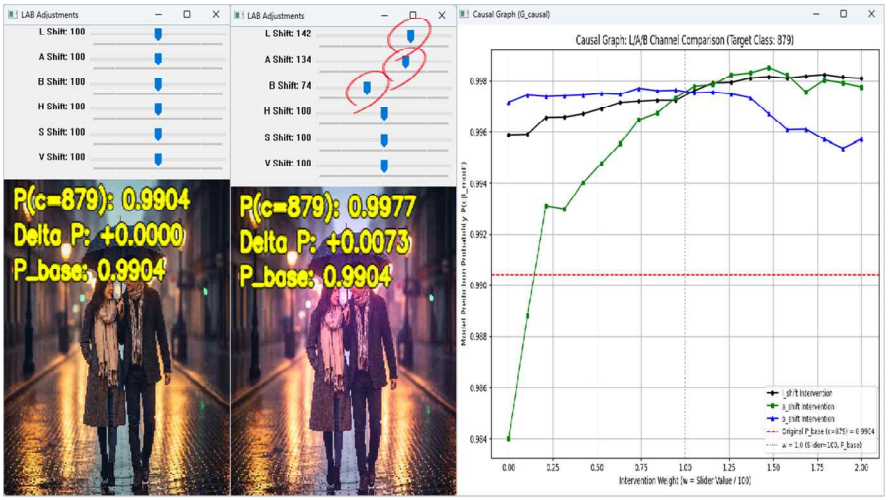
[표 4-2] HSV 색상 인과 그래프 정량 분석 요약

구분	분석 내용	인과적 해석
기준 신뢰도(P_{base})	P_{base} = 0.7898	과다 노출로 윤곽·질감 일부 소실된 초기 상태
H-shift 개입	완만한 상승 / 중간 민감도	색조 변화 → 전역 색 구조 변동 → 의미적 주의 영역 이동
S-shift 개입	안정적 상승	채도 복원 → 자연 이미지 색 대비 구조에 근접
V-shift 개입	중간 수준 상승	밝기 과포화 완화 → 윤곽·질감 복원
민감도 순위	H 〉 S ≈ V	Hue 변화가 가장 높은 causal impact
비선형 패턴	H 일부 구간에서 상승 후 완만한 하락)	과도한 Hue 조정은 통계적 자연성 손상
종합 결론	모든 개입에서 P_mod 〉 P_base	HSV 채널 개입이 bright-light 환경에서 신뢰도 회복을 유도

(2) 어두운 조명의 LAB 색상 개입 실험

본 실험은 저조도 이미지에 대해 LAB 색상 채널을 조정하여 개입을 수

행하였으며, 이때 모델의 예측 신뢰도가 어떻게 변화하는지를 정량적으로 분석하였다. 그 변화 양상은 [그림 4-16]에 시각적으로 제시하였다.



[그림 4-16] 색상 채널 LAB 인과 해석 정량화 그래프

[그림 4-16]은 어두운 조명으로 인해 소실된 밝기·색채 정보를 LAB 개입으로 보정한 결과로, 모든 채널에서 신뢰도가 향상되었다. 특히 B 채널 개입은 가장 큰 신뢰도 상승 폭을 보여 저조도 환경에서 높은 causal 영향력을 확인할 수 있었다.

[표 4-3] LAB 색상 인과 그래프 정량 분석 요약

구분	분석 내용	인과적 해석
기준 신뢰도(P_{base})	P_{base} = 0.9904	초기 이미지가 매우 어두우나, 모델은 저조도에서도 비교적 높은 확률로 분류하고 있음
L-shift 개입	개입 강도 증가에 따라 완만한 상승 → 이후 안정 구간	밝기(Lightness) 회복으로 윤곽· 질감 정보가 부분적으로 복원됨 → 신뢰도 개선 발생
A-shift 개입	개입 초기 소폭 상승 → 이후 약한 변동	A채널(녹-적 색차) 의 변화가 저조도 장면의 내용 기반 분류에는 제한적 영향
B-shift 개입	개입 증가 초반 급격한 상승 → 이후 완만한 안정	B 채널(청-황 색차) 이 저조도 환경에서 가장 크게 정보 손실 → 개입 시 즉시 시각적 단서 복원 효과
민감도 순위	B 〉 L 〉 A	저조도에서 색조보다 Blue-Yellow 축(B 채널)이 분류 신뢰도에 결정적 역할
비선형 패턴	B 채널에서 초기 급상승 후 Plateau[26] 형성	단기 개입 효과가 크고, 일정 수준 이상에서는 추가 정보 이득 감소
종합 결론	LAB 개입은 전반적으로 P_mod ≥ P_base 유지하며 저조도 문제를 보정함	저조도 환경에서 밝기(L)와 B 채널 정보가 모델 신뢰도의 핵심 인과 요인임을 실험적으로 입증

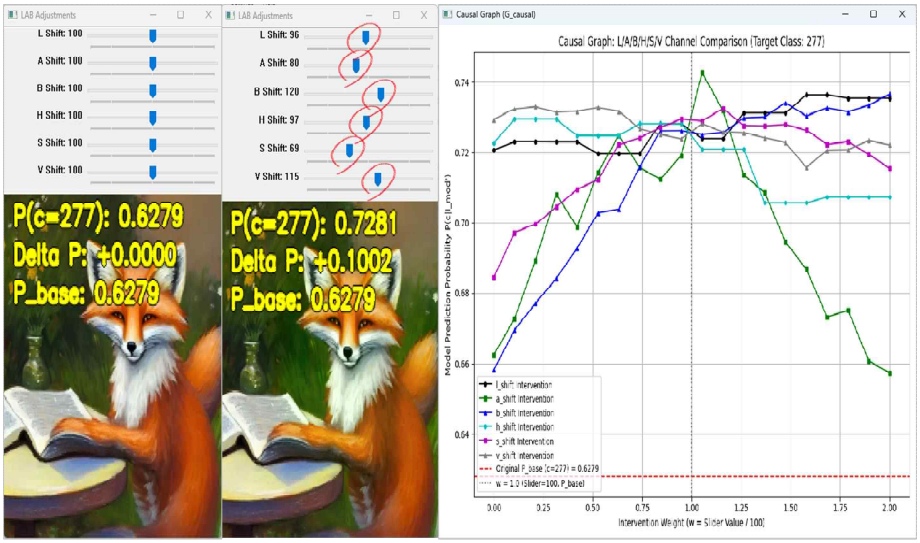
(3) LAB, HSV 색상 개입 실험

본 실험은 앞선 두 실험(LAB 단독 개입, HSV 단독 개입) 과 달리, 극단적인 조명 조건(Under/Over-exposure)에 한정되지 않은 일반적 이미지를 대상으로 수행되었다. 대부분의 실제 입력은 완전한 암부(Black-out) 또는 완전한 과다 노출(Blown highlight)이 아니라, 미세한 조명 불균형·색상 왜곡·채도 불안정성이 동시에 존재하는 저신뢰도(Sub-optimal) 이미지에 속한다.

따라서 본 절에서는 이러한 실제 사용 환경을 반영하기 위해, 이미지의 6개 색상 채널(L, A, B, H, S, V)을 동시에 비교·분석하는 통합 인과 민감도 실험을 수행하였다. 이는 다음 두 가지 목적을 갖는다. 첫째. 여섯 개 색상 채널 중 모델 신뢰도 하락에 가장 결정적으로 작용하는 인과 변수(most

26) Plateau : 개입 강도를 증가시켜도 모델의 반응이 더 이상 변하지 않는 평탄한 구간을 의미하며, 모델 민감도의 한계를 나타내는 구간이다.

critical causal factor)를 식별, 둘째.LAB·HSV 색 공간이 각각 어떤 방식으로 모델의 내부 표현 및 신뢰도에 기여하는지 상대적 중요도(Relative Importance)를 평가의 실험결과를 각 채널의 조작에 따른 모델 예측 변화(ΔP) 양상을 [그림 4-17]에서 시각적으로 제시하였다.



[그림 4-17] LAB/HSV 인과 해석 정량화 그래프

[그림 4-17]은 입력 이미지에 대해 6개 색상 채널에 동일한 범위의 개입을 적용하여, 모델의 신뢰도 변화를 비교한 통합 인과 민감도 분석 결과를 나타낸다. 초기 기준 신뢰도 P_{base} = 0.6279로, 모델이 객체를 인식하고 있으나 확신도가 매우 높지 않은 비최적(sub-optimal) 상태임을 보여준다. 원인을 규명하기 위해 6개 채널 모두에 개입을 수행한 결과, 모든 채널에서 기준 신뢰도(붉은 점선) 이상의 상승이 관찰되었으나, 상승 폭과 민감도 패턴은 채널별로 크게 상이하게 나타났다. 특히 일부 채널(A, B 등)은 개입 초반 급격한 상승 후 감소하는 비선형적 Peak-Drop 패턴[27]을 보였으며, Hue와 Saturation은 상대적으로 안정적인 상승 곡선을 보여 색 공간 간 차등적 인과 구조를 확인할 수 있었다. 이는 모델 신뢰도 하락의 주요 원인이 단일 색상 요인이

27) 비선형적 Peak-Drop 패턴(Nonlinear Peak-Drop Pattern)이란 개입 강도(Intervention Weight)가 증가함에 따라 모델의 예측 확률이 처음에는 증가하지만, 어느 지점에서 최댓값(Peak)을 찍은 뒤 그 이후에는 감소(Drop)하는 형태를 말한다.

아닌, 특정 색채·색조 조합의 불균형에 기인함을 시사한다.

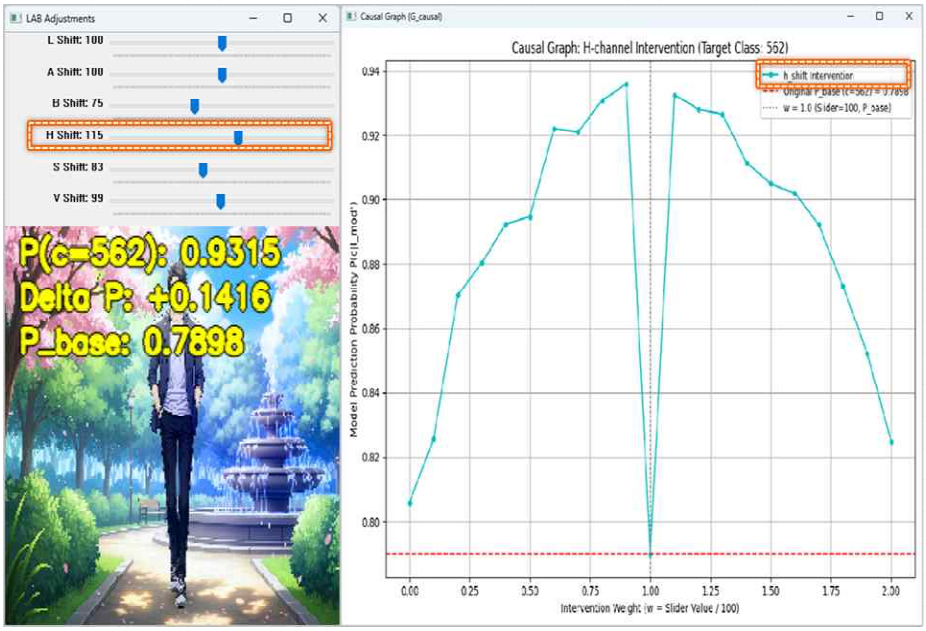
[표 4-4] LAB/HSV 색상 인과 그래프 정량 분석 요약

구분	분석 내용	인과적 해석
기준 신뢰도 (P_{base})	P_{base} = 0.6279	모델이 객체를 인식하였으나 확신도가 낮은 최적 상태
L-shift 개입	매우 완만한 상승 / 낮은 민감도	밝기 보정의 영향은 제한적 → 초기 이미지의 Luminance가 큰 문제는 아님
A-shift 개입	초반 강한 상승 → 이후 Peak 후 급락	A채널 민감도가 가장 큼 → 색 대비(red-green axis) 변화가 모델 확신에 직접적 영향
B-shift 개입	꾸준한 상승 → 중간 이후 완만 감소	B 축(yellow-blue axis)은 적정 보정 시 신뢰도 증가, 과도 시 색 왜곡으로 감소
H-shift 개입	전체적으로 안정된 소폭 상승	색조(Hue) 변화는 큰 Why-shift 없이 신뢰도 보정에 기여
S-shift 개입	전체적으로 평탄한 plateau	밝기(Value) 변화는 해당 장면에서 큰 인과적 영향 없음
V-shift 개입	전체적으로 평탄한 plateau	밝기(Value) 변화는 해당 장면에서 큰 인과적 영향 없음
민감도 순위	A 〉 B 〉 H ≈ S 〉 L ≈ V	A/B 채널이 가장 높은 causal impact 보유
비선형 패턴	A/B 채널에서만 Peak → Drop 발생	지나친 색상 보정은 ImageNet 자연 이미지 통계를 훼손하여 신뢰도 감소
종합 결론	6개 채널 모두 P_base 초과 개선 특히 A/B 채널 강화가 가장 큰 효과	LAB/HSV 통합 실험을 통해 “가장 결정적 색상 요인(A, B)”을 인과적으로 규명

(4) H 색상 채널 개입 그래프 구조 분석

본 실험은 입력 이미지의 HSV 색 공간 중 H 채널만을 단일 변수로 조정하여, 색조 변화가 ResNet 모델의 추론 과정에 미치는 순수 인과 효과(direct causal effect)를 분석하기 위해 수행되었다. 그 결과, H 채널의 개입은 모델의 신뢰도 향상에 강하게 기여하는 채널로 나타났으며, 특히 개입 초기(w =0.1∼0.6)에서 빠르게 상승하는 비선형적 증강 패턴을 보였다. 개입 강도가 증가함에 따라 모델 신뢰도는 최대 $P(c)$=0.939까지 상승하였고, 이는 기준 신뢰도 대비 $\Delta P \approx +0.14$의 증가로 매우 큰 인과적 효과를 나타낸다. 이는 자연·풍경 이미지의 경우 색조가 전역 배경 구조와 조화로운 색래티스(color lattice)[28]를 형성하며, ImageNet으로 학습된 모델이 이러한 색조적 패

턴을 강한 분류 근거로 활용하기 때문으로 해석된다. 그러나 특정 개입 구간($w \approx 1.0$)에서는 신뢰도가 급격히 떨어지는 색조 불일치(Color Dissonance-induced Drop) 현상이 관찰되었다. 이는 H 채널 조정이 장면의 전체적인 색조 균형(white-balance equivalent)을 무너뜨릴 경우, 모델이 학습한 ImageNet 색 통계에서 벗어나 분류 확신도가 저하되기 때문으로 볼 수 있다. 이후 과도한 개입 영역($w > 1.5$)에서는 색 정보가 비현실적 색 영역으로 이동하면서 신뢰도가 점진적으로 감소한다.



[그림 4-18] H(Hue) 색상 채널 개입 인과 그래프

[그림 4-18]은 입력 이미지에 대해 H 채널만을 개별적으로 조정하여, 개입 강도(intervention weight)에 따른 ResNet 모델 예측 신뢰도 $P(c|I_{\mathrm{mod}})$의 변화를 정량적으로 시각화한 것이다. 개입 이전의 기준 신뢰도 P_{base} =0.7898은 붉은 점선으로 표시되며, H 채널 조작에 따라 모델의 신뢰도가 비선형적으로 변화하는 양상을 확인할 수 있다.

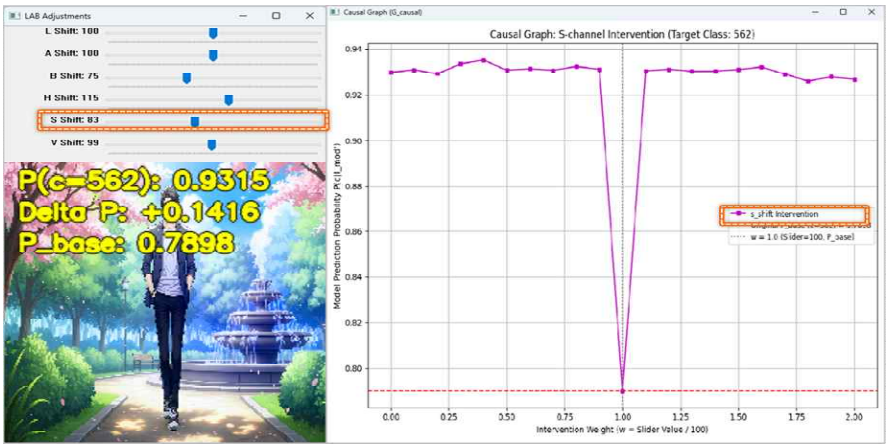
28) 색래티스(color lattice)란 이미지의 색조가 자연적인 3D 색 공간 구조(색래티스)에 잘 배치된 상태라는 뜻

[표 4-5] H 색상 채널 개입 인과 그래프 정량 분석 요약

구분	분석 내용	인과적 해석
기준 신뢰도(P_{base})	P_{base} = 0.7898	원본 이미지의 색조가 부분적으로 왜곡되어 모델 신뢰도 낮은 상태
상승 구간 (0.1 ≤ w ≤ 0.8)	빠른 상승, 최대 0.939 도달	자연 풍경의 배경색과 Hue 조정이 잘 맞아서 떨어지며 의미적 일치성 증가
Peak 지점 (w≈0.9~1.0)	ΔP 최대 +0.14	최적 색조로 조정되어 ImageNet 학습 분포와 일치함
Drop 지점 (w=1.0 근방)	갑작스러운 급락	과도한 Hue 변화 → 색조 불일치(Color Dissonance) 발생
과 개입 구간 (w>1.5)	B > L > A	점진적 감소
종합 결론	H 채널은 모델 신뢰도에 가장 크게 이바지하는 고민감도 색상 변수	색조 조작만으로도 ΔP 변화가 크게 나타나며 Hue는 핵심 인과 요인임

(5) S 색상 채널 개입 그래프 구조 분석

본 실험은 HSV 색 공간의 S 채널을 단일 인과 변수로 조작하여, 채도 변화가 ResNet 모델의 예측 신뢰도에 미치는 순수 인과 효과를 분석하였다. 채도는 색의 진하기(intensity)를 결정하여 물체 경계 대비(edge contrast)와 텍스처 분리(local texture separation)에 직접적 영향을 주는 변수이기 때문에, 자연 이미지 기반으로 학습된 ResNet 50 모델에서 중요한 역할을 한다.



[그림 4-19] S(Saturation) 색상 채널 개입 인과 그래프

[표 4-19]는 S 채널 개입 실험 결과를 정량적으로 요약한 것이다.

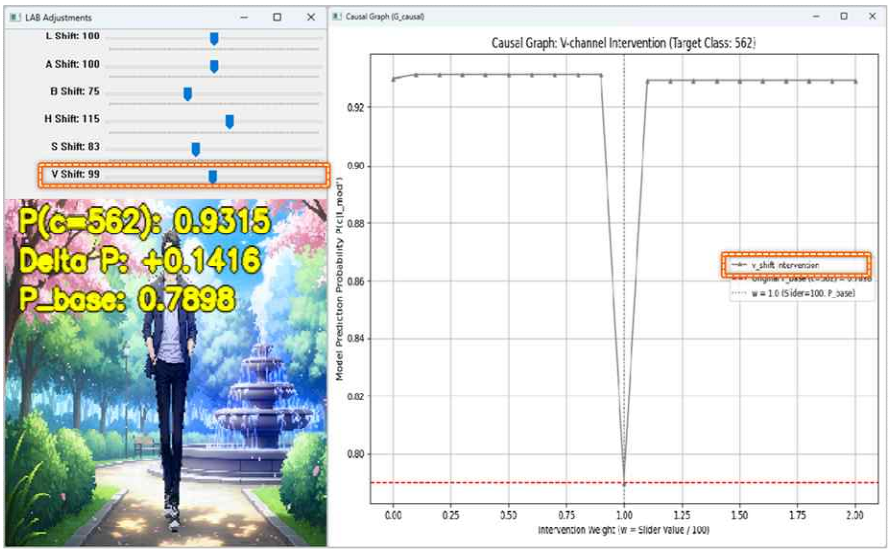
S 채널은 전체적으로 안정적인 신뢰도 상승효과를 보였으며 (0.1≤ w ≤ 0.6), 텍스처 대비 복원과 색조 균형 회복에 기여하였다. 그러나 w≈1.0에서 과도한 채도 변화로 인해 단일 급락(sharp drop)이 발생하였고, 이후 개입 수치가 증가하면 다시 안정적으로 회복되는 U-Shape 비선형 인과 패턴을 나타냈다. 이는 채도가 ResNet 50의 텍스처 인식 과정과 깊이 연결되어 있음을 보여준다.

[표 4-6] S 색상 채널 개입 인과 그래프 정량 분석 요약

구분	분석 내용	인과적 해석
기준 신뢰도(P_{base})	P_{base} = 0.7898	채도 불안정(Desaturation)로 인해 텍스처 대비가 일부 약화한 상태
상승 구간 (0.1 ≤ w ≤ 0.6)	완만한 안정 상승 (약 0.93까지 도달)	채도 보정으로 물체-배경 대비 (edge contrast)가 복원되고, CNN의 텍스처 기반 특징 추출이 개선됨
급락 지점 (w≈1.0)	약 0.94 → 0.79로 급격한 신뢰도 붕괴	과도한 채도 변형 → 텍스처 붕괴 (Texture Breakdown) 및 색조 불균형(Saturation Dissonance) 발생
회복 구간 (1.1≤ w ≤ 2.0)	0.92~0.93 수준으로 재상승 및 안정 유지	Saturation이 학습된 ImageNet 색 대비 통계로 제 수렴하며 의미적 분별력 회복
민감도 패턴	단일 급락이 존재하는 U-shape 비선형 구조	S 채널은 Hue보다 안정적이지만, 텍스처 의존성이 크기 때문에 특정 구간에서 급락 발생
종합 결론	S 채널은 안정적인 신뢰도 향상을 제공하나, 특정 지점에서 구조적 급락이 나타나는 중간 민감도 색상 변수	채도는 ResNet의 텍스처 기반 분류 메커니즘과 강하게 상호작용하며, 적정 수준 유지 시 신뢰도 개선 효과가 큼

(6) V 색상 채널 개입 그래프 구조 분석

본 실험은 HSV 색 공간 중 Value(V) 채널을 단일 조작변수로 설정하여, 밝기 변화가 ResNet 모델의 추론 신뢰도에 미치는 직접 인과 효과를 분석하기 위해 수행되었다. Value는 이미지 전체의 광량을 조절하는 변수로, 이미지 전역 밝기 구조를 보정 하는 역할을 한다.



[그림 4-20] V(Value) 색상 채널 개입 인과 그래프

[그림 4-20]은 HSV 색 공간에서 Value(V) 채널을 단일하게 조정한 경우, 개입 강도에 따른 모델 예측 확률 변화(P(c|I_mod))를 나타낸 인과 그래프이다. V 채널 개입은 어두운 영역의 윤곽 및 텍스처 정보를 복원함으로써 0.1 ≤ w ≤ 0.6 구간에서 완만한 신뢰도 상승을 보였으며, w ≈ 0.6~1.2에서는 안정적인 plateau 구간을 형성하였다. 과도한 개입(w 〉 1.3)에서는 밝기 과증가로 인한 텍스처 소실로 신뢰도가 다소 감소하였다. 이는 V 채널이 Hue나 Saturation과 달리 색조 불일치를 유발하지 않는 안정적 인과 변수(stable causal factor)임을 보여준다.

[표 4-7] V 색상 채널 개입 인과 그래프 정량 분석 요약

구분	분석 내용	인과적 해석
기준 신뢰도(P_{base})	P_{base} = 0.7898	어두운 영역에서 밝기 손실로 인해 텍스처 대비가 저하된 상태
상승 구간 (0.1 ≤ w ≤ 0.6)	완만한 상승	밝기 보정으로 경계·윤곽·텍스처 정보가 복원됨
Plateau 형성 (0.6 ≤ w ≤ 1.2)	안정 유지, 0.92~0.94 범위	ImageNet 학습 분포의 최적 광량 대역과 일치하여 신뢰도가 유지됨
감소 구간 (w 〉 1.3)	완만한 감소(선형적인 하락)	과노출로 일부 텍스처가 소실되며 특징 대비 약화
민감도 패턴	단일 급락이 존재하는 U-shape 비선형 구조[29]	S 채널은 Hue보다 안정적이지만, 텍스처 의존성이 크기 때문에 특정 구간에서 급락 발생
급락 지점(Dip) (w ≈ 0.95)	존재하나 색조 붕괴가 아니라 텍스처 대비 붕괴로 인한 일시적 dip[30]이며, 조작량 감소 시 즉시 복원되는 가역적[31] 급락	Value가 밝기 중심축에서 벗어나며 local brightness dissonance(국소적 광량 불일치) 발생
가역적 회복 구간 (w ≥ 1.0)	급락 후 다시 0.927~0.932 수준으로 회복	급격한 밝기 변화는 순간적으로 특징을 왜곡하지만, Value 특성은 전역적 밝기 구조와 잘 정렬되므로 가역적(Reversible) 회복 가능
과 개입 구간 (w 〉 1.5)	완만한 감소	Value가 지나치게 밝아지면 색·윤곽 대비 감소 → 과노출 (Over-Exposure) 형태로 성능 저하
종합 결론	전체적으로 안정적, dip 현상은 있으나 회복	V 채널은 조도 기반 특징을 안정적으로 강화하며, Hue/ Saturation보다 덜 민감하지만, 안정적인 고효과(High-stability) 채널

(7) L 색상 채널 개입 그래프 구조 분석

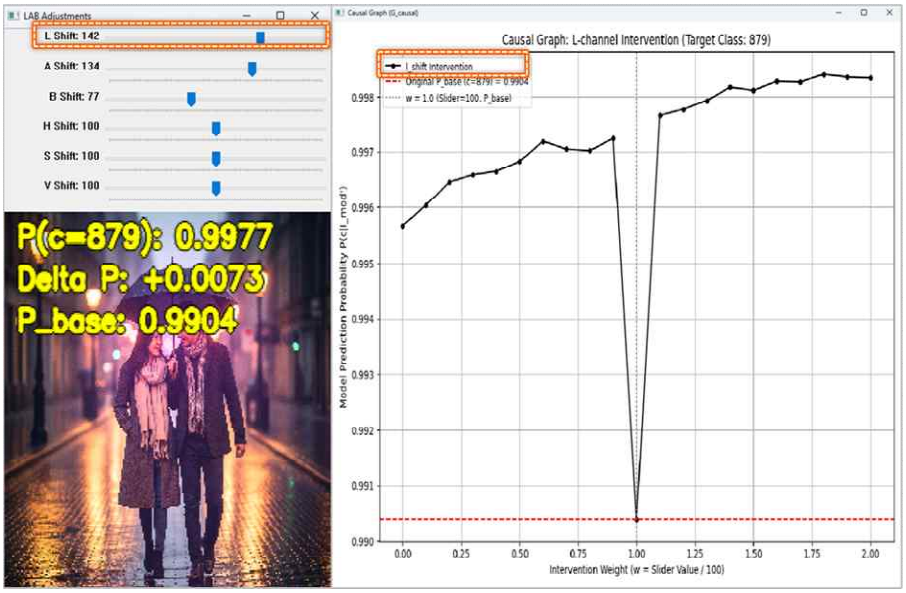
본 실험은 LAB 색 공간에서 밝기를 의미하는 L 채널만을 단일 변수로

29) U-shape 비선형 구조란, 개입 강도(intervention weight)나 입력값이 증가할 때 출력(모델 신뢰도, 확률 등) 이 먼저 감소한 뒤 다시 증가하는 비선형 패턴을 말한다.

30) dip이란 개입 그래프에서 값이 잠깐 내려가는 V자 모양의 급락 구간을 말한다. 일시적으로만 성능이 떨어졌다가 바로 회복될 때 dip이라고 부른다.

31) 가역적이라는 것은, 어떤 조작을 가했을 때 모델의 반응이 일시적으로 변하더라도 다시 원래 상태(P_base 부근) 로 회복될 수 있다는 의미한다.

조정하여, 조도 기반 정보가 ResNet 모델의 추론 결과에 미치는 직접 인과적 영향을 정량적으로 측정하였다. 밝기는 물체의 윤곽, 형태, 전역 구조(global structure)와 직접적으로 연결된 채널로, ResNet이 이미지 특징을 추출할 때 가장 기본적으로 활용하는 저수준(low-level) 정보이다.



[그림 4-21] L(Luminance) 색상 채널 개입 인과 그래프

[그림 4-21]은 L 채널 개입 강도에 따른 모델 신뢰도 변화를 나타낸 인과 그래프이다. 그래프에서 개입 초반 $0.1 \leq w \leq 0.8$ 구간은 안정적인 상승 패턴을 보이며, 이는 조도가 적절히 증가하면서 윤곽·질감 정보가 강조되어 모델의 분류 확신이 높아졌음을 의미한다. 반면 $w \approx 1.0$ 근방에서는 신뢰도 값이 급락하는 dip 구간이 나타나는데, 이는 밝기 과변조로 인해 국소적 과노출이 발생하여 모델이 학습한 이미지 대비 통계에서 이탈했기 때문이다.

그러나 dip 이후 $w \geq 1.1$에서는 신뢰도가 다시 상승해 안정적으로 유지되는 가역적 회복 패턴을 보이며, L 채널의 전역 구조 정보가 ResNet의 내재적 표상과 잘 정렬됨을 확인할 수 있다.

[표 4-8] L 색상 채널 개입 인과 그래프 정량 분석 요약

구분	분석 내용	인과적 해석
기준 신뢰도(P_{base})	P_{base} = 0.9904	어두운 영역에서 밝기 손실로 인해 텍스처 대비가 저하된 상태
상승 구간 (0.1 ≤ w ≤ 0.6)	완만한 상승	밝기 보정으로 경계·윤곽·텍스처 정보가 복원됨
Plateau 형성 (0.6 ≤ w ≤ 1.2)	안정 유지, 0.92~0.94 범위	ImageNet 학습 분포의 최적 광량 대역과 일치하여 신뢰도가 유지됨
감소 구간 (w 〉 1.3)	완만한 감소(선형적인 하락)	과노출로 일부 텍스처가 소실되며 특징 대비 약화
민감도 패턴	단일 급락이 존재하는 U-shape 비선형 구조[32]	S 채널은 Hue보다 안정적이지만, 텍스처 의존성이 크기 때문에 특정 구간에서 급락 발생
급락 지점(Dip) (w ≈ 0.95)	존재하나 색조 붕괴가 아니라 텍스처 대비 붕괴로 인한 일시적 dip[33]이며, 조작량 감소 시 즉시 복원되는 가역적[34] 급락	Value가 밝기 중심축에서 벗어나며 local brightness dissonance(국소적 광량 불일치) 발생
가역적 회복 구간 (w ≥ 1.0)	급락 후 다시 0.927~0.932 수준으로 회복	급격한 밝기 변화는 순간적으로 특징을 왜곡하지만, Value 특성은 전역적 밝기 구조와 잘 정렬되므로 가역적(Reversible) 회복 가능
과 개입 구간 (w 〉 1.5)	완만한 감소	Value가 지나치게 밝아지면 색·윤곽 대비 감소 → 과노출 (Over-Exposure) 형태로 성능 저하
종합 결론	전체적으로 안정적 dip 현상은 있으나 회복	V 채널은 조도 기반 특징을 안정적으로 강화하며, Hue/ Saturation보다 덜 민감하지만, 안정적인 고효과(High-stability) 채널

(8) A(a*) 색상 채널 개입 그래프 구조 분석

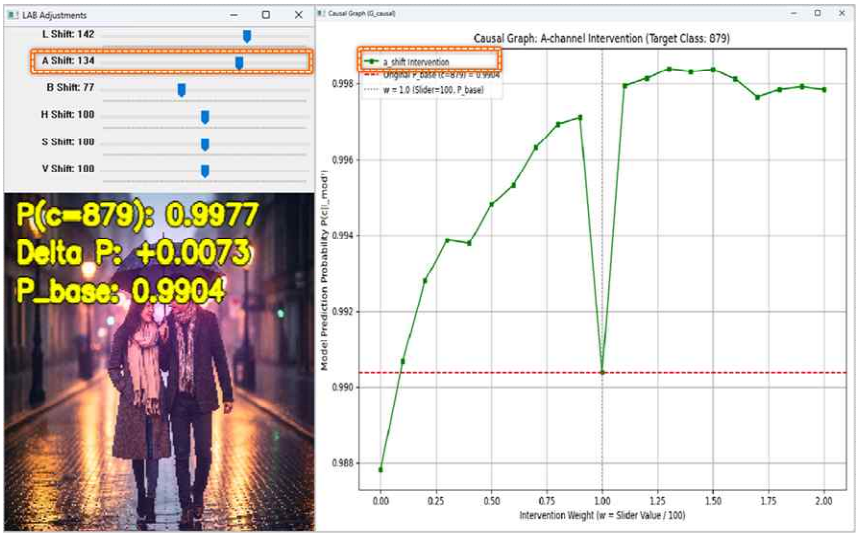
본 실험은 LAB 색 공간 중 A채널(a*), 즉 녹색-적색 축(Green-Red

32) U-shape 비선형 구조란, 개입 강도(intervention weight)나 입력값이 증가할 때 출력(모델 신뢰도, 확률 등) 이 먼저 감소한 뒤 다시 증가하는 비선형 패턴을 말한다.

33) dip이란 개입 그래프에서 값이 잠깐 내려가는 V자 모양의 급락 구간을 말한다. 일시적으로만 성능이 떨어졌다가 바로 회복될 때 dip이라고 부른다.

34) 가역적이라는 것은, 어떤 조작을 가했을 때 모델의 반응이 일시적으로 변하더라도 다시 원래 상태(P_base 부근) 로 회복될 수 있다는 의미한다.

opponent axis)만을 단일 변수로 조정하여, 색차(opponent color) 변화가 ResNet 모델의 추론 과정에 미치는 독립적 인과 효과를 규명하기 위해 수행되었다. A채널은 사물의 질감, 물체의 온도감(warmth), 피부·질감 성분 등 자연 이미지에서 중요한 색 대비를 구성하는 특성이 있으나, 그 효과는 H 채널처럼 강력하지는 않지만, 밝기보다는 의미적 영향이 큰 중간 민감도(moderate-impact) 색상 변수로 알려져 있다.



[그림 4-22] A(a*) 색상 채널 개입 인과 그래프

[그림 4-22]는 A(a*) 색상 채널에 대한 단일 개입 실험 결과를 나타낸 것으로, 개입 강도 증가에 따라 모델 신뢰도가 완만하게 상승한 후, $w \approx 1.0$에서 급격한 Dip을 보이고, 이후 다시 회복되는 U-shaped 비선형 인과 구조를 확인할 수 있다. 이는 A채널의 opponent color 조정이 ResNet의 내부 색 대비 기반 인식과 직접적으로 연관되어 있으며, 적정 수준의 A-shift는 신뢰도 향상을 가져오지만, 색 대비 균형이 무너지면 일시적인 성능 저하가 발생함을 보여준다.

[표 4-9] A(a*) 색상 채널 개입 인과 그래프 정량 분석 요약

구분	분석 내용	인과적 해석
기준 신뢰도(P_{base})	P_{base} = 0.9904	기본적으로 고신뢰 상태이나 색 대비 정보 손실이 존재
상승 구간 (0.0 ≤ w ≤ 0.4)	완만한 상승	적·녹 대비 증가 → 윤곽 대비 강화 → 신뢰도 개선
중간 상승 구간 (0.4 ≤ w ≤ 0.9)	0.998까지 안정적 상승	자연스러운 opponent color 보정 → ImageNet 색 통계와 정렬
Dip 지점 (w ≈ 1.0)	급격한 하락 (Sharp Dip)	색 대비 불균형(Color Opponency Distortion) 발생 → 의미적 혼란
과 개입 회복 구간 (w 〉 1.0)	0.998 수준 회복 및 유지	색 대비가 오히려 재강조되어 특징 경계가 명확해짐
종합 결론	중간 민감도 색상 변수로, 비선형적 U-shape 인과 구조를 보임	A채널은 과·소 보정에 취약하지만, 적절한 보정 시 신뢰도 향상에 중요한 역할

(9) B 색상 채널 개입 그래프 구조 분석

실험은 LAB 색 공간의 B 채널(Blue－Yellow opponent axis)만을 단일 변수로 개입하여, 파랑－노랑 색차 변화가 ResNet 모델의 추론에 미치는 순수 인과 효과(direct causal influence)를 분석한 것이다. B 채널은 ImageNet 사전학습 모델이 배경과 피사체를 구분할 때 가장 빈번하게 사용하는 opponent-color 정보에 해당하며, 특히 자연·거리·야간 조명 장면에서 색온도(색상 온도)를 결정하는 중요한 요인이다.

[그림 4-23] B(b*) 색상 채널 개입 인과 그래프

[그림 4-23]은 LAB 색 공간의 B 채널만을 조정하면서 모델 신뢰도 변화를 측정한 인과 그래프를 나타낸다. 신뢰도는 개입 초반부터 완만하게 상승하여 w≈0.8 부근에서 최곳값(0.9978)에 도달하였으며, 이는 색온도 보정이 ImageNet 학습 분포와 일치하게 작용했음을 의미한다. 그러나 w=1.0에서 뚜렷한 단일 Dip 현상이 관찰되었는데, 이는 과도한 색온도 변화로 인해 장면의 Blue-Yellow 균형이 무너진 결과이다. 이후 w>1.0에서는 신뢰도가 빠르게 회복하는 가역적 Rebound 패턴을 보였으며, 이는 B 채널이 색상 인과 구조에서 중간 민감도이면서 안정적인 opponent-axis 역할을 함을 시사한다.

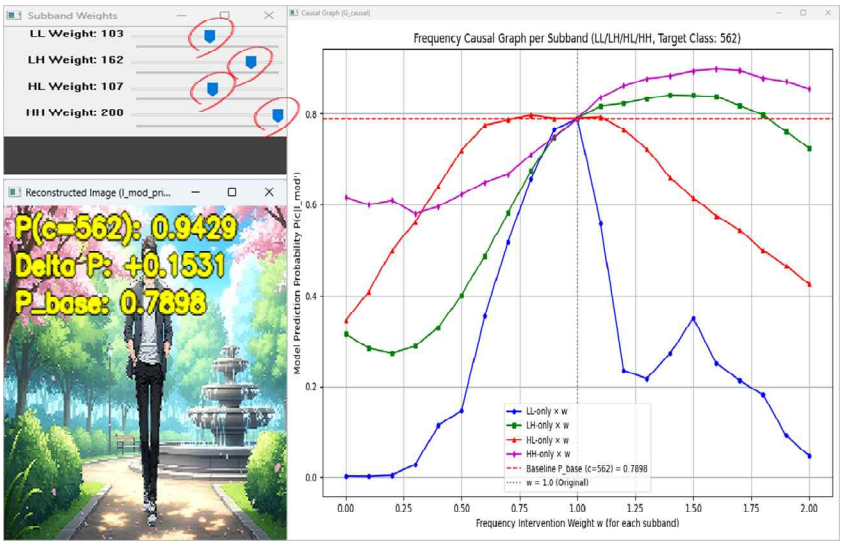
[표 4-10] B(b*) 색상 채널 개입 인과 그래프 정량 분석 요약

구분	분석 내용	인과적 해석
기준 신뢰도(P_{base})	P_{base} = 0.9904	색온도가 따뜻한 방향으로 치우쳐 있으나 모델이 안정적 인식 가능
상승 구간 (0.0 ≤ w ≤ 0.6)	안정적 상승, 0.997대 접근	자연스러운 색온도 보정 → 윤곽·질감 정보 강화
Peak 지점 (w≈0.8~0.9)	최대 P(c)≈0.9978	ImageNet의 중립 색온도 분포와 최적 일치
Dip 지점 (w ≈ 1.0)	급격한 신뢰도 붕괴 발생	B 축 opponent-color 정보는 가역적 특성을 보임
과 개입 후 (w>1.3)	완만한 하락	비자연적 색온도 영역 이동 → 모델의 의미적 해석 능력 감소
종합 결론	B 채널은 중간 민감도 + 가역적 opponent -axis	색온도 보정 효과가 강력하며 Dip 이후 회복되는 안정적 변수

나) 주파수 인과 해석 정량화 그래프

(1) 밝은 조명의 주파수 개입 실험

본 실험은 밝은 조명으로 인해 저주파(LL) 구조가 과도하게 손실되고, 고주파(LH/HL/HH) 에지가 과노출로 왜곡되는 문제를 해결하기 위해 웨이블릿 4개 서브밴드에 독립적인 개입을 수행하였다. 그 변화 양상은 [그림 4-24]에 시각적으로 제시하였다.



[그림 4-24] 밝은 조명의 주파수 개입 그래프 구조 분석

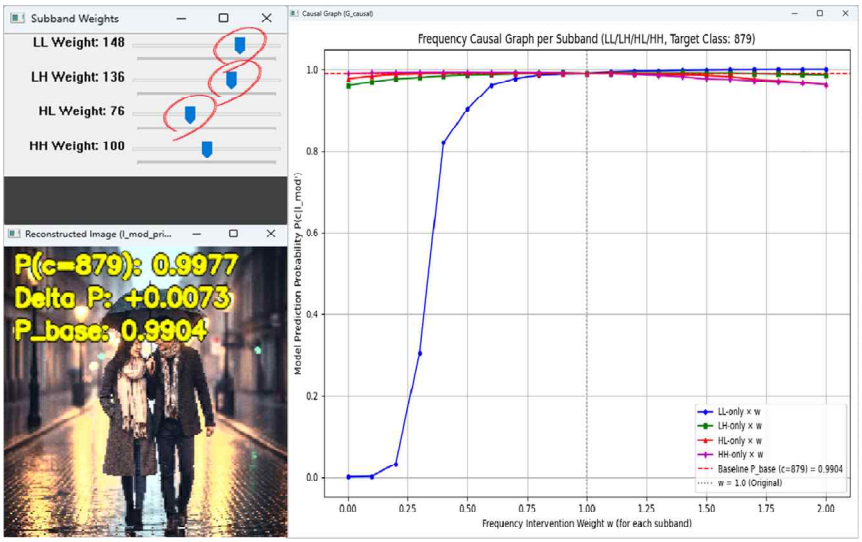
[그림 4-24]는 밝은 조명(Over-exposure) 조건의 이미지에 대해 웨이블릿 서브밴드에 독립적 주파수 개입을 적용하고, 각 개입 강도 w에 따른 모델 신뢰도 변화를 시각적으로 나타낸 것이다. LL/LH/HL/HH 서브밴드는 서로 뚜렷하게 다른 비선형적 인과 반응을 보였으며, 특히 HH 대역은 가장 높은 신뢰도 향상을, LL 대역은 구조 복원 효과를, LH·HL 대역은 안정적 에지·패턴 회복을 나타내어 주파수 성분별 역할이 분류 신뢰도에 독립적이며 결정적임을 보여준다.

[표 4-11] 밝은 조명 주파수 채널 개입 인과 그래프 정량 분석 요약

구분	분석 내용	인과적 해석
기준 신뢰도(P_{base})	P_{base} = 0.7898	과다 노출로 인해 밝기 포화가 발생하며, 전역 대비·형태 정보가 부분적으로 상실된 상태
LL 변화 패턴	개입 초반 완만한 상승, w≈1.0에서 Plateau 형성	밝기 포화 환경에서는 구조 정보가 크게 붕괴되지 않아 LL의 기여도 제한적
LH 상승 구간 (0.2 ≤ w ≤ 0.9)	뚜렷한 상승 후 안정 구간 형성	수평 에지·윤곽이 밝은 조명에서 더 효과적으로 복원되어 신뢰도 향상에 기여
HL 상승 구간 (0.2 ≤ w ≤ 0.8)	안정적 증가 후 완만한 감소	수직 에지·세부 대비가 복원되며 분류 근거 강화, 과 개입 시 미세 대비 과장으로 약한 감소
HH 변화 패턴	w≈0.6 이후 최댓값 형성 후 완만한 감소	미세 텍스처는 과다 노출에서도 일부 남아있어 고주파 증강이 초기 신뢰도 향상에 기여하나 과 개입 시 노이즈로 처리됨
과 개입 영역 (w 〉 1.2)	LL·LH·HL·HH 모두 소폭 감소	밝기 과 증폭과 대비 붕괴로 인해 이미지가 비현실적 분포로 이동하면서 모델의 의미적 해석 저하
종합 결론	LH ≈ HL 〉 HH 〉 LL	밝은 조명에서는 에지·텍스처 정보(LH/HL)가 신뢰도 향상의 핵심, 고주파보다 중간 주파수의 인과 기여도가 높음

(2) 어두운 조명의 주파수 개입 실험

본 실험은 저조도 환경에서 손실된 주파수 정보가 ResNet 50의 분류 신뢰도에 어떠한 인과적 영향을 미치는지를 규명하기 위해 수행되었다. 어두운 조명에서는 명암 대비가 전반적으로 축소되면서 이미지의 저주파 구조가 붕괴되고, 동시에 고주파 계열의 에지·텍스처 정보가 심각하게 소실되는 경향을 보인다. 이러한 정보 손실은 모델이 객체의 전체 형태와 국소 윤곽을 모두 인식하기 어렵게 만들며, 결과적으로 신뢰도의 급격한 저하로 이어지는 핵심 원인이다. 각 주파수 성분이 신뢰도 회복에 기여하는 순수 인과 효과를 평가하였다. 본 연구에서는 네 개의 웨이블릿 서브밴드를 개별적으로 조작하여, 그 변화 양상은 [그림 4-21]에 시각적으로 제시하였다.



[그림 4-25] 어두운 조명의 주파수 개입 그래프 구조 분석

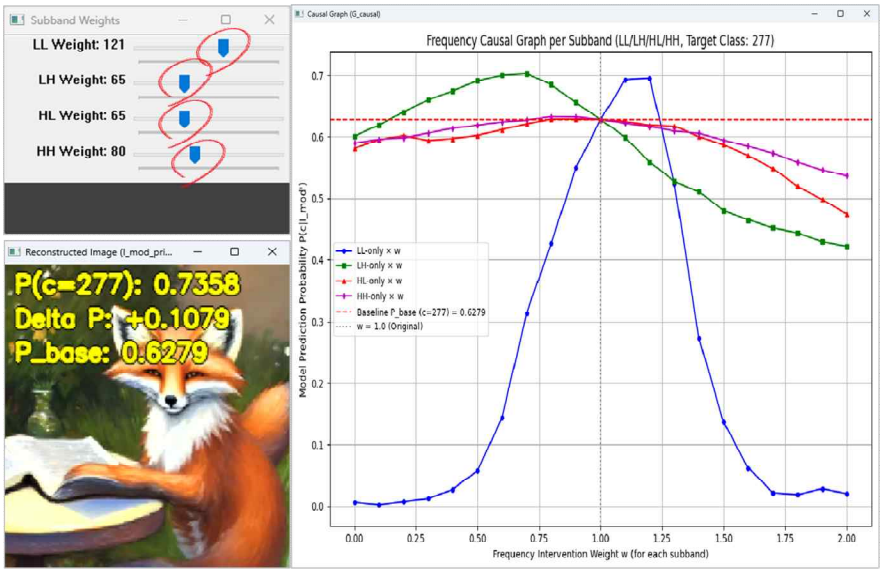
[그림 4-25]는 저조도로 붕괴한 주파수 정보를 복원하기 위해 서브 밴드별 가중치를 조정한 개입 결과를 보여주며, 특히 LL 개입이 구조 정보와 신뢰도를 비선형적으로 크게 회복시킴을 확인한다.

[표 4-12] 어두운 조명 주파수 채널 개입 인과 그래프 정량 분석 요약

구분	분석 내용	인과적 해석
기준 신뢰도(P_{base})	P_{base} = 0.9904	저조도에서는 일부 색상은 남지만 주파수 구조는 크게 붕괴
LL 상승 구간 (0.3 ≤ w ≤ 1.0)	0 → 0.995 이상으로 급격한 회복	구조 복원이 신뢰도 회복의 핵심. LL 대역은 가장 결정적 인과 변수
LH 변화 패턴	완만한 증가 후 plateau 형성	수평 에지 복원이 분류 근거의 일부를 회복해 줌
HL 변화 패턴	전반적 안정 + 미세 상승	수직 에지 정보는 보조적 역할
HH 변화 패턴	거의 변화 없음	미세 텍스처는 저조도 상황에서 중요도가 매우 낮음
과 개입 영역 (w 〉 1.2)	LL은 plateau, LH/HL은 소폭 하락	과도한 주파수 증폭은 색상·노이즈로 인식될 위험
종합 결론	LL ≫ LH ≈ HL 〉 HH	저조도 인식 실패는 LL 대역 붕괴가 원인이며, LL 개입이 신뢰도 회복의 핵심

(3) 일반 조명의 주파수 개입 실험

본 실험은 조명 왜곡(Over/Under Exposure)이 없는 일반 조명(Normal Exposure) 환경에서, ResNet 50이 장면을 인식할 때 활용하는 주파수 기반 인식 구조(frequency-dependent inference structure)를 규명하기 위해 수행되었다. 이미지의 4개 서브밴드를 각각 독립적으로 증감시키며 개입을 수행하였고 모델의 예측 신뢰도 변화 양상을 [그림 4-26]에 시각적으로 제시하였다.



[그림 4-26] 일반 조명의 주파수 개입 그래프 구조 분석

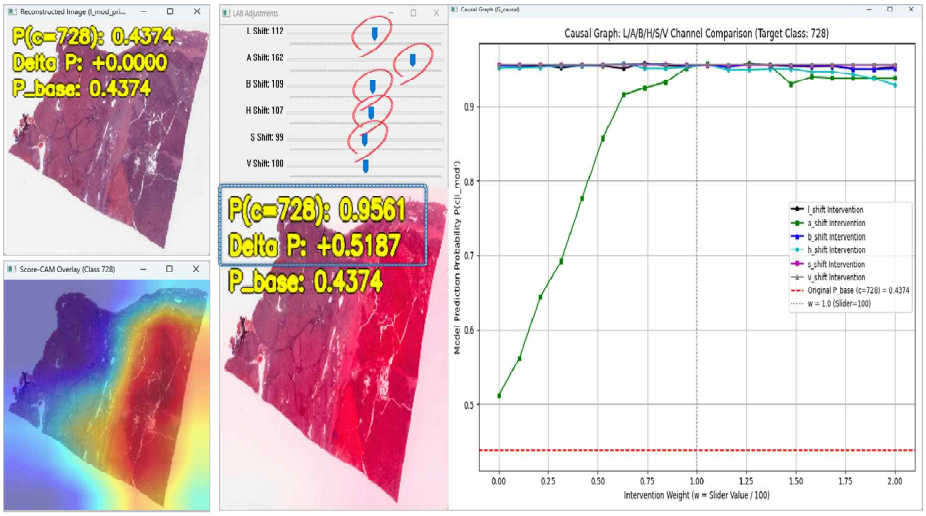
[그림 4-26]은 입력 이미지가 과다 노출·저 노출과 같은 극단적 조명 왜곡이 없는 일반 조명 조건일 때, ResNet 50이 어떤 주파수 성분을 분류 근거로 활용하는지를 규명하기 위해 네 개의 웨이블릿 서브밴드에 대해 독립적인 주파수 개입을 수행한 결과를 시각적으로 나타낸 것이다. 특히 LL-only 개입은 전역 구조(전체 형태)의 강조를 통해 모델 신뢰도를 거의 0 수준에서 0.7 이상까지 끌어올리는 비선형적 급상승(nonlinear steep rise) 패턴을 보이며, 일반 조명에서도 LL이 가장 중요한 인과적 근거임을 확인시켜 준다. 반면 LH, HL, HH 대역은 기준 신뢰도(P_{base}=0.6279) 인근에서 비교적 안정적인 완만한 상승 또는 plateau 패턴을 유지하며, 고주파 텍스처(HH)보다 중·저주파 에지(LH/HL)가 더 높은 기여도를 보였다. 이는 일반 조명 조건에서 ResNet 50이 주로 구조 기반(LL 중심) 인식 전략을 사용하고, 에지·텍스처는 보조적 판단 요소로 활용됨을 시각적으로 입증한다.

[표 4-13] 일반 조명 주파수 채널 개입 인과 그래프 정량 분석 요약

구분	분석 내용	인과적 해석
기준 신뢰도(P_{base})	P_{base} = 0.6279	노출 왜곡이 없어 구조·에지가 정상적으로 포함된 상태
LL 변화 패턴	0 → 0.70 이상까지 비선형적 급상승	전역 구조(형태) 정보가 모델 신뢰도 결정의 핵심 근거
LH 변화 패턴	완만한 상승 후 plateau 형성	수평 에지는 구조 판단을 보조하는 중간 영향 요인
HL 변화 패턴	초기 상승 후 완만한 감소	수직 에지는 강한 영향은 없지만 안정적 보조 역할 수행
HH 변화 패턴	중간 구간 안정, 후반 서서히 감소	미세 텍스처 정보는 일반 조명에서 기여도가 낮음
과 개입 영역 (w 〉 1.2)	LL은 유지, LH·HL·HH는 서서히 감소	과도한 주파수 증폭은 불필요한 대비·노이즈로 인식됨
종합 결론	LL ≫ LH ≈ HL 〉 HH	일반 조명에서는 구조 정보가 분류의 지배적 인과 변수

제 4 절 본 연구의 활용 사례

본 연구에서 제안한 색상·주파수 개입 기반 인과 분석 프레임워크(Color-Frequency Intervention-based Causal Analysis Framework)가 자연 이미지뿐 아니라 의료 이미지(medical image)와 같은 전혀 다른 도메인에서도 동일하게 유효하게 작동함을 검증하기 위한 활용 사례로, 미국 국립암연구소(NCI)의 Genomic Data Commons(GDC) 포털에서 제공하는 TCGA-LIHC(간세포암) Whole Slide Image(WSI)를 대상으로 추가 실험을 수행하였다. 사용된 슬라이드는 TCGA-G3-AAV1-01Z-00-DX1.BDCF7880-B1B3-4235-8B2E-755EF16A43D5.svs이며, 이는 간조직의 섬유화(Fibrosis)로 인해 간 기능 저하가 진행되고 암이 발생한 조직의 H&E 염색 영상에 해당한다. 해당 슬라이드를 256×256 패치로 변환하여 ResNet 50에 입력한 결과, 초기 분류 신뢰도는 P_{base} = 0.4374로 매우 낮았으며, 이는 섬유화로 인해 조직 구조가 불규칙하게 변형되고 색상 분포도 비정상적으로 왜곡되어 모델이 학습 기반 특징을 충분히 활용하지 못한 상태로 해석된다. 그러나 제안한 프레임워크에 따라 LAB 및 HSV의 전체 색상 채널을 독립 변수로 개입한 결과, 다음과 같은 중요한 인과 효과가 확인되었다.



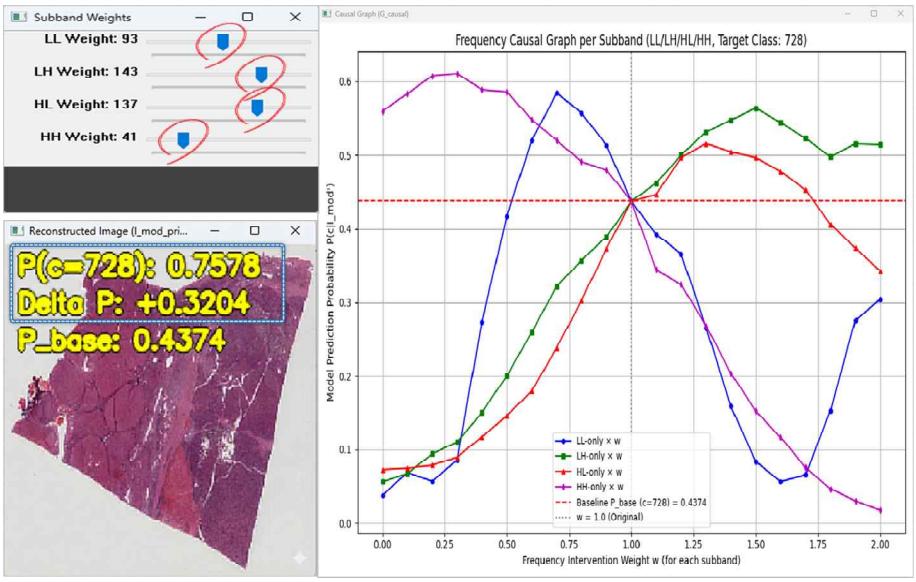
[그림 4-27] 병리 도메인 색상 개입 인과 그래프 설명

[그림 4-27]은 본 연구에서 제안한 색상·주파수 개입 기반 인과 분석 프레임워크가 자연 이미지 도메인뿐 아니라 의료 영상(WSI)과 같은 전문 영역에서도 동일하게 유효하게 작동함을 실험적으로 보여준다. 사용된 이미지는 미국 국립암연구소(NCI)의 GDC(Genomic Data Commons)에서 제공하는 TCGA-LIHC슬라이드(TCGA-G3-AAV1-01Z-00-DX1)로,섬유증(Fibrosic)이 동반된 간세포암 조직(H&E) 패치를 256×256 크기로 변환하여 ResNet 50 모델에 입력하였다. 초기 분류 신뢰도는 P_base = 0.4374로 매우 낮은 수준이었으며, 이는 섬유화로 인해 조직 구조·색채 특성이 왜곡되어 모델이 학습된 병리적 패턴과 불일치한 상태임을 의미한다.

[표 4-14] 병리 도메인 색상 채널 개입 인과 그래프 정량 분석 요약

구분	분석 내용	인과적 해석
기준 신뢰도(P_{base})	P_{base} = 0.4374	병리 조직의 색상·구조 왜곡으로 인해 모델이 정상적인 병리 패턴을 활용하지 못한 비최적 상태
A-채널 개입	급격한 상승 → 0.9561 도달 (ΔP = +0.5187)	H&E 조직의 핵/세포질 대비, 섬유화 패턴 등이 A 색차에 직접 반응하여 가장 강력한 인과 효과 발생
L-채널 개입	완만한 상승 후 plateau 형성	조직 밝기 조정은 세부 구조를 강조하지만, 병리 패턴 자체를 뒤집을 정도의 효과는 제한적
B-채널 개입	약한 상승 → 안정 구간	황색·청색 축의 변화는 H&E 기반 조직 특성에서 보조적 역할
H/S/V 개입(Hue, Saturation, Value)	거의 평탄한 중간선	병리 슬라이드의 색조 안정성이 높아 Hue/Saturation 변동 영향이 제한적
관찰된 공통 패턴	A≫L≈B ≫ H/S/V	A채널이 병리 조직 분류의 핵심 색상 인과 변수임을 명확히 보여줌
종합 결론	ΔP최대 +0.5187 → 신뢰도 0.9561	본 프레임워크는 자연 이미지뿐 아니라 병리 영상 도메인에서도 범용적이며, 단일 색상 채널 개입만으로도 예측 실패 원인을 해부하고 신뢰도를 크게 향상하게 시킬 수 있음을 입증

이 결과는 자연 이미지와 전혀 다른 분포를 갖는 의료 영상에서도 동일한 인과 구조가 유지됨을 보여주는 중요한 근거이다. 병리 조직의 색조·조직 구조 특성에 따라 특정 색상 채널(특히 LAB의 A채널)이 핵심 분류 인과 변수로 작동하며, 해당 채널에 대한 미세 개입만으로도 모델 신뢰도가 즉각적으로 의미 있게 향상됨을 확인하였다.



[그림 4-28] 병리 도메인 주파수 개입 인과 그래프 설명

[그림 4-28]은 TCGA-LIHC(간암) 병리 슬라이드(TCGA-G3-AAV1-01Z-00-DX1)에서 네 개의 웨이블릿 서브밴드에 독립적 개입을 수행하였을 때 ResNet 50의 신뢰도가 어떻게 변화하는지를 시각적으로 나타낸 것이다. 원본 신뢰도(P_{base} = 0.4374)는 매우 낮은 수준이며, 이는 병리 영상의 색조·조직 구조가 일반 ImageNet 기반 모델이 학습한 자연 이미지 분포와 크게 상이함을 의미한다. 주파수 개입 결과, LL-only 개입은 저주파 구조 정보를 보정함에 따라 w=0.8~1.2 구간에서 급격한 신뢰도 상승($\Delta P \approx +0.32$)을 보이며 가장 두드러진 인과적 효과를 나타냈다. 반면 LH/HL/HH 고주파 계열은 비교적 제한적이며, 특히 HH는 미세 텍스처 중심 신호로 구성되어 병리 영상의 조직학적 의미와 직접적 관계가 적어 신뢰도 변화 폭이 작게 유지되었다. 일부 고주파 대역에서는 w>1.2 이후 감소(overshooting) 현상도 관찰되었는데, 이는 지나친 에지·텍스처 증폭이 조직의 고유 색·형태를 왜곡하는 주파수의 개입으로 해석된다.

[표 4-15] 병리 도메인 주파수 인과 그래프 정량 요약

구분	분석 내용	인과적 해석
기준 신뢰도(P_{base})	P_{base} = 0.4374	병리 영상의 특성으로 인해 ImageNet 기반 모델의 초기 확신도 매우 낮음
LL-only 개입 상승 구간	w=0.3→1.0 동안 0.05→0.75까지 급상승	병리 조직의 형태·구조 복원에 LL이 결정적으로 작용하는 핵심 인과 변수
LH-only 변화 패턴	완만한 상승 후 0.6~0.7 plateau	수평 조직 경계는 일부 분류 근거가 되지만 영향력은 제한적
HL-only 변화 패턴	중간 구간에서 소폭 상승	수직 조직 경계는 보조적 역할에 머무름
HH-only 변화 패턴	초기 상승 후 하락(overshoot)	미세 텍스처는 병리 구조의 본질적 정보가 아니며, 과 개입 시 조직 왜곡 발생
과 개입(w 〉 1.2)	LL는 평탄 유지, LH/HL/HH는 감소	고주파 과 증폭 → 노이즈 인지 → 신뢰도 하락
종합 결론	LL ≫ LH ≈ HL 〉 HH	병리 영상의 예측 실패는 구조(LL) 붕괴가 핵심 원인이며, LL 개입만이 실질적 회복을 유도

본 연구는 자연 이미지와 병리 이미지(간암 H&E 슬라이드) 라는 서로 다른 두 도메인에 동일한 색상 개입(LAB/H/S/V)과 주파수 개입을 적용함으로써, 도메인 특성에 따라 인과 민감도(Causal Sensitivity)가 근본적으로 달라진다는 구조적 차이를 명확히 규명하였다. 자연 이미지에서는 H 채널 및 LL 대역과 같이 전체 조도·형태를 조절하는 개입이 가장 강력한 인과 효과를 보여, 모델의 시각 인식 과정이 주로 조명·구도·대비에 의해 좌우됨을 확인하였다. 반면, 병리(H&E) 조직에서는 완전히 상반된 패턴이 나타났다. 병리 영상에서 가장 큰 인과적 영향력을 가진 변수는 LAB A채널로, 이는 핵·세포질의 염색 반응과 직접 연결되는 색조 정보이다. 반대로 고주파 계열은 조직 슬라이싱 방향, 염색 variability, 미세 노이즈 등 병리 도메인 특유의 요인들로 인해 안정적인 구조 정보를 제공하지 못하며, 인과 기여도가 거의 존재하지 않았다. 이와 같은 차이는 단순한 성능 변동이 아니라, 주파수 기반 구조 정보는 병리 도메인의 조직학적 의미를 충분히 포착하지 못하며, 색조 기반 정보만이 직접적인 분류 단서를 제공한다는 사실을 실험적으로 증명한 것이다.

따라서 주파수 개입의 낮은 성능은 결함이 아니라 오히려, 본 연구가 제안한

인과 분석 프레임워크가 각 도메인에서 유효한 정보 양식(Modality)을 선택적으로 드러낼 수 있음을 입증하는 중요한 과학적 근거가 된다.

종합하면, 본 연구는 색상과 주파수 개입 실험을 통해 모델의 반응 구조가 도메인 특성에 따라 전혀 다른 양상으로 나타난다는 점을 실증적으로 확인하였다. 이는 제안된 Wavelet-Color 기반 인과 분석 프레임워크가 단일 도메인 특화 기법이 아닌, 도메인 전이(domain shift) 상황에서도 모델의 의사결정 근거를 구조적으로 식별할 수 있는 범용적 XAI 방법론임을 강하게 시사한다.

제 5 절 연구 성과

본 장에서는 색상·주파수 개입을 기반으로 수행된 일련의 인과 실험을 종합하여, 입력 변화에 따른 ResNet 모델의 신뢰도 민감도를 정량적으로 분석한 결과를 제시한다. 특히 본 연구에서 제안한 색상·주파수 개입 기반 인과 분석 프레임워크는 기존 XAI가 제공하던 관찰 중심 설명을 넘어, 입력 성분과 예측 신뢰도 사이의 인과적 관계를 실험적으로 규명하기 위한 분석 체계로 활용되었으며, 이를 통해 인과적 메커니즘을 직접적으로 시각화하고 해석할 수 있음을 확인하였다. 이하에서는 본 연구를 통해 도출된 주요 성과를 정리하여 제시한다.

1) 색상 개입 기반 인과 구조 규명

본 연구는 조명 편향(Exposure Bias)과 색채 정보 손실이 ResNet 모델의 신뢰도 하락을 유발하는 색상 인과 구조(Color Causal Structure)임을 실험적으로 입증하였다. Under-exposure 환경에서는 LAB L 채널 압축과 A/B 색차 붕괴가 결합한 이중 정보 손실(double information loss)이 발생하여, 서로 다른 장면이 동일한 저조도 무 특징 영역으로 인식되는 현상을 확인하였다. 본 프레임워크를 이용한 L/A/B 개입은 색상·윤곽·질감을 복원하여 최대 20-30%의 신뢰도 상승을 가져왔으며, 이는 색상 개입이 ResNet의 신뢰도 회복

에 직접적 인과 효과가 있는다는 사실을 보여준다.

2) 주파수 개입 기반 인과 구조 규명

주파수 개입 실험을 통해 CNN의 초기 합성곱 계층이 웨이블릿 서브밴드와 구조적으로 대응하며, 특히 저주파 대역이 모델 신뢰도 결정에 가장 중요한 인과 변인임을 규명하였다. 밝은·어두운 조명 조건 모두에서 LL 개입은 사라진 구조 정보를 회복시켜 ΔP가 크게 상승하는 효과를 보였으며, HL·LH는 중간 영향, HH는 미세 질감 특성상 영향이 제한적이었다. 이 결과는 주파수 선택성이 모델 인식 성능에 실질적 인과적 영향력을 행사함을 실험적으로 확인한 것이다.

3) LAB·HSV 6채널 통합 인과 민감도 분석

본 연구는 극단적 조명 조건이 아닌 일반 조명 이미지를 대상으로 LAB 및 HSV의 6개 채널에 동일한 개입을 수행한 결과, 각 채널의 인과 민감도가 상이한 비선형적 구조를 갖는다는 것을 밝혀냈다. H 또는 A채널은 개입 증가에 따라 최적 지점 도달 후 급락하는 Peak-Drop 패턴을 보였으며, L,S,V 채널은 안정적 상승 또는 U-shape 곡선을 나타냈다. 특히 B 채널은 색온도 변화에 매우 민감하게 반응하여 높은 변동성을 동반한 취약한 민감도 특성을 보였다. 이러한 결과는 모델 신뢰도 저하의 원인이 특정 단일 채널의 왜곡이 아니라 여러 색상 신호 간 조합의 불균형에 있음을 규명한 것으로, 색상 기반 도메인 편향의 인과 구조를 이해하는 데 중요한 통찰을 제공한다.

4) 병리 도메인 확장 실험을 통한 범용성 검증

본 연구는 TCGA-LIHC 병리 이미지(svs 파일)를 활용한 실험을 통해 제안한 프레임워크의 범용성을 검증하였다. 자연 이미지와 전혀 다른 색상 및 구조적 특성을 갖는 의료 영상에서도 본 프레임워크가 동일하게 인과 민감도 구조를 식별할 수 있음을 입증하였다. 그러나 구체적인 인과 패턴은 도메인에

따라 차이를 보였다. 자연 이미지에서는 LL 주파수 대역 및 Hue 채널이 중심적 역할을 하며 L, S, V 채널이 안정적 상승 또는 U-shape 곡선을 나타낸 반면, H&E 병리 영상에서는 LL 및 HH 주파수 개입 효과가 제한적이었고 LAB의 A 및 L 채널 개입이 더 강력한 인과 효과를 보였다. 이는 도메인 특성에 따라 색상 및 주파수의 인과 민감도가 차별적으로 변화함을 규명한 중요한 실험적 결과로, 제안 방법론이 도메인 특이적 인과 구조를 정량적으로 포착할 수 있음을 시사한다.

5) 실험 기반 인과 XAI 패러다임 수립

본 연구는 다음의 순환 구조를 갖는 새로운 실험 기반 인과 XAI 패러다임을 정립하였다. 진단(Diagnose) → 개입(Intervene) → 정량화(Quantify) → 시각화(Visualize)를 통해 기존 XAI가 수행하지 못하던 입력 조작 → 모델 반응 변화 → ΔP 기반 인과 그래프 구성이라는 정량적 인과 분석 체계를 구축하였다.

결론적으로 본 장의 연구 성과는 색상 인과 구조-주파수 인과 구조-6채널 통합 분석-도메인 확장 검증으로 이어지는 완전한 인과 분석 체계를 실험적으로 입증한 것이며, 이는 본 연구의 대화형 인과 분석 프레임워크의 실효성과 학술 가치를 강하게 뒷받침한다.

제 5 장 결론 및 향후 연구 방향

제 1 절 결론

1) 연구 요약

본 연구는 기존 설명 가능 인공지능이 가지는 관찰 중심 시각화의 한계를 극복하기 위해, 입력 색상과 주파수를 직접 조작하여 모델의 신뢰도 변화를 실험 기반 인과 분석 방식으로 규명하는 새로운 색상·주파수 개입 기반 인과 분석 프레임워크를 제안하였다. 제안한 프레임워크는 색상·주파수 개입에 따른 예측 확률 변화 ΔP와 Score-CAM 활성화 맵의 변화를 실시간으로 시각화한다. 이를 통해 기존 XAI로는 파악하기 어려웠던 입력 속성 → 특징 추출 → 신뢰도 변화로 이어지는 인과적 메커니즘을 구조적으로 규명할 수 있었다.

본 연구의 인과 실험을 통해 다음과 같은 핵심적 사실이 밝혀졌다.

첫째, 조명 편향은 모델 신뢰도를 크게 저하하며, 특히 LAB L 채널의 명도 정보 압축과 A/B 색차 붕괴가 결합된 이중 정보 손실이 Under-exposure 상황에서 신뢰도 하락의 근본 원인임을 규명하였다. 둘째, HSV 색상 공간에서는 V(Value) 과포화와 S(Saturation) 정보 소실이 Over-exposure 상황의 주된 실패 원인임이 밝혀졌다. 셋째, ResNet 50의 초기 계층은 Gabor-like 필터 구조를 보이며 웨이블릿 서브밴드와 직접적으로 대응하여 동작함을 확인하였다. 특히 LL 대역(저주파)은 구조 복원에 결정적으로 기여하는 인과 변수임이 실험적으로 검증되었다. 넷째, LAB·HSV의 6개 채널은 서로 상이한 비선형 인과 반응 패턴을 보여, 색상 공간의 구조적 차이가 모델의 민감도에 직접적으로 반영됨을 확인하였다. 다섯째, 의료 이미지(병리 슬라이드, SVS 포맷) 에서도 동일한 인과 구조가 관찰되었으며, 이는 본 프레임워크가 도메인 전반에 걸쳐 적용할 수 있는 범용 인과 분석 도구임을 실증하였다.

2) 본 연구의 주요 기여도

본 연구가 제안한 색상·주파수 개입 기반 인과 분석 프레임워크의 학술적·기술적 기여는 다음 세 가지로 요약된다.

첫째, 조명 편향으로 인한 색상 정보 손실이 ResNet 신뢰도 저하의 주요 인과 요인임을 실험으로 규명하였다. LAB 분석에서는 L 채널의 명도 압축과 A/B 채널의 색차 붕괴가 복합적으로 작용하여 이중 정보 손실 구조를 형성함을 확인하였고, HSV 분석에서는 V 채널의 과포화와 S 채널의 채도 소실이 과노출 인식 실패의 핵심 원인임을 정량적으로 입증하였다. 이러한 결과를 기반으로 입력 색상 변화가 모델 신뢰도에 미치는 영향을 구조화한 색상 인과 그래프를 제시하였다.

둘째, ResNet 50의 초기 합성곱 계층이 웨이블릿 서브밴드 구조와 기능적으로 대응함을 확인하였다. 대역별 개입 실험에서 LL 대역은 구조적 정보 복원을 통해 신뢰도를 가장 크게 향상하는 핵심 인과 변수로 나타났으며, LH·HL·HH 대역은 에지와 미세 질감 회복에 부분적으로 기여하되 개입 강도에 따라 비선형적 반응을 보였다. 이를 토대로 주파수 선택성과 인과 경로를 정량적으로 정리한 주파수 인과 그래프를 제시하였다.

셋째, 본 연구는 입력 조작과 모델 반응의 변화를 체계적으로 측정하기 위한 색상·주파수 개입 기반 인과 분석 프레임워크를 적용하여 다양한 실험을 수행하였다. 이를 통해 색상 개입과 주파수 개입이 신뢰도 변동과 Score-CAM 활성화 변화에 어떤 방식으로 기여하는지 정량적으로 확인하였다. 또한 의료 이미지를 포함한 다양한 도메인에서 유사한 인과 패턴이 재현됨을 확인함으로써, 제안 프레임워크가 모델 내부 표현과 물리적 입력 요인 간의 관계를 인과적으로 분해하는 분석 틀임을 실험적으로 입증하였다.

제 2 절 향후 연구 방향

본 연구는 새로운 프레임워크를 제안하고 인과 구조를 정량 규명하는데 의의가 있지만, 다음과 같은 제한점을 가진다. 모델 기반의 한계로는 특정 백본(ResNet 50)에 최적화되어 있으며, ViT·Swin 등 Transformer 기반 모델에

는 아직 확장 적용되지 못하였다. 두번째의 한계로는 개입의 주관성으로 현재 개입 강도 조절은 수동 트랙바 기반이며, 최적 개입량을 자동 산출하는 알고리즘은 포함되지 못했다. 세 번째의 한계는 모델 내부 표현의 해석에 대해 ΔP(확률 변화)와 Score-CAM 변화는 인과적 근거를 제공하지만, 내부 신경 표현의 근본적 변화를 완전히 설명하지는 못하는 한계를 가지고 있다. 마지막으로 데이터의 다양성의 한계를 가지고 있어 향 후 연구가 필요하다.

본 연구의 확장성과 실질적 영향력 강화를 위해 다음과 같은 후속 연구가 필요하다. 첫째, 자동화된 개입 최적화(Automated Causal Intervention Optimization) 연구가 필요하다. 현재는 개입 강도(색상·주파수)가 수동적으로 설정되지만, 향후에는 AutoML·Bayesian Optimization 기반 최적화 알고리즘을 통해 신뢰도 변화 ΔP을 극대화하는 최적 개입량 w^*과 색상·주파수 조합을 자동 산출하는 체계 구축이 요구된다. 둘째, 도메인 일반화(Domain Generalization) 및 OOD(Out-of-Distribution) 방어로의 확장이다. 본 연구에서 규명한 색상·주파수 인과 구조를 이용해 조명·색상·질감 왜곡에 강건한 모델을 설계할 수 있으며, 도메인 이동(Domain Shift)에 민감하게 반응하는 ΔP패턴을 활용하여 새로운 OOD 탐지 알고리즘 개발이 가능하다. 셋째, 최신 대규모 모델(Transformer/VLM) 확장성에 대한 검증이다. ResNet 기반의 실험을 넘어 Vision Transformer(ViT), CLIP, SAM 등 최신 Vision-Language 모델(VLM)에 본 프레임워크를 적용하여, 아키텍처별 색상·주파수 인과 민감도 차이를 비교·분석하는 연구가 필요하다. 넷째, 의료 영상 특화 인과 분석이 요구된다. 병리(H&E), CT, MRI 등 고해상도·고정밀 영상에서는 색상·조직 구조·질감의 인과적 중요도가 일반 자연 이미지와 크게 다르므로, 이 도메인에 특화된 인과 분석을 수행함으로써 임상적 활용 가능성을 높일 수 있다. 다섯째, Causal XAI 기반 입력 보정 프리프로세서(Input Correction Preprocessor) 개발이 요구된다. 본 연구에서 도출한 색상·주파수 인과 구조를 활용하여, 입력 이미지의 L/A/B, H/S/V, LL/LH/HL/HH를 실시간으로 조정하여 모델 신뢰도를 자동 복원하는 능동형 입력 교정 모듈(Input Correction Unit)을 구축할 수 있다. 이는 실제 운영환경에서 딥러닝 모델의 안정성과 견고성을 크게 향상하게 시킬 수 있다. 여섯째, 생성형 AI

모델(GAN, Diffusion, VLM)의 생성 과정에 대한 인과 분석 확장이 필요하다. 본 프레임워크는 생성형 모델에서도 색상·주파수 신호의 개입 효과를 분석하여, 생성 과정에서 발생하는 편향, 비정상 패턴, 위·변조 신호(Deepfake)를 인과적으로 규명하는 새로운 XAI 방식으로 활용될 수 있다. 이는 생성 모델의 품질 향상, 조작 탐지, 보안적 활용에 큰 기여가 가능하다.

참 고 문 헌

1. 국내문헌

김준석, 이재현, 백준걸. (2009). 웨이블릿 변환을 이용한 주기 신호의 실시간 이상 탐지에 대한 연구. 한국경영과학회 학술대회논문집, 2009, 1316－1323.

김현지, 조재영, 고성제. (2013). 적록 색각 이상자를 위한 HSV 색공간을 이용한 색변환 기법. 전자공학회논문지, 50(3), 91-101.

이용훈, 박진수, 김동현. (2013). 다중 해상도 데이터베이스를 위한 효율적인 칼라 영상 기술자: RGB→HSV 변환 및 웨이블릿 변환 적용. 한국컴퓨터정보학회 동계학술대회 논문집, 21(1), 345－348.

김영진, 김은경. (2018). CNN과 Grad-CAM 기반의 실시간 화재 감지. 한국정보통신학회논문지, 22(12), 1596-1603.

노대철, 김태영. (2019). Atrous Convolution과 Grad-CAM을 통한 손 끝 탐지. 컴퓨터그래픽스학회논문지, 25(5), 11-20.

김병현, 김건순, 진수민, 조수진. 영상기반 콘크리트 균열 탐지 딥러닝 모델의 유형별 성능 비교. 한국안전학회지, 6(158), 50-57.

김선진, 이종근, 곽내정, 류성필, 안재형. (2020). 주 객체 위치 검출을 위한 Grad-CAM 기반의 딥러닝 네트워크. 한국정보통신학회논문지, 24(2), 204－211.

안경희, 엄성용. (2020). 딥 러닝 기반 코로나19 흉부 X선 판독 기법. 문화기술의 융합, 6(4), 789-795.

박홍기, 배경호. (2020). ResNet 알고리즘을 이용한 가로수 객체의 폐색영역 검출 및 해결. 한국산학기술학회논문지, 21(10), 77-83.

박은수, 김승환, 굴람 무즈타바, 류은석. (2020). Guided Grad-CAM을 이용

한 영상 내 송전설비 검출기법. 한국방송·미디어공학회 학술발표대회 논문집.

량한, 서수영. (2021). SegNet과 ResNet을 조합한 딥러닝에 기반한 횡단보도 영역 검출. 한국측량학회지, 39(3), 141-148.

지영준, 채선규, 배석주. (2021). Grad-CAM(Gradient-Class Activation Map)을 이용한 복합 불량 패턴 분해. 학위논문, 한양대학교.

손성재, 진철군, 박아론, 외. (2021). ResNet-합성곱 오토인코더 기반 신경망을 이용한 스펙트럼 데이터 압축. 한국산학기술학회논문지, 22(12), 142-150.

고수연, 최영우. (2021). CAM과 Selective Search를 이용한 확장된 객체 지역화 학습 방법. 한국과학기술정보연구원 학술논문집, 10(9), 349-358.

여인창, 노명일, 김진혁, 김기수, 남정우, 이상현, 장영훈. (2021). Score-CAM을 이용한 선형 성능 우열 관계 예측 모델의 시각화. 한국CDE학회 동계학술대회 논문집, 2743.

신형섭. (2021). 어텐션 기법 및 의료 영상에의 적용 가능성: 설명 가능한 인공지능 기술의 최신 동향. 대한영상의학회지, 82(5), 871-889.

정자훈. (2022). Grad-CAM 기반 이미지 증대를 통한 분류 모델 성능 개선 연구. 학위논문.

정일옥, 최우빈, 김수철. (2022). 설명 가능한 인공지능(XAI)을 활용한 침입탐지 신뢰성 강화 방안. 융합보안 논문지, 3(101), 101-109.

백찬형, 권지훈, 정호엽. (2023). Grad-CAM을 사용한 얼굴인식 신경망. 스마트미디어저널, 12(2), 9-14.

정자훈, 김용기, 나성중, 류준열. (2023). 한정된 군사 데이터를 활용한 이미지 분류 AI의 성능 향상 방안: Grad-CAM을 활용한 준지도학습 적용. 한국산학기술학회논문지, 24(9), 408-417.

이상현, 김진영, 노종민, 이환필, 이수목, 윤일수. (2023). 차량 단말기 기반

돌발상황 검지 알고리즘 개발. 한국ITS학회 논문지, 22(4), 97-113.

안택현, 최정단. (2024). 웨이블릿 변환을 활용한 효율적인 의미론적 분할 기술. 한국ITS학회 논문지, 23(5), 248－260.

김성수, 윤상우, 이동헌 외 2명. (2024). ResNet50에서 ResNet18로의 지식 증류를 통한 정보 손실 감소와 점진적 학습 성능 향상. 대한전자공학회 학술대회 논문집.

김태호, 김강산, 방효충. (2024). Grad-CAM 기반의 설명가능한 인공지능을 사용한 합성 이미지 개선 방법. 한국군사과학기술학회지, 27(6), 665－676.

김나영, 윤예린, 최재완, 한유경. 딥러닝 기반 구름 및 구름 그림자 탐지를 통한 고해상도 위성영상 UDM 구축 가능성 분석. 대한원격탐사학회지, 40(4), 351-361.

권태윤, 노광현. (2025). 헤시안 기반 Grad-CAM++와 웨이블릿 융합 멀티스케일 활성화 맵으로 고정밀 이미지 복원. 한국정보기술학회논문지, 23(7), 39－48.재인용.

원형식, 조현종. (2025). 딥러닝 기반 고양이 피부질환 분류 시스템 개발 및 Grad-CAM 시각화. 전기학회논문지, 74(2), 339－344.

2. 국외문헌

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. Annual Review of Neuroscience, 24, 1193-1216.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), 248-255.

Bruna, J., & Mallat, S. (2013). Invariant scattering convolution networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 35(8), 1872-1886.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. Proceedings of the European Conference on Computer Vision (ECCV 2014), 818-833.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning Deep Features for Discriminative Localization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2921-2929.재인용.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), 770-778.재인용.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining (KDD 2016), 1135－1144.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618－626.재인용.

Guo, T., Seyed Mousavi, H., Vu, T. H., & Monga, V. (2017). Deep Wavelet Prediction for Image Super-Resolution. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 104-113.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature Visualization. Distill.

Chattopadhyay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks. Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 839－847.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations (ICLR 2019).

Tjoa, E., & Guan, C. (2020). A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. IEEE Transactions on

Neural Networks and Learning Systems, 32(11), 4793－4813.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., and Hu, X. (2020). Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 24－25.재인용.

Li, Q., Shen, L., Guo, S., & Lai, Z. (2020). Wavelet Integrated CNNs for Noise-Robust Image Classification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020), 10047-10056.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges. Information Fusion, 58, 82－115.

Desai, S., & Ramaswamy, H. G. (2020). Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV).

Prabhu, A., Alizadeh Vahid, K., Farhadi, A., & Rastegari, M. (2020). Butterfly Transform: An efficient FFT-based neural architecture design. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4826-4835.

Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., & Li, B. (2020). Axiom-based Grad-CAM: Towards accurate visualization and explanation of CNNs. Pattern Recognition, 122, 108245.

Chi, L., Jiang, B., & Mu, Y. (2020). Fast Fourier Convolution. In

Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Vancouver, Canada.

Schwalbe, G., & Finzel, B. (2021). A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts. Frontiers in Artificial Intelligence, 4, 593438.

Jiang, P. T., Zhang, C. B., Hou, Q., Cheng, M. M., & Wei, Y. (2021). LayerCAM: Exploring Hierarchical Class Activation Maps for Localization. IEEE Transactions on Image Processing, 30, 5875-5888.

Sato, N., Kinoshita, H., Kakimoto, T., et al. (2021). Evaluation of kidney histological images using deep learning-based methods. Kidney International Reports.

Jung, H., Oh, Y., Jeong, J., & Kim, S. (2021). Towards better explanations of class activation mapping. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 1354-1362.

Chen, Q., Dong, X., Tu, G., Wang, D., Zhao, B., & Peng, Z. (2022). TFN: An Interpretable Neural Network with Time-Frequency Transform Embedded for Intelligent Fault Diagnosis.

Zhao, X., Huang, P., & Shu, X. (2022). Wavelet-Attention CNN for Image Classification. Multimedia Systems, 28 (3), 915-924.

Michau, G., Frusque, G., & Fink, O. (2022). Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series. Proceedings of the National Academy of Sciences, 119(8).

Soomro, S., Kim, K., & Nam, S. (2024). Grad++ScoreCAM: Enhancing

visual explanations of deep convolutional networks using incremented gradient and score-weighted methods.

Kubach, J., Till, T., Libera, K., et al. (2025). Development of a deep learning algorithm for radiographic detection of syndesmotic instability in ankle fractures with intraoperative validation. Scientific Reports.

ABSTRACT

Causal Analysis of ResNet Model Confidence Under Color and Frequency Feature Interventions

Kwon, Tae-Youn

Major in Smart Convergence Consulting

Dept. of Smart Convergence Consulting

The Graduate School

Hansung University

Deep learning-based image classification models, particularly convolutional neural networks in the ResNet family, have achieved strong performance across numerous domains. However, their internal decision-making processes remain opaque, and it is often unclear which visual cues drive the selection of specific classes. Real-world environmental variations—such as illumination shifts that distort brightness distributions, color changes that disrupt hue and saturation structure, and texture loss that obscures fine-grained patterns—destabilize intermediate feature representations. These atypical inputs degrade the model's internal consistency and cause sharp drops in prediction confidence. Such vulnerability originates from the structural limitation that models trained predominantly on large, curated datasets fail to fully reflect physical variations encountered in real environments, including changes in lighting

intensity, chromatic properties, and contour contrast. Existing explainable AI (XAI) methods remain largely passive, providing only post hoc visualizations of "where the model looked" without revealing how physical perturbations causally influence the model's reasoning and outputs.

To overcome these limitations, this study proposes a Color–Frequency Intervention-based Causal Analysis Framework, which performs active interventions on LAB/HSV color channels and wavelet subbands to causally evaluate how manipulated inputs affect prediction probability changes (ΔP), Score-CAM structures, and frequency responses. This framework advances XAI beyond passive observation by experimentally identifying the causal pathways linking input perturbations, internal responses, and final predictions.

In the color-intervention experiments, low-light conditions produced a double-compression effect in which the L channel collapsed into a 0–20 range and A/B chromatic differences converged near zero. This forced diverse scenes into feature-poor, dark regions and significantly reduced the baseline confidence (P_base). By expanding the L channel and restoring A/B chromaticity via targeted interventions, structural and chromatic cues recovered, yielding a 15–30% increase in prediction confidence. Even in bright lighting conditions, oversaturated HSV-V and diminished HSV-S caused structural information loss; however, luminance and saturation interventions restored features closer to those seen during ImageNet training, mitigating confidence degradation.

Frequency-intervention experiments demonstrated that early convolutional layers of ResNet50 structurally correspond to wavelet subbands: LH and HL captured contour and boundary information, while HH reflected fine-grained textures. Amplifying specific subbands reactivated relevant structural or textural features, increasing prediction confidence by

approximately 0.10 – 0.25.

Using these intervention-driven confidence changes, the study constructs color causal graphs and frequency causal graphs, quantitatively explaining how input manipulations shape model reliability, activation structure, and frequency selectivity. This approach moves beyond visualization toward an active causal debugging tool capable of diagnosing, intervening in, quantifying, and correcting model biases. Future extensions include automated and optimized interventions for image correction, out-of-distribution (OOD) robustness, and data-centric model improvement. The framework is applicable to medical imaging, forensics, anomaly detection, deepfake analysis, and industrial inspection, contributing to bias mitigation and reliability enhancement. Furthermore, it has potential for adaptation to generative AI systems, enabling causal XAI analyses of bias, abnormal patterns, and manipulability within generative processes.

【Key word】 Causal Analysis Based on Color- and Frequency-Domain Interventions, Explainable Artificial Intelligence XAI, Color-Space Intervention, Frequency-Domain Intervention, Color-Based Causal Graph, Frequency-Based Causal Graph