Journal of KIIT. Vol. 23, No. 9, pp. 201-210, Sep. 30, 2025. pISSN 1598-8619, eISSN 2093-7571 
http://dx.doi.org/10.14801/jkiit.2025.23.9.201

# 멀티모달 비전-언어 표현 학습 기반의 해충 이미지 분류

김도현*, 손창환**

# Pest Image Classification Via Multimodal Vision-Language Representation Learning

Do-Hyun Kim*, Chang-Hwan Son**

본 성과물은 2025년도 정부(과학기술정보통신부)의 재원으로 연구개발특구진흥재단-(군산) 2025년 강소특구 지역 특성화 육성 사업의 지원을 받아 수행된 연구임(RS-2025-02312252)

## 요 약

해충 이미지 분류는 해충 간 텍스처와 색상이 유사하기 때문에 높은 수준의 분류 정밀도가 요구된다. 본 연구에서는 기존 유니모달 모델이 갖는 한계를 극복하기 위해, 비전-언어 모델의 사전 학습된 멀티모달 정보를 활용하는 새로운 해충 이미지 분류 모델을 제안한다. 구체적으로, CLIP 모델을 해충 이미지 데이터에 적용하여 텍스트 인코더와 이미지 인코더를 정렬한 뒤, 각 해충 클래스에 대한 텍스트 사전 정보를 추출하여 기존 이미지 분류 모델에 통합하였다. 이를 위해, 입력 해충 종류에 적응적으로 텍스트 정보를 재구성하고, 시각 특징과 결합할 수 있도록 멀티모달 특징 퓨전 모듈을 설계하였다. 실험 결과, 제안한 모듈은 시각 특징 구별력 상향에 효과적임을 검증하였으며, 기존 CNN, ViT 및 하이브리드 기반의 대표 이미지 분류 모델보다 더 우수한 인식 성능을 달성하였다.

## Abstract

Pest image classification requires a high level of precision due to the strong similarity in texture and color among different species. To address the limitations of existing unimodal models, this study proposes a novel pest image classification model that leverages pretrained multimodal information from vision-language models. Specifically, the CLIP model is applied to pest image data to align the text and image encoders, after which class-specific textual prior information is extracted and integrated into conventional image classification models. To achieve this, we designed a multimodal feature fusion module that reconstructs adaptive textual information based on the input pest type and combines it with visual features. Experimental results demonstrate that the proposed module effectively enhances the discriminative power of visual features and achieves superior recognition performance compared to representative image classification models based on CNNs, ViTs, and hybrid architectures.

Keywords
pest image classification, multimodal learning, prompt learning, vision-language model

* 국립군산대학교 소프트웨어학과 학사과정
- ORCID: https://orcid.org/0009-0006-3037-7150
** 국립군산대학교 소프트웨어학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0001-7077-3074

· Received: Aug. 25, 2025, Revised: Sep. 05, 2025, Accepted: Sep. 08, 2025
· Corresponding Author: Chang-Hwan Son
Department of Software Science and Engineering, Kunsan National University, Republic of Korea
Tel.: +82-63-469-8915, Email: cson@kunsan.ac.kr

## Ⅰ. 서 론

해충은 농업 생산성과 작물 품질 저하의 주요 원인으로 인식되고 있다. 이에 따라 해충을 정확히 식별할 수 있는 자동화 기술의 필요성이 제기되고 있으며, 특히 인공지능 기술의 발전과 맞물려 그 중요성이 더욱 부각되고 있다. 해충 분류 결과는 병해충 방제 전략 수립과 데이터 기반 농업 의사결정에 핵심적인 정보를 제공한다. 최근 이미지 기반 딥러닝 해충 분류 기술이 발전하고 있으나, 해충 간 유사한 외형, 다양한 자세 변화, 조명 조건의 변화, 원거리 촬영에 따른 해상도 저하 등으로 인해 여전히 해결이 요구되는 기술적 과제들이 존재한다.

해충 이미지 분류를 위한 딥러닝 모델은 크게 합성곱 신경망(CNN, Convolutional Neural Network), 비전 트랜스포머(ViT, Vision Transformer), 그리고 두 구조를 결합한 하이브리드 모델로 분류할 수 있다. CNN은 학습 가능한 필터를 통해 해충의 국부적 시각 특징을 자동으로 추출할 수 있다는 강점을 갖는다. 반면, ViT는 입력 이미지의 전역 정보를 기반으로 특징을 추출하여, 보다 넓은 문맥 이해에 효과적이다. 하이브리드 모델은 두 구조의 장점을 결합함으로써, 국부 및 전역 특징을 동시에 추출하고 융합할 수 있다는 이점을 제공한다.

최근 딥러닝 분야에서는 비전-언어 모델(Vision-Language models)이 큰 주목을 받고 있다. 비전-언어 모델은 이미지와 텍스트 데이터를 각각의 인코더를 통해 공유 임베딩 공간으로 변환한 뒤, 두 모달리티 간 특징을 정렬하는 방식으로 학습된다. 대표적인 모델로는 CLIP(Contrastive Language-Image Pretraining)[1]과 ALIGN(A Large-scale ImaGe and Noisy-text embedding)[2] 등이 있다. 사전 학습된 두 종류의 이미지 및 텍스트 인코더는 이후 이미지 분류, 객체 검출 등과 같은 하위 작업(Downstream tasks)에 효과적으로 활용될 수 있다.

본 연구에서는 이러한 추세에 발맞추어, CLIP 비전-언어 모델을 활용한 해충 이미지 분류 모델을 새롭게 제안한다. 먼저, CLIP 모델을 해충 이미지-텍스트 데이터 쌍으로 미세 조정(Fine-tuning)하여, 이미지와 텍스트 간의 의미적 정합을 학습한다. 이후, 사전 학습된 텍스트 정보를 CNN의 대표 모델인 ResNet[3]에 통합하여 시각적 특징의 구별력을 강화하고자 한다. 이를 위해, 본 논문에서는 사전 학습된 텍스트 정보를 ResNet에 효과적으로 결합할 수 있는 '멀티모달 특징 퓨전 모듈'(MFFM, Multimodal Feature Fusion Module)을 새롭게 설계하였다. MFFM은 ResNet에서 추출된 시각적 특징과 텍스트 사전 정보를 융합하는 역할을 수행하며 그 적용 위치는 애블레이션 스터디(Ablation study)를 통해 결정되었다. 실험 결과, 제안한 MFFM을 결합한 ResNet 모델은 기존보다 향상된 분류 정확도를 달성하였으며, 이는 사전 학습된 텍스트 정보가 해충 이미지 분류 성능 개선에 효과적임을 시사한다. 또한, 제안한 MFFM 기반 ResNet은 기존의 일반 이미지 분류 모델 대비 우수한 성능을 달성하였다.

## Ⅱ. 관련 연구

해충 이미지 분류 모델은 스케일(Scale)[4] 문제, 관심 영역(ROI, Regions of Interests)[5] 추출 및 긴꼬리 분포(LTD, Long-tailed distribution)[6] 문제를 해결하기 위해, 기존의 CNN 및 ViT 계열의 대표 모델을 변형하여 개발되었다.

첫째, 스케일 문제는 이미지 분류와 객체 검출을 포함한 컴퓨터 비전 분야에서 오랜 난제로 인식되고 있다. 동일한 객체라도 촬영 거리나 각도에 따라 크기와 형태가 변형되어 물체 인식에 어려움을 초래할 수 있기 때문이다. 이를 해결하기 위해 다중 스케일 표현을 딥러닝 프레임워크에 통합하려는 다양한 시도가 이루어졌다. 대표적인 모델로는 FPN(Feature Pyramid Network)[7]과 UNet[8] 등이 있다. 또 다른 연구 방향으로는 표준 합성곱 연산을 변형하는 방법들이 있는데, 예를 들어 서로 다른 크기의 필터를 갖는 이중 분기 구조[9]를 적용해 다양한 수용 영역(Receptive field)을 확보하거나, 변형 합성곱(Deformable convolution)[10] 및 확장 합성곱(Dilated convolution)[11]과 같이 특징 추출 영역을 확장하거나 변경한 합성곱 기법들도 있다.

둘째, 긴꼬리 분포 문제란 수집된 데이터셋에서 클래스별 데이터 개수가 크게 불균형한 현상을 의

미한다. 실제로 다양한 해충의 발생 빈도는 지역, 시기, 기상 조건 등에 따라 다르기 때문에, 해충 데이터셋에서는 특정 클래스에 데이터가 집중되는 긴 꼬리 분포 현상이 자주 발생한다.

이러한 긴꼬리 분포 문제를 해결하기 위해, 데이터 균형을 맞추기 위한 데이터 증강(Data augmentation) 기법을 적용하거나, 쉽게 분류되는 샘플의 가중치를 줄이고 어려운 샘플에 더 큰 가중치를 부여하는 Focal[12] 손실과 같은 손실 함수 변형을 사용할 수 있다. 최근에는 ViT와 CNN을 결합한 하이브리드 모델[13]이 전역 및 국부 특징 학습을 동시에 활용하여 긴꼬리 분포 문제를 해결하는 연구들이 보고되고 있다. 그러나 단순하면서도 효과적인 데이터 증강과 손실 함수 변형 기법이 여전히 긴꼬리 분포 문제 해결의 주류를 이루고 있다.

셋째, 관심 영역 추출은 클래스 간 특징 구별력을 향상시킬 수 있는 영역을 검출하여 이미지 분류 성능을 개선하는 접근 방법이다. 예를 들어, 잎사귀 질병의 경우 질병이 발생한 잎사귀와 해당 질병 부위를 배경과 구분하고, 해충 분류의 경우에는 해충 간 식별력을 높이기 위해 해충 영역만을 검출할 수 있다. 관심 영역 기반 모델은 크게 두 가지 방식으로 나눌 수 있다. 첫째는 슈퍼픽셀(Super-pixels)[14], 클래스 활성화 맵[15], 영역 분할(Segmentation)[16] 등을 통해 관심 영역을 명시적으로 검출하고, 이를 CNN 모델에 통합하는 방식이다. 둘째는, 공간 및 채널 어텐션 모듈[17]을 활용하여 모델이 중요한 영역에 스스로 주의를 집중하도록 학습하는 방식이다. 그러나 최근에는 이러한 관심 영역 추출과 어텐션 기능이 비전 트랜스포머(ViT) 구조에 통합되는 추세다. 대표적인 ViT 기반 모델로는 MViT[18], Cross-ViT[19], ROI-ViT 등이 있으며, 이들 모델은 관심 영역 정보를 효과적으로 반영하면서도 전역적인 문맥 정보를 함께 고려할 수 있는 장점을 제공한다.

## III. 제안한 비전-언어 정렬 기반 해충 이미지 분류

본 연구는 해충 이미지 분류의 정확도를 향상시키기 위해, 이미지의 시각적 특징뿐만 아니라 사전 학습된 텍스트 특징까지 함께 활용할 수 있는 멀티모달 비전-언어 통합 모듈 설계 방식을 새롭게 제안한다. 기존의 유니모달 기반 접근과는 달리, 본 연구에서는 CLIP 모델을 미세 조정하여 생성한 텍스트 프롬프트 정보를 ResNet 기반 이미지 분류 모델에 효과적으로 통합할 수 있는 방안을 제시한다.

### 3.1 비전-언어 정렬 모델의 사전 학습 과정

그림 1은 해충 분류를 위해 비전-언어 모델인 CLIP을 미세 조정하는 과정을 보여준다. CLIP은 크게 텍스트 인코더와 이미지 인코더로 구성되며, 각각 입력된 텍스트 문장과 해충 이미지를 공유 임베딩 공간으로 변환하여 텍스트 특징과 시각 특징을 생성한다. 이때 이미지 인코더의 파라미터는 고정된(Frozen) 상태이고 텍스트 인코더의 파라미터만 훈련된다. 텍스트 인코더의 입력 문장은 'a photo of a [class]'와 같은 템플릿 형태로 구성되며, [class]에는 입력 이미지에 해당하는 해충의 종류가 삽입된다. 예를 들어, "rice leaf roller", "rice leaf caterpillar" 등의 해충명이 사용된다. 본 연구에서는 이미지 인코더로 CNN 계열 모델인 ResNet-50을 사용하였고, 텍스트 인코더로는 트랜스포머 기반의 언어 모델을 사용하였다. 텍스트 인코더는 프롬프트 형태의 문장을 입력받아 토큰화한 뒤, 트랜스포머 인코더를 거쳐 512차원의 벡터로 변환한다. 이미지 인코더는 해충 이미지를 입력받아, ResNet-50의 마지막 단을 어텐션 풀링(Attention pooling) 구조로 개조하여 최종적으로 512차원의 시각 특징을 생성한다. 이렇게 변환된 텍스트 특징과 시각 특징은 각각 공유 임베딩 공간으로 변환된 후, 대조 학습(Contrastive learning)을 통해서 정렬된다. 대조 손실은 다음과 같이 정의된다.

$$L_{i\to t} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp((I_i\cdot T_i)/\tau)}{\sum_{j=1}^{N}\exp((I_i\cdot T_j)/\tau)} \quad (1)$$

$$L_{t\to i} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp((T_i\cdot I_i)/\tau)}{\sum_{j=1}^{N}\exp((T_i\cdot I_j)/\tau)} \quad (2)$$

$$L = \frac{1}{2}(L_{i\to t} + L_{t\to i}) \quad (3)$$

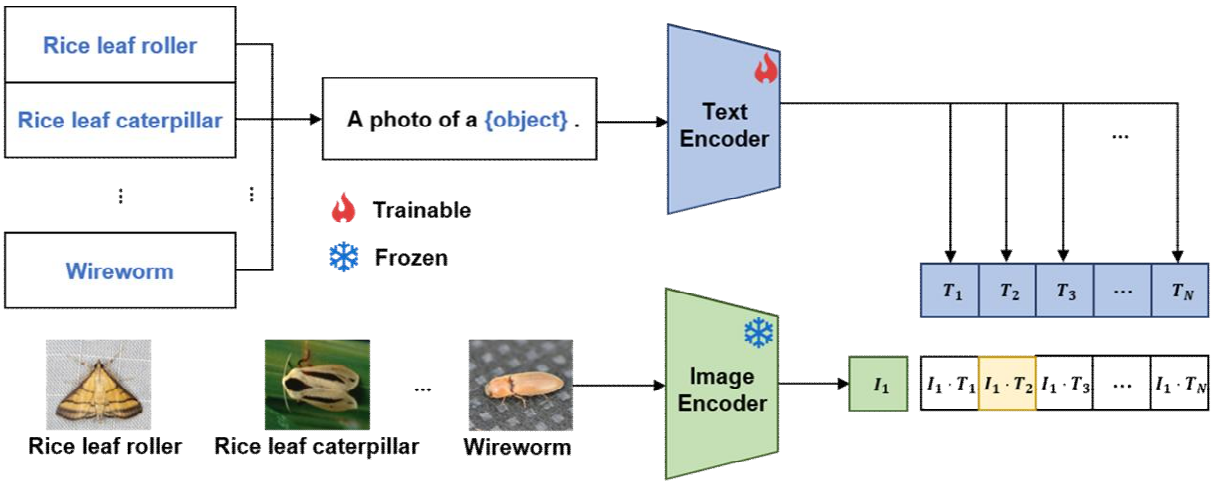


그림 1. 해충 분류를 위한 CLIP 모델 사전 학습 과정
Fig. 1. Pre-training procedure of the CLIP model for pest classification

식 (1)과 (2)는 각각 이미지에서 텍스트로, 텍스트에서 이미지로 정렬하기 위한 대조 손실을 의미한다. 식에서 $N$은 배치 크기를 의미하며, $T_j$와 $I_i$는 배치(Batch) 내에서 부정 쌍(Negative pair)을, $T_i$와 $I_i$는 긍정 쌍(Positive sample)을 의미한다. 따라서 식 (1)과 (2)는 긍정 쌍의 유사도를 총 부정 쌍의 유사도로 정규화한 비율을 의미한다. 그리고 식 (3)의 최종 손실 $L$은 두 손실의 평균으로 정의되며 $\tau$는 학습 가능한 온도 스케일링 파라미터로 내적 유사도의 분포를 조정하여 학습 안정성을 향상시킨다.

## 3.2 제안한 멀티모달 특징 퓨전 모듈

사전 학습된 CLIP 모델은 해충 분류 작업에 유용한 텍스트 사전 정보를 제공한다. 텍스트 인코더에서 추출된 텍스트 특징은 시각 특징과 정렬되어 있어 시각 특징을 강화하는 데 효과적으로 활용될 수 있다. 이에 본 연구에서는 사전 학습된 텍스트 특징을 해충 분류 모델에 통합하기 위한 MFFM을 제안한다.

그림 2의 상단은 해충 분류를 위한 개선된 ResNet-50 모델의 구조를 나타내고, 하단은 ResNet-50 내부에서 추출된 시각 특징과 사전 학습된 텍스트 정보를 결합하기 위한 MFFM의 세부 구조를 보여준다. 그림 상단에서 기존 ResNet-50 모델과 달리, 제안한 구조에서는 MFFM이 새롭게 추가된 것을 확인할 수 있다. 이 모듈은 ResNet-50 내부에서 추출된 시각 특징을 강화하기 위해, 텍스트 인코더로부터 추출된 텍스트 사전 정보를 활용한다. 특히 클래스별로 추출된 텍스트 특징의 상대적 중요도를 학습하여, 이를 하나의 통합된 텍스트 특징으로 생성한다. 이후 통합된 텍스트 특징을 입력 시각 특징과 결합함으로써 특징 표현의 구별력을 향상시킨다.

제안한 MFFM은 그림 2의 하단과 같이, '텍스트 프롬프트 어텐션 블록'(Text Prompt Attention Block, TPAB)과 '텍스트-이미지 특징 퓨전 블록'(TIFFB, Text-Image Feature Fusion Block)으로 구성된다. 먼저, TPAB는 식 (4)와 같이 ResNet-50 내부 특징인 $F$에 대해 전역 평균 풀링(GAP, Global Average Pooling) 계층, 합성곱 계층 및 소프트맥스 계층을 적용하여 최종 가중치 $w$를 계산한다.

$$w = Softmax(Conv_{1\times1}(GAP(F))) \tag{4}$$

여기서 $GAP$는 전역 평균 풀링 계층이고 $Conv_{1\times1}$과 $Softmax$는 각각 $1\times1$ 합성곱 계층 및 소프트맥스 계층을 나타낸다.

가중치 $w$가 주어진 경우, TPAB는 사전 학습된 텍스트 특징 $T$와 선형 결합(Linear combination)하여 입력 해충 이미지에 가장 최적화된 하나의 텍스트 특징을 뽑아낸다. 이 텍스트 특징은 시각 특징과 사전 정렬되어 있기 때문에 입력 해충 이미지의 시각 특징을 대표한다고 볼 수 있다.

$$P = Conv_{3\times3}(w \odot T) \quad (5)$$

여기서 $\odot$는 선형 결합을 나타내고 $Conv_{3\times3}$는 $3\times3$ 합성곱 계층을 의미한다. 식 (5)는 결국 입력 해충 이미지의 시각 특징에 적응적인 텍스트 프롬프트인 $P$가 생성됨을 알려준다.

TIFFB는 ResNet-50 내부 특징인 $F$와 식 (5)를 통해 생성된 텍스트 프롬프트 $P$를 융합하여 ResNet-50 모델의 특징 추출 능력을 강화한다.

$$F_c = Concat(F, P) \quad (6)$$

여기서 $Concat$은 채널 축으로 입력 특징 맵을 쌓는 연산을 의미한다. 그리고 $F_c$는 텍스트 및 이미지 특징을 결합한 멀티모달 특징을 나타낸다.

멀티모달 특징 $F_c$는 트랜스포머 블록을 통해 전역 문맥 고려하여 특징을 강화한 후, 두 단계의 합성곱 계층을 거쳐 다음 ResNet-50 계층의 크기와 동일한 출력 특징 $\hat{F}$를 산출한다.

$$\hat{F} = Conv_{3\times3}(Conv_{1\times1}(Transformer(F_c))) \quad (7)$$

여기서 $Transformer$는 전역적 상호작용을 학습하기 위한 어텐션 모듈을 포함하고 있으며, $Conv_{1\times1}$과 $Conv_{3\times3}$는 각각 채널 축소 및 지역 특징 보정을 위한 합성곱 계층을 의미한다. 따라서, 식 (7)는 텍스트 사전 정보를 포함한 멀티모달 특징을 입력으로 받아서 최종 ResNet-50 모델의 내부 특징 구별력을 강화하는 과정으로 해석할 수 있다.

### 3.3 MFFM 적용 위치에 따른 분류 모델 구조

제안한 MFFM을 기존 분류 모델인 ResNet-50에 통합하기 위해서는, 먼저 MFFM을 적용할 계층 위치를 선정하는 과정이 필요하다. 그림 3은 MFFM을 서로 다른 계층에 적용한 ResNet-50 구조를 보여준다. ResNet-50은 그림 2에서 확인할 수 있듯 총 4개의 계층으로 구성되어 있으며, 문헌에 따라서는 이를 4개의 합성곱 블록으로 설명하기도 한다.

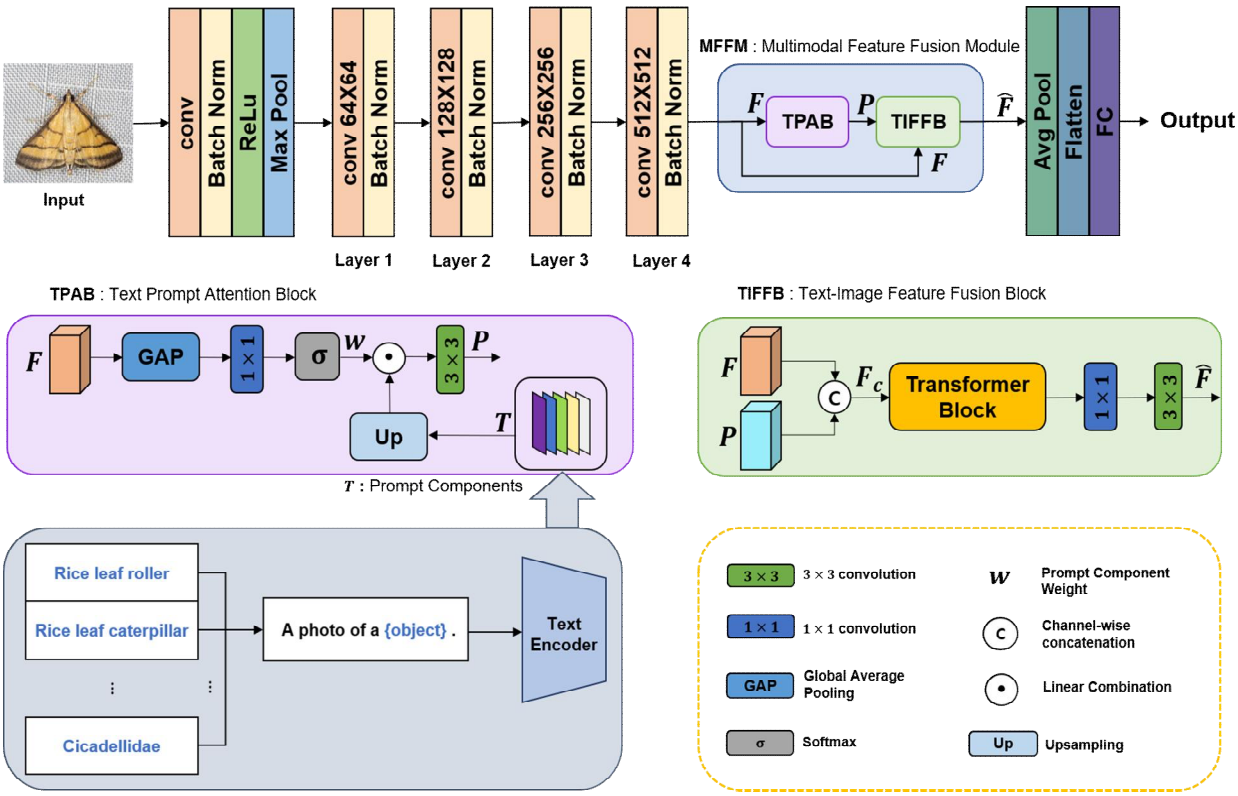


그림 2. 제안한 MFFM을 포함한 해충 이미지 분류 모델 구조

Fig. 2. Architecture of the proposed pest image classification model with MFFM

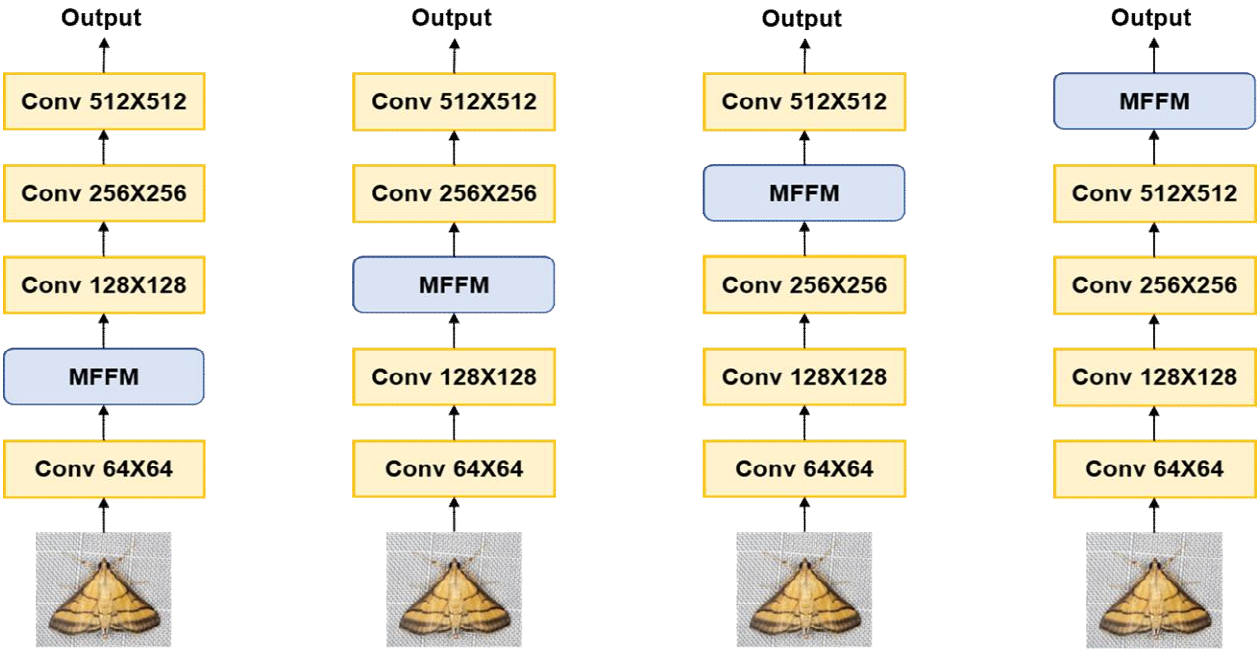


그림 3. MFFM 결합 위치; (a) 첫 번째 계층 후, (b) 두 번째 계층 후, (c) 세 번째 계층 후, (d) 네 번째 계층 후
Fig. 3. Fusion points of MFFM: (a) after the first layer, (b) after the second layer, (c) after the third layer, and (d) after the fourth layer

본 연구에서는 4개 계층 각각 뒤에 MFFM을 적용하여 해충 이미지 분류 정확도를 비교하고, 이를 바탕으로 최종 MFFM 적용 위치를 결정하고자 한다.

## IV. 실험 및 결과

### 4.1 실험 환경

본 연구에서는 해충 이미지 분류 성능 평가를 위해 IP102 공개 데이터셋[20]을 사용하였다. IP102는 총 102개의 해충 클래스로 구성되어 있으며, 실제 농업 환경에서 촬영된 다양한 해상도, 배경, 객체 크기를 포함한다. 그러나 본 연구와 관련 없는 애벌레 이미지와 워터마크가 포함된 합성 이미지도 일부 존재한다. 따라서 본 연구에서는 해충 외의 불필요한 이미지를 제거하여 총 86개 클래스로 정제하였으며, 최종 실험 데이터셋으로 35,296장의 해충 이미지를 사용하였다.

그림 4는 실험에 사용된 해충 이미지 예시를 보여준다. 그림에서 알 수 있듯, 클래스 간 텍스처 및 색상과 같은 외관적 유사도가 높아, 해충 이미지 분류가 도전적임을 시사한다.

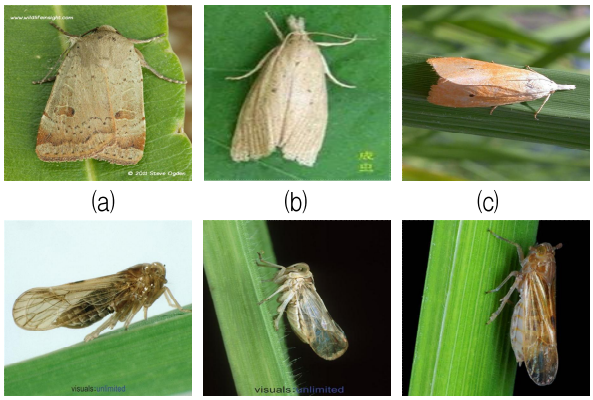


그림 4. 해충 이미지 데이터;
(a) 노랑밤나방, (b) 이화명나방, (c) 노랑줄기나방,
(d) 갈색벼멸구, (e) 흰등멸구, (f) 벼매미충
Fig. 4. Pest image samples;
(a) Yellow cutworm, (b) Asiatic rice borer,
(c) Yellow rice borer, (d) Brown plant hopper,
(e) White backed plant hopper, (f) Rice leaf hopper

학습을 위해, 훈련과 테스트 데이터의 비율은 7:3으로 설정하였고, 입력 이미지의 크기는 $224 \times 224$로 통일하였다. 학습률은 0.0001, 배치 크기는 10으로 설정하였다. 최적화 기법으로는 확률적 경사 하강법을 사용하였고 총 에폭 수는 80이다. 모든 실험은 NVIDIA GeForce RTX 3080 GPU, Python 3.8.13, CUDA 11.8, PyTorch 2.1.1 환경에서 수행되었다.

### 4.2 MFFM 적용 위치에 따른 정확도 분석

먼저, MFFM의 적용 위치에 따른 모델 성능을 평가하였다. 그림 3에서 제시된 총 4개의 적용 위치에 따라 ResNet-50 모델을 학습하고 테스트하였다. 표 1은 MFFM 적용 위치별 모델 정확도를 보여준다. 표에서 알 수 있듯, 네 번째 계층 뒤에 MFFM을 적용하는 것이 가장 우수한 성능을 나타냈다. 이는 텍스트 사전 정보가 고수준 시각 특징과 정렬되어 있기 때문에, ResNet-50에서도 고수준 시각 특징을 추출하는 네 번째 계층에 MFFM을 적용하는 것이 가장 적합함을 시사한다.

표 1. MFFM 적용 위치에 따른 모델 정확도 비교
Table 1. Comparison of model accuracy according to MFFM application positions

| Models | Accuracy |
|---|---|
| ResNet-50[3] | 79.60% |
| Proposed model(layer 1) | 78.44% |
| Proposed model(layer 2) | 75.37% |
| Proposed model(layer 3) | 79.78% |
| **Proposed model(layer 4)** | **80.23%** |

### 4.3 해충 이미지 모델 정확도 비교

표 2는 제안한 모델과 기존 이미지 분류 모델들의 정확도 평가 결과를 보여준다. 비교 대상 모델에는 CNN 계열, ViT 계열, 그리고 하이브리드 계열에 속한 DeiT[21], CrossViT[19], PVT[22], MobileNetv3[23], EfficientNet[24], ResNet-50[3], CoatNet[13]이 포함되었다. 비교 실험을 위해, 동일한 데이터셋과 학습 조건이 모든 모델에 적용되었다. 표 2에서 보듯이, 제안한 모델은 80.23%이라는 가장 높은 정확도를 획득하였다. 특히, 트랜스포머 계열 모델인 CrossViT 대비 약 4.13%, CNN 계열 모델인 MobileNetv3 대비 약 9.23%, 하이브리드 계열인 CoatNet 대비 약 26.67% 향상된 성능을 보였다. 이는 멀티모달 기반의 MFFM 모듈이 기존 유니모달이 갖는 정보의 한계를 보완하고, 세밀한 해충 이미지 분류 작업의 성능 개선에 아주 효과적임을 말해준다. MFFM 적용으로 인해 ResNet-50 대비 이미지당 처리 시간이 약 0.4ms 증가하여 연산 속도가 다소 저하되었으나, 이러한 차이는 실시간 대응에 큰 영향을 미칠 정도는 아니다. 반면, 해충 분류 정확도는 뚜렷한 향상을 보였기 때문에, MFFM의 추가는 모델 성능 측면에서 충분한 의의가 있다고 판단된다.

표 2. 해충 이미지 분류 정확도 비교
Table 2. Comparison of classification accuracy for pest images

| Models | Accuracy |
|---|---|
| DeiT[21] | 77.59% |
| CrossViT[19] | 76.10% |
| PVT[22] | 76.24% |
| MobileNet(v3)[23] | 71.00% |
| EfficientNet[24] | 79.01% |
| ResNet-50[3] | 79.60% |
| CoatNet[13] | 53.56% |
| **Proposed model** | **80.23%** |

## V. 결 론

본 연구에서는 해충 간 미세한 시각적 차이를 효과적으로 구분하기 위해, 시각 정보와 텍스트 사전 정보를 결합할 수 있는 멀티모달 특징 퓨전 모듈을 새롭게 제안하였다. 먼저, 비전-언어 모델인 CLIP을 해충 이미지를 대상으로 사전 학습하여 이미지 인코더와 텍스터 인코더를 공유 임베딩 공간에서 정렬하였다. 그리고 사전 학습된 텍스트 인코더를 통해 해충 클래스별 텍스트 특징을 추출하여 기존 ResNet-50 모델의 시각 특징과 결합할 수 있는 멀티모달 특징 퓨전 모델을 개발하였다. 실험 결과에서 제안한 멀티모달 시각 특징 퓨전 모듈이 기존 해충 이미지 분류 모델과 통합되어 시각 특징 강화에 효과적임을 검증하였다. 또한 기존 CNN, ViT 및 하이브리드 계열의 대표 모델보다 더 우수한 해충 분류 결과를 획득할 수 있었다.

본 연구에서 사용한 텍스트 프롬프트는 해충 클래스명에 기반하고 있어, 텍스트 정보가 제한적이라는 한계로 인해 융합 효과가 충분히 발휘되지 못할 수 있다. 이러한 한계를 보완하기 위해, 향후 연구에서는 앙상블 프롬프트 학습 모델이나 생성형 AI를 활용한 프롬프트 확장 기법을 적용할 예정이다.

## Acknowledgement

'2025년도 한국정보기술학회 하계종합학술대회에서 발표한 논문(텍스트 프롬프트 사전 정보를 활용한 해충 이미지 분류)[25]을 확장한 것임'

## References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision", Proc. International Conference on Machine Learning, Vienna, Austria, pp. 8748-8763, Jul. 2021. https://doi.org/10.48550/arXiv.2103.00020.

[2] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision", Proc. International Conference on Machine Learning, Virtual, Vol. 139, pp. 4904-4916, Jul. 2021. https://doi.org/10.48550/arXiv.2102.05918.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", Proc. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, pp. 770-778, Jun. 2016. https://doi.org/10.1109/CVPR.2016.90.

[4] J.-H. Lee and C.-H. Son, "Trap-based pest counting: Multiscale and deformable attention CenterNet integrating internal LR and HR joint feature learning", Remote Sensing, Vol. 15, No. 15, pp. 3810, Jul. 2023. https://doi.org/10.3390/rs15153810.

[5] H.-J. Yu, C.-H. Son, and D. H. Lee, "Apple leaf disease identification through region-of-interest aware deep convolutional neural network", Journal of Imaging Science and Technology, Vol. 64, No. 2, pp. 20507-1–20507-10, Jan. 2020. https://doi.org/10.2352/j.imagingsci.technol.2020.64.2.020507.

[6] C. Wang, J. Zhang, J. He, W. Luo, X. Yuan, and L. Gu, "A two-stream network with complementary feature fusion for pest image classification", Engineering Application of Artificial Intelligence, Vol. 124, pp. 106563, Sep. 2023. https://doi.org/10.1016/j.engappai.2023.106563.

[7] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection", Proc. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 2117-2125, Jul. 2017. https://doi.org/10.1109/CVPR.2017.106.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation", Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, pp. 234–241, Oct. 2015. https://doi.org/10.1007/978-3-319-24574-4_28.

[9] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks", Proc. IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, pp. 510-519, Jun. 2019. https://doi.org/10.1109/CVPR.2019.00060.

[10] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks", Proc. IEEE International Conference on Computer Vision, Venice, Italy, pp. 764-773, Oct. 2017. https://doi.org/10.1109/ICCV.2017.89.

[11] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions", arXiv, online, Nov. 2015. https://doi.org/10.48550/arXiv.1511.07122.

[12] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection", Proc. IEEE International Conference on Computer Vision, Venice, Italy, pp. 2999-3007, Oct. 2017. https://doi.org/10.1109/ICCV.2017.324.

[13] Z. Dai, H. Liu, Q. V. Le, and M. Tan,

"CoAtNet: marrying convolution and attention for all data sizes", arXiv, online, Jun. 2021. https://doi.org/10.48550/arXiv.2106.04803.

[14] E. C. Tetila, B. B. Machado, G. K. Menezes, A. D. S. Oliveira, M. Alvarez, W. P. Amorim, N. A. de S. Belete, G. G. da Silva, and H. Pistori, "Automatic recognition of soybean leaf diseases using UAV images and deep convolutional neural networks", IEEE Geoscience and Remote Sensing Letters, Vol. 17, No. 5, pp. 903-907, May 2020. https://doi.org/10.1109/LGRS.2019.2932385.

[15] G.-E. Kim, C.-H. Son, and S. Lee, "ROI-aware multiscale cross-attention vision transformer for pest image identification", Computers and Electronics in Agriculture, Vol. 237, pp. 110732, Oct. 2025. https://doi.org/10.1016/j.compag.2025.110732.

[16] H.-J. Yu and C.-H. Son, "Leaf spot attention network for apple leaf disease identification", Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, pp. 229-237, Jun. 2020. https://doi.org/10.1109/CVPRW50498.2020.00034.

[17] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module", Proc. European Conference on Computer Vision, Munich, Germany, pp. 3-19, Sep. 2018. https://doi.org/10.1007/978-3-030-01234-2_1.

[18] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers", Proc. IEEE International Conference on Computer Vision, Montreal, QC, Canada, pp. 6804-6815, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.00675.

[19] C. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification", Proc. IEEE/CVF International Conference on Computer Vision, Montreal, Canada, pp. 357-366, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.00041.

[20] X. Wu, C. Zhan, Y.-K. Lai, M.-M. Cheng, and J. Yang, "IP102: A large-scale benchmark dataset for insect pest recognition", Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, pp. 8787-8796, Jul. 2019. https://doi.org/10.1109/CVPR.2019.00899.

[21] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention", Proc. International Conference on Machine Learning, Virtual, Vol. 139, pp. 10347-10357, Jul. 2021. https://doi.org/10.48550/arXiv.2012.12877.

[22] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions", Proc. IEEE/CVF International Conference on Computer Vision, Montreal, Canada, pp. 568-578, Oct. 2021. https://doi.org/10.1109/ICCV48922.2021.00061.

[23] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3", Proc. IEEE International Conference on Computer Vision, Seoul, Korea, pp. 1314-1324, Oct. 2019. https://doi.org/10.1109/ICCV.2019.00140.

[24] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks", Proc. International Conference on Machine Learning, Long Beach, CA, USA, pp. 6105-6114, Jun. 2019. https://doi.org/10.48550/arXiv.1905.11946.

[25] D. H. Kim and C. H. Son, "Pest image classification with text prompt-driven prior information", Proc. Korea Information Technology Conference, Jeju, Korea, pp. 671-674, Jun. 2025.

## 저자소개


### 김 도 현 (Do-Hyun Kim)

2022년 3월 ~ 현재 :
국립군산대학교 소프트웨어학과
학사과정
관심분야 : 컴퓨터 비전, 영상처리,
기계학습, 딥 러닝

### 손 창 환 (Chang-Hwan Son)

2002년 2월 : 경북대학교
전자전기공학부(공학사)
2004년 2월 : 경북대학교
전자공학과(공학석사)
2008년 8월 : 경북대학교
전자공학과(공학박사)
2017년 4월 ~ 현재 :
국립군산대학교 소프트웨어학과 교수
관심분야 : 컴퓨터 비전, 영상처리, 기계학습, 딥 러닝