

An Ensemble Approach for Outdoor Fire Detection Using DINOv2 and Swin Transformer to Minimize Unwanted Fire Alarms

Miseo Kim*, Eun-Jin Kim**, Young-Seob Jeong***

*Researcher, FS Co., Daejeon, Korea

**Student, Dept. of Computer Engineering, Chungbuk National University, Cheongju, Korea

***Professor, Dept. of Computer Engineering, Chungbuk National University, Cheongju, Korea

[Abstract]

This study proposes a DINOv2-Swin ensemble model based on Transformers to improve outdoor fire detection accuracy and address the persistent issue of Unwanted Fire Alarms (UFAs). Outdoor environments pose challenges for precise detection due to vast areas and complex backgrounds, leading to high false alarm rates. To overcome this, we combined the strengths of DINOv2 for global visual feature extraction and the Swin Transformer for localized pattern recognition. To evaluate practical applicability, 1-FPR — i.e., specificity (true negative rate), computed as one minus the false positive rate — was adopted as a key metric alongside accuracy to assess the suppression of false alarms. Experimental results using the FLAME (Fire Luminosity Analysis in Multispectral Images) dataset demonstrated that the DINOv2-Swin model achieved an accuracy of 81.21% and a 1-FPR of 94.31%, representing a 4.98%p improvement in accuracy over the previous Xception-based approach. Furthermore, additional evaluation on the external D-Fire dataset provided evidence of the model's cross-dataset generalization potential. Overall, this research suggests that an image-based Transformer ensemble can help reduce UFAs, with potential applicability to firefighting response efficiency and future intelligent fire surveillance systems.

► **Key words:** Outdoor Fire Detection, Transformer Ensemble, DINOv2, Swin Transformer, Unwanted Fire Alarms

-
- First Author: Miseo Kim, Corresponding Author: Young-Seob Jeong
 - *Miseo Kim (oasiwm@fscom.kr), FS Co.
 - **Eun-Jin Kim (dw0815@chungbuk.ac.kr), Dept. of Computer Engineering, Chungbuk National University
 - ***Young-Seob Jeong (ysjay@chungbuk.ac.kr), Dept. of Computer Engineering, Chungbuk National University
 - Received: 2026. 07. 21, Revised: 2026. 09. 10, Accepted: 2026. 09. 14.

[요 약]

본 연구에서는 야외 화재 탐지의 정확도를 제고하고 빈번한 비화재보 문제를 해결하기 위해 Transformer 기반의 DINOv2-Swin 앙상블 모델을 제안한다. 야외 환경은 광범위한 영역과 복잡한 배경으로 인해 오경보 발생률이 높다는 한계가 있다. 이를 극복하고자 전역적 시각 특징 추출에 강점을 가진 DINOv2와 지역적 패턴 포착에 최적화된 Swin Transformer의 장점을 결합하여 탐지 정밀도를 극대화하였다. 특히 실용성을 고려해 정확도뿐만 아니라 오경보 억제율을 나타내는 1-FPR(specificity, $TN/(TN+FP)$)을 주요 지표로 설정하였다. 야외 화재 데이터셋인 FLAME(Fire Luminosity Analysis in Multispectral Images)를 활용한 실험 결과, 제안 모델은 81.21%의 정확도와 94.31%의 1-FPR을 달성하였으며, 이는 기존 Xception 모델 대비 약 4.98%p 향상된 성능이다. 아울러 외부 데이터셋인 D-Fire에 대한 추가 평가를 실시하여 교차 데이터셋 일반화 성능(cross-dataset generalization)에 대한 근거를 제시하였다. 본 연구는 이미지 기반 앙상블 기법이 비화재보를 최소화하고 향후 소방 대응 효율성을 개선할 수 있는 가능성을 시사한다는 점에서 의의가 있으며, 향후 지능형 화재 감시 시스템 분야에서의 활용 가능성을 확인하였다.

▶ **주제어:** 야외 화재 탐지, 트랜스포머 앙상블, 비화재보

I. Introduction

화재는 인명 및 재산에 심각한 피해를 야기하는 치명적인 재난 중 하나이다. 특히 야외 환경에서 발생하는 화재는 건조한 기후와 강풍 등 기상 요건에 따라 불길이 급격히 확산되는 특성을 지니며, 이는 산불이나 대형 화재로 이어져 통제 불가능한 수준의 피해를 야기할 수 있다. 이러한 특성상 화재 발생 초기 대응의 중요성이 더욱 강조된다. 이에 따라 화재 징후를 신속히 포착하기 위한 인공지능 기반 화재 예측 시스템 연구가 활발히 진행되고 있다.

그러나 기존의 화재 예측 기술은 탐지 정확도 향상에 치중한 나머지, 실제 화재가 아님에도 경보가 발생하는 ‘비화재보’ 문제를 초래하고 있다. 소방청 통계[1]에 따르면 2024년 전체 화재 신고 392,603건 중 약 41.97%인 164,793건이 비화재보로 집계되었으며, 이는 2019년 (25,020건) 대비 약 558.65% 급증한 수치이다.

이러한 비화재보의 급증은 단순한 기기 오작동의 문제를 넘어 심각한 사회·경제적 부작용을 낳는다. 잦은 오경보로 인한 불필요한 출동은 한정된 소방 인력과 장비의 운영 효율성을 저해하며, 실제 긴급 상황에 투입되어야 할 대응 체계의 피로도를 누적시킨다. 더욱 우려되는 지점은 시스템에 대한 신뢰도 저하이다. 반복되는 오경보는 관리자와 시민들의 경각심을 무디게 만들어, 정작 실제 화재가 발생했을 때 경보를 무시하거나 대응을 지연시키는 치명적인 결과를 초래할 수 있다. 따라서 현대의 화재 탐지 기술은 높은 탐지 정확도뿐만 아니라, 오경보율(False

Alarm Rate)을 획기적으로 낮출 수 있는 정밀성을 동시에 확보해야 하는 과제에 직면해 있다.

본 연구에서는 이러한 한계를 극복하고자 최신 딥러닝 아키텍처 중 탁월한 시각적 특징 추출 능력을 입증한 트랜스포머(Transformer) 기반 비전 모델 2종을 결합한 앙상블 모델을 설계하였다. 최근 센서 데이터와 이미지 데이터를 통합하여 화재 여부를 판단하는 멀티모달(Multimodal) 화재 예측 연구가 주목받고 있으나, 본 연구에서는 야외 환경이라는 지리적 특수성에 주목하였다. 야외 환경은 감시 범위가 광활하여 모든 구역에 센서를 설치하고 유지보수하기에는 경제적·환경적 제약이 크다. 따라서 본 논문에서는 시각적 정보만을 활용하되, 앙상블 기법을 통해 단일 모델의 한계를 극복하고 오경보율을 최소화하는 정밀한 화재 탐지 알고리즘을 제안하고자 한다.

본 연구의 주요 기여는 다음과 같다.

- 전역적 특징(DINOv2)과 지역적 특징(Swin Transformer)의 결합 방식이 비화재보 억제에 미치는 영향을 실증적으로 분석한다. Concat, Weighted Sum, Cross-Attention 세 가지 융합 전략을 동일 조건에서 비교하여, 정확도가 가장 높은 Cross-Attention이 오히려 1-FPR을 6.09%p 저하시킴을 확인하고, 오경보 억제 관점에서는 단순 Concat 방식이 가장 효과적임을 규명한다.

- 단순 정확도 외에 비화재보 억제율(1-FPR)을 핵심 평가 지표로 채택하여, 실제 소방 행정 시스템 관점에서의 실용적 성능 평가 체계를 제시한다.
- FLAME 데이터셋 실험을 통해 기존 Xception[15] 기반 모델 대비 정확도 4.98%p 향상(81.21%)과 높은 1-FPR(94.31%)을 달성함을 검증하고, 재학습 없이 독립적인 D-Fire[25] 데이터셋에 적용한 zero-shot cross-dataset 평가에서도 비교 모델 중 가장 우수한 일반화 성능(Accuracy 74.90%, F1 70.33%)을 확보하여 실제 배포 환경에서의 신뢰성 향상 가능성을 시사한다.

본 논문의 구성은 다음과 같다. 2장에서는 야외 화재 탐지 분야의 관련 연구를 소개하고, 기존 연구의 한계를 분석한다. 3장에서는 제안하는 DINOv2-Swin 앙상블 모델의 아키텍처를 상세히 기술한다. 4장에서는 사용 데이터셋, 실험 설정, 성능 비교 결과와 융합 방식 비교(Fusion Ablation) 및 교차 데이터셋 일반화 평가를 제시하며, 5장에서 결론 및 향후 연구 방향을 논의한다.

II. Preliminaries

1. Related works

최근 야외 화재 예측·감지 분야에는 센서 데이터와 영상 데이터를 융합하는 멀티모달 접근의 연구가 주목받고 있다. 그러나 야외 환경의 특성상, 센서를 넓은 범위에 설치 및 유지보수를 하는 것은 힘들기 때문에 야외 화재에서는 여전히 RGB 영상 데이터를 이용한 연구가 활발히 진행되고 있다[2-4].

이러한 영상 데이터를 사용한 단일 모달 기반 연구는 정확도 향상에만 집중되어 있어 실제 야외 환경에서 발생하는 오탐 문제는 상대적으로 충분히 다루어지지 않았다. 대표적으로, [2]에서는 모델의 일반화 성능을 평가하여 일부 오탐이 발생할 수 있음을 보여주었다. 이러한 오탐 문제는 야외 환경에서 특히 심각한데, 산불 연기 탐지를 위해 CNN(Convolutional Neural Network)[5], LSTM(Long Short-Term Memory)[6], ViT(Vision Transformer)[7]를 결합한 이미지 분류 모델인 SmokeyNet 연구[8]에 따르면, 이미지에서 연기의 경계가 모호하며, 구름, 안개 등으로 인해 오탐이 발생하는 문제가 있다고 보고하였다. 최근 딥러닝 기반 화재·연기 탐지 방법론에 대한 종합 리뷰[16]와 Gao 등[17]의 연구에서도 지적되었듯, 햇빛 반사, 안개, 노을 등 환경적 교란 요인으로 인한 비화재보(Unwanted Fire Alarm, UFA) 억제는 실제 야외 화재 감지 시스템의

실용성을 결정짓는 핵심 과제로 떠오르고 있다.

이러한 시각적 교란 및 오탐을 극복하기 위해, 최근에는 계층적 구조와 이동 창(Shifted Window) 주의집중 메커니즘을 갖춘 Swin Transformer[11] 기반 연구들(DynamicSwin-Fire[18], FireNet-KD[19], Swin-Faster R-CNN[20])이 활발히 시도되고 있다. 이들 연구는 Swin Transformer가 복잡한 야외 배경 속에서 국소적(localized) 화염 및 연기의 세밀한 경계 패턴을 포착하는데 뛰어남을 입증하였다. 한편, 자기지도학습 기반 파운데이션 모델인 DINOv2[10]는 대규모 무라벨 데이터로부터 강력한 전역적 시각 표현(global visual representation)을 추출하여 조명 변화 및 복잡한 배경 노이즈에 강건한 특성을 보인다. 최근 RF-DETR[21] 연구는 DINOv2 백본이 야외 환경의 다양한 시각적 간섭 속에서도 우수한 특징 표현력을 제공함을 확인하였다.

Table 1. Comparison of related outdoor fire detection studies

Study	Methodology	Dataset	Key Limitations
[2]	CNN-based classification	Wildfire	Lack of false alarm rate evaluation
[3]	Deep learning method review	Various	Lack of outdoor-specific focus
[8]	CNN+LSTM+ViT	HPWREN	Confusion between smoke and clouds
[9]	RGB-Thermal fusion	UAV	Limitations in heterogeneous sensor synchronization
[12]	Xception	FLAME	False alarm rate not considered
[18]	DynamicSwin-Fire	Self-collected	Focus on local features, lack of global context
[21]	RF-DETR (DINOv2)	Custom Outdoor	Ensemble not applied
[22]	UAV-FlameNet	UAV Fire	Lack of systematic evaluation of false-alarm suppression metrics
Proposed model	DINOv2-Swin Ensemble	FLAME	-

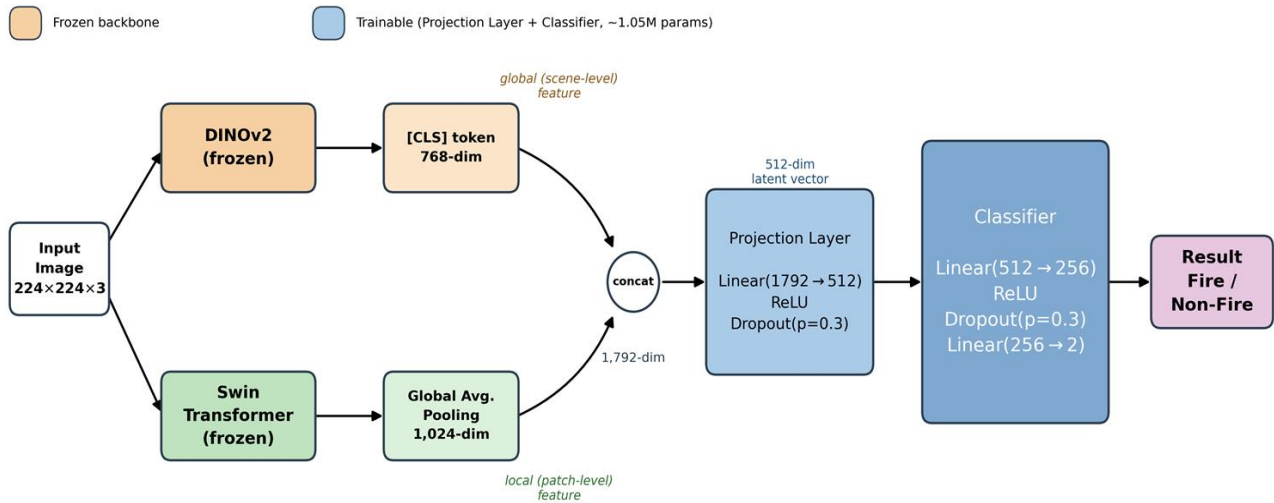


Fig. 1. Model Architecture.

최근에는 RGB 영상의 시각적 한계를 보완하고자 UAV 기반 RGB-Thermal 데이터셋인 RGBT-3M[9]과 같이 광학 및 열화상 영상을 융합하는 접근법도 연구되고 있다. 또한, FASDD[23]나 DetectiumFire[24], UAV-FlameNet[22], D-Fire[25] 등 다양한 환경을 포괄하는 대규모 공개 데이터셋 구축과 주의집중 기반 융합 기법들이 제안되고 있으나, 이러한 방식들 또한 공개 데이터셋의 부족, 서로 다른 이기종 센서 간의 시간 및 공간적 동기화 문제, 그리고 실시간 연산 부담 등의 기술적 난점을 수반한다.

Table 1은 야외 화재 탐지 관련 주요 연구들의 방법론과 한계를 요약한 것이다. 기존 연구들은 탐지 정확도 향상에 초점을 맞추거나, 멀티모달 RGB-Thermal 등 추가 센서를 활용하는 방향으로 발전해 왔으나, 각각 환경적 제약, 오경보를 미고려, 데이터 구축 및 센서 동기화의 한계라는 문제점을 안고 있다.

이에 본 연구에서는 RGB 이미지 기반의 단일 모달 접근법을 유지하면서도, Vision Transformer(ViT) 기반의 앙상블 학습 구조를 도입하여 오탐률을 획기적으로 낮추는 방안을 제안한다. 본 연구는 앙상블 기법을 통해 정확도와 오탐 제어 능력을 동시에 확보함으로써 화재 경보의 신뢰성을 높였다는 점에서 기존 연구와 명확한 차별성을 가진다.

III. The Proposed Scheme

1. Method

1.1 Model Architecture

본 연구에서는 DINOv2[10]와 Swin Transformer[11]의 앙상블 구조를 설계하였다. 제안 모델의 전체 구조는 Fig. 1과 같다.

DINOv2는 self-distillation 기반의 self-supervised learning을 통해 전역적 특징을 효과적으로 학습하며, 오탐을 억제하는 능력이 뛰어나다. Swin Transformer는 hierarchical 구조와 shifted window 기반의 self-attention을 통해 이미지의 지역적 특징을 세밀하게 포착하며, 전반적인 분류 정확도 향상에 기여한다.

제안된 앙상블 모델은 다음과 같이 구성된다. DINOv2의 [CLS] 토큰에서 추출한 전역적 특징(768차원)과 Swin Transformer의 마지막 stage에서 수행한 global average pooling을 통해 추출한 지역적 특징(1,024차원)을 연결(concatenation)하여 1,792차원의 결합 벡터를 구성한다.

이후 이 결합 벡터는 Projection layer를 통과한다. Projection layer는 1계층 MLP(Multi-Layer Perceptron)로 구성되며, Linear(1,792→512) - ReLU - Dropout(p=0.3)의 순서로 이루어져 512차원의 잠재 표현으로 투영된다. 최종 분류는 2계층 MLP로 구성된 Classifier를 통해 수행되며, Linear(512→256) - ReLU - Dropout(p=0.3) - Linear(256→2)의 구조를 가진다. 이 설계는 전역적 정보와 지역적 정보를 동시에 활용하여 안정적이고 신뢰도 높은 분류 성능을 제공한다.

Swin Transformer는 각 stage에서 윈도우 단위 self-attention을 적용하며, shifted window 기법을 통해 인접 윈도우 간 정보를 효율적으로 통합한다. DINOv2는 전체 이미지를 대표하는 [CLS] 토큰 임베딩을 활용하여 장면 전체의 문맥적 이해를 수행한다.

두 백본 모델은 학습 중 freeze 상태로 유지되며, Projection layer와 Classifier의 파라미터만 업데이트된다. 전체 모델의 파라미터 수는 약 175M개이며, 이 중 학습 가능한 파라미터는 약 1.05M개이다. 단일 이미지(224x224)

기준 연산량은 약 74.24G FLOPs로 측정되었다.

IV. Experiments

1. Dataset

본 연구에서는 공개 데이터셋인 The FLAME dataset: Aerial Imagery Fire burn detection using drones (UAVs)[12]를 사용하였다. FLAME 데이터셋은 미국 애리조나 북부 지역의 소각 실험 과정에서 드론(UAV)을 통해 수집된 항공 영상 데이터이다. 해당 데이터셋은 RGB 광학 비디오와 적외선 열화상(Thermal) 영상을 동시에 제공하며, 화재의 시각적 형태뿐만 아니라 온도 분포에 따른 열적 특성을 통합적으로 분석할 수 있다는 장점이 있다.

FLAME 데이터셋은 연구 목적에 따라 화재 분류(Fire Classification)와 화재 분할(Fire Segmentation) 두 가지 과제로 구분되어 설계되었으며, 본 연구에서는 이미지 내 화재 발생 여부를 판별하기 위해 화재 분류 데이터셋을 실험에 사용하였다.

화재 분류 데이터는 “Fire”와 “Non-Fire”의 이진 분류가 가능하도록 라벨링 되어 있으며, 학습용으로 39,375장, 테스트용으로 8,617장의 이미지가 포함되어 있다. 분류 데이터는 열화상 이미지를 포함하지 않고 RGB 이미지로만 구성되어 있다.

2. Composition of Training Data

실험에 사용되는 원본 FLAME 분류 데이터셋은 39,375장의 학습용 데이터, 8,617장의 테스트 데이터로 구성되어 있다. 본 연구에서는 학습용 데이터의 10%를 검증용 데이터로 사용하여, 학습용 데이터 35,439장, 검증용 데이터 3,936장, 테스트 데이터 8,617장을 사용하였다.

학습용(39,375장)/테스트용(8,617장) 분할은 FLAME 데이터셋 제공자[12]가 제공한 공식 split을 그대로 사용하였으며, 본 연구에서 자체적으로 추가한 것은 학습용 데이터 내에서의 학습/검증 분할뿐이다. 해당 학습/검증 분할은 random seed 42로 고정된 프레임 단위 무작위 분할(frame-level random split)로 수행되었다. FLAME 데이터셋은 복수의 영상 시퀀스에서 추출된 프레임으로 구성되어 있어 시간적으로 인접한 프레임이 학습과 검증 세트에 동시에 포함될 수 있으나, 이 분할은 checkpoint 선택(early stopping)에만 관여하며 테스트 세트는 원 데이터셋 제공자가 지정한 분할을 그대로 유지하여 학습/검증 과정과 완전히 독립적이다. 다만 validation 세트 내 시간적

중복이 모델 선택 과정에 간접적인 영향을 줄 가능성은 배제할 수 없으며, 이는 향후 sequence-disjoint 분할을 통해 검증할 필요가 있다.

3. Experimental Setup

본 연구에서는 Meta AI에서 제안한 DINOv2-base(이하 DINOv2)와 Microsoft에서 제안한 Swin Transformer-base(이하 Swin Transformer)를 결합한 앙상블 모델을 사용하였으며, Hugging Face에 공개된 “facebook/dinov2-base”와 “microsoft/swin-base-patch4-window7-224” 가중치를 적용하였다.

두 backbone은 freeze한 채로 projector와 classifier만 학습하는 방식으로 fine-tuning을 진행하였다. 입력 이미지는 다양한 해상도로 구성되어 있기 때문에 PyTorch의 transforms를 이용하여 224x224로 resize하였으며, 채널별 평균(mean=[0.485, 0.456, 0.406])과 표준편차(std=[0.229, 0.224, 0.225])를 사용하여 정규화 후 모델에 입력하였다. Resize에는 BICUBIC 보간법을 적용하였다. Projector는 1계층 MLP, classifier는 2계층 MLP로 구성하였으며, 활성화 함수로는 ReLU를 사용하였다.

손실 함수로는 CrossEntropyLoss를 사용하였으며, optimizer는 AdamW를 적용하였다. 초기 학습률은 $1e-5$ 로 설정하고 Hugging Face의 linear learning rate scheduler를 적용하여 학습 중 유동적으로 변경되도록 하였으며, 학습 초반 5%를 warmup 단계로 설정하였다. 배치 크기는 64로 설정하였으며, 총 100 epoch의 학습 중 5 epoch 동안 validation loss가 감소하지 않으면 학습을 멈추는 early stopping(patience=5) 기법을 적용하였다. 최적의 모델은 validation loss가 가장 낮을 때 체크포인트를 저장하는 방식으로 선택하였다.

실험의 재현성을 위해 random seed를 42로 고정하였으며, Python random, NumPy, PyTorch의 모든 난수 생성기에 동일한 seed를 적용하였다. 실험은 NVIDIA RTX4090 GPU 2장(CUDA 12.2, Driver 535.183.01) 환경에서 진행되었다. 제안 모델(DINOv2-Swin)의 FLAME 데이터셋 기준 총 학습 시간은 약 143분(48 epoch, epoch당 평균 약 3분)이었다.

Table 2. Performance comparison of vision models on FLAME test set.

Model	Acc.	Precision	Recall	F1	1-FPR
DINOv2-base	78.49	92.35	69.71	79.48	91.46
CLIP-ViT-B/16	71.74	79.91	70.24	74.80	73.93
SigLIP-ViT-B/16	64.49	70.40	69.77	70.09	56.70
Swin-Base	81.16	80.30	90.62	85.17	67.18
ConvNeXt-Base	86.12	93.07	82.89	87.69	90.89
EfficientNet V2-M	76.79	82.45	77.57	79.94	75.63
EVA-CLIP-ViT-B/16	89.60	94.35	87.81	90.97	92.24
DINOv2-Swin (Ours)	81.21	94.94	72.34	82.10	94.31

4. Results

본 실험은 정확도뿐만 아니라 오경보율을 줄이는 것을 목표로 한다. 따라서 정확도(Accuracy), Precision, Recall, F1-Score, 그리고 1-FPR(specificity, $TN/(TN+FP)$)을 평가 지표로 활용하였다.

본 연구에서는 화재 탐지의 정확도 향상뿐만 아니라 오경보율을 최소화하기 위해 다양한 비교 실험을 수행하였다. 비교 실험에는 DINOv2, CLIP[13], SigLIP[14], Swin Transformer, 그리고 지도학습(supervised) 기반의 최신 CNN 계열 백본인 ConvNeXt-Base[26], EfficientNetV2-M[27], 대조학습(contrastive learning) 기반의 대규모 사전학습 모델 EVA-CLIP-ViT-B/16[28]을 포함한 총 7개 모델에 대해 진행하였다(Table 2). 각 모델은 Hugging Face에 공개된 체크포인트 "facebook/dinov2-base", "openai/clip-vit-base-patch16", "google/siglip-base-patch16-224", "microsoft/swin-base-patch4-window7-224", "timm/convnext_base.fb_in1k", "timm/tf_efficientnetv2_m.in21k", "timm/eva02_base_patch16_clip_224.merged2b"를 사용하였다. 모든 모델에 대해 실험은 backbone은 freeze한 상태로 projector와 classifier만 학습하도록 하였고, 동일한 FLAME 데이터셋 하에서 공정하게 평가되었다.

Fig. 2는 제안 모델(DINOv2-Swin)의 테스트셋에 대한 혼동 행렬(Confusion Matrix)을 나타낸다. 비화재(Non-Fire) 이미지의 94.31%를 정확히 분류하였으며, 화재(Fire) 이미지에 대해서는 72.34%의 재현율을 달성하였다.

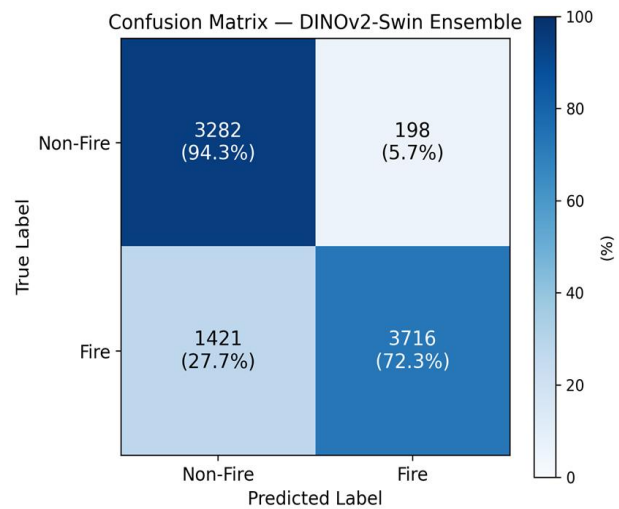


Fig. 2. Confusion Matrix of the proposed DINOv2-Swin model on the FLAME test set.

비교 결과, Accuracy는 EVA-CLIP-ViT-B/16이 89.60%로 가장 우수하였고 ConvNeXt-Base(86.12%)가 뒤를 이었다. 그러나 두 모델의 1-FPR은 각각 92.24%, 90.89%로 제안 모델(94.31%)에는 미치지 못하였으며, 이는 Recall이 각각 87.81%, 82.89%로 매우 높은 대신 화재와 유사한 비화재 패턴에 대한 오탐도 함께 증가하는 경향을 보였기 때문이다. EfficientNetV2-M(Accuracy 76.79%, 1-FPR 75.63%)은 상대적으로 가벼운 CNN 계열 모델로, 오탐 억제와 탐지 정확도 모두 중간 수준에 그쳤다.

Swin Transformer는 1-FPR이 67.18%로 가장 낮게 나타났는데, 이는 Recall이 90.62%로 매우 높아 화재 이미지 대부분을 탐지하는 반면, 연기·노을·강한 조명 등 화재와 시각적으로 유사한 비화재 패턴도 화재로 오분류하는 경향이 있기 때문이다(False Positive 1,142건). 이는 Swin Transformer의 지역적 패턴 포착 능력이 오히려 세밀한 국소 특징에 과도하게 반응하여 발생하는 현상으로 해석된다.

반면, DINOv2는 self-distillation 기반의 전역 표현 학습을 통해 오탐을 효과적으로 억제하며 1-FPR 91.46%로 높은 성능을 보였다. 이러한 결과는 Swin Transformer와 DINOv2가 각각 지역적 시각 정보와 전역적 맥락에 강점을 가진다고 해석할 수 있다. 따라서 하나의 모델만으로는 정확도와 1-FPR을 동시에 최적화하기 어렵다는 한계가 존재함을 확인할 수 있다.

제안하는 DINOv2-Swin 앙상블 모델은 Accuracy 측면에서는 EVA-CLIP-ViT-B/16, ConvNeXt-Base 등 대형 모델에 미치지 못하지만, 두 backbone의 상호보완적 표현을 결합함으로써 비교한 7개 모델 중 가장 높은

1-FPR(94.31%)과 Precision(94.94%)을 달성하여 오경보 억제 관점에서 가장 우수한 성능을 보였다.

Table 3은 본 연구의 DINOv2-Swin 앙상블 모델과 FLAME 데이터셋을 사용한 기존 연구[12]에서 제시한 Xception 모델의 성능을 비교한 결과이다.

Table 3. Performance comparison between the proposed DINOv2-Swin and Xception[12] model.

Model	Acc.	Precision	Recall	F1	1-FPR
DINOv2-Swin (Ours)	81.21	94.94	72.34	82.10	94.31
Xception[12]	76.23	-	-	-	-

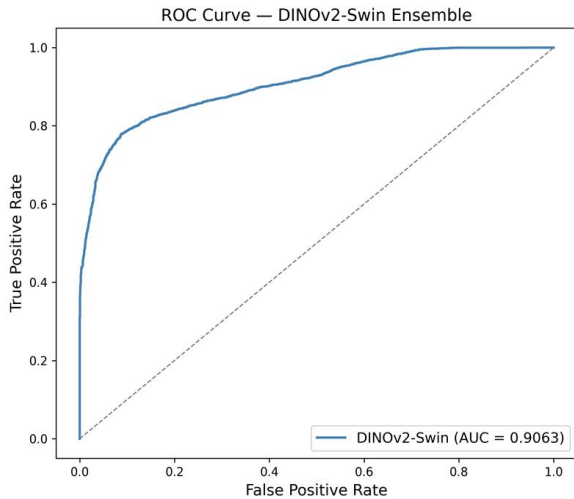


Fig. 3. ROC Curve of the proposed DINOv2-Swin model on the FLAME test set.

기존 연구[12]의 Xception 모델은 학습 데이터에서 화재와 비화재 클래스가 불균형하여 비화재 프레임이 화재로 잘못 분류되는 False Positive가 상대적으로 높게 나타났으며, 최종 분류 정확도는 76.23%였다. 본 연구에서는 데이터 증강 기법 없이도 정확도 81.21%를 달성하며 기존 연구보다 4.98%p 향상된 성능을 보였고, 1-FPR 또한 94.31%로 높은 비화재보 억제 성능을 보였다.

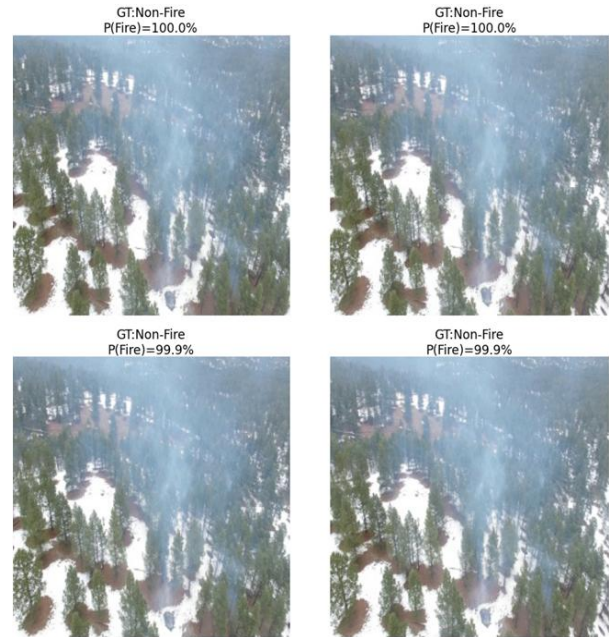


Fig. 4. Representative False Positive cases of the proposed DINOv2-Swin model.

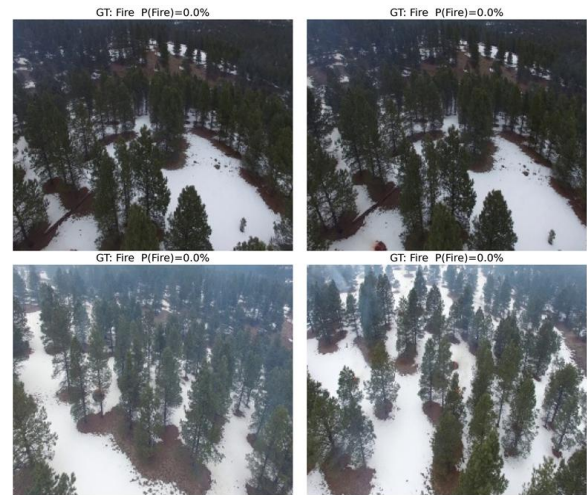


Fig. 5. Representative False Negative cases of the proposed DINOv2-Swin model.

Fig. 3은 제안 모델의 ROC(Receiver Operating Characteristic) 곡선을 나타낸다. AUC(Area Under the Curve) 값은 0.906으로, 다양한 분류 임계값에서도 안정적인 화재 탐지 성능을 유지함을 확인할 수 있다.

제안 모델의 오분류 사례에 대한 정성적 분석을 수행하였다. Fig. 4는 오탐(False Positive, FP) 대표 사례를, Fig. 5는 미탐(False Negative, FN) 대표 사례를 나타낸다.

FP 198건을 분석한 결과, 대부분의 케이스에서 나무 사이로 연기와 시각적으로 유사한 연무가 포착되어 화재로 오분류된 것을 확인하였다. FLAME 데이터셋은 실제 소각

실험 현장에서 수집된 특성상, Non-Fire로 레이블된 프레임에도 잔류 연무가 포함될 수 있어 시각적 판별을 어렵게 한다. 이러한 FP 케이스는 21개의 연속 프레임 클러스터로 집중되어 있어, 실질적인 오탐 장면 수는 개별 프레임 수보다 현저히 적다.

FN 1421건의 분석에서는 두 가지 경향이 확인되었다(Fig. 5). 대부분의 케이스는 화재 강도가 낮은 소강(smoldering) 단계에서 촬영된 프레임으로, 가시적인 불꽃이나 연기가 두드러지지 않아 비화재 장면과 시각적으로 구별이 어렵다($\text{fire_prob} \approx 0.000$). 또한 일부 케이스는 화재 예측 확률이 0.42~0.44로 분류 임계값(0.5)에 근접하여, 임계값 조정을 통해 탐지 가능성이 있는 경계 케이스에 해당한다. FN 케이스 또한 50개의 연속 클러스터로 발생하는 경향을 보인다.

이러한 분석은 향후 연구에서 자연 연무와 실제 연기를 구별하는 능력을 강화하고, 소강 화재 단계에 특화된 학습 전략을 도입할 필요성을 시사한다.

5. Fusion Method Ablation

Table 4는 두 backbone의 특징을 결합하는 융합 방식이 성능에 미치는 영향을 분석하기 위해, 동일한 DINOv2-Swin 이중 backbone 구조에 세 가지 융합 방법을 적용하여 비교한 결과이다. 모든 실험은 동일한 하이퍼파라미터($\text{lr}=1\text{e-}5$, $\text{batch}=64$, $\text{patience}=5$, $\text{seed}=42$)로 수행되었으며, 두 backbone은 모두 freeze하였다. 세 방법 모두 최종 classifier 구조($\text{Linear}(512 \rightarrow 256) - \text{ReLU} - \text{Dropout}(p=0.3) - \text{Linear}(256 \rightarrow 2)$)와 latent 차원(512)을 동일하게 유지하여, 융합 방식 자체의 효과만을 공정하게 비교할 수 있도록 하였다.

세 가지 융합 방법의 구체적인 구조는 다음과 같다.

- Concat (제안 방법): DINOv2 CLS 토큰(768-dim)과 Swin GAP feature(1,024-dim)를 직접 연결하여 1,792-dim 벡터를 구성한 후, $\text{Linear}(1792 \rightarrow 512) - \text{ReLU} - \text{Dropout}(p=0.3) - \text{Linear}(512 \rightarrow 256) - \text{ReLU} - \text{Dropout}(p=0.3) - \text{Linear}(256 \rightarrow 2)$ 로 구성된 분류기에 입력한다.
- Weighted Sum: DINOv2 특징(768-dim)과 Swin 특징(1,024-dim)을 각각 독립적인 Linear projection으로 동일 차원(512-dim)으로 변환한 후, 학습 가능한 스칼라 α ($=\text{sigmoid}(\alpha_{\text{logit}}$, 초기값 0.5)로 가중 합산한다: $f = \alpha \cdot f_{\text{DINO}} + (1 - \alpha) \cdot f_{\text{Swin}}$. 이후 $\text{ReLU} - \text{Dropout}(p=0.3) - \text{Linear}(512 \rightarrow 256) - \text{ReLU} - \text{Dropout}(p=0.3) - \text{Linear}(256 \rightarrow 2)$ 로 분류한다.

- Cross-Attention: DINOv2 특징(768-dim)과 Swin 특징(1,024-dim)을 각각 Linear projection으로 512-dim으로 변환한 후, DINOv2 투영 벡터를 Query(Q), Swin 투영 벡터를 Key(K) 및 Value(V)로 사용하는 Multi-Head Attention($\text{embed_dim}=512$, $\text{num_heads}=8$, $\text{head_dim}=64$)을 적용한다. Attention 출력에 residual connection과 LayerNorm을 적용한 후, $\text{Dropout}(p=0.3) - \text{Linear}(512 \rightarrow 256) - \text{ReLU} - \text{Dropout}(p=0.3) - \text{Linear}(256 \rightarrow 2)$ 로 분류한다. 단, 본 구조에서 Q와 K/V는 각각 단일 벡터(1개 토큰)이므로, attention score는 head당 스칼라 1개로 계산되고 이에 대한 softmax 정규화는 값에 관계없이 항상 1이 된다. 즉 attention weight가 두 특징 간 선택적 게이팅 기능을 수행하지 못하며, 결과적으로 이 구조는 일반적인 multi-token Cross-Attention이 아니라 추가적인 선형 변환과 정규화를 포함한 특징 결합기(feature combiner)에 가깝게 동작함에 유의해야 한다.

Table 4. Fusion method ablation study on FLAME test set.

Fusion Method	Accuracy	F1	1-FPR
Weighted Sum	81.17	82.32	92.41
Cross-Attention	83.95	85.76	88.22
Concat (Ours)	81.21	82.10	94.31

Cross-Attention은 Accuracy(+2.74%p)와 F1(+3.65%p) 측면에서 Concat보다 우수한 수치를 보인다. 그러나 본 연구의 핵심 목표인 비화재보(UFA) 억제 관점에서 Cross-Attention의 1-FPR은 88.22%로, Concat(94.31%)보다 6.09%p 낮다. 이는 Cross-Attention 적용 시 비화재 이미지를 화재로 오분류하는 False Positive가 유의미하게 증가함을 의미한다. 오경보는 소방 인력 출동, 시스템 신뢰도 저하 등 직접적인 사회적 비용을 수반하므로, UFA 최소화를 목적으로 하는 본 연구에서는 1-FPR이 우선 지표가 된다.

Weighted Sum 실험에서 학습된 α 값은 0.503(DINOv2)으로, 두 backbone이 거의 동등하게 기여함을 확인하였다. 이는 단순 가중 합산보다 두 특징 공간을 명시적으로 결합하는 Concat이 오경보 억제에 더 효과적임을 시사한다.

따라서 Concat 방식이 본 논문의 응용 목적(오경보 최소화)에 가장 부합하는 융합 전략임을 확인하였다.

6. Cross-Dataset Generalization Evaluation on D-Fire

Table 5는 제안 모델의 교차 데이터셋 일반화 성능(Cross-dataset generalization)과 미학습 야외 환경에서의 적용 가능성을 검증하기 위해, FLAME 데이터셋으로 학습된 모델들을 추가 재학습(Fine-tuning) 없이 독립적인 야외 화재 데이터셋인 D-Fire[25]에 직접 적용하여 평가를 수행한 결과이다.

D-Fire 데이터셋은 21,527장의 야외 화재 이미지로 구성된 공개 데이터셋으로, 공식 학습/테스트 분할(17,221/4,306장)을 제공한다. 원본 레이블은 YOLO 형식의 경계 박스(bounding box) 주석으로 제공되며, 본 실험에서는 화재(class 0) 경계 박스가 하나 이상 존재하는 이미지를 Fire(1), 그렇지 않은 이미지를 Non-Fire(0)로 변환하여 이진 분류 과제에 적용하였다. FLAME으로 학습된 모델을 추가 fine-tuning 없이 D-Fire 테스트셋(4,306장)에 그대로 적용하였으며, 테스트셋은 Fire 2,081장, Non-Fire 2,225장으로 구성된다.

Table 5. Cross-Dataset generalization.

Model	Accuracy	F1	1-FPR
DINOv2-base	67.70	66.02	70.29
CLIP-ViT-B/16	62.91	60.56	66.65
SigLIP-ViT-B/16	52.46	17.02	92.09
Swin-Base	69.67	59.57	91.60
ConvNeXt-Base	51.83	0.86	99.91
EfficientNetV2-M	51.83	1.71	99.51
EVA-CLIP-ViT-B/16	73.13	69.12	83.33
DINOv2-Swin (Ours)	74.90	70.33	87.37

실험 결과, 제안하는 DINOv2-Swin 앙상블 모델은 도메인 전환(Domain Shift) 조건하에서도 74.90%의 정확도(Accuracy)와 70.33%의 F1-Score를 기록하여, 모든 비교 baseline 모델 중 가장 우수한 종합 분류 성능을 달성하였다.

특히 SigLIP 모델의 경우 1-FPR이 92.09%로 높게 측정되었으나, 이는 시각적 도메인 변화로 인해 대부분의 프레임을 비화재(Non-Fire)로 편향되게 판별하는 극단적인 과소 예측(Under-detection) 현상에 기인한다. 이로 인해 실제 화재 프레임을 거의 포착하지 못하여 F1-Score가 17.02%로 급격히 붕괴하는 한계를 보였다. Swin-Base 역시 높은 1-FPR(91.60%)에 비해 F1-Score는 59.57%에 그쳐 화재 감지력과 오탐 억제력 간의 불균형을 나타냈다.

D-Fire 교차 데이터셋 평가에서 흥미로운 결과가 관찰되었다. FLAME 데이터셋 내 정확도에서 제안 모델을 능가했던 ConvNeXt-Base와 EfficientNetV2-M은 D-Fire로 도메인이 전환되자 각각 Recall 0.43%, 0.86%로 완전

히 붕괴하여, 사실상 모든 이미지를 Non-Fire로 예측하는 수준(Accuracy가 D-Fire 테스트셋의 클래스 비율과 일치하는 51.83%로 수렴)에 그쳤다. 이러한 현상은 Table 2(FLAME)에서는 관찰되지 않다가 D-Fire로 도메인이 전환된 이후에만 나타났다는 점에서, 두 모델의 학습 실패나 성능 저하가 아니라 frozen feature의 분포가 도메인에 따라 크게 달라진다는 것을 알 수 있다. 이는 지도학습 기반 사전학습 표현이 self-supervised/contrastive 표현에 비해 도메인 전환에 더 취약할 수 있음을 시사하는 결과이나, 본 연구에서는 그 근본 원인을 직접 규명하지는 않았으며 향후 연구 과제로 남긴다.

반면 self-distillation 기반의 DINOv2와 대규모 대조학습 기반의 EVA-CLIP은 붕괴 없이 각각 Accuracy 67.70%, 73.13%를 유지하였다. 이러한 대비는 사전학습 방식(supervised vs. self-supervised/contrastive)의 차이가 frozen backbone의 도메인 일반화 능력에 영향을 줄 가능성을 시사한다. 다만 두 모델군은 구조, 사전학습 데이터 규모, 파라미터 수 등에서도 동시에 차이가 있어, 본 실험만으로 사전학습 방식의 영향을 직접적인 원인으로 단정하기는 어렵다.

붕괴하지 않은 모델들만을 대상으로 비교하면, 제안하는 DINOv2-Swin 앙상블은 Accuracy 74.90%, F1 70.33%로 EVA-CLIP-ViT-B/16(73.13%, 69.12%)과 Swin-Base(69.67%, 59.57%)를 모두 상회하며, 비교 모델 중 가장 안정적인 cross-domain 일반화 성능을 보였다. 종합하면, 본 연구의 제안 모델은 FLAME 데이터셋 내 단일 최고 정확도를 달성하지는 못하였으나(Table 2), 도메인이 전환되는 실제 배포 환경을 고려한 cross-dataset 평가에서는 Accuracy·F1 지표에서 가장 우수한 일반화 성능을 보여, 실용적 관점에서 더 신뢰할 수 있는 모델임을 확인하였다.

V. Conclusions

본 연구에서는 야외 환경에서 화재를 안정적으로 탐지하고, 고질적인 문제인 비화재보를 개선하기 위해 Transformer 기반의 DINOv2-Swin 앙상블 모델을 구성하였다. 야외 환경은 광활한 범위와 복잡한 배경 특성상 단일 모델만으로는 정밀한 특징 추출에 한계가 있다. 이를 극복하고자 전역적 시각 표현 학습에 강점을 가진 DINOv2와 계층적 어텐션 구조를 통해 지역적 세부 패턴을 효과적으로 포착하는 Swin Transformer를 결합하였다. 이러한 하이브리드 접근 방식은 화재의 연기나 불꽃이

갖는 정형화되지 않은 시각적 패턴을 다각도에서 분석할 수 있게 하여 모델의 일반화 성능을 향상시켰다.

실용적인 소방 행정 및 안전 관리 시스템으로의 적용 가능성을 평가하기 위해, 본 연구에서는 정확도뿐만 아니라 오경보 억제율을 나타내는 1-FPR을 핵심 평가지표로 채택하였다. 야외 화재 RGB 이미지 데이터셋인 FLAME을 활용하여 성능을 평가한 결과, 제안 모델인 DINOv2-Swin은 81.21%의 정확도와 94.31%의 1-FPR을 달성했다. 이는 기존 선행 연구에서 제시된 Xception 모델과 비교했을 때 정확도 측면에서 약 4.98%p의 향상을 기록한 결과이다.

본 연구의 성과는 단일 모달 환경에서도 Transformer 앙상블 기법을 통해 정확도와 비화재보 억제 성능을 동시에 향상시킬 수 있음을 FLAME 데이터셋 상에서 확인하였다는 점에서 의의가 있다. 또한 D-Fire 데이터셋을 활용한 cross-dataset 평가에서 fine-tuning 없이도 비교 모델 대비 가장 높은 일반화 성능(Accuracy 74.90%, F1 70.33%)을 달성하여, 제안 모델의 도메인 이전(domain transfer) 가능성을 확인하였다. 향후 연구에서는 극심한 조명 변화나 기상 악조건 하에서의 성능 강건성 확보, 실시간 처리 효율성 개선, YOLO, RT-DETR, GroundingDINO 등 객체 탐지 기반 모델과의 비교 검증, 그리고 sequence-disjoint 분할을 통한 validation leakage 재검증 등을 과제로 남겨둔다.

ACKNOWLEDGEMENT

This research was supported by the Regional Innovation System & Education(RISE) program through the (Chungbuk Regional Innovation System & Education Center), funded by the Ministry of Education(MOE) and the (Chungcheongbuk-do), Republic of Korea.(2026-RISE-11-014-03) This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2020-NR049604).

REFERENCES

- [1] National Fire Agency. (2025). *2025 Statistical Yearbook of National Fire Agency (Technical Report No. 11-1661000-000010-10)*. National Fire Agency, Sejong, Korea.
- [2] I. El-Madafri, M. Peña, and N. Olmedo-Torre, "The Wildfire Dataset: Enhancing Deep Learning-Based Forest Fire Detection with a Diverse Evolving Open-Source Dataset Focused on Data Representativeness and a Novel Multi-Task Learning Approach," *Forests*, Vol. 14, No. 9, p. 1697, September 2023. DOI: 10.3390/f14091697
- [3] A. Saleh, M. A. Zulkifley, H. H. Harun, F. Gaudreault, I. Davison, and M. Spraggon, "Forest Fire Surveillance Systems: A Review of Deep Learning Methods," *Heliyon*, Vol. 10, No. 1, p. e23127, January 2024. DOI: 10.1016/j.heliyon.2023.e23127
- [4] R. N. Vasconcelos, W. J. S. Franca Rocha, D. P. Costa, S. G. Duverger, M. M. M. d. Santana, E. C. B. Cambui, J. Ferreira-Ferreira, M. Oliveira, L. d. S. Barbosa, and C. L. Cordeiro, "Fire Detection with Deep Learning: A Comprehensive Review," *Land*, Vol. 13, No. 10, p. 1696, October 2024. DOI: 10.3390/land13101696
- [5] K. O'Shea, and R. Nash, "An Introduction to Convolutional Neural Networks," *arXiv preprint arXiv:1511.08458*, November 2015. DOI: 10.48550/arXiv.1511.08458
- [6] C. B. Vennerød, A. Kjærran, and E. S. Bugge, "Long Short-Term Memory RNN," *arXiv preprint arXiv:2105.06756*, May 2021. DOI: 10.48550/arXiv.2105.06756
- [7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, October 2020. DOI: 10.48550/arXiv.2010.11929
- [8] A. Dewangan, Y. Pande, H.-W. Braun, F. Vernon, I. Perez, I. Altintas, G. W. Cottrell, and M. H. Nguyen, "FigLib & SmokeyNet: Dataset and Deep Learning Model for Real-Time Wildland Fire Smoke Detection," *Remote Sensing*, Vol. 14, No. 4, p. 1007, February 2022. DOI: 10.3390/rs14041007
- [9] Y. Zhang, X. Rui, and W. Song, "A UAV-Based Multi-Scenario RGB-Thermal Dataset and Fusion Model for Enhanced Forest Fire Detection," *Remote Sensing*, Vol. 17, No. 15, p. 2593, 2025. DOI: 10.3390/rs17152593
- [10] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning Robust Visual Features without Supervision," *arXiv preprint arXiv:2304.07193*, April 2023. DOI: 10.48550/arXiv.2304.07193
- [11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *arXiv preprint arXiv:2103.14030*, March 2021. DOI: 10.48550/arXiv.2103.14030
- [12] A. Shamsoshoara, F. Afghah, A. Razi, L. Zheng, P. Z. Fulé, and E. Blasch, "Aerial Imagery Pile Burn Detection Using Deep

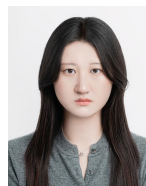
- Learning: The FLAME Dataset," *Computer Networks*, Vol. 193, p. 108001, July 2021. DOI: 10.1016/j.comnet.2021.108001
- [13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *arXiv preprint arXiv:2103.00020*, March 2021. DOI: 10.48550/arXiv.2103.00020
- [14] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid Loss for Language Image Pre-Training," *arXiv preprint arXiv:2303.15343*, March 2023. DOI: 10.48550/arXiv.2303.15343
- [15] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800-1807, Los Alamitos, USA, July 2017. DOI: 10.1109/CVPR.2017.195
- [16] A. Elhanashi, S. Essahraoui, P. Dini, and S. Saponara, "Early Fire and Smoke Detection Using Deep Learning: A Comprehensive Review of Models, Datasets, and Challenges," *Applied Sciences*, Vol. 15, No. 18, p. 10255, 2025. DOI: 10.3390/app151810255
- [17] Q. Gao, H.-H. Wang, Z. Qin, L. Fan, K. Li, C. Wang, and Q. Zhang, "Using Multimodal Large Language Models for False Alarm Reduction in Image-based Fire Detection," *Preprint*, 2026. DOI: 10.21203/rs.3.rs-8847038/v1
- [18] N. Ahmad, M. Akbar, E. H. Alkhamash, and M. M. Jamjoom, "DynamicSwin-Fire: A Dynamic Swin Transformer With Fire-Specific Attention and Edge-Guided Gradient Fusion for Robust Fire Segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 63, Art no. 4110815, 2025. DOI: 10.1109/TGRS.2025.3618172
- [19] N. Ahmad, M. Akbar, E. H. Alkhamash, and M. M. Jamjoom, "FireNet-KD: Swin Transformer-Based Wildfire Detection with Multi-Source Knowledge Distillation," *Fire*, Vol. 8, No. 8, p. 295, 2025. DOI: 10.3390/fire8080295
- [20] S. Choi, S. Kim, and H. Jung, "Optimized Faster R-CNN with Swintransformer for Robust Multi-Class Wildfire Detection," *Fire*, Vol. 8, No. 5, p. 180, 2025. DOI: 10.3390/fire8050180
- [21] I. Robinson, P. Robicheaux, M. Popov, D. Ramanan, and N. Peri, "RF-DETR: Neural Architecture Search for Real-Time Detection Transformers," *arXiv preprint arXiv:2511.09554*, 2026. DOI: 10.48550/arXiv.2511.09554
- [22] J. Hu, Q. Lin, W. Jiao, H. Liu, K. Jia, J. Liu, and J. Lv, "UAV-FlameNet: A Lightweight and High-Precision Flame Detection Model for UAV Aerial Fire Monitoring," *Frontiers in Plant Science*, Vol. 17, Art no. 1892035, 2026. DOI: 10.3389/fpls.2026.1892035
- [23] M. Wang, L. Jiang, P. Yue, D. Yu, and T. Tuo, "An Open Flame and Smoke Detection Dataset for Deep Learning in Remote Sensing Based Fire Detection," *Geo-Spatial Information Science*, Vol. 28, No. 2, pp. 511-526, 2025. DOI: 10.1080/10095020.2024.2347922
- [24] Z. Liu, S. H. Khajavi, and G. Jiang, "DetectiumFire: A Comprehensive Multi-modal Dataset Bridging Vision and Language for Fire Understanding," *Proceedings of the 39th Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026.
- [25] P. V. A. B. de Venâncio, A. C. Lisboa, and A. V. Barbosa, "An Automatic Fire Detection System Based on Deep Convolutional Neural Networks for Low-Power, Resource-Constrained Devices," *Neural Computing and Applications*, Vol. 34, No. 20, pp. 17823-17844, 2022. DOI: 10.1007/s00521-022-07467-z
- [26] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *arXiv preprint arXiv:2201.03545*, January 2022. DOI: 10.48550/arXiv.2201.03545
- [27] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," *arXiv preprint arXiv:2104.00298*, April 2021. DOI: 10.48550/arXiv.2104.00298
- [28] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "EVA-CLIP: Improved Training Techniques for CLIP at Scale," *arXiv preprint arXiv:2303.15389*, March 2023. DOI: 10.48550/arXiv.2303.15389

Authors



Miseo Kim received the B.S. and M.S. degrees in Computer Engineering from Chungbuk National University, Cheongju, Republic of Korea, in 2024 and 2026, respectively.

She is currently working as a researcher in the Systems Development Department at FS Co., Ltd.



Eun-Jin Kim received the B.S. degree in Computer Engineering from Chungbuk National University, Cheongju, Republic of Korea, in 2023, respectively. She is currently pursuing an integrated M.S. and Ph.D. degree

in Computer Engineering at Chungbuk National University, Cheongju, Republic of Korea. Her current research interests include mitigating prompt sensitivity in LLMs to improve robustness and malware detection using deep learning techniques.



Young-Seob Jeong received the M.S. and Ph.D. degrees in computer science from KAIST, Daejeon, South Korea, in 2012 and 2016, respectively. He is currently a Faculty Member with the Department of Computer

Engineering, Chungbuk National University, Cheongju, South Korea. His current research interests include malware detection using deep learning techniques, language models for lowresourced languages, healthcare systems for patients, and pharmaceuticals.