# COSE474-2024F: Final Project Proposal
# "CBAM과 FEM을 활용한 YOLO 기반 VisDrone 객체 탐지 성능 향상 "

**Youngmin Kim**

## 1. Introduction

본 연구에서는 YOLO 모델에 CBAM(Convolutional Block Attention Module)과 FEM(Feature Enhancement Module)을 통합하여 이러한 한계를 극복하고자 했습니다. CBAM은 중요 특징에 대한 가중치를 부여하고, FEM은 다양한 스케일의 특징을 효과적으로 학습할 수 있도록 설계되었습니다.

제안된 방법은 VisDrone 데이터셋에서 다음과 같은 성과를 보였습니다:

1. **모델 구조 개선**: YOLO에 CBAM과 FEM을 통합하여 특징 표현력 강화.
2. **최적화된 학습 전략**: 이미지 크기 조정, 반정밀도(Half-Precision) 학습, 데이터 증강 기술 적용.
3. **성능 평가**: mAP, Precision, Recall 등 주요 지표에서 성능 향상 확인.

실험 결과, 제안된 방법은 특히 소형 객체와 복잡한 장면에서 탐지 정확도를 크게 향상시켰으며, 드론 기반 객체 탐지의 새로운 가능성을 제시합니다.

## 2. Related Works

객체 검출 분야는 빠른 속도와 정확도를 동시에 추구하며 발전해 왔습니다. R-CNN 계열과 같은 이단계(two-stage) 방법은 높은 정확도를 보였으나 계산량이 많았고, 이후 YOLO나 SSD와 같은 단발(single-stage) 검출기들은 연산 효율성을 높여 실시간 검출을 가능하게 하였습니다. 최근에는 YOLOv8 등 더 가벼우면서도 정확도를 향상시킨 모델이 등장하며 실용성이 한층 강화되고 있습니다. 특징 맵의 효율적 활용을 위해 채널 및 공간 어텐션 기법이 활발히 연구되고 있습니다. 예를 들어 SENet은 채널 중요도를 재조정하여 다양한 비전 태스크에서 성능을 높였으며, CBAM(Convolutional Block Attention Module)은 채널과 공간 정보를 모두 활용하는 주의 메커니즘을 제안해 특징 표현력 향상에 기여하였습니다. 또한 ECA[9]와 같은 경량 어텐션 모듈도 높은 효율성을 보여주었고, DETR[10]나 Deformable DETR[11]과 같이 Transformer 기반 검출 방식은 전역적 관계 파악을 통해 새로운 가능성을 제시하고 있습니다. 그러나 경량성과 속도를 중요시하는 응용 환경에서는 여전히 CNN 기반 어텐션 모듈 도입이 선호되는 경향이 있습니다.

여기에 FPN[12], PANet[13], BiFPN[14] 등 멀티 스케일 특징 추출 기법이 적용되면서 다양한 크기의 객체에 대한 검출 성능이 향상되었습니다. 특히 VisDrone[15], UAVDT[16] 등 드론 영상에서는 복잡한 배경, 다양한 스케일, 작은 객체가 등장하여 더욱 강건한 특징 표현이 필요합니다. 이러한 상황에서 어텐션 모듈과 경량 특화 블록을 활용한 네트워크 구조 개선이 효과적인 해결책으로 제안되고 있습니다. 본 연구는 이러한 흐름을 반영하여, YOLO 기반 모델에 경량화된
Feature Enhancement Module(FEM)을 적용하고, CBAM과 같은 어텐션 기법을 결합하여 드론 영상 환경에서의 객체 검출 성능 향상을 달성하는 방법을 제안합니다.

## 3. Methods

### 1. 데이터 전처리

VisDrone 데이터셋은 드론 촬영 영상으로부터 다양한 객체(보행자, 차량, 자전거 등)를 포함하고 있으며, 각 이미지에 대해 경계상자(annotation) 정보가 제공됩니다. 본 연구에서는 YOLO 형식의 라벨로 변환하기 위해 별도의 전처리 스크립트를 개발하였습니다. 해당 스크립트는 다음 과정을 거칩니다.

1. **폴더 구조 생성**:
   `dataset/train/images`, `dataset/train/labels`, `dataset/val/images`, `dataset/val/labels` 등의 디렉토리 구조를 만들어 YOLOv8에서 요구하는 표준적인 폴더 구조를 갖추었습니다.
2. **어노테이션 변환**:
   VisDrone의 주어진 `.txt` 어노테이션 파일을 읽어, 각 객체의 좌표를 YOLO 형식(정규화된 중심 좌표와 폭 · 높이 정보)으로 변환하였습니다. 이 과정에서 필요 없는 클래스나 에러 라인을 무시하고, 채널 수와 크기를 고려하여 정규화 연산을 수행하였습니다. 또한 멀티프로세싱을 통해 변환 속도를 개선하였습니다.
3. **이미지 복사 및 정리**:
   변환된 어노테이션에 대응하는 이미지 파일을 목적지 디렉토리로 병렬 처리(멀티프로세싱)하여 복사함

으로써 대용량 데이터 처리 시간을 단축하였습니다.

이와 같은 전처리 과정을 통해 VisDrone 데이터셋을 YOLO 호환 형식으로 효율적으로 준비하였으며, 이는 본 연구에서 제안하는 모델 개선 방법을 안정적으로 적용하고 평가할 수 있는 기반이 됩니다.

### 2. 모델 구조 개선

전처리 완료된 데이터셋을 활용하여, 본 연구는 YOLOv8n과 같은 경량 Backbone을 기본으로 하여 Neck 구간에 Feature Enhancement Module(FEM)과 CBAM(Convolutional Block Attention Module)을 적용하였습니다.

1. **Feature Enhancement Module(FEM)**:

   FEM은 입력 특징 맵의 채널 수를 줄이고(group convolution을 사용) 3x3, 5x5 커널을 통해 다중 스케일 특징을 추출한 뒤 병합합니다. 이를 통해 연산량 증가를 최소화하면서 풍부한 공간적 정보를 얻을 수 있습니다. 이후 Batch Normalization과 CBAM을 적용하여 특징 맵을 한 번 더 정제합니다.

2. **CBAM 적용**:

   CBAM은 채널 및 공간 어텐션을 모두 고려하는 모듈로, 중요한 채널 및 영역에 집중하도록 유도합니다. 이를 통해 모델이 객체 위치와 형태 정보를 더 정확히 반영할 수 있으며, 특히 드론 영상에서 빈번히 나타나는 작은 객체 검출에 유리한 특성을 갖습니다.

3. **최종 Detection Head**:

   FEM과 CBAM을 통해 강화된 특징 맵은 YOLOv8 Detection Head에 전달되어 최종적으로 객체 위치와 범주를 예측합니다. 이는 기존 YOLO 모델 대비 더욱 정교한 특징 표현을 기반으로 동작하므로, mAP, Recall, Precision 등 평가 지표에서 향상된 성능을 나타낼 수 있습니다.

**Algorithm 1** VisDrone to YOLO Format Conversion

VisDrone dataset $(A_{train}, A_{val}, I_{train}, I_{val})$ YOLO-formatted dataset Initialize directories: `dataset/{train,val}/{images,labels}` Define $C = \{1 \mapsto 0, ..., 10 \mapsto 9\}$ Class mapping $S \in \{train, val\}$ $a \in A_S$ Process annotations $(x, y, w, h, k)$ in $a$ $k \in C$ $x_c \leftarrow (x + w/2)/W$ Normalize coordinates $y_c \leftarrow (y + h/2)/H$ $w' \leftarrow w/W$, $h' \leftarrow h/H$ Write $(C[k], x_c, y_c, w', h')$ to output Copy image from $I_S$ in parallel

**Algorithm 2** Enhanced YOLO Training with FEM

Base weights $W$, Dataset $D$, Epochs $E$ Trained weights $W_{final}$ Initialize model with $W$ Add FEM to model neck Enable mixed precision training $e = 1$ to $E$ batch $(X, Y)$ in $D$ Apply augmentations to X Forward pass through enhanced model Compute loss and backpropagate Update learning rate Evaluate on validation set **return** $W_{final}$

## 4. Experiments

### 1. 데이터셋 및 전처리

VisDrone 데이터셋은 복잡한 도시 환경에서 촬영한 드론 영상을 포함하고 있으며, 다양한 크기와 형태의 객체들이 등장합니다. 본 연구에서는 YOLO 형식 어노테이션을 사용하기 위해 자체 개발한 전처리 스크립트를 통해 VisDrone의 주어진 어노테이션과 이미지를 YOLOv8 표준 디렉토리 구조(`train/images`, `train/labels`, `val/images`, `val/labels`)로 변환하였습니다. 이 과정에서 멀티프로세싱을 활용해 처리 속도를 높였고, 추후 재현을 위해 전처리 완료된 데이터셋을 압축 백업한 뒤 필요할 때마다 병렬 압축 해제를 통해 신속히 복원할 수 있도록 구현하였습니다.

### 2. 모델 구성 및 학습 설정

기본 모델(Baseline)은 YOLOv8n을 사용하였습니다. 제안한 방법에서는 이 기본 모델의 Neck 부분에 경량화된 Feature Enhancement Module(FEM)과 CBAM(Convolutional Block Attention Module)을 적용하였습니다. FEM은 다중 스케일 특징 추출을 통해 풍부한 공간 정보를 확보하고, 채널 축소와 그룹 컨볼루션을 통해 연산량 증가를 최소화합니다. CBAM은 채널 및 공간 어텐션을 통해 중요한 특징에 집중할 수 있도록 하여 최종 특징 맵의 표현력을 극대화합니다.

학습 과정에서는 half precision(mixed precision) 모드를 사용하여 GPU 메모리 사용량과 연산 시간을 절감하였으며, Mosaic 및 Mixup 등의 Augmentation을 통해 데이터 다양성을 확보하였습니다. 이미지 해상도는 1024로 설정하였으며, Batch Size는 16으로 하여 실험을 진행하였습니다. Epoch 수는 30으로 설정하였으나 필요 시 더 많은 Epoch로 연장하여 FEM+CBAM 조합의 장기적 학습 효과를 검증할 수 있습니다.

### 3. 성능 평가 지표 및 비교 실험

성능 평가는 mAP(Mean Average Precision), Precision, Recall 등을 활용하였습니다. 특히 mAP@[0.5:0.95]는 다양한 IoU 임계값에서의 전반적인 검출 성능을 종합적으로 확인할 수 있어, 제안한 기법(FEM+CBAM)과 Baseline(YOLOv8n) 간의 성능 차이를 명확히 보여줍니다.

비교 대상은 다음 두 가지입니다.

- **Baseline**: FEM 및 CBAM 적용 전의 순수 YOLOv8n

모델

- **FEM+CBAM**: FEM과 CBAM을 모두 적용한 제안 모델

이러한 단순 비교를 통해 제안한 모듈 추가가 실제 성능 개선으로 이어지는지 명확히 확인할 수 있습니다.

**4. 실험 결과** 본 절에서는 기본 YOLOv8n 모델(Baseline)과 제안한 FEM+CBAM 적용 모델의 학습 과정 및 최종 성능을 비교한 결과를 제시합니다. 이를 위해 총 4개의 결과 그래프를 활용하였습니다. 두 모델 각각에 대해 학습 및 검증 손실 및 지표 변화를 담은 학습 곡선과, 클래스별 Precision-Recall 곡선을 제시하여 성능 변화를 종합적으로 확인하였습니다.

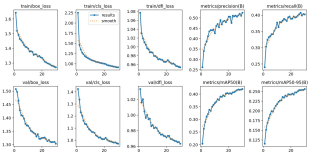

*Figure 1.* Precision-Recall Curve before tuning

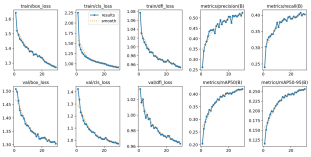

*Figure 2.* Precision-Recall Curve after tuning

figure 1과 figure 2는 각각 Baseline 모델과 FEM+CBAM 적용 모델에 대한 최종 Precision-Recall 곡선을 나타내며, mAP@0.5를 비롯한 전반적인 검출 성능을 확인할 수 있습니다. Baseline 모델 (figure 1): Baseline(YOLOv8n) 모델의 Precision-Recall 곡선에서는 복잡한 드론 영상 환경으로 인해 일부 클래스에서 매우 낮은 Precision 또는 Recall을 보이는 경향이 있었습니다. 전체 mAP@0.5는 약 0.301로 측정되어, VisDrone 데이터셋의 난이도를 반영합니다. FEM+CBAM 적용 모델 (figure 2): FEM과 CBAM을 적용한 제안 모델은 동일한 학습 조건 하에서 전반적인 Precision-Recall 특성이 개선되었습니다. 특히 car, bus 등 상대적으로 뚜렷한 형태를 가진 객체 클래스에서 Precision 및 Recall이 명확하게 상승하였으며, 전체 mAP@0.5는 0.406으로 약 0.105 포인트 향상하였습니다. 이는 제안 모듈들이 특징 맵 강화 및 주의 메커니즘을 통해 모델이 중요한 영역과 채널에 더욱 집중하도록 유도한 결과로 볼 수 있습니다. 이로써 Precision-Recall 곡선을 통한 일차적 비교에서 FEM+CBAM 적용 모델이 Baseline 대비 명확한 성능 개선을 달성했음을 확인할 수 있습니다.

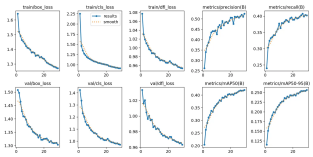

*Figure 3.* Learning curves before tuning

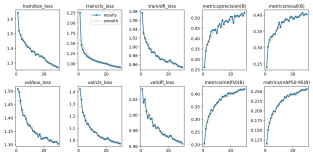

*Figure 4.* Learning curves after tuning

이후 figure 3과 figure 4는 각각 Baseline 모델과 FEM+CBAM 적용 모델의 학습 과정을 나타내는 손실(손실 값: Box, Class, DFL) 및 mAP, Precision, Recall 변화 곡선을 제시합니다. 이 곡선을 통해 다음과 같은 점들을 확인할 수 있습니다.Baseline 대비 FEM+CBAM 적용 모델은 초기 Epoch부터 손실값이 안정적으로 감소하며, mAP 등 성능 지표가 더 빠르게 상승하는 경향을 보입니다.성능 향상이 Epoch 진행에 따라 누적되는 모습을 통해 제안한 모듈 도입이 학습 안정성과 최종 성능 모두에 긍정적인 기여를 한다는 점을 재확인할 수 있습니다.

## 5. Conclusion

본 연구에서는 드론 영상 객체 검출 분야에서 널리 활용되는 YOLOv8n 모델을 기반으로, Feature Enhancement Module(FEM)과 Convolutional Block Attention Module(CBAM)을 결합하여 성능을 향상시키는 방법을 제안하였습니다. VisDrone 데이터셋을 활용한 실험 결과, 제안한 모델은 기본 YOLOv8n 대비 mAP@0.5 지표에서 유의미한 상승(약 0.105 포인트 개선)을 달성하였으며, 특정 클래스(예: 차량, 버스)에서 Precision과 Recall 모두를 고르게 향상시켰습니다. 이러한 성능 개선은 FEM을 통한 다중 스케일 특징 맵 강화와 CBAM을 통한 채널 · 공간 어텐션 기법이 어우러져, 네트워크가 중요한 특징에 더욱 효과적으로 주목할 수 있었기 때문입니다. 또한 Half Precision 학습, Augmentation 적용, 경량 Backbone 사용 등을 통해 연산 효율성도 어느 정도 유지하면서 성능을 높일 수 있음을 확인하였습니다. 본 연구는 드론 영상 환경에서의 객체 검출 성능 개선을 위한 하나의 방향을 제시하였으며, 향후 연구에서는 Transformer 기반 구조나 NAS(Neural Architecture Search) 기법을 결합하거나, 추가적인 Attention 모듈 적용, 다양한 Backbone 네트워크 검증 등을 통해 더욱 폭넓은 성능 향상을 모색할 수 있을 것입니다. 이를 통해 점차 복잡해지는 드론 환경 감지 과업에 보다 안정적이고 효율적인 솔루션을 제공할 수 있으리라 기대합니다.

## 6. Reference

## References

[1] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788. https://doi.org/10.1109/CVPR.2016.91

[2] Bochkovskiy, A., Wang, C., & Liao, H. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint arXiv:2004.10934*. https://arxiv.org/abs/2004.10934

[3] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

[4] Zhu, P., Wen, L., Bian, X., Ling, H., & Hu, Q. (2018). Vision Meets Drone Object Detection in Image Challenge Results. *Lecture Notes in Computer Science (LNCS)*, 263–279. https://doi.org/10.1007/978-3-030-11009-3_38

[5] Du, D., et al. (2019). The VisDrone Benchmark: The Vision Meets Drone Object Detection and Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(8), 2021–2039. https://doi.org/10.1109/TPAMI.2020.3007741