https://doi.org/10.7840/kics.2023.48.1.1

고해상도 특징맵 추출기와 Attention 기법을 사용한 항공 영상 특화 객체 검출기

김 해 문•, 안 종 식*, 이 태 영*, 최 병 인°

The Object Detector for Aerial Image Using High Resolution Feature Extractor and Attention Module

HaeMoon Kim•, JongSik Ahn*, Tae-Young Lee*, Byungin Choi°

요 약

YOLOv5와 같은 일반적인 객체 검출기는 일상 환경에서 획득 가능한 이미지로 구성된 COCO 데이터세트에서는 큰 성과를 보였지만 항공 영상에서는 높은 성능을 도출하지 못하였다. 이는 항공 영상의 특성을 반영하지 못한 네트워크 구조 때문이다. 본 논문에서는 항공 영상의 특성을 분석하고 이를 통해 항공 영상에 특화된 객체 검출기를 제안한다. 항공 영상은 첫째, 이미지 내 객체들의 크기가 소형이고 밀집도가 높은 양상을 보인다. 둘째, 넓은 시야각으로 인해 객체 정보보다 배경 정보가 많이 존재하며 이로 인해 복잡한 배경이 많아 밀집도가 높은 객체의 구분이 어렵다. 이러한 특성을 반영하여 높은 해상도를 유지하며 특징을 추출하는 SB 네트워크와 객체와 배경의 분리를 위한 Triplet Attention을 사용한 TA 네트워크를 제안한다. 제안하는 네트워크인 YOLOv5l-TA는 YOLOv5l 네트워크 대비 $mAP_{0.5}$ 성능이 11.2% 향상하였고 객체의 크기별로 계산한 mAP_{vt}, mAP_t, mAP_s, mAP_m 성능이 각각 280%, 55%, 36.1%, 4.8% 증가하였다.

키워드 : 항공 영상, 고해상도 특징맵, Attention 네트워크, 소형 객체 탐지기
Key Words : Aerial Image, High Resolution Feature Map, Attention Network, Tiny Object Detection

ABSTRACT

Object detectors, such as YOLOv5, achieve high performance on datasets that consist of objects in everday scenes, like the COCO dataset. However, it shows poor performance in aerial images because the detectors did not consider the size of the objects. First, the aerial images contain very tiny objects and these objects are densely located in a image. Second, because of wide FOV, most of images has a lot of complex background information. It makes object detector very difficult to recognize object and background. In this paper, we propose an object detector that focuses on tiny objects with high resolution feature maps and attention network. We densely located in a image. Second, because of wide FOV, most of images has a lot of complex background information. It makes object detector very difficult to recognize object and background. In this paper, we propose an object detector that focuses on tiny objects with high resolution feature maps and attention network. We design SB network which is feature extractor through high resolution feature map. Also

◆ First Author : Hanwha Systems, haemoon1205@hanwha.com, 정회원
° Corresponding Author : Hanwha Systems, byungin.choi@hanwha.com, 정회원
* Hanwha Systems, jongsik.ahn@hanwha.com; ty.lee@hanwha.com
논문번호 : 202211-281-A-RU, Received November 18, 2022; Revised November 29, 2022; Accepted December 5, 2022

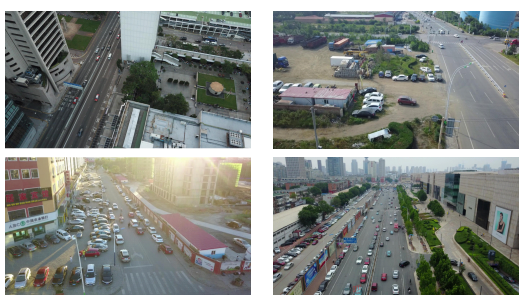
we adopted Triplet Attention to TA network for distinguish between objects and background. The proposed YOLOv5l-TA network and achieves $mAP_{0.5}$ 11.2% higher than YOLOv5l baseline network and 280%, 55%, 36.1%, 4.8% in mAP_{vt}, mAP_t, mAP_s, mAP_m metrics.

Ⅰ. 서 론

객체 검출(Object Detection)은 영상 내 다수의 객체 위치를 경계 상자(Bounding Box)로 표현하고 해당 경계 상자 내 객체의 종류(Class)를 분류하는 컴퓨터 영상 처리 분야 중 하나이다. 현재 객체 검출 알고리즘들은 다양한 산업 분야에서 활용되고 있으며 특히 고정된 감시장비 또는 드론(Drone)에서 촬영된 항공 영상을 사용하는 감시 시스템 분야에서 활용도가 높다. 감시 시스템 분야의 객체 검출은 표적 객체 미탐지로 인해 발생할 수 있는 감시 공백을 방지하기 위하여 높은 성능의 객체 검출 알고리즘이 요구된다. 이에 따라 2015년 이후 인간의 인지능력을 뛰어넘었다고 평가받는[1] 인공 신경망 기반 딥러닝(Deep Learning) 알고리즘을 활용한 객체 검출 알고리즘이 최근 감시 시스템 분야에서 사용되고 있다.

최근 객체 검출기는 인공 신경망을 사용한 VGG (Very Deep Convolutional Networks)[2], ResNet (Residual Neural Network)[1], EfficientNet (Efficient Network)[3] 등의 CNN (Convolutional Neural Networks) 기반 특징 추출 네트워크를 사용하여 COCO (Microsoft Common Objects in Context) 데이터세트를 대상으로 높은 검출정확도 성능을 달성하였다. 이러한 객체 검출 방법에는 크게 2-Stage 방식과 1-Stage 방식이 존재한다. 2-Stage 방식의 객체 검출기는 대표적으로 R-CNN (Region Based Convolutional Neural Networks) 계열[4-6]의 네트워크가 있으며 다수의 객체 후보 영역(Region Proposal)을 선정하고 해당 영역에 대하여 객체 종류 분류 및 영역 보정을 수행한다. 이러한 네트워크는 높은 객체 검출 정확도를 달성하지만, 네트워크 계산 비용(Computation Cost)이 높아 1-Stage 방식에 비해 추론 속도가 느리다. 반면 1-Stage 객체 검출기는 대표적으로 YOLO (You Only Look Once) 계열[7-10]이 있다. 1-Stage 객체 검출기는 객체 검출 문제를 하나의 회귀(Regression) 문제로 정의하여 종단간(End-to-End) 네트워크로 구성해 객체의 위치 추정과 클래스 분류 수행한다. 1-Stage 객체 검출기는 계산 비용이 2-Stage에 비해 적고 검출 성능이 낮았으나 최근 연구들로 인해 현재는 2-Stage와 비슷한 수준의 검출 성능을 보이고 있고 빠른 추론 속도로 인해 실시간 영상에서 활용될 수 있어 산업분야에 활발하게 사용되고 있다. CSPDarkNet (Cross Stage Partial Dark Network) 특징 추출 네트워크는 YOLOv5[11]에 사용되는 특징 추출 네트워크로서 YOLOv3의 특징 추출 네트워크 DarkNet에 CSPNet (Cross Stage Partial Network)[12] 네트워크 경량화 기법을 적용해 파라미터 개수를 줄이고 효율적으로 객체 특징을 추출하여 검출정확도 향상에 기여하였다. 하지만 이러한 검출정확도 향상은 일반적인 생활 반경 내에서 구성된 데이터세트에는 성과를 거두기 어렵다. 이는 항공 영상의 특성을 고려하지 못하였기 때문이다.

항공 영상 데이터세트의 객체 크기에 따른 분석을 위하여 W_{gt}는 정답 경계 상자의 가로 길이, H_{gt}는 세로 길이일 때 표 1과 같이 객체 크기에 따라 객체 크기 종류를 정의하였다. 그림 2는 드론에서 촬영된 대표적인 항공 영상 데이터세트, VisDrone-DET2019 Challenge[13] 데이터세트에서 객체 종류별로 객체 크기를 분석한 결과이며 전체 객체의 81.59%의 객체들이 Small 이하에 해당하는 소형 객체인 것을 확인하였다. 이처럼 항공 영상은 그림 1-a와 같이 대부분의 객체가 매우 작은 Pixel영역으로 표현되어 전체 영상 크기 대비 매우 작은 영역에 위치하고 다수 객체들이 작은 영역에 조밀하게 모여 있는 경우가 빈번하다. 두 번째로 그림 1-b와 같이 드론의 비행 고도, 영상 촬영 각도에 따라 객체는 다양한 크기와 형상을 가진다. 또한 FOV (Field Of View)가 넓어 다량의 배경 정보가



(a) A tiny and dense objects (b) Object of various sizes on a complex background

그림 1. 항공 영상의 특성
Fig. 1. The properties of drone view images

표 1. 객체 크기별 분류
Table 1. The categories by object size

The classification by object size	Objects size
Very Tiny	$W_{gt} \times H_{gt} < 8^2$
Tiny	$8^2 \leq W_{gt} \times H_{gt} < 16^2$
Small	$16^2 \leq W_{gt} \times H_{gt} < 32^2$
Medium	$32^2 \leq W_{gt} \times H_{gt} < 96^2$
Large	$96^2 \leq W_{gt} \times H_{gt}$

영상 내 존재하여 촬영된 영상 다수가 복잡한 배경 정보를 가진다.

항공 영상에서의 객체들은 적은 Pixel 개수로 표현되어 배경과의 구분이 모호하고 공간(Spatial) 정보가 부족하여 객체를 식별하기 어렵다. 이러한 특징 때문에 딥러닝 기반의 특징 추출 네트워크에 입력 시 일반적인 객체 검출기처럼 동작할 경우 각 Stage의 Down-sampling 계층을 지나 특징맵의 해상도가 낮아지며 공간 정보가 소실되어 검출정확도의 하락으로 이어진다. 그림 3에서는 항공 영상이 YOLOv5의 특징 추출 네트워크인 CSPDarkNet을 통해 특징이 추출되는 과정을 시각화하였다. 특징 추출 네트워크는 각 Stage의 입력 특징맵이 통과한 이전 Stage 개수에 따라 Shallow, Middle, Deep 특징맵으로 분류되고 Shallow 특징맵은 높은 해상도를 유지하여 객체와 배경의 구분이 가능하며 공간 정보가 풍부하다는 특징이 있다. 반면 Deep 특징맵은 해상도가 낮지만 의미론적(Semantic) 정보가 풍부하여 개별 객체 클래스 분류에 유리하다. Middle 특징맵은 앞서 언급한 2개의 특징맵의 특성이 고르게 나타난다. 따라서 CSPDarkNet 특징 추출 네트워크에서 소형 객체의 공간 정보는 Shallow 특징맵에서 가장 풍부하고 Deep

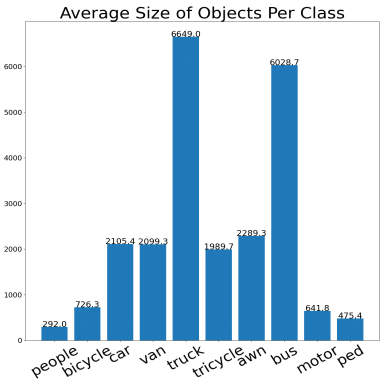
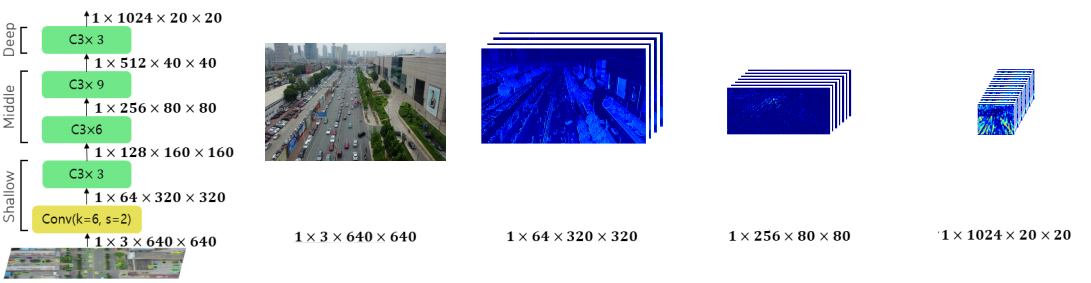


그림 2. VisDrone-DET2019 Challenge 데이터세트 객체 크기 분석
Fig. 2. The analysis of object size in VisDrone-DET2019 dataset

특징맵으로 진행할수록 의미론적 정보가 증가하는 것에 반해 공간 정보는 감소한다. 특히 YOLOv5 네트워크의 Deep 특징맵에서 소형 객체의 절대적 크기는 입력 영상 대비 1/32 으로 그림 4에서 확인할 수 있듯 객체의 크기가 32x64 Pixel인 객체는 Stage 5에서 1x2로 대부분의 공간 정보가 소실된다.

이러한 특징을 반영하여 본 논문에서는 YOLOv5



(a) CSPDarkNet (b) Input image (c) Shallow feature map (d) Middle feature map (e) Deep feature map

그림 3. YOLOv5l 특징 추출 네트워크의 특징맵 분류 및 시각화
Fig. 3. The Visualization of YOLOv5l backbone network feature map

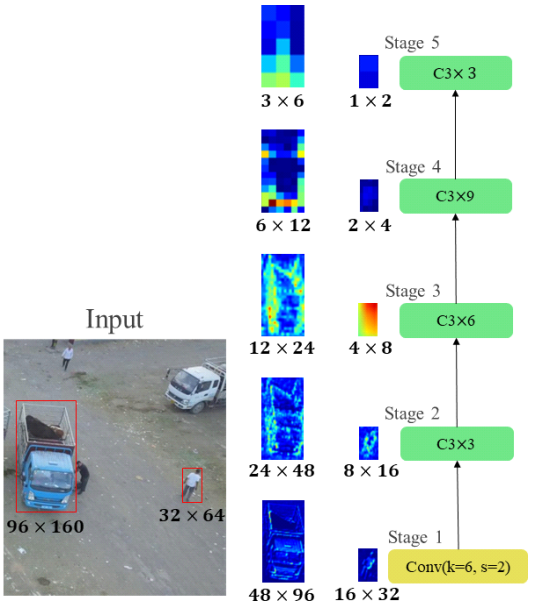


그림 4. CSPDarkNet Stage 진행에 따른 객체의 크기 및 특징맵 변화
Fig. 4. The changes of object size and feature map in CSPDarkNet

구조를 중심으로 항공 영상에 특화된 객체 검출 네트워크를 제안한다. 제안하는 네트워크는 작고 밀집된 객체의 특징을 지니는 항공 영상의 특성을 반영하기 위하여 기존 네트워크를 항공 객체 검출에 적합하도록 커널 개수, 채널, 깊이 등의 네트워크 구조 스케일링(Scaling)을 진행하였다. 공간 정보가 소실되어 작은 객체 탐지에 불리한 Deep 특징맵이 출력되는 네트워크 계층을 제거하였으며 Shallow 특징맵을 중심으로 소형 객체 검출을 하였다. 두번째, 객체와 복잡한 배경과의 분리를 위하여 Triplet Attention Network[14]를 객체 검출기의 FPN (Feature Pyramid Network)[15] 네트워크 구조에 결합하였다. 제안하는 방법으로 소형 객체 검출정확도를 효과적으로 향상하였으며 VisDrone-DET2019 Challenge 데이터세트에서 YOLOv5l 네트워크 대비 $mAP_{0.5}$ 성능이 11.2% 향상되는 것을 확인하였다.

Ⅱ. 본 론

2.1 YOLOv5l

YOLOv5 네트워크는 YOLOv5l (YOLOv5 Large) 네트워크를 바탕으로 네트워크 구조의 채널, 깊이(Depth) 스케일링을 통해 Nano, Small, Medium, Large, Xlarge 네트워크로 분류하였다. 그림 5는 YOLOv5 네트워크 중 베이스라인 네트워크 YOLOv5l 네트워크의 구조이다. 특징 추출 네트워크는 CSPDarkNet을 사용하였으며 여러 크기의 객체들에 대응하고 특징 추출 능력 향상을 위하여 일반적으로 사용하는 FPN 네트워크 대신 PAN (Path Aggregation Network)[16] 을 적용해 Top-Down 및 Bottom-up 경로에서 상위, 하위 Stage 특징맵을 결합하였다. 최종적으로 객체 탐지 및 분류를 위하여 PAN 네트워크 구조 내 3개의 Stage에 Head라고 불리는 객체 탐지 및 분류 네트워크를 구성하여 객체 위치 추정과 분류를 진행하였다.

YOLOv5l 네트워크에서 특징 추출 네트워크는 그림 5처럼 5개의 Stage로 이루어져 있다. 각 Stage 통과 시 입력 이미지의 해상도는 절반이 되므로 가로, 세로 32 Pixel 이하에 해당하는 Small 객체들은 Stage 5에서 1 Pixel 크기로 감소한다. 이는 객체를 구분하기 위한 공간정보가 대부분 소실된 상태이고 따라서

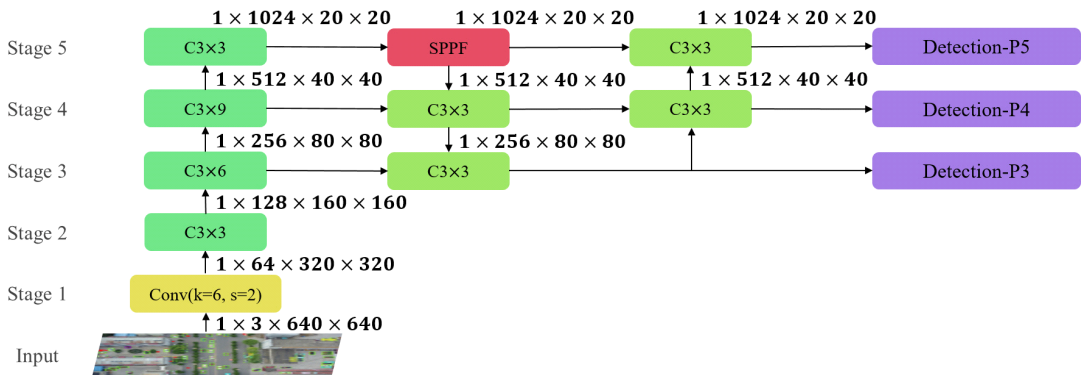


그림 5. YOLOv5l 네트워크 구조
Fig. 5. The architecture of YOLOv5l Network

YOLOv5l 구조는 대부분의 객체가 Small 이하의 객체들로 구성된 항공 영상 데이터세트에 적합하지 않은 구조이다.

2.2 YOLOv5l-SB

특징 추출 네트워크의 특징 추출 과정에서 높은 해상도로 특징을 추출할 경우 요구되는 계산량은 많아지고 이에 따라 네트워크 추론 속도 성능이 감소하게 된다. 때문에 기존 YOLOv5의 특징 추출 네트워크 CSPDarkNet은 매 Stage 진행 시 특징맵이 Down-sampling 되어 Stage 별로 해상도를 낮추었고 이러한 구조를 통해 연산량을 줄였다. 하지만 소형 객체 탐지를 위해서는 공간 정보가 소실되지 않은 높은 해상도의 특징맵이 필요하다. 따라서 네트워크 연산량 증가의 최소화와 해상도 감소에 따른 공간 정보 손실 최소화를 목표하는 특징맵 분석 바탕의 Stage 개수 선정이 필요하다. 그림 4와 같은 특징맵 분석 결과, VisDrone 데이터세트 전체 객체의 81.59%에 해당하는 소형 객체는 Stage 4, 5 에서 공간 정보 대부분이 소실되었다. 따라서 네트워크 연산량, 공간 정보 손실이 최소화된 Stage 개수를 최대 3으로 설정하여 네트워크 YOLOv5l의 구조에서 특징 추출 네트워크 스케일링을 진행하였다. 본 논문에서는 먼저 Stage 1, 2 특징맵만을 사용하는 단일 스케일 특징맵만을 객체 검출에 사용하는 YOLOv5l-SB (Spatial Backbone) 네트워크를 설계하였다.

$$(W_F, H_F) = (W_I \times \frac{1}{2^i}, H_I \times \frac{1}{2^i}), (i = 1,2,3..5) \quad (1)$$

기존 YOLOv5l의 네트워크에서 32x32 Pixel 크기의 객체가 입력될 경우, Stage 1에서 5까지 객체의 크기는 식 1에 따라 16x16, 8x8, 4x4, 2x2, 1x1 Pixel이 된다. 이때, 수식 1의 W_F와 H_F는 특징맵의 해상도, W_I와 H_I는 입력 이미지의 해상도를 의미하며 i는 i번째 Stage를 의미한다. 최소한의 공간 정보를 포함하고 객체의 유형을 분류할 수 있는 한계를 8x8 Pixel로 선정하여 Stage 1, 2만을 이용해 그림 6과 같이 제안하는 Spatial Backbone을 설계하였다. 추가적으로 공간 정보를 더욱 확보하기 위해 Stage 2의 컨볼루션 연산에서 커널 크기와 Stride를 기존 6, 2에서 7, 1로 조절하여 고해상도의 특징맵을 유지하는 SB + HR (High Resolution) 기법 네트워크와 컨볼루션 커널 개수와 채널 개수를 늘려 네트워크 파라미터 개수를 크게 한 SB + CH (Channel) 기법 네트워크 또한 설계하였다. 제안하는 YOLOvl-SB 네트워크 평가 결과 Very Tiny 객체에 대한 mAP 성능이 YOLOv5l 대비 최소 2.4배, 최대 3.6배로 크게 상승하는 것을 확인하였다. 하지만 Very Tiny 객체를 제외한 나머지 객체 크기에 대한 mAP 성능은 감소하는 경향을 보였다.

2.3 YOLOv5l-TA

제안하는 YOLOv5l-SB 네트워크에서 Very Tiny를 제외한 크기의 성능감소는 다중 스케일이 아닌 단일 스케일에서 객체를 검출한 네트워크 구조의 한계 때문이다. 더불어 하나의 객체 분류 및 검출 네트워크가 연결되어 다양한 크기의 객체에 대응하지 못하였다. 따라서 다양한 크기의 객체에 대응하기 위하여 YOLOv5l-SB의 네트워크에 Stage를 추가한 YOLOv5l-TA (Triplet Attention) 네트워크를 제안하였다.

그림 7과 같이 기존 YOLOv5l-SB의 특징 추출 네트워크에 1개의 Stage를 추가하였고 추가한 Stage 계층에 객체 분류 및 검출 네트워크를 부착하여 다양한 크기의 객체에 대응하고자 하였다. 또한 PAN 구조를 적용하여 Top-Down 경로에서 다중 해상도 특징맵을 결합하여, 소형 객체의 특징 추출 능력을 보완하였다. 추가적으로 공간 정보는 Stage 2의 높은 해상도의 특징맵에 비해 부족하지만 의미론적 정보가 풍부한 Stage 3의 특징맵 정보를 Triplet Attention 통한 특징 결합 경로를 설계하여 객체 및 배경 구분을 용이한 고해상도 특징맵과 의미론적 정보가 풍부한 Deep 특징맵을 결합하도록 설계하였다.

Attention 네트워크는 특징맵의 채널, 공간 등 특정

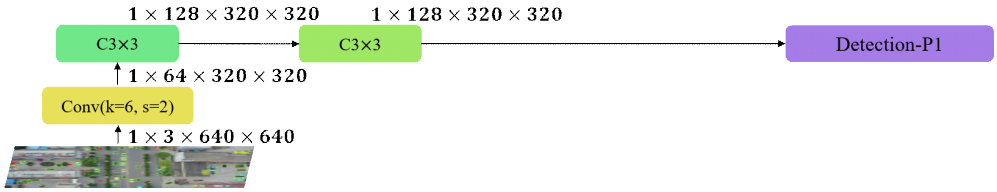


그림 6. YOLOv5l-SB 네트워크 구조
Fig. 6. The architecture of YOLOv5l-SB Network

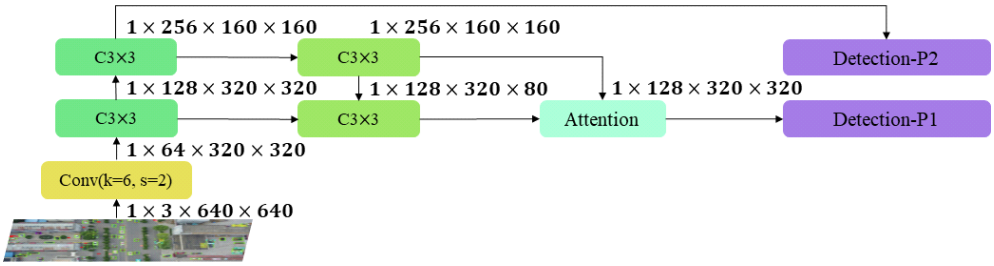


그림 7. YOLOv5l-TA 네트워크 구조
Fig. 7. The architecture of YOLOv5l-TA Network

차원 내 특정 부분의 정보를 강조하는 네트워크이며 가중치를 스칼라 곱하는 방법으로 입력 특징맵을 특정 부분을 강조한다. 객체 검출기 네트워크에서 공간 방향 정보는 위치 정보를 의미하며 채널 방향 정보는 클래스 정보를 의미한다. CBAM (Convolutional Block Attention Module)[17] 네트워크는 MLP (Multi Layer Perceptron)와 컨볼루션 연산을 사용하여 적은 네트워크 파라미터 개수로 채널, 공간 방향 특징을 효율적으로 강조하였다. 본 논문에서 적용한 Triplet Attention 네트워크는 그림 8과 같이 3개의 경로를 나눠 특징을 강조한다. 2개의 경로에서는 채널 방향 정보를 강조하고 하나의 경로에서는 CBAM 네트워크와 동일한 방법으로 공간 방향 정보를 강조한다. 2개의 경로에서는 입력 특징맵 Tensor를 높이, 너비 방향으로 회전 후 Z-풀링(Zeroth Pool) 연산을 사용하여 차원을 줄이고 컨볼루션과 시그모이드 연산을 사용하여 특징맵 가중치를 계산한다. 최종적으로 입력 특징맵에 가중치를 결합하며 다시 기존의 입력 특징 맵으로 높이, 너비 방향으로 회전한다. 이러한 Triplet Attention은 객체의 크기가 작고 밀집도가 높으며 객체와 배경

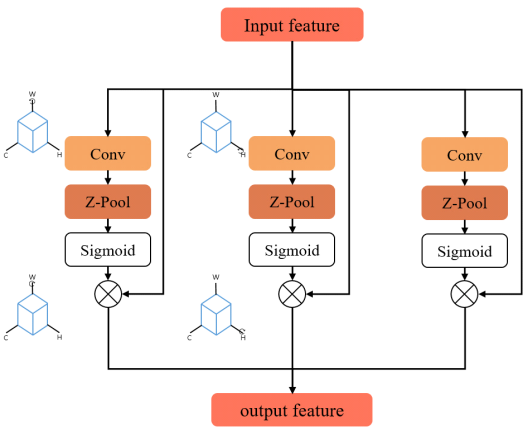


그림 8. Triplet Attention 네트워크 구조
Fig. 8. The architecture of Triplet Attention Network

구분이 어려운 항공 영상에서 효과적으로 객체의 특징을 강조하여 제안하는 YOLOv5l-TA 네트워크가 YOLOv5l 대비 11.2%의 $mAP_{0.5}$ 성능 향상되는 것을 확인하였다.

Ⅲ. 실험 결과

3.1 실험 환경 및 평가지표

본 논문에서 제안하는 네트워크는 Pytorch 프레임워크를 사용하여 설계하였다. 네트워크 학습에는 Nvidia GeForce RTX 2080ti GPU 2개를 사용하며 평가에는 Nvidia GeForce RTX 2080ti GPU 1개를 사용하였다. SGD(Stochastic Gradient Descent) 최적화기법을 사용하고 네트워크 학습률(Leaning Rate)은 0.01으로 초기화하며 Lambda LR 스케줄러(Scheduler)를 사용하여 300 Epoch 동안 학습을 진행하였다. 네트워크 입력 이미지는 Mosaic 데이터 증강 기법을 적용하고 Mosaic 이미지를 구성하는 개별 이미지는 Rotation, Shear, Perspective, HSV 채널 변환, 히스토그램 평활화 기법을 적용하였다. YOLOv5l-SB 네트워크 학습에는 COCO 데이터세트에서 사전에 학습된 YOLOv5l 네트워크 가중치로 전이 학습(Transfer Learning)을 진행하고 YOLOv5l-TA 네트워크 학습에는 사전 학습된 YOLOv5l-SB 네트워크 가중치로 전이 학습을 진행하였다.

$$Precision = \frac{TP}{TP+FP} \quad Recall = \frac{TP}{TP+FN}$$
$$AP = \int_0^1 Precision(Recall)dr \tag{2}$$

본 논문에서 제안하는 네트워크의 검출정확도 성능의 평가지표에는 개별 IoU (Intersection over Union)의 임계값에 대한 mAP (mean Average Precision)를 사용하였다. mAP는 정밀도(Precision)와 재현율

(Recall) 그래프의 개별 클래스 면적을 의미하며 식 2와 같다. 객체 크기별 검출정확도 성능 분석을 위해 표 1과 같은 기준으로 mAP를 mAP_{vt}부터 mAP_l까지 분할하였다. 제안하는 네트워크의 크기, 계산 비용, 추론 속도 성능의 평가지표에는 Param (Parameters), FLOPs (FLoating point Operations Per Second), FPS (Frames Per Second)을 사용하였다.

3.2 제안하는 네트워크의 성능 평가

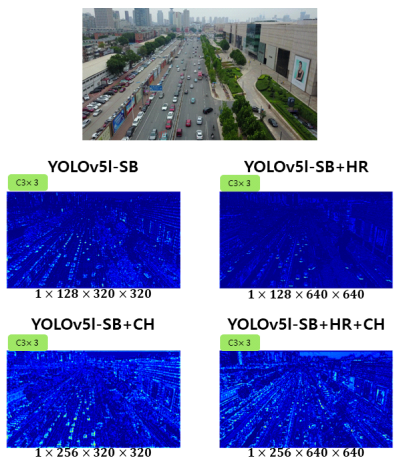
표 2에서 확인할 수 있듯 제안하는 YOLOv5-SB 네트워크를 항공 영상 데이터세트인 VisDrone-DET2019 Challenge에서 평가하였을 때 YOLOv5l 대비 mAP_{vt} 성능이 향상하는 것을 확인할 수 있다. YOLOv5l은 mAP_{vt}가 0.5인 반면, YOLOv5l-SB와 YOLOv5-SB + CH는 1.2로 2.4배 성능 향상을 확인할 수 있고 YOLOv5-SB + HR은 mAP_{vt} 1.5로 3배의 성능 향상, YOLOv5-SB + HR + CH는 mAP_{vt}는 1.8로 3.6배의 성능 향상을 확인하였다.

개별 네트워크들의 효과에 대해 분석하기 위해 그림 9와 같이 특징맵을 기반으로 분석하였다. YOLOv5-SB 네트워크와 YOLOv5-SB + CH 네트워크는 작고 밀집도가 높은 객체에 대한 특징이 잘 추출되었지만 배경의 특징도 함께 강조되었다. 반면 YOLOv5-SB + HR은 높은 해상도로 인해 객체와 배경이 분리되며 특징이 추출되었다. 각 네트워크들의 특징맵에 미치는 영향들로 인해 소형 객체에 대한 성능이 향상되지만 소형 객체를 제외한 객체의 성능은 대부분 YOLOv5l 대비 하락한 것을 확인할 수 있다. 이는 YOLOv5-SB 네트워크가 소형객체의 특징 추출 능력은 우수하나 단일 해상도 특징맵을 사용하여 다양한 객체 크기에 대응하지 못하기 때문이다.

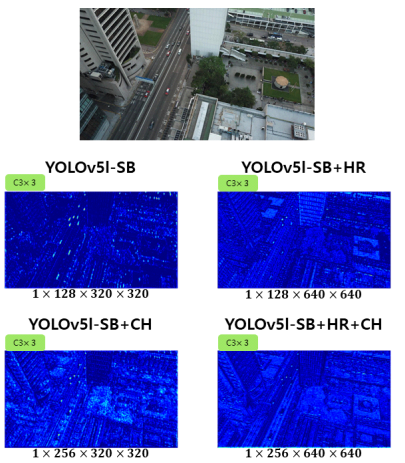
SB + CH와 SB + HR 기법을 적용한 네트워크들은 기존 YOLOv5l-SB 네트워크보다 mAP_{vt} 수치가 높았다. 또한 SB + CH와 SB + HR 기법을 함께 적용한 YOLOv5l-SB + HR + CH 네트워크의 경우 mAP_{vt} 성능이 가장 높은 것을 확인할 수 있다. SB +

표 2. YOLOv5l-SB 네트워크들의 성능 평가 결과
Table 2. The results of YOLOv5l-SB Network

Model	Params.	FLOPs	FPS	$mAP_{0.5}$	mAP	mAP_{vt}	mAP_t	mAP_s	mAP_m	mAP_l
YOLOv5l	**46.27M**	107.9G	65	**31.2**	**17.1**	0.5	4.0	11.9	**26.8**	**40.8**
YOLOv5l-SB	0.32M	65G	**67**	15.6	6.2	1.2	4.2	9.3	8.2	1.5
YOLOv5l-SB + CH	1.109M	226.6G	59	24.7	8.6	1.2	**5.1**	**13.2**	17.4	5.6
YOLOv5l-SB + HR	0.32M	261.9G	21	12.3	4.7	1.5	3.4	7.4	6.2	1.0
YOLOv5l-SB + HR + CH	1.22M	**996.9G**	16	16.0	6.5	**1.8**	4.6	9.8	8.7	1.6



(a) Feature map for tiny objects



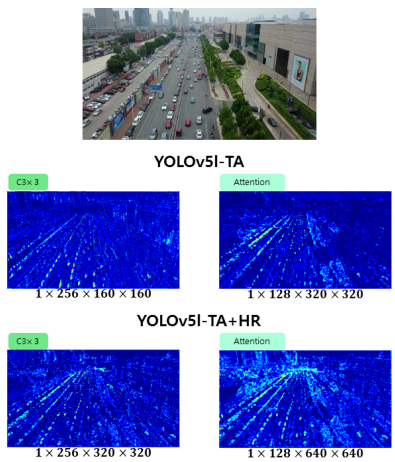
(b) Feature map for complex background and dense objects

그림 9. YOLOv5l-SB 네트워크 특징맵 시각화
Fig. 9. The visualization of the YOLOv5l-SB network feature map

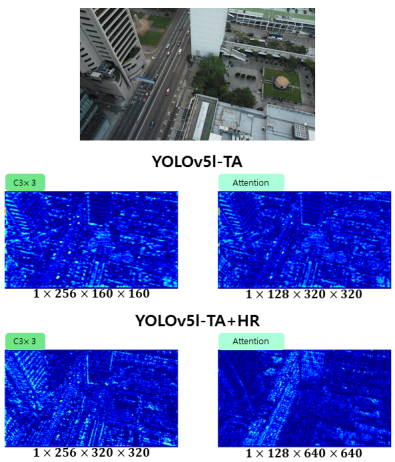
CH와 SB + HR 기법을 적용한 2개의 네트워크는 연산 요구량인 FLOPs가 YOLOv5l-SB 대비 약 3.5배 이상 증가하였다. 따라서 연산량 증가를 최소화하며 다양한 크기의 객체 검출이 가능하도록 제안한 YOLOv5l-TA 네트워크 성능평가에는 기존 YOLOv5l-TA 네트워크 평가와 함께 SB + HR 기법을 적용한 YOLOv5l-TA + HR 네트워크만을 평가하였다. 평가 결과, 표 3과 같이 YOLOv5l-TA 네트워크는 YOLOv5l 대비 $mAP_{0.5}$가 증가하여 11.2%의 성능이 향상하였으며 객체 크기별 mAP 또한 mAP_l를 제외한 나머지 수치는 향상하였다. 이러한 mAP_l 성능감소는 대형 객체에 대한 탐지 및 분류를 수행하는 Head 네트워크의 부재 때문이다. 기존 YOLOv5l 네트워크는 Stage 4, 5 에 Head 네트워크를 결합하여 대형 객체를 검출하였다. 반면 YOLOv5l-TA 네트워크는 소형 객체의 검출정확도 향상과 고해상도 특징맵에 따른 계산 비용 증가의 최소화를 고려하여 Stage 2, 3특징맵에 Head 네트워크를 결합하여 소형, 중형 객체를 검출하였다. 따라서 mAP_l 성능은 감소하였지만 25 FPS 성능을 달성하며 소형, 중형 객체의 효과적 검출이 가능하였다. SB + HR을 적용한 YOLOv5l-TA + HR 네트워크는 Small 객체 이하의 크기 객체들에 대하여 YOLOv5l 대비 성능향상을 확인할 수 있지만 전체 mAP 성능은 하락함을 확인할 수 있다. 그림 10에서 확인할 수 있듯 YOLOv5l-TA + HR과 같이 특징 추출 네트워크에서 1개의 Down-sampling 계층을 사용하여 높은 해상도 특징맵을 유지하는 것은 배경 정보의 불필요한 특징 또한 강조되는 현상이 발생할 수 있으며 이에 따라 전체 성능이 감소할 수 있다. 또한 특징 추출 네트워크에 Stage를 추가한 것은 다양한 객체 크기에 대응하기 위함이지만 높은 해상도를 유지하는 것은 본래의 목적에 적합하지 않은 방법이다. 따라서 Small 이상의 객체 크기를 갖는 객체들의 검출정확도, mAP_m와 mAP_l는 증가하지 않는다. 더불어 FLOPs 또한 YOLOv5l-TA 대비 4배 이상으로 크게 증가하여 네트워크 추론 속도

표 3. YOLOv5l-TA 네트워크들의 성능 평가 결과
Table 3. The results of YOLOv5l-TA Network

Model	Params.	FLOPs	FPS	$mAP_{0.5}$	mAP	mAP_{vt}	mAP_t	mAP_s	mAP_m	mAP_l
YOLOv5l	**46.27M**	107.9G	65	31.2	17.1	0.5	4.0	11.9	26.8	**40.8**
YOLOv5l-SB	0.32M	65G	**67**	15.6	6.2	1.2	4.2	9.3	8.2	1.5
YOLOv5l-SB + HR	0.32M	261.9G	21	12.3	4.7	1.5	3.4	7.4	6.2	1.0
YOLOv5l-TA	2.23M	214.7G	25	**34.7**	**17.7**	**1.9**	6.2	**16.2**	**28.1**	19.1
YOLOv5l-TA + HR	2.24M	**855.2G**	9	29.6	14.1	1.4	**6.6**	15.7	22.8	7.1

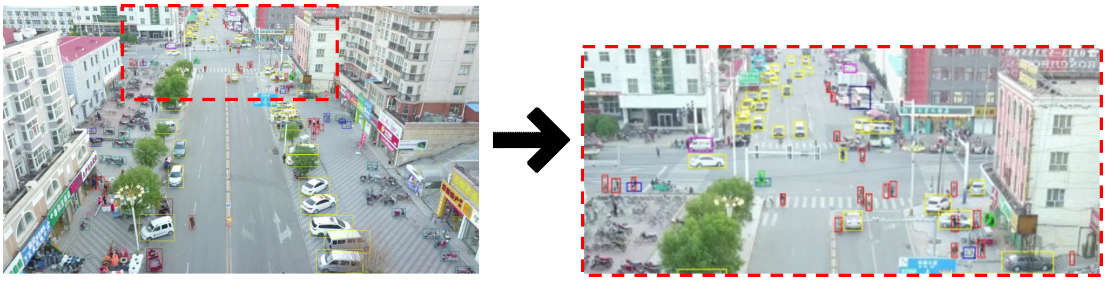


(a) Feature map for tiny objects

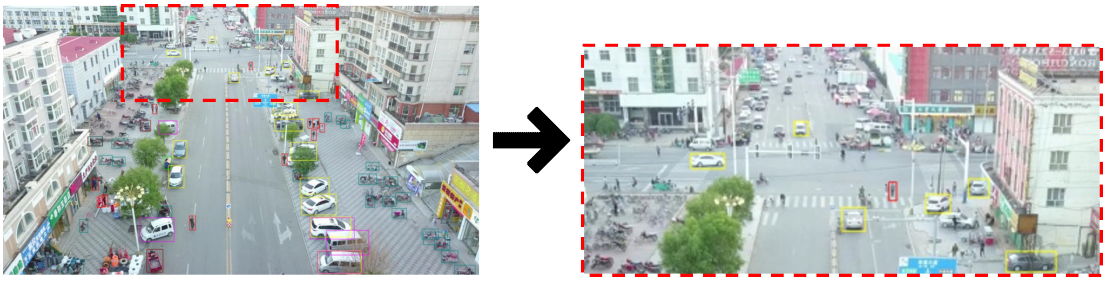


(b) Feature map for complex background and dense objects

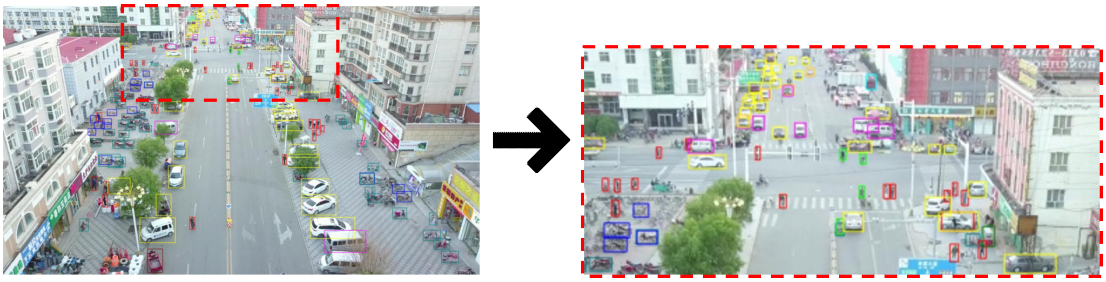
그림 10. YOLOv5l-TA 네트워크 특징맵 시각화
Fig. 10. The visualization of the YOLOv5l-TA network feature map



(a) Ground Truth



(b) YOLOv5l



(c) YOLOv5l-TA

그림 11. 객체 검출기 네트워크의 검출 결과
Fig. 11. The results of object detection

성능이 좋지 못하였다.

3.3 제안하는 네트워크 객체 검출 결과

그림 11에서는 YOLOv5l 네트워크와 제안하는 YOLOv5l-TA 네트워크의 정성적인 결과를 비교하였다. 해당 그림에서 확인할 수 있듯 비교 모델인 YOLOv5l 대비 작고 밀집된 객체의 탐지 능력과 복잡한 배경에서의 객체 탐지 능력이 크게 비교되는 것을 확인할 수 있다. 특히 중형 객체의 검출 성능은 유지함과 동시에 소형 객체 탐지 능력은 큰 수준으로 증가하여 제안하는 네트워크가 항공 영상에 최적화된 네트워크임을 확인할 수 있다.

Ⅳ. 결 론

본 논문에서는 YOLOv5를 기반으로 감시정찰 분야에서 활용될 수 있는 항공 영상에 특화된 YOLOv5l-TA 네트워크를 제안하였다. 제안하는 네트워크는 공간 정보를 보전하기 위한 Spatial Backbone 구조와 객체와 배경 구분에 용이하도록 하는 Triplet Attention을 사용하여 다수의 소형객체가 밀집된 형태로 배경 정보가 많은 항공 영상에 특화되어 $mAP_{0.5}$ 성능을 향상시켰다. 하지만 mAP_{vt}, mAP_{t}, mAP_{s}, mAP_{m}의 전반적인 객체 크기에서 성능 향상을 확인하였지만 mAP_{l}은 감소한 것을 확인할 수 있다. 이러한 문제는 Stage 3으로만 이루어진 작은 특징 추출 네트워크의 한계 때문이며 향후 모든 객체 크기에 대응

하여 성능이 향상하는 네트워크를 개발할 예정이다.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. CVPR*, pp. 770-778, Las Vegas, USA, Jun. 2016.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Int. Conf. Mach. Learn.*, pp. 6105-6114, California, USA, Jun. 2019.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. CVPR*, pp. 580-587, Columbus, USA, Jun. 2014.

[5] R. Girshick, "Fast R-CNN," in *Proc. IEEE Conf. CVPR*, pp. 1440-1448, Boston, USA, Jun. 2015.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in NIPS*, pp. 91-99, Montreal, Canada, Dec. 2015.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. CVPR*, pp. 779-788, Las Vegas, USA Jun. 2016.

[8] J. Redmon and A. Farhadi "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. CVPR*, pp. 7263-7271, Honolulu, USA, Jul. 2017.

[9] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[10] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[11] Ultralytics, yolov5, Retrieved Feb. 18. 2022, from https://github.com/ultralytics/yolov5

[12] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Int. Conf. CVPR Wkshps.*, pp. 390-391, Seattle USA, Jun. 2020.

[13] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang, et al., "Visdrone-det2019: The vision meets drone object detection in image challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vision Wkshps.*, pp. 0-0, Seoul, Korea, Oct. 2019.

[14] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to Attend: Convolutional triplet attention module," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vision*, pp. 3139-3148, Virtual, Jan. 2021.

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. CVPR*, pp. 2117-2125, Honolulu, USA, Jul. 2017.

[16] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. CVPR*, pp. 8759-8768, Salt Lake City USA, Jun. 2018.

[17] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Eur. Conf. Computer Vision*, pp. 3-19, Munich Germany, 2018.

김 해 문 (HaeMoon Kim)

2020년~2022년 : 한양대학교 인공지능융합학과 석사
2022년~현재 : 한화시스템
<관심분야> 영상 처리, 물체 탐지 추적, 영상 변환, 기계학습

이 태 영 (Tae-Young Lee)

2015년~2019년 : 한국과학기술연구원(KIST) 내 실감교류인체감응솔루션연구단
2019년~현재 : 한화시스템
<관심분야> 영상처리, 초해상도 복원, 물체 탐지 추적, 기계학습

안 종 식 (JongSik Ahn)

2017년~2020년 : 에이치비테크놀러지
2021년~현재 : 한화시스템
<관심분야> 영상처리, 물체 탐지 추적, 기계학습, 딥러닝 학습 방법론

최 병 인 (Byungin Choi)

2006년~현재 : 한화시스템
<관심분야> 영상처리, 물체 탐지 추적, 머신러닝, 딥러닝