YOLO를 이용한 Face Detector

- **YOLO5FACE: Why Reinventing a Face Detector**

YOLO5Face: Why Reinventing a Face Detector

Delong Qi, Weijun Tan*, Qi Yao, Jingfeng Liu
Shenzhen Deepcam Information Technologies
Shenzhen, China
{delong.qi,weijun.tan,qi.yao,jingfeng.liu}@deepcam.com
*LinkSprite Technologies, USA, weijun.tan@linksprite.com

***Abstract*—Tremendous progress has been made on face detection in recent years using convolutional neural networks. While many face detectors use designs designated for the detection of face, we treat face detection as a general object detection task. We implement a face detector based on YOLOv5 object detector and call it YOLO5Face. We add a five-point landmark regression head into it and use the Wing loss function. We design detectors with different model sizes, from a large model to achieve the best performance, to a super small model for real-time detection on an embedded or mobile device. Experiment results on the WiderFace dataset show that our face detectors can achieve state-of-the-art performance in almost all the Easy, Medium, and Hard subsets, exceeding the more complex designated face detectors. The code is available at https://www.github.com/deepcam-cn/yolov5-face.**

***Index Terms*—Face detection, convolutional neural network, YOLO, real-time, embedded device, object detection**

I. Introduction

Face detection is a very important computer vision task. Tremendous progresses have been made since deep learning, particularly convolutional neural network (CNN), has been used in this task. As the first step of many tasks, including face recognition, verification, tracking, alignment, expression analysis, face detection attracts many researches and developments in the academia and the industry. And the performance of face detection has improved significantly over the years. For a survey of the face detection, please refer to the benchmark results [1], [2]. There are many methods in this field from different perspectives. Research directions include design of CNN network, loss functions, data augmentations, and training strategies. For example, in the YOLOv4 paper, the authors explore all these research directions and propose the YOLOV4 object detector based on optimizations of network architecture, selection of bags of freebies, and selection of bags of specials [3].

In our approach, we treat the face detection as a general object detection task. We have the same intuition as the TinaFace [4]. Intuitively, face is an object. As discussed in the TinaFace [4], from the perspective of data, the properties that faces has, like pose, scale, occlusion, illumination, blur and etc., also exist in other objects. The unique properties in faces like expression and makeup can also correspond to distortion and color in objects. Landmarks are special to face, but they are not unique either. They are just key points of an object. For example, in license plate detection, landmarks are also used. And adding landmark regression in the object prediction head is straightforward. Then from the perspective of challenges encountered by face detection like multi-scale, small faces and dense scenes, they all exist in generic object detection. Thus, face detection is just a sub task of general object detection.

In this paper, we follow this intuition and design a face detector based on the YOLOv5 object detector [5]. We modify the design for face detection considering large faces, small faces, landmark supervision, for different complexities and applications. Our goal is to provide a portfolio of models for different applications, from very complex ones to get the best performance to very simple ones to get the best trade-off of performance and speed on embedded or mobile devices.

Our main contributions are summarized as following,

- We redesign the YOLOV5 object detector [5] as a face detector, and call it YOLO5Face. We implement key modifications to the network to improve the performance in terms of mean average precision (mAP) and speed. The details of these modifications will be presented in Section III.
- We design a series of models of different model sizes, from large models, to medium models, to super small models, for needs in different applications. In addition to the backbone used in YOLOv5 [5], we implement a backbone based on ShuffleNetV2 [6], which gives the state-of-the-art (SOTA) performance and fast speed for mobile device.
- We evaluate our models on the WiderFace [1] dataset. On VGA resolution images, almost all our models achieve the SOTA performance and fast speed. This proves our goal, as the tile of this paper claims, we do not need to reinvent a face detector since the YOLO5Face can accomplish it.

II. Related Work

A. Object Detection

General object detection aims at locating and classifying the pre-defined objects in a given image. Before deep CNN is used, traditional face detection uses hand crafted features, like HAAR, HOG, LBP, SIFT, DPM, ACF, etc. The seminal work by Viola and Jones [7] introduces integral image to compute HAAR-like features. For a survey of face detection using hand crafted features, please refer to [8], [9].

Since the deep CNN shows its power in many machine learning tasks, face detection is dominated by deep CNN

- **YOLO-FaceV2: A Scale and Occlusion Aware Face Detector**

YOLO-FaceV2: A Scale and Occlusion Aware Face Detector

Ziping Yu[1], Hongbo Huang*[2], Weijun Chen[3], Yongxin Su[4], Yahui Liu[5], and Xiuying Wang[2]

[1]*School of Instrument Science and Opto-electronic Engineering, Beijing Information Science and Technology University, Beijing, China*
[2]*Computer School, Beijing Information Science and Technology University, Beijing, China*
[3]*Data Algorithm, NIO, Shanghai, China*
[4]*School of Mechanical and Electrical Engineering, Beijing Information Science and Technology University, Beijing, China*
[5]*School of Information Management, Beijing Information Science and Technology University, Beijing, China*

Abstract

In recent years, face detection algorithms based on deep learning have made great progress. These algorithms can be generally divided into two categories, i.e. two-stage detector like Faster R-CNN and one-stage detector like YOLO. Because of the better balance between accuracy and speed, one-stage detectors have been widely used in many applications. In this paper, we propose a real-time face detector based on the one-stage detector YOLOv5, named YOLO-FaceV2. We design a Receptive Field Enhancement module called RFE to enhance receptive field of small face, and use NWD Loss to make up for the sensitivity of IoU to the location deviation of tiny objects. For face occlusion, we present an attention module named SEAM and introduce Repulsion Loss to solve it. Moreover, we use a weight function Slide to solve the imbalance between easy and hard samples and use the information of the effective receptive field to design the anchor. The experimental results on WiderFace dataset show that our face detector outperforms YOLO and its variants can be find in all easy, medium and hard subsets. Source code in https://github.com/Krasjet-Yu/YOLO-FaceV2

***Keywords*—** Face detection, YOLO, Scale-Aware, Loss function, Imbalance problem

1 Introduction

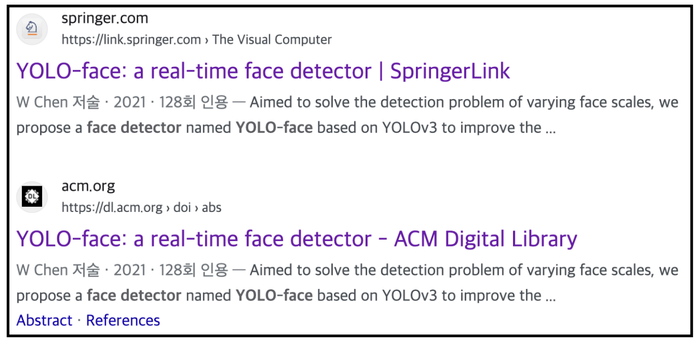
Face detection is an essential step in many face-related applications, such as face recognition, face verification and face attribute analysis, etc. With the booming of deep convolutional neural networks in recent years, the performance of face detectors has been greatly improved. Many high-performance face detection algorithms based on deep learning have been proposed. Generally, these algorithms can be divided into two branches. One branch of typical deep-learning-based face detection algorithms [1, 2, 3] uses cascading means of neural networks as

*Corresponding Author: hhb@bistu.edu.cn

YOLO-FaceV2 : A Scale and Occlusion Aware Face Detector

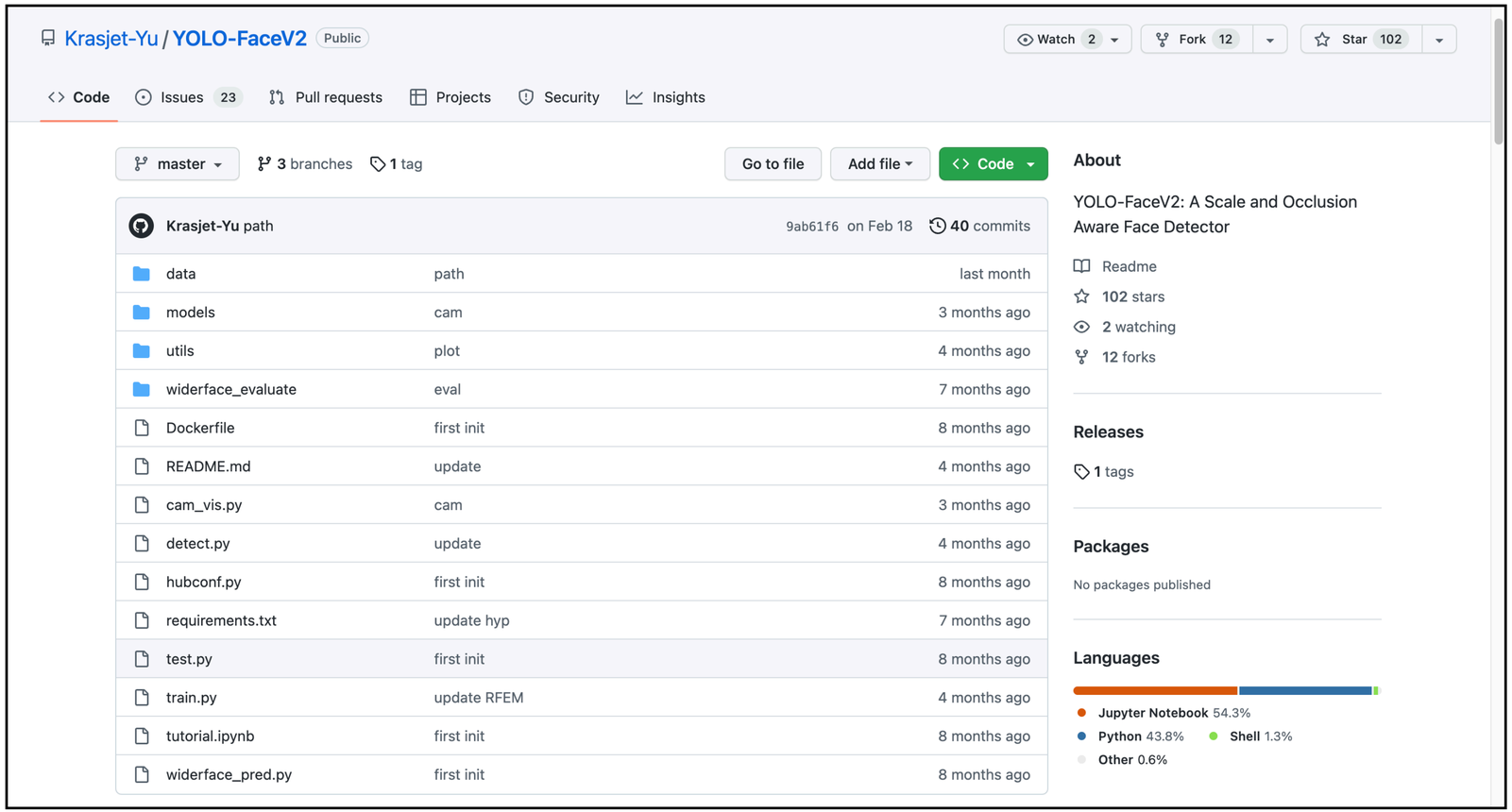
Previous ver과 비교

- YOLO-Face (Previous Ver)
  - an improved face detector based on YOLOv3
    - challenge: focused on the problem of scale variance
    - proposed method
      1. design anchor ratios suitable for human face
      2. utilized a more accurate regression loss function
  - published: 2021 3월, 128회 인용 - but 유료
- YOLO-FaceV2
  - an improved face detector based on YOLOv5
    - challenge
      1. small objects (varying face scales)
      2. imbalance of positive and negative samples
      3. face occlusions
  - published: 2022 8월, 1회 인용 - 무료
  - 소스코드: https://github.com/Krasjet-Yu/YOLO-FaceV2



YOLO-FaceV2

저자 Github



YOLO-FaceV2

데이터셋

- **Dataset: WIDER FACE**
  - **WIDER FACE dataset은 기존에 존재하는 dataset보다 더 challenging함**
    - **annotation 추가**
      - **3개의 event (scale, occlusion, pose)**
        - **Small/Medium/Large scale subsets**
        - **None/Partial/Heavy occlusion subsets**
        - **Typical/Atypical pose subsets**
      - **levels of difficulty : Easy, Medium, Hard subsets**
    - **각 요소에 다한 level에 따라 성능을 체크할 수 있음**

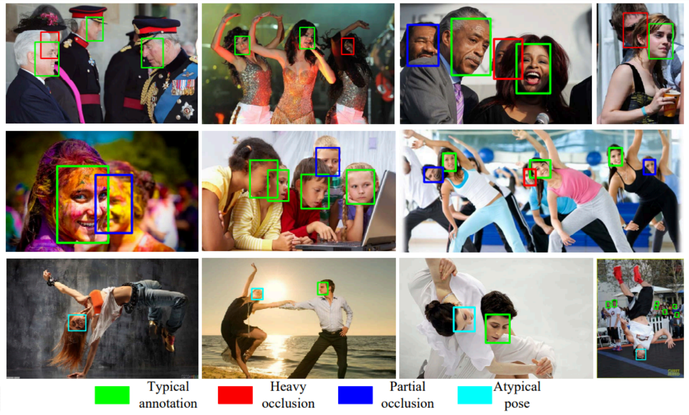
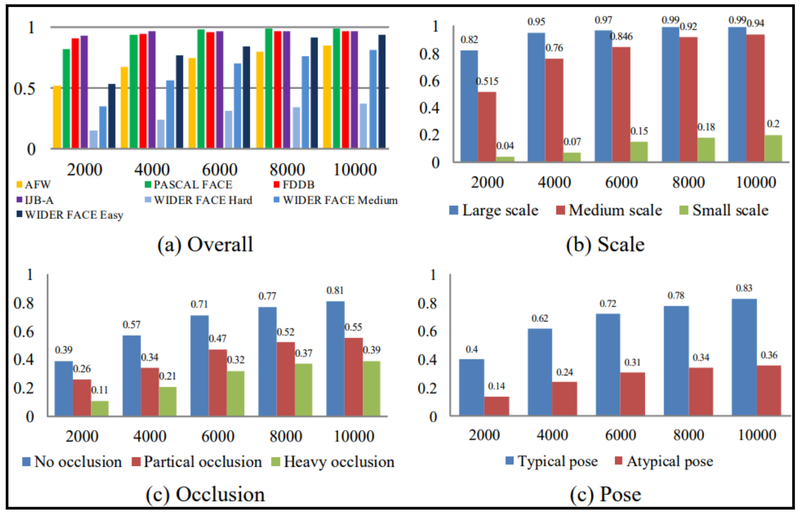


Figure 2. Examples of annotation in WIDER FACE dataset **(Best view in color).**



Wider Face 외 데이터셋

FDDB (Face Detection Dataset and Benchmark)

Edit

Introduced by Jain et al. in Fddb: A benchmark for face detection in unconstrained settings

The **Face Detection Dataset and Benchmark** (**FDDB**) dataset is a collection of labeled faces from Faces in the Wild dataset. It contains a total of 5171 face annotations, where images are also of various resolution, e.g. 363x450 and 229x410. The dataset incorporates a range of challenges, including difficult pose angles, out-of-focus faces and low resolution. Both greyscale and color images are included.

Source: A Comparison of CNN-based Face and Head Detectors for Real-Time Video Surveillance Applications

VGGFace2

Edit

Introduced by Cao et al. in VGGFace2: A dataset for recognising faces across pose and age

The **VGGFace2** dataset is made of around 3.31 million images divided into 9131 classes, each representing a different person identity. The dataset is divided into two splits, one for the training and one for test. The latter contains around 170000 images divided into 500 identities while all the other images belong to the remaining 8631 classes available for training. While constructing the datasets, the authors focused their efforts on reaching a very low label noise and a high pose and age diversity thus, making the VGGFace2 dataset a suitable choice to train state-of-the-art deep learning models on face-related tasks. The images of the training set have an average resolution of 137x180 pixels, with less than 1% at a resolution below 32 pixels (considering the shortest side).

CAUTION: Authors note that the distribution of identities in the VGG-Face dataset may not be representative of the global human population. Please be careful of unintended societal, gender, racial and other biases when training or deploying models trained on this data.

Usage

YOLO-FaceV2

challenge & proposed method

inspired Model: YOLOv5 + TridentNet + Attention Network in FAN

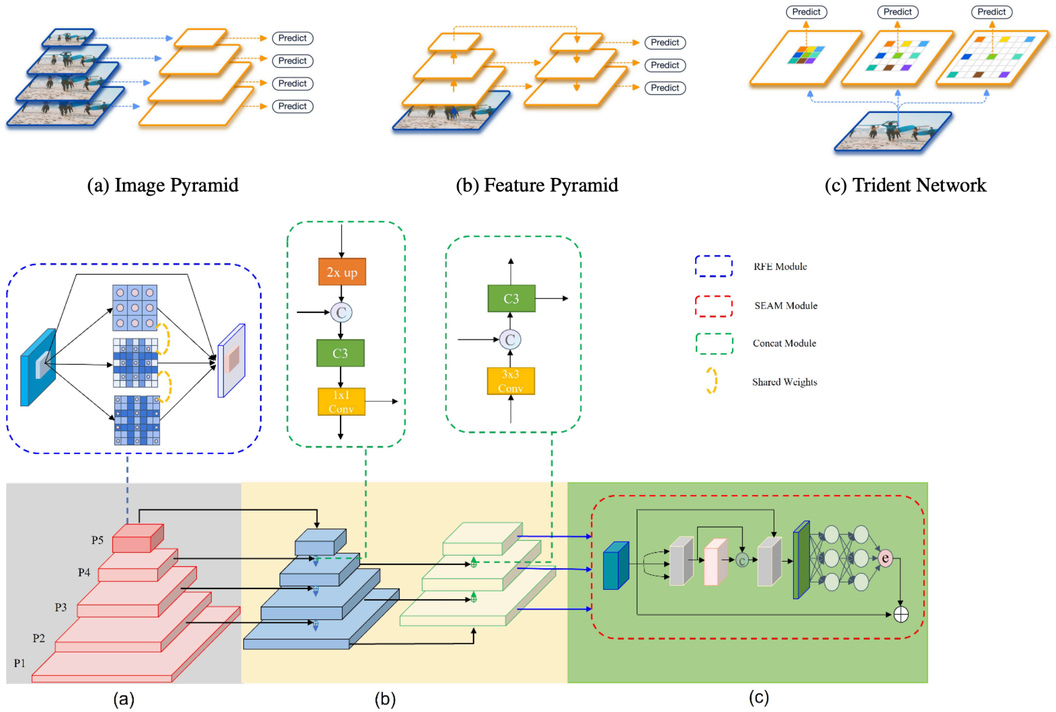
1. **Detecting multiscale faces**
   - **rceptive field & resolution**

     > design a Receptive Field Enhancement module (RFE) to learn different receptive fields of the feature map

     + enhance the feature pyramid representation
2. face occlusions (the occlusion between different faces, the occlusion of faces by other objects)
   - the occlusion between different faces: Repulsion Loss (to improve the recall of intra-class occlusions (class 간 x, class 내부))
   - the occlusion of faces by other objects: attention module SEAM
3. problem of imbalance between hard and easy samples

YOLO-FaceV2

FPN for Multi scale fusion



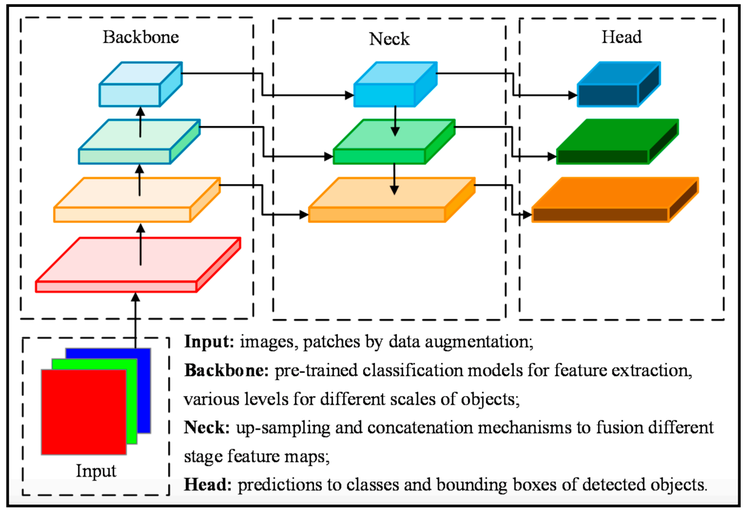
YOLO-FaceV2

method 1 for Multi scale fusion

- **the main method to solve the problem of varying scales**
  : constructing a pyramid to fuse the multi-scale features of faces.
- **in YOLOv5, FPN fuses the features of P3, P4 and P5 layers**
  - **문제점: 작은 물체의 경우, 더 얕은 P3 (번호가 내려갈수록 고해상도)에서도 유지되는 픽셀 정보가 매우 적음**
    (작은 물체의 경우, 쉽게 정보 손실됨)
  - **해결: 해상도 올림 for detecting small objects**
- **fuses P2 layer information of FPN**
  - **목적: to obtain more pixel-level information and compensate the information of small face**
  - **문제점: the detection accuracy of large and medium targets will be slightly reduced because the output feature map perceptual field becomes smaller**
  - **해결: P5 is replaced by the RFE (To expand the receptive field by using dilated convolution)**
    - **RFE - TridentNet 에서 영감을 얻음**
- 공부하다가 헷갈림*: multi-scale objects != small objects*

Yolo v5
네트워크 구조

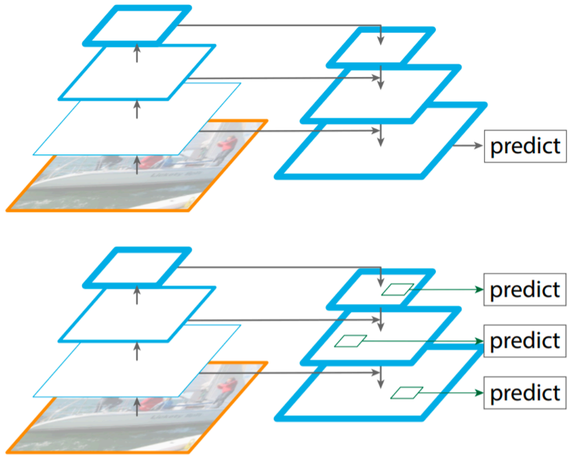
- 이전에 리뷰한 Yolov1-Yolov3와 네트워크 구조가 매우 다름
  - Yolov4에서 네트워크 구조 변경
    - 3개의 구조: Backbone + Neck + Head
    - 다양한 scale의 object들을 검출를 위함
  - Backbone
    - Input -> feature map 추출 (input을 여러 conv layer에 통과)
    - 다수 Pretrained model을 사용
  - Neck
    - Backbone에서 extract된 feature들을 적절하게 조화시킴. (up-sample, concat 등) ex) FPN, PANet
  - Head
    - Localization 및 classification 수행
      - 다양한 크기의 feature map 덕분에, 서로 다른 사이즈의 object를 검출 가능

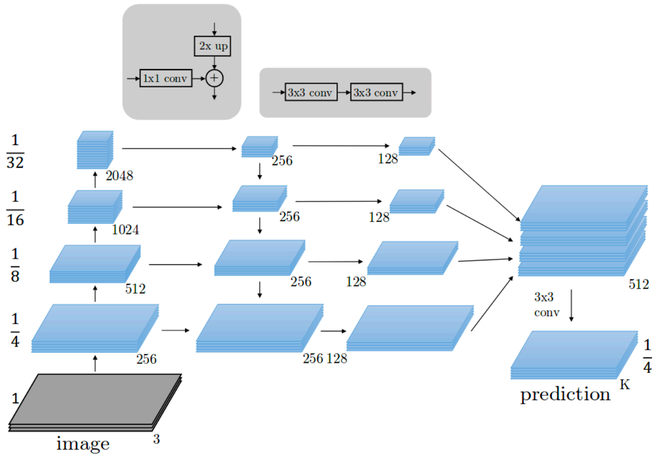


Feature Pyramid Networks for Object Detection

보충자료

- 이미지 내 존재하는 다양한 크기의 객체를 인식하는 것
  - 추론 속도, 메모리 적게 사용하면서 다양한 크기의 객체를 인식하는 방법을 제시

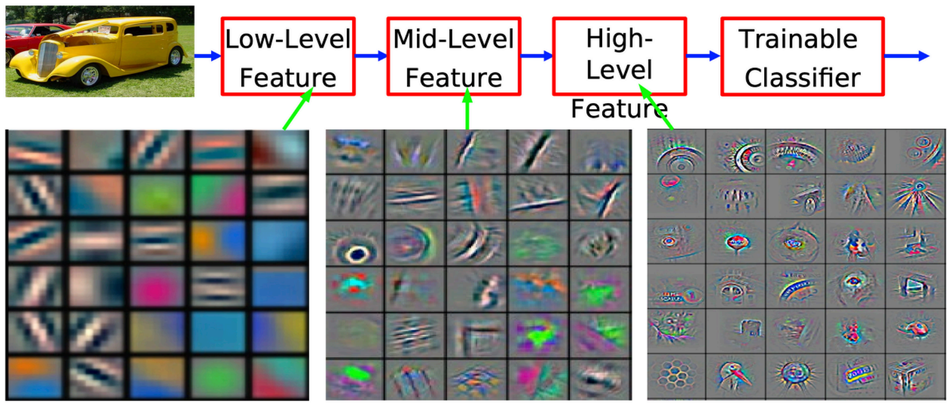




  - Bottom-up pathway: 원본 이미지에서 Conv 연산 수행 후, 각 stage마다 서로 다른 scale을 가지는 4개의 feature map을 추출
  - Top-down pathway: 채널수(channels = 256)를 맞춰주기 위해 각 feature map에 1x1 적용하고, upsampling을 수행한 후, Lateral connections 과정 수행
    - Lateral connections: pyramid level 바로 아래 있는 feature map과 element-wise addition 연산을 수행 후, 4개의 서로 다른 feature map에 3x3 conv 연산을 적용

FPN
특징

- convolutional network에서 얻을 수 있는 서로 다른 해상도의 feature map을 쌓아올린 형태



- 입력층에 가까울수록 (얕을수록):
  - feature map은 높은 해상도 & 가장자리, 곡선 등과 같은 저수준 특징(low-level feature)을 보유
- 입력층에 멀수록 (깊을수록):
  - feature map은 낮은 해상도 & 질감과 물체의 일부분 등 class를 추론할 수 있는 고수준 특징(high-level feature) 보유