検出対象数に応じた物体検出アルゴリズムの動的切り替え手法の提案

Object Detection Scheme Switching for Improving Video Analysis Performance

米田 一成 † 水谷 后宏 ‡,§
Issei Komeda Kimihiro Mizutani

1. はじめに

画像解析技術の発展により，リアルタイムにて対象物の行動を検知，解析出来るようになってきている．画像解析技術の中でも，物体検出という画像内の対象物の位置と，人や車などのカテゴリを指すクラスを特定する手法が提案されている [1]，[2]．これらの手法で利用される物体検出アルゴリズムは，検出精度と処理速度がトレードオフの関係になっているため，それらを両立させることが難しい．そのため，検出精度と処理速度の片方が利用者の設定する性能要件を満たさない場合がある．また，採用される物体検出アルゴリズムは一つであることが多いため，画像内の対象物の数が急激に増加し検出することが出来ない場合など，状況によっては性能要件を満たさない場合がある．そこで，本研究では性能の異なる複数の物体検出アルゴリズムを採用し，それらを映像内の対象物の数に応じて動的に切り替えることで，検出精度と処理速度のトレードオフを緩和した物体検出アルゴリズムを提案する．提案手法では，性能の異なる 3 つの物体検出アルゴリズムとして YOLOv3 [3]，YOLOv5 [4]，MobileNetV2 [5] を採用する．YOLOv3 は 3 つの物体検出アルゴリズムの中で最も検出精度が高いが，処理速度は最も遅いアルゴリズムであり，一方で，MobileNetV2 は 3 つの物体検出アルゴリズムの中で検出精度が最も低いが，最速のアルゴリズムである．そして，YOLOv5 は検出精度が MobileNetV2 より高く，処理速度は YOLOv3 より高いアルゴリズムである．これらの特徴を持った物体検出アルゴリズムを処理速度の速い順に状況に応じて切り替えていく．切り替えのタイミングはそれぞれのアルゴリズムの検出できる対象物の数の最大値とし，その最大値以下の範囲内ではより処理速度の速い物体検出アルゴリズムを採用する．こうすることにより検出精度を落とすことなく最速のアルゴリズムを採用することができる．提案手法の評価では，単体の物体検出アルゴリズムとして YOLOv5 と複数の物体検出アルゴリズムを採用した提案手法を比較し，評価する．また，さらなる処理速度向上のため，最も遅い YOLOv3 を採用しない提案手法を改良した手法についても同様に評価を行う．本稿の構成として，第 2 章にて物体検出に関する関連研究について述べる．第 3 章では提案手法の概要と詳細，実験環境について述べる．第 4 章では提案手法と提案手法を改良した手法を単体の物体検出アルゴリズムの YOLOv5 を用いて比較を行い，評価する．第 5 章では，本稿のまとめと今後の課題について述べる．

† 近畿大学大学院総合理工学研究科, Graduate School of Science and Engineering Research, Kindai University
‡ 近畿大学情報学部, Faculty of Informatics (KDIX), Kindai University
§ 近畿大学情報学研究所, Cyber Informatics Research Institute, Kindai University

2. 関連研究

本章では物体検出に関する研究として，提案手法にて採用する YOLOv3，YOLOv5，MobileNetV2 について説明を行い，それらの手法に共通する課題を述べる．

2.1 YOLOv3

YOLOv3 は物体検出手法の一つである YOLOv2 [6] の改良版である．YOLO [1] とは，You Only Look Once の略であり物体検出手法の一つである．従来の物体検出手法では物体の検出の後に識別の処理を行うような end-to-end な処理構成になっており，処理速度の遅延が発生していた．そこで，YOLO は検出と識別の処理を同時に行うことで処理速度の遅延を解消した．この YOLO の改良版となるのが YOLOv2 であり，YOLOv2 をさらに改良したものが YOLOv3 である．YOLOv3 は ResNet [7] と FPN 構造 [8] を採用することで検出精度を更に向上させたアルゴリズムである．ResNet は Convolutional Neural Network(CNN) のある層で求める最適な出力を学習するのではなく，前層の入力を参照した残差関数を学習することで特徴量学習を促進させ，その結果，層の多重化が可能となり精度の向上に役立つことが知られている．FPN 構造とは，物体検出向けにマルチスケールの画像特徴集約を効率的に行うための CNN の拡張機構である．異なる特徴マップで物体検出を行い，その特徴マップまでに取得した情報を利用することで，高レベル特徴だけでなく中レベルや小レベルの特徴を利用することが可能となり，小さな物体の未検出を減らすことができる．提案手法では YOLOv3 を物体検出アルゴリズムの一つとして採用する．

2.2 YOLOv5

YOLOv5 は YOLOv3 の改良版であり，YOLOv4 [9] の構造と類似している．主要な 3 つの構造として，Back-

bone には Cross Stage Partial network(CSPNet [10]) という特徴マップの一部のみを畳み込み演算し，残りを連結させることで精度を落とさずに高速化する手法が用いられる．Neck には Spatial pyramid pooling(SPP) という CNN において任意サイズの入力画像から固定サイズの特徴ベクトルを出力する特徴を持つ CNN の層と，Path Aggregation Network(PAN [11]) という Mask R-CNN [12] をベースに異なるレベルの空間的特徴量を利用したネットワークが用いられる．なお，Head には YOLOv3 の Head が用いられている．YOLOv4 と YOLOv5 の違いは，YOLOv4 では C で記述された Darknet フレームワークを利用しているが，YOLOv5 は Python のフレームワークに基づいて点である．提案手法ではこの YOLOv5 を物体検出アルゴリズムの一つとして採用する．

2.3 MobileNetV2

MobileNetV2 は MobileNet[13] の後継モデルとして，MobileNet の構想，およびそのモデルを基礎にしながらモジュールを大幅に改良したものである．MobileNet はスマートフォンなど小型デバイスのようにリソースに制約のある環境下でも使用でき，かつ高精度であるように設計された CNN である．MobileNet では標準的な畳み込みではなく，Depthwise Separable Convolution という計算方法を採用している．通常の畳み込みでは，空間方向とチャンネル方向の畳み込みを同時に行うが，Depthwise Separable Convolution では空間方向の畳み込みを行った後に，チャンネル方向の畳み込みを行う．こうすることにより，空間方向とチャンネル方向を乗算した計算コストが，和算分で済むようになりリソースに制約のある環境でも高速に動くことができる．この MobileNet の改良版である MobileNetV2 は，活性化関数である ReLU の入力値が 0 以下の場合，伝搬情報を消失させてしまう問題に対して，Inverted Residual Block を採用することで情報の消失を防ぐことで，より検出精度を向上させた CNN である．この Inverted Residual Block は通常の Residual Block と異なり，Block の中間部分の次元数を増やすことで ReLU により，本来，消失してしまう情報を他の次元に持たせ，情報の消失を防いでいる．提案手法では Single Shot MultiBox Detecter(SSD [2]) という物体検出手法の backbone に MobileNetV2 を利用した物体検出アルゴリズムの MobileNetV2 を採用する．

2.4 関連研究の課題

これらの関連研究にて提案された物体検出手法は，測定した検出精度と処理速度以上の性能は見込めない．また，一般的には採用した物体検出アルゴリズムは継続的に使用され，状況に応じて変えられることはない．したがって，測定した検出精度と処理速度の片方または両方が利用者の設定する性能要件を満たせない場合，他の物体検出手法を検討する必要がある．そこで，本提案手法は複数の物体検出アルゴリズムを採用し，状況に応じて切り替えることで検出精度と処理速度のトレードオフを緩和し，単体の物体検出アルゴリズムを採用する場合よりも性能要件を満たすことができる物体検出手法を提案する．

3. 提案手法

本章では提案手法で採用する 3 つの物体検出アルゴリズムの詳細と提案手法の概要，物体検出アルゴリズムを切り替えるタイミング設定，提案手法を評価する実験環境として作成したシステムについて述べる．

3.1 採用する 3 つの物体検出アルゴリズムの詳細

提案手法では，異なる検出精度と処理速度を持つ物体検出アルゴリズムとして，YOLOv3，YOLOv5，MobileNetV2 の 3 つの物体検出アルゴリズムを採用し，これらを状況に応じて切り替えることを目指す．表 1 は，映像に写る対象物の数に応じた 3 つの物体検出アルゴリズムの検出精度を測定した結果を示す．A から E のアルファベットは映像に写る対象物 (本研究では対象物を人間のみとする) の数が異なる動画を表しており，A は 1 人から 2 人，B は 5 人から 6 人，C は 9 人から 10 人，D は 20 人から 25 人，E は 30 人から 35 人写っている動画を用いて精度を測定した．表 1 より，MobileNetV2 はこの 3 つの物体検出アルゴリズムの中で最も精度が低いことが分かる．しかし，B までの検出率は 90%以上を保っている．つまり，映像内に写る対象物の数が少ない時、MobileNetV2 でも検出可能であることが分かる．同様に YOLOv5 は D までの検出率を 90%以上保っており，映像に写る対象物の数が 20 人程度までは YOLOv5 で検出可能であることが分かる．そして，最も精度の高い物体検出アルゴリズムは YOLOv3 である．表 2 は同じ実験環境で処理速度を測定した結果を示す．処理速度は FPS を用いる．MobileNetV2 は約 43 fps，YOLOv5 は約 11 fps，YOLOv3 は約 5 fps であった．したがって，MobileNetV2 は 3 つのアルゴリズムの中で検出精度が最も低いが，最速のアルゴリズムである．YOLOv3 は 3 つのアルゴリズムの中で検出精度が最も高いが，最も遅いアルゴリズムである．そして，YOLOv5 は MobileNetV2 より検出精度が高く，YOLOv3 より処理速度が速いアルゴリズムである．以上の結果より，提案手法で採用する 3 つの物体検出アルゴリズムの検出精度と処理速度はトレードオフの関係にあることが分かる．また，MobileNetV2 のような精度の低い物体検出アルゴリズムでも，映像に写る対象物の数によっては高い検出精度を保つことができることも分かった．この 3 つの物体検出ア

表 1: A から E の映像に写る人数が異なる動画に対して測定した，100%を 1.00 とする 3 つのアルゴリズムの対象物の数に応じた精度

Model	A	B	C	D	E
YOLOv3	1.00	1.00	0.99	0.97	0.92
YOLOv5	1.00	1.00	0.95	0.91	0.79
MobileNetV2	1.00	0.93	0.78	0.59	0.54

表 2: A から E の映像に写る人数が異なる動画に対して測定した 3 つのアルゴリズムの処理速度 (fps)

Model	A	B	C	D	E
YOLOv3	5	5	5	3	3
YOLOv5	12	12	10	10	10
MobileNetV2	45	45	43	43	43

ルゴリズムを提案手法では採用し，状況に応じて切り替える．

3.2 提案手法の概要

提案手法の概要図を図 1 に示す．性能の異なる 3 つの物体検出アルゴリズムである YOLOv3，YOLOv5，MobileNetV2 は検出精度と処理速度が異なる．3.1 章で述べたように，YOLOv3 は最も精度が高いが，処理速度が最も遅いアルゴリズムであり，MobileNetV2 は最も精度が低いが，最速のアルゴリズムである．そして，YOLOv5 は平均的な精度と速度を持つ．また，映像に写る対象物の数が少ない場合，検出精度の低いアルゴリズムでも高い検出率を保つことが分かった．したがって，提案手法ではこれら 3 つのアルゴリズムを映像に写る対象物の数に応じて，処理速度の速い順に物体検出アルゴリズムを切り替える．具体的には，表 1 より MobileNetV2 は B まで検出率を 90%以上保っており，YOLOv3 や YOLOv5 と同等の検出精度であり，3 つのアルゴリズムの中で最速であるため，映像に写る対象物の数が 5,6 人までは MobileNetV2 を採用することで，高い検出精度を保ちながら処理速度の向上を図ることが出来る．同様に YOLOv5 は D まで YOLOv3 と同等の高い検出精度を保つことができ，YOLOv3 よりも処理速度が速い．このことから映像に写る対象物の数が 20 人程度までで，かつ MobileNetV2 が対象物を検出できる範囲を超えている場合，YOLOv5 を採用することとした．なお，対象物の数が約 20 人より多い場合には，MobileNetV2 や YOLOv5 では検出精度が落ちるため，YOLOv3 を採用し検出精度の向上を図る．つまり，提案手法は，検出精度が同程度の範囲内では，より処理速度の速いアルゴリズムを採用する．また，表 2 より YOLOv3 は約 5 fps であるためリアルタイム処理には向いていない．よってさらなる処理速度向上のため，YOLOv3 を採用せず MobileNetV2 と YOLOv5 のみを採用する手法も提案（以下、改良手法）する．

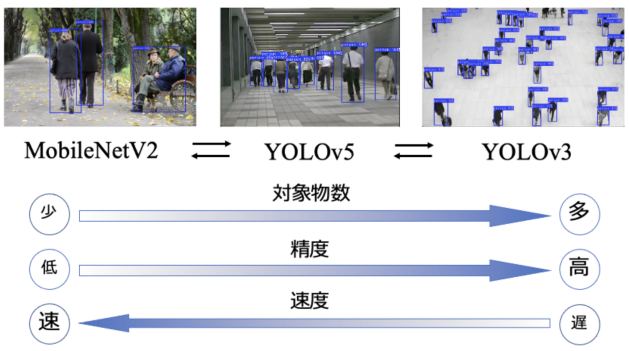

図 1: 提案手法の概要図

3.3 切り替えタイミングの設定

物体検出アルゴリズムを切り替えるタイミングは，画像ごとに物体検出アルゴリズムが検出できる対象物の数の最大値に依存する．図 2 の左の図は表 1 の結果から，検出精度が 90%を超えた場合の，それぞれのアルゴリズムの検出可能な対象物の数の最大値を示しており，右の図は 3 つの物体検出アルゴリズムを採用した提案手法の切り替えタイミングと，MobileNetV2 と YOLOv5 の 2 つの物体検出アルゴリズムのみを採用した改良手法の切り替えタイミングを示している．左の図に示した通り，MobileNetV2 は検出できる対象物の数が約 5 人以下であり，YOLOv5 は約 25 人以下，YOLOv3 は約 25 人以上検出することが出来ると分かる．提案手法は検出精度が同程度の範囲内において，より処理速度の速い物体検出アルゴリズムを採用するため，対象物の数が 5 人以下の場合は最速の MobileNetV2 を採用する．続いて対象物の数が 5 人以上 25 人未満までは YOLOv3 より速い YOLOv5 を採用する．そして対象物の数が 25 人以上の場合は YOLOv3 を採用する．改良手法では YOLOv3 を採用しないため，25 人以上も YOLOv5 を採用することとする．また，図 3 は画像ごとに提案手法で採用する物体検出アルゴリズムを切り替えるフローを表している．直前に物体検出を行った画像に写る対象物の数から，対象となるアルゴリズムを判断し切り替え，次の画像の物体検出を行う．図 3 の場合，一つ前の画像に写る対象物の数が 8 人で，5 人以上 25 人未満を満たすため YOLOv5 を採用し，次の画像の物体検出を YOLOv5 が行う．

3.4 実験環境

提案手法の実験環境として，撮影した映像をサーバ上で解析し，解析結果を Web ページ上で配信する映像処

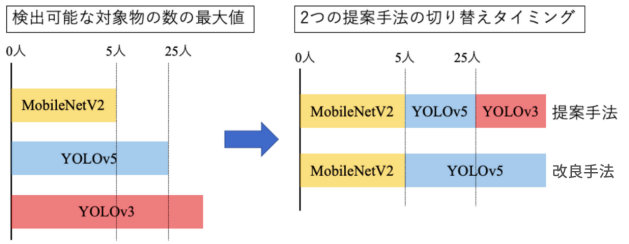

図 2: 左図:3 つのアルゴリズムの検出可能な対象物の数の最大値，右図:2 つの提案手法で採用するアルゴリズムの切り替えのタイミング

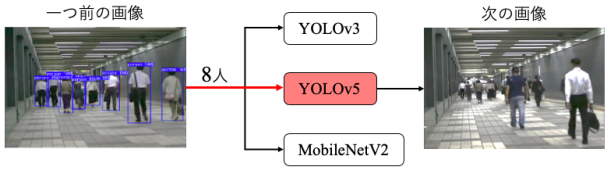

図 3: 画像ごとに物体検出アルゴリズムを切り替える際のフロー

理システムを作成した．図 4 は作成したシステムの構成を示している．カメラデバイスを用意し，映像データをサーバへ送信する．この時 FFmpeg というビデオファイルに関する録音や変換などの機能を提供するマルチメディアフレームワークのストリーミング機能を利用して映像データを送信し続ける．送信先のサーバ上ではカメラデバイスから送信された映像データに対して，3 つの物体検出アルゴリズムの切り替えタイミングとなるパラメータを設定する．このパラメータは採用する物体検出アルゴリズムの検出することができる，対象物の数の最大値を指す．そして，そのパラメータを基に物体検出アルゴリズムを状況に応じて切り替え，出力結果を Web ページ上で配信し続ける．提案手法の評価はこのシステム上で行う．

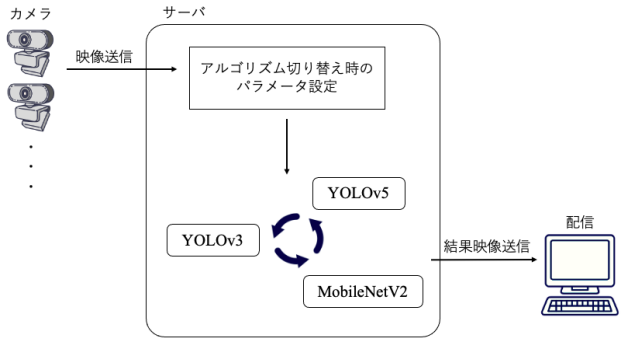

図 4: 映像処理システムの構成

図 5: A から E の映像

4. 提案手法の評価

本章では提案手法を評価した結果について述べる．複数の物体検出アルゴリズムを採用する提案手法が，一つの物体検出アルゴリズムのみを採用した手法と比べ，検出精度と処理速度のトレードオフがより緩和されているかを明らかにするために，提案手法とその改良手法，そして単体の物体検出アルゴリズムとして YOLOv5 の検出精度と処理速度を比較する．この時表 1 で使用した A から E の映像を用いる．

4.1 A から E の映像

Pixabay [14] と言う写真や映像の共有サイトから人が写っている映像を取得する．図 5 は映像に写る人の数が異なる 5 つの A から E の映像であり，提案手法を評価する際に使用する．図 5 の左上の A の映像には 1 人から 2 人写っており，右上の B の映像には 5 人から 6 人，左中段の C の映像には 9 人から 10 人，右中段の D の映像には 20 人から 25 人，左下の F の映像には 30 人から 35 人写っている．これら 5 つの映像の FPS は 20 fps で，それぞれの映像の長さは 5 秒ずつである．提案手法を評価する際にはこれらの映像を組み合わせた 1 つの映像を使用する．この映像は前述の表 1，2 より物体検出アルゴリズムの検出精度と処理速度を測定した際に用いた映像である．映像処理システムのサーバ上に A から E の映像を配置し，提案手法の評価を行う．

4.2 提案手法の評価結果

図 6 は提案手法，改良手法，YOLOv5 の検出精度を比較した結果を示している．横軸は A から E のそれぞれ 5 秒ずつの映像を 1 つに組み合わせた映像の時間である．提案手法とその改良手法は A から B の映像にあた

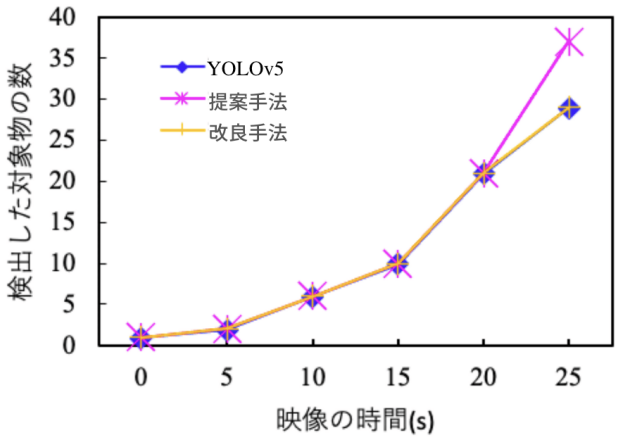

図 6: 提案手法，改良手法，YOLOv5 の検出精度を比較した結果

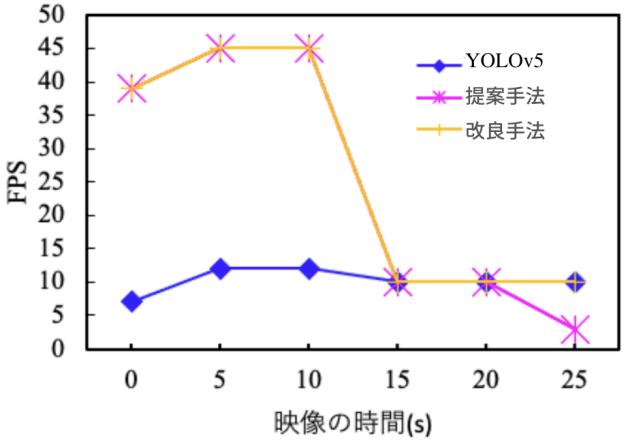

図 7: 提案手法，改良手法，YOLOv5 の処理速度 (FPS) を比較した結果

表 3: 100%を 1.0 とした動画全体の精度と処理時間 (s)

モデル	検出精度	処理時間
YOLOv5	0.93	50.62
提案手法	0.94	45.51
提案手法の改良手法	0.93	30.25

る 0 秒から 10 秒まで MobileNetV2 を採用しているが，YOLOv5 の精度と比較すると差は小さいことが分かる．また，提案手法は E の映像にあたる 20 秒から 25 秒の間で YOLOv3 を採用しているため検出精度が他二つよりも高くなっている．表 3 の検出精度より，改良手法と YOLOv5 は同じ精度であり，提案手法は YOLOv3 を採用しているため他二つより高い結果となった．図 7 は提案手法，改良手法，YOLOv5 の処理速度を比較した結果を示している．提案手法と，その改良手法は 0 秒から 10 秒まで MobileNetV2 を採用しているため，YOLOv5 よりも速くなっていることが分かる．そして，提案手法は 25 秒で YOLOv3 を採用しているため速度が落ちているが，表 3 の処理時間をみると，動画全体の処理時間は YOLOv5 より短いことが分かる．つまり，提案手法は精度と速度の両方において，単体で採用した YOLOv5 よりも高くなった．改良手法の精度は YOLOv5 と同等でありながら，処理時間を約 40%削減することができた．以上の結果より，性能の異なる複数の物体検出アルゴリズムを状況に応じて切り替える本提案手法は，従来の単体で採用された物体検出アルゴリズムの検出精度と処理速度間のトレードオフをより緩和し，優れた性能を発揮することが分かった．

5. まとめと今後の課題

本稿では，単体で採用される物体検出アルゴリズムの検出精度と処理速度間にあるトレードオフを緩和し，より性能要件を満たすため，性能の異なる複数の物体検出アルゴリズムを採用し，それらを映像内の対象物の数に応じて切り替える手法を提案した．提案手法と単体の物体検出アルゴリズムの YOLOv5 を比較した結果，検出精度と処理速度の両方が向上した．また，改良手法は YOLOv5 と同等の検出精度を保ちながら，処理速度が約 40%向上した．今後の課題として，物体検出アルゴリズムを切り替えるタイミングとなる，検出可能な対象物の数の最大値を求めるには，事前に対象の映像を用いてそれぞれの物体検出アルゴリズムの検出可能な最大値を測定する必要があった．そこでリアルタイム映像処理を実現するために，事前に対象の映像を用いなくても物体検出アルゴリズムの切り替えタイミングを設定できる新たな手法を研究する予定である．また，提案手法は 1 つの映像に対して評価を行ったが，複数の映像に対してもそれぞれに見合った物体検出アルゴリズムを選択できるシステムの構築を目指す．

参考文献

[1] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, ”You Only Look Once: Unified, Real-Time Object Detection ” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.779788, 2016.

[2] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg, ”SSD: Single Shot MultiBox

Detector ” in Proc. European Conference on Computer Vision (ECCV), pp. 21-37, 2016.

[3] Joseph Redmon et al, ” YOLOv3: An Incremental Improvement ” arXiv preprint arXiv:1804.02767.

[4] G Jocher et al, ”ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements ” Oct. 2020.

[5] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen, ” MobileNetV2: Inverted Residuals and Linear Bottlenecks ” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4510-4520, 2018.

[6] Joseph Redmon, Ali Farhadi, “ YOLO9000: Better, Faster, Stronger ” arXiv preprint arXiv: 1612.08242, 2016.

[7] K. He, X. Zhang, S. Ren, and J. Sun, ”Deep residual learning for image recognition” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), arXiv preprint arXiv: 1512-03385, 2016.

[8] Tsung-Yi. Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie, ”Feature pyramid networks for object detection”, in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2117–2125, 2017.

[9] Alexey Bochkovskiy et al, ” YOLOv4: Optimal Speed and Accuracy of Object Detection ” arXiv preprint arXiv: 2004.10934, 2020.

[10] C.Y. Wang, H.Y.M. Liao, Y.H. Wu, P.Y. Chen, J.W. Hsieh, I.H. Yeh, ”CSPNet: A New Backbone that can Enhance Learning Capability of CNN” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops, pp. 390–391, 2020.

[11] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, Jiaya Jia, ”Path Aggregation Network for Instance Segmentation”, in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), arXiv preprint arXiv: 1803.01534, 2018.

[12] Kaiming He, Georgia Gkioxari, Piotr Dolla´r, Ross Girshick, ”Mask R-CNN”, in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2961–2969, 2017.

[13] Andrew.G.Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, ” MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications ” arXiv preprint arXiv:1704.04861v1, 2017.

[14] pixabay : https://pixabay.com/