様々な特徴抽出器を用いた動画像の連続フレームからの物体検出の評価

工学部工学科　情報通信工学プログラム

60211682　新西　拓斗

指導教員　椋木雅之

研究背景

物体検出:

　動画や画像に含まれる特定の物体の位置と範囲を推定する技術

深層学習を用いた物体検出手法

　大量のデータを学習することで高精度の物体検出が可能

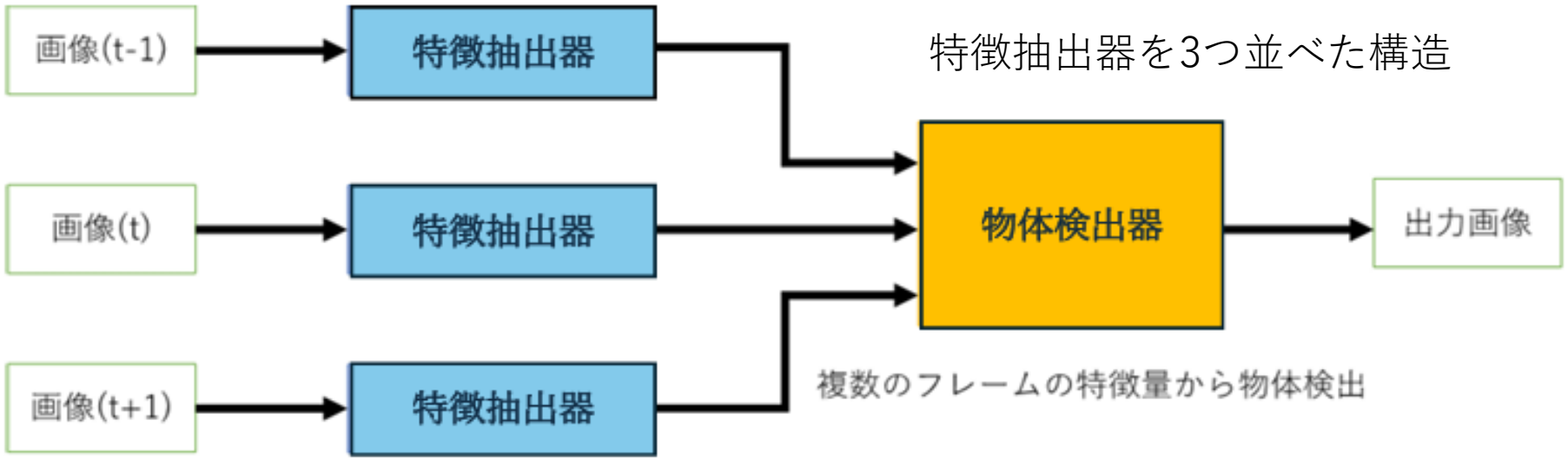
問題点

　連続するフレームでの画像内容は変化が小さいのにも関わらず、連続での検出に失敗する

田邊[1]の手法

動画像の連続する3フレームを入力として物体検出する手法

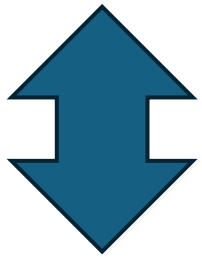
- 時間方向に安定して検出できた
- 物体検出の精度自体は高くなかった



[1]田邊英介, ”深層学習による動画像の連続フレームからの物体検出”, 宮崎大学工学院工学研究科令和4年度修士論文 (2022)

問題点

特徴抽出器:
- 画像の持つ特徴を特徴量として抽出
- 検出対象固有の特徴を抽出できれば、検出精度が向上
- CNNやTransformerが活躍



田邊の研究では
- ResNet[2]を特徴抽出器として使用
- 特徴抽出器の影響を十分検討していなかった

[2] Kaiming He , Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition”, CVPR (2016)

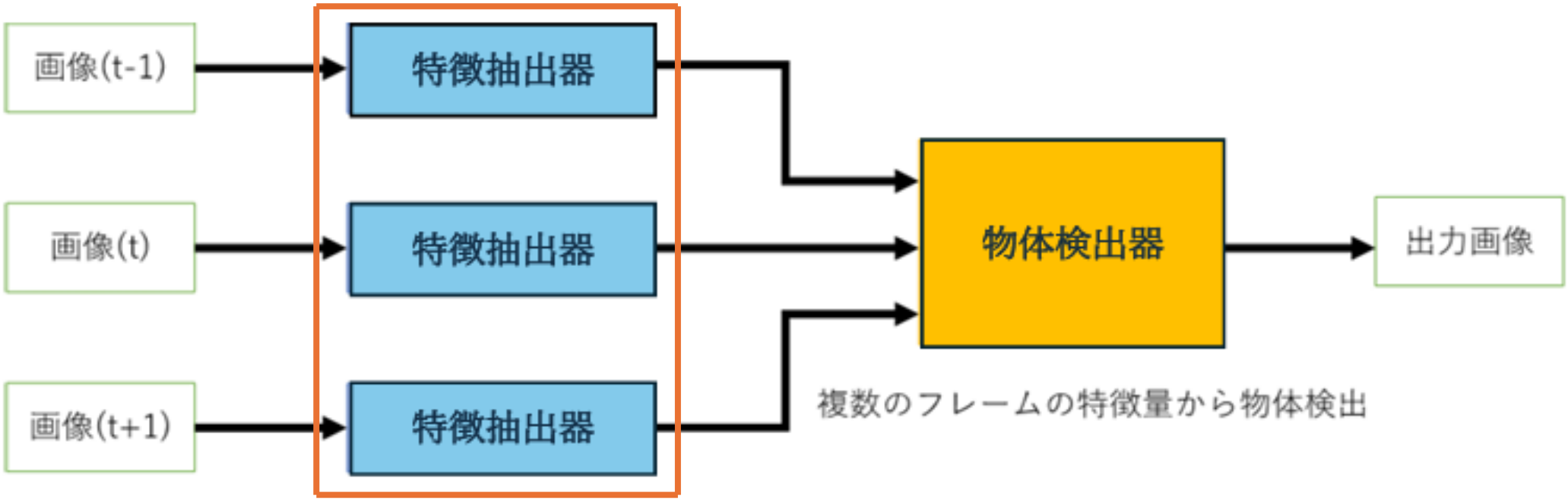
研究目的

特徴抽出器の違いによる検出精度の変化を分析し最適な構成を探索する

特徴抽出器を様々なモデルに変更し物体検出精度を評価する

提案手法のネットワーク構造

- 基本構造はYOLOv3
- 特徴抽出器を様々なモデルに変更



特徴抽出器

●ResNet

- スキップ接続を導入
- 深いニューラルネットワークの勾配消失問題を解決
- 本研究ではResNet18とResNet152を使用

●EfficientNet[3]

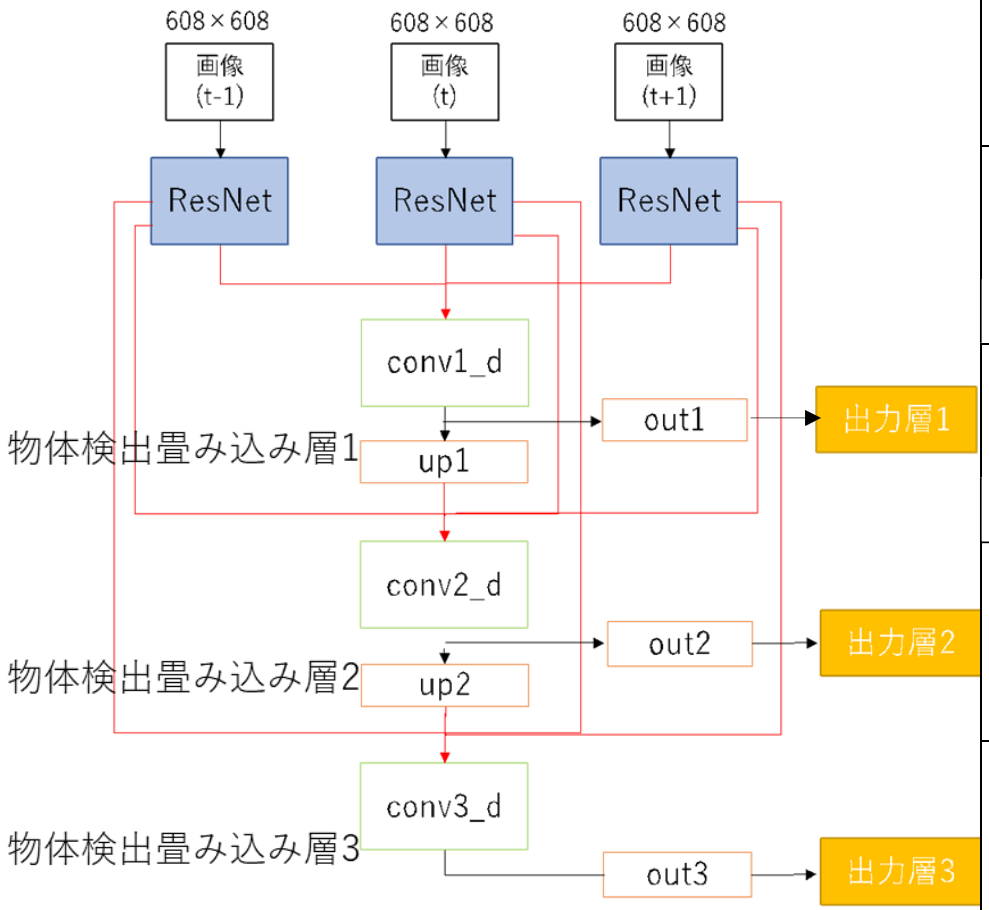
- 高精度と計算効率を両立したCNNモデル
- 本研究ではEfficientNetB0とEfficientNetB7を使用

●SwinTransformer[4]

- Transformerベースのモデル
- スライディングウィンドウ機構を採用
- 局所性と計算効率を向上

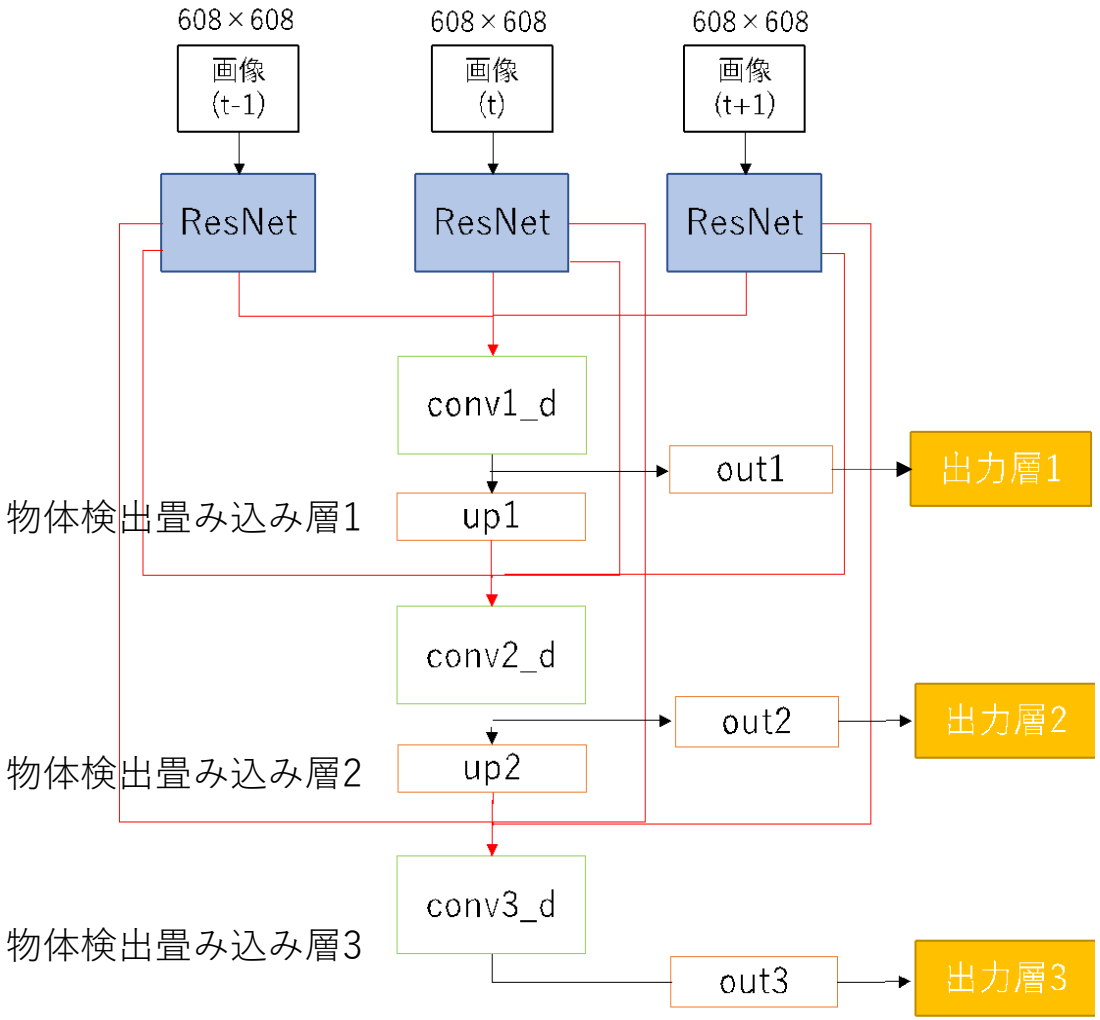
[3] Tan, Mingxing, and Quoc V. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.” ICML (2019)
[4] Ze Liu(Microsoft Research Asia) et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”, ICCV (2021)

物体検出器



	層	サイズ	チャネル
ResNet152	Conv3_x	76×76	512
	Conv4_x	38×38	1024
	Conv5_x	19×19	2048
ResNet18	Conv3_x	76×76	128
	Conv4_x	38×38	256
	Conv5_x	19×19	512
EfficientNet-B0	4 層目	152×152	40
	5 層目	76×76	80
	7 層目	38×38	192
EfficientNet-B7	4 層目	38×38	80
	5 層目	76×76	160
	7 層目	152×152	384
SwinTransformerB	Stage2	28×28	256
	Stage3	14×14	512
	Stage4	7×7	1024

物体検出器



出力

- 物体の位置、範囲
- 種類
- 評価値

出力結果を統合して、物体検出結果とする

実験

- 各特徴抽出器を用いた検出精度を比較する

- 検出する対象物体は馬の1クラス
- 実験データとして35本の馬の動画像を使用
- 合計で1400枚の画像を使用

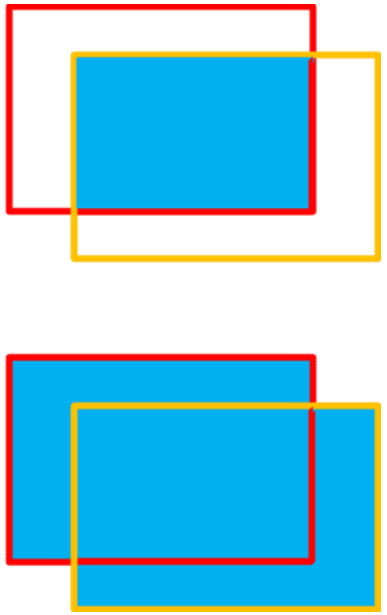
評価方法

平均適合率(AP)を使用

AP：検出した正解ラベルの割合の平均

正解の判定にはIntersection over Union (IoU)を使用

$$\mathrm{IoU} = \frac{\text{重複領域}}{\text{全体領域}}$$



閾値:0.5以上を正解とする

評価方法

TP：正解したBBox
FP：正解でないBBox
FN：どの検出したBBoxとも紐づいていない正解の矩形

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

PrecisionからAPを求める

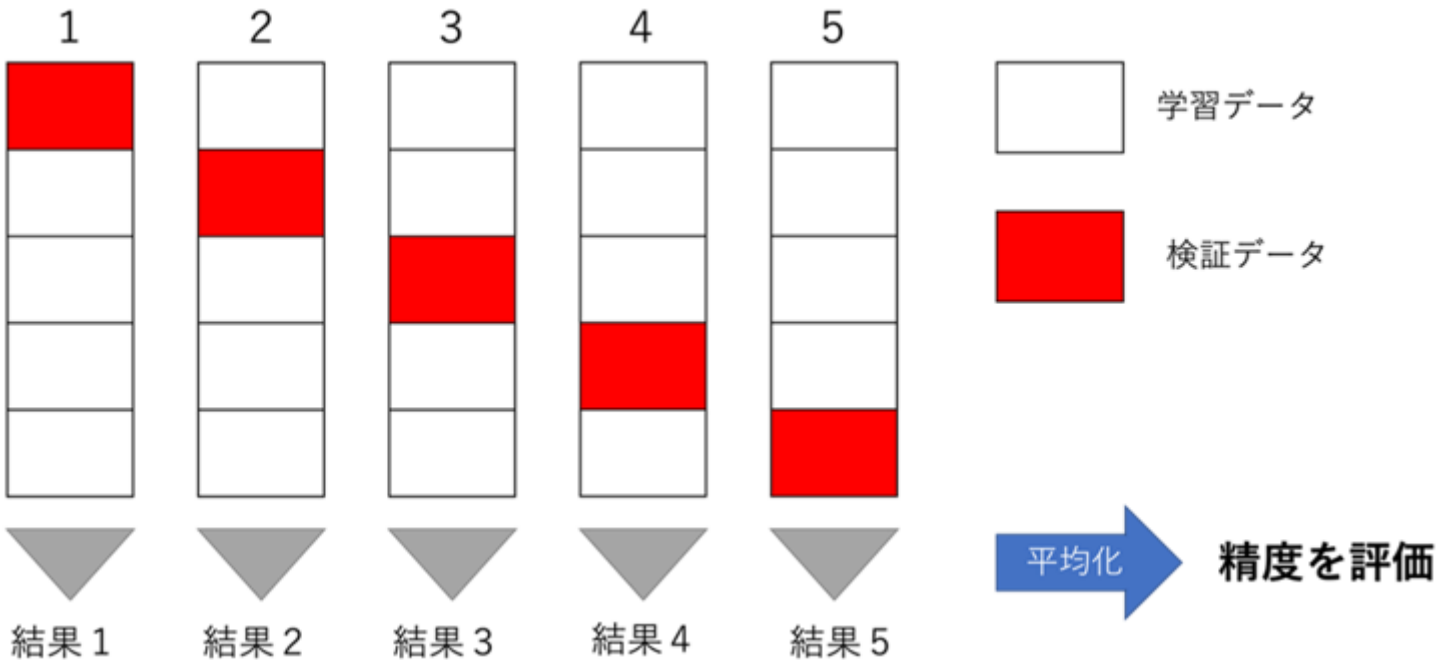
$$\text{AP} = \int_0^1 P(r)dr$$

$P(r)$:信頼度の閾値がrの時のPrecision

APの最大値は1となり、値が大きいほど検出精度が高い

クロスバリデーション

- 5つのパターンで検証
- 学習回数は10000回



実験の結果(平均AP)

	パターン１	パターン２	パターン３	パターン４	パターン５	平均
Resnet152	**0.7050**	**0.7240**	**0.6362**	**0.5533**	0.4325	**0.6102**
(再実験)ResNet152	0.6089	0.5955	0.6012	0.3821	0.2809	0.4937
ResNet18	0.4147	0.3816	0.4194	0.3365	0.1847	0.3474
EfficientNet-B0	0.2788	0.4643	0.3777	0.2273	0.2752	0.3247
EfficientNet-B7	0.4860	0.6069	0.5179	0.3982	**0.5599**	0.5138
SwinTransformerB	0.3819	0.3367	0.5007	0.3156	0.3649	0.3800

まとめ

特徴抽出器の違いによる検出精度の変化を調査

- 複数フレームを用いた物体検出手法の特徴抽出器を変更
- 各特徴抽出器の物体検出精度を比較

田邊の手法でも用いられていたResNet152が優れていた

今後の課題

- データ拡充による精度向上
- ネットワーク深度の最適化
- SwinTransformerの入力解像度の統一