スマホ撮影画像への応用を見据えたナンバープレート検出モデル性能に関する予備調査

宮木　笙伍† 　河合 由起子†† 　栗　達†

† 京都産業大学情報理工学部　〒 603-8555 京都府京都市北区上賀茂本山
†† 関西大学　〒 565-8585 大阪府吹田市山田南 50-2
E-mail: †{g2354558, lida}@cc.kyoto-su.ac.jp, ††ykawa@kansai-u.ac.jp

あらまし　近年, スマホによる SNS などへの風景画像投稿の増加に伴い, 人物や居住場所, ナンバープレートなど個人を特定し得る情報は, データ収集・分析・可視化において重要なリスク要因となる. 一方, 交通管理や防犯分野において, 定点カメラによる車両情報の自動取得を目的としたナンバープレート検出システムの重要性が高まっている. 本研究では, スマホによる撮影画像に対するナンバープレート検出システム構築に向けた予備検討として, 定点カメラによる公開データセットとスマートフォン撮影画像データセットを用い, 物体検出モデルによるナンバープレート検出性能の比較を行った. なお, 都市回遊促進のためのデータ収集を目的とした研究の一部として, 実環境で取得される画像データに含まれるナンバープレートなどのプライバシー情報を自動的に検出する手法の予備調査に位置づけられる. 比較対象として,YOLO シリーズに属する YOLOv11 および YOLO-World, ならびに Transformer 系モデルである RT-DETR の 3 手法を採用した. 各モデルは,Epoch 数 100, 入力画像サイズ 640 などの統一した学習条件下で学習および評価を行い,Precision（適合率）, 再現率（Recall）,F1 スコア,mAP および推論速度を指標として性能を比較した. さらに, 視覚-言語モデルである YOLO-World に対しては, テキストプロンプトの変更が検出精度に与える影響を検証する追加実験も実施した. その結果,YOLO-World は Precision 0.8155，mAP@50-95 で 0.5197 を記録し, 高精度な検出性能を示した.RT-DETR は F1 スコアおよび Recall で最高値を示した. しかし YOLO 系と比較しリアルタイム性に課題が残る結果となった.YOLOv11 は推論速度において最も優れていたが,mAP および F1 スコアでは他のモデルに及ばなかった. また,YOLO-World におけるプロンプトの変更による顕著な精度向上は確認されなかった. 本予備実験では, 学習データが少ない場合には YOLO-World が, 見逃しを減らすことを重視する場合には RT-DETR が有効であることが示された.

キーワード　物体検出, 機械学習, ナンバープレート検出

1 はじめに

スマートフォンの普及により, ほとんどすべての国民が SNS を利用できるようになった. 総務省の調査などをもとにした 2025 年のデータによれば, 日本国内の SNS 利用者は総人口の約 78.1%に当たる約 9,600 万人に達しており, インターネットユーザー全体で見てもその普及率は極めて高い [1]. 特に, 画像や動画を主体とする「ビジュアルコミュニケーション」の拡大が著しい. 月間アクティブユーザー数 (MAU) 7,370 万人を擁する YouTube や, 若年層を中心に利用率が高い Instagram(MAU 5,545 万人以上),TikTok(MAU 2,600 万人以上) など, 視覚的情報を共有するプラットフォームが日常的に利用されている. また, スマートフォンのカメラ性能は, 近年劇的な進化を遂げている. 最新のハイエンドモデルでは,2 億画素を超える超高解像度センサーや, 多くの光を取り込める 1 インチ大型センサーの搭載が進んでおり,「一眼レフ並み」と評される水準にある [2]. これらの進化は, 撮影者が意図した被写体だけでなく, 撮影に写り込んだ情報までも極めて鮮明に記録されることを意味する.SNS 上の膨大な画像データにおいて, こうした高精細な映り込み情報はプライバシー保護の観点から看過できないリスク要因となっており, 自動的かつ高精度にこれらを検出し, 適切に処理する技術の重要性が高まっている. 特に, 車両のナンバープレートは, 所有者の特定や移動履歴の把握に直結するため, プライバシー保護において極めて重要な情報である. 交通管理や犯罪捜査の現場では, 定点カメラを用いた自動ナンバープレート認識 (LPR) システムの導入が進んでいるが, これらの既存技術が, 撮影条件の多様なスマートフォン画像に対しても同様に機能するかは自明ではない. 手振れ, 角度, 照明条件が一致しないスマートフォン画像における検出性能を明らかにすることは, プライバシー保護技術の向上に寄与するものである.

そこで本研究では, スマートフォンによる撮影画像に対する検出システム構築に向けた予備検討として, 定点カメラによる公開データセットを用いた物体検出モデルの比較評価を行う. 具体的には, 最新の YOLO シリーズである YOLOv11 及び YOLO-World, さらに Transformer ベースのアーキテクチャを持つ RT-DETR の 3 手法を対象とする. これらを統一された条件下で学習・評価し,F1 スコアや mAP などの客観的指標に基づいて各モデルの検出特性を分析する. 本検証により, 各アーキテクチャの長所と短所を明らかにし, 実環境におけるナンバー

プレート検出に最適なモデルの選定の指標を示す.

本論文は以下の構成とする. 2 章では関連研究を紹介する. 3 章では物体検出モデルを比較するための実験設定を述べる. 4 章では実験結果を示し, 定量的および定性的な評価を行う. 最後に 5 章で本論文のまとめと今後の課題について述べる.

2 関連研究

本章では, 関連研究として物体検出モデルに関する研究とナンバープレート検出に関する研究を紹介する.

2.1 リアルタイム物体検出モデルの進化

Sapkota ら [3] は,Ultralytics YOLO シリーズ（v5,v8,11,26）のアーキテクチャの進化と性能評価に関する包括的なレビューを行っている. 彼らは,YOLOv8 におけるアンカーフリー予測と分離型ヘッドの導入,YOLO11 における C3k2 モジュールによる特徴抽出の効率化, そして最新の YOLO26 における NMS（Non-Maximum Suppression）および DFL（Distribution Focal Loss）の削除による推論の高速化とエッジデバイスへの適合性向上について詳述している

一方,Zhao ら [4] は,YOLO シリーズが NMS による後処理に依存している点が速度と精度のトレードオフに悪影響を与えていると指摘し, リアルタイム・エンドツーエンド物体検出器である RT-DETR を提案している. この研究では, マルチスケール特徴を効率的に処理するハイブリッドエンコーダと, 不確実性を最小化するクエリ選択手法を導入することで,RT-DETR が同規模の YOLO モデル（YOLOv5,v8 など）を速度と精度の両面で上回ることを実験により示している

2.2 オープン語彙物体検出への拡張

従来の物体検出器が固定されたカテゴリに限定されるという課題に対し,Cheng ら [5] は,YOLO-World と名付けたリアルタイム・オープン語彙物体検出器を提案している. 彼らは, 視覚と言語の情報を融合するための RepVL-PAN（Re-parameterizable Vision-Language Path Aggregation Network）と, 領域テキスト対照損失を用いた大規模な事前学習スキームを導入した. これにより, 推論時にはテキストエンコーダを再パラメータ化することで計算コストを削減しつつ, ゼロショットで多様な物体を検出できることを確認し,LVIS データセットにおいて高い精度と FPS を達成している

2.3 特定タスクへの応用：ナンバープレート検出

Fu [6] は, ナンバープレート認識（LPR）における深層学習技術の適用について調査を行っている.CNN（特に YOLO シリーズ）を用いた検出が主流である一方で, 文字認識における RNN（LSTM や GRU）の活用, 低解像度画像の超解像やデータ拡張における GAN や Diffusion Model の利用. そして大域的な特徴抽出に優れた Transformer の導入など, 各モデルのアーキテクチャごとの利点と課題を比較・整理している

2.4 関連研究のまとめと本調査の位置づけ

これらの知見を踏まえると, 物体検出技術は, 高速・軽量な CNN ベースのモデル,Transformer を用いたエンドツーエンド型モデルおよび視と言語を統合した事前学習モデルへと多様化していることが分かる.

一方で, これらのモデルが定点カメラで学習された条件からスマートフォンによる撮影画像のような撮影条件が大きく異なる環境へ適用された場合の検出性能の違いについては, 十分に検証されていない.

そこで本調査では, 事前学習の有無やモデルアーキテクチャの違いに着目し, 定点カメラ画像で学習したモデルを用いて, スマートフォン撮影画像に対するナンバープレート検出性能を比較評価する. 本検証により, ドメイン乖離環境下における各モデルの特性を明らかにすることを目的とする.

3 実験設定

本研究では, 各モデルの性能を公平に比較するため, 統一されたデータセット・学習条件下で実験を行った.

3.1 使用データセット

本実験で使用するデータは, ネガティブ画像（ナンバープレートなし）, ポジティブ画像（ナンバープレートあり）, および評価用のスマートフォン撮影画像の 3 種とする. 定点カメラで撮影された車両画像・ナンバープレート画像と, スマートフォンで撮影された日常的な風景画像の異なるドメイン間におけるモデルの検出性能を評価するため, 学習用と評価用で異なるデータセットを使用した.

3.1.1 学習用データセット

学習用データセットの構築にあたり, ポジティブ画像として Roboflow Universe で公開されている **License-plate-japan dataset**[1] および **Number Plate in Japan dataset**[2] を使用した（元データでポジティブ画像 1000 枚）. ネガティブ画像には, 背景画像として一般的によく用いられる **Microsoft COCO**[3] を採用した. ポジティブ画像に対して回転（90°, 180°, 270°）および反転（上下・左右）のデータ拡張を行い, 拡張後のポジティブ画像 6,000 枚とネガティブ画像 600 枚を用いて, 総枚数 6,600 枚のデータセットを作成した. ネガティブ画像と拡張済みポジティブ画像により,YOLOv11,YOLO-World,RT-DETR の 3 モデルに対してファインチューニングを行った.

3.1.2 評価用データセット

実環境での適用を想定し, 独自に収集したスマートフォンを用いて撮影されたデータセットを評価用として採用した. これには手ブレや多様な照明条件が含まれており,Positive(ナンバー

1 : [5] Roboflow Universe – license-plate-japan_1. https://universe.roboflow.com/new-workspace-vijtn/license-plate-japan_1 (accessed 2026-01-25).

2 : [6] Roboflow Universe – Number Plate in Japan. https://universe.roboflow.com/moriken/number-plate-in-japan (accessed 2026-01-25).

3 : [7] Microsoft COCO: Common Objects in Context. https://cocodataset.org/ (accessed 2026-01-25).

プレートあり)52 枚,Negative(なし)52 枚の計 104 枚で構成される.

3.2 比較モデルと学習条件

比較対象として, 実験当時の物体検出モデルである YOLOv11,YOLO-World, および Transformer ベースの RT-DETR の 3 種類を選定した. 学習は Google Colab 上の L4 GPU 環境で実施し, ネガティブ画像と拡張済みポジティブ画像を用いたファインチューニングを行った. ハイパーパラメータは Epoch を 100,Image Size を 640 に統一し, それ以外のハイパーパラメータは各モデルのデフォルト設定を採用した. 実験環境のハードウェア構成を表 1 および表 2 に, 各モデル共通の学習ハイパーパラメータを表 3 に示す.

表 1 学習環境

Item	Specification
GPU	NVIDIA Tesla L4
Platform	Google Colab
Framework	Ultralytics 8.3, PyTorch 2.x

表 2 推論環境 (M2 MacBook Air)

Item	Specification
Chip	Apple M2 (8-core GPU)
Memory	8 GB Unified Memory
Platform	macOS
Framework	Ultralytics 8.3, PyTorch 2.x (MPS)

表 3 学習ハイパーパラメータ

Parameter	Value
Input Image Size	640 × 640
Batch Size	16 (Default)
Epochs	100
Optimizer	Auto (SGD)
Comparison Models	
YOLOv11	`yolo11s.pt` (Small)
YOLO-World	`yolov8s-world.pt` (Small)
RT-DETR	`rtdetr-l.pt` (Large)

3.3 評価指標

性能評価には, 適合率 (Precision), 再現率 (Recall),F1 スコア, および mAP(mean Average Precision) を用いた. また, 実用性の観点から推論時間も合わせて計測した.

4 実験結果

表 4 に, 各モデルの評価結果を示す.

4.1 定量評価

実験の結果, **YOLO-World** が mAP@50 において 0.6980,Precision において 0.8155 と最も高い値を示し, 最も高精度な位置検出が可能であることが確認された. 一方, **RT-DETR** は F1 スコア (0.7271) および Recall(0.6731) で最高値を記録し見逃しが最も少なかったが, 推論時間が 523.4 ms と他の YOLO 系モデルの約 6–8 倍となり, リアルタイム性には課題が残る結果となった. **YOLOv11** は推論速度が最速（65.4 ms）であったものの,mAP および F1 スコアでは他の 2 モデルに及ばなかった.

4.2 定性評価

各モデルの実際の検出結果を図 1 に示す. 定量評価では YOLO-World が高い適合率を示したが, 実際の検出画像を確認すると, 全てのモデルにおいて共通の傾向が見られた. 正面で大きく写る対象では全モデルが検出に成功した一方, 看板などでは誤検出が, 小さい対象では見逃しが確認された. また, 明るさや大きさが十分でブレの少ない画像では全モデルが正確に検出できる一方,「強い手ブレ」などが含まれる画像では, どのモデルも検出に失敗するか, あるいは看板の文字を誤検出するケースが確認された. これは, 学習データと評価データの間で画像のドメインが乖離しており, モデルのアーキテクチャの差ではなく, 特徴抽出そのものが困難であったためと考えられる. ただしその中でも,YOLO-World はバウンディングボックスの追従性が比較的良好であり, 他のモデルよりもロバストな挙動を示した.

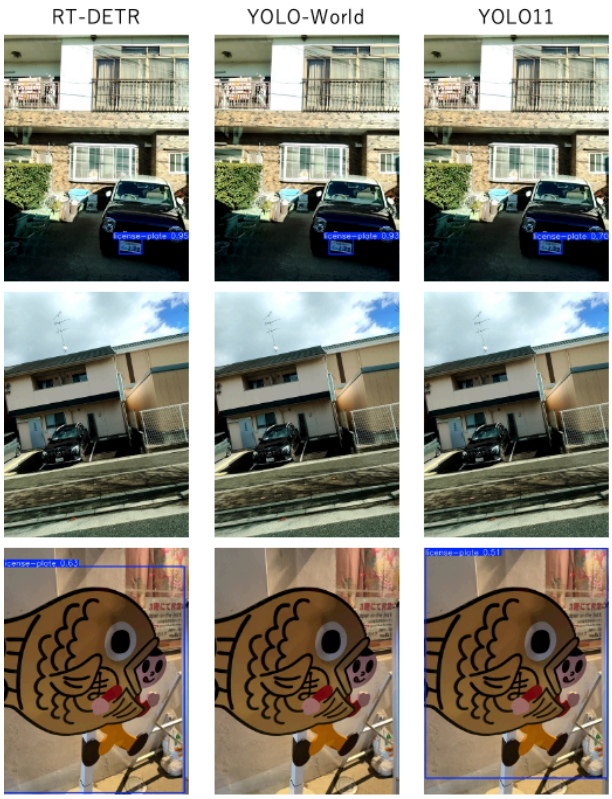

図 1 スマートフォン撮影画像に対する各モデルの検出結果比較

表 4　各モデルの性能比較（最高値を太字で示す）

Model	Precision	Recall	F1 Score	mAP@50	mAP@50-95	Inference (ms)
YOLOv11s	0.7980	0.5000	0.6148	0.5994	0.4343	**65.4**
YOLO-World	**0.8155**	0.5962	0.6888	**0.6980**	**0.5197**	81.0
RT-DETR	0.7905	**0.6731**	**0.7271**	0.6668	0.4653	523.4

4.3　追加実験（YOLO-World のプロンプト比較）

YOLO-World は学習データに含まれないクラスをテキスト（プロンプト）で指定して検出できる. そこで, プロンプトの違いが検出精度に与える影響を比較する追加実験を行った. 使用したプロンプトとそれぞれの意図を表 5 に示す. ファインチューニング済み YOLO-World に対し, スマートフォン撮影画像を入力として, 各プロンプトで推論した.

表 5　追加実験で使用したプロンプトとその意図

プロンプト	意図・特徴
`license plate`	最もシンプル
`vehicle license plate`	vehicle 追加により誤検出低減の可能性
`license plate on a vehicle`	空間制約による高精度化の可能性

表 6 に, プロンプト別の評価結果を示す. プロンプトを変更しても,Precision,Recall,F1,mAP@50,mAP@50-95 のいずれも大きな差はなく, プロンプト変更による精度の差は小さいと結論づけられる.

表 6　プロンプト別の検出性能（追加実験）

Prompt	Precision	Recall	F1	mAP@50	mAP@50-95
license plate	0.8141	0.5962	0.6888	0.6995	0.5229
vehicle license plate	0.8340	0.5798	0.6840	0.7031	0.5261
license plate on a vehicle	0.8342	0.5808	0.6848	0.6968	0.5192

4.4　考　　察

4.4.1　視覚-言語モデルによる汎化性能

実験結果より,YOLO-World が mAP および Precision において最も優れた性能を示した. この要因として学習手法の違いが挙げられる可能性がある.YOLOv11 や RT-DETR が画像のみから特徴を学習するのに対し,YOLO-World は画像とテキストのペアを用いた事前学習を行なっている. 一般に, 言語情報を補助タスクとして学習されたモデルは, 物体の意味的な概念をより深く獲得し, 未知のドメインや小規模なデータセットに対する**汎化性能（Generalization capability）**の向上に寄与すると考えられている. 本実験のような「定点カメラで学習し, スマートフォン画像で推論する」というドメインの乖離がある環境において,YOLO-World が持つ意味的な理解が, 未知の特徴に対する安定性として寄与したことが示唆される. また, 車両検出の後にナンバープレート検出を行う二段階構成とすれば, 改善の余地があると考えられる. 追加実験（4.3 節）では,`license plate` への適合が強く,`vehicle` を追加しても精度の改善は見られなかった.

4.4.2　データ量とモデル構造

一方,RT-DETR は YOLO 系モデルと比較して推論速度が遅く, 精度面でも YOLO-World に及ばなかった. これは RT-DETR の性能限界というよりも, 学習データ不足に起因する可能性が高い.Transformer ベースのモデルは,CNN ベースのモデルと比較して帰納的バイアスが弱い傾向にあり, 高い性能を発揮するにはより大規模なデータを必要とすることが多くの研究で指摘されている. 本実験のデータ規模ではドメインの乖離を解決できなかったと考えられる. YOLOv11 はシンプルなアーキテクチャにより推論速度が最速であり, したがって小規模データセットでの運用やリアルタイム性が求められる環境では YOLO シリーズ（特に YOLO-World）が適しているが, データセットの拡充が可能であれば,RT-DETR の採用も再考の余地がある.

5　ま　と　め

本研究では, スマートフォンによる撮影画像を対象としてナンバープレート検出システムの構築に向けた予備実験として, 最新の物体検出モデル 3 種の比較評価を行った. 6,000 枚の中規模データセットを用いた実験の結果, 視覚-言語事前学習を取り入れた **YOLO-World** が,Precision で 0.8155 および mAP@50-95 で 0.5197 を記録し, 最も高い検出性能と汎化能力を示した. 結論として, **学習データが少ない場合には YOLO-World が, 見逃しを減らすことを重視する場合には RT-DETR** が有効であることが示された.

今後の課題として, 以下の 5 点が挙げられる. 第一に, **車両検出の後にナンバープレート検出を行う二段階構成**の検討である. 第二に, **モバイル端末での性能評価**である. 実用化に向けては精度だけでなく, スマートフォン（エッジデバイス）上でのリアルタイム動作が求められる. 本研究では実装の安定性を重視し YOLOv11 を採用したが, 今後のモバイル実装フェーズにおいては, 最新の YOLO26 を含む複数の軽量モデルを対象に, 処理速度とバッテリー消費の観点からも比較検討を進める予定である. これに併せて, モデルの軽量化（量子化や枝刈り）と実機での遅延評価を行う. 第三に, **スマートフォン撮影画像を用いたファインチューニング**の検討である. 第四に, **エッジデバイス向け検出パイプライン**の検討である.

謝　　辞

本研究の一部は, 京都産業大学先端科学技術研究所（人間情報学研究センター）共同研究プロジェクト（M2301）の助成を受けたものである. ここに記して謝意を表す.

文　　献

[1]　いいね AI. 日本の主要 sns プラットフォーム徹底分析: 2025 年

最新データから見る利用動向, 2025. Accessed: 2026-01-04.
[2] 株式会社ノジマ 家電小ネタ帳編集部. 【一眼レフ並み！】カメラ性能が高いスマホをランキング形式でご紹介. Accessed: 2026-01-04.
[3] Ranjan Sapkota and Manoj Karkee. Ultralytics YOLO evolution: An overview of YOLO26, YOLO11, YOLOv8, and YOLOv5 object detectors for computer vision and pattern recognition. *arXiv preprint*, October 2025.
[4] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10, 2024.
[5] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–15, 2024.
[6] Zhuoqun Fu. Deep learning for license plate number recognition: A survey. In *2024 IEEE International Conference*, pages 1–6. IEEE, 2024.