車種別交通量計測AIを対象としたドメイン適応に基づく自己学習アルゴリズムの提案

尾原　健太郎[1]・柳沼　秀樹[2]・寺部　慎太郎[3]・海野　遥香[4]・鈴木　雄[5]

[1]学生非会員　東京理科大学大学院　創域理工学研究科社会基盤工学専攻（〒 278-8510 千葉県野田市山崎 2641）
E-mail: 7623507@ed.tus.ac.jp
[2]正会員　東京理科大学准教授　創域理工学部社会基盤工学科（〒 278-8510 千葉県野田市山崎 2641）
E-mail: yaginuma@rs.tus.ac.jp (Corresponding Author)
[3]正会員　東京理科大学教授　創域理工学部社会基盤工学科（〒 278-8510 千葉県野田市山崎 2641）
E-mail: terabe@rs.tus.ac.jp
[4]正会員　茨城大学大学院助教 理工学研究科（〒 316-8511 茨城県日立市中成沢町 4-12-1）
E-mail: haruka.uno.uj38@vc.ibaraki.ac.jp
[5]正会員　東京理科大学助教　創域理工学部社会基盤工学科（〒 278-8510 千葉県野田市山崎 2641）
E-mail: yusuzuki@rs.tus.ac.jp

近年，深層学習を活用した AI の導入により，自動車交通量観測の自動化が進められている．しかし，特定の観測地点で計測精度が著しく低下するという課題が存在する．精度改善のためには，学習データの追加が効果的だが，作業コストは非常に大きい．そこで，本研究では地点毎に生じる追加学習データの作成コスト低減と精度向上を目的とした AI モデルの自己学習手法を提案する．さらに，複数地点の CCTV 映像を用いて提案手法の性能評価に取り組む．提案手法は，ドメイン適応に基づいて学習データと推論データの分布距離を近づけるように，観測地点の背景画像に検知対象の物体情報を合成する自己学習アルゴリズムを提案する．複数地点の CCTV 映像で計測精度を検証した結果，人手によるアノテーションを必要とせずに精度向上が確認された．本提案手法により，コストを抑えた地点別の計測精度向上が可能となり，実務的な課題解決に大きく寄与する．

Key Words: *Traffic count, Convolutional Neural Network (CNN), Self-Supervised Learning, Domain Adaptation*

1. はじめに

近年，深層学習を活用した AI の導入により，自動車交通量観測の自動化が進められている．交通分野での実務的な利用を想定すると，車種分類の細分化や日本でいう軽自動車などの国ごとの独自規格に対応することが求められる．そのため，AI の再学習が不可欠だが大量の学習画像の作成といった作業が生じる．学習画像の作成においては，異なる地点や状況からの画像を多く用意することが AI モデルの性能向上に有効であるが，学習画像の作成における人的コストが大きい．また，大量の学習画像を作成したのにも関わらず，一部の地点では想定した性能を大きく下回ると言った事象が AI モデルには存在し，実用上の課題となっている．このような課題は、観測地点の状況と学習データの分布が一致しないことから生じ、いわゆる「ドメインの不一致」と呼ばれる．主な原因は，学習画像に含まれる対象物体や背景画像のバリエーションの不足である．しかしながら，学習を行うにあたっての十分な画像枚数などが示されていない点が現状の課題である．

本研究では背景画像のバリエーションを増加させることに注目し，背景画像に対してドメイン適応が実施されるような学習画像の生成手法を提案する．背景画像に対するドメイン適応の効果については，Grad-CAM(Gradient-weighted Class Activation Mapping) という CNN の予測の根拠について示す手法を用いて検証する．加えて本手法では，学習画像生成においてインスタンスセグメンテーションなどの手法を取り入れることで人的コストを低減した自己学習システムの提案を目指す．

2. 本研究で用いる理論とモデル

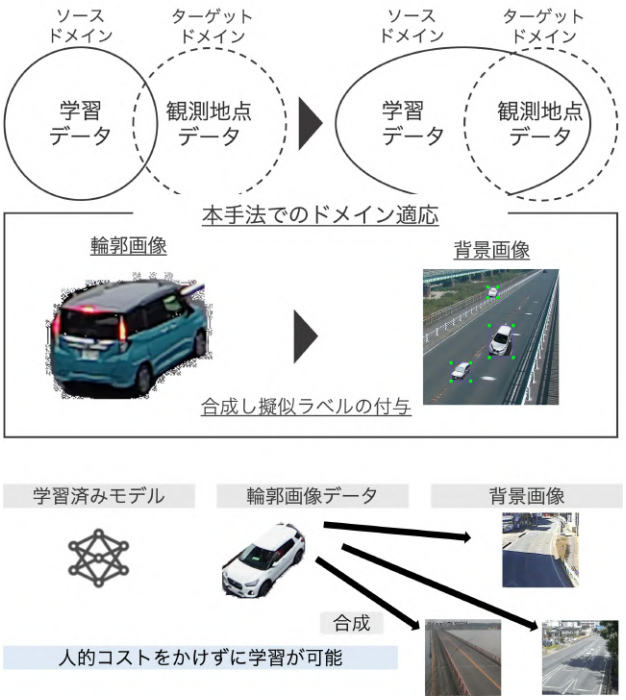
(1) 本研究で用いる理論

a) ドメイン適応

機械学習モデルは，学習データと観測データの分布に差が生じている場合，適切な性能を発揮できない場合がある．通常の機械学習においては，観測データであるターゲットドメインと学習データであるソースドメインの分布が同様と仮定するが，ドメイン適応の考えではターゲットとソースの分布が異なるものと仮定する．そのため，ターゲットの分布に近づけるように学習を進め，推論の性能を高めようとする手法がドメイン適応 (domain adaptation) である．

ソースドメインのデータが正解ラベル付きのデータ $\mathcal{D}_s = \{(x_n, d_n)\}_{n=1,\ldots,N_s}$，ターゲットドメインのデータが正解ラベルのないデータ $\mathcal{D}_t = \{(x_n)\}_{n=1,\ldots,N_t}$，といった条件の場合の問題を，教師なしドメイン適応 (unsupervised domain adaptation, 以下 UDA) と呼ぶ．本研究においては，観測地点であるターゲットのラベルを作成することは現実的ではないことから UDA の問題に対処していくこととする．

ドメイン適応にもいくつか種類が存在し，入力 $\mathbf{x}$ の空間でのソースドメインの分布 $p_s(\mathbf{x})$ とターゲットドメインの分布 $p_t(\mathbf{x})$ の分布の位置を合わせることを目的とした手法，ターゲットドメインのデータ $\mathcal{D}_t$ に擬似ラベルを作成してモデルを学習させる手法がある．前者の分布の位置を合わせる手法は Tzeng ら[1] により提案された．特徴抽出用のネットワーク $\mathbf{g}(\mathbf{x})$ と，入力 $\mathbf{x}$ がソースとターゲットのどちらのドメインであるかを判定する識別器 $h(\mathbf{g}(\mathbf{x}))$ を考える．ソースの分布の特徴 $p_s(\mathbf{g}(\mathbf{x}))$ とターゲットの分布の特徴 $p_t(\mathbf{g}(\mathbf{x}))$ に差がなければ，どんな高性能な識別器 h を使用してもドメインの識別は難しい．よって，高性能な識別器 h でも判別できない場合，二つの分布は同様だといえるため，h と g を互いに競争させながら学習させる．一方，後者の擬似ラベルを作成する手法は，ソースドメインのデータ $\mathcal{D}_s$ を学習させたモデルをターゲットドメインのデータ $\mathcal{D}_t$ に適用し，ラベルが付与された $\mathcal{D}_t$ 使用してモデルの再学習を行う手法である．この手法は半教師あり学習の構造に類似しており，三本ら[2] は，果実の位置推定というタスクにおいて，COCO データセットを用いた学習済みモデルを使用してラベルが付与されていない画像に擬似ラベルを付与した．擬似ラベルが付与された画像を学習データとして学習を実施した結果，精度の改善が確認された．しかしながら，この手法は学習済みモデルの精度に大きく依存し，擬似ラベルの精度が低いとモデル学習の際のノイズとなりうる点が課題である．そこで，擬似ラベルの精度が担保される手法のひとつとして合成画像データの使用が挙げられる．Richter ら[3] は，正確な正解ラベルを大量に得るために GTA5 というゲームでが合成画像の生成を行った．GTA5 とは，都市空間が再現可能なオープンワールドコンピュータゲームである．ゲームの画面出力結果をセグメンテーションの位置ラベルとして使用することで，低コストで学習データを作成することに取り組んだ．自動車の車載カメラによって撮影された学習データである，CamVid というデータセットに合成画像を加えたモデルの精度比較を実施した．その結果，CamVid のデータのみで学習されたモデルと比較して，合成画像を加えたモデルのセグメンテーションにおける精度が上昇した．



図–1 本手法でのドメイン適応の流れ

本研究でのドメイン適応の流れを**図–1** に示す．本研究のドメイン適応では，ターゲットドメインが観測地点地点と明確であるため，ターゲットドメインのデータに擬似ラベルを付与する手法を取り入れる．具体的には，検出対象物体の輪郭情報が取得可能なインスタンスセグメンテーションの学習済みモデルを使用する．ここでカメラの画質や設置条件などが良い地点を使用することによって，取得する対象物体の輪郭情報の精度を担保する．検出対象物体の輪郭情報取得後，観測地点の画像に輪郭情報を用いて対象物体の合成を行う．物体の合成位置についてはこちらで指定可能であるため，擬似ラベルの位置情報の精度が悪化せず学習時のノイズにならない点が本手法の最大の利点である．また，実際の観測地点の画像を使用しているため，デジタル生成画像と現実画像との質感の違いなどを考慮せずに済むため，モデルの精度改善に寄与すると考えられる．

本手法の想定される利用方法について全体の流れについて説明する．はじめに，可能な限り多くの地点の多くの時間帯の学習データを作成し基本となるモデルの学習を行う．その後，学習後の基本となるモデルを用いて観測対象地点について推論を行い，問題が無い地点については基本モデルを使用して交通量計測を実施する．検出精度が悪い地点については，基本モデルで用意した画像に対象地点の合成画像を加えて再学習させ地点別モデルの作成を実施することで物体検出精度の向上を試みる．



図–2 過剰な bounding-box の検出



図–3 ドメイン適応に基づいた合成画像

(2) 本研究で用いるモデル

本手法の基準となる物体検出のモデルは YOLOv7[4] を使用する．YOLOv7 は 2022 年に発表された YOLO モデルであり，畳み込みニューラルネットワーク (Convolutional Neural Network, 以下 CNN) を使用した深層学習ベースのモデルである．高速で物体検出が可能な点が本モデルの特徴である．また，擬似ラベル作成のための物体の輪郭情報の取得は Mask-R-CNN というインスタンスセグメンテーションモデルを使用した．インスタンスセグメンテーションとは，物体検出に加えて物体の形状まで抽出し，画像をピクセル単位で分類を行うタスクである．

本提案手法の評価方法の一部には，CNN の予測の根拠について示す Grad-CAM[5][6] というモデルを用いる．通常の物体検出タスクにおいては，画像データセットを使用してモデルの評価を行うことが一般的であるが，本研究においては観測地点の正解データは存在しない．そのため，Grad-CAM を使用した出力結果を評価項目に加えることとする．

3. 本研究で用いる学習データ

基本モデルの開発のために学習データの作成を実施した．学習データについては，4 地点で 1455 枚の画像のデータの作成し，作成した学習画像を用いて YOLOv7 の学習を行った．その結果，一部地点においては**図–2** に示すように画像全体に過剰に bounding-box が検出されてしまい，検出精度が悪い状況が確認された．このような検出状況では，車種分類のみならず交通量の計測さえも困難である．そのため，同様の検出結果となった 3 地点においてドメイン適応に基づいた合成画像を作成し，擬似ラベルの付与を行った．検出対象物体の合成位置については，物体が道路上に存在するように位置を設定し，物体の大きさについても周囲の物体の大きさと一致するように調整を行った．作成した画像の例を**図–3** に示す．

表–1 学習画像の詳細

データの種類	地点名	画像サイズ	枚数
通常学習画像	千住大橋	1920×1080	298
	藤代大橋	1920×1080	472
	城北	1920×1080	322
	京都	1920×1080	363
合成学習画像	湯舟橋	1280×720	28
	田之倉	1280×720	28
	神宮橋	1920×1080	63

また，今回作成したデータセットは**表–1** に示す通りである．合成画像を作成した 3 地点においては検出精度が悪い地点を選択したが，通常の学習画像の地点と比較してカメラの解像度が低いという特徴がある．

4. 合成画像を用いた学習の結果と考察

(1) 自己学習による物体検出精度の改善

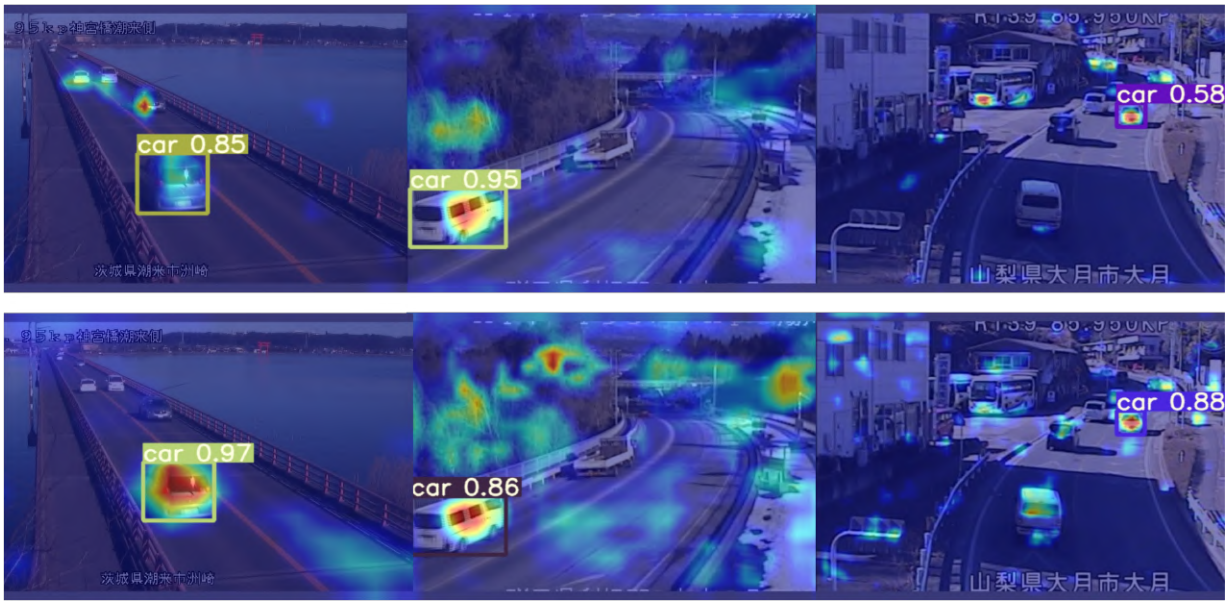
本手法による物体検出精度の向上を検証するために，通常の学習画像のみ使用したモデルと通常の学習画像に合成画像を組み合わせたモデルの処理結果を比較した．いずれのモデルも YOLOv7 の学習済みモデルについて，バッチ数 16，エポック数 300 で backbone モデルの入力に近い 50 層を固定した転移学習を行っている．推論結果を**図–4** に示す．通常の学習画像のみを使用したモデルは，物体が存在しない部分に対しても過剰に検出されていることが確認できる．一方，合成画像を加えたモデルでは，過剰な検出が改善され．対象物体のみが検出されている．このように観測地点の背景画像を使用した合成画像を加えることで，計測精度が改善されることが視覚的に確認できた．

(2) CNN の判断根拠においての背景画像の影響

本手法では，ターゲットである観測地点の背景画像に対象物体を合成することでドメイン適応を行った．観



図–4 推論結果の比較 (通常学習画像のみ (上) 通常学習画像と合成学習画像 (下))



図–5 Grad-CAM による出力結果 (通常学習画像のみ (上) 通常学習画像と合成学習画像 (下))

測地点の背景画像を学習させることで物体検出精度が向上することが示されたが，CNN の判断根拠として背景画像が与えている影響を Grad-CAM を用いて検証を行う．Grad-CAM による出力結果を**図–5** に示す．出力結果より確認できることとして，合成画像を加えたモデルは通常画像のみのモデルと比較して，ヒートマップが反応している範囲が拡大していることが挙げられる．このような結果より，CNN は物体のみではなく背景画像までも判断根拠として用いていることが確認され，背景画像においてのドメイン適応が適切に実施されたと考えられる．しかしながら，Grad-CAM による検証は定性的な判断になってしまう点が課題である．

そこで，定量的に背景画像の影響を判断する方法として，合成画像にのみ通常のラベルとは異なる別ラベルを付与して学習させ，合成画像の背景画像として使用した地点と使用していない地点の推論結果を比較した．今回，別ラベルを dummy という名前とし，推論後のクラス分類において dummy ラベルが占める割合を**表–2** に整理した．分類割合より，dummy ラベルは背景画像として使用していない地点に関してはほとんど存在しない一方で，背景画像として利用している地点に関しては，半分以上が dummy ラベルとなった．この結果は，CNN が物体のみではなく背景も判断根拠として含めている可能性が高いことを示しているため，基本

表–2 学習画像の詳細

背景画像の利用	地点名	画像サイズ	dummy ラベルの割合 (%)
背景画像として利用していない	千住大橋	1920×1080	0.208
	藤代大橋	1920×1080	0.156
背景画像として利用	湯舟橋	1280×720	83.9
	田之倉	1280×720	59.2
	神宮橋	1920×1080	1.88

となる学習データの収集方法などについて再度検証する必要がある．しかしながら,「神宮橋」など背景画像が学習画像に含まれているのにも関わらず，dummy ラベルの割合が低い地点も存在するため，完全な CNN の解釈に至っていない点については，今後の課題である．

5. おわりに

本研究では，背景画像に対してドメイン適応用いた自己学習手法を取り入れることで検出精度の改善を達成した．今回は，通常学習画像の学習のみでは検出精度が悪い地点において合成画像の作成を行った．検出精度が悪かった地点の特徴として，通常の学習データを作成した地点と比較して CCTV 画像の画質が低いといった点が挙げられる．今後は，画質が悪い CCTV カメラの入れ替えが進むことでこのような課題の解消が期待されるが，現在国土交通省が管理している CCTV カメラのみでも数千台存在するため，入れ替えにはある程度の期間が必要になることが想定される．そのため，画質が悪い CCTV においても交通量計測については一定の精度を確保可能な点が本手法の有用性といえる．

本手法では，通常学習画像に対象地点の合成画像を加えることで，通常の学習データで作成した車種分類の特徴を引き継ぎつつ物体検出精度の改善が可能になるといった特徴がある．また，検出精度が悪い地点について人的コストをかけずに物体検出精度の向上が可能になることが，実務利用を想定した際の大きな利点になると考えられる．

今後の課題としては，まず車種分類の精度向上が挙げられる．本手法は対象地点の画像を含めて学習することで，背景画像についてのドメイン適応を行ったが，対象物体の特徴のバリエーションの分布は考慮できていない．今後は，合成する対象物体のバリエーションを持たせることへの効果や，十分な学習のために必要な合成画像枚数などについても検証する必要がある．また，国内全ての CCTV への適用を見据えると，再学習させるモデルの個数についても検証が必要である．本研究では合成画像を作成した地点が 3 地点と少なかったため，単一のモデルで対処できていたが国内全ての観測地点への適用を目指すとなると複数個モデルを作成することが想定される．今後は，観測地点のクラスタリングなどを実施した上での各クラスタごとのモデルの学習が必要になると考えられる．

謝辞

本研究は国土交通省新道路技術会議「カメラ画像および複数の観測データを融合した次世代交通計測手法に関する研究開発」の助成を受けたものです．ここに記して感謝致します．

参考文献

1) Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T.: Adversarial discriminative domain adaptation, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.
2) 三本拓未, 四宮友貴, and 吉田真一: 果実位置検出における半教師あり深層学習とデータ拡張の適用, **日本知能情報ファジィ学会 ファジィ システム シンポジウム 講演論文集 第 *37* 回ファジィシステムシンポジウム**, pp. 241–244, 日本知能情報ファジィ学会, 2021.
3) Richter, S. R., Vineet, V., Roth, S., and Koltun, V.: Playing for data: Ground truth from computer games, *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 102–118, Springer, 2016.
4) Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7464–7475, 2023.
5) Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization, *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.
6) 芳賀柚希, 柳沼秀樹, 寺部慎太郎, 海野遥香, and 鈴木雄: 説明可能な ai を活用した cnn 型交通計測モデルの戦略的高度化, **土木計画学研究・講演集** *(CD-ROM), 2022*, 土木計画学会, 2022.

(Received June 16, 2023)
(Accepted ?)

Proposal of Self-Learning Algorithm Based on Domain Adaptation for Vehicle-Specific Traffic Measurement AI

Kentaro OBARA, Hideki YAGINUMA, Shintaro TERABE, Haruka UNO and Yu SUZUKI

In recent years, the introduction of AI using deep learning has led to the automation of vehicle traffic volume measurement. However, there is a problem that measurement accuracy is significantly lower at certain observation points. To improve accuracy, it is effective to add learning data, but the cost is very high. Therefore, this study proposes a self-learning method for AI models to reduce the cost of creating additional training data for each observation point and to improve the accuracy. Furthermore, the performance of the proposed method is evaluated using CCTV images from multiple locations. The proposed method proposes a self-learning algorithm that composites the object information to be detected into the background image of the observation point so that the distribution distance between the training data and the inferred data is brought closer based on domain adaptation. The measurement accuracy was verified using CCTV images from multiple locations, and the accuracy improvement was confirmed without the need for manual annotation. The proposed method enables cost-effective improvement of measurement accuracy at different locations and contributes significantly to solving practical problems.