画像電子学会
The Institute of Image Electronics
Engineers of Japan

年次大会予稿
Proceedings of the Media Computing Conference

アノテーションロスを克服する ラベル付き画像セグメントの合成によるデータ拡張

Data Augmentation by Synthesizing Labeled Image Segments Overcomes Annotation Loss

加納 一馬† 森 裕輝† 東浦 圭亮† Tahera HOSSAIN†
片山 晋† 浦野 健太† 米澤 拓郎† 河口 信夫†

Kazuma KANO† Yuki MORI† Keisuke HIGASHIURA† Tahera HOSSAIN†
Shin KATAYAMA† Kenta URANO† Takuro YONEZAWA† and Nobuo KAWAGUCHI†

†名古屋大学 大学院工学研究科 †Graduate School of Engineering, Nagoya University

E-mail: kazuma@ucl.nuee.nagoya-u.ac.jp

1. はじめに

カメラによる物体検出は至るところで利用されており，産業現場における生産性の向上を目的とした作業分析への応用も期待されている．現在の物体検出技術の多くは深層学習に基づいて高い精度を実現しており，人々は公開された学習済みモデルによって手軽に恩恵を享受できる．しかし，これらの汎用モデルは特殊な状況では十分に機能しない可能性がある．例えば，物流倉庫には多種多様な物品やハンドパレットなどの専用機器が存在するが，モデルはこれらに関する知識を備えていない場合がある．また，背の高い物体による遮蔽を防ぎつつ広い空間を監視するには広角カメラの高所への設置が有効だが，歪みのある画像や頭上からの視点は一般に想定されていない．このような場合，ファインチューニングや転移学習のためのデータセットが必要になり，アノテーションにかかるコストが物体検出システムの実用化を妨げている．

我々はこれまでに，オプティカルフローで抽出した動体の埋め込み表現をクラスタリングによってグループ化する半自動アノテーションフレームワーク[1]を開発し，バウンディングボックスの自動決定と複数オブジェクトへの一括ラベリングにより人的労力を格段に削減させてきた．しかし，この手法には静止物体や上手くグループ化されなかったオブジェクトに対してアノテーションできないという課題があった．そこで本研究ではこのアノテーションのロスを克服するために画像合成を導入する．手法の概要を図 1 に示す．提案手法では，オプティカルフローによって与えられる候補の中から検出対象が含まれないフレーム（本稿では無対象フレームと呼ぶ）を選び出し，その上に半自動フレームワークによってラベル付けしたセグメントを貼り付けて，全ての検出対象がアノテーションされている画像を合成する．このとき，前景と背景のコンテキストを揃えるため，同じ定点カメラの映像から収集したフレーム上の元の位置にセグメントを配置する．物流倉庫の天井に取り付けられた広角カメラの映像から作業員についてアノテーションされたデータセットを作成して検出モデルの性能を評価したところ，アノテーションロスの排除によりモデルの性能が向上し，提案手法は手動アノテーションに迫る高い精度を 4 分の 1 未満の作業時間で達成した．

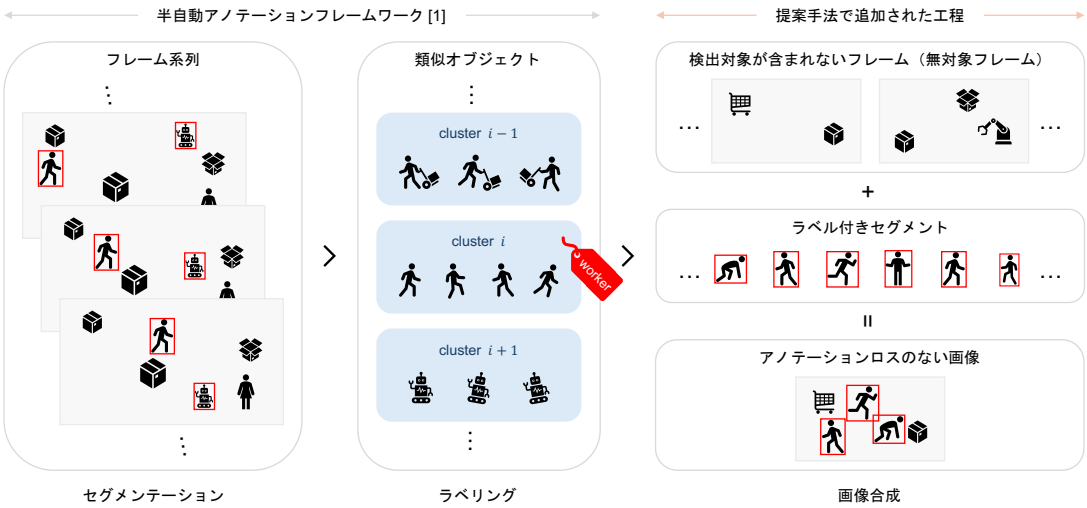


図 1 ラベル付きセグメントを用いたアノテーションロスのない画像の合成

2. 関連研究

2.1 能動学習

多くの研究が様々な切り口でアノテーションコストの削減に取り組んできた．能動学習は優先的にアノテーションすべきデータをモデルに自ら選択させる技術である．Yang らは Fully Convolutional Network（FCN）によって提供される類似性と不確実性の情報を活用し，アノテーションが足りていない代表的なデータを特定した[2]．Yoo らはネットワークに小さなパラメトリックモジュールを付け加え，ラベルのないデータに対する損失関数の値，すなわちどれだけ間違えそうかを予測した[3]．これらの研究は学習の効率を向上させたが，依然としてかなりのデータ量を必要としており，さらなるコストの削減が求められる．

2.2 アノテーションの自動化

いくつかの研究ではアノテーション作業の一部を自動化して人的労力を低減している．Lu らは肺結節の悪性度や属性の予測に自己教師あり対照学習を導入し，アノテーションされていないデータで部分的にモデルを学習させた[4]．Elangovan らはキッチンでの行動に対するアノテーション作業において，機械学習技術を用いていくつかのサブタスクを自動化した[5]．しかしこれらの手法は特定のドメインに依存しており，多品種倉庫のような環境には適用できない．そこで我々は人やロボットおよびそれらの搬送物を含む可動クラスの物体検出を対象とした汎用的な半自動アノテーションに取り組んできた[1]．これにより大量のアノテーションにかかるコストが大幅に削減されたが，その反面アノテーションの抜け漏れや不良を生じやすかった．

2.3 画像合成によるデータ拡張

入手困難なデータを再現したり既存のデータを拡張したりするために画像合成がしばしば用いられる．Dwibedi らは物体のセグメントをシーン画像上に貼り付けて多様なシーンを生成した[6]．ただし簡素な背景を持つ複数視点からの物体画像の利用を前提としており，様々な容姿や体勢をとりうる作業員などを対象とする場合にそのような画像を用意するのは手間がかかる．さらにセグメントを背景画像上のランダムな位置に配置しているため，特に広角カメラによる歪んだ画像を使用する場合に不自然さを生じる．セグメントを適切に配置するために専用のモデルで背景画像のコンテキストを推定する研究もあるが[7, 8]，複雑な環境で正確にコンテキストを認識するのは難しい．

3. 提案手法

3.1 システムの概観

アノテーションロスをなくしつつデータの多様性を確保するため，以前開発した半自動アノテーションフレームワーク[1]に画像合成を適用する．提案手法におけるデータセットの作成手順を図 2 に示す．青色は自動，橙色は人手による作業を伴う工程を表す．ステップ 1 では深層学習による密なオプティカルフロー手法である RAFT [9]で算出した各画素の動きの大きさに基づいて物体をセグメンテーションする．ステップ 2 ではシャムネットワークによる自己教師あり表現学習手法である SimSiam [10]を用いてセグメントを固定長ベクトルにエンコードする．なお，似たオブジェクトが近くにエンコードされるよう事前にモデルを学習させておく．ステップ 3 では K-Means でベクトルをクラスタリングする．ステップ 4 ではグループ化された類似オブジェクトに対してまとめてラベリングする．ステップ 5 では RAFT で動きが検知されなかったフレームの中から無対象フレームを選別する．ステップ 6 では無対象フレーム上にラベル付きセグメントを配置してアノテーション付き画像を合成する．

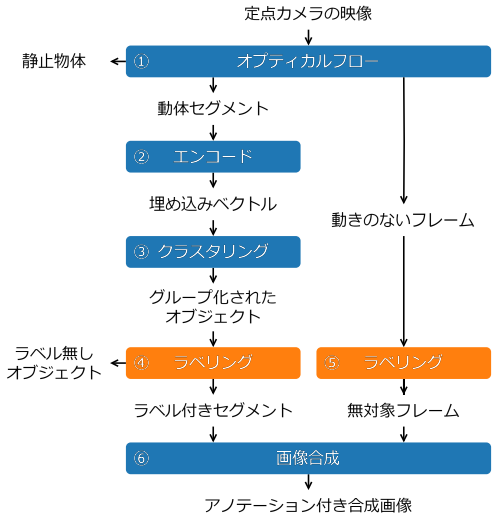


図 2　データセットの作成手順

ステップ 1 から 4 については以前の手法[1]にならい，新たにステップ 5 と 6 を追加している．以前の手法ではステップ 1 で見逃された静止物体とステップ 4 で上手くグループ化されなかったオブジェクトにアノテーションが付与されないのに対して，提案手法ではラベル付けされたオブジェクトのみを使用して全ての検出対象がアノテーションされた画像を作成する．また，きれいにセグメンテーションされなかった物体を意図的に除外できるため，バウンディングボックスの質の向上が期待できる．この手法は定点カメラを用いた任意の可動クラスの物体検出に適用できる．ステップ 1 から 4 についての詳細な説明は以前の論文に譲り，以降はステップ 5 と 6 について述べる．

3.2 無対象フレームの収集

全ての検出対象がアノテーションされた画像を合成するために検出対象が含まれない背景画像が必要になる．提案手法では，検出対象は動く傾向にあると仮定し，RAFT が動きを検知しなかったフレームの中からランダムに候補をサンプリングして，無対象フレームを効率的に収集する．データの多様性のためこれらのフレームには検出対象でない物体が多く含まれているのが望ましい．この工程では各カメラのいくつかのフレームを人が確認する必要があるものの，手動アノテーションと比べるとはるかに少ない負担で済む．なお，用意する無対象フレームが多いほど生成される画像は多様になるがその分だけ必要な労力も増える．

3.3 合成画像の生成

ラベル付きセグメントと無対象フレームからアノテーションロスのない合成画像を生成する．まず，ラベル付きセグメントを 1 つランダムに取り出す．次に，それが採用済みのセグメントと重複していないかを調

べる．2 つのバウンディングボックスの重なる領域と小さい方のボックスとの面積比が閾値 $R_{overlap}$ より大きいとき重複していると判定する．本稿では $R_{overlap} = 0.25$ とした．どの採用済みセグメントとも重複していなければそのセグメントを採用する．これらの操作を $\mathrm{round}(\mathrm{gauss}(\mu,\sigma))$ 個のセグメントが集まるまで繰り返す．ここで μ と σ は 1 回の合成に使用するセグメント数の平均と標準偏差に相当するパラメータである．続いて，無対象フレームをランダムに 1 つ取り出す．最後に，フレーム上にセグメントを貼り付けて画像を合成する．このとき，元の位置および向きのままカメラから遠い順にセグメントを配置して，前景と背景のコンテキストを揃えるとともに実際の遠近感を模倣する．我々の環境ではカメラが真下を向いていたため，カメラとの距離の代わりに画像の中心との距離を計算した．以上の操作をカメラごとに $\mathrm{round}(CN/\mu)$ 回繰り返す．ここで N はそのカメラにおけるラベル付きセグメントの総数を表し，C は各セグメントが合成に使用される回数の平均に相当するパラメータである．

4. 評価

4.1. データセットの作成

物流倉庫の入出荷場の天井に鉛直下向きに設置された広角カメラ（H.View HV-800G2A5）で撮影された映像内の作業員に対して半自動的にアノテーションした．動画はフル HD，8000 [kbps]，5 [fps]で記録され，事前に Double Sphere カメラモデル[11]を用いて歪みを補正されている．アノテーションの質を保証するために設けたラベリングのガイドラインを以下に記す．

- 複数作業員または非作業員を含むオブジェクトは作業員オブジェクトとみなさない．
- 身体の一部のみで構成されるまたは過剰な余白を含むオブジェクトにはラベルを付与しない．

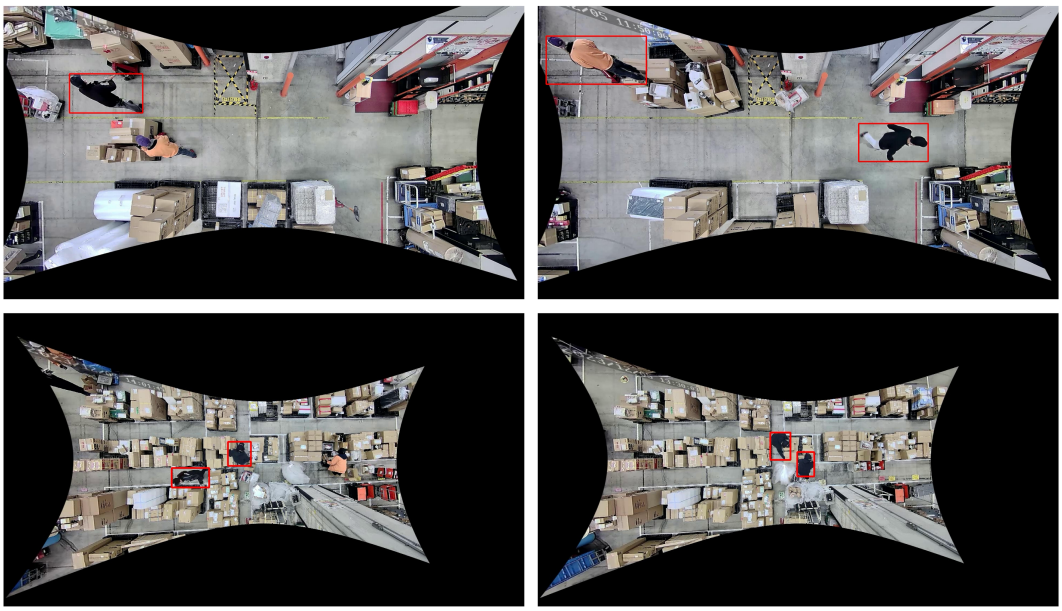
最終的に 22 台のカメラで撮影された 13709 フレームに渡る 22000 個の動体オブジェクトのうち 4066 オブジェクトを作業員としてラベリングした．また，無対象フレーム，すなわち作業員が含まれないフレームをカメラごとに 5 枚ずつ収集した．

続いて，これらのデータを基に 2 つの方法でデータセットを作成した．1 つ目は以前の半自動手法[1]，すなわちラベル付きオブジェクトを含むフレームを寄せ集める方法，2 つ目は提案手法，すなわち合成画像を生成する方法である．提案手法における合成画像生成に関するハイパーパラメータを**表 1** にまとめる．また，以前の手法と提案手法によるデータの例を**図 3** (a)と(b)にそれぞれ示す．赤色の矩形はアノテーションを表す．以前の手法ではアノテーションされていない作業員が見受けられるのに対して，提案手法ではそれらがなくなっている．提案手法では上の画像の左上に見られるように非現実的な状況が生成される場合もあるものの，遠近感には一貫性があり概ね自然に見える．

さらに，2 人の手慣れたアノテータにほぼ同数のアノテーションを手動で付けてもらい，それを基にデータセットを作成した．また，手動でアノテーションされたデータセットをテスト用としてもう 1 つ用意した．これらのデータセットに含まれるアノテーションおよび画像の数とラベリングやアノテーションにかかった時間を**表 2** にまとめる．なお，提案手法のデータセットに含まれるアノテーション数が多いのは元のアノテーションが C 倍に拡張されているためである．提案手法と以前の手法における作業時間は手動アノテーションの 4 分の 1 未満であった．ただし，無対象フレームの収集や自動的な処理の実行にかかった時間は含まれない．

表 1　合成画像生成のハイパーパラメータ

画像当たりのオブジェクト数の平均 μ	2
画像当たりのオブジェクト数の標準偏差 ρ	0.5
各オブジェクトの使用回数の平均 C	2



(a) Semi-automated [1]　　(b) Proposed

図 3　データセットに含まれる画像とアノテーションの例

表 2　データ量と作業時間

	アノテーション数	画像数	時間 [分]
提案手法	8131	4066	94
半自動[1]	4066	3786	94
手動	4062	1322	432
テスト	863	467	–

4.2. モデルの学習

初めに,テストデータを除く各データセットを約 8:2 の比で学習用と検証用のサブセットに分割した．学習用のサブセットは hue, saturation, brightness, translation, scale, shear, perspective, flip についてランダムに拡張した．その後，各データセットを用いて YOLOv8 の medium サイズの学習済みモデル[12]をファインチューニングした．本実験のタスクは単一クラスの物体検出であり，モデルは作業員を囲うようなバウンディングボックスを推定するように学習される．バッチサイズは 128，最大エポック数は 300 とし，検証用サブセットに対する平均適合率（AP: Average Precision）が最大となったエポックにおける重みを採用した．

4.3. 結果と考察

テストデータに対する PR 曲線を**図 4** に示す. また，このときの AP を**表 3** にまとめる．まず，以前の半自動手法[1]との違いについて考察する．提案手法は以前の手法と比べて全体的に優れた結果が得られた．以前の手法との差分であるアノテーションロスの排除がモデルの性能向上に寄与したと推察される．以前の手法では再現率が 0.85 辺りで急激に低下しているが，これは学習データ中のアノテーションされていない作業員に起因すると考えられる．モデルは特に尤もらしいときにのみ作業員だと判断するように学習され，偽陰性が増加したと推測される．

次に手動アノテーションと比較しながらさらなる性能向上の余地を検討する．提案手法の AP は手動アノテーションのそれに迫っており，再現率は特に優れていた．本実験は単一クラスを対象としており，背景画像の限られる提案手法では作業員以外の物体がデータセットに十分含まれていなかった．モデルはそれらしき物体を作業員とみなしやすいように学習され，適合率を犠牲に再現率が高くなったと推測される. 実際，適合率は約 0.9 で頭打ちになってしまっている．提案手法は複数クラスにも容易に適用できるが，その場合各クラスは互いに負例として働くため適合率の向上が期待できる．また，検出対象でないオブジェクトの合成も背景画像を多様化し適合率の向上につながる可能性がある．

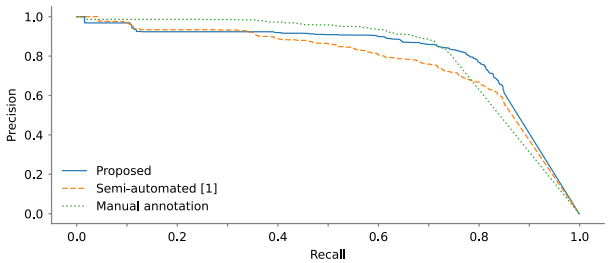


図 4　テストデータに対する PR 曲線

表 3　テストデータに対する平均適合率

	AP50	AP75	AP50:95
提案手法	0.81	0.62	0.56
半自動[1]	0.77	0.57	0.53
手動	0.82	0.64	0.57

謝辞　本研究の一部は JSPS 科研費（JP22K18422），NEDO 委託研究（JPNP23003），NICT 委託研究（222C01, 22609），内閣府 CSTI SIP3（JPJ012495），JST BOOST（JPMJBS2422）の支援を受けています．また，実験環境を提供いただいたトラスコ中山株式会社様に感謝いたします．

文　献

[1] K. Higashiura et al., “Semi-Automated Framework for Digitalizing Multi-Product Warehouses with Large Scale Camera Arrays,” 2024 IEEE International Conf. on Pervasive Computing and Communications (PerCom), pp.98-105, 2024.

[2] L. Yang et al., “Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation,” Medical Image Computing and Computer Assisted Intervention - MICCAI 2017: 20th International Conf., no.3, pp.399-407, 2017.

[3] D. Yoo and I. S. Kweon, “Learning Loss for Active Learning,” 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp.93-102, 2019.

[4] J. Lu et al., “Reducing Annotation Need in Self-explanatory Models for Lung Nodule Diagnosis,” Interpretability of Machine Intelligence in Medical Image Computing: 5th International Workshop, pp.33-43, 2022.

[5] N. Elangovan et al., “On Human Grasping and Manipulation in Kitchens: Automated Annotation, Insights, and Metrics for Effective Data Collection,” 2023 IEEE International Conf. on Robotics and Automation (ICRA), pp.11329-11335, 2023.

[6] D. Dwibedi et al., “Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection,” 2017 IEEE International Conf. on Computer Vision (ICCV), pp. 1310-1319, 2017.

[7] G. Georgakis et al., “Synthesizing Training Data for Object Detection in Indoor Scenes,” Proc. of Robotics: Science and Systems, 2017.

[8] N. Dvornik et al., “Modeling Visual Context Is Key to Augmenting Object Detection Datasets,” Computer Vision - ECCV 2018: 15th European Conf., no.12, pp. 375-391, 2018.

[9] Z. Teed and J. Deng, “RAFT: Recurrent All-Pairs Field Transforms for Optical Flow,” Computer Vision - ECCV 2020: 16th European Conf., no.2, pp.402-419, 2020.

[10] X. Chen and K. He, “Exploring Simple Siamese Representation Learning,” 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp.15745-15753, 2021.

[11] V. Usenko et al., “The Double Sphere Camera Model,” 2018 International Conf. on 3D Vision (3DV), pp.552-560, 2018.

[12] G. Jocher et al., “Ultralytics YOLOv8,” 2023.