

敵対的訓練を用いたドメイン不変な特徴抽出

藤井 一磨¹ 計良 宥志² 川本 一彦²

概要: 深層学習を用いた物体検出では、ソースドメインとターゲットドメインの間に背景やスタイルなどに大きな乖離があるとき、ターゲットドメインでの検出性能が低下してしまう。このドメインシフト問題は、深層モデルがソースドメインに固有な特徴を抽出するために起こる。本研究では、深層モデルに対する敵対的攻撃に対してロバストな特徴がドメイン不変な性質を持つことを利用し、教師無しドメイン適応のための特徴抽出法を提案する。提案手法では、ソースドメインでの深層物体検出モデルの敵対的訓練によりロバストな特徴を抽出しつつ、ソースドメインとターゲットドメイン間の特徴アライメントを加えることでドメイン適応を実現している。さらに、提案手法のターゲットドメインでの検出精度は、ソースドメインからの Fréchet 距離が大きいほど向上することを発見し実験的に検証している。この結果を用いれば、提案手法の有効性を深層学習前に判定することができる。検証用データセットを用いた実験で、ベースラインと比較しつつ提案手法の有効性を示している。

キーワード: ドメイン適応, 物体検出, 敵対的訓練

1. はじめに

物体検出は画像に映る物体の種類と位置を推定する技術で、車の自動運転や防犯カメラからの異常検知などに応用される。近年の深層学習の発展に伴い、数多くの物体検出モデルが提案されている [1], [2], [3]。これらは、PASCAL Visual Object Classes (VOC) [4] や Microsoft Common Objects in Context (MS COCO) [5] のような大規模な正解ラベル付きデータセットで学習することで高い検出精度を実現している。しかし、学習データとテストデータで背景や画像スタイルに大きな乖離がある場合、テストデータでの検出精度が低下するという問題がある。このような学習データとテストデータ間のドメインの違いはドメインシフトと呼ばれる。ドメインシフトにより検出精度が低下する例を図 1 に示す。物体検出の応用場面において、検出精度の低下は重大な事故を引き起こす可能性があるため、ドメインシフト問題への対策が求められる。

ドメインシフトによる精度の低下を防ぐ素朴な方法は、テストデータと同じドメインのデータに正解ラベルを付与して新しい学習データを作成することである。しかし、大量の画像に対して手作業で正解ラベルを付与するのは、コストが非常に大きく現実的ではない。そのため、正解ラベルが豊富にあるソースドメインで学習した深層モデルを、



図 1 ドメインシフトによって検出精度が低下する例。

正解ラベルが存在しないターゲットドメインに適応させる教師無しドメイン適応が注目を集めている [6]。最近の研究では、物体検出のための教師なしドメイン適応手法として様々なアプローチが提案されている [7]。最も汎用的なアプローチは、ソースドメインとターゲットドメインから得られる特徴を近づける特徴アライメントである [8], [9]。しかし、特徴アライメントだけでは十分なドメイン適応性能を達成するのは難しい。その他に、ターゲットドメインでの自己訓練 [10] や画像変換を利用する方法 [11] などがあるが、これらはターゲットドメインの検出難易度や種類によっては適用が難しいという問題がある。

そこで本研究では、ターゲットドメインの検出難易度や種類によらず適用可能な、ドメイン不変な特徴抽出による教師無しドメイン適応アプローチを提案する。具体的には、敵対的攻撃に対してロバストな特徴 [12] がドメイン不変な性質を持つことに着目し、ソースドメインでの敵対的訓練によって教師無しドメイン適応が可能になることを実証する。これは、敵対的訓練された深層モデルが転移学習

¹ 千葉大学大学院融合理工学府

² 千葉大学大学院工学研究院

に有効であることを示した研究 [13], [14] から着想を得たものである。これらの研究では目標のタスクに少量の正解ラベルが存在する転移学習に焦点を当てているのに対し、本研究ではターゲットドメインに正解ラベルが存在しない教師無しドメイン適応に敵対的訓練を応用する。さらに、ソースドメインでの敵対的訓練によってドメイン不変な特徴を抽出しつつ、ターゲットドメインとの特徴アライメントを加えるドメイン適応手法を提案する。

画像スタイル、天候、および場所のドメインシフトを想定した実験で、提案手法による教師なしドメイン適応の有効性を確認する。物体の外観や背景の変化が大きい、画像スタイルおよび場所のドメインシフトでは、ソースドメインでの敵対的訓練によってターゲットドメインの検出精度が向上する。さらに、特徴アライメントを加えることで、教師無しドメイン適応性能の向上を確認する。一方で、物体の外観の変化が小さい天候のドメインシフトでは、ソースドメインでの敵対的訓練によってターゲットドメインの検出精度が低下する。この結果を踏まえて、ドメインシフトの大きさとターゲットドメインでの検出精度の関係を実験的に明らかにする。ドメインシフトの大きさは Fréchet 距離によって測定することができ、これによって提案手法の有効性を深層学習前に判定できることを示す。

本研究における主な貢献は以下の通りである。

- 物体検出の教師なしドメイン適応に対して、敵対的訓練によるドメイン不変な特徴抽出が有効であることを実証する。実験により、ソースドメインでの敵対的訓練は、ドメインシフトが大きい場合にターゲットドメインでの検出精度を向上させることを示す。
- 敵対的訓練によるドメイン不変な特徴抽出に加えて、ソースドメインとターゲットドメイン間の特徴アライメントを適用した手法を提案する。実験により、それぞれを単独で適用した場合よりも、提案手法がターゲットドメインでの検出精度を向上させることを示す。
- Fréchet 距離によってドメインシフトの大きさを定量化することで、提案手法の有効性を深層学習前に判定できることを示す。

2. 関連研究

2.1 物体検出における教師無しドメイン適応

物体検出における教師無しドメイン適応の代表的なアプローチとして、特徴アライメント [8], [9], ターゲットドメインでの自己訓練 [10], および画像変換 [11] が挙げられる。特徴アライメントはソースドメインとターゲットドメインから得られる特徴を近づける方法で、特徴抽出器とドメイン判別器による敵対的な学習 [15] が広く使われている。この特徴アライメントを利用した手法は数多く提案されており、様々なドメインシフトに対して有効であることが実証されている。ターゲットドメインでの自己訓練は、ソー

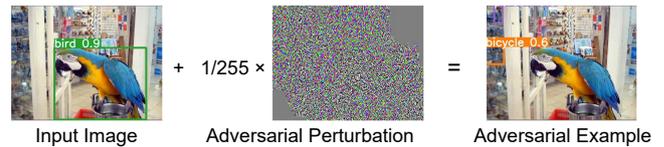


図 2 物体検出モデルに対する敵対的攻撃の例。物体検出モデルは左の元画像で鳥を正しく検出するが、元画像に敵対的摂動を加えた右の敵対的サンプルでは検出に失敗する。

ソースドメインで学習した知識に基づいてターゲット画像に疑似ラベルを付与することで深層モデルを学習する方法である。しかし、ターゲットドメインの検出難易度が高いと、誤った疑似ラベルで学習する可能性が高くなるという問題がある。画像変換によるドメイン適応は、CycleGAN [16] のような画像変換技術を用いてソース画像をターゲットドメインのスタイルに変換したのちに、元の正解ラベルを利用してモデルを学習する方法である。画像変換では物体の形状を変えるような変換は難しいため、ドメインシフトの種類によっては適用できない場合がある。

本研究では、深層モデルがソースドメインに固有な特徴を抽出することを防ぎ、ドメインシフトによる性能低下を緩和するというアプローチをとる。そのために、ソースドメインでの敵対的訓練によるドメイン不変な特徴抽出を利用する。提案アプローチはターゲット画像を使わずに学習できるため、ターゲットドメインの検出難易度や種類によらず適用でき、他のアプローチとも容易に組み合わせることが可能である。

2.2 敵対的訓練

深層モデルの脆弱性の一つとして、敵対的攻撃が挙げられる。敵対的攻撃は、深層モデルに誤りを生じさせるような敵対的摂動を入力データに加える攻撃である [17]。図 2 に物体検出モデルに対する敵対的攻撃の例を示す。これに対し敵対的訓練では、学習データに敵対的摂動を加えることで敵対的攻撃に対してロバストな深層モデルを学習する。

敵対的訓練に関する最近の研究で、敵対的攻撃に対して非ロバストな特徴とロバストな特徴の存在が明らかにされている [12]。通常訓練されたモデルは、高い精度を達成するために非ロバストな特徴に依存する。この非ロバストな特徴は人間には知覚できないような特徴であり、敵対的攻撃が成功する要因となる。一方、敵対的訓練されたモデルは、敵対的摂動に騙されないために人間の視覚に近いロバストな特徴を主に抽出するようになる [18]。このような抽出される特徴の違いに基づき、敵対的訓練されたモデルが通常訓練されたモデルよりも新しいデータセットへの転移学習に有効であることが実証されている [13], [14]。

以上の研究から、敵対的訓練によって深層モデルが抽出するロバストな特徴はドメイン不変な特性を持つことが考えられる。本研究ではこの性質に着目し、物体検出にお

ける教師無しドメイン適応に対して、ソースドメインでの敵対的訓練が有効であることを示す。さらに、敵対的訓練によるドメイン不変な特徴抽出に、ソースドメインとターゲットドメイン間の特徴アライメント [9] を加えたドメイン適応手法を提案する。

3. 提案手法

本研究では、図 3 に示すようにソースドメインでの敵対的訓練（図上段）と特徴アライメント（図下段）によるドメイン適応手法を提案する。本節では最初に、物体検出における教師無しドメイン適応の問題設定を定式化する。次に、ドメイン不変な特徴抽出を実現するための、ソースドメインでの敵対的訓練について説明する。最後に、ドメイン不変な特徴をターゲットドメインに整合させるために、特徴アライメントを加えたドメイン適応手法を提案する。

3.1 問題設定

物体検出の教師無しドメイン適応では、ソースドメインのデータ分布 D_s から正解ラベルつきデータ (x_s, y_s) と、ターゲットドメインのデータ分布 D_t から正解ラベルなしデータ x_t が利用できる。ここで、 x_s と x_t は入力画像、 y_s は正解ラベルを表し、 D_s と D_t は異なるデータ分布を持つ。教師無しドメイン適応の目標は、ソースドメインの正解ラベルつきデータとターゲットドメインの正解ラベルなしデータのみを利用して、ターゲットドメインでの検出精度を向上させることである。

3.2 ソースドメインでの敵対的訓練

本研究の主な目的は、物体検出モデルの教師無しドメイン適応に対して、ソースドメインでの敵対的訓練が有効であることを実証することである。敵対的訓練によって抽出されるロバストな特徴はドメイン不変な性質を持つことから、ターゲットドメインでの検出精度の向上に寄与することが期待される。

まず、ソースドメインでの通常訓練を式で表すと、

$$\min_F \mathbb{E}_{(x_s, y_s) \sim D_s} [\mathcal{L}_{\text{det}}(F(x_s), y_s)], \quad (1)$$

となる。ここで、 \mathcal{L}_{det} は物体検出の損失関数、 F は物体検出モデルのネットワークを表す。これに対し、敵対的訓練では、図 3 上段のように入力画像 x_s に敵対的摂動を加えて物体検出モデルを学習する。本研究では、敵対的摂動の生成方法として代表的な fast gradient sign method (FGSM) [19] を使用する。FGSM によって敵対的摂動 δ^* は以下のように生成される。

$$\delta^* = \epsilon \text{sign}(\nabla_{x_s} \mathcal{L}_{\text{det}}(F(x_s), y_s)). \quad (2)$$

ここで、 ϵ は摂動の大きさを決める値で、通常は人間の視覚では判別できないほど小さい値を用いる。式 (2) では、

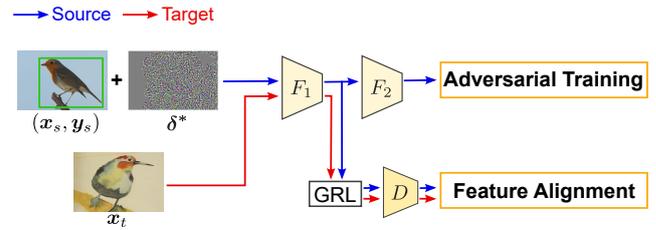


図 3 提案手法による学習の流れ。 F_1 および F_2 は物体検出モデルのネットワーク、 D はドメイン判別器、GRL は gradient reversal layer を表す。ソース画像 x_s に敵対的摂動 δ^* を加えて物体検出モデルを敵対的訓練すると同時に、ソース画像 x_s とターゲット画像 x_t による特徴アライメントを適用する。

\mathcal{L}_{det} の x_s に関する勾配の符号を使って、 \mathcal{L}_{det} を最大化するような δ^* が計算される。この δ^* を用いると、ソースドメインでの敵対的訓練は以下の式で表される。

$$\min_F \mathbb{E}_{(x_s, y_s) \sim D_s} [\mathcal{L}_{\text{det}}(F(x_s + \delta^*), y_s)]. \quad (3)$$

式 (3) の敵対的訓練によって、物体検出モデルは敵対的摂動による誤った予測を防ぐために、人の視覚に近いロバストな特徴を抽出するようになる。

3.3 ドメイン不変な特徴抽出と特徴アライメントによるドメイン適応

ソースドメインでの敵対的訓練によって、物体検出モデルはドメイン不変な性質を持つロバストな特徴を抽出することが期待される。しかし、敵対的訓練ではソースドメインのデータのみを使用するため、ロバストな特徴はターゲットドメインで不整合を引き起こす可能性がある。そこで本研究では、敵対的訓練に加えてソースドメインとターゲットドメイン間の特徴アライメント（図 3 下段）を適用する手法を提案する。ドメイン不変な特徴をターゲットドメインに整合させることで、ドメイン適応性能の向上を目指す。

提案手法では、色やテクスチャといった局所的な特徴のアライメントを促す手法 [9] を採用している。この手法は、特徴アライメント手法の中でも比較的単純な構造で、物体検出モデルのアーキテクチャに依存せず実装することができる。物体検出モデル F の前半の特徴抽出部を F_1 、残りの特徴抽出部と検出部を F_2 とする。 F_1 の出力は、逆伝播時に勾配の符号を反転する役割を持つ gradient reversal layer (GRL) [15] を通して、ドメイン判別器 D に入力される。 F_1 は幅 W 、高さ H の特徴マップを出力し、 D は入力された特徴マップのセルごとにドメイン予測を出力する。 D はソース画像に対するドメイン予測が 0、ターゲット画像に対するドメイン予測が 1 に近づくように学習されるのに対して、 F_1 は D のドメイン予測が逆になるように学習される。GRL の勾配の符号を反転させる機能により、 D と F_1 における敵対的な特徴アライメントは、以下に示す共通の損失関数を使用することができる。



図 4 画像スタイルのドメインシフトの例。

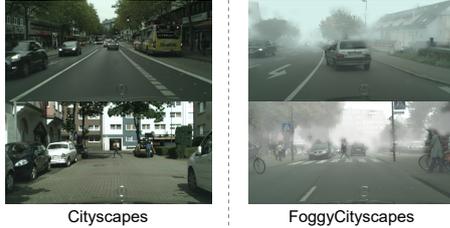


図 5 天候のドメインシフトの例。

$$\mathcal{L}_{fa_s}(D(F_1(\mathbf{x}_s + \delta^*))) = \frac{1}{WH} \sum_{w,h} D(F_1(\mathbf{x}_s + \delta^*))_{wh}^2, \quad (4)$$

$$\mathcal{L}_{fa_t}(D(F_1(\mathbf{x}_t))) = \frac{1}{WH} \sum_{w,h} (1 - D(F_1(\mathbf{x}_t)))_{wh}^2. \quad (5)$$

ここで、 $D(\cdot)_{wh}$ はセル座標 (w, h) における D のドメイン予測を表す。式 (4) でソース画像 \mathbf{x}_s に敵対的摂動 δ^* を加える点に注意されたい。

式 (3) で示したソースドメインでの敵対的訓練と合わせて、提案手法の学習は以下の式で表すことができる。

$$\begin{aligned} \max_{F_1} \min_{F, D} \mathbb{E}_{\substack{(\mathbf{x}_s, \mathbf{y}_s) \sim \mathcal{D}_s \\ \mathbf{x}_t \sim \mathcal{D}_t}} [\mathcal{L}_{\det}(F(\mathbf{x}_s + \delta^*), \mathbf{y}_s) \\ + \lambda(\mathcal{L}_{fa_s}(D(F_1(\mathbf{x}_s + \delta^*))) + \mathcal{L}_{fa_t}(D(F_1(\mathbf{x}_t))))]. \end{aligned} \quad (6)$$

ここで、 λ は敵対的訓練と特徴アライメントのバランスを調整するハイパーパラメータである。

4. 実験

本節では、教師無しドメイン適応の検証用データセットを使用した実験で、提案手法の有効性を示す。その後、提案手法によるドメイン適応の性能がドメインシフトの大きさに依存することを実験的に明らかにする。

4.1 データセット

本研究では、三種類のドメインシフトについて提案手法の有効性を検証する。一つ目は画像スタイルのドメインシフトで、ソースドメインに PASCAL VOC [4]、ターゲットドメインに Clipart1k, Watercolor2k, および Comic2k [20] のデータセットを用いる (図 4)。PASCAL VOC は実世界の画像データセットである。16551 枚の訓練データを持ち、



図 6 場所のドメインシフトの例。

20 種類の物体のクラスを含む。Clipart1k はベクター画像やスケッチ等のデータセットである。500 枚の訓練データと 500 枚のテストから成り、PASCAL VOC と同じ 20 種類の物体のクラスを持つ。Watercolor2k および Comic2k は、それぞれ水彩画および漫画のデータセットである。どちらも 1000 枚の訓練データと 1000 枚のテストデータから成り、PASCAL VOC に含まれる 6 種類の物体のクラスを持つ。画像スタイルのドメインシフトでは、ソースドメインとターゲットドメイン間で物体の外観が大きく異なる。

二つ目は天候のドメインシフトで、ソースドメインに Cityscapes [21]、ターゲットドメインに FoggyCityscapes [22] のデータセットを用いる (図 5)。Cityscapes は街中を走行する車の車載カメラから撮影された画像データセットである。2975 枚の訓練データと 500 枚のテストデータから成り、8 種類の物体のクラスを含む。FoggyCityscapes は Cityscapes の画像に対して人工的に霧を付与したデータセットである。天候のドメインシフトでは、ソースドメインとターゲットドメイン間で天候状況が大きく異なるが、物体の外観はあまり変わらない。

三つ目は場所のドメインシフトで、ソースドメインに MS COCO [5] と一般的な工事現場での人検出データセット、ターゲットドメインに特殊な工事現場での人検出データセットを用いる (図 6)。MS COCO は実世界の画像データセットである。このドメインシフトでは人物の検出を対象とするため、MS COCO の訓練データから人物が含まれる 64115 枚の画像を使用する。工事現場での人検出データセットは、重機視点の画像データセットである。ソースドメインには、複数の一般的な工事現場で撮影された 8887 枚の訓練データから成るデータセットを使用する。ターゲットドメインには、土、コンクリタイル、および雪の工事現場で撮影されたデータセットを使用する。土およびコンクリタイルの工事現場のデータセットは、それぞれ 502 枚および 514 枚のテストデータから成り、訓練データは存在しない。雪の工事現場のデータセットは 479 枚の訓練データと 1108 枚のテストデータから成る。場所のドメインシフトでは、ソースドメインとターゲットドメイン間で物体の映り方や背景が大きく異なる。

表 1 PASCAL VOC から Clipart1k, Watercolor2k, および Comic2k へのドメイン適応の結果. 各データセットのテストデータでの mAP (%) を報告している. ST, AT, および FA はそれぞれソースドメインでの通常訓練, ソースドメインでの敵対的訓練, および特徴アライメントを表す.

Method	Clipart1k	Watercolor2k	Comic2k
ST	37.6	49.2	26.7
ST + FA	43.7	44.3	31.5
AT (ours)	45.3	55.3	29.5
AT + FA (ours)	49.4	55.6	33.4

4.2 実験設定

モデル 本研究では, 物体検出モデルとして You Only Look Once v3 (YOLOv3) [3] を使用する. YOLOv3 の 26 個の畳み込み層を含む前半のネットワークを 3.3 節で示した F_1 とし, それ以降のネットワークを F_2 とする. 局所的な特徴アライメントで使用するドメイン判別器 D は, 元論文 [9] で提案されたモデルを使用する.

学習方法 訓練データには四枚の画像を結合して一枚の画像にするモザイクデータ拡張 [23] を適用し, 入力画像サイズは 416×416 に固定する. 全ての実験で物体検出モデルは, MS COCO で事前学習済みの YOLOv3 を初期重みとして, ソースドメインのデータセットサイズで 50 エポック学習される. 学習率は warmup によって最初の 3 エポックで 1.0×10^{-2} まで徐々に上げた後, cosine annealing によって 2.0×10^{-3} までなめらかに下げる. ソースドメインでの通常訓練および敵対的訓練では, バッチサイズは 16 とする. 特徴アライメント適用時には, ターゲットドメインの正解ラベルなし訓練データを 16 枚加えて, バッチサイズは 32 とする. 式 (2) の敵対的摂動の大きさは $\epsilon = 1/255$ とし, 式 (6) の学習のバランスを調整するハイパーパラメータは $\lambda = 1.0$ とする.

評価方法 テスト時の入力画像は, 画像の長辺が 416 になるようにリサイズされる. 評価指標として, ターゲットドメインのテストデータで mean average precision (mAP) を計算する. 予測されたバウンディングボックスと正解バウンディングボックスの intersection over union (IoU) が 0.5 より大きい場合に正しく検出されたとみなす. 全ての実験で物体検出モデルを三回ずつ学習し, 結果にはその平均値を報告する.

4.3 結果

画像スタイルのドメインシフト 画像スタイルのドメインシフトでの実験結果を報告する. PASCAL VOC から Clipart1k, Watercolor2k, および Comic2k へのドメイン適応の結果を表 1 に示す. 提案アプローチであるソースドメインでの敵対的訓練 (AT) は通常訓練 (ST) より Clipart1k, Watercolor2k, および Comic2k の mAP をそれぞれ 7.7%, 6.1%, および 2.8% 向上させる. AT に特徴アライメントを

表 2 Cityscapes から FoggyCityscapes へのドメイン適応の結果. テストデータでの mAP (%) を報告している.

Method	FoggyCityscapes
ST	27.2
ST + FA	31.1
AT (ours)	9.5
AT + FA (ours)	13.5

適用した提案手法 (AT+FA) は, ST より mAP をそれぞれ 11.8%, 6.4%, および 6.7% 向上させる. これは ST に特徴アライメントを適用した場合 (ST+FA) や AT よりも上がり幅が大きい.

以上の結果をまとめると, 画像スタイルのドメインシフトでは, AT はターゲットドメインでの検出精度を向上させ, FA を適用することでより高い検出精度を達成する. FA のような一般的なドメイン適応手法ではターゲットドメインの訓練データを使用するのに対し, AT ではソースドメインの訓練データだけを使用してターゲットドメインの検出精度を向上できるという特徴がある. これらの結果は以下のように説明できる. 画像スタイルのドメインシフトでは物体の外観が大きく変化するため, ST で抽出されるソースドメイン固有の非ロバストな特徴は, ターゲットドメインでは有用な情報を持たない. 結果として, ST はターゲットドメインでの精度低下を引き起こす. 一方で, AT で抽出されるロバストな特徴はドメイン不変な性質を持つことから, ターゲットドメインでの精度低下を緩和することができる. AT に FA を適用することによって, ロバストな特徴がターゲットドメインでの整合性を持ち, さらなる精度向上につながる.

天候のドメインシフト 次に, 天候のドメインシフトでの実験結果を報告する. Cityscapes から FoggyCityscapes へのドメイン適応の結果を表 2 に示す. 画像スタイルのドメインシフトでの結果とは対照的に, 提案手法である AT および AT+FA の mAP は, ST の mAP よりそれぞれ 17.7% および 13.7% 低下する. このドメイン適応の設定では, ST+FA が全クラスの物体で最も高い AP を獲得する.

天候のドメインシフトでは, ソースドメインとターゲットドメイン間で物体の外観はあまり変わらない. そのため, ST で抽出される非ロバストな特徴はターゲットドメインでも十分に有用な情報を持つことが考えられる. ドメインシフトを考慮しない場合, 非ロバストな特徴は深層モデルの高い予測精度に貢献することから, AT による非ロバストな特徴抽出の抑制は精度の低下を引き起こすことが知られている [12], [18]. これと同様に, ST で抽出される非ロバストな特徴がターゲットドメインでも有用な情報を持つと考えられる今回の設定では, AT によってターゲットドメインでの検出精度が低下する. 以上の結果より, ソースドメインとターゲットドメイン間のドメインシフトの大きさが AT の検出精度に影響を及ぼすことが考えられる. ド

表 3 MS COCO から土, コンクリタイル, および雪の工事現場へのドメイン適応の結果. 各テストデータでの mAP (%) を報告している.

Method	土	コンクリタイル	雪
ST	75.9	96.0	13.1
ST + FA	-	-	10.1
AT (ours)	87.2	98.1	28.1
AT + FA (ours)	-	-	26.5

表 4 一般的な工事現場から土, コンクリタイル, および雪の工事現場へのドメイン適応の結果. 各テストデータでの mAP (%) を報告している.

Method	土	コンクリタイル	雪
ST	99.6	99.7	16.8
ST + FA	-	-	24.4
AT (ours)	99.6	99.7	38.0
AT + FA (ours)	-	-	43.6

メインシフトの大きさと検出精度の関係については, 4.4 節で調査する.

場所のドメインシフト 最後に, 場所のドメインシフトでの実験結果を報告する. MS COCO から土, コンクリタイル, および雪の工事現場へのドメイン適応の結果を表 3 に示す. 土およびコンクリタイルの工事現場では訓練データが存在しないため, ST と AT の結果のみを報告している. AT は土, コンクリタイル, および雪の工事現場での人物検出の mAP をそれぞれ 11.3%, 2.1%, および 15.0% 向上させる. 雪の工事現場では, ST+FA は ST より mAP を 3.0% 低下させ, AT+FA は AT より mAP を 1.6% 低下させる. MS COCO のような汎用的な画像データセットと工事現場の画像データセット間では, 人物の映り方や背景の変化によるドメインシフトが大きい. AT で抽出されるロバストな特徴はこのようなドメインシフトに対しても堅牢であり, ターゲットドメインでの検出精度の向上に貢献する. 一方で, 本研究で使用した FA は, 色やテクスチャといった局所的な特徴をアライメントするものであり, 構図が大きく変わるようなドメインシフトに対しては有効でないことが示唆される. 局所的な特徴アライメントの代わりに, 深層モデルのより深い層での大域的な特徴をアライメントする方法を採用することで結果が改善する可能性がある.

次に, 一般的な工事現場から土, コンクリタイル, および雪の工事現場へのドメイン適応の結果を表 4 に示す. 土およびコンクリタイルの工事現場では, ST と AT による mAP に差がないことが分かる. これは, 一般的な工事現場とこれらのデータセット間では人物の映り方や背景の変化が小さく, さらにテストデータの検出難易度も非常に低いためである. このように, ドメインシフトが小さい場合やターゲットドメインの検出難易度が低い場合は, AT による大幅な精度向上は見込まれない. 一方で, 雪の工事現場では, AT および AT+FA はそれぞれ ST より 21.2% お

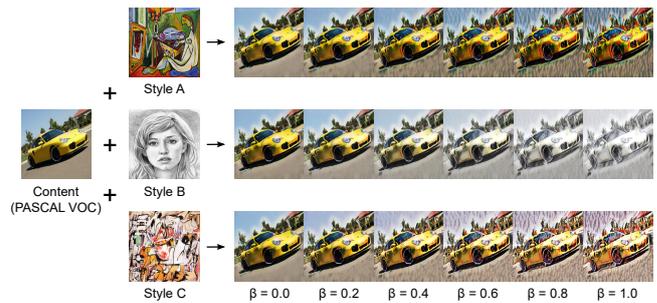


図 7 使用した三枚のスタイル画像と AdaIN によるスタイル変換の例. コンテンツ画像とスタイル画像のバランス β を変化させながら PASCAL VOC のテストデータをスタイル変換する.

よび 26.8% 高い mAP を獲得する. 一般的な工事現場と雪の工事現場のデータセット間では, 構図の変化は小さいが背景の変化によるドメインシフトが大きく, さらにテストデータでの検出難易度も高い. このような場合は, AT によるドメイン不変な特徴抽出が教師なしドメイン適応の手法として特に有効であり, FA を加えることで検出精度がさらに向上することが分かる.

4.4 ドメインシフトの大きさと検出精度の関係

画像スタイルや場所のような大きなドメインシフトでは AT によりターゲットドメインでの検出精度が向上するが, 天候のような小さなドメインシフトでは AT により検出精度が低下する. そのため, 提案手法の有効性を事前に判定するには, ドメインシフトの大きさを定量的に測定する必要がある. 画像集合の分布間の距離を測る方法として Fréchet Inception Distance (FID) [24] が知られている. FID は画像認識モデル Inception-v3 [25] の特徴空間で二つの分布間の Fréchet 距離を計算することで求められる. 本研究では, Inception-v3 の代わりに YOLOv3 の特徴空間で Fréchet 距離を計算することで, ドメインシフトの大きさを測定する. これは, 画像認識モデルの Inception-v3 よりも物体検出モデルの YOLOv3 の方が, 本研究で扱う物体検出タスクに適した特徴抽出が可能であるためである. 以下の実験では, MS COCO で事前学習された YOLOv3 を使用し, そのバックボーンネットワークから抽出される特徴マップを利用して Fréchet 距離を計算する.

人工的なドメインシフトで Fréchet 距離の測定 最初に, PASCAL VOC で人工的にドメインシフトを生じさせる実験によって, Fréchet 距離がドメインシフトの大きさを測る方法として相応しいかを調べる. PASCAL VOC の訓練データとテストデータ間にドメインシフトを生じさせるために, テストデータに対して adaptive instance normalization (AdaIN) [26] によるスタイル変換を適用する. AdaIN ではスタイル変換時にコンテンツ画像とスタイル画像のバランス (content-style trade-off) を調整することができ, 実験ではこのバランス β を 0.0 から 1.0 まで 0.1

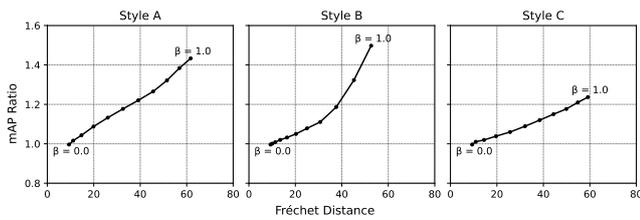


図 8 Fréchet 距離と敵対的訓練 (AT) によるドメイン適応性能の関係. 横軸: PASCAL VOC の訓練データと, 各スタイル画像および β を用いてスタイル変換されたテストデータ間の Fréchet 距離. 縦軸: 訓練データで通常訓練 (ST) されたモデルの mAP に対する AT されたモデルの mAP の比.

おきに変化させることでドメインシフトの大きさを制御する. β が大きいほど, PASCAL VOC の訓練データとテストデータ間のドメインシフトが大きくなる. 実験に使用した三枚のスタイル画像とスタイル変換された PASCAL VOC のテストデータの例を図 7 に示す. このように, スタイル画像毎に 11 段階の強さでスタイル変換を適用した PASCAL VOC のテストデータを作成する.

スタイル変換されていない PASCAL VOC の訓練データと, それぞれのスタイル画像および β でスタイル変換された PASCAL VOC のテストデータ間で Fréchet 距離を計算する. さらに, 訓練データで ST されたモデルと AT されたモデルを用意し, スタイル変換されたテストデータの mAP を調べる. ST と AT による検出精度の違いを調べるために, ST の mAP に対する AT の mAP の比を計算する. 実験結果を図 8 に示す. β が大きくなるにつれて訓練データとスタイル変換されたテストデータ間の Fréchet 距離が大きくなるのが分かる. それに伴い, ST の mAP に対する AT の mAP の比も大きくなる. これらの結果から, Fréchet 距離によってドメインシフトの大きさを測ることができ, ドメイン適応における AT の有効性を事前に予測できることが分かる.

ドメイン適応データセットで Fréchet 距離の測定 次に, 本研究で使用したドメイン適応データセットでドメインシフトの大きさを測定する. 具体的には, ソースドメインの訓練データとターゲットドメインのテストデータ間で Fréchet 距離を計算する. 各データセットでの Fréchet 距離, ST の mAP, AT の mAP, および ST の mAP に対する AT の mAP の比を表 5 に示す. PASCAL VOC から Clipart1k, Watercolor2k, および Comic2k のドメインシフトでは, Fréchet 距離はそれぞれ 44.2, 42.1, および 59.0 であり, いずれも AT による精度向上が確認される.

Cityscapes から FoggyCityscapes のドメインシフトでは, Fréchet 距離は 9.6 で他のデータセットと比べて非常に小さく, AT による精度低下が確認される. このように Fréchet 距離が小さい場合は, ST で抽出される非ロバストな特徴がターゲットドメインでも十分に有用であることから, AT によるドメイン適応が必要でないことが事前に

分かる. mAP 比が 0.35 で, 図 8 の人工的なドメインシフトの結果と比較して小さい原因として, FoggyCityscapes に小さな物体が多いことが挙げられる. PASCAL VOC のような様々なシーンでの画像データセットと比べると, Cityscapes や FoggyCityscapes のような車載カメラからの画像データセットでは遠くに小さく映る物体が多い. このような小さな物体の検出には, ST で抽出される非ロバストな特徴が特に重要であると考えられる. したがって, 非ロバストな特徴を抑制する AT は, FoggyCityscapes で大幅な精度低下を引き起こす.

MS COCO から土, コンクリタイル, および雪の工事現場のドメインシフトでは, Fréchet 距離はそれぞれ 84.9, 81.0, および 107.3 と比較的大きく, いずれも AT による精度向上が確認される. 一般的な工事現場から土およびコンクリタイルの工事現場のドメインシフトでは, Fréchet 距離はそれぞれ 19.9 および 20.8 と比較的小さく, ST と AT の mAP は変わらない. 最後に, 一般的な工事現場から雪の工事現場のドメインシフトでは, Fréchet 距離は 65.2 と比較的大きく, AT による大幅な精度向上が確認される.

以上の結果より, ドメイン適応データセットにおいても, ソースドメインとターゲットドメイン間の Fréchet 距離が大きいほど, AT によってターゲットドメインでの検出精度が向上する傾向にあることが分かる. 4.3 節で示したように, データセットによって AT が有効である場合とそうでない場合が存在するが, Fréchet 距離を計算することによって AT の有効性を深層学習前に判定することができる.

5. おわりに

本研究では, 物体検出における教師なしドメイン適応に取り組んだ. 敵対的訓練された深層モデルによって抽出されるロバストな特徴がドメイン不変な性質を持つことに着目し, ソースドメインでの敵対的訓練によるドメイン適応アプローチを提案した. さらに, 敵対的訓練に特徴アライメントを加えた学習方法を提案し, ドメイン適応性能の向上を目指した. 画像スタイルおよび場所のドメイン適応実験では, 提案手法によってターゲットドメインの検出精度が向上することを示した. 一方で, 天候のドメイン適応実験では, 敵対的訓練はターゲットドメインの検出精度の低下を引き起こした. これらの結果を踏まえ, Fréchet 距離によってドメインシフトの大きさを測定した結果, Fréchet 距離が大きいほど提案手法によってターゲットドメインでの検出精度が向上する傾向にあることを発見した. これを用いれば, 深層学習前に提案手法の有効性を予測することができる. 本研究では, 敵対的訓練により教師無しドメイン適応が可能になることを示したが, 実験に使用した多くのデータセットでは十分に信頼できるほどの検出精度には至っていない. ドメイン適応に特化した敵対的訓練の方法などを検討して, より高い検出精度を目指す必要がある.

表 5 ソースドメインの訓練データとターゲットドメインのテストデータ間の Fréchet 距離と、ソースドメインで通常学習 (ST) および敵対的訓練 (AT) したモデルのターゲットドメインでの mAP (%) とその比。

ソースドメイン	ターゲットドメイン	Fréchet 距離	ST の mAP	AT の mAP	mAP 比
PASCAL VOC	Clipart1k	44.2	37.6	45.3	1.20
PASCAL VOC	Watercolor2k	42.1	49.2	55.3	1.12
PASCAL VOC	Comic2k	59.0	26.7	29.5	1.10
Cityscapes	FoggyCityscapes	9.6	27.2	9.5	0.35
MS COCO	土の工事現場	84.9	75.9	87.2	1.15
MS COCO	コンクリタイルの工事現場	81.0	96.0	98.1	1.02
MS COCO	雪の工事現場	107.3	13.1	28.1	2.14
一般的な工事現場	土の工事現場	19.9	99.6	99.6	1.00
一般的な工事現場	コンクリタイルの工事現場	20.8	99.7	99.7	1.00
一般的な工事現場	雪の工事現場	65.2	16.8	38.0	2.26

謝辞 本研究は住友建機株式会社からの支援および JSPS 科研費 JP20K23341 の助成を受けたものです。

参考文献

- [1] Ren, S., He, K., Girshick, R. and Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *NeurIPS* (2015).
- [2] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A. C.: SSD: Single Shot MultiBox Detector, *ECCV*, pp. 21–37 (2016).
- [3] Redmon, J. and Farhadi, A.: YOLOv3: An Incremental Improvement, *arXiv preprint arXiv:1804.02767* (2018).
- [4] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge, *IJCV*, Vol. 88, No. 2, pp. 303–338 (2010).
- [5] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: Microsoft COCO: Common Objects in Context, *ECCV*, pp. 740–755 (2014).
- [6] Zhao, S., Yue, X., Zhang, S., Li, B., Zhao, H., Wu, B., Krishna, R., Gonzalez, J. E., Sangiovanni-Vincentelli, A. L., Seshia, S. A. and Keutzer, K.: A Review of Single-Source Deep Unsupervised Visual Domain Adaptation, *TNNLS*, pp. 1–21 (2020).
- [7] Oza, P., Sindagi, V. A., VS, V. and Patel, V. M.: Unsupervised Domain Adaptation of Object Detectors: A Survey, *arXiv preprint arXiv:2105.13502* (2021).
- [8] Chen, Y., Li, W., Sakaridis, C., Dai, D. and Van Gool, L.: Domain Adaptive Faster R-CNN for Object Detection in the Wild, *CVPR*, pp. 3339–3348 (2018).
- [9] Saito, K., Ushiku, Y., Harada, T. and Saenko, K.: Strong-Weak Distribution Alignment for Adaptive Object Detection, *CVPR*, pp. 6949–6958 (2019).
- [10] Kim, S., Choi, J., Kim, T. and Kim, C.: Self-Training and Adversarial Background Regularization for Unsupervised Domain Adaptive One-Stage Object Detection, *ICCV*, pp. 6091–6100 (2019).
- [11] Hsu, H.-K., Yao, C.-H., Tsai, Y.-H., Hung, W.-C., Tseng, H.-Y., Singh, M. and Yang, M.-H.: Progressive Domain Adaptation for Object Detection, *WACV*, pp. 738–746 (2020).
- [12] Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B. and Madry, A.: Adversarial Examples Are Not Bugs, They Are Features, *NeurIPS* (2019).
- [13] Salman, H., Ilyas, A., Engstrom, L., Kapoor, A. and Madry, A.: Do Adversarially Robust ImageNet Models Transfer Better?, *NeurIPS* (2020).
- [14] Utrera, F., Kravitz, E., Erichson, N. B., Khanna, R. and Mahoney, M. W.: Adversarially-Trained Deep Nets Transfer Better: Illustration on Image Classification, *ICLR* (2021).
- [15] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M. and Lempitsky, V.: Domain-Adversarial Training of Neural Networks, *JMLR*, Vol. 17, No. 59, pp. 1–35 (2016).
- [16] Zhu, J.-Y., Park, T., Isola, P. and Efros, A. A.: Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks, *ICCV*, pp. 2242–2251 (2017).
- [17] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J. and Fergus, R.: Intriguing properties of neural networks, *ICLR* (2014).
- [18] Tsipras, D., Santurkar, S., Engstrom, L., Turner, A. and Madry, A.: Robustness May Be at Odds with Accuracy, *ICLR* (2019).
- [19] Goodfellow, I. J., Shlens, J. and Szegedy, C.: Explaining and Harnessing Adversarial Examples, *ICLR* (2015).
- [20] Inoue, N., Furuta, R., Yamasaki, T. and Aizawa, K.: Cross-Domain Weakly-Supervised Object Detection Through Progressive Domain Adaptation, *CVPR*, pp. 5001–5009 (2018).
- [21] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding, *CVPR*, pp. 3213–3223 (2016).
- [22] Sakaridis, C., Dai, D. and Van Gool, L.: Semantic Foggy Scene Understanding with Synthetic Data, *IJCV*, Vol. 126, No. 9, pp. 973–992 (2018).
- [23] Bochkovskiy, A., Wang, C.-Y. and Liao, H.-Y. M.: YOLOv4: Optimal Speed and Accuracy of Object Detection, *arXiv preprint arXiv:2004.10934* (2020).
- [24] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, *NeurIPS* (2017).
- [25] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z.: Rethinking the Inception Architecture for Computer Vision, *CVPR*, pp. 2818–2826 (2016).
- [26] Huang, X. and Belongie, S.: Arbitrary Style Transfer in Real-Time With Adaptive Instance Normalization, *ICCV*, pp. 1510–1519 (2017).