

分類性能による制約を考慮した 敵対的不変表現学習によるドメイン汎化

Adversarial Invariant Feature Learning with Accuracy Constraint for Domain Generalization

阿久澤 圭^{*1} 岩澤 有祐^{*1} 松尾 豊^{*1}
Kei Akuzawa Yusuke Iwasawa Yutaka Matsuo

^{*1} 東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo

Learning domain-invariant representation is a dominant approach for domain generalization, where we need to build a classifier that is robust toward domain shifts. However, previous methods based on domain invariance overlooked the underlying dependency of classes on domains, which is responsible for the trade-off between classification accuracy and domain invariance. This study proposes a novel method *adversarial feature learning under accuracy constraint (AFLAC)*, which maximizes domain invariance within a range that does not interfere with classification accuracy. The reason for the constraint is that the primary purpose of domain generalization is to classify unseen domains rather than the invariance itself, and improving the invariance can negatively affect that performance. Empirical validations show that the performance of AFLAC is superior to that of baseline methods, supporting the importance of considering the dependency and the efficacy of the proposed method to overcome the problem.

1. はじめに

教師あり学習では典型的にデータが訓練時・テスト時で i.i.d. であるという仮定を置いているが、これが成り立たないとき機械学習モデルのテストデータに対する予測精度が著しく低下する [Torralba 11]. ドメイン汎化は、このような状況を想定した研究領域の一つであり、複数のソースドメインからラベル付き訓練データが得られるという設定のもと、それらの訓練データを利用してターゲットドメインから得られるテストデータへの予測を行う。ドメイン汎化は手書き文字認識 [Shankar 18], そしてユーザーにロバストな加速度センサーからの行動認識 [Erfani 16] など、様々な応用先を持つ技術である。

本研究では、ドメイン d とクラスラベル y が何かしらの共通要因 z によって統計的に従属する状況下 (図 1-(c)) でドメイン汎化に取り組む。そのような状況の例に WISDM Activity Prediction データセット ([Kwapisz 11], 以降 WISDM) がある。WISDM では人間がドメイン、人間の行動がクラスに該当するが、いくつかの行動 (ジョギング, 階段を上る) が激しい運動を必要とするため、そうした運動を敬遠するユーザーの存在によって従属性が生じる。また別のいくつかの行動 (座る, 立つ) がデータセット収集作業の途中から付け加えられたために、あるユーザーらについてはそれらの行動に関するデータを全く得ることができない。このような状況は現実のデータセットに一般的でありドメイン適応では取り組まれている [Zhang 13] にも関わらずドメイン汎化では無視されてきた。

ドメイン汎化に取り組むために、多くの手法が不変表現学習を利用している [Muandet 13, Erfani 16, Xie 17]. 不変表現学習では、入力データ x を特徴量 h に写像したとき、 h がドメイン d に関する情報を持たなくなる、あるいは複数のドメインの分布を特徴量空間で近づけるような制約を置く。このようにして得た h から y への予測を行うことで、予測が特定の d に対して過剰適合しなくなり、テスト時に現れるターゲットドメインに対しても正しい予測を行えることが期待できる (図

1-(b)). しかし、ドメインとクラスが従属する時、単にドメインに不変な表現を学習しようとする分類性能を妨げてしまう危険性がある。直感的に言えば、ドメインとクラスが従属するとき y は d に関する情報を持っているので、 h が d に関する情報を持つことは y に対する予測精度の向上につながる。しかし、不変表現学習は h が d に関する情報を全く持たなくなるような制約をかけるので、不変性と分類性能のトレードオフが生じる。このトレードオフは (訓練時にはソースドメインのみを用いるため) ソースドメインに対して生じるが、ターゲットドメインに対する分類性能を損なう危険性も抱えている。例えばもしターゲットドメインがソースドメインに似ている時、あるいは極端な場合ターゲットドメインとあるソースドメインが全く同じ性質を持つとき、ソースドメインに対する分類性能を損なうことは明らかにターゲットドメインに対する分類性能を損なうことにつながる (図 1-(d)).

本研究では、不変性を優先することはドメイン汎化性能を損なう危険があることから、分類性能を損なわない範囲内で不変性を最大化することを提案する。提案手法 AFLAC (Adversarial Feature Learning under Accuracy Constraint) は、既存の敵対的不変表現学習手法の Domain Adversarial Networks (DAN, [Ganin 16, Xie 17]) の改善手法であり、分類性能による制約付きドメイン不変性を達成するように意図されている。分類性能による制約付きドメイン不変性は分類性能を妨げない範囲内で達成できる $H(d|h)$ (H はエントロピーを指す) の最大値であり $H(d|y)$ に等しい。実験では AFLAC がベースライン手法に比べて高いドメイン汎化性能を持つことを示し、ドメインとクラスの従属関係を考慮することの重要性、および AFLAC がその従属性が引き起こすトレードオフの問題に対処する能力を持つことを示す。

2. 関連研究

不変表現学習のドメイン汎化に対する有効性は [Muandet 13] によって初めて示された。DAN は敵対的学習に基づいて End-to-End に不変表現学習を行う手法であり、本研究の提案手法 AFLAC の基盤となっている。DAN は当初ドメイン適応のために提案された [Ganin 16] が、[Xie 17] によってドメイン汎

連絡先: 阿久澤 圭, 東京大学大学院工学系研究科技術経営戦略学専攻, 〒113-8656 東京都文京区本郷 7-3-1, akuzawa-kei@weblab.t.u-tokyo.ac.jp

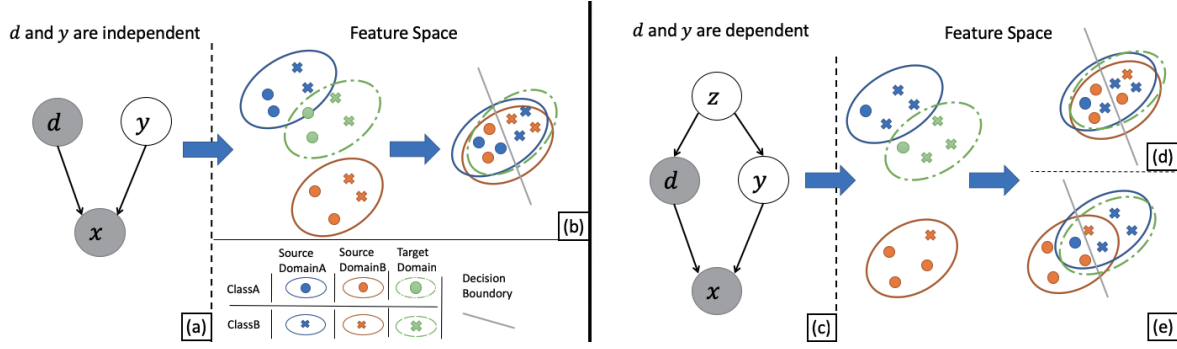


図 1: ドメインとクラスの従属性が引き起こす分類性能と不変性のトレードオフの概要. (a) ドメインとクラスが独立なとき, (b) 完全なドメイン不変性と最適な分類性能は同時に達成することができる. (c) ドメインとクラスが従属するとき, ドメイン不変性と分類性能の間にトレードオフが存在する. すなわち, (d) 完全なドメイン不変性が達成されているときは最適な分類性能を達成することができず, また (e) 逆も同じである.

化における有効性が示されている. [Xie 17] はまた, 本研究が取り組む分類性能と不変性のトレードオフについて指摘しているが, その問題への対処法については十分に議論していない.

不変表現学習を用いずにドメイン汎化に取り組む研究もいくつか存在する. 例えば [Li 18] は Semantic Alignment と呼ばれるアプローチでドメイン汎化に取り組んでいる. Semantic Alignment では, クラスで条件づけた潜在表現が全てのソースドメインで同じになるような制約をかけるが, トレードオフの問題に対処できるかは明らかではない. CrossGrad [Shankar 18] は近年 state-of-the-art の性能を示したドメイン汎化手法の一つであり, 敵対的例を生成することでデータ拡張を行う. しかしこの手法は y と d が統計的に独立であることを仮定しているため, 本研究の設定にそのまま適用可能ではない.

ドメイン適応においては [Zhang 13] が $p(y)$ がソース, ターゲットドメインで変化するような設定に取り組んでいる. [Zhang 13] は $p(y)$ の変化を推定し, ソースドメインに対する分類精度を犠牲にすることでその分布の変化を矯正し, ターゲットドメインに対する分類精度を高めている. しかしそのような手法はドメイン汎化には適用可能でない, または適用する必要がない. なぜなら, ドメイン汎化においてはターゲットドメインのデータは全く得ることができないためソースとターゲットの分布の違いには対処しようがなく, また本研究が対象としているのは複数あるソースドメイン間の分布の変化である. 代わりに本研究では複数あるソースドメインに対する不変性を高めつつ分類性能を最大化する手法を提案する.

3. モデル

3.1 既存手法: DAN

本節では, 提案モデルの基盤となる DAN に関する説明を行う. DAN は敵対的学習に基づいて不変表現学習を行う手法であり, エンコーダーが出力した潜在表現からドメインを識別しようとするドメイン識別器を持つ. ドメイン識別器がドメインを予測しようとする一方で, エンコーダーがその識別器を騙すように学習することで, 潜在表現は「ドメイン識別器が正しく識別を行うことができないような表現」, すなわちドメインに関する情報を全く持たないような表現になる.

$f_E(x)$, $q_M(y|h)$, $q_D(d|h)$ (E, M, D はパラメータ) をそれぞれエンコーダー, クラスラベルの分類器 (以降分類器), ドメインラベルの識別器 (以降識別器) とする. DAN の目的関数

は以下のように書ける.

$$\begin{aligned} \min_{E, M} \max_D J(E, M, D) &= \mathbb{E}_{x, d, y \sim p(x, d, y)} [-\gamma L_d + L_y] \\ &= \mathbb{E}_{x, d, y \sim p(x, d, y)} [\gamma \log q_D(d|h = f_E(x)) \\ &\quad - \log q_M(y|h = f_E(x))] \end{aligned} \quad (1)$$

ここで, 式 1 の L_y と L_d の最小化は通常の分類問題と同様, 単に分類器と識別器の分類誤差を最小化しているだけである. ただし, 第一項はエンコーダーと識別器の間のミニマックスゲームに相当し, 識別器は潜在表現 h から d を当てようとするのに対して, エンコーダーは識別器を騙そうとしている. DAN の訓練の様子は図 2-(a) に示されている.

3.2 提案手法: AFLAC

本研究では DAN の改善手法 AFLAC を提案する. DAN の正則化項は完全なドメイン不変性, すなわち $H(d|h) = H(d)$ を達成するような正則化を持つ [Xie 17] が, それに対して AFLAC は分類性能による制約付きドメイン不変性 $H(d|h) = H(d|y)$ を達成するような正則化を持つ. $H(d|h) = H(d|y)$ は学習された表現 h がクラスラベル y と同じだけドメイン d に関する情報を含むことを意味するが, そのような表現は分類性能を損なうことがないと考えられる.

AFLAC は DAN と同様にエンコーダー, 分類器, 識別器を持つ. そして以下の二つの最適化問題を, 交互に勾配法によるパラメータの更新を行うことによって解く.

$$\begin{aligned} \min_{E, M} V(E, M) &= \mathbb{E}_{x, d, y \sim p(x, d, y)} [\gamma L_{D_{KL}} + L_y] \\ &= \mathbb{E}_{x, d, y \sim p(x, d, y)} [\gamma D_{KL}[p(d|y)|q_D(d|h = f_E(x))] \\ &\quad - \log q_M(y|h = f_E(x))] \end{aligned} \quad (2)$$

$$\begin{aligned} \min_D W(E, D) &= \mathbb{E}_{x, d \sim p(x, d)} [L_d] \\ &= \mathbb{E}_{x, d \sim p(x, d)} [-\log q_D(d|h = f_E(x))] \end{aligned} \quad (3)$$

DAN と同様に, 式 2 の第二項は q_M , 式 3 は q_D の尤度最大化を表している. 一方式 2 の第一項は DAN と異なり, Kullback-Leibler divergence (KLD) の最小化を通してすべての確率が 0 ではない y と h のペアに関して $q_D(d|h) = p(d|y)$ を成立させるような働きを持つ. ここで $q_D(d|h)$ は L_d の最小化によって $p(d|h)$ を近似するように学習するが, その近似が十分に行われているとき, $D_{KL}[p(d|y)|q_D(d|h)]$ の最小化は $H(d|h) = H(d|y)$ を達成する.

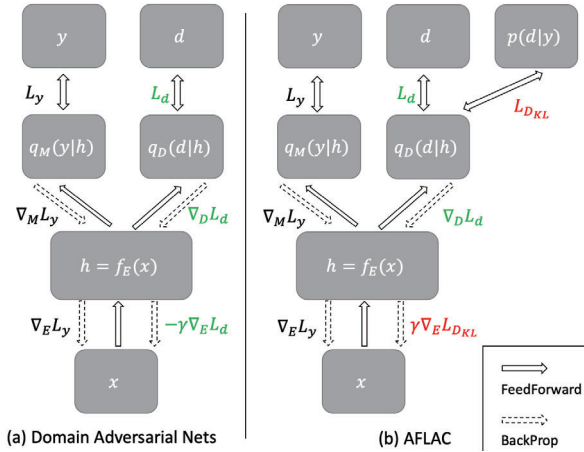


図 2: (a)DAN と (b)AFLAC の比較. (a) 分類器と識別器は L_y と L_d を最小化するように、エンコーダーは L_y を最小化、 L_d を最大化するように学習する. (b) 識別器は L_d の最小化によって $p(d|h)$ を近似するように、エンコーダーは L_{DKL} の最小化によって $p(d|h)$ と $p(d|y)$ を近づけるように学習する.

実際の訓練時には真の $p(d|y)$ を手に入れることはできないが、その最尤推定量や MAP 推定量を用いることができる。また、式 2 の $D_{KL}[p(d|y)|q_D(d|h)]$ に関しては他の分布間距離、例えば L1 距離や $D_{KL}[q_D(d|h)|p(d|y)]$ も考えられるが、大きな精度の改善が見られなかったため本研究では考えないこととする。AFLAC の訓練の概要は、図 2-(b) に示されている。

4. 実験

4.1 データセット

BMNISTR Biased Rotated MNIST (以降 BMNISTR) はドメイン汎化のための標準的なデータセットである MNISTR[Ghifary 15] を元に、本研究がドメインとクラスが従属するようにサンプルサイズに修正を加えたものである。MNISTR や BMNISTR では、それぞれのクラスは 0 から 9 のアラビア数字に対応し、それぞれのドメインは画像の傾き (0, 15, 30, 45, 60, 75 度) に対応する。またそれぞれの傾きは M0, M75 のように表記する。それぞれの画像は [Ghifary 15] の実験設定と同様に 16 x 16 の大きさを持つ。本研究は BMNISTR-1 から BMNISTR-4 までの、それぞれ異なるドメインとクラスの従属性を持つ 4 つのデータセットを作り出した。表 1 が示すように、BMNISTR-1, -2, -3 は似たような傾向の従属性を持つがその強さが異なっている。一方で、BMNISTR-4 はその他のデータセットとは異なる傾向の従属性を持っている。訓練では、一つのドメインをテストドメインとして、残り全てのドメインを訓練データとする設定で行った。また BMNISTR に対しては、エンコーダーを二層の畳み込み層と二層の全結合層、クラス分類機を三層の全結合層、ドメイン識別器を二層の全結合層からなるディープニューラルネットとした。

WISDM WISDM データセットは 36 人のユーザーによる 6 つの行動 (walking, jogging, upstairs, downstairs, sitting, and standing) を加速度計によって計測したセンサーデータによって構成される。このデータセットではユーザーがドメイン、行動がクラスに相当し、ユーザーに対して頑健な予測を行うことが目標となる。WISDM は 1. 章で述べた理由からドメインとクラスの従属性を持つ。WISDM では、ランダムに選択した $<10 / 26>$, $<26 / 10>$ 人のユーザーを $<$ ソース/ター

表 1: BMNISTR における各ドメイン・クラスに対するサンプルサイズ。クラス 0 から 4 に対するサンプルサイズはドメインによって異なる一方で、クラス 5 から 9 に対するサンプルサイズは全てのドメインで同一となっている。

Dataset	Class	M0	M15	M30	M45	M60	M75
BMNISTR-1	0~4	100	85	70	55	40	25
	5~9	100	100	100	100	100	100
BMNISTR-2	0~4	100	80	60	40	20	0
	5~9	100	100	100	100	100	100
BMNISTR-3	0~4	100	90	80	70	60	50
	5~9	100	100	100	100	100	100
BMNISTR-4	0~4	100	25	100	25	100	25
	5~9	100	100	100	100	100	100

表 2: BMNISTR でターゲットドメインを M0 としたときのクラス 0 から 4 と 5 から 9 に対する平均の F 値。RI は AFLAC-Abl から AFLAC の相対改善率を表す。

Dataset	Class	CNN	DAN	CIDDG	AFLAC-Abl	AFLAC	RI
BMNISTR-1	0~4	83.86	84.54	87.50	87.46	90.62	3.6%
	5~9	83.90	85.24	87.46	86.46	88.10	1.9%
BMNISTR-2	0~4	84.76	86.20	88.52	86.42	89.58	3.7%
	5~9	83.36	85.22	87.02	85.62	86.86	1.4%
BMNISTR-3	0~4	82.54	85.30	87.64	88.60	89.64	1.2%
	5~9	82.18	85.80	86.74	87.60	89.04	1.6%
BMNISTR-4	0~4	71.26	79.22	76.76	76.56	80.02	4.5%
	5~9	78.62	83.14	82.64	82.94	82.80	-0.2%

ゲット > ユーザーとして用いた。データの前処理として、60 フレーム (3 秒間に相当) を一つのサンプルとしてデータセットを構築したところ、合計で 18210 サンプルとなった。モデル構造はエンコーダーを三層の畳み込み層と一層の全結合層、クラス分類器を一層の全結合層、ドメイン識別器を二層の全結合層とした。

4.2 ベースライン

提案手法の有効性を示すために、本研究では以下の手法との比較実験を行う。(1) **CNN** は通常の畳み込み層で形成されたディープネットである。(2) **DAN** [Xie 17] は 3.1 節で説明した、敵対的学習による不変表現学習を利用したドメイン汎化手法である。(3) **CIDDG** は [Li 18] の提案モデルを我々が再現実装したものであり、Semantic Alignment を利用している。(4) **AFLAC-Abl** は ablation study のために用意した AFLAC の変種である。AFLAC-Abl は AFLAC の目的関数である式 2 の $D_{KL}[p(d|y)|q_D(d|h)]$ を $D_{KL}[p(d)|q_D(d|h)]$ によって置き換えたモデルであり、DAN と同様に完全にドメイン不変な表現を得る、すなわち $H(d|h) = H(d)$ を成立させるような正則化項を持つ。AFLAC と AFLAC-Abl を比べることで、ドメインとクラスの従属性が引き起こすトレードオフの問題を考慮することがドメイン汎化性能に与える影響を観察することができる。AFLAC と AFLAC-Abl の訓練では、2 式の KLD の計算に必要な真の $p(d|y)$ と $p(d)$ を得ることができないため、それらの最尤推定量を代わりに用いた。

4.3 様々な種類の従属の下での比較

本節ではドメインとクラスの従属性が、ドメイン不変性を利用したドメイン汎化手法の性能に与える影響について確認する。表 2 は BMNISTR でターゲットドメインを M0 とした際の、クラス 0 から 4 と 5 から 9 のそれぞれに対する F 値の平均を示している。ここで、クラス 0 から 4 のサンプルサイズはドメインごとに異なるが、クラス 5 から 9 は全てのドメインで同じであることに注意する (表 1)。表 2 が示すよ

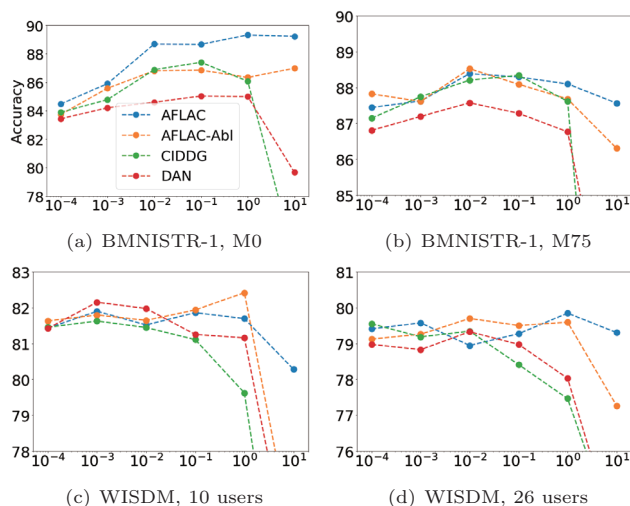


図 3: 様々な γ の値の下での正解率の比較. 各キャプションはデータセットの名前とターゲットドメインを表している.

うに, AFLAC の性能はベースライン手法の性能をほとんどのデータセットおよびクラスについて上回っているが, これはドメイン不変性を利用した DAN や AFLAC-Abl といった手法の性能がドメインとクラスの従属性によって損なわれること, そして AFLAC がその問題を緩和できることを示唆している. また AFLAC の AFLAC-Abl に対する相対改善率を見ると, BMNISTR-1, -2, -4 ではクラス 0 から 4 に対する相対改善率の方がクラス 5 から 9 に対するものよりも大きい, これは AFLAC がドメインとクラスの従属性が発生しているクラスについてより正確に予測を行えることを示唆している. また, BMNISTR-1 の相対改善率は BMNISTR-3 の相対改善率よりも大きい, これはドメインとクラスの従属性が強くなるほど, DAN や AFLAC といったドメイン不変性に基づいた手法の性能が損なわれることを示唆している. 最後に, BMNISTR-1 と BMNISTR-4 は異なる傾向のドメインとクラスの従属性を持つが, AFLAC はそのどちらのデータセットに対しても F 値を改善させている.

4.4 ハイパーパラメータに対する頑健性

次に本節では, 正則化の強さとドメイン汎化性能の関係について調べる. 図 3 は DAN, CIDDG, AFLAC-Abl, AFLAC を様々なハイパーパラメータ γ を用いて訓練した際の, y に対する正解率を示している. これらの画像から, 以下のことが示唆される. (1) 正則化項の重みを強くしたときに, AFLAC の訓練は DAN や CIDDG の訓練より安定する傾向がある. 図 3-(a, b, c, d) からは, γ の値を 1 や 10 にした時に AFLAC や AFLAC-Abl が DAN よりも高い性能を示す傾向があることが読み取れる. その理由はおそらく, AFLAC や AFLAC-Abl の制約項が KLD であり 0 によって下から抑えられているため, DAN の制約項のように重みを大きくした時に発散する心配がなく, 訓練が安定するからだと考えられる. (2) AFLAC は, 正則化を強めても分類精度が損なわれず, したがってハイパーパラメータ選択に頑健な傾向にある. 図 3-(b, c, d) は, γ を 10 のような大きな値にした時に AFLAC-Abl の y に対する正解率が大きく低下する一方で AFLAC のものはそれほど大きく低下しないことを示しているが, これは AFLAC の正則化が分類精度を損なわない範囲でドメイン不変性を最大化するように意図されているからであると考えられる.

5. 結論

本研究では, 既存研究が検討してこなかった「ドメインとクラスが統計的に従属する」という状況下でドメイン汎化を行うため, 分類性能を妨げることのない範囲でドメイン不変性を最大化する新しい手法 AFLAC を提案した. 実験では AFLAC がベースライン手法よりも優れた性能を発揮することを確認し, ドメインとクラスの従属性が引き起こすトレードオフを考慮することのドメイン汎化にける重要性, およびその問題に対する AFLAC の有効性を確認した.

参考文献

- [Erfani 16] Erfani, S., Baktashmotlagh, M., Moshtaghi, M., Nguyen, V., Leckie, C., Bailey, J., and Kotagiri, R.: Robust domain generalisation by enforcing distribution invariance, in *25th International Joint Conference on Artificial Intelligence* (2016)
- [Ganin 16] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V.: Domain-adversarial Training of Neural Networks, *J. Mach. Learn. Res.* (2016)
- [Ghifary 15] Ghifary, M., Bastiaan Kleijn, W., Zhang, M., and Balduzzi, D.: Domain Generalization for Object Recognition With Multi-Task Autoencoders, in *Proc. of the IEEE International Conference on Computer Vision (ICCV)* (2015)
- [Kwapisz 11] Kwapisz, J. R., Weiss, G. M., and Moore, S. A.: Activity Recognition Using Cell Phone Accelerometers, *SIGKDD Explor. Newsl.* (2011)
- [Li 18] Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., and Tao, D.: Deep Domain Generalization via Conditional Invariant Adversarial Networks, in *The European Conference on Computer Vision (ECCV)* (2018)
- [Muandet 13] Muandet, K., Balduzzi, D., and Schölkopf, B.: Domain Generalization via Invariant Feature Representation, in *Proc. of the 30th International Conference on Machine Learning* (2013)
- [Shankar 18] Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., and Sarawagi, S.: Generalizing Across Domains via Cross-Gradient Training, in *Proc. International Conference on Learning Representations* (2018)
- [Torralba 11] Torralba, A. and Efros, A. A.: Unbiased Look at Dataset Bias, in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition* (2011)
- [Xie 17] Xie, Q., Dai, Z., Du, Y., Hovy, E., and Neubig, G.: Controllable Invariance through Adversarial Feature Learning, in *Proc. of the 30th International Conference on Neural Information Processing Systems* (2017)
- [Zhang 13] Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z.: Domain Adaptation under Target and Conditional Shift, in *Proc. of the 30th International Conference on Machine Learning* (2013)