Domain Generalization のための MixStyle-based Contrastive Test-time Adaptation

山下 滉太[1,a)]　堀田 一弘[1,b)]

要旨

本論文は Computer Vision and Pattern Recognition 分野におけるトップカンファレンスである CVPR2024(The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024) における Workshop のフルペーパー査読を通過し，採択された論文を再構成したものである．近年，Domain Generalization 研究では，テスト時にモデルを適応する Test-time Adaptation TTA）手法への関心が高まっている．TTA の代表的手法である Test-time Training TTT）の問題点を改善するために，MixStyle-based Contrastive Test-time Adaptation MCTTA）を提案する．3 つのデータセットを用いた実験において，TTT やその他 TTA 手法，従来の Domain Generalization 手法と比較して，提案する MCTTA の有効性を示した．

1. はじめに

ディープラーニングでは，異なるドメインや環境において同等の性能を維持することは依然として難しい課題である．この課題への対処には，Domain Generalization の研究 [1][2] が重要な役割を果たしている．Domain Generalization は，学習ドメインとは異なるドメインのデータに対しても，高精度で汎用性の高いモデルを開発することを目指している．Domain Generalization に対するアプローチは，大きく以下の三つのカテゴリーに分類される．第一カテゴリーとして，Data Augmentation をはじめとするデータレベルのアプローチがある．例えば，Mixup[3] は異なるドメインの画像を重ね合わせて新しい学習画像を生成する手法であり，モデルがドメイン固有の特徴に過剰適合するのを防ぐ．Mixup が画像単位でドメイン情報を混合しているのに対し，我々の研究で利用する既存手法の MixStyle[4] では特徴量単位でドメイン情報を混合する．具体的には，CNN[5] モデルの浅い層から出力される特徴量の平均と標準偏差を Instance Normalization[6] を通じてバッチ間で混合し，ドメイン情報が混在した特徴量を生成する．第二カテゴリーとしてモデルレベルのアプローチがある．例えば SagNet[7] は content-biased network，style-biased network と呼ばれる 2 つのネットワークを通じて，画像のスタイル（例えば，色調やテクスチャ）から形状などのコンテンツを分離し，モデルがスタイルの違いに左右されずにコンテンツに基づいて判断できるようにする．第三カテゴリーとしてアルゴリズムレベルのアプローチがある．例えば，Meta-Learning for domain generalization（MLDG）[8] は，Domain Generalization のために設計された Meta-learning 手法であり，複数のドメインのデータを用いて Meta-learning を行い，モデルが未知のドメインに対しても良好なパフォーマンスを発揮する．また，近年注目されている Test-time Adaptation（TTA）[9] もこのカテゴリーに属する．TTA の主な目的は，モデルをテスト時にリアルタイムで適応することにより，学習時には未学習であったテストドメインのデータに対しても一貫した性能を発揮させることである．Test-time Training（TTT）[10] は TTA の代表的な手法であり，まず学習データを用いて解きたいタスクと Rotation Prediction のような自己教師ありタスクのマルチタスク学習を行う．その後，テストデータに対して自己教師あり学習によりモデルをテストデータに適応させ，学習データとテストデータのドメインの差異を吸収する．TTT は Domain Generalization に有効な手法であるが，いくつかの問題がある．第一に，Rotation Prediction などの一般的な自己教師タスクが，Domain Generalization に特化したタスクではなく，モデルの汎化に直接貢献しないこと，第二に，学習ドメインやテストドメインにより最適な自己教師ありタスクが異なり，不適切なタスクを選択するとモデルの性能が低下することである．自己教師ありタスクの選択は定量的な基準ではなく直観に頼る必要があり，これは適応を必要とする TTA において大きな課題である．一方，TENT[11] は自己教師タスクではなく教師なしタスクを利用する TTA である．学習時に TTT のようなマルチタスク学習を行わず，メインタスクのみを学習しておき，テストデータ時に教師なし学習である Entropy Minimization を用い，モデルの Batch

1 名城大学
a) 200442179@ccalumni.meijo-u.ac.jp
b) kazuhotta@meijo-u.ac.jp

Normalization[12] 層の更新を行うことによりモデルをテストデータに適応させる．TENT はテストデータに対する Entropy Minimization のみを行えば良く，非常に手軽な手法であるが，ドメインによっては TTT よりも精度が劣ることがある．

これらの問題に対処するため，TTT を改善した新しい TTA である MixStyle-based Contrastive Test-time Adaptation（MCTTA）を提案する．MCTTA では，MixStyle を用いた新たな Contrastive Learning である MixStyle-based Contrastive Learning（MCL）を用い，異なるドメインに対して一貫した特徴量を抽出できるよう特徴抽出器を学習する．MCL では，Data Augmentation に特徴空間での MixStyle を使用する，Domain Generalization に特化した独自の Contrastive Learning[13][14] であり，同一構造で初期値の異なる 2 つのモデルを使用する．MCTTA の学習プロセスは，Training と TTA フェーズに分かれており，Training フェーズでは最終的に解きたいタスク（以下 Classification と仮定）の Ensemble Learning と MCL を用いて Domain Generalization に特化したマルチタスク学習 [15] を行う．その後，TTA フェーズにてテストデータに MCL を実行し，特徴抽出器をテストデータに適応させる．

DomainBed[16] ベンチマークライブラリと PACS[17], Office-Home[18], Colored MNIST[19] の 3 つデータセットを用いた実験により，提案手法は TTT に対して平均 1.7% の精度向上に成功し，その他 TTA 手法，従来の Domain Generalization 法と比較しても最も高精度を達成した．また，MCL が Domain Generalization に関して従来の Contrastive Learning よりも優れていることと，2 つのモデルを用いて Ensemble Classification とモデル間で MCL を行うことの有効性も確認した．

本論文の構成は以下の通りである．2 節で提案手法を説明し，3 節で実験結果を示す．最後に，4 節で結論を述べる．

2. 提案手法

本稿では，TTT の課題に対する非常にシンプルかつ強力な解決策として MCTTA を提案する．第一の課題に対する解決策として MCTTA では，Domain Generalization に特化した独自の Contrastive Learning である MCL を用いる．色変換やランダムなトリミングによる従来の Contrastive Learning とは異なり，MCL は Data Augmentation として，特徴空間で MixStyle を使用する．この MCL により，モデルは様々なドメインにわたって一貫した特徴量で識別ができるようになる．なお，MCL は一般的な Contrastive Learning とは異なり，2 つの特徴抽出器を使用するが，これらの特徴抽出器は Ensemble Learning により，MCL だけでなく Classification の安定性と精度を向上させるためにも有効活用される．また第二の課題に対しては，MCTTA が様々なドメインに対して普遍的に使用することができることを 3 節の評価実験にて示す．なお，MCTTA は 1 節で記述した，Domain Generalization に対する三つのアプローチを全て兼ね備えているにもかかわらず，非常にシンプルな手法であることは特筆すべき点である．データレベルのアプローチには MixStyle，モデルレベルのアプローチには 2 つの特徴抽出器を用いた Ensemble Classification，アルゴリズムレベルのアプローチには MCL によるテストドメインへの適応を用いている．

2.1 節では学習データを使用して，Ensemble Classification と MCL によるマルチタスク学習を行う Training フェーズについて説明し，2.2 節ではテストデータを使用した MCL により，特徴抽出器をテストドメインに適応する TTA フェーズについて説明する．

2.1 Training フェーズ

MCTTA の Training フェーズの目的は Ensemble Classification と MCL のマルチタスク学習を通じて，様々なドメインに対して一貫した特徴量を抽出することができるよう特徴抽出器を学習することである．学習データを x とする（前処理は行わない）．x は同一構造で異なる初期値を持つ 2 つの特徴抽出器に入力され，2 種類の出力が得られる．図 2.1 に示すように，一方の出力は単純な特徴抽出から得られる特徴量であり，他方出力は MixStyle を適用し，バッチ間でドメイン情報を混合した特徴量である．学習の各ステップで MixStyle を適用する層は，特徴抽出器の中間の Convolution よりも入力に近い側の層からランダムに選択される．浅い層に限定しているのは，ドメイン情報が CNN の浅い層で主に表現される [4] ためであり，ランダムに選択を行う理由は，ドメイン情報の混合に多様性を持たせるためである．得られた全ての特徴量は，Ensemble Classification と MCL の両方に利用される．Ensemble Classification は，Classification の安定性と精度を向上させ，さらに MixStyle を使用しているため，通常の Classification に比べて Domain Generalization も促進される．MCL では，MixStyle を行っていない特徴量と，MixStyle を行ってバッチ間でドメイン情報が混ざった特徴量の間で Cosine Similarity に基づく Loss が算出され，特徴量間の類似度を最大化することにより，特徴抽出器が様々なドメインに対して一貫した特徴を抽出できるようになり，Domain Generalization が促進される．

まず，Ensemble Classification を定式化する．$\mathcal{P}_{ori}$ は一方の特徴抽出器のうち MixStyle を実行していない特徴量から得たクラスの確率分布，$\mathcal{P}_{mix}$ は一方の特徴抽出器のうち MixStyle を実行した特徴量から得たクラスの確率分布を表す．同様に，他方の特徴抽出器から得たクラスの確率分布を $\mathcal{P}'_{ori}$，$\mathcal{P}'_{mix}$ とする．また，y は正解ラベルを，CE は Cross Entropy Loss を表している．

$$Pre = \frac{\mathcal{P}_{ori} + \mathcal{P}'_{ori}}{2}, Pre' = \frac{\mathcal{P}_{mix} + \mathcal{P}'_{mix}}{2} \quad (1)$$

$$\mathcal{L}_{ce} = CE(Pre, y) + CE(Pre', y) \quad (2)$$

次に MCL の定式化を行う．$\mathcal{Z}_{ori}$ は一方の特徴抽出器のうち MixStyle を実行していない特徴量，$\mathcal{Z}_{mix}$ は，一方の特徴抽出器のうち，MixStyle を実行した特徴量を表している．同様に，他方の特徴抽出器から得た特徴量を $\mathcal{Z}'_{ori}$，$\mathcal{Z}'_{mix}$ とする．また，$\cdot$ はベクトルの内積，$\| \|$ はベクトルのノルムを表している．

$$\mathcal{L}_{cos} = 2 - \left(\frac{\mathcal{Z}_{\mathbf{ori}} \cdot \mathcal{Z}'_{\mathbf{mix}}}{\|\mathcal{Z}_{\mathbf{ori}}\|\|\mathcal{Z}'_{\mathbf{mix}}\|} + \frac{\mathcal{Z}'_{\mathbf{ori}} \cdot \mathcal{Z}_{\mathbf{mix}}}{\|\mathcal{Z}'_{\mathbf{ori}}\|\|\mathcal{Z}_{\mathbf{mix}}\|}\right) \quad (3)$$

そして，Training フェーズの最終的な Loss は，

$$\mathcal{L}_{train} = \mathcal{L}_{ce} + \alpha \mathcal{L}_{cos} \quad (4)$$

である．Training フェーズでは $\mathcal{L}_{train}$ を最小化するよう 2 つの特徴抽出器と 2 つの分類器を学習する．なお，α はハイパーパラメータである．

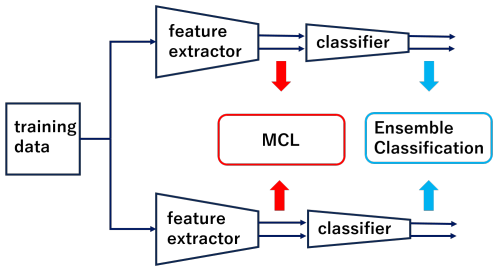

図 1　MCTTA の Training フェーズにおける全体図

図 2　MCTTA における特徴抽出過程

2.2　TTA フェーズ

MCTTA の TTA フェーズの目的は，MCL を用いて特徴抽出器をテストデータに適応させることである．TTT では，Training フェーズで回転予測などの一般的な自己教師ありタスクと Classification のマルチタスク学習を行い，TTA フェーズで自己教師タスクを行うことにより，学習データとテストデータのドメインの差異を特徴抽出器が吸収することができる．MCTTA では，ドメイン汎化に特化した自己教師ありタスクである MCL を用いるため，テストデータを使用してドメインの差異を吸収するだけでなく，モデルのドメインに対する汎化をさらに向上させることが可能である．TTA フェーズは，テストデータを推論する際にオンライン方式で実施され，更新したパラメータはそれ以降のテストでも保持される．テストデータに対して Training フェーズと同じ特徴抽出が行われるが，テストデータにはクラスラベルが存在しないため，Ensemble Classification は行われず，MCL のみでモデルを学習する．なお，推論には Ensemble Classification を用いる．

TTA フェーズの最終的な Loss は，式（3）を用いて，$\mathcal{L}_{tta} = \mathcal{L}_{cos}$ と表される．TTA フェーズでは $\mathcal{L}_{tta}$ を最小化するように 2 つの特徴抽出器のみを学習する．ただし，学習率は Training フェーズの β 倍する．ここで，β はハイパーパラメータである．

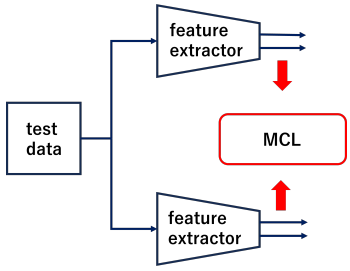

図 3　MCTTA の TTA フェーズの概要

3.　評価実験

評価実験には，客観性を担保するために DomainBed ベンチマークライブラリを使用し，データセットには，PACS，Office-Home，Colored MNIST の 3 つを用いる．PACS には 7 クラスの 9,991 枚の画像，そして Photo，Art Painting，Cartoon，Sketch の 4 つのドメインが含まれており，現在 Domain Generalization の研究分野で最も広く使用されている．Office-Home には，15,588 枚の画像，65 のクラス，そして Art，Clipart，Product，Real World の 4 つのドメインが含まれている．Colored MNIST は，70,000 枚の画像，2 つのクラスを持ち，MNIST[20] に色を付け，3 つのドメインを作成したデータセットである．

学習画像はドメイン毎に 8:2（train:validation）の割合で分割される．この分割は試行ごとに DomainBed により動的に選択される．モデルはテストドメインを除く全てのドメインを用いて 5,000 イテレーション学習され，テストドメインでテストを行う．テストに使用するモデルの選択には，DomainBed において training-domain validation set と呼ばれるモデル選択法が採用される．この方法では，DomainBed で定められたステップ数ごとに全学習ドメインの validation データにおける平均精度を計算し，この平均精度が最も高いモデルが選択される．また全実験において，特徴抽出器には ImageNet[21] で事前学習された ResNet18[22] のバックボーンが使用される．試行は初期値を変えて 10 回行い，試行のシード値，学習率，および拡張スキルは DomainBed に従って動的に設定され，バッチサイズは 32 で統一する．提案手法独自のハイパーパラメータ α，β（2.1, 2.2 節参照）は，全実験において $\alpha = 1$,

$\beta = 0.1$ とする．$\alpha = 1$ とするのは Ensemble Classification と MCL を同じ割合でモデルの学習に活用するためであり，$\beta = 0.1$ とするのは，モデルがテストデータに対して過剰に適応することを防ぐためである．過剰な適応が続くと，様々なドメインに対して一貫した特徴量を抽出することができる特徴抽出器がその能力を失ってしまう可能性がある．なお，TTT の自己教師タスクには Rotation Prediction を用いる．

表 1　3 つのデータセットを用いた実験結果

Algorithm	PA	OH	CM	Avg	RC ↓
ERM	80.6	58.0	51.3	63.3	44
Fish	81.9	59.5	52.0	64.5	27
GroupDRO	80.2	57.7	51.6	63.2	44
Mixup	79.3	61.3	51.5	64.0	40
MLDG	82.1	57.7	52.1	64.0	27
CORAL	82.1	62.9	51.9	65.6	20
MTL	80.0	57.8	51.6	63.1	44
SagNet	82.7	60.6	51.9	65.1	22
ARM	82.1	57.4	53.4	64.3	24
VREx	81.2	58.8	52.2	64.1	29
SD	82.0	63.0	52.0	65.7	20
ANDMask	77.0	57.1	52.2	62.1	40
SelfReg	81.7	63.5	51.7	65.6	26
TRM	81.9	59.5	50.0	63.8	38
TENT	83.0	62.6	51.9	65.8	19
SHOT	83.6	63.5	51.9	66.3	12
TTT	83.4	60.6	52.3	65.4	14
ours	84.6	63.5	53.2	67.1	4

表 1 の 3 つのデータセット名の下に初期値を変えた 10 回の試行によるテストドメインの平均精度（%）を示す．例えば PACS（PA）の場合，Photo をテストドメインとし，その他 3 つのドメインでモデルを学習した場合の精度と，Art Painting をテストドメインとした場合の精度，Cartoon をテストドメインとした場合の精度，Sketch をテストドメインとした場合の精度の 4 つのテストドメインの平均精度を示している．Avg は 3 つのデータセットの平均値を表し，Rank Score（RC）は，3 つのデータセットに対する手法の順位を加算したスコアである．例えば，提案手法（ours）では PA で 1 位，OH で 1 位，CM で 2 位のため RC は 4 である．なお同一順位の手法が存在する場合は，最も少ないスコアが加算される．例えば OH では同率 1 位が 3 手法存在するため，これら 3 手法には 1 スコア，次に精度が良い手法は 4 位として 4 スコア加算される．なお，TENT，SHOT[23]，TTT，ours は TTA 手法であり，その他の手法はテスト時にモデルを適応しない従来の Domain Generalization 手法である．

提案手法は，TTT と比較して PA では 1.2%，OH では 2.9%，CM では 0.9%，Avg では 1.7%の精度向上に成功しており，様々なドメインに対して有効な学習であることが分かる．TTT のようなマルチタスク学習を行わず，教師なし学習のみでモデルをテストドメインに適応する TENT と比較した場合，提案手法は PA では 1.6%，OH では 0.9%，CM では 1.3%，Avg では 1.3%優っている．また，TENT は TTT に対して，Avg では 0.4%優っているが，CM では 0.4%劣っており，1 節で述べたように，全てのドメインに対して TTT より有効な手法であるとまでは言えないこが分かる．また，疑似ラベル付けを用いた TTA である SHOT は，Avg が 66.3%であり，TTT を 0.9%上回る高精度を達成しているが，我々の提案手法は SHOT よりもさらに平均 0.8%優っている．その他，1 節において述べた，Mixup や SagNet，MLDG などの，テスト時にモデルを適応しない従来の Domain Generalization 手法と比較しても，MCTTA は最も優れた平均精度を達成していることが分かる．さらに RC で比較した場合においても，最も優れているのは提案手法であることが分かる．

表 2　MCL の有効性

Color	MixStyle	PA	OH	CM	Avg
✓	-	74.5	59.6	54.9	63.0
-	✓	84.6	63.5	53.2	67.1

表 2 では，MCL の有効性の確認として，ランダムな色変換を画像単位で行う従来の Contrastive Learning（表 2 における Color）と，MixStyle を特徴量単位で行う MCL（表 2 における MixStyle）の精度を比較する．なお，その他のアルゴリズムは全て提案手法と同一である．MixStyle を用いた MCL は，ランダムな色変換と比較して，Avg において 4.1%精度を向上させており，MCL が非常に有効に働いていることが分かる．ただし，MNIST に色を付けて異なるドメインを作成した CM では 1.7%劣った結果となっており，これは色が異なるドメイン間で一貫した特徴を抽出することができるよう学習行う従来の Contrastive Learning（Color）が CM と相性が極端に良いためである．

表 3　2 つのモデルを使用することの有効性

Ensemble Learning	PA	OH	CM	Avg
-	83.4	62.7	52.9	66.3
✓	84.6	63.5	53.2	67.1

表 3 では，同一構造で初期値の異なる 2 つのモデルを使用することの有効性を示すために，1 つのモデルで Classification と MCL を行う場合と，提案手法のように 2 つのモデルを用いて Ensemble Classification とモデル間で MCL を行う場合を比較する．2 つのモデルを使用することにより，PA では 1.2%，OH では 0.8%，CM では 0.3%，Avg では 0.8%精度が向上しており，その有効性が確認できる．

4. おわりに

TTT の問題点を改善するために, Domain Generalization に特化した独自の Contrastive Learning である MCL を用いた新しい TTA である MCTTA を提案した．DomainBed ベンチマークライブラリと 3 つのデータセットを用いた実験を行い，MCTTA は TTT から平均 1.7%の精度向上に成功し，従来の Domain Generalization 法と比較して最も優れた精度を達成した．また，MCL と 2 つのモデルを使用することの有効性も確認した．

参考文献

[1] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[2] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[3] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ICLR 2018*, 2017.

[4] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *ICLR 2021*, 2021.

[5] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 2021.

[6] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[7] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8690–8699, 2021.

[8] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32, 2018.

[9] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *arXiv preprint arXiv:2303.15361*, 2023.

[10] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

[11] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *ICLR 2021 Spotlight*, 2020.

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

[13] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, Vol. 9, No. 1, p. 2, 2020.

[14] Pranjal Kumar, Piyush Rawat, and Siddhartha Chauhan. Contrastive self-supervised learning: review, progress, challenges and future research directions. *International Journal of Multimedia Information Retrieval*, Vol. 11, No. 4, pp. 461–488, 2022.

[15] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[16] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[17] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

[18] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.

[19] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324, 1998.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[23] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pp. 6028–6039. PMLR, 2020.