２０２３年度

博士学位論文

深層距離学習における
解釈可能性とドメイン適応に関する研究

中部大学大学院

工学研究科　ロボット理工学専攻

鵜飼　祐生

論　文　要　旨

距離学習（類似度学習）は画像認識における基本技術の一つである．距離学習はクラス分類手法と異なり，サンプル間の類似度に基づき推論を行う．そのため，距離学習は学習データとテストデータのクラスラベルが一致しない ‘open-set’ な問題設定に適用できる．特に人物間の照合や商品認識などのタスクでは，深層学習技術を応用した深層距離学習技術が，非常に高い精度を達成出来ることが示されてきた．そのため深層距離学習技術はスマートリテールなどへの応用が期待される，重要な基盤技術の一つとなっている．一方で深層距離学習には社会実装上の二つの課題が存在する．

第一の課題は学習データを収集した環境（ソースドメイン）とは異なる環境（ターゲットドメイン）にモデルを適用した際，性能が大幅に低下するドメインシフトに関する課題である．性能劣化に対処するため，導入先の環境ごとにデータセットを作成することは人的および時間的コストの観点から実用的ではない．加えて人物画像間の照合を目的とする人物再同定タスクでは，プライバシーの観点からデータを導入先の環境から持ち出すことが難しい場合がある．そのため，人物再同定では答え入れなしにターゲットドメインでモデルを学習する教師なしドメイン適応手法が多く提案されてきた．しかし，従来の人物再同定における教師なしドメイン適応手法では服の色など画像の大部分を占める大域的な特徴のみが考慮されており，人物間を見分けるために重要な腕時計などの局所的な特徴が十分考慮されていない．そのため，従来手法では導入先環境において答え入れを行った場合と比較し十分な性能を達成できない課題が存在する．

そこで本論文では局所的特徴と大域的特徴の両方を考慮する新規の教師なしドメイン適応手法を提案する．提案手法では各特徴を個別にクラスタリングして得られる疑似ラベルとそれらの積集合を用いて学習を行う．積集合ラベルで同一ラベルを持つ人物画像は両方の特徴で同一の人物と判断されたと言える．そこで積集合ラベルによる教示により，よく似た異なる人物の画像を同一人物として学習することを防ぐことが出来る．これより提案手法は従来手法と比較し高い精度を達成出来る．

第二の課題は何故深層学習モデルが二つの画像を似ていると推論したかを説明できない解釈可能性に関する課題である．深層学習モデルがその推論根拠を説明できることはモデルを信用し，またモデルが誤った根拠に基づき判断を行っていないか確認する上で非常に重要である．そのため，人間の生活に影響を及ぼす ‘ハイリスク’ な分野では深層学習モデルの解釈可能性が必要不可欠となる．この課題に対処するため，クラス分類タスクでは高い精度と解釈可能性を両立する深層学習モデルを構築する手法が多く提案されている．中でも case-based な推論による解釈可能性と深層学習による特徴抽出を組み合わせた Prototypical Part Network (ProtoPNet) による ‘This looks like that’ フレームワークが効果的な手法として幅広い注目を集めている．しかし，ProtoPNet は事前に定義された Prototype とクラスラベルの関係性に依存した学習手法を用いるため，類似度学習に適用することが難しい．

そこで本研究では ProtoPNet を類似度学習に適用可能となるよう拡張した ProtoMetric を提案する．提案手法ではサンプル間で各 Prototype との類似度を比較することによりクラスと Prototype の関係を推定し，その推定に基づき学習を行う．これより，ProtoMetric は Prototype とクラスラベルの関係を事前に定義することなく学習出来る．また ProtoMetric により従来手法を適用することが難しい画像検索タスクにおいて，初めて ‘This looks like that’ フレームワークに基づく解釈可能性が実現される．

目 次

図目次

表 目 次

第1章

序論

本章では，本研究の背景及び目的，本論文の構成について述べる．

1.1 研究の背景

距離学習（類似度学習）は与えられたサンプルペアがどの程度意味的に類似するかを推定する，画像認識分野における基本技術である．距離学習ではクラスを直接回帰する代わりに，推定されたサンプル間の類似度を用いて推論を行う．そのため，距離学習技術は（画像）分類手法と異なり，学習データに存在しない未知のクラスが含まれる状況に適用することが出来る．特に深層学習技術が画像分類タスクにおいて目覚ましい発展を遂げて以降，深層学習技術を取り入れた深層距離学習技術が，画像検索や人物再同定タスクにおいて高い精度を達成することが示されてきた．加えて深層距離学習技術は商品認識など一クラス当たりのサンプル数が少ない一方，膨大なクラス数を含む場合においても有効性が確認されている．そのため，深層距離学習技術は，スマートリテールなどの幅広いアプリケーションで採用される，重要な基盤技術となっている．

一方で深層距離学習技術には二つの課題があることが知られている．第一の課題は学習データを収集した環境（ソースドメイン）とは異なる環境（ターゲットドメイン）に深層学習モデルを適用した際精度が大幅に劣化する，ドメインシフトに関する課題である．深層学習モデルは一般に大量の学習データから推論に有効な特徴を学習し，高い精度を達成する．従って，高い精度を達成する深層学習モデルの実現には学習データの慎重な選定とアノテーションが必要となる．そのため，深層学習モデルの学習データセットの構築には膨大な時間的および人的コストを必要とする．このような背景から，適用先の環境ごとに学習データを収集しデータセットを構築することは非実用的であり，システムを水平展開する上で大きな障害となり得る．そのため，人手によるアノテーションなしに適用先の環境に深層学習モデルを適合させる教師なしドメイン適応技術は，深層学習システムを広く社会実装する上で重要な技術となる．特に人物再同定タスクではドメインの変化に対し精度を大幅に低下させる事に加え，プライバシーの観点からターゲットドメインのデータを持ち出すことが難しい．そのため人物再同定タスクでは答え入れなしにターゲットドメインへ深層学習モデルを適応させる技術が必要不可欠となっており，様々な教師なしドメイン適応手法が提案されている．一方で，これらの手法では服の色などの画像内の大部分を占める大域的な特徴のみが考慮されており，時計や靴の違いといった局所的な特徴が考慮されていない．時計や靴などの違いは人物間を見分けるために重要な特徴であり，よく似た異なる人物画像を識別するためには大域的特徴だけでなくこれらの局所的な特徴を考慮する必要がある．

第二の課題は深層学習モデルの推論過程があまりに複雑であるために，人間に理解できない事実上の「Black-box」となっている解釈可能性に関する課題である．機械学習システムがその推論根拠を出力できることは，システムの出力を人間が信頼し，システムを運用するうえで必要不可欠である．加えて推論根拠を理解することなく機械学習システムに判断を委ねることは，倫理上の重大なリスクに繋がる．例えばアメリカ司法で仮釈放や保釈の決定に広く使われている COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) は予測が人種にバイアスされているとの批判がなされている [17] [1]．このような機械学習システムの抱えるリスクのため，欧州では AI 規制に関

[1] 人種にバイアスされているとの批判は Surrogate Model を構築した結果得られた説明を根拠としており，COMPAS が真に人種にバイアスされていることを確かめるものではない [74]．ここで重要なのは black-box モデルである COMPAS が人種にバイアスされているとの疑惑を招き，システムに

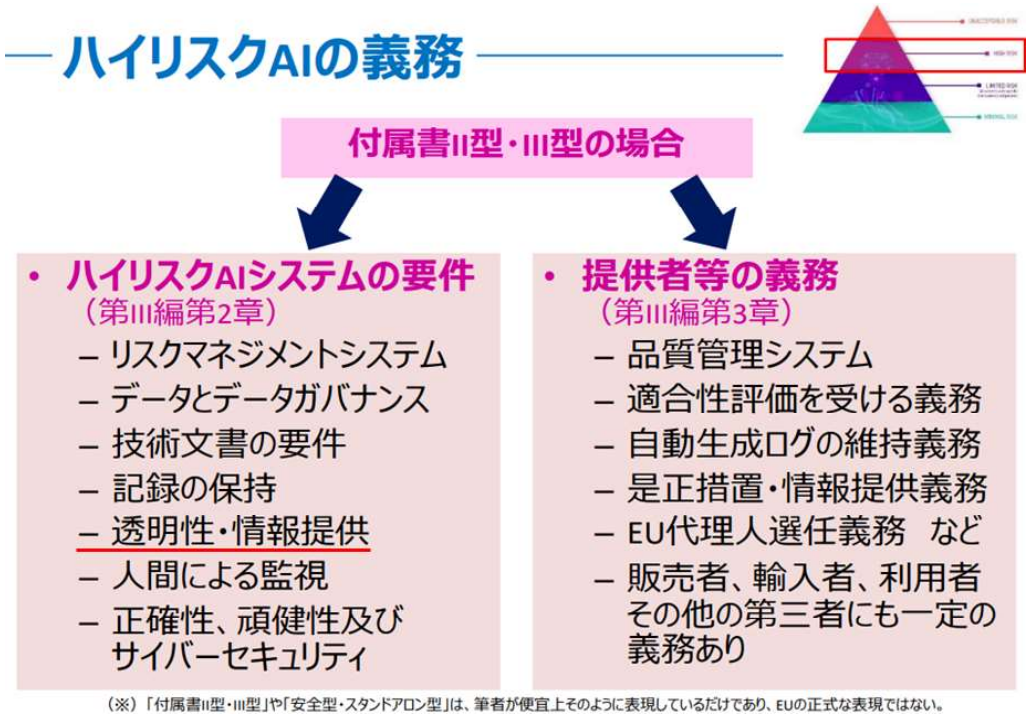

図 1.1: ハイリスク AI に要求されるシステム要件．図は総務省「諸外国における AI 規制の動向に関する調査研究」より抜粋．

する法的な議論が実施されている．特に人間の社会生活に影響を及ぼす「ハイリスク AI」にはシステム出力の「透明性」がシステム要件として課されている[2]．従って機械学習システムの解釈可能性はシステム出力の正当性および妥当性を監査出来ることを求める法的な要請の意味においても必要不可欠となりつつある．このような社会背景から，深層学習における解釈可能性は幅広い注目を集めており，特に画像分類タスクを対象として様々な研究が活発に行われている．深層学習における解釈可能性の研究は学習済みのモデルを解析するアプローチと解釈可能なモデルを構築するアプローチに分割される．類似度学習を対象とする解釈可能性の研究では前者のアプローチが採用されており，後者のアプローチに関する研究はほとんどない．一方学習済みのモデルを解析するアプローチには，モデルの推論過程とは何ら関係のない説明が生成されることが指摘されており [57]，忠実性に課題を抱えている．そのため，解釈可能な類似度学習モデルの構築に関する研究が重要となる．

1.2　研究目的

前節で説明した研究背景から，本研究では以下に示す二つの項目の実現を目的とする．

対する不信が提起された事実である．

[2]2023 年 6 月現在，透明性の要件はどの場面でどのように AI を使用しているかの開示にとどまるが，透明性に対する要求は高まる傾向にある．

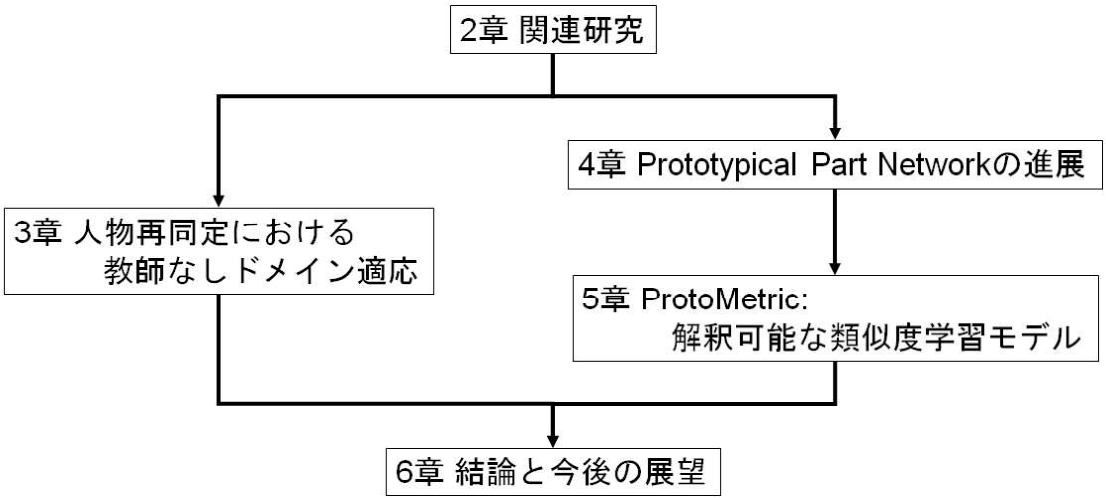

図 1.2: 本論文の構成.

・ 局所的な特徴と大域的な特徴を考慮した教師なしドメイン適応手法.

・ 本質的に解釈可能な類似度学習モデル.

第一の項目では特に人物再同定を対象として，服の色等の大域的な特徴と腕時計などの局所的な特徴を考慮した，高い精度を達成する教師なしドメイン適応手法を提案する．第二の項目では学習データ内に含まれる画像パッチと類似する画像パッチを与えられた画像ペアがどの程度共通して持っているかを基に類似度を算出することで，推論根拠を出力可能な類似度学習モデルを実現する．以下では各項目における本研究の目的を詳細に説明する．

局所的な特徴と大域的な特徴を考慮した教師なしドメイン適応手法　人物再同定タスクでは服の色やテクスチャといった画像内に大きく映る大域的な特徴に加えて，腕時計や靴などのほんの僅かな画像領域にのみ存在する局所的な特徴が同一人物と他人を区別するための重要な特徴となる．そのため，人物再同定では大域的特徴と局所的特徴の両方を考慮した手法が有効と考えられ，教師あり学習の研究においてその有効性が確認されている．一方で，ほとんどの教師なしドメイン適応の研究では大域的な特徴を用いることのみが考慮され，局所的な特徴を用いることの有効性はほとんど考慮されていない．加えて，局所的な特徴を考慮する方法においても局所的な特徴と大域的な特徴とはそれぞれ独立に扱われており，それらを組み合わせて用いることの有効性は考慮されていない．そこで本研究では，大域的な特徴と局所的な特徴を組み合わせることにより，高い精度を達成する人物再同定における教師なしドメイン適応手法を提案する．より具体的には深層学習モデルより出力される大域的な特徴および局所的な特徴それぞれをクラスタリングして得られる疑似ラベル集合に加え，それらの積集合を用いて学習することで精度向上を達成できることを示す.

本質的に解釈可能な類似度学習モデル　画像分類タスクを対象とした深層学習の解釈可能性に関する研究は多くなされている一方で，距離学習（類似度学習）を対象とした解釈可能性の研究は少な

い．また既存の距離学習を対象とする解釈可能性の研究では，学習済みの Black-box モデルに対し，画像中のどの領域が画像間類似度に寄与するかが推論根拠の説明として提示されるのみである．そのため，既存研究では類似度学習モデルがどのような特徴を捉えた結果，二つの画像を似ている，もしくは似ていないと判断したかを解釈することは出来ない．加えて，Black-box モデルを解釈する手法では，推論根拠の説明とモデルの推論過程が独立しているために，モデル推論とは全く関係のない説明が生成される場合があるとの批判がなされている [57]．そこで本研究ではモデルがどのような特徴を捉えた結果二枚の画像を似ていると判断したか説明できる，本質的に解釈可能な深層類似度学習モデルの実現を目指す．より具体的には画像分類タスクにおいて提案された ‘This looks like that’ フレームワークを類似度学習へ適用可能となるよう拡張した ProtoMetric を提案する．ProtoMetric では学習データ内に含まれる画像パッチと類似する画像パッチを，入力された画像ペアがどの程度共通して持っているかを基に類似度を算出する．すなわち入力画像ペアのある画像領域と似た学習データ内の画像パッチが二枚の画像を似ていると推論した根拠として出力される．これよりモデルがどのような画像特徴を捉えた結果，入力画像ペアの類似度を算出したかを解釈することが可能となる．

1.3 本論文の構成

図 1.2 に本論文の構成をまとめる．2 章では本論文に関連する深層距離学習，特に人物再同定における教師なしドメイン適応に関する研究および深層学習における解釈可能性に関する研究について説明する．続いて 3 章では提案する大域的な特徴と局所的な特徴の両方を考慮した人物再同定における教師なしドメイン適応手法について説明する．次に 4 章では，従来の ProtoPNet 派生手法について詳細に説明し，従来手法には類似度学習に適用することが難しい課題がある事を説明する．その後 5 章では画像検索タスクへ適用可能となるよう ProtoPNet を拡張した ProtoMetric の説明を行う．最後に 6 章では本研究の総括を行い，今後の展望について説明する．

第2章

関連研究

本章では深層距離学習および深層学習における解釈可能性に関する従来研究について説明する．

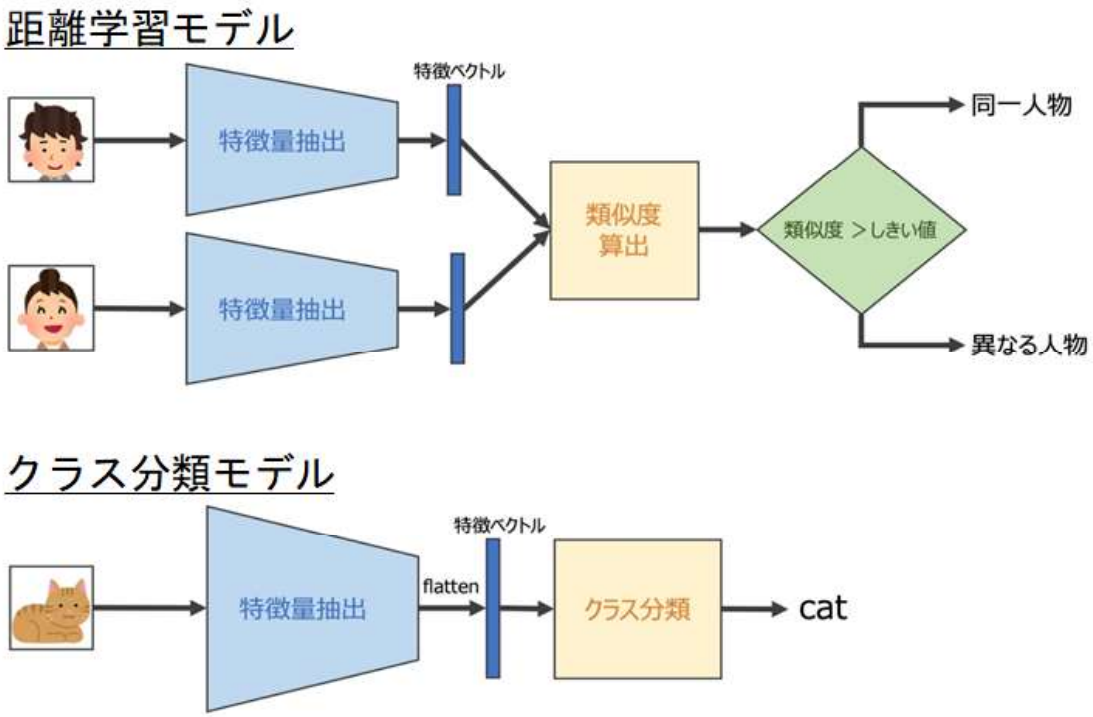

図 2.1: 距離学習モデルによる推論方式とクラス分類モデルによる推論方式の違い．図は https://tech-blog.optim.co.jp/entry/2021/10/01/100000 より抜粋し改変.

2.1 深層距離学習についての関連研究

先述したように，深層距離学習ではクラスラベルを直接回帰する代わりにサンプル間の類似度を基に推論を行う（図 2.1）．この目的のため，深層距離学習では特徴空間上で同一クラスのサンプル間距離が小さく，異なるクラスのサンプル間距離が大きくなるようにモデルを最適化する．深層距離学習手法はサンプル間距離の最適化の仕方により Ranking-base なアプローチと Proxy-base なアプローチに分類される．本章では各アプローチの手法について，それぞれ詳細を説明する．

2.1.1 Ranking-base な手法

Ranking-base なアプローチはミニバッチ内でサンプルペアを構成し，サンプルペア間の距離を最適化するアプローチである．ミニバッチ内でサンプルの 3 つ組 [29] あるいは 4 つ組 [27] を構成し，各組内でのサンプルペア間の相対的な距離を最適化する手法に加え，クラス内およびクラス間サンプルの距離が一定のマージン以上離れるように最適化を行う手法が提案されている [33]．また，どのようにサンプル間距離を最適化するかだけではなく，どのようにミニバッチを構成するかが精度に大きな影響を与えることが実験的に示されており，バッチ内サンプルの構成方法についても検証が行われている [34]．Ranking-base なアプローチは深層距離学習における伝統的なアプローチとして，

深層距離学習の発展に大きく寄与してきた．しかし，サンプルの組み合わせ数はサンプルサイズに対し指数的に増大するため，Ranking-base な手法の学習プロセスは複雑になりやすく sub-optimal な学習となりやすい課題がある．

2.1.2 proxy-base な手法

Ranking-base な手法の学習プロセスが複雑になりやすい課題に対応するため，Proxy-base な手法が提案されている [31][56][67][126]．Proxy-base な手法ではサンプル間の距離の代わりに，クラス重心となる仮想的な重み（Proxy）とバッチ内サンプルを深層学習モデルに入力した結果得られる特徴ベクトルとの距離を最適化する．これより，Proxy-base な手法ではバッチ内でサンプルペアを構成する必要がなくなるため，学習が複雑化する課題を回避することが出来る．またバッチ内にサンプリングされたクラスの種別に関わらず，入力サンプルと全てのクラスとの距離を最適化する事が出来る．そのため，Proxy-base な手法はクラス数が大きく細かな特徴を捉える必要のある問題設定においても高い性能を達成し得る．この利点のため，Proxy-base の手法は顔認証技術にも応用されている [43][46].

一般に Proxy-base の手法は Ranking-base の手法と比較し，学習が早く高い精度を達成できる．一方で Ranking-base の手法ではクラスラベルを陽に必要としないため，クラスラベルが利用できない状況においても学習できる利点がある．Ranking-base の手法と Proxy-base の手法では学習によって獲得される特徴表現が異なる可能性が示唆されている [34]．加えて，深層距離学習は損失関数以外にもデータ拡張手法やモデル構造に大きく性能を影響される [72][73]．そのため，いずれの手法が有効であるかはタスクや得られるデータの種別，また目的に応じた適切な評価指標 [72] のもとで慎重に検証する必要がある．

2.2 人物再同定における教師なしドメイン適応

人物再同定は与えられた人物画像（Query）と同一の人物画像を所与の人物画像群（Gallery）から検索する画像検索タスクである．深層距離学習の発展により人物再同定手法は人手で作成した特徴量を用いる手法と比較して大幅な精度向上を達成してきた．一方で人物再同定手法は学習データを収集した環境（ソースドメイン）とは異なる環境（ターゲットドメイン）に適用した場合に大きく精度が劣化する課題がある．そこで，コストの高い人手によるアノテーションなしにターゲットドメインで高い精度を達成するため，多くの教師なしドメイン適応手法が提案されてきた．教師なしドメイン適応手法は教師ラベルのアノテーションされたソースドメインのデータセット $D_s = \{x_i^s, y_i^s\}_{i=0,1,...}$ 及び教師ラベルのアノテーションされていないターゲットドメインのデータセット $D_t = \{x_i^t\}_{i=0,1,...}$ が与えられた時ターゲットドメインにおいて高い精度を達成するモデルを学習するタスクと定式化される．教師なしドメイン適応の手法はソースドメインの画像をターゲットドメインの画像にスタイル変換した画像を学習に用いる「生成画像を用いるアプローチ」とターゲットドメインの画像を

表 2.1: 人物再同定における実画像データセットの統計．Market1501 では誤検出画像が gallery データセット内に含まれるため，Market1501 では各サブセットの人物画像数の総計と全体の人物画像数は一致しない．

データセット名	サブセット	カメラ数	人物 ID 数	人物画像数
Market1501 [24]	全体	6	1,501	32,668
	training	6	750	12,936
	query	6	751	3,368
	gallery	6	751	19,732
DukeMTMC [20]	全体	8	1,812	36,411
	train	8	702	16,522
	query	8	702	2,228
	gallery	8	1110	17,661
MSMT17 [44]	全体	15	4,101	126,901
	train	15	1,041	32,621
	query	15	3,060	11,659
	gallery	15	3,060	82,161

クラスタリングすることで得られる疑似ラベルを教師信号として学習を行う「疑似ラベルを用いるアプローチ」に大別される．以下では人物再同定におけるデータセットおよび評価指標について説明した後，各アプローチの関連研究について説明する．

2.2.1 人物再同定におけるデータセット

人物再同定では一般的に Market1501[24]，DukeMTMC[20]，および MSMT17[44] の三つのデータセットが手法の評価に用いられる．人物再同定の実験では各データセットをトレーニングデータセット，クエリデータセット，ギャラリデータセットの 3 つのサブセットに分割する．ここで，トレーニングデータセットはモデルを学習するために用いられるデータセットであり，クエリデータセット及びギャラリデータセットはモデル性能評価のために用いられるテストデータセットである．人物再同定の評価ではクエリデータセットに含まれる人物と同一人物をギャラリデータセットから検索し，後述する評価指標により検索結果を評価することでモデル性能を評価する．本論文における実験ではソースドメインにおけるトレーニングデータセットで学習したモデルをターゲットドメインにおけるトレーニングデータセットを用いてドメイン適応を行う．その後ドメイン適応の結果得られたモデルに対し，ターゲットドメインにおけるクエリ，ギャラリデータセットを用いることでターゲットドメインにおけるモデル性能を評価する．なお本論文におけるこれらサブセットへの分割は各データセットの作成者による分割に従った．以下では Market1501，DukeMTMC，および MSMT17 データセットについて説明する．

■ **Market1501**

Market-1501[24] は，6 台のカメラで撮影された 1,501 人の 32,668 枚の画像を，トレーニングデータセット，クエリデータセット，ギャラリデータセットの 3 つのセットに分割したものである．各セットには，751 人分の画像が 12,936 枚，750 人分の画像が 3,368 枚，750 人分の画像が 19,372 枚含まれている．

■ **DukeMTMC**

DukeMTMC[20] は，8 台のカメラで撮影された 1,812 人の 36,411 枚の画像が含まれるデータセットである．DukeMTMC ではトレーニングセットには 702 人の 16,522 枚の画像，クエリデータセットにはトレーニングデータセットとは異なる 702 人の 2,228 枚の画像，またギャラリデータセットにはクエリデータセット内の 702 人を含む 1,110 人の 17,661 枚の画像が含まれるように分割される．

DukeMTMC は Market1501 と同規模のデータセットであり人物再同定における教師なしドメイン適応の研究では頻繁に用いられてきた．しかし，DukeMTMC はプライバシー保護の問題から Duke 大学により公開が停止されており，これ以上用いるべきではないとの提言がなされている [65]．よって本論文では DukeMTMC を用いた実験は掲載しない．

■ **MSMT17**

MSMT17[44] は屋内外に設置された 15 台のカメラで撮影した 4,101 人の人物画像を含むデータセットである．MSMT17 は 32,621 枚の画像を持つ 1,041 人のトレーニングデータセット，11,659 枚の画像を持つ 3,060 人のクエリデータセット，82,621 枚の画像の中に 3,060 人のギャラリデータセットに分割される．MSMT17 は，現在最大規模のデータセットであるだけでなく，照明のばらつきや，画像解像度の違いなど画像の多様性にも富んだ，最も難易度の高いデータセットでもある．

MSMT17 データセットはデータセット使用に関する同意書に署名し，データセット著者に送付することで取得できる．MSMT17 データセット使用に関する同意書内には「データセット内の画像を論文掲載に使用しない」ことが明記されている．そのため，本論文では MSMT17 データセットの画像を掲載しない．MSMT17 データセットの画像例についてはデータセット著者らの文献 [44] を参照されたい．

2.2.2 人物再同定における評価指標

人物再同定では一般に Cumulative Match Characteristic (CMC) curve[5] および mean average precision (mAP)[24] の二つの評価指標が採用される．ここで CMC-curve は各クエリ画像に対し最も類似する人物画像を k 枚ギャラリデータセットより取得した内に同一人物画像が含まれるクエリ画像数のクエリデータセットに対する割合であり，同一人物画像を取得するまでにどの程度の枚数を確認しなけ

ればならないかの指標となる．したがって CMC-curve は各人物再同定手法が各データセットにおいてどの程度機能するかを示す直感的に分かりやすい尺度となる．しかし，CMC-curve は埋め込み空間の違いを効果的に捉えることができないため，妥当性に疑問が示されている [72]．そのため本論文では mAP を主な評価指標とし，CMC-curve は各手法の有効性を計る直感的な尺度として Rank-1 Accuracy (R1) および Rank-5 Accuracy (R5) を報告するにとどめる．ここで R1 は下式のように定義される．

$$\mathrm{R1} = \frac{1}{|\mathbb{Q}|}\sum_{q\in\mathbb{Q}} \mathbf{1}\left(y_q = y_{g_*} | g_* = \arg\max_{g\in\mathbb{G}} \frac{f(x_q)\cdot f(x_g)}{\|f(x_q)\|\|f(x_g)\|}\right) \tag{2.1}$$

ただし $\mathbb{Q}$ および $\mathbb{G}$ はクエリ，ギャラリデータセットに含まれるサンプルインデックスの集合であり，$\mathbf{1}(\cdot)$ は括弧内の条件が満たされる場合に 1，その他の場合に 0 を取る特性関数である．

また，mAP は各 Query 画像に対し算出される Precision-Recall 曲線の Area Under Curve (AUC) として定義される．すなわち mAP は

$$\mathrm{mAP} = \frac{1}{|\mathbb{Q}|}\sum_{q\in\mathbb{Q}}\sum_{n} (R_n^q - R_{n+1}^q) P_n^q \tag{2.2}$$

と定式化される．ただし，R_n^q および P_n^q はそれぞれ n 番目の閾値に対する Recall 値および Precision 値である．そこで mAP は Gallery 画像内に含まれる全ての Query 画像と同一の人物画像を検索した際，検索結果がどの程度 Query 画像と同一人物の画像によって占有されるかを示す指標と言える．

上では人物再同定手法の評価に用いるデータセット及び評価方法について説明した．以下では人物再同定における教師なしドメイン適応手法の二つのアプローチ，生成画像を用いるアプローチと疑似ラベルを用いるアプローチについてそれぞれ詳細に説明する．

2.2.3 生成画像を用いるアプローチ

生成画像を用いるアプローチでは，ソースドメインにおける画像スタイルをターゲットドメインにおける画像スタイルへ変換する．その後，変換された画像を用いてモデルを再学習することで，モデルをファインチューニングする．すなわち生成画像を用いるアプローチではソースドメインのデータをターゲットドメインへスタイル変換した結果得られるデータセット $D_{s\to t} = \{x_i^{s\to t}, y_i^s\}_{i=0,1,\ldots}$ を用いて教師あり学習を行う．深層学習による画像スタイル変換手法の進歩により，画像変換の質が向上したこともあり，生成画像を用いるアプローチに基づく手法が多く提案されてきた [39][44][51]．しかし，画像変換は一般的に処理負荷が大きく，またその品質においてもモデルを学習するために十分とは言えない．このため，クラスタリングベースのアプローチと比べ，生成画像を用いるアプローチでは一般的に精度が低いことが知られている．

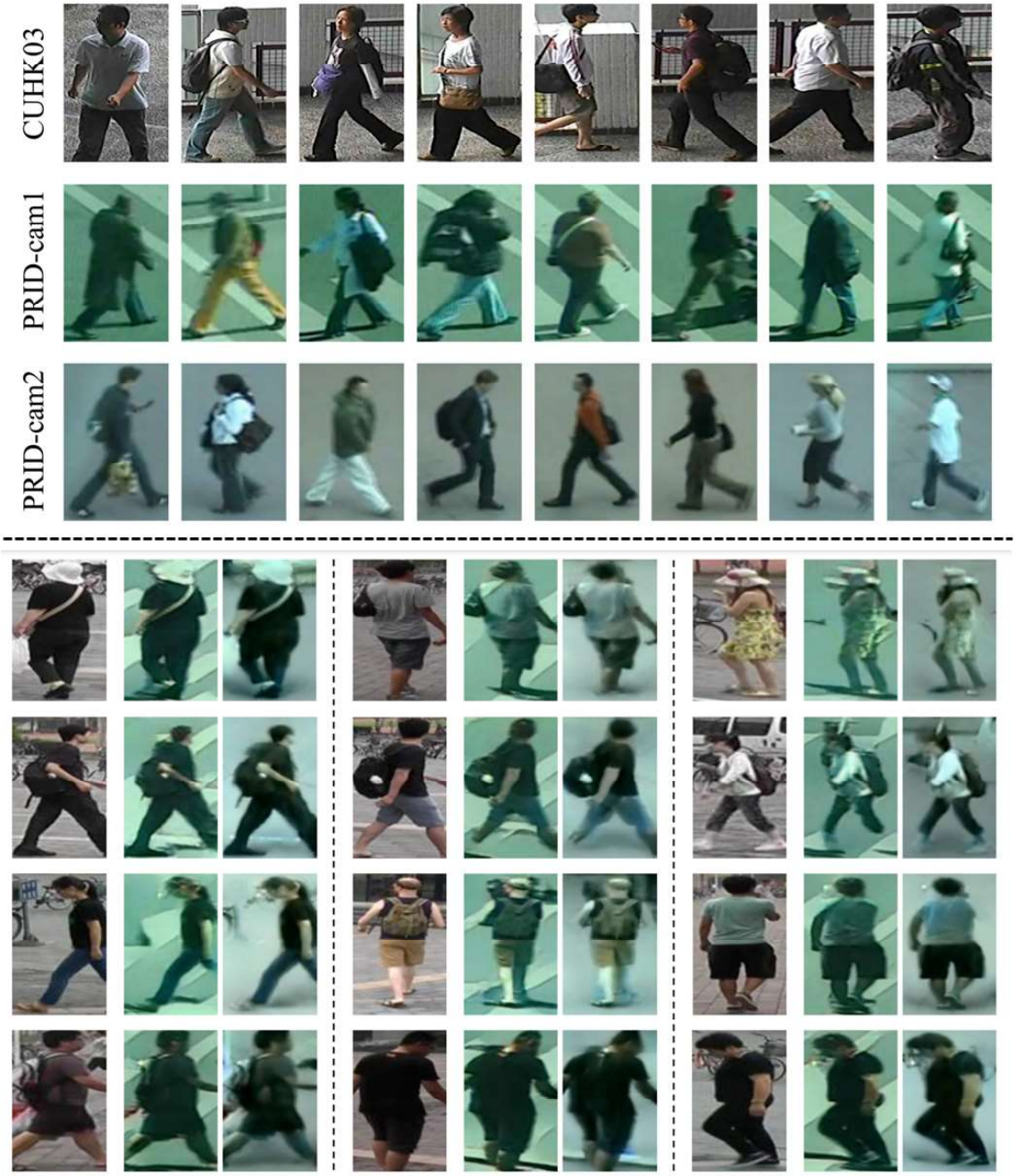

図 2.2: 上：ソースドメイン（CUHK03）とターゲットドメイン（PRID）の画像例．下：スタイル変換により生成された画像の例．各列において，左端が変換元画像であり右側二枚が変換後画像である．画像は文献 [44] より抜粋．

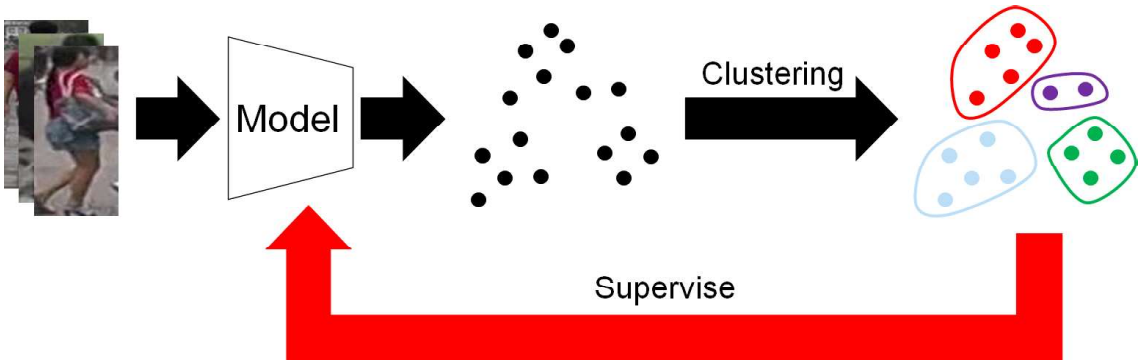

図 2.3: 疑似ラベルを用いるアプローチにおける教師なしドメイン適応の学習フレームワーク．Model はソースドメインで学習されたものを用いる．

2.2.4 疑似ラベルを用いるアプローチ

クラスタリングに基づくアプローチでは，k-means や DBSCAN などのクラスタリング手法を用いてターゲットドメインのデータに疑似ラベル $\hat{y}^t$ を付与した後，擬似ラベルを用いてモデルをファインチューニングする．すなわち疑似ラベルを用いるアプローチでは初めにソースドメインでモデルを学習した後，学習済みモデルを用いて疑似ラベル付きのターゲットドメインデータセット $\hat{D}_t = \{x_i^t, \hat{y}_i^t\}$ を生成した後 $\hat{D}_t$ により教師あり学習を行う．疑似ラベルを用いるアプローチは高い精度を達成することが出来るため，多くの教師なしドメイン適応手法に採用されてきた [65][75][113]．このアプローチにおいて解決すべき問題の一つに，擬似ラベルの間違いによる誤差拡大をいかに低減するかということがある．疑似ラベルの間違いそのものを低減する観点からは，UDAP[75] において詳しい解析がなされており，k-reciprocal encodings[36] を用いた DBSCAN アルゴリズム [3] により疑似ラベルを生成することで高い精度を達成している．他の観点としては，注意機構を利用した手法 [66] や，相互学習を利用した手法 [64][78][113]，疑似ラベルの不確実性を利用する手法 [63][65][79] 等が提案されている．また，ソースドメインで学習した後ターゲットドメインで学習する 2-stage の学習の代わりに，ターゲットドメインとソースドメイン両方のデータを同時に用いて学習する 1-stage の手法が提案されている [65][87]．1-stage の手法はターゲットドメインのみならずソースドメインにおいても高い精度を達成することが確認されており，非常に有効な教師なしドメイン適応手法である．しかし，1-stage の手法では両ドメインのデータを同時に学習する必要性からターゲットドメインへの適用時に多くの計算コストを必要とする．加えて，人物再同定のデータはプライバシーの観点から非常に慎重に扱わなければならない場面が多く，実用上ソースドメインからデータを持ち出すことは難しい．そのため学習時に両ドメインのデータを必要とする 1-stage の手法にはアプリケーションとして実用化する上での課題が存在していると言える．

疑似ラベルを用いる手法では一般的に深層学習モデルより出力される特徴マップに Global Average Pooling (GAP) を適用し得られる特徴ベクトルをクラスタリングすることにより疑似ラベルを取得する．GAP は特徴マップ全体を平均化するため，特徴マップ内における平均的な特性（大域的な特徴）を出力する傾向にある．そのため，GAP は画像内の一部の領域にのみ存在する局所的な特徴を弱め

てしまう．一方で，人物再同定では帽子や靴などの局所的な特徴が人物間を見分ける重要な特徴となる．そのため，人物再同定における教師あり学習・教師なしドメイン適応両分野において局所的な特徴を利用する研究がなされており，その有効性が確認されている．以下では大域的な特徴および局所的特徴の両方を教師なしドメイン適応に利用した手法である SSG[48] および ABMT[113] について説明する．

■ Self Similarity Grouping (SSG)

SSG[48] の学習フレームワークを図 2.4 に示す．SSG では空間方向に特徴マップを分割し，全特徴マップ及び分割された特徴マップそれぞれに対し個別に GAP を適用する．これより，SSG では一枚の画像に対し画像全体を表す大域的な特徴と複数の画像内のある領域に着目する局所的な特徴を表現する特徴ベクトルが抽出される．その後，これらの特徴ベクトルをそれぞれ個別にクラスタリングすることで得られる疑似ラベルを用いて，各特徴ベクトルを教示する．SSG は人物再同定における教師なしドメイン適応において大域的特徴と局所的特徴の両方を用いることが有効であると示した初の論文である．一方で，SSG では特徴マップを空間方向に決められた領域で分割しているため，画像内における人物姿勢変動の効果を受けやすい．そのため SSG はより局所的な特徴を捉えるため特徴マップの分割数を大きくした場合に性能が劣化することが実験により確認されている [48].

■ Asymmetric Branch Mutual Teaching (ABMT)

ABMT[113] の学習フレームワークを図 2.5 に示す．ABMT では Global Max Pooling (GMP) を特徴マップに適用することで局所的な特徴を得る．GMP は GAP とは異なり，各チャンネルごとに特徴マップ内の最大値を出力する．そのため，GMP は画像内のあるピクセルを捉える局所的な特徴を抽出する傾向にあると言える．ABMT ではこれら二つの GAP 及び GMP から出力される特徴を結合した特徴ベクトルを用いてクラスタリングを行う．その後，クラスタリングにより生成された疑似ラベルを用いて各ブランチをそれぞれ個別に教示する．また学習時，学習により更新される Student モデルとは別に指数移動平均により更新されるモデル（Mean Teacher）を用意し，Mean Teacher の各ブランチより出力されるロジットを用いて Student モデルのもう一方のブランチを教示する．ABMT は異なるブランチの出力を用いることにより，特徴が均一化するために学習とともに効果が失われていく Mutual Training の課題を解決した．一方で ABMT は学習中に二つのモデルを用いるため計算コストが高い課題がある．また，各ブランチより算出された特徴ベクトルを結合することにより一つの疑似ラベルを生成しており，各特徴をどのように組み合わせるかは暗黙的なモデルの学習に依存している．そのため，各特徴を明示的に組み合わせることにより，よく似た異なる人物画像を同一人物として学習することを防ぐことの有効性は検討されていない．

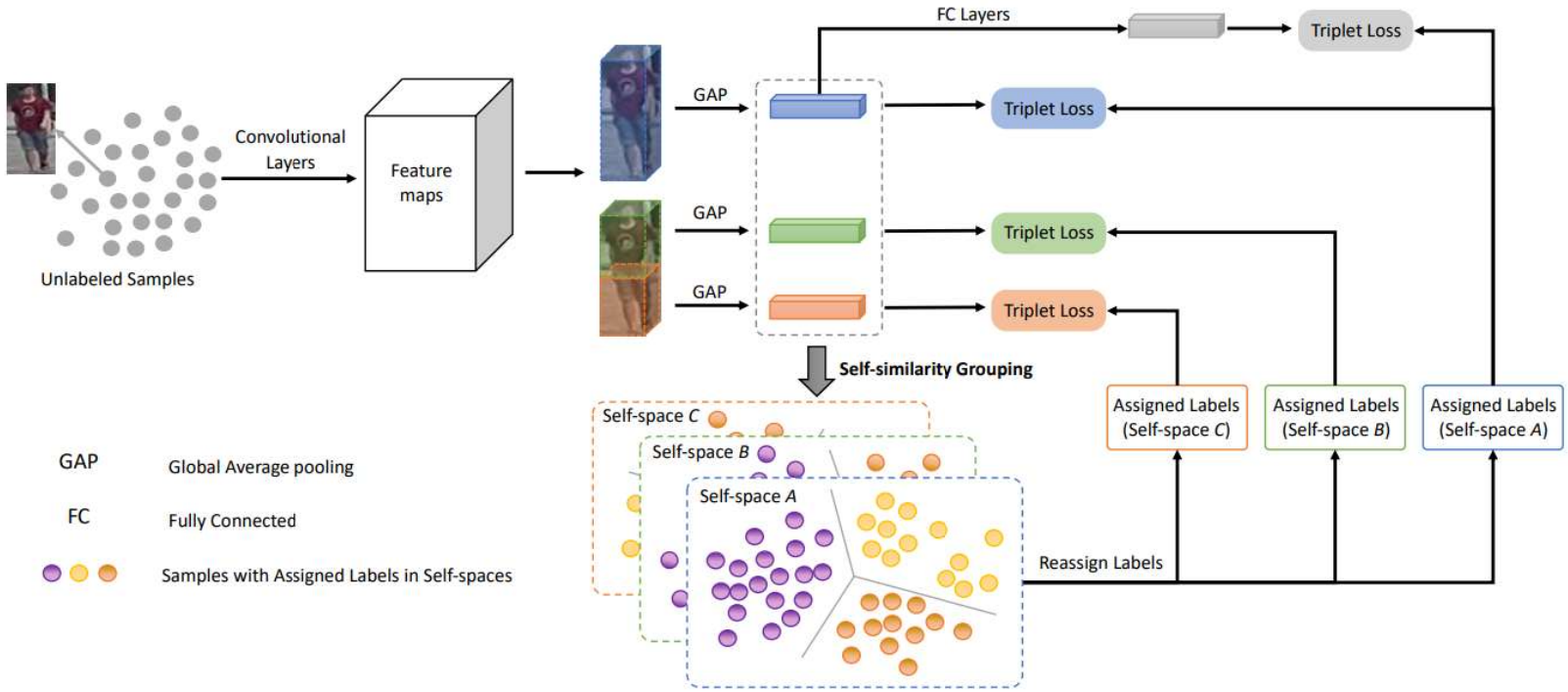

図 2.4: SSG の学習フレームワーク．図は文献 [48] より引用．

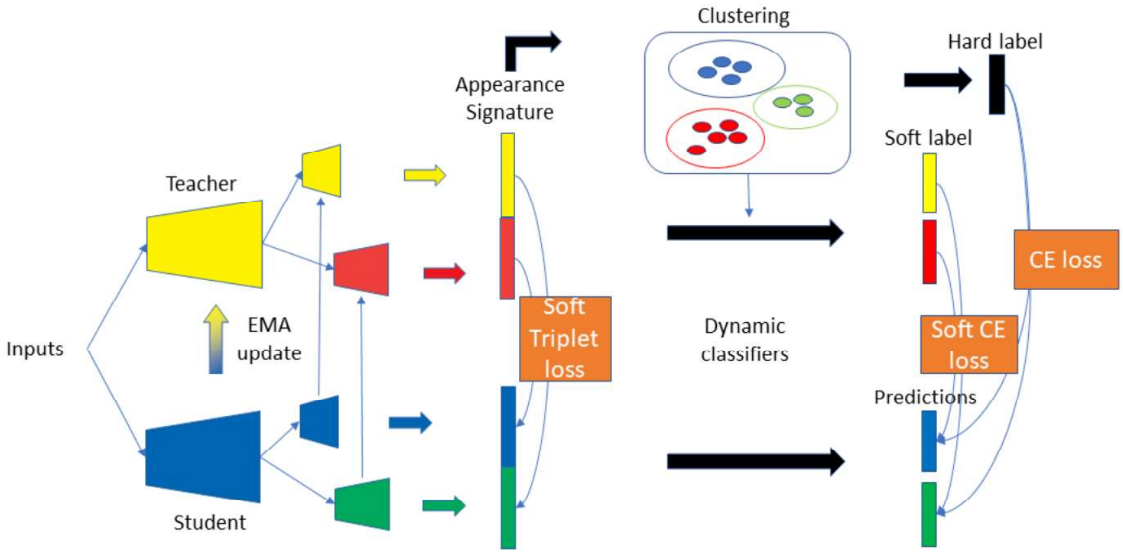

図 2.5: ABMT の学習フレームワーク．図は文献 [113] より引用．

以上のように従来の教師なしドメイン適応手法では一つの特徴ベクトルのみを用いて，もしくは大域的な特徴及び局所的な特徴それぞれより得られた疑似ラベルを個別に用いてモデルを学習する．各特徴より得られた疑似ラベルにはそれぞれの特徴抽出の観点から各画像に写る人物が同一人物か否かを判断した情報が埋め込まれていると考えられる．そのため，従来の教師なしドメイン適応手法では一つの特徴抽出の観点のみがモデル教示として用いられていると言える．一方で，各人物画像が同一人物か否かを判断するには大域的・局所的などの一つの観点のみではなく，複数の観点を組み合わせることが必要となる．本論文では，大域的な特徴及び局所的な特徴それぞれより得られた疑似ラベルを組み合わせた結果を明示的にモデルに教示を与えることで，従来手法より高い教師なしドメイン適応精度を達成できることを示す．

2.2.5　ソースドメインにおける学習の検討

一般にドメイン適応の設定ではソースドメインにおいて，いかに良い特徴表現を獲得したかによりターゲットドメインでの精度が大きく変化する．そのため，実画像大規模データセットにおける教師なし学習 [85][117] や，CG データを活用した大規模データセットによる事前学習手法 [77][105][114][125] が研究されてきた．特に CG データの活用は，プライバシー保護の観点から大規模な人物画像データの収集・保管が難しい課題を解決出来るため非常に有用である．本論文では CG データを活用した事前学習により，提案手法においてもターゲットドメインでの性能を向上させることが出来ることを示す．

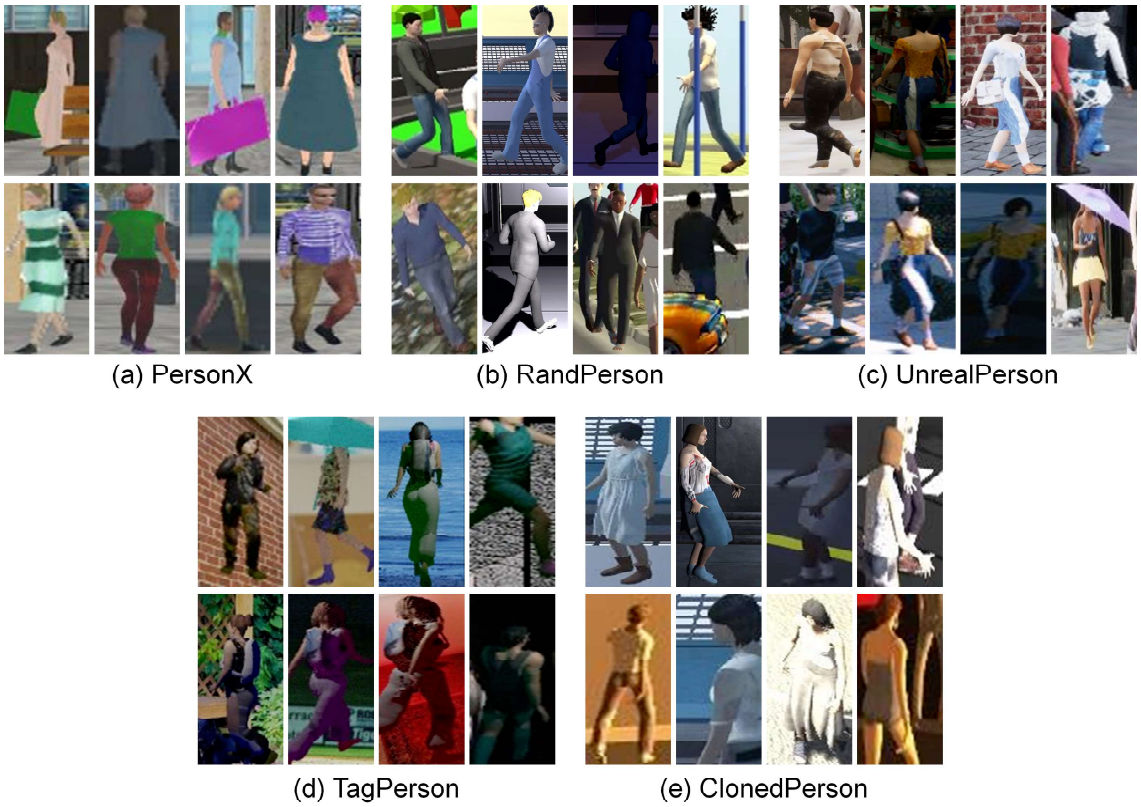

図 2.6: CG 画像を用いた人物再同定データセットの画像例．画像は各データセットより抜粋したものを使用．

表 2.2: 人物再同定における CG 学習データセットの統計．TagPerson は COCO データセットの画像に人物画像を埋め込みデータセットを生成するためカメラ数を ∞ と表現した．

データセット名	カメラ数	人物 ID 数	画像枚数
PersonX[58]	4	410	9,840
RandPerson[77]	19	8,000	1,801,816
UnrealPerson[105]	34	6,799	1,256,381
TagPerson[114]	∞	2,954	71,580
ClonedPerson[125]	24	5,621	887,766

2.3 深層学習モデルにおける解釈可能性

深層学習の解釈可能性に関する研究は学習済みモデルを説明する Post-hoc なアプローチと，解釈可能なモデルを学習する Ante-hoc なアプローチに分類される．以下では各アプローチの関連研究を詳細に説明する．

2.3.1 Post-hoc なアプローチ

Post-hoc なアプローチは，学習済みモデルを解析することにより推論根拠を解釈しようとするアプローチである．Post-hoc な手法では伝統的にモデルが入力画像内のどの領域を重視したかを出力することによりモデルの推論根拠を説明する Saliency-base なアプローチが採用されてきた [19][23][32][42]．Saliency-based な手法にはヒートマップという直感的に理解しやすい形式で説明を行うことが出来る利点がある．Saliency-based な手法は画像認識分野以外にも画像検索 [109] や人物再同定 [60][62] など多くの分野で深層学習モデルの推論根拠を説明する手法として発展を遂げてきた．しかし，出力されたヒートマップは深層学習モデルの推論における各画像領域の重要度を正確に反映しない場合があり，忠実性に課題を抱えていると知られている [38]．この忠実性の課題に対しては Shapley 値 [1] を活用した手法が提案されており [30]，画像分類のみならず深層距離学習 [118] や点群 [99] 等様々なタスクにおける有効性が示されている．Shapley 値は協力ゲームにおいて，プレイヤー間の相互作用を含めた各プレイヤーの貢献度を評価するため提案された指標であり，説明の忠実性を理論的に保証することが出来る．しかし，Shapley 値の計算は組み合わせ数に応じ指数的に増加するため，生画像などの入力次元数が大きなデータに適用することは難しい．加えて Saliency-based な手法ではヒートマップ以上の情報を提示することが出来ないため，モデルの推論過程をより深く理解することが難しい課題がある．

ヒートマップ以上の情報を提示することが出来ない課題に対して，モデルがどのような特徴を捉えたかを解析することによりモデルの推論過程を解釈しようとする Concept-base な手法が提案されている．Concept-based な手法では，ニューロン発火のパターンと画像特徴（コンセプト）に対応関係が存在することを仮定し，ニューロン発火パターンをコンセプトにより記述することでモデル推

論の説明を行う．ニューロン発火パターンとコンセプトの対応関係を求めるため，ニューロン発火パターンを再現するように入力画像を最適化する Feature Visualization[15] と呼ばれる手法が提案されている．Feature Visualization は解析対象となるモデルのみを用いて任意の層における任意のニューロン発火パターンの意味を解析できる利点がある．しかし，Feature Visualization は初期値に大きく依存するため性質の良い結果を安定して得ることが難しい課題がある．また Dravid らは複数のモデルにおいて共通のコンセプトを示すニューロンが存在することを実験的に示し，生成モデルを介してニューロンの意味を可視化した [132]．しかし，共通のコンセプトを示すニューロンは全体の一部であり全てのニューロンの意味を理解することは難しい．一方で Kim らはある画像特徴をもつ入力画像群とそうでない画像群を解析対象モデルに入力した結果得られる，ニューロン発火パターンの違いをもとにニューロン発火パターンの意味を求める Testing Concept Activation Vector (TCAV) を提案した [41]．また Bau らはコンセプトに対するセグメンテーションマスクとニューロン発火の Intersection of Union (IoU) を計算する事で各ニューロンがどのような画像に発火するかを推定する手法を提案した [26]．しかし，これらの手法にはコンセプトに対してアノテーションされたデータセットを必要とする課題が存在する．この課題に対しては，Super pixel に対するニューロン発火パターンをクラスタリングする [50] ことによりアノテーションされたデータセットなしにコンセプトを抽出する手法が提案されている．また大規模言語モデルや CLIP[96] などの大規模視覚言語モデルを利用する [131][138][140][145] ことにより，ニューロン発火パターンの意味を自然言語で記述する手法が提案されている．

Concept-base の手法では Saliency-base の手法と異なり，深層学習モデルの中間層における特徴表現を解析することが出来る．そのため，Concept-base な手法を通して深層学習モデル内部の特徴表現や推論過程に対する様々な知見が提供されてきた．加えて Santurkar らは Concept-base の手法を用いてニューロン発火パターンを解釈したモデルに対し，意図したように深層学習モデルの挙動を変更できる（深層学習モデルのデバッグが可能となる）ことを示した [98]．しかし，一般に深層学習モデルは様々な特徴を複雑に組み合わせることで高い精度を達成しているため，Concept-base の手法で同定されるコンセプト数は膨大となる．そのため，Concept-base の手法で生成される説明は煩雑になりやすく，コンセプトの修正も容易なものではない．加えてニューロンの意味を求めるために用いられるデータセット次第で推定されるニューロンの意味は異なったものになり [146]，またデータセットに付与された僅かな摂動により同一のニューロンに対し全く異なるコンセプトが推定される [150] など忠実性に課題があることも知られている．

さらに異なるアプローチとして，解釈可能なモデル（Surrogate Model）を用いて Black-box モデルの出力をある入力の周りで近似することにより Black-box の挙動を理解しようとする手法 [86] が提案されている．Surrogate Model の構築は解析対象とするモデルに依存しないため，任意のモデルに適用することが出来る利点がある．また類似度学習タスクを対象としても，画像の属性情報を利用し，属性情報に付属する Saliency Map を出力することにより画像間類似度の推論根拠を説明しようとする手法が提案されている [83][121]．しかし，Surrogate Model はあくまで Black-box モデルの出力を近似するのみであり，解析対象とするモデルと推論根拠を同一にするとは限らない [57]．従って，Surrogate Model を構築する手法においても忠実性の課題が存在すると言える．

Post-hoc なアプローチは深層学習モデルの推論過程について理解を深める上で非常に大きな役割を果たしてきた．また一般に Post-hoc なアプローチでは既存の学習モデルを再学習する必要がないため，モデルに変更を加えることなく製品に採用された深層学習モデルの解釈可能性を向上させることが出来る利点がある．しかし，Post-hoc な手法で生成される説明と深層学習モデルの推論過程はそれぞれ独立しているために，Post-hoc なアプローチで生成される説明はモデル推論となんら関係がない場合があり，忠実性に課題を抱えていることが知られている [57].

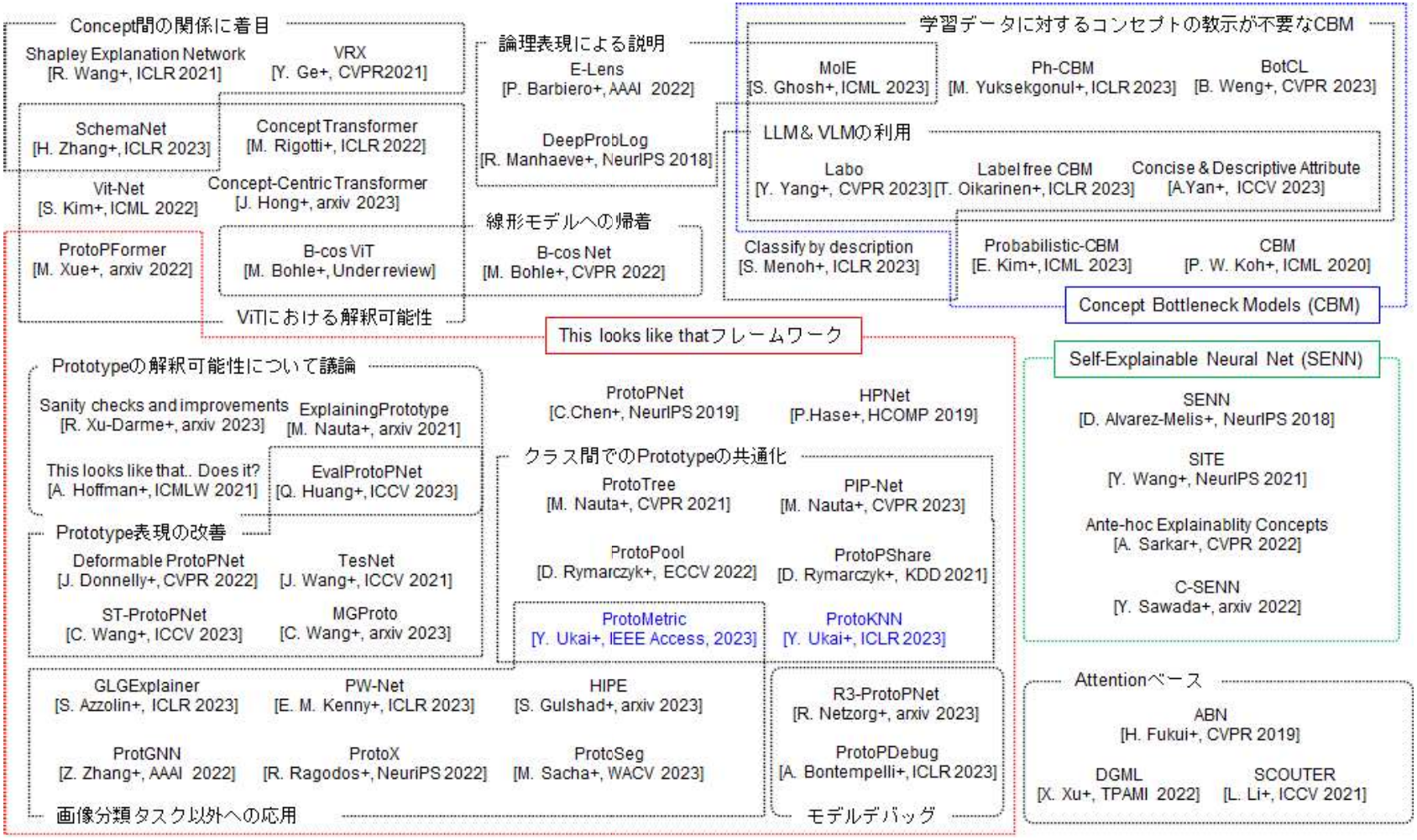

図 2.7: Gray-box モデルを構築する研究に関するカオスマップ．筆者の研究を青字で示す．

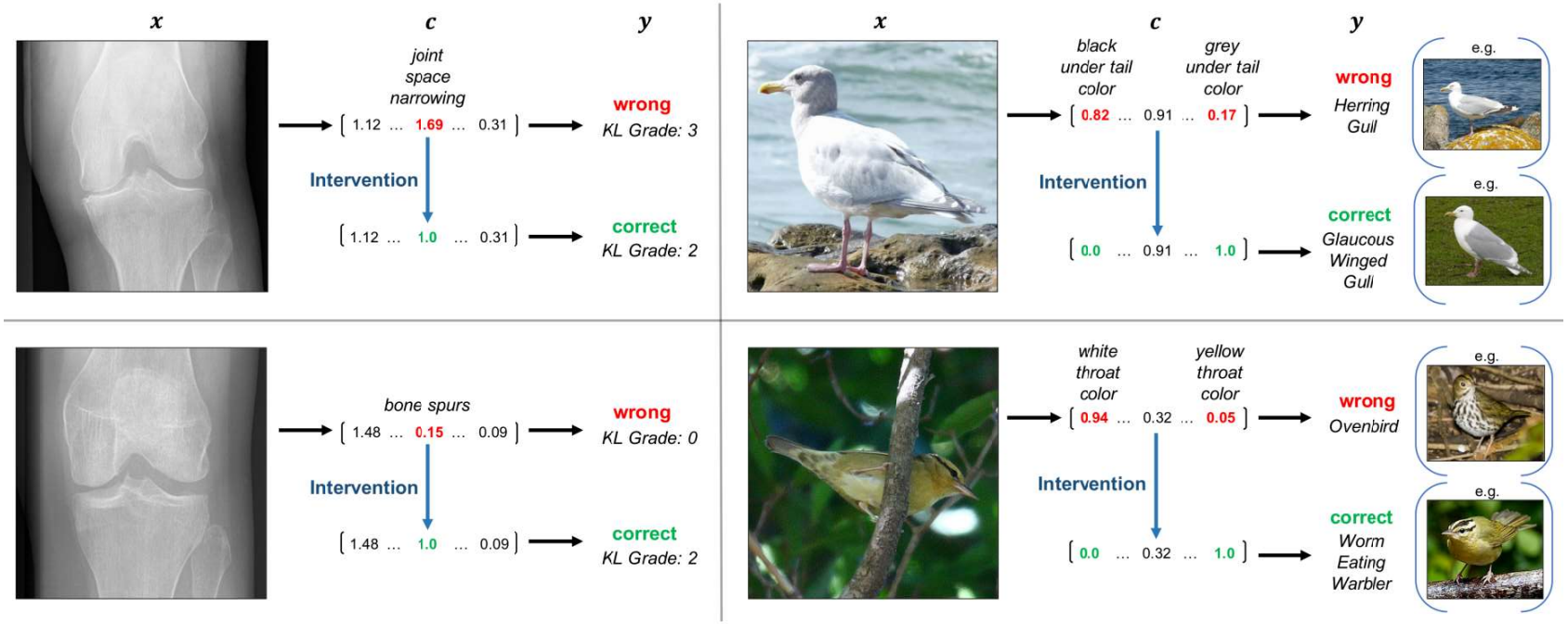

図 2.8: Concept Bottleneck Model の概要およびモデル推論の修正例．中間表現が言語として表現されるため，間違ったコンセプトの回帰を修正することでモデル推論を修正出来る．図は文献 [68] より引用．

2.3.2 Ante-hoc なアプローチ

Ante-hoc なアプローチは本質的に解釈可能なモデルを構築・学習することにより解釈可能性の実現を目指すアプローチである．Ante-hoc なアプローチではモデルの推論過程と説明の生成が密接に関連するため，説明の忠実性が保証される利点がある．しかし，一般的に，本質的に解釈可能な White-box モデルは Black-box モデルと比較し精度が低い課題がある．この課題に対処するため，深層学習による特徴抽出と解釈可能性を組み合わせた 'Gray-box' モデルを構築する手法が提案されている（図 2.7）．Gray-box モデルはその出力する説明により，Attention-based な手法及び Concept-based な手法に分類することが出来る．Attention-based な手法 [49][107][153] では注意機構を用いることにより，モデルが入力画像中のどの領域に着目するかを可視化し，着目領域のみをモデル後段に入力する．これより，Attention-based な手法では視覚的説明の実現及びモデル性能の改善を両立することが出来る．またクラス分類タスクでは注視領域の修正という直感的な方法によりサンプルに対するモデル推論を修正できることが知られている [92]．一方で，Attention-based な手法には Post-hoc なアプローチにおける Saliency-based な手法と同様にヒートマップ以上の情報を提示することが出来ない課題が存在する．

ヒートマップ以上の情報を提示することが出来ない課題に対し．コンセプトを用いて推論過程を説明する Concept-based の Gray-box モデルが提案されている．Concept-based の Gray-box モデルではまず意味的に接地された中間特徴表現を深層学習モデルにより学習する．その後，得られた中間特徴表現を本質的に解釈可能な White-box モデルで分類することにより解釈可能性を実現する．したがって，Concept-based の Gray-box 手法ではどのように「意味的に接地された中間特徴表現」が獲得されるように深層学習モデルを学習させるかが課題となる．以下では Concept-based な Gray-box モデルを実現する上で特に注目を集めている二つのアプローチ，Concept bottleneck model[68] および 'This looks like that' フレームワーク [45] について説明する．

■ Concept Bottleneck Models

図 2.8 に示すように，Concept Bottleneck Model では回帰によって得られた（言語化された）コンセプトの存在度を解釈可能な中間特徴表現として用いる．すなわち Concept Bottleneck Model では

$$\mathcal{X} \to \mathcal{C} \to \mathcal{Y} \tag{2.3}$$

という推論プロセスを採用する．ただし，$\mathcal{X}, \mathcal{C}, \mathcal{Y}$ はそれぞれ入力データ空間，コンセプトの存在度を基底とする空間，およびラベル空間である．ここでコンセプトの存在度合いは各画像もしくは画像クラスに対し付与されたアノテーションを教示として学習した深層学習モデルにコンセプトの存在度を回帰させることで取得する．そのため，Concept Bottleneck Model の学習には多数のコンセプトの追加教示を必要とする課題がある．この課題に対処するため，Yuksekgonul らは TCAV の考え方を応用し，学習データとは異なるコンセプトに対する教示が与えられたデータセット（プローブデータセット）を用いることで，学習データに対するコンセプトの教示の必要性を解消した [157]．また，

大規模言語モデルを用いて，クラスに対するコンセプトの教示を自動的に得られるように改良した手法も提案されている [144][154]．特に Wang らは学習データ数が少ない条件において大規模言語モデルを活用した Concept Bottleneck Model が Black-box モデルと比較し高い性能を達成できることを示した [154]．これより，大規模言語モデルに埋め込まれた事前知識を活用した，解釈可能かつ高い精度を達成するモデル構築の実現が期待される．

Concept Bottleneck Model には中間特徴表現が自然言語として記述される特性から事前知識を埋め込みやすく，またサンプルに依らずモデルを修正することが出来る利点がある．一方で，コンセプトの特定はどのプローブデータセットを用いるかに大きく依存し，かつコンセプトを回帰するようモデルを学習することはクラスラベルを学習することよりも難しい場合があることが示唆されている [146]．そのため，Concept Bottleneck Model の学習では深層学習モデルに回帰させるべきコンセプトをどのように正しく選択すべきかを考慮する必要がある．また，著者の知る限り Concept Bottleneck Model を画像分類以外のタスクへ応用した手法は存在しておらず，幅広い画像認識タスクへ応用する上でも課題を抱えていると言える．

■ ‘This looks like that’ フレームワーク

‘This looks like that’ フレームワークでは Prototype と入力画像の類似度を解釈可能な中間特徴表現として採用する．ここで，Prototype はクラス毎に複数個用意されるランダムな値で初期化された特徴ベクトルであり，学習データ内のある画像パッチに接地するよう学習される．そのため，Prototype と入力画像の類似度は，学習データ中のある画像パッチと類似する画像パッチが入力画像内にどの程度存在するかという意味で解釈可能となる．すなわち ‘This looks like that’ フレームワークでは case-based な推論による解釈可能性が実現されることとなる．Chen らは ‘This looks like that’ フレームワークを実現する深層学習モデルのネットワークとして Prototypical Part Network (ProtoPNet) を提案した [45]．ProtoPNet では Prototype と入力画像の類似度は以下の式によって記述される．

$$s_{a,i} = Sim(x_a, \boldsymbol{p}_i) = \max_{\boldsymbol{z} \in \boldsymbol{Z}_a} \log\left(1 + \frac{1}{\|\boldsymbol{z} - \boldsymbol{p}_i\|_2^2 + \epsilon}\right) \tag{2.4}$$

ただし x_a は入力画像であり，下付き文字の a はデータインデックスを表す．また $\boldsymbol{p}_i$ は Prototype であり，$\boldsymbol{Z}_a$ は入力画像 x_a を深層学習モデルに入力した結果得られる特徴マップ，ϵ は数値計算の安定化のため導入される定数である．最後に ProtoPNet では Prototype と入力画像の類似度を線形分類器に入力することでクラス分類を行う．すなわち，入力 x_a に対する予測ラベル $\hat{y}_a$ は

$$\hat{y}_a = \arg\max_y \sum_j W_j^y s_{a,j} \tag{2.5}$$

と表される．そこで，図 2.10 に表されるように，線形分類層の重みおよび Prototype と入力画像の類似度を用いて，入力画像に対する分類予測の推論根拠が説明されることとなる．

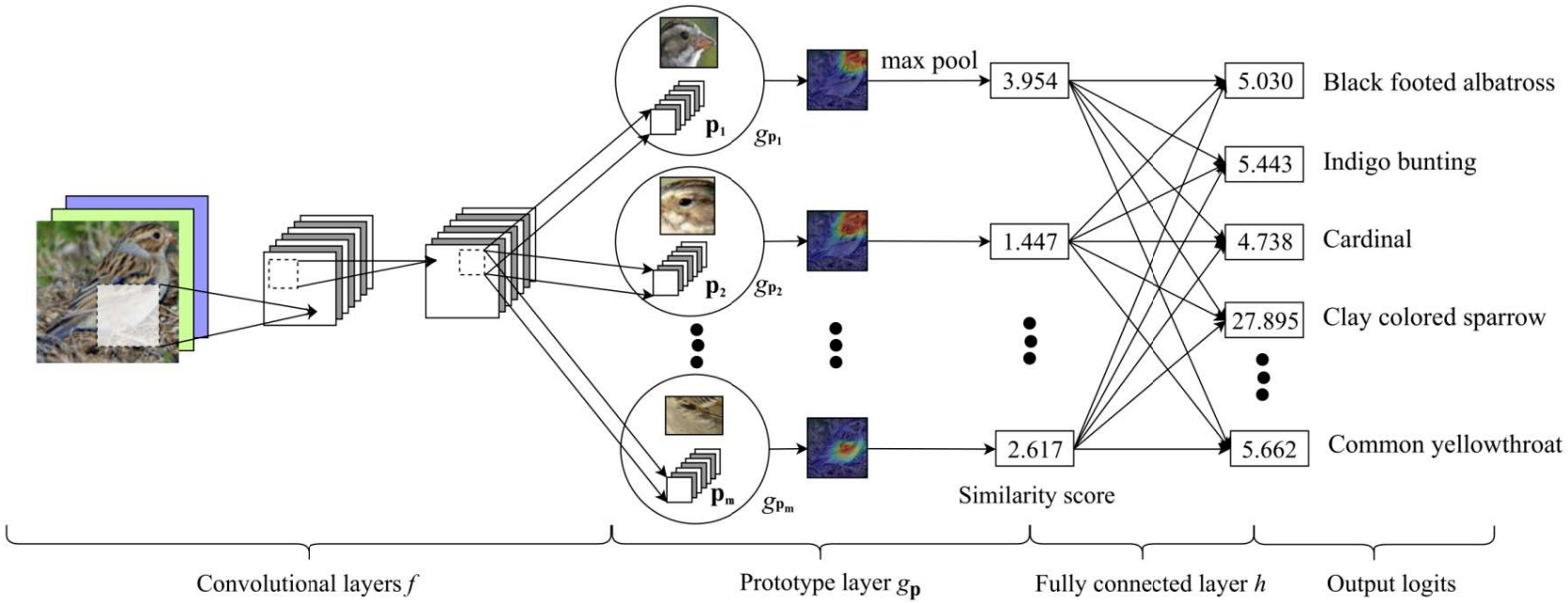

図 2.9: ProtoPNet のモデル構造．図は文献 [45] より引用．

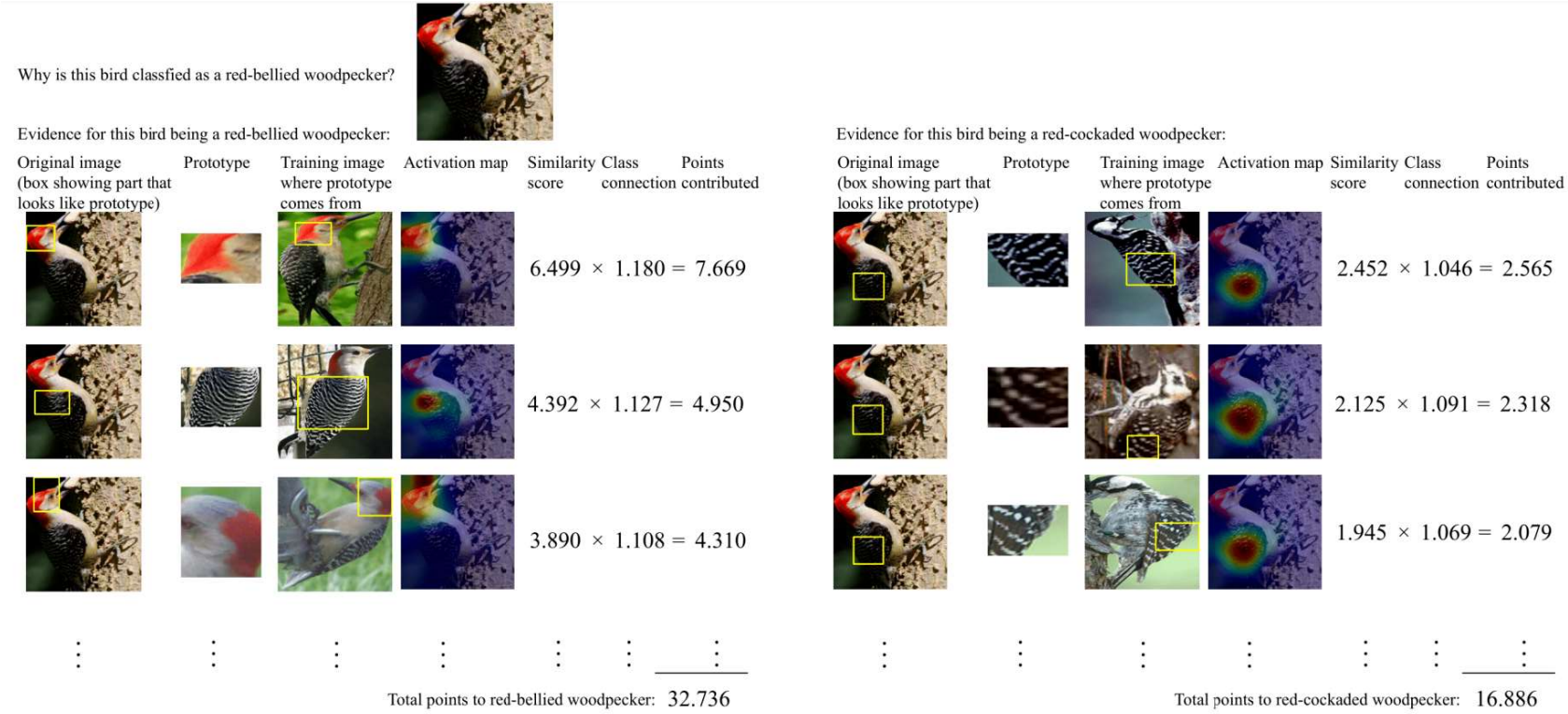

図 2.10: ProtoPNet による説明の例．図は文献 [45] より引用．

Chen らによって ProtoPNet が提案されて以降，様々な ‘This looks like that’ フレームワークに基づく本質的に解釈可能なモデルが提案されてきた．性能改善の観点では Grassmann 多様体上で Prototype を学習することにより Prototype の redundancy を改善しようとした Tesnet[102] や，Prototype の画像内での変形に対応することを目的として Deformable Convolution を導入した Deformable ProtoPNet[116] が提案されている．また Wang らはクラス境界に近い Prototype を学習し，Support Vector Machine の Support として利用することにより大幅な精度向上を達成できることを示した [151].

さらにメモリ効率の観点からはクラス毎に用意された Prototype をクラス間で共有することにより Prototype 数を削減する手法が提案されている [93][97][124][142]．特に Nauta は後段の White-box な分類器として決定木を採用することによりクラス間で共通した Prototype を学習することを初めて可能とした [93]．また Rymarczyk は Prototype をどのクラスへ割り当てるかを学習により決定することで，線形分類器を用いる場合においてもクラス間で共通した Prototype を学習可能となることを示した [124].

‘This looks like that’ フレームワークの学習にはクラスラベル以上の教示を必要としない利点がある．加えて Prototype による case-based な推論による解釈可能性の実現は画像認識以外のタスクにおいても有効となる．そのため ‘This looks like that’ フレームワークはグラフ分類 [127] や強化学習 [139]，Semantic Segmentation[149] 等の画像分類以外のタスクへの応用が進んでいる．また Prototype の Purity が低い（Prototype が特定の部位以外にも発火する）課題に対しても Prototype の学習において自己教師あり学習を活用する手法が有効であると確認されている [142]．加えてモデルデバッグの観点からは学習された Prototype を修正する手法が提案されており [130][143]，その有効性が実験的に確認されている．以上の背景から ‘This looks like that’ フレームワークは解釈可能な深層学習モデルを実現するための有効なアプローチとして幅広い注目を集めている．

以上のように，解釈可能な深層学習モデルの実現を目的として現在に至るまで多くの Gray-box なモデルが提案されてきた．これらの手法により深層学習モデルの性能を損なうことなく，モデルの解釈可能性は大きく向上されてきた．一方で，これらの手法は一般にクラス分類問題を対象としており，画像認識タスクの重要な基礎技術である類似度学習手法への適用はほとんど検討されてこなかった．本論文では，類似度学習に適用可能な ‘This looks like that’ フレームワークに基づく Gray-box モデルを実現する．

2.3.3 本研究で実現する解釈可能性

深層学習の解釈可能性の研究では実現すべき目的の違いにより，異なる意味合いで解釈可能性が論じられる事に注意する必要がある．例えば Post-hoc なアプローチにおける Saliency-based な手法や Gray-box モデルを構築する手法では，モデルの推論根拠に対する説明可能性が主な目的となる．また，これらの手法ではモデルの推論根拠を人間が理解出来るように説明することを目的とするため，出力される説明は直感的に分かりやすいことが重要となる．一方でこれらの手法では，深層学習モデル内部に対する解釈可能性—すなわち，なぜそのような中間表現が発現したかや，何故その領域

が注視されたかに関する説明—は実現されない．他方で Post-hoc な Concept-based の手法ではモデル全体の解析を行うことを目的に解釈可能性が実現される．そのため Post-hoc な Concept-based の手法により，モデル内部でどのような特徴表現がなされているかについて（部分的ではあるものの）直感的な解釈が得られる．しかし，モデル全体を解析する目的のため，Post-hoc な Concept-based の手法では説明の分かりやすさより説明の詳細さが追及される．そのため，Post-hoc な Concept-based の手法により生成される説明は一般的に煩雑で分かりにくいものとなる．また，先述したように Post-hoc な手法で生成される説明には忠実性に課題を抱えている事が知られている [57]．加えてニューロン発火パターンの意味の特定はその手続きに依存する点に注意が必要である [146]．先述した解釈可能性に関する研究の他にも ‘mathematically interpretable’ な White-box モデル [155] の提案がなされており，これらのモデルではモデル各層の役割が明白なため，各層が意図した動作をしているかを検証することが可能となる．また，‘mathmatically interpretable’ なモデルでは，特別に設計された損失関数を用いることなしに Segmentation-like なよい特徴表現が発現することが実験的に確認されている [156]．しかし，‘mathematically interpretable’ なモデルでは個別のサンプルに対する（直感的に分かりやすい）深層学習モデル推論根拠の説明はなされない．

説明が社会的な要素を多分に含み状況や対象に応じて適切な説明が変化することを踏まえれば，全ての問題設定・状況において統一的に適用可能な解釈可能性は存在しないと考えられる．したがって求められる解釈の程度や種別等，どのような解釈可能性を実現するかは個々の問題設定に応じて適切に決定されるべきと言える．本研究ではエンドユーザに推論根拠を説明可能な類似度学習モデルの構築を目的とする．そのため，本論文ではモデル推論根拠の（直感的な）説明が可能という意味で解釈可能な類似度学習モデルを ‘This looks like that’ フレームワークに基づき実現する．

第3章

人物再同定における大域的特徴と局所的特徴を利用した教師なしドメイン適応

本章では局所的な特徴と大域的な特徴を組み合わせた，人物再同定における教師なしドメイン適応手法を提案する．2 章で説明したように，人物再同定における教師なしドメイン適応では疑似ラベルを用いるアプローチに基づく手法が多く提案されている．しかし，これらの手法のうちの多くは GAP より出力される大域的な特徴のみを用いており，人物間を見分けるために重要な局所的特徴を考慮していない．また局所的な特徴を考慮する手法 [48] においても，個別にクラスタリングした結果得られる疑似ラベルをそのまま用いており，疑似ラベルを組み合わせることの有効性は検証されていない．

各特徴をクラスタリングした結果得られる疑似ラベルは，各特徴抽出の観点から各画像が同一人物か否かをデータセット全体を考慮し判定した結果と考えることが出来る．したがって，異なる観点から抽出された特徴ベクトルをもとに算出される疑似ラベルを組み合わせることで，より積極的に外見のよく似た異なる人物（Hard-Negative）に対する弁別性を高めるよう学習することが可能となると考えられる．そこで本章では GAP，GMP 出力をもとに算出された疑似ラベルを組み合わせる新規の教師なしドメイン適応手法を提案する．より具体的には，二つの疑似ラベルセットに加えそれらの積集合セットを用いて学習を行う教師なしドメイン適応手法を提案する．積集合セットにおいて同一の疑似ラベルが付与された画像は，二つの異なる観点の両方で同一人物と判断された画像と捉えることが出来る．そこで積集合セットを用いて学習を行うことで，ターゲットドメインへの適応を行う際の学習過程において外見のあまり似ていない同一人物画像（Hard-Positive）を取り込むために多くの Hard-Negative が混在した状況でも Hard-Negative に対する弁別性を高めるよう学習することが出来ると考えられる．

本章の構成は以下のとおりである．まず 3.1 章では提案手法について説明する．続く 3.2 章では提案手法の有効性を示すため実施した実験結果について説明し，最後に 3.3 章で本章をまとめる．

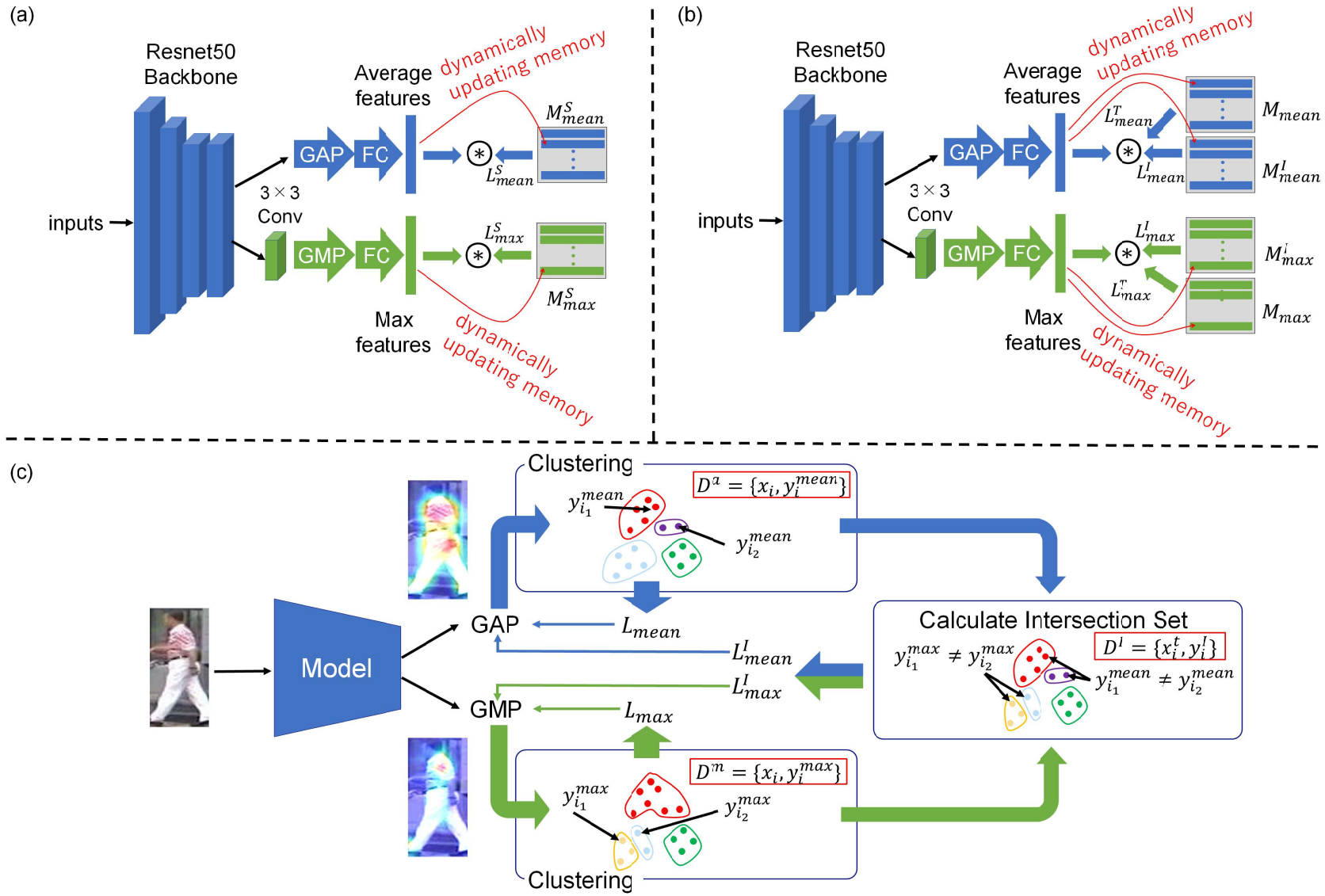

図 3.1: 提案手法のフレームワークとロススキーム．(a) ソースドメインにおける提案手法のロススキーム．ソースドメインでは正解ラベルを用いて，3.2 式で定義される損失関数により GAP，GMP ブランチをそれぞれ個別に教示する．(b)，(c) ターゲットドメインにおける提案手法のロススキームおよびフレームワーク．ターゲットドメインでは GAP，GMP ブランチの各出力を個別にクラスタリングして得た疑似ラベルセットと両疑似ラベルの積集合をとり作成した積集合セットを用いて学習を行う．

3.1 提案手法

2 章で説明したように，ターゲットドメインへの適応時にソースドメインの学習データを必要とする 1-stage の方法は，プライバシーの問題など実用上の課題を抱える．そこで提案手法では 2-stage のアプローチによるドメイン適応を実施する．すなわち，ソースドメインにおけるラベル付きの学習データを用いて深層学習モデルの学習を実施した後，ターゲットドメインにおけるラベルの付与されていない学習データのみを用いてターゲットドメインへのドメイン適応を実施する．以下では，ソースドメインでの学習及びターゲットドメインへのドメイン適応で用いたノンパラメトリックな分類器による教示について説明した後（3.1.1 節），ソースドメインにおける学習について説明し（3.1.2 節），最後にターゲットドメインへの適応について説明する（3.1.3 節）．

3.1.1 ノンパラメトリックな分類器による教示

提案手法では SpCL[65] と同様にノンパラメトリックな分類器を用いて学習した．これは大域・局所特徴両方を共有する人物画像数は少数で，高々画像数枚のみから構成される場合があるためである．本節では，このノンパラメトリックな分類器による教示について説明する．ノンパラメトリックな分類器による学習では，ラベル付きデータセット $\mathcal{D} = \{x_i, y_i\}_{i=0,1,...}$ が与えられた際，外部メモリ $\boldsymbol{M}$ を構築し各ラベルに含まれる特徴量の平均値によりメモリ $\boldsymbol{M}$ を初期化する．ここで，x_i はデータセットに含まれるデータサンプルであり，y_i はサンプル x_i に付与されたラベルである．また下付き文字の i はデータのインデックスである．すなわちメモリ $\boldsymbol{M}$ は各エポックの最初に下式により初期化される:

$$\boldsymbol{M}|_{\mathcal{D},F}[k] \leftarrow \frac{\frac{1}{N_k}\sum_{(x_i,y_i)\in\mathcal{D}} \mathbf{1}(y_i = k)\hat{F}(x_i)}{\|\frac{1}{N_k}\sum_{(x_i,y_i)\in\mathcal{D}} \mathbf{1}(y_i = k)\hat{F}(x_i)\|_2} \tag{3.1}$$

ここで F はサンプル x_i を特徴量 $F(x_i)$ に変換するエンコーダであり，N_k は疑似ラベル k を持つサンプル数，$\hat{F}(x_i)$ は特徴量 $F(x_i)$ を L2 正規化したベクトル，$\mathbf{1}(\cdot)$ は括弧内の条件が真のとき 1，偽のとき 0 を返す特性関数である．また，$\boldsymbol{M}|_{\mathcal{D},F}$ はメモリ $\boldsymbol{M}$ がデータセット $\mathcal{D}$ についてエンコーダ F を用いて構築されたことを明示しており，$[k]$ は k 番目の列要素であることを示している．このように構築されたメモリ $\boldsymbol{M}$，データセット $\mathcal{D}$，およびエンコーダ F が与えられた時，サンプル x_i に対する損失は，

$$L_{mem}(x_i, y_i|\boldsymbol{M}, \mathcal{D}, F) = -\frac{\exp(\boldsymbol{M}[y_i]^T \cdot \hat{F}(x_i)/\tau)}{\sum_{y\in\mathbb{Y}} \exp(\boldsymbol{M}[y]^T \cdot \hat{F}(x_i)/\tau)} \tag{3.2}$$

と定義される．ただし，$\mathbb{Y}$ はデータセット $\mathcal{D}$ に含まれる全てのクラスラベルの集合であり，τ は温度パラメータである．3.2 式は特徴量 $F(x_i)$ とメモリに格納された各特徴量とのコサイン類似度を温度パラメータで除算した値に Softmax を適用した結果得られるベクトルの，正解クラスに対応する成分に -1 を乗算した値とわかる．よって 3.2 式は特徴量 $F(x_i)$ とメモリに格納された各ラベルに対

応する特徴量とのコサイン類似度から，正解ラベルに x_i が属する確率を算出し，その最大化を行う損失関数と捉えられる．温度パラメータは確率算出時の Softmax の鋭さを調節するパラメータであり，温度パラメータを低くするほど Softmax の出力は argmax に近づき，高くするほど Softmax の出力は一様分布に近づく．したがって 3.2 式における温度パラメータは，正解ラベルに x_i が属する確率をどの程度 Soft に算出するかを調節するパラメータであると言える．また，メモリ $\boldsymbol{M}$ は学習時，下式により動的に更新される:

$$\begin{aligned} \boldsymbol{M}[y_i] &\leftarrow m \cdot \boldsymbol{M}[y_i] + (1-m) \cdot \hat{F}(x_i) \\ \boldsymbol{M}[y_i] &\leftarrow \frac{\boldsymbol{M}[y_i]}{\|\boldsymbol{M}[y_i]\|_2} \end{aligned} \tag{3.3}$$

ただし，m はメモリの更新量を決めるハイパーパラメータである．本論文では SpCL[65] に従い二つのハイパーパラメータ τ, m をそれぞれ，0.05 および 0.2 で固定した．

3.1.2 ソースドメインにおける学習

続いてソースドメインにおける事前学習について説明する．ソースドメインでは，正解ラベル付きデータセット $\mathcal{D}^S = \{x_i^s, y_i^s\}_{i=0,1,\ldots}$ が与えられる．ここで，上付き文字の s はソースドメインのデータであることを明示している．提案手法ではソースドメインでの学習により，人物再同定において有効な大域的・局所的特徴を抽出するモデルを作成する．以下では提案手法で用いたモデルのモデル構造について説明した後，ソースドメインにおける事前学習の詳細を説明する．

まず提案手法で用いたモデルの構造について説明する．提案手法では図 3.1（a）に示すように最終プーリング層が GAP および GMP である，二つのブランチ構造を持つモデルを用いる．モデルバックボーン出力に GAP と GMP を各々適用することにより，各ブランチから大域的特徴量と局所的特徴量を抽出することが可能となる．ただし，GAP ブランチから流入する勾配と比較して強すぎる勾配流入を緩和するため，GMP の直前には，3×3 畳み込み層と batch normalization（BN）層を追加した．また，GAP，GMP 直後に全結合層と BN 層を追加することで，次元数を 2048 次元から 1024 次元に削減し正規化を行った．モデルのバックボーンには，ImageNet で事前学習した Resnet 50[13] を使用し，[53] に従い，Layer4 における stride は 1 に設定し直した．

次に，ソースドメインにおける事前学習の詳細を説明する．ソースドメインでの学習では，図 3.1（a）に示すように，ノンパラメトリックな分類器により学習を進める．ここで，図 3.1 では GAP および GMP 出力に対する処理を青および緑色で示した．ソースドメインにおける学習では，先述したモデルによりサンプル x_i^s が大域的（GAP）特徴 $F_a(x_i^s)$ と局所的（GMP）特徴 $F_m(x_i^s)$ へと変換される．ここで，F_a および F_m はそれぞれサンプル x_i を GAP 特徴 $F_a(x_i^s)$ および GMP 特徴 $F_m(x_i^s)$ へと変換するエンコーダを表している．続いて，ラベル付きデータセット D^S について，大域・局所的特徴それぞれに対してメモリ M_{mean}^S（$\equiv M|_{D^S,F_a}$）および，M_{max}^S（$\equiv M|_{D^S,F_m}$）を構築する．その後，各ブランチを 3.2 式により教示することで学習を実施する．すなわち，ソースドメインにおいては，以下の損失関数を用いて学習を進めることとなる：

Algorithm 1 ターゲットドメインへのドメイン適応における学習プロセス

Require: An unlabelled training set $\mathcal{D} = \{x_i^t\}$;

Require: Source pretrained model $F_\theta \equiv \{F_a, F_m\}$;

Require: Temperature τ and momentum m;

for n in [1, num_epochs] **do**
  Encode all of the training samples $\{x_i^t\}$ to GAP and GMP features $\{F_a(x_i^t)\}$, $\{F_m(x_i^t)\}$ by F_θ;
  Generate two pseudo-labeled sets $\mathcal{D}^a = \{x_i^t, y_i^a\}$, $\mathcal{D}^m = \{x_i^t, y_i^m\}$ by clustering $\{F_a(x_i^t)\}$, $\{F_m(x_i^t)\}$ respectively;
  Generate intersection sets $\mathcal{D}^I = \{x_i^t, y_i^I\}$ with Eq.3.5;
  Initialize four external memories $\boldsymbol{M}_{mean}$, $\boldsymbol{M}^I_{mean}$, $\boldsymbol{M}_{max}$, $\boldsymbol{M}^I_{max}$ with Eq.3.1;
  for each mini-batch $\{x_i^t\} \in \mathcal{D}$ **do**
    Encode features $\{F_a(x_i^t)\}$, $\{F_m(x_i^t)\}$ with F_θ;
    Compute loss with non-parametric classifier with Eq.3.6 and update F_θ with back-propagation;
    Update memories with Eq.3.3;
  end for
end for

$$
\begin{aligned}
L_{src} &= \sum_{(x_i^s, y_i^s) \in \mathcal{D}^S} L_{mem}(x_i^s, y_i^s | \boldsymbol{M}^S_{mean}, \mathcal{D}^S, F_a) + \sum_{(x_i^s, y_i^s) \in \mathcal{D}^S} L_{mem}(x_i^s, y_i^s | \boldsymbol{M}^S_{max}, \mathcal{D}^S, F_m) \\
&\equiv L^S_{mean} + L^S_{max}
\end{aligned}
\tag{3.4}
$$

3.1.3 ターゲットドメインへの適応

次にターゲットドメインでの学習について説明する．図 3.1（b）および（c）に示すように提案手法では三つの疑似ラベルセットを用いてモデルをターゲットドメインへと適応させる．これらは GAP, GMP から出力される特徴を個別にクラスタリングした結果得られる疑似ラベルセットとそれらの積集合セットである．ここで，個別に生成される疑似ラベルセットを用いて各ブランチを学習することにより，それぞれの特徴の特性を個別に強化することが可能となる．加えて，個別に生成されたデータセットは各特徴の特性を反映しているため，積集合でラベルが同じ画像は大域的・局所的という両方の観点で似た画像と言え，積集合セットによる教示ではこれらの人物画像のみが同一人物として教示される．言い換えると一方の特徴抽出の観点でのみ同一人物となる人物画像は積集合セットの教示により，異なる人物としてそれらの弁別性を向上するように学習がなされる．そのため，積集合による教示により，Hard-Positive を積極的に紐づけるために Hard-Negative が多く混在した状況でも，Hard-Negative に対するモデルの弁別性を十分に高めることが可能となる．

まずターゲットドメインへの適応における手法全体の概要を説明する．ターゲットドメインでは，サンプル x_i^t のみが与えられる．ここで上付き文字の t はターゲットドメインのデータであることを

示している．提案手法では，図 3.1（c）に示すように，まずターゲットドメインのデータに対し疑似ラベルを付与することで 3 つの疑似ラベルデータセットを作成する．その後，図 3.1（b）に示すように，各疑似ラベルセットに対して，モデル外部に構築したメモリによるノンパラメトリックな分類器を用いて学習を実施する．以下では手法の詳細を説明する．また，アルゴリズムの全体を Algorithm 1 に示す．

まず初めに，学習に用いる 3 つの疑似ラベルセットの構築方法について説明する．ラベル付けされていないデータ集合 $\mathcal{D}^t = \{x_i^t\}_{i=0,1,...}$ が与えられたとき，提案手法ではまず $\mathcal{D}^t$ の全てのデータを大域（GAP）特徴 $\{F_a(x_i^t)\}_{i=0,1,...}$ と局所（GMP）特徴 $\{F_m(x_i^t)\}_{i=0,1,...}$ に変換する．その後，GAP, GMP 特徴に対し，それぞれ個別に DBSCAN クラスタリングアルゴリズムを適用することで二つの疑似ラベルセット $\mathcal{D}^a = \{x_i^t, y_i^a\}_{i=0,1,...}$，$\mathcal{D}^m = \{x_i^t, y_i^m\}_{i=0,1,...}$ を生成する．ここで，y_i^a，y_i^m はそれぞれサンプル x_i^t から抽出された GAP，GMP 特徴量に対しクラスタリングによって付与された疑似ラベルである．次に生成された二つの疑似ラベルセットから，下式に基づき隣接行列 $\boldsymbol{K}^I$ を計算する．

$$K_{i,j}^I = \begin{cases} 1 & (\text{if } y_i^m = y_j^m \text{ and } y_i^a = y_j^a) \\ 0 & (\text{otherwise}) \end{cases} \tag{3.5}$$

その後，隣接行列における値が 1 となるペアに対し同一のラベルを，その他のペアに異なるラベルを付与することで，積集合データセット $\mathcal{D}^I = \{x_i^t, y_i^I\}_{i=0,1,...}$ を計算する．y_i^I は積集合データセットに付与されたラベルを表しており上付き文字の I は積集合データセットに対するラベルであることを明示している．ここで疑似ラベルセット $\mathcal{D}^a$，$\mathcal{D}^m$ には各特徴間の特性の違いがデータセット全体という意味でグローバルに反映されている．一方，$\mathcal{D}^a$，$\mathcal{D}^m$ の積集合セット $\mathcal{D}^I$ では一方のデータセットでは同一人物とされたペアでも，もう一方のデータセットでは別人とされたペアは別人としてラベル付けされる．そこで，積集合セット $\mathcal{D}^I$ による教示は $\mathcal{D}^a$，$\mathcal{D}^m$ それぞれにおいて，もう一方の特徴表現の特性を考慮し，特徴間の特性の違いをもとにクラス内サンプルの特徴表現の多様性を保持・強化するよう学習する事に繋がる．そのため，積集合セット $\mathcal{D}^I$ を用いることで，大域的・局所的な特徴間の特性の違いを考慮した特徴表現の学習を実施することが可能となる．

その後図 3.1（b）に示すように，生成した各データセットに基づき，4 つの外部メモリ $\boldsymbol{M}_{mean}$，$\boldsymbol{M}_{mean}^I$，$\boldsymbol{M}_{max}$，$\boldsymbol{M}_{max}^I$ を構築する．ここで，$\boldsymbol{M}_{mean}$，$\boldsymbol{M}_{mean}^I$，$\boldsymbol{M}_{max}$，$\boldsymbol{M}_{max}^I$ はそれぞれ，3.1 式の表記を用いて，$\boldsymbol{M}|_{\mathcal{D}^a,F_a}$，$\boldsymbol{M}|_{\mathcal{D}^I,F_a}$，$\boldsymbol{M}|_{\mathcal{D}^m,F_m}$，$\boldsymbol{M}|_{\mathcal{D}^I,F_m}$ と表される．すなわち $\boldsymbol{M}_{mean}$，$\boldsymbol{M}_{mean}^I$，$\boldsymbol{M}_{max}$，$\boldsymbol{M}_{max}^I$ はそれぞれ $M|_{\mathcal{D}^a,F_a}$，$M|_{\mathcal{D}^I,F_a}$，$M|_{\mathcal{D}^m,F_m}$，$M|_{\mathcal{D}^I,F_m}$ の添え字で表されるデータセット $\mathcal{D}^*$（$* \in \{a, m, I\}$）に対してエンコーダ F_*（$* \in \{a, m\}$）を用いて 3.1 式により構築されたメモリである．その後，図 3.1（b）に示すように，構築された疑似ラベルセットおよびメモリを用いて，3.1.1 節で説明した損失（3.2 式）を各ブランチ出力，各サンプルに対して計算し，学習を実施する．

よって，最終的に最小化すべき損失関数は以下で与えられる：

$$
\begin{aligned}
L_{tgt} = & \sum_{(x_i^t, y_i^a) \in \mathcal{D}^a} L_{mem}(x_i^t, y_i^a | M_{mean}, \mathcal{D}^a, F_a) + \sum_{(x_i^t, y_i^m) \in \mathcal{D}^m} L_{mem}(x_i^t, y_i^m | M_{max}, \mathcal{D}^m, F_m) \\
& \lambda_I \cdot \sum_{(x_i^t, y_i^I) \in \mathcal{D}^I} L_{mem}(x_i^t, y_i^I | M_{mean}^I, \mathcal{D}^I, F_a) + \lambda_I \cdot \sum_{(x_i^t, y_i^I) \in \mathcal{D}^I} L_{mem}(x_i^t, y_i^I | M_{max}^I, \mathcal{D}^I, F_m) \quad (3.6) \\
& \equiv L_{mean}^T + L_{max}^T + \lambda_I \cdot L_{mean}^I + \lambda_I \cdot L_{max}^I
\end{aligned}
$$

ここで λ_I は積集合セットによる教示の強さを調節するハイパーパラメータであり，疑似ラベルセット $\mathcal{D}^a$，$\mathcal{D}^m$ におけるクラス内サンプルに対する特徴表現の多様性をどの程度強化するかを調節する．本論文ではすべての実験で $\lambda_I = 1.0$ とした．

3.2 評価実験

本章では提案手法の有効性を評価するため実施した評価実験の内容について説明する．本章における実験では実画像データセット間におけるドメイン適応の評価として，Market1501[24] から MSMT17[44] へのドメイン適応，および MSMT17 から Market1501 へのドメイン適応の二つの実験設定で評価した．また，CG データで学習したモデルから実画像データセットへのドメイン適応の評価として，PersonX[58] および UnrealPerson[105] から Market1501 および MSMT17 へのドメイン適応の四つの実験設定で評価した．以下ではまず，実験設定の詳細を説明し，実画像データセット間のドメイン適応における提案手法と他手法との比較結果を説明する．次に CG 画像から実画像データセットへのドメイン適応の結果について述べ，最後に Ablation Study による提案手法の各モジュールの有効性を検証した結果を説明する．

3.2.1 実験設定の詳細

以下では，まずソースドメインにおける事前学習の詳細について述べたのち，ターゲットドメインにおける学習の詳細について述べる．ソースドメインにおける学習では，optimizer には Adam を採用し，weight decay は 0.0005 に設定した．また学習率は 0.00035 とし，20 エポック毎に学習率を 1/10 に減衰させ，50epoch 学習した．各エポックには 200 回の iteration が含まれるようにし，各ミニバッチには 16 の ID と各 ID に 4 枚の画像が含まれるようなサンプリング戦略を採用した．データの構成として，画像サイズは 256×128 にリサイズし，データ拡張として，Random Flipping，Cropping，Erasing を用いた．

ターゲットドメインにおける学習においても optimizer には Adam を採用し，weight decay は 0.0005 に設定した．また，学習率は 0.00035 とし，20epoch 毎に学習率を 1/10 に減衰させ，50epoch 学習した．ただし，各エポックには 1600 回の iteration が含まれるようにした．またバッチサイズは 16 に設定し，各ミニバッチには ID 毎に 4 枚の画像が含まれるようにした．これは GPU 一枚に含まれるバッチサイズを 16 とすることで高い精度が得られるという経験則 [64] のためである．しかし，ノン

表 3.1: Market1501[24] および MSMT17[44] データセット間での教師なしドメイン適応における従来手法と提案手法の比較結果．ただし表内において ‘MS’ は MSMT17 を表し，‘M’ は Market1501 を表す．また，2-stage の手法において最も高い精度となる値を太字，二番目に高い精度となる値に下線，三番目に高い精度となる値をイタリックで示し強調した．

Approach	Methods	MS→M			M→MS		
		mAP	R1	R5	mAP	R1	R5
2-stage	ECN[61]	-	-	-	8.5%	25.3%	36.3%
	SSG[48]	-	-	-	13.2%	31.6%	-
	MMCL[76]	-	-	-	15.1%	40.8%	53.1%
	JVTC[69]	-	-	-	19.0%	42.1%	53.4%
	NRMT[79]	-	-	-	19.8%	43.7%	56.5%
	D-MMD[71]	50.8%	72.8%	88.1%	13.5%	29.1%	46.3%
	DG-Net++[80]	*64.6%*	*83.1%*	*91.5%*	22.1%	48.4%	60.9%
	MMT-1500[64]	-	-	-	22.9%	49.2%	63.1%
	MMT-dbscan[64]	75.6%	89.3%	95.8%	*24.0%*	*50.1%*	*63.5%*
	ABMT[113]	-	-	-	23.2%	49.2%	-
	UNRN[108]	-	-	-	25.3%	52.4%	64.7%
	GCL [82]	-	-	-	21.5%	45.0%	57.1%
	Ours	**76.8%**	**89.9%**	**95.9%**	**30.8%**	**58.1%**	**71.0%**
1-stage	SpCL[65]	77.5%	89.7%	96.1 %	26.8%	53.7%	65.0%
	TDRL[87]	-	-	-	32.9%	61.8%	-

パラメトリックな分類器による学習ではメモリ内に含まれる特徴量が up-to-date であることが不可欠であるが，これは小さなバッチサイズと相容れない．このため，モデルの更新は 2 回の iteration の後に 1 回のみ実施した．これにより，1 枚の GPU で 2 枚の GPU を学習に用いるのと同等の効果が得られる．

またクラスタリングでは，DBSCAN[3] を用い，メトリックには k-reciprocal encoding[36] を採用した．k-reciprocal encoding 計算時のハイパーパラメータ k_1 および k_2 は SpCL[65] に基づきそれぞれ 30，6 で固定した．また，より積極的に Hard-Positive を紐づけるため，DBSCAN の閾値は 0.68 に設定した．

3.2.2 実画像データセット間のドメイン適応

表 3.1 に人物再同定における教師なしドメイン適応の state-of-the-art 手法と提案手法の比較結果を示す．表中における 1-stage のアプローチとはターゲットドメインでの学習においてもソースドメインのデータを利用しながら学習する手法である．人物再同定で扱う人物姿画像はプライバシーに配慮した慎重な取り扱いが必要となるため，ターゲットドメインにおける学習においてもソースドメインのデータを利用できると仮定することは重大な実用上の課題となりうる．そこで，本論文では 1-stage は参考値として掲載し，2-stage の手法間での比較することとした．表内において，2-stage のアプローチに基づく手法の中で 1，2，3 番目に mAP および R1 の高い手法をそれぞれ太字，下線，イタリックで強調した．以下では各設定での比較結果の詳細を説明する．

MSMT17 から Market-1501 へのドメイン適応は，人物再同定における教師なしドメイン適応において一般的な設定ではなく，従来手法においても結果を示したものは少ない．これは，MSMT17 が最も大規模なデータセットであり，Market-1501 および DukeMTMC が同規模のデータセットのため，DukeMTMC から Market-1501，もしくはその逆の設定が手法の有効性を示すために，一般的に用いられてきたためである．しかし，DukeMTMC は公開停止されており，これ以上用いるべきではない．そこで，本論文では MSMT17 から Market-1501 へドメイン適応する設定を採用する事とした．

さて，提案手法は，2-stage のアプローチをとるすべての手法に比べ，最も高い精度を達成しており，この設定における提案手法の有効性が確認出来る．一方，高精度を達成する 1-stage の手法との比較では SpCL[65] にはわずかに劣る結果となっている．これは SpCL は閾値を変更した複数回のクラスタリングを実施した結果から，生成されたクラスタの安定性・信頼性を評価し，信頼されるクラスタのみを同一人物画像として採用することで，データセットにおいて重要となる Hard-Negative の除去を効果的に行っているためと考えられる．実際，Market-1501 ではデータセット内の環境が均一であるためにカメラ間での同一人物の見え方がよく似通っており，Hard-Positive の紐づけを強く行なわずとも学習後期において同一人物の画像を紐づけることが可能となる．そのため，Hard-Negative の除去を行い誤差拡大を防ぐことが Market-1501 のようなデータセットへのドメイン適応では高精度の達成に重要と考えられる．また，SpCL は高精度を達成するためにターゲットドメインでの学習においてもソースドメインのデータを用いており，先述したように重大な実用上の課題を抱えている点に注意が必要である．

Market-1501 から MSMT17 へのドメイン適応はより照明変化が激しく，多くの人物画像が含まれる環境へのドメイン適応であり，最も困難な実験設定である．Market-1501 データセットはほぼ同一の時間帯の屋外環境で撮影された，ほぼ同様の照明環境の人物画像が含まれる一方で，MSMT17 では様々な時間帯，屋内・屋外といった多様な照明環境における人物画像が多く含まれている[1]．このような実験設定ではソースドメインでは存在しない照明変動に対し，異なる照明環境下における同一人物画像を紐づけつつ，同様の照明環境下（例えば逆光環境下）における別人を異なる人物と見分ける必要があり，この実験設定を非常に困難なものとしている．

[1]本論文においてもデータセットの画像を掲載すべきところではあるが，論文への MSMT17 データセット画像掲載は利用規約上認められていない．各データセットにおける画像例は [24][44] を参照されたい．

この設定においても，提案手法は 2-stage のアプローチに基づく手法の中で，すべての state-of-the-art 手法より高い精度を達成している．加えて，1-stage のアプローチに基づく手法との比較においても，SpCL を mAP において 4%上回る大きな精度向上を達成している．これらの結果より提案手法の MSMT17 という大規模なデータセットへのドメイン適応における有効性を確認できたと言える．一方，提案手法は TDRL[87] より mAP において 2%下回る精度となっているが，TDRL はターゲットドメインでの学習時ソースドメインのデータを利用できることを仮定しており，先述した実用上の課題を抱えている．加えて，両ドメインのデータを段階的に重みを変えながら学習を実施しているため，多くのハイパーパラメータを調整する必要がある．

CMC スコアの変遷を確認すれば，提案手法は他の 2-stage 手法と比較し，全ての Rank において約 6%向上していることが確認できる．すなわち他手法と比較し提案手法では上位類似画像を取得した際に同一人物がそのうちに含まれる割合を 5%向上することに成功しており，提案手法の改善幅がこの実験設定において大きいことが読み取れる．また，実験結果において注記すべき点が 2 点ある．第 1 に提案手法は 2 回のクラスタリングを実施する一方，SpCL[65] は 3 回クラスタリングを実施する必要があり，また SpCL はターゲットドメインへの適応時にも学習データを必要とするにも関わらず，提案手法は SpCL より約 4%の mAP 向上を示している．第 2 に，ABMT[113] も提案手法同様に GAP と GMP を併用するモデル構造を採用し，大域的特徴と局所的特徴の両方を利用しているが，提案手法は ABMT より，約 7%と大幅な mAP 向上を達成している．これら 2 点は，提案手法の大域的な特徴および局所的な特徴に加え，それら特徴間の違いを利用することでさらなる精度向上が可能であるという本論文の主張を裏付けており，提案手法の有効性を示している．state-of-the-art 手法の内いくつかの手法は自己蒸留により高い精度を達成するため，学習時に複数のモデルを必要とする [64][79][113]．特に MMT[64] は学習時に 4 つのモデルを同時に必要とするため，多大な計算資源・計算コストを必要とする．一方提案手法は 1 つのモデルを用いるのみで，計算コストも少ないながらも，より高い精度を達成している点は注目に値する．

以上まとめると，提案手法は人物再同定における二つのデータセット間における教師なしドメイン適応の設定において state-of-the-art 手法と同等の精度を達成した．特に最も困難な MSMT17 へのドメイン適応では 2-stage の従来手法と比べ，大幅な精度向上を達成している．これより従来手法に対する提案手法の有効性・優位性が示されたと考えられる．

3.2.3 CG データセットから実画像データセットへのドメイン適応

表 3.2 に CG 画像から実画像データセットへの教師なしドメイン適応における提案手法の結果を示す．結果より PersonX をソースドメインとして利用する場合には，実画像データセットを用いて教師なしドメインを実施した場合と比較し，大幅に性能が劣化していることが確認できる．これは，PersonX データセットに含まれる人数が少ないことに加えて，CG 画像に含まれる背景バリエーションが少ないために，ターゲットドメインへの適応に良い初期値を獲得できていないためと考えられる．実際 2 章で述べたように，PersonX の学習データに含まれる人物数は 410 人と Market1501 と比較しても小規模であり，カメラ数も 4 つと少ないために背景バリエーションも少ない．一方で UnrealPerson

表 3.2: CG 画像から実画像データセットへの教師なしドメイン適応における提案手法の結果．また表中の括弧内には実画像データセットをソースドメインとして用いた場合との精度差を記した．

Target / Source	Market1501[24]		MSMT17[44]	
	mAP	R1	mAP	R1
PersonX[58]	71.0% (−5.8%)	86.8% (−3.1%)	21.4% (−9.4%)	45.3% (−12.6%)
UnrealPerson[105]	77.3% (+0.5%)	89.8% (−0.1%)	38.0% (+7.2%)	66.0% (+7.9%)

では異なる 4 つの CG 環境における計 34 個のカメラを用いた多様な背景バリエーションに加え，かつ人物数も 6,799 人と大規模なデータセットを構築している．結果，UnrealPerson をソースドメインとして用いた場合には，Market1501 および MSMT17 へのドメイン適応において，実データを用いてドメイン適応した場合と比較し大幅な精度向上を達成している．これはソースドメインにおける学習がターゲットドメインへの適応に大きな影響を与えることを示しており，人物再同定における CG データによる学習の有効性を示している．どのような CG データで学習した場合にどのようなターゲットドメインへ上手くドメイン適応可能かについては活発に研究が進められており，ターゲットドメインへの適応に適切なソースドメインでの学習設計は今後の課題である．

3.2.4 Ablation Study

さらに提案手法の有効性を確認するため，Ablation Study を実施した．結果を表 5.4 に示す．まず，表内の単語を説明する．表内において ‘M’ は Market-1501，‘MS’ は MSMT17 を表している．‘Model’ の列において，‘GAP only’ は人物再同定において一般的に用いられるモデル構造 [53] を用いた結果であり，‘GMP only’ は ‘GAP only’ のモデル構造において GAP を GMP に変更した結果である．また，‘our model’ は提案モデルを用いた場合の結果を示している．次に ‘Condition’ における単語の定義を説明する．‘Oracle’ はターゲットドメインへの適応を実施する際に，疑似ラベルの代わりに真ラベルを用いた結果であり，精度の上限である．また，‘Direct transfer’ はソースドメインにて学習したモデルを直接ターゲットドメインで評価した結果であり，精度の下限である．‘w/ Target Adaptation’ は ‘GAP only’，’GMP only’ モデルに対し，ノンパラメトリックな分類器により 2-stage のアプローチによる教師なしドメイン適応を実施した結果である．すなわち，各モデルについて損失関数を提案手法と同様のノンパラメトリックな分類器とし，提案手法と同様の実験設定にて教師なしドメイン適応を実施した結果である．これは UDAP[75] で提案された，クラスタリングにより疑似ラベルを生成し学習を実施するという学習手法において，損失関数および実験設定を本論文における実験と揃え実行した結果と等価である．以後では本結果をそれぞれ単に ‘GAP only’，‘GMP only’ と表す．また，‘w/o L^I_{mean}，L^I_{max}’ は提案手法において積集合による教示を用いない場合の結果であり，‘ours full model’ は提案手法の全ての要素を用いた場合の結果である．すなわち表 5.4 に示す ‘w/o L^I_{mean}，L^I_{max}’ および ‘ours full model’ の結果は ‘GAP only’，‘GMP only’ に対し，GAP および GMP から出

表 3.3: Ablation study の結果．ただし表内において ‘MS’ は MSMT17[44] を表し，‘M’ は Market1501[24] を表す．また，‘Direct Transfer’ はソースドメインで学習したモデルをドメイン適応することなくターゲットドメインで評価した結果であり，‘Oracle’ は疑似ラベルとして正解ラベルを用いることでドメイン適応を実施した結果である．その他の実験条件および実験に対する考察等の詳細は本文を参照されたい．

Models	Conditions	MS→M		M→MS	
		mAP	R1	mAP	R1
GAP only	Oracle	85.1%	93.5%	47.6%	73.8%
	Direct transfer	29.0%	55.6%	3.2%	10.2%
	w/ Target Adaptation	75.2%	89.4%	28.0%	57.0%
GMP only	Oracle	84.1%	93.3%	48.6%	73.8%
	Direct transfer	30.5%	55.0%	4.2%	12.6%
	w/ Target Adaptation	72.7%	87.6%	25.2%	53.0%
our model	Oracle	84.9%	94.1%	52.5%	76.7%
	Direct transfer	23.4%	45.7%	2.9%	8.9%
	w/o L^I_{mean}, L^I_{max}	74.1%	88.8%	27.9%	54.4%
	ours full model	**76.8%**	**89.9%**	**30.8%**	**58.1%**

力される特徴それぞれより構築した疑似ラベルセットによる教示，および本論文の新規点である，それらの積集合セットによる教示それぞれの効果について検証した結果である．

以下では，各設定における結果の詳細を説明する．MSMT17 から Market-1501 の設定では，‘GMP only’ ではドメイン適応の結果は（‘w/ Target Adaptation’）他のモデル構造と比較し低い精度となっている．これは局所的特徴のみを考慮した結果，一部のみが似た多くの Hard-Negative の人物画像を紐づけてしまい，誤差拡大が発生してしまったためと考えられる．一方 ‘GAP only’ と比べ，‘w/o L^I_{mean}，L^I_{max}’ では僅かに精度が劣化している．これは Market-1501 のような比較的小規模のデータセットでは，Hard-Negative を除去することが Hard-Positive を紐づけるより重要となるためである．この仮説を裏付けるため Market-1501 へのドメイン設定における最終エポックで生成された疑似ラベルの特性を調査した．結果を表 3.4 に示す．表 3.4 において，‘clusters’ はクラスタ数を示しており ‘outliers’ は DBSCAN によるクラスタリングで疑似ラベルが付与されなかったサンプルの数を示している．また，‘our model’ 行の各セルにおいて ‘/’ 左側の結果は GAP 出力より生成された疑似ラベルに関する結果を示しており，右側の結果は GMP 出力より生成された疑似ラベルに関する結果を示している．表 3.4 の結果において ‘GAP only’ のモデルにおける ‘outlier’，‘clusters’ はそれぞれ 14，403 であるのに対し，‘w/o L^I_{mean}，L^I_{max}’ は GAP，GMP いずれのブランチにおいても ‘outlier’，‘clusters’ がそれぞれ 10，3 および 390，388 であり，少なくなっていることがわかる．‘outlier’ が少ないことは多くのサンプルが疑似ラベルを付与され，あるサンプル画像と同一人物の画像として疑

表 3.4: MSMT17[44] から Market1501[24] への教師なしドメイン適応において，最終エポックで生成された疑似ラベルの特性．表内において，‘clusters’ は疑似ラベルのクラスタ数を表し，‘outliers’ は疑似ラベルを付与されなかったサンプル数を表す．また ‘our model’ では GAP ブランチ（左）および GMP ブランチ（右）のそれぞれで算出された疑似ラベルの結果を報告した．

Models	Conditions	clusters	outliers
GAP only	w/ Adaptation	403	14
our model	w/o L^I_{mean}, L^I_{max}	390 / 388	10 / 3
	ours full model	428 / 425	7 / 7

似ラベルデータセット内に取り込まれたことを示している．また ‘clusters’ が少ないことは多くの人物画像が同一人物の画像として疑似ラベルデータセット内に取り込まれたことを示している．そこで ‘w/o L^I_{mean}，L^I_{max}’ において ‘outlier’，‘clusters’ が ‘GAP only’ のモデルより少ないことは提案モデルが Hard-Positive と共に多くの Hard-Negative を紐づけたことを示している．一方表 5.4 において，‘ours full model’ の行を確認すれば，表 5.4 における，‘w/o L^I_{mean}，L^I_{max}’，‘GAP only’，‘GMP only’ の全てと比較して，精度向上していると分かる．加えて表 3.4 において，‘ours full model’ の行における ‘clusters’ が大きく増加していることが確認される．そこで積集合による教示により学習過程において疑似ラベルデータセットにおけるクラス内サンプルの特徴表現の多様性を強化した結果，最終エポックにおける ‘clusters’ の増加に示されるように Hard-Negative に対する弁別性が獲得されたものと考えられる．したがって，積集合による教示により Hard-Negative に対する弁別性が得られた結果，表 5.4 において，‘ours full model’ の精度が大きく向上したと結論付けられる．ここで，2 章で述べたように，Market-1501 に含まれる人数は 751 人であり，表 3.4 におけるどの ‘clusters’ の値より大きいことを注意しておく．

続いて Market-1501 から MSMT17 への設定における結果について述べる．この設定では表 5.4 において，‘GAP only’ と ‘w/o L^I_{mean}，L^I_{max}’ の間にほとんど精度差は見られない．また ‘GMP only’ は両者に比べて約 2%精度が劣化していることが確認できる．これは，MSMT17 から Market-1501 と同様，多くの Hard-Negative が混在した結果，誤差拡大が発生したためと考えられる．

一方で，積集合を用いた教示により，‘w/o L^I_{mean}，L^I_{max}’，‘GAP only’，’GMP only’ いずれと比較しても約 3%と大幅に mAP が向上していることが表 5.4 より確認できる．Market-1501 へドメイン適応する場合と比べ，より大幅な精度向上が得られたことは，MSMT17 のような大規模データセットにおいて高精度を達成するには，Hard-Negative を弁別するだけでなく，Hard-Positive をより多く取り込むことが必要であることを示している．これはまた，提案手法のように大域的特徴と局所的特徴を同時に捉え，両者の違いを学習に応用することがこの目的達成に有効であることも示している．

以上より，GAP による大域的特徴と GMP による局所的特徴の両方を用いながら，その違いを効果的に利用することにより本論文の目的である，Hard-Positive を積極的に紐づけながらも，Hard-Negative を弁別することを達成出来たと言える．

3.3 まとめ

本章では大域的特徴および局所的特徴に加え，それら特徴間の違いを利用する新しい教師なしドメイン適応手法を提案した．特に提案手法では大域的・局所的特徴量から生成される二つの疑似ラベルセットの積集合を用いることにより，両特徴間の違いを活用した．積集合による教示により，より積極的に Hard-Positive を取り込みつつも，Hard-Negative を弁別するように学習することが可能であることを実験的に確かめた．また複数の公開データセットを用いた実験により，提案手法の有効性が定量的に示された．特に Market-1501 から MSMT17 へドメイン適応する設定では，2-stage の従来手法である MMT を 6.8%上回る mAP を達成した．

第4章

Prototypical Part Networkの進展

本章では従来の Prototypical Part Network （ProtoPNet）の派生手法について詳細に説明する．ProtoPNet 派生手法では学習データ内のある画像パッチに対応する特徴ベクトル（画像パッチ特徴）を代表するよう学習可能なパラメータである Prototype を学習する．その後，入力画像と Prototype との類似度を基に White-box な分類器を用いて推論を行う．これより，どのような学習データに含まれる特徴と同一の特徴を捉えたために推論クラスに分類されたかを提示可能となり，case-based な推論による解釈可能性が実現される．ProtoPNet は case-based な推論による解釈可能性と深層学習による特徴抽出を組み合わせることにより，推論過程を解釈可能性としながら高い精度を達成した．この利点のため ProtoPNet による ‘This looks like that’ フレームワークは幅広い注目を集めており，精度やメモリ効率を改善する手法が多く提案されている．また Prototype がどのような画像特徴を捉えているか [94] や，特徴マップ上の Prototype が入力画像上のどの画像領域に真に対応するかを推定する手法 [134] が提案されており，Prototype の解釈可能性についても改善がなされている．

他の ‘Gray-box’ なモデルを構築する手法と比較して，‘This looks like that’ フレームワークにはクラスラベル以上の教師ラベルを必要としない利点がある．そのため解釈可能性が重要な医療画像 [81][88][100][123] や自動運転 [119]，DeepFake の検出 [101]，また化合物生成を応用分野に持つグラフ分類 [115][127][128] およびグラフ回帰 [147] など幅広い分類タスクへの応用が研究されている．また，学習データ内のある（画像）領域との類似度により推論根拠の説明を行う ‘This looks like that’ フレームワークは解釈可能性を実現するフレームワークとして非常に汎用性が高い．そのため他の説明手法ではあまり検討されてこなかった時系列データ [54][95] や動画像分類 [135]，Semantic Segmentation[149]，強化学習 [122][139] や教師なしドメイン適応 [133]，また継続学習 [148] への応用に関しても研究が進められている（図 4.1）．加えて，僅かな人手による修正によりモデル性能を向上するモデルデバッグに関する手法 [130][143] についても提案がなされている．しかし，これらの ‘This looks like that’ フレームワークの応用に関する研究は学習データとテストデータのクラスが一致する ‘closed-set’ な問題設定に限定される．これは従来の ProtoPNet 派生手法が学習データとテストデータでクラスラベルの異なる ‘open-set’ な問題に有効な類似度学習に適用することが難しい課題を抱えているためである．

本章の目的は提案手法の基礎となる ProtoPNet およびその派生手法の進展を概説することにより，従来研究のコンテキストにおける本研究の意義を確認することにある．本章では ProtoPNet 派生手法を (1) クラス毎に固有の Prototype を学習する手法と (2) クラス間で共通した Prototype を学習する手法の二つのグループに分類し，各グループの手法をそれぞれ詳細に説明する．各グループの手法の説

明では特に従来の ‘This looks like that’ フレームワークは類似度学習に適用することが出来ない理由を詳細に説明する．類似度学習はサンプル間の類似度を基に推論を行う画像認識における基盤技術の一つであり，学習データとテストデータでクラスが異なる状況への適用等重要な応用分野を持つ．そのため，類似度学習に適用可能となる解釈可能なモデルの構築は実用上重要となり，ProtoPNet を類似度学習へ拡張する本研究の意義が確認される．

本章の構成は以下のとおりである．まず 4.1 章ではクラス毎に固有の Prototype を学習する手法について説明する．次に 4.2 章ではクラス間で共通した Prototype を学習する手法を説明し，最後に 4.3 章で本章のまとめを行う．本章では特に断りのない限り，バッチ内のデータインデックスの集合を $\mathbb{B}$，全てのクラスラベルの集合を $\mathbb{Y}$，全ての Prototype インデックスの集合を $\mathbb{P}$，クラス y に所属する Prototype インデックスの集合を $\mathbb{P}_y$ と記述する．また入力画像およびそのクラスラベルを x_i および y_i と表し，x_i をモデルバックボーンに入力した結果得られる特徴マップを $\mathbf{Z}_i$，画像 x_i と Prototype $\boldsymbol{p}_j$ との類似度を $s_{i,j}$ と表す．$s_{i,j}$ を列方向に並べて得られる列ベクトルを $\boldsymbol{s}_i$ と表す．

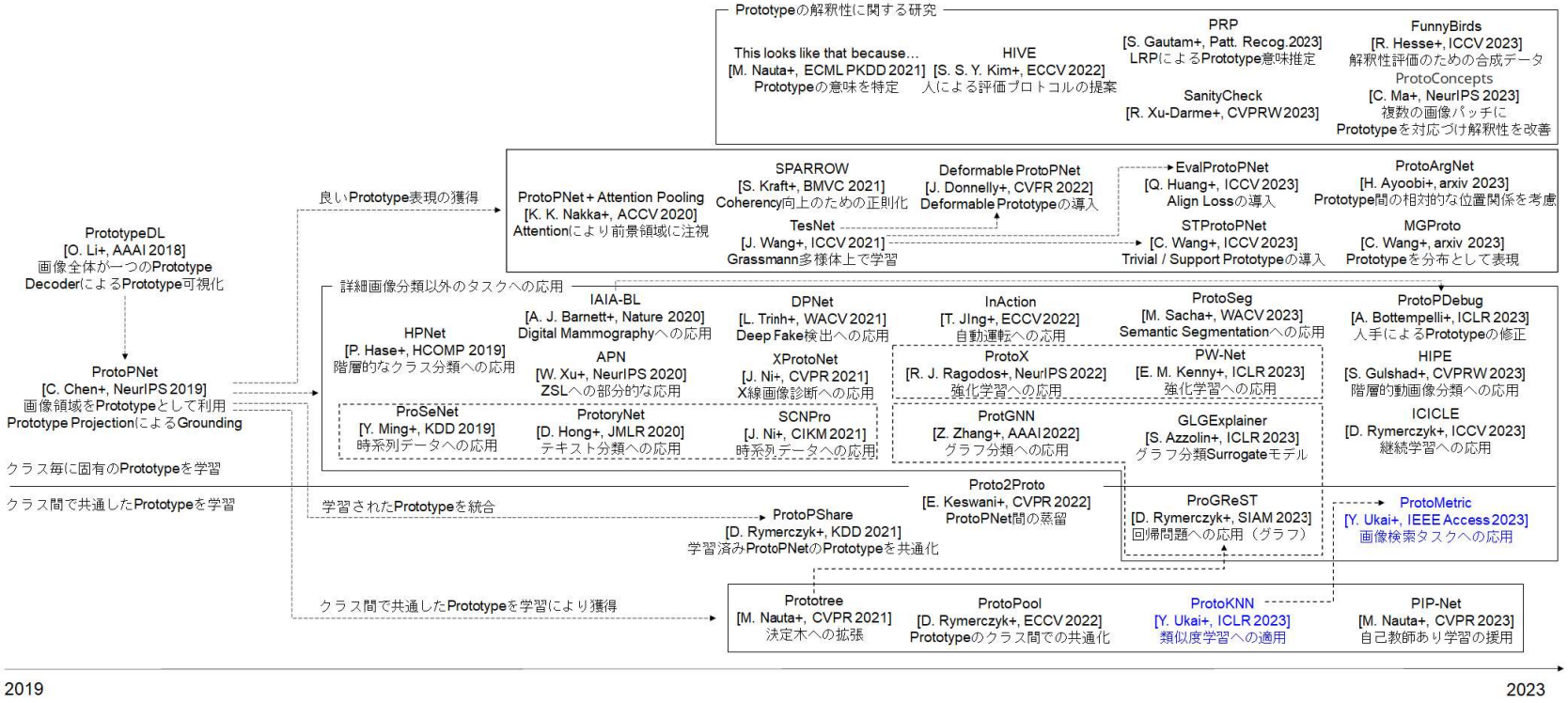

図 4.1: ProtoPNet 派生手法の推移．筆者らの研究を青字で示す．

4.1 クラス毎に固有の Prototype を学習する手法

本章では ProtoPNet と同様に各クラスに固有の Prototype を学習する手法について説明する．ProtoPNet は case-based な推論による解釈可能性を実現しつつ，深層学習の特徴抽出能力を活用することで高い精度を達成することに成功した．一方で，ProtoPNet によって学習された Prototype は重複や背景領域を含むなど冗長な表現となることが知られている．また，同一の Prototype が複数の意味の異なる画像領域に反応する等 Prototype の Purity (Consistency) に課題を抱えていることが報告されている．この課題に対処し，「良い」Prototype 表現を獲得するための様々な手法が提案されており，大幅な精度の改善が達成されてきた．以下では，まず ProtoPNet について説明した後，これらの良い Prototype 表現の獲得を目的に提案されてきた手法について詳細に説明する．

■ ProtoPNet

ProtoPNet[45] のモデル構造を図 4.2 に再謁する．ProtoPNet は各クラス毎に複数個の Prototype を定義し，各 Prototype がそれらの所属するクラスに特徴的な画像特徴を代表するよう学習する．この目的のため ProtoPNet の学習では Cross Entropy Loss L_{ce} に加え，Cluster Loss L_{clst} および Separation Loss L_{sep} と呼ばれる二つの損失関数を最小化する．ここで Cluster Loss は入力画像を変換することで得られる，特徴マップ内のあるピクセルに含まれる特徴量（画像パッチ特徴）と入力画像クラスに所属する Prototype とを近づける損失関数である．また，Separation Loss は Cluster Loss とは対照的に特徴マップ内の画像パッチ特徴と入力画像クラスに所属しない Prototype とを遠ざける損失関数である．各損失関数は下式のように定式化される．

$$L_{ce} = -\frac{1}{|\mathbb{B}|}\sum_{i\in\mathbb{B}} \log P(y_i|x_i) \text{ where } P(y_i|x_i) = \sigma\big(\boldsymbol{W}^T \boldsymbol{s}_i\big)_{y_i} \tag{4.1}$$

$$L_{clst} = -\frac{1}{|\mathbb{B}|}\sum_{i\in\mathbb{B}} \min_{j\in\mathbb{P}_{y_i}, \boldsymbol{z}\in\mathbf{Z}_i} \|\boldsymbol{z} - \boldsymbol{p}_j\|_2^2 \tag{4.2}$$

$$L_{sep} = \frac{1}{|\mathbb{B}|}\sum_{i\in\mathbb{B}} \min_{j\notin\mathbb{P}_{y_i}, \boldsymbol{z}\in\mathbf{Z}_i} \|\boldsymbol{z} - \boldsymbol{p}_j\|_2^2 \tag{4.3}$$

ただし，$\boldsymbol{W} \in \mathbb{R}^{|\mathbb{P}|\times|\mathbb{Y}|}$ はクラス分類層である線形分類器の重みであり $\sigma(\cdot)$ は Softmax 関数を表す．ProtoPNet は学習データ中のある画像パッチ特徴に接地された Prototype と入力画像の画像パッチ特徴との類似度をもとに推論を行うことで case-based な推論による解釈可能性を実現している．そのため ‘This looks like that’ フレームワークを実現するためには Prototype が学習データ内のある画像パッチを代表するように（ある画像パッチに接地されるように）学習する必要がある．そこで Prototype が学習データ内のある画像パッチを代表するよう課される Cluster Loss は ‘This looks like that’ フレームワークを実現するために ProtoPNet の学習において必要不可欠な損失関数と言える．

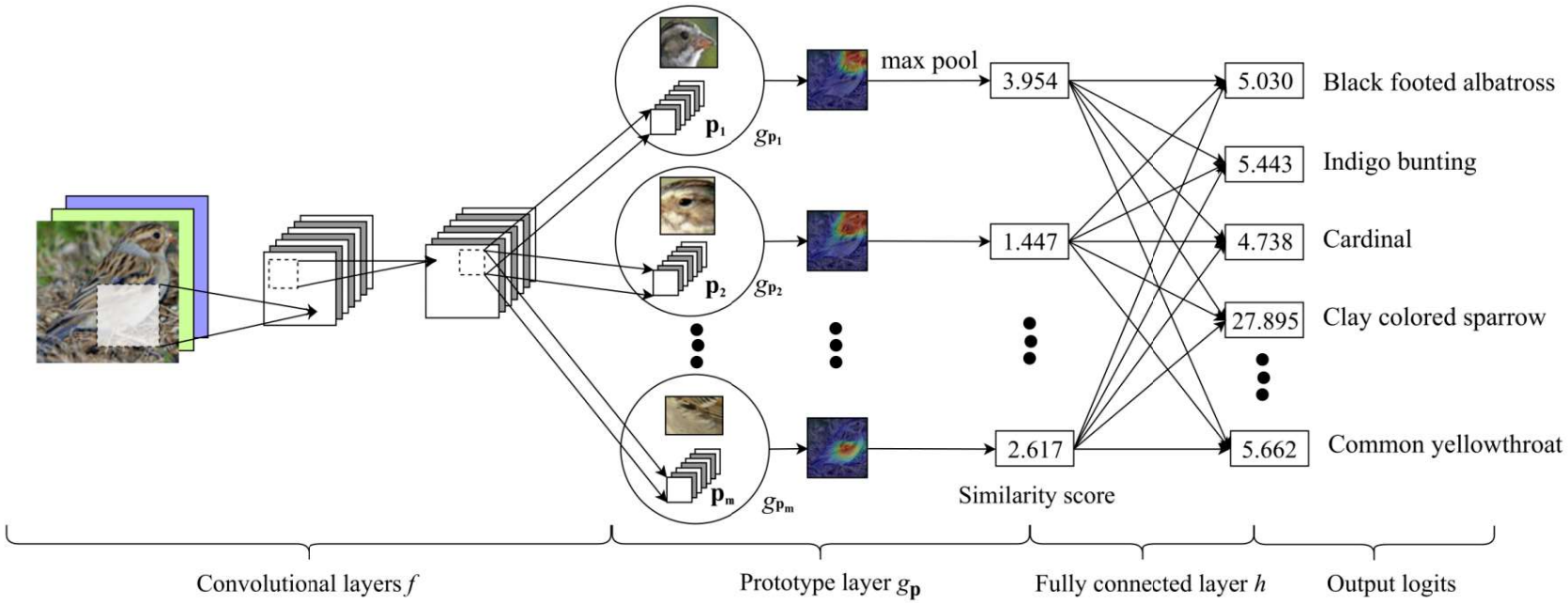

図 4.2: ProtoPNet のモデル構造．図は文献 [45] より引用．

ProtoPNet の学習ではまず Prototype Layer とモデルバックボーンを上記 3 つの損失関数を最小化するよう学習した後，線形層をファインチューニングする二段階の学習方式を採用する．ここで線形層のファインチューニングでは異なるクラスに属する Prototype が推論に寄与しないように L_{ce} に加え $L_{l1} = \sum_{y \in \mathbb{Y}} j \notin \mathbb{P}_y \|W_{y,j}\|$ が課される．また Prototype Layer とモデルバックボーンの学習時には線形分類器の重み W は Prototype の所属するクラスに対しては 1.0，その他のクラスに対しては -0.5 で固定される．すなわち，$\boldsymbol{W}$ はその (i, j) 成分を $W_{i,j}$ として

$$W_{j,y} = \begin{cases} 1.0 & j \in \mathbb{P}_y \\ -0.5 & \text{otherwise} \end{cases} \tag{4.4}$$

と初期化される．4.4 式のように初期化された線形重みを用いて ProtoPNet を学習することで，各 Prototype がそれらの所属するクラスを認識するために重要な特徴を代表するように学習出来る．また，Prototype およびモデルバックボーンの学習終了後には，Prototype が学習データ内のある画像パッチ内に接地されるように Prototype Projection が実行される．ここで Prototype Projection は Prototype を特徴空間上で学習データ内のもっとも類似する画像パッチ特徴と置き換える操作であり，下式のように表現される．

$$\boldsymbol{p}_j \leftarrow \underset{\boldsymbol{z} \in \mathbf{Z}_i, i \in \mathbb{D}}{\arg\max}\, \theta(\boldsymbol{z}, \boldsymbol{p}_j) \tag{4.5}$$

ただし，$\mathbb{D}$ は学習データセット内に含まれる全てのデータインデックスの集合である．また，$\theta(\boldsymbol{z}, \boldsymbol{p})$ は特徴ベクトル $\boldsymbol{z}$ と Prototype $\boldsymbol{p}$ の類似度であり，ProtoPNet では $\theta(\boldsymbol{z}, \boldsymbol{p}) = -\|\boldsymbol{z} - \boldsymbol{p}\|_2^2$ と定義される．

■ TesNet

図 4.3 に Tesnet[102] のモデル構造を示す．Tesnet では ProtoPNet とは異なり，クラス y に所属する K 個の Prototype の集合を Grassmann 多様体 $\mathrm{Gr}(K, V)$ 上の点 $\boldsymbol{B}_y$ として表現する．ここで，Grassmann 多様体 $\mathrm{Gr}(K, V)$ は n 次元ベクトル空間 V における K 次元部分ベクトル空間の集合として定義される可微分多様体である．すなわちクラス y に所属する Prototype のインデックスの集合を $\mathbb{P}_y$ とすれば $\boldsymbol{B}_y$ は $\boldsymbol{B}_y = \{\boldsymbol{p}_j\}_{j \in \mathbb{P}_y}$ と表現される．Tesnet では Grassmann 多様体上の点として Prototype の集合を表現するため，Prototype $\boldsymbol{p}_j$ と特徴ベクトル $\boldsymbol{z}$ の類似度 $\theta(\boldsymbol{z}, \boldsymbol{p}_j)$ および入力画像 x_i と Prototype $\boldsymbol{p}_j$ の類似度 $s_{i,j}$ を

$$s_{i,j} = \max_{\boldsymbol{z} \in \mathbf{Z}_i} \theta(\boldsymbol{z}, \boldsymbol{p}_j) = \max_{\boldsymbol{z} \in \mathbf{Z}_i} \boldsymbol{z} \cdot \boldsymbol{p}_j \tag{4.6}$$

により定義する．ただし $\mathbf{Z}_i$ は画像 x_i を Model Backbone に入力した結果得られる特徴マップである．また Tesnet では Cluster Loss L_{clst} および Separation Loss L_{sep} は下式のように書き直される．

$$L_{clst} = -\frac{1}{|\mathbb{B}|} \sum_{i \in \mathbb{B}} \min_{j \in \mathbb{P}_{y_i}, \boldsymbol{z} \in \mathbf{Z}_i} \frac{\boldsymbol{p}_j \cdot \boldsymbol{z}}{\|\boldsymbol{p}_j\|}, \quad L_{sep} = \frac{1}{|\mathbb{B}|} \sum_{i \in \mathbb{B}} \min_{j \notin \mathbb{P}_{y_i}, \boldsymbol{z} \in \mathbf{Z}_i} \frac{\boldsymbol{p}_j \cdot \boldsymbol{z}}{\|\boldsymbol{p}_j\|} \tag{4.7}$$

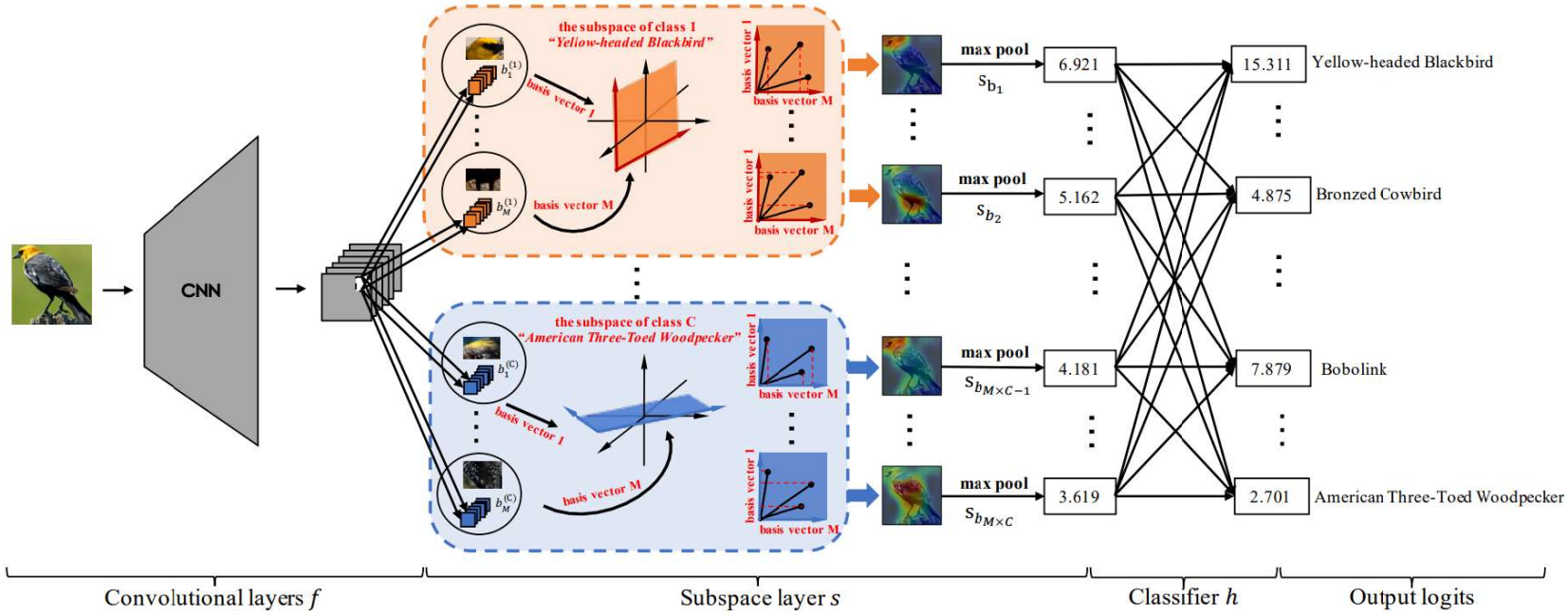

図 4.3: Tesnet のモデル構造．図は文献 [102] より引用．

さらに Tesnet では異なるクラスに属する Prototype が異なる画像特徴を表現するように，以下の損失関数を課すことで各クラスに所属する Prototype の集合を Grassmann 多様体上で遠ざける．

$$L_{ss} = -\sum_{y1,y2\in\mathbb{Y},y_1<y_2} \frac{1}{\sqrt{2}} \left\| \boldsymbol{B}_{y_1}^T \boldsymbol{B}_{y_1} - \boldsymbol{B}_{y_2}^T \boldsymbol{B}_{y_2} \right\|_F \tag{4.8}$$

加えて，同一クラスに所属する異なる Prototype が同一の特徴を表現することを避けるため，各 Prototype が $\boldsymbol{B}_y$ として定義される部分空間の異なる基底となるように以下の損失関数が課される．

$$L_{orth} = \sum_{y\in\mathbb{Y}} \left\| \boldsymbol{B}_y^T \boldsymbol{B}_y - \boldsymbol{I}_K \right\|_F \tag{4.9}$$

ただし，$\boldsymbol{I}_K$ は K 行 K 列の単位行列である．そこで，Tesnet の学習において最小化される全体の損失関数は下式により定義される．

$$L_{tesnet} = L_{ce} + \lambda_{clst} L_{clst} + \lambda_{sep} L_{sep} + \lambda_{ss} L_{ss} + \lambda_{orth} L_{orth} \tag{4.10}$$

ただし，$\lambda_*, (* \in \{clst, sep, ss, orth\})$ は各損失関数の重みを調整するハイパーパラメータである．

Tesnet においても ProtoPNet と同様に二段階の学習がなされる．すなわち，4.4 式により初期化されたクラス分類層の重み W のもとで，L_{tesnet} を最小化するよう Model Backbone および Prototype Layer を学習，Prototype Projection を行った後，L_{ce} および L_{l1} を最小化するよう線形分類層が学習される．以上の学習方法により，Tesnet は効果的に Prototype 間の冗長性を削減し ProtoPNet の性能を大幅に向上させることに成功した．

■ Deformable ProtoPNet

図 4.4 に Deformable ProtoPNet[116] のモデル構造を示す．Deformable ProtoPNet では固定サイズの画像領域を Prototype として扱う代わりに，可変サイズの画像領域を Prototype とする Deformable Prototype を導入することにより画像特徴の空間的な変形に対応する．ここで Deformable Prototype $\boldsymbol{P}^j$ は複数の特徴ベクトル $\{\boldsymbol{p}_{m,n}^j\}_{m,n=0,1,\ldots}$ を用いて構成される．入力画像 x_i をモデルバックボーンに入力することで得られる特徴マップ $\mathbf{Z}_i$ の位置 (h, w) における特徴ベクトルを $\boldsymbol{z}_{i,h,w}$ とすれば，Deformable Prototype $\boldsymbol{P}^j$ と入力画像 x_i の類似度 $s_{i,j}$ は下式により定義される．

$$s_{i,j} = \max_{\boldsymbol{z}_{i,h,w}\in\mathbf{Z}_i} \theta\left(\boldsymbol{P}^j, \boldsymbol{z}_{i,h,w}\right) \equiv \max_{\boldsymbol{z}_{i,h,w}\in\mathbf{Z}_i} \left(\frac{1}{|\boldsymbol{P}^j|} \sum_m \sum_n \frac{\boldsymbol{p}_{m,n}^j \cdot \boldsymbol{z}_{i,h+m+\Delta_{h,m},w+b+\Delta_{w,n}}}{\|\boldsymbol{p}_{m,n}^j\|_2 \|\boldsymbol{z}_{i,h+m+\Delta_{h,m},w+b+\Delta_{w,n}}\|_2} \right) \tag{4.11}$$

ただし，$|\boldsymbol{P}^j|$ は $\boldsymbol{P}^j$ に含まれる Prototype 数である．また $\Delta_{h,m}$ および $\Delta_{w,n}$ はモデルバックボーンの後に追加される畳み込み層 δ により算出されるオフセット値である．ここで $\Delta_{h,m}$ および $\Delta_{w,n}$ は特徴マップ $\mathbf{Z}_i$ および位置 h, w, m, n にのみ依存し，Prototype $\boldsymbol{p}_{m,n}^j$ には依存しないことに注意されたい．

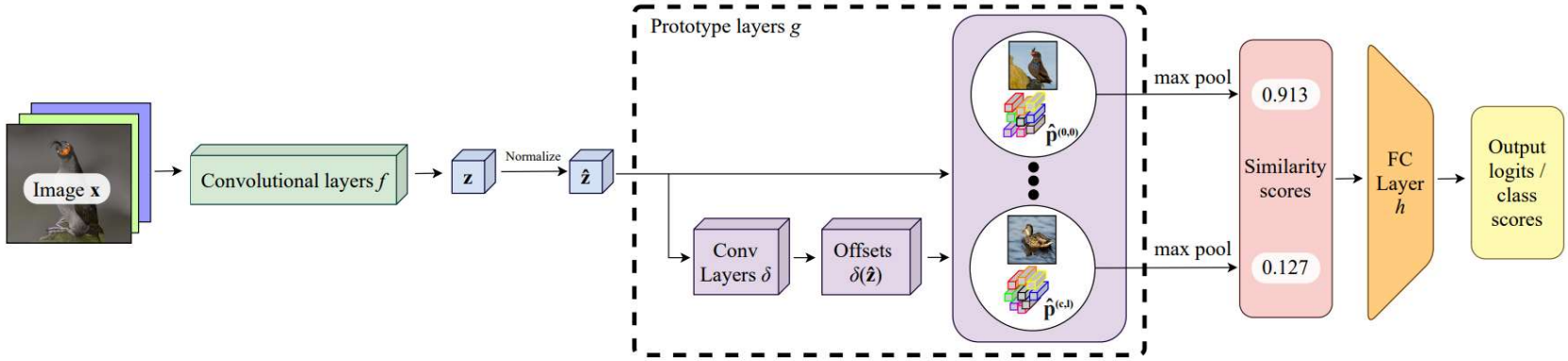

図 4.4: Deformable ProtoPNet のモデル構造．図は文献 [116] より引用．

Deformable ProtoPNet では Cross Entropy Loss L_{ce} は，4.11 式で定義された $\theta(\boldsymbol{P}^j, \boldsymbol{z}_{i,h,w})$ を用いて下式のように書き直される．

$$L_{ce} = -\sum_{i\in\mathbb{B}} \log\sigma(\boldsymbol{W}^T \boldsymbol{s}_i^{(-)})_{y_i},\ \text{where}\ \boldsymbol{s}_i^- = \begin{cases} \max_{\boldsymbol{z}_{i,h,w}\in\mathbf{Z}_i} \theta\left(\boldsymbol{P}^j, \boldsymbol{z}_{i,h,w}\right) & \text{if } j \in \mathbb{P}_{y_i} \\ \max_{\boldsymbol{z}_{i,h,w}\in\mathbf{Z}_i} \left[\theta\left(\boldsymbol{P}^j, \boldsymbol{z}_{i,h,w}\right) - \phi\right]_+ & \text{otherwise} \end{cases} \tag{4.12}$$

ただし，$[\cdot]_+$ は ReLU 関数であり，ϕ は定数値をとるマージンである．また，Cluster Loss L_{clst} および Separation Loss L_{sep} は $s_{i,j}$ を用いて下式のように書き直される．

$$L_{clst} = -\frac{1}{|\mathbb{B}|}\sum_{i\in\mathbb{B}} \max_{j\in\mathbb{P}_{y_i}} s_{i,j}, \quad L_{sep} = \frac{1}{|\mathbb{B}|}\sum_{i\in\mathbb{B}} \max_{j\notin\mathbb{P}_{y_i}} s_{i,j} \tag{4.13}$$

さらに Deformable ProtoPNet ではクラス y に属する Deformable Prototype を構成する全ての Prototype $\{\boldsymbol{p}^j_{m,n} | j \in \mathbb{P}_y\}$ が異なる画像特徴を表現するよう以下の損失関数を課す．

$$L_{ortho} = \sum_{y\in\mathbb{Y}} \left\| \boldsymbol{M}^{(y)} \boldsymbol{M}^{(y)^T} - \boldsymbol{I}_M \right\|_F \tag{4.14}$$

ここで，$\boldsymbol{M}^{(y)} \in \mathbb{R}^{M\times C}$ は $\{\boldsymbol{p}^j_{m,n} | j \in \mathbb{P}_y\}$ を行方向に並べた結果得られる行列であり，$\boldsymbol{I}_M$ は M 行 M 列の単位ベクトルである．ただし M は $\{\boldsymbol{p}^j_{m,n} | j \in \mathbb{P}_y\}$ の個数であり，C は Prototype $\boldsymbol{p}^j_{m,n}$ の次元数である．

Deformable ProtoPNet は以上 4 つの損失関数を最小化するようにモデルバックボーンおよび Prototype を学習し，Prototype Projection を行った後，L_{ce} を最小化するよう線形層を最適化する二段階の学習戦略を採用する．ここで線形層の最適化では ProtoPNet と同様に L_{ce} に加え $L_{l1} = \sum_{y\in\mathbb{Y}} j \notin \mathbb{P}_y \|W_{y,j}\|$ が課される．Deformable ProtoPNet は画像特徴の空間的な変形に対応することで，特に前景領域を囲む矩形で画像を切り抜かない実験設定において高い精度を達成した．一方で，複数の特徴ベクトルにより Prototype を構成し，また高解像度の特徴マップを必要とするため他の ProtoPNet 派生手法と比較し多くのメモリが必要となる課題を抱えている．

■ ST-ProtoPNet

Wang らは ProtoPNet の学習により獲得される Prototype は特徴空間上でクラス中心に近い Trivial な特徴表現となっていることを発見した [151]．一方で Support Vector Machine (SVM) に代表されるように，クラス分類には分類境界付近の特徴が有効となる [70]．そこで Wang らは Trivial な特徴に加え，分類境界付近の特徴（Support）の両方を Prototype の表現として獲得する ST-ProtoPNet を提案した．図 4.5 に示すように ST-ProtoPNet は Trivial 特徴を学習するブランチと Support 特徴を学習するブランチをもつ Siamise 構造をとる．ST-ProtoPNet では Prototype $\boldsymbol{p}_j$ と特徴ベクトル $\boldsymbol{z}$ 間の類似度 $\theta(\boldsymbol{z}, \boldsymbol{p}_j)$ を $\theta(\boldsymbol{z}, \boldsymbol{p}_j) = \boldsymbol{z} \cdot \boldsymbol{p}_j$ と定義する．そこで ST-ProtoPNet では Cluster Loss L_{clst} および Separation Loss L_{sep} は下式のように書き直される．

$$L_{clst} = -\max_{j\in\mathbb{P}^*_y} \max_{\boldsymbol{z}\in\mathbf{Z}^*_i} \boldsymbol{z} \cdot \boldsymbol{p}_j, \quad L_{sep} = \max_{j\notin\mathbb{P}^*_y} \max_{\boldsymbol{z}\in\mathbf{Z}^*_i} \boldsymbol{z} \cdot \boldsymbol{p}_j \tag{4.15}$$

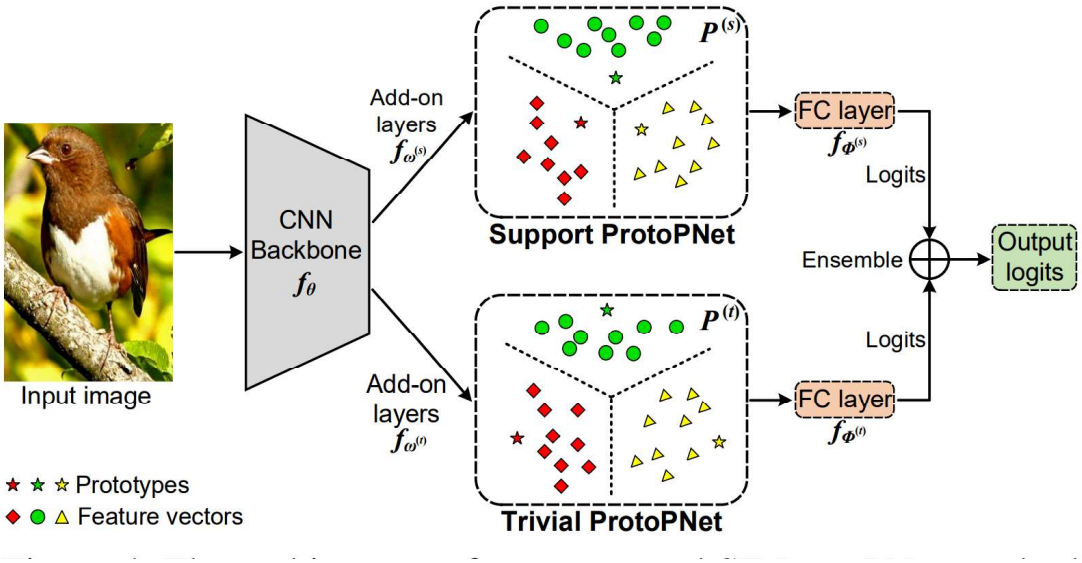

図 4.5: ST-ProtoPNet のモデル構造．図は文献 [151] より引用．

ここで，$\mathbb{P}_y^*(* \in \{s,t\})$ は Support，Trivial ブランチにおけるクラス y に所属する Prototype のインデックスの集合であり，$\mathbf{Z}_i^*(* \in \{s,t\})$ は x_i を入力したとき Support，Trivial ブランチより出力される特徴マップである．また Tesnet と同様にクラス内の Prototype が同一の特徴を学習しないようにするため下式で定義される損失関数が課される．

$$L_{orth} = \sum_{y \in \mathbb{Y}} \|\boldsymbol{P}_y \boldsymbol{P}_y^T - \boldsymbol{I}_K\|_F \tag{4.16}$$

ただし，$\boldsymbol{P}_y \in \mathbb{R}^{K \times C}$ はクラス y に属する K 個の Prototype $\boldsymbol{p}_j \in \mathbb{R}^C$ を行方向に結合して得られる行列であり，$\boldsymbol{I}_K$ は K 行 K 列の単位行列である．

ST-ProtoPNet では以上の損失関数に加え，Support ブランチに属する Prototype が Support 特徴を，Trivial ブランチに属する Prototype が Trivial 特徴を学習するようにするため，各ブランチに Closeness Loss L_{cls} および Discrimination Loss L_{dsc} を課す．ここで各損失関数は下式のように定式化される．

$$L_{cls} = -\sum_{y_1,y_2 \in \mathbb{Y}, y_1 < y_2} \min_{j \in \mathbb{P}_{y_1}, k \in \mathbb{P}_{y_2}} \boldsymbol{p}_j \cdot \boldsymbol{p}_k, \quad L_{dsc} = \sum_{y_1,y_2 \in \mathbb{Y}, y_1 < y_2} \max_{j \in \mathbb{P}_{y_1}, k \in \mathbb{P}_{y_2}} \boldsymbol{p}_j \cdot \boldsymbol{p}_k \tag{4.17}$$

各損失関数の式形より，Closeness Loss は異なるクラスに属する Prototype のうち最も小さな類似度を持つ Prototype 間の類似度を増大する損失関数となっており，Discrimination Loss は異なるクラスに属する Prototype のうち最も大きな類似度を持つ Prototype 間の類似度を低減する損失関数となっているとわかる．そこで L_{cls} および L_{dsc} により，各ブランチの Prototype が分類境界およびクラス中心近傍の特徴を代表するよう学習されると分かる．以上まとめると，ST-ProtoPNet では Support,

Trivial ブランチそれぞれに下式で定義される損失 L_s および L_t を与えて学習を行う．

$$\begin{aligned} L_s &= L_{ce}^s + \lambda_{clst}L_{clst} + \lambda_{sep}L_{sep} + \lambda_{cls}L_{cls} + \lambda_{orth}L_{orth} \\ L_t &= L_{ce}^t + \lambda_{clst}L_{clst} + \lambda_{sep}L_{sep} + \lambda_{dsc}L_{dsc} + \lambda_{orth}L_{orth} \end{aligned} \tag{4.18}$$

ここで，$L_{ce}^*(* \in \{s,t\})$ は

$$L_{ce}^* = -\sum_{i\in\mathbb{B}} \log\sigma\Big({\boldsymbol{W}_*}^T \boldsymbol{s}_i^*\Big)_{y_i},\ \text{where}\ s_{i,j}^* = \max_{\boldsymbol{z}\in\mathbf{Z}_i^*} \boldsymbol{z}\cdot\boldsymbol{p}_j \tag{4.19}$$

と定義される．ただし，$\boldsymbol{W}_*(* \in \{s,t\})$ は Support，Trivial ブランチにおける線形層の重みである．ST-ProtoPNet においても 4.18 式を用いてモデルバックボーンおよび Prototype の学習および Prototype Projection を行った後，L_{ce} を最小化するように線形層を最適化する二段階の学習戦略を採用する．以上の学習手法により，ST-ProtoPNet はクラス中心に近い Trivial な特徴に加えて分類に重要な Support 特徴を Prototype の特徴表現として獲得することに成功した．加えて，ST-ProtoPNet は Support および Trivial な Prototype の両方を用いることにより性能改善が達成出来る事を実験的に証明した．

■ EvalProtoPNet

先述したように ProtoPNet で学習された Prototype は異なる画像内で意味的に異なる部位に反応する場合がある．また同一の画像内であっても僅かに加えられた画像ノイズにより，Prototype の反応位置が変化することが知られている．これらの課題に対処するため Huang らは Shallow-Deep Feature Alignment (SDFA) モジュールを導入した EvalProtoPNet を提案した [137]．図 4.6 に EvalProtoPNet のモデル構造を示す．EvalProtoPNet では ST-ProtoPNet と同様に特徴ベクトル $\boldsymbol{z}$ と Prototype $\boldsymbol{p}_j$ 間の類似度が定義される．そのため，Cluster Loss，Separation Loss，および Orthogonal Loss は ST-ProtoPNet と同様に定義される（4.15 式および 4.16 式を参照されたい）．以下では EvalProtoPNet で新たに導入された Score Aggregation (SA) Module によるクラスロジットの算出と SDFA Module における Align Loss L_{align} について説明する．

SA Module　SA Module では線形分類層の重み $\boldsymbol{W}_{i,j}$ を 4.4 式により初期化および固定する代わりに，

$$W_{j,y} = \begin{cases} \frac{\exp(w_j)}{\sum_{k\in\mathbb{P}_y}\exp(w_k)} & \text{if } j \in \mathbb{P}_y \\ 0 & \text{otherwise} \end{cases} \tag{4.20}$$

と定義する．ただし $\boldsymbol{w} \in \mathbb{R}^{|\mathbb{P}|}$ は学習パラメータである．これより EvalProtoPNet は Prototype $\boldsymbol{p}_j, (j \in \mathbb{P}_y)$ のクラス y に対する最終的な線形分類層の重み $W_{j,y}$ が負の値を取りえる従来手法の課題に対処した．また SA Module により，クラスに属する Prototype のみがクラスロジットに正の寄与を与えることが保証されるため，EvalProtoPNet では線形分類層，モデルバックボーンおよび Prototype を同時に学習することが可能となる．

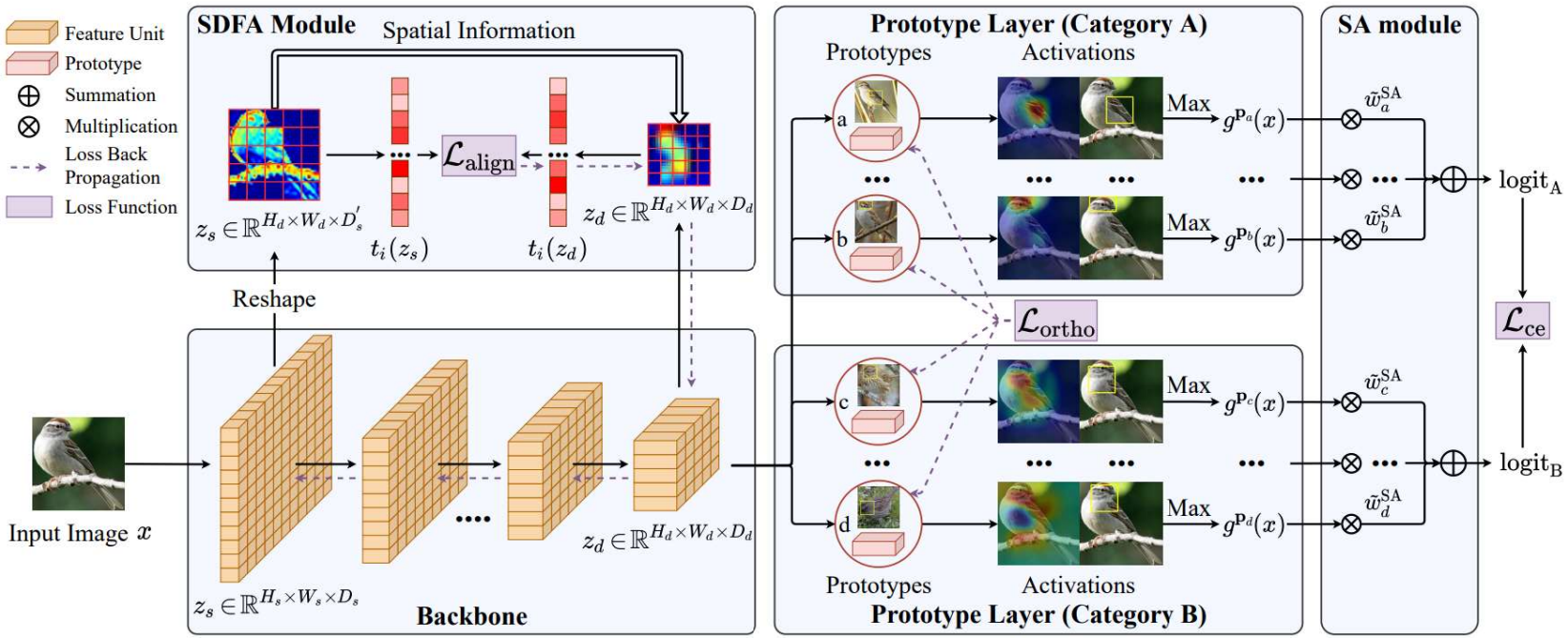

図 4.6: EvalProtoPNet のモデル構造．図は文献 [137] より引用．

SDFA Module　SDFA モジュールでは浅い層における特徴マップ出力 $\mathbf{Z}_i^S$ および深い層における特徴マップ出力 $\mathbf{Z}_i^D$ を入力とし，各特徴マップにおける空間的な構造が類似するように Align Loss L_{align} を課す．ここで L_{align} は下式のように表される．

$$L_{align} = \frac{1}{|\mathbb{M}|^2} \sum_{(h,w)\in\mathbb{M}} \sum_{(h',w')\in\mathbb{M}} \left[\left| \frac{\boldsymbol{z}_{i,h,w}^D \cdot \boldsymbol{z}_{i,h',w'}^D}{\|\boldsymbol{z}_{i,h,w}^D\|_2 \|\boldsymbol{z}_{i,h',w'}^D\|_2} - \mathrm{sg}\left(\frac{\boldsymbol{z}_{i,h,w}^S \cdot \boldsymbol{z}_{i,h',w'}^S}{\|\boldsymbol{z}_{i,h,w}^S\|_2 \|\boldsymbol{z}_{i,h',w'}^S\|_2} \right) \right| - \gamma \right]_+ \tag{4.21}$$

ただし，$\mathbb{M}$ は特徴マップにおける (h, w) で表される全てのピクセルインデックスの集合であり，$\boldsymbol{z}_{i,h,w}^S$ および $\boldsymbol{z}_{i,h,w}^D$ は浅い層における特徴マップ出力 $\mathbf{Z}_i^S$ および深い層における特徴マップ出力 $\mathbf{Z}_i^D$ の位置 (h, w) における特徴ベクトルである．また，$\mathrm{sg}(\cdot)$ は括弧内の変数に対して勾配を計算しない事を示す．

Huang らは深い層における特徴マップでは位置的な情報が失われる一方で浅い層では位置的情報が保たれることを観察により発見し，そのため浅い層と深い層の空間的な構造が同一となるよう正則化をかけることにより特徴抽出が改善することを主張する．以下に SDFA Module の効果に関する筆者の考察を説明する．一般に ProtoPNet 派生手法においてモデルバックボーンとして用いられる Resnet，VGG，DenseNet では ReLU 関数が用いられるため，モデルバックボーンにより表現される関数は Lipchitz 連続となる．加えてモデルバックボーンにおける Lipchitz 定数はモデルバックボーン内部の畳み込み層や線形層のため，層を重ねるごとに大きなものとなる．そのため，L_{align} により深い層における特徴マップの空間的構造を浅い層と類似させることはモデルバックボーンにおける Lipshitz 定数の増加を抑制することに繋がり，そのためノイズに対する頑健性の向上につながっていることが推測される．また Neural Network では浅い層では色やエッジ等の低次の特徴が表現される一方，深い層では頭などの Semantic な特徴表現が獲得されることが実験的に知られている [28][131]. そのため，4.21 式による二次統計量に関する正則化は，低次の特徴にサポートされない高次の特徴表現を抑制する効果を持つことが期待される．すなわち，Align Loss により異なる低次の特徴を持つ頭や胸などが深い層において同一の特徴となることを防ぐことが可能となると期待される．いずれにせよ，Huang らにより SDFA Module が Prototype が意味的に同一の部位に反応し，ノイズに対しても頑健となることが実験的に確認されており，SDFA Module により ProtoPNet の学習により獲得される Prototype の解釈性を向上できることが知見として得られている．

以上見たようにクラス毎に固有の Prototype を用いる手法では，学習の結果獲得される Prototype 表現の課題を解決することにより精度向上を達成してきた．これらの手法では事前に定義された Prototype とクラスラベルの関係性を効果的に利用することにより，各クラスに固有の特徴を Prototype が表現するように学習を行っている．実際，これらの手法では事前に定義された Prototype とクラスラベルの関係性に基づき線形分類層を初期化し学習を行う．また，事前に定義された Prototype とクラスラベルの関係性は Prototype と画像パッチ特徴を近づけるため課される Cluster Loss の算出にも利用される．一方で事前に定義された Prototype とクラスラベルの関係性を学習に必要とするため，推論時には学習時に存在しないクラスラベルを持つサンプルを対象とする類似度学習にこれらの手法を採用する事は難しい．

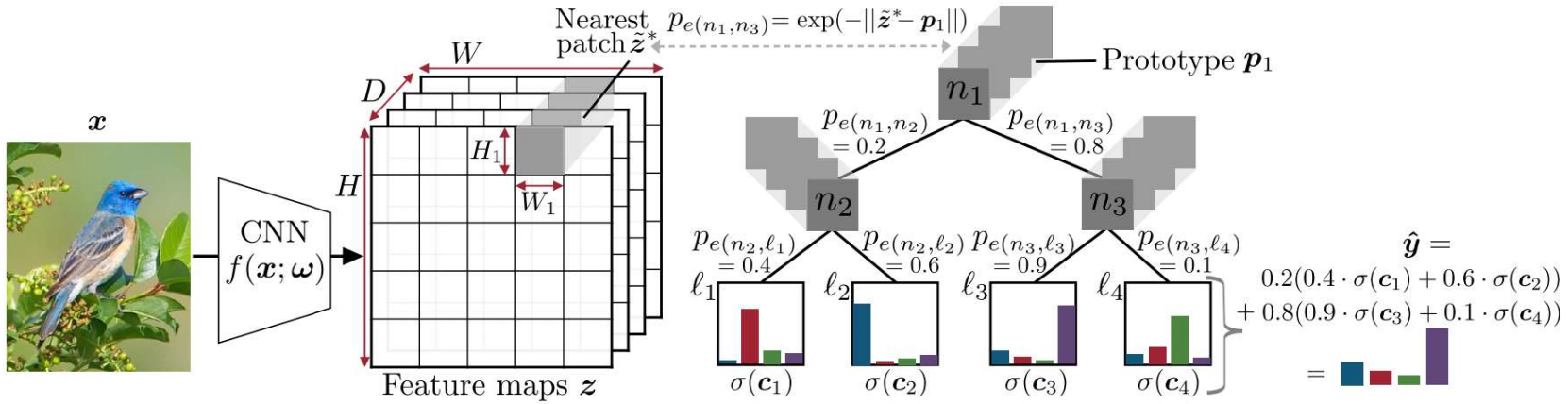

図 4.7: ProtoTree のモデル構造．図は文献 [93] より引用．

4.2 クラス間で共通したPrototypeを用いる手法

本章ではクラス間で共通する Prototype を学習によって獲得する ProtoPNet 派生手法について説明する．これらの手法はクラスに特有の Prototype を学習する手法と比較し Prototype 数を 10%程度に削減するなど，メモリ効率を大幅に改善出来る．各手法はそれぞれ異なる学習フレームワークを用いることで，クラス間で共通する Prototype を学習により獲得する目的を効果的に達成してきた．以下では，各手法の学習フレームワークを詳細に説明する．

■ Prototree

Prototree[93] は Deep Neural Decision Forests (dNDF) [14] と ‘This looks like that’ フレームワークを組み合わせた手法であり，線形分類器の代わりに決定木を後段の分類器として採用する．これより Prototree はクラス間で共通した Prototype を用いて ‘This looks like that’ フレームワークを構築することに初めて成功した．図 4.7 に示すように，Prototree では決定木の各ノードに割り当てられた Prototype と入力画像の類似度を用いて決定木の分岐を決定する．すなわち，入力画像 x_i が葉ノード l に到達する確率 $\mu_l(x_i)$ は，

$$\mu_l(x_i) = \prod_{n \in \mathbb{N}_l} s_{i,n}^{\mathbf{1}(l \swarrow n)} (1 - s_{i,n})^{\mathbf{1}(n \searrow l)} \tag{4.22}$$

と表される．ここで $\mathbf{1}(\cdot)$ は括弧内の中身が真であるとき 1，偽となるとき 0 を取る指示関数であり，$l \swarrow n$（$n \searrow l$）は葉ノード l がノード n で分割される部分木の左（右）側に存在する時真値をとる．また $\mathbb{N}_l$ は根より葉ノード l への経路上に存在する全てのノードインデックスの集合であり，$s_{i,n}$ は下式によって定義される入力画像 x_i と n 番目のノードに割り当てられた Prototype $\boldsymbol{p}_n$ との類似度である：

$$s_{i,n} = \max_{\boldsymbol{z} \in \mathbf{Z}_i} \exp\left(-\|\boldsymbol{z} - \boldsymbol{p}_n\|_2^2\right) \tag{4.23}$$

また，Prototree により入力画像 x_i がクラス y に分類される確率 $P(y|x_i)$ は l 番目の葉ノードに割り当てられたクラス y に対するクラス確率を $\pi_{l,y}$ として

$$P(y|x_i) = \sum_{l \in \mathbb{L}} \mu_l(x_i) \pi_{l,y} \tag{4.24}$$

と表される．ただし，$\mathbb{L}$ は全ての葉ノードインデックスの集合である．

Prototree の学習は dNDF の学習手法をそのまま引き継いだ学習手法を採用する．すなわち各葉ノードへの入力 x_i の到達確率 $\mu_l(x_i)$ を最適化するステップと葉ノードに割り当てられるクラス確率 $\boldsymbol{\pi}_l$ を最適化するステップの二つのステップを各イタレーション毎に交互に繰り返す EM アルゴリズムにより最適化がなされる．以下では各ステップにおける最適化について説明する．

$\mu_l(x_i)$ を最適化するステップでは Cross Entropy Loss を用いて Prototype およびモデルパラメータが誤差逆伝搬法により更新される．すなわちモデルパラメータ θ および Prototype $\{p_i\}_{i=0,1,...}$ は，t

ステップ目における両パラメータの集合を ω^t と表せば ω^t は

$$\omega^{t+1} \leftarrow \omega^t - \eta \frac{\partial}{\partial \omega^t} \left(- \sum_{i \in \mathbb{B}} \log P(y_i|x_i) \right) \tag{4.25}$$

によって更新されることとなる．ただし，η は学習率である．

$\boldsymbol{\pi}_l$ を最適化するステップでは勾配降下法による更新ではなく繰り返しアルゴリズムの 1 ステップとして $\boldsymbol{\pi}_l$ を更新する．すなわち t ステップ目における $\boldsymbol{\pi}_l$ の第 y 成分を $\pi^t_{l,y}$ と表せば，$\pi^t_{l,y}$ を

$$\pi^{t+1}_{l,y} = \frac{1}{Z^{t+1}_l} \sum_{i \in \mathbb{B}} \frac{\mathbf{1}(y = y_i) \pi^t_{l,y} \mu_l(x_i)}{\sum_{m \in \mathbb{L}} \pi^t_{m,y} \mu_m(x_i)} \tag{4.26}$$

と更新する．この更新により必ず Cross Entropy Loss $\sum_{i \in \mathbb{B}} - \log P(y_i|x_i)$ が減少すると証明できる（証明は付録を参照されたい）．ただし，Z^{t+1}_l は $\sum_y \pi^{t+1}_{l,y} = 1$ となるよう課される正規化定数である．

dNDF の学習では各ノードに対する出力，すなわち Prototree における 4.23 式の値は学習とともに 0 または 1 のどちらかの値を取るようになる事が実験的に確認されている．従って，Prototree の学習フレームワークでは Cluster Loss を用いることなく Prototype と画像パッチ特徴とを近づける事が可能となる．そのため，Prototree の学習では Prototype とクラスラベルの関係性を事前に定義することなく学習を行うことが可能となる．しかし，以上見たように Prototree の学習フレームワークは 4.26 式による最適化など決定木を後段の分類器を用いることを前提として構築されている．そのため，Prototree の学習フレームワークを類似度学習に適用することは難しい．また Prototree の推論過程にはある Prototype を含まないという推論が含まれるため直感的に理解しにくい説明を生成する可能性が示唆されている．

■ ProtoPool

図 4.8 に ProtoPool のモデル構造を示す．ProtoPool は Prototype とクラスラベルの関係性を定義する代わりにクラス毎に K 個用意された slot への Prototype の割り当て方を学習する．ここで，クラス y に所属する k 番目のスロットに Prototype p_j が割り当てられる確率 $P(p_j|y,k)$ は，学習可能なパラメータ $q \in \mathbb{R}^{|\mathbb{Y}| \times |\mathbb{P}| \times K}$ を用いて下式によって定義される．

$$P(p_j|y,k) = \frac{\exp\left(\frac{q_{y,j,k}}{\tau}\right)}{\sum_{l<K} \exp\left(\frac{q_{y,k,l}}{\tau}\right)} \tag{4.27}$$

ただし，τ は温度パラメータである．また，$P(p_j|y,k)$ は学習中には下式のように標準 Gumble 分布に従う確率変数を加えた結果得られる値が採用される．

$$P(p_j|y,k) = \frac{\exp\left(\frac{q_{y,j,k}+\gamma_k}{\tau}\right)}{\sum_{l<K} \exp\left(\frac{q_{y,k,l}+\gamma_l}{\tau}\right)} \tag{4.28}$$

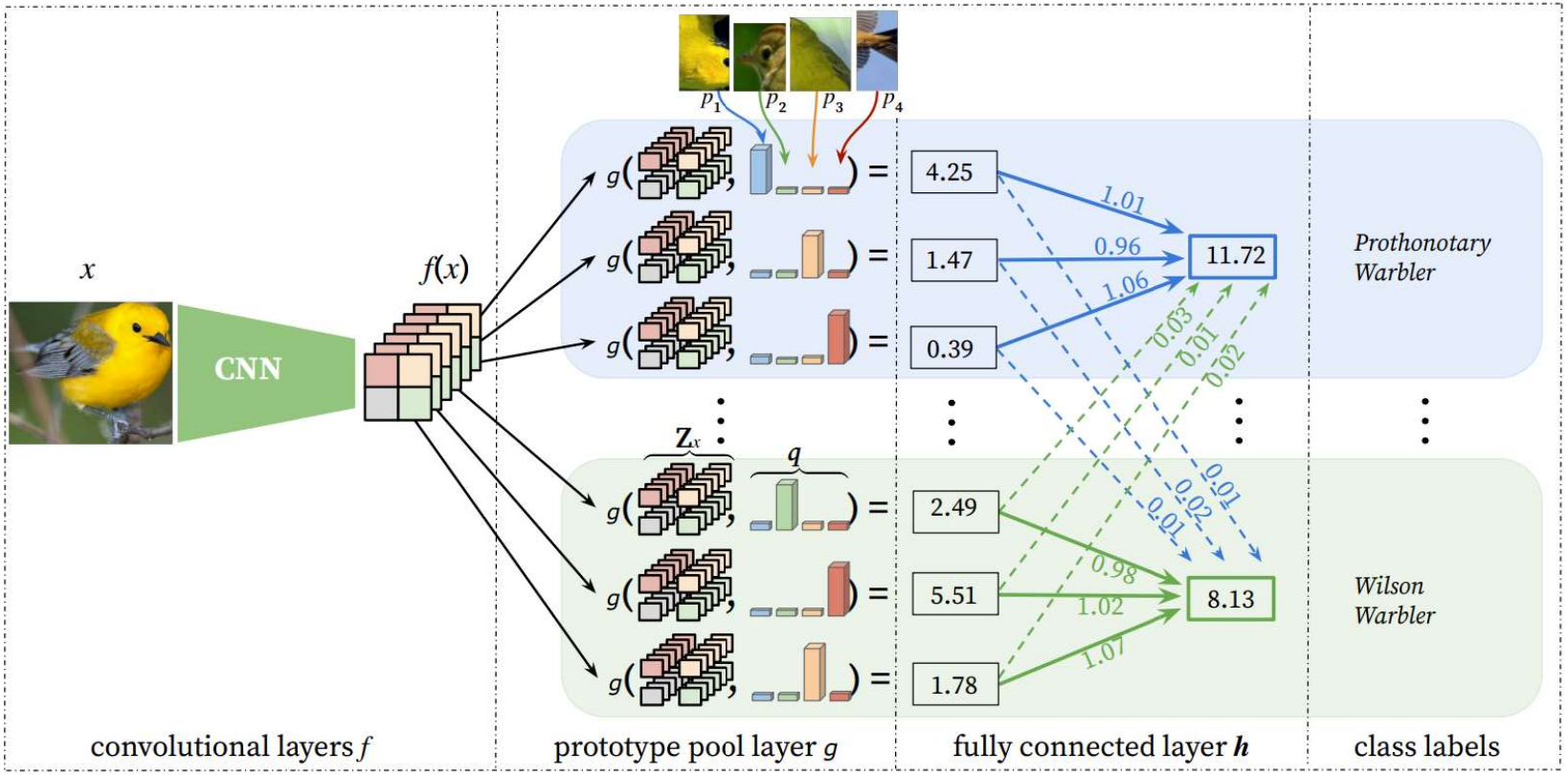

図 4.8: ProtoPool のモデル構造．図は文献 [124] より引用．

ここで $\{\gamma_l\}_{l=0,1,\ldots}$ は標準 Gumbel 分布に従う確率変数である．ProtoPool では 4.27 式あるいは 4.28 式で定義される Prototype の Assignment に基づきクラスラベルが予測される．すなわち，入力画像 x_i がクラス y に分類される確率 $P(y|x_i)$ は $\mathbb{K}_y$ をクラス y に所属するスロットのインデックスの集合として，

$$P(y|x_i) = \frac{\exp(c_y)}{\sum_{y' \in \mathbb{Y}} \exp(c_{y'})}, \quad \text{where} \quad c_y = \sum_{y' \in \mathbb{Y}} \sum_{k \in \mathbb{K}_{y'}} W_{k,y} \sum_{j \in \mathbb{P}} q_{y',j,k} s_{i,j} \tag{4.29}$$

となる．ただし，$W_{k,y}$ は下式によって初期化される線形分類層における重みである．

$$W_{k,y} = \begin{cases} 1.0 & \text{if } k \in \mathbb{K}_y \\ 0.0 & \text{otherwise} \end{cases} \tag{4.30}$$

そこで，ProtoPool ではクラス分類損失は $P(y_i|x_i)$ を用いて 4.1 式と同様に定義される．また，ProtoPool では $q_{y,j,k}$ を用いて ProtoPNet と同様に Cluster Loss L_{clst} および Separation Loss L_{sep} が定義される．すなわち，ProtoPool では $\mathbb{P}_y = \{j | j \in \arg\operatorname{topk}_l(\sum_k q_{y,l,k})\}$ として，

$$L_{clst} = -\frac{1}{|\mathbb{B}|} \sum_{i \in \mathbb{B}} \min_{j \in \mathbb{P}_{y_i}} \left(\min_{\boldsymbol{z} \in \mathbf{Z}_i} \|\boldsymbol{z} - \boldsymbol{p}_j\|_2^2 - \frac{1}{|\mathbf{Z}_i|} \sum_{\boldsymbol{z} \in \mathbf{Z}_i} \|\boldsymbol{z} - \boldsymbol{p}_j\|_2^2 \right) \tag{4.31}$$

$$L_{sep} = \frac{1}{|\mathbb{B}|} \sum_{i \in \mathbb{B}} \min_{j \notin \mathbb{P}_{y_i}} \left(\min_{\boldsymbol{z} \in \mathbf{Z}_i} \|\boldsymbol{z} - \boldsymbol{p}_j\|_2^2 - \frac{1}{|\mathbf{Z}_i|} \sum_{\boldsymbol{z} \in \mathbf{Z}_i} \|\boldsymbol{z} - \boldsymbol{p}_j\|_2^2 \right) \tag{4.32}$$

と定義される．ただし，$\arg\operatorname{topk}_l$ は l について括弧内の値のうち上位 k 個の値をとるインデックスの集合と定義した．また括弧内において，Prototype と特徴マップ内の各特徴ベクトルとの距離の平均値が減算されているのはより局所的な特徴を Prototype として学習するためである．ProtoPool では以上の損失関数に加え，各クラス内に所属する Prototype の割り当てが互いに異なるものとするため以下の損失関数が課される．

$$L_{orth} = \frac{1}{|\mathbb{Y}|} \sum_{y \in \mathbb{Y}} \sum_{k,l \in \mathbb{K}_y} \frac{\boldsymbol{q}_{y,\cdot,k} \cdot \boldsymbol{q}_{y,\cdot,l}}{\|\boldsymbol{q}_{y,\cdot,k}\|_2 \|\boldsymbol{q}_{y,\cdot,l}\|_2} \tag{4.33}$$

すなわち，ProtoPool では L_{ce}，L_{clst}，L_{sep} および L_{orth} の 4 つの損失関数を最小化するように学習が実行される．以上の学習フレームワークにより，ProtoPool は事前に定義された Prototype とクラスラベルの関係性を必要とすることなく ‘This looks like that’ フレームワークを実現することに成功した．しかし，ProtoPool の学習は著者らによって述べられる通り，温度パラメータを学習中にどのように変化させるかに関する慎重な設定を必要とする [124]．加えて，クラスと Prototype の関係性が事前に定義された slot に学習を依存するため線形分類器以外を用いた学習に適用することが難しい．従って，以上より ProtoPool を類似度学習に適用することは難しいと言える．

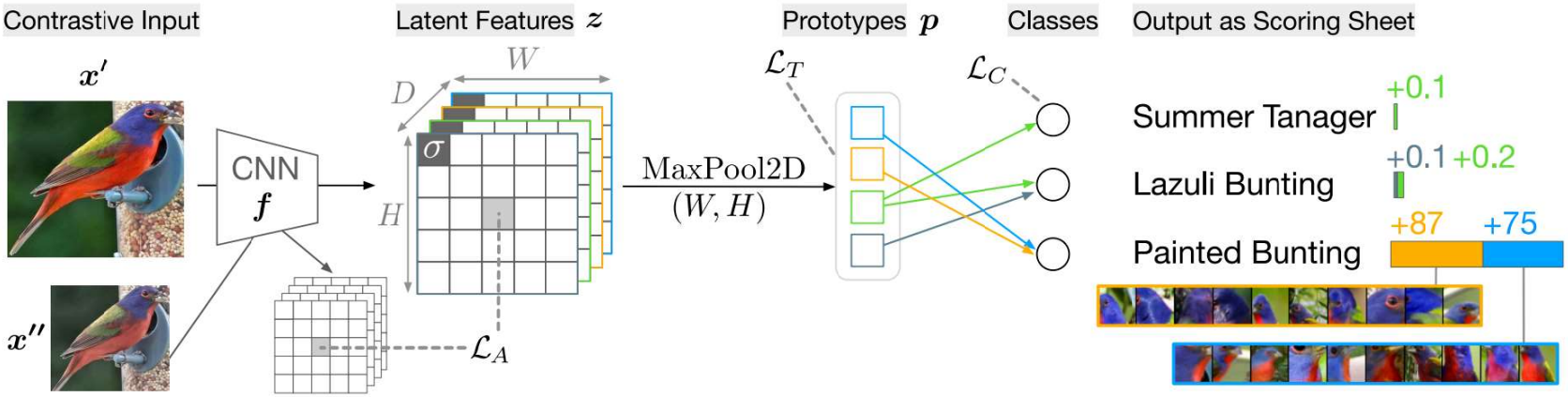

図 4.9: PIP-Net のモデル構造．図は文献 [142] より引用．

■ **PIP-Net**

PIP-Net[142] の学習フレームワークを図 4.9 に示す．PIP-Net は自己教師あり学習を学習フレームワークに採用することにより，Prototype の Purity を向上しつつ Prototype とクラスラベルの関係性を事前に定義する必要なく ‘This looks like that’ フレームワークを実現した手法である．PIP-Net では Prototype $\boldsymbol{p}_j$ と入力画像 x_i 間の類似度 $s_{i,j}$ は下式で与えられる．

$$s_{i,j} = \max_{\boldsymbol{z} \in \mathbf{Z}_i} a(\boldsymbol{z})_j, \text{ where } a(\boldsymbol{z})_j = \frac{\exp(\boldsymbol{z} \cdot \boldsymbol{p}_j)}{\sum_{k \in \mathbb{P}} \exp(\boldsymbol{z} \cdot \boldsymbol{p}_k)} \tag{4.34}$$

そこで PIP-Net における入力画像 x_i のクラス y に対する予測確率 $P(y|x_i)$ は線形分類層の重みを $\boldsymbol{W}$，$\sigma(\cdot)$ を Softmax 関数として $P(y|x_i) = \sigma(\boldsymbol{W}^T \boldsymbol{s}_i)_y$ で与えられる．PIP-Net では Cross Entropy Loss $-\frac{1}{|\mathbb{B}|} \sum_{i \in \mathbb{B}} \log P(y_i|x_i)$ に加え，以下の二つの損失関数を最小化するよう学習がなされる．

$$L_{align} = -\frac{1}{|\mathbb{B}|HW} \sum_{i \in \mathbb{B}} \sum_{(h,w) \in (H,W)} \sum_{j \in \mathbb{P}} \log \left(a(\boldsymbol{z}_{i,h,w})_j a(\boldsymbol{z}'_{i,h,w})_j \right) \tag{4.35}$$

$$L_{tanh} = -\frac{1}{|\mathbb{P}|} \sum_{j \in \mathbb{P}} \log \left(\tanh \left(\sum_{i \in \mathbb{B}} \sum_{(h,w) \in (H,W)} a(\boldsymbol{z}_{i,h,w})_j \right) + \epsilon \right) \tag{4.36}$$

ただし，H および W は特徴マップ $\mathbf{Z}_i$ の高さおよび幅方向のピクセル数であり，$\boldsymbol{z}_{i,h,w}$（$\boldsymbol{z}_{i,h,w}$）は特徴マップ $\mathbf{Z}_i$（$\mathbf{Z}'_i$）の位置 (h,w) に格納される特徴ベクトルである．また $\mathbf{Z}'_i$ は入力画像 x_i に異なるデータ拡張を施した入力画像 x'_i をモデルに入力した結果得られる特徴マップである．ここで，Align loss L_{align} は異なるデータ拡張を施した同一画像パッチに同一の Prototype が割り当てられるよう課される損失関数であり，tanh-loss L_{tanh} は少なくとも一つの画像パッチに各 Prototype が割り当てられるように課される損失関数である．ここで 4.35 式に注目すれば Align loss は任意のサンプル x_i および位置 (h,w) に対して $\forall j, a(\boldsymbol{z}_{i,h,w})_j = a(\boldsymbol{z}_{i,h,w})_j$ かつ，$\boldsymbol{a}(\boldsymbol{z}_{i,h,w})$ が one-hot ベクトルとなるとき最小値をとると分かる．したがって，PIP-Net では Cluster loss や Separation loss を用いることなしに Prototype と画像パッチ特徴を近づける（離す）ことが可能となる．これより，PIP-Net では事前にクラスと Prototype の関係性を定義することなしに ‘This looks like that’ フレームワークを実現することが可能となる．

以上見たように，PIP-Net は一見すると類似度学習に適用可能なように思える．しかし，図 4.10 および図 4.11 に示すように PIP-Net の採用する自己教師あり学習のフレームワークでは全ての特徴ベクトルが同一の値を取る自明解に陥る Feature Collapse を十分防ぐことが出来ない．このような特徴表現のため PIP-Net では各サンプルに対し非常に少ない Prototype 数でその推論を説明することが出来る．しかし，ごく限られた特徴のみしか抽出しない学習方法のために，PIP-Net を類似度学習に適用することは難しい．これは類似度学習の重要な応用先である画像検索タスクでは学習時と推論時で異なるクラスラベルを対象とするため，クラス識別に有効なごく限られた特徴だけではなく多様な特徴を画像内より抽出する必要があるからである．

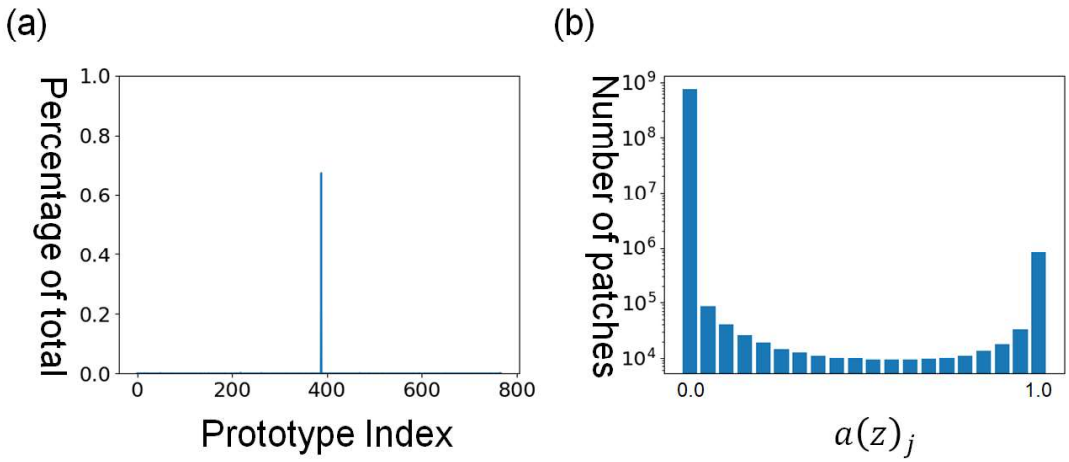

図 4.10: PIP-Net の Prototype 発火に関する評価の結果．ただし，モデルバックボーンには ConvNext-tiny を採用した．(a) 各画像パッチにおいて各 Prototype が Prototype 間で最大の発火となる頻度の割合．一つの Prototype が全体の 70%近くの画像パッチで最大値を取っていることが確認され，Feature Collapse が起こっていることが確認される．(b) Prototype の発火値の頻度分布．Prototype の発火値はほとんど 0 または 1 に二値化されていることが確認できる．

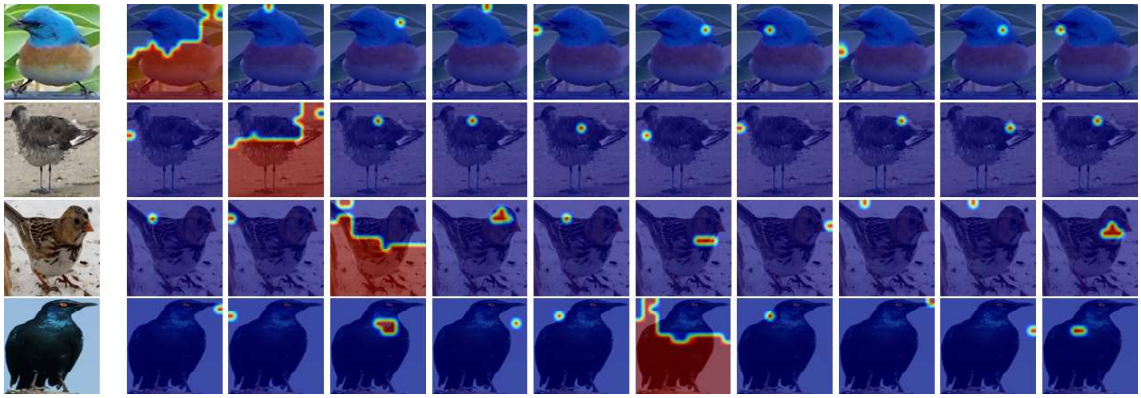

図 4.11: PIP-Net における各入力画像（左端）に対して最も大きな $s_{i,j}$ をとる 10 個の Prototype のヒートマップ $a(Z_{i,h,w})_j$．ただし，モデルバックボーンには ConvNext-tiny を採用した．一つの共通した Prototype が画像内の大半の領域で発火しており，その他の Prototype は高々数個の画像パッチ内でのみ発火していることが確認される．

4.3 まとめ

Limitation: Prototype の解釈性　解釈可能性の研究では非常に多くの研究が提案されており，定量的な解釈可能性の評価が手法間の比較のため重要性を増している．これは ProtoPNet 派生手法においても同様であり，人手を介した実験プロトコルの提案 [120] や CUB200-2011 データセット内にアノテーションされた keypoint annotation[91][137][142]，もしくは合成データ [136] を利用した評価方法が提案されている．しかし，人手を介した実験は倫理審査を必要とする等評価に時間がかかる事に加え，主観的になりやすく再現が非常に難しい．またデータセット内のアノテーションを用いる方法は定量評価による手法間の公正な評価を行うことが出来る一方で，評価指標に対する過適合を引き起こす恐れがある．加えて，解釈可能な Prototype の意味がデータセット内のアノテーションと必ずしも対応するわけではない点にも注意する必要がある（例えば青い体毛に反応する Prototype は青い頭だけでなく青い腹にも反応するため低い解釈性を持つと評価される．一方ですべての腹に反応する Prototype は赤い腹にも青い腹にも反応するため，説明として出力された場合混乱を招くが，腹のみに反応するため高い解釈性を持つと評価されることになる）．解釈可能性をどのように定量的に評価するかは 2023 年現在活発な議論のもと研究が進展している分野であり，本論文のスコープを大きく超える．解釈可能性の定量的な評価手法に関する検討やその改善は今後の課題である．

結論　本章では ProtoPNet 派生手法の進展について説明した．クラス毎に固有の Prototype を学習する手法では ProtoPNet の学習により獲得された Prototype の特徴表現上の課題を解決することで精度の向上や解釈性の向上が達成されてきた．本章で説明した手法により，Prototype の特徴表現に関する知見が多く得られており，より解釈性の高い効率的な Prototype の学習による精度の向上が期待される．一方，クラス毎に固有の Prototype を学習する手法ではクラス毎に複数個の Prototype を用意する必要があるため，非常に多くの Prototype を必要とする課題がある．この課題に対してはクラス間で共通した Prototype の学習を行う手法が提案されており，メモリ効率の改善がなされている．

一方で，従来の ProtoPNet 派生手法のいずれもが類似度学習に適用することが難しい課題を抱えている．類似度学習は学習データとテストデータでクラスラベルの異なる ‘open-set’ な問題設定に有効な手法であり，実用上重要となる．また，この課題のため ‘This looks like that’ フレームワークの応用は学習データとテストデータでクラスラベルの一致する ‘closed-set’ な問題設定に限定されている．そこで，‘This looks like that’ フレームワークの応用を幅広い画像認識タスクへ広げるためには類似度学習に適用可能となるよう ProtoPNet を拡張する手法が重要となると言える．

第5章

ProtoMetric：解釈可能な深層類似度学習モデル

本章では本質的に解釈可能な類似度学習モデルである ProtoMetric を提案する．4 章で説明したように，Prototypical Part Network (ProtoPNet) による ‘This looks like that’ フレームワークは解釈可能なモデルを構築するための有効な手法として大きな注目を集めている．加えて，その汎用性の高さから画像分類タスク以外のタスクへの ‘This looks like that’ フレームワークの応用に関しても研究が進められている．

しかし，ProtoPNet の学習では事前に定義された Prototype とクラスラベルの関係性を必要とするため，ProtoPNet およびその派生手法には類似度学習に適用することが難しい課題がある．Prototype とクラスラベルの関係性は case-based な推論による解釈可能性を実現するために必要な Cluster Loss の算出に用いられる．一方でクラス間で共通した Prototype を用いて ‘This looks like that’ フレームワークを実現する手法では Prototype とクラスラベルの関係性を必要としない．しかし，特定の分類器を前提とした学習フレームワークや入力画像より限られた特徴のみを抽出する学習のため，これらの手法も類似度学習へ適用することは難しい．この課題のため，‘This looks like that’ フレームワークの応用は学習データに含まれるクラスとテストデータに含まれるクラスが同一となる closed-set なタスクに制限されている．

本章ではこの課題に対処するため，類似度学習に適用可能となるよう ProtoPNet を拡張した ProtoMetric を提案する．特に，本章ではクラスラベルと Prototype の関係性を学習中に推定しながら，その推定結果に基づき特徴マップ内のある特徴ベクトル（画像パッチ特徴）と Prototype を近づける新規の Cluster Loss を提案する．

本章の構成は以下の通りである．まず，5.1 章では提案手法である ProtoMetric について説明する．次に 5.2 章では ProtoMetric の詳細画像認識タスクおよび画像検索タスクにおける実験結果についてそれぞれ説明し，最後に 5.3 章で本章をまとめる．また，本章では 4 章と同様にバッチ内のデータインデックスの集合を $\mathbb{B}$，全てのクラスラベルの集合を $\mathbb{Y}$，入力画像およびそのクラスラベルを x_i および y_i と表し，全ての Prototype インデックスの集合を $\mathbb{P}$，クラス y に所属する Prototype インデックスの集合を $\mathbb{P}_y$ と記述する．また x_i をモデルバックボーンに入力した結果得られる特徴マップを $\mathbf{Z}_i$，画像 x_i と Prototype $\boldsymbol{p}_j$ との類似度を $s_{i,j}$ と表し，$s_{i,j}$ を列方向に並べて得られる列ベクトルを $\boldsymbol{s}_i$ と表す．

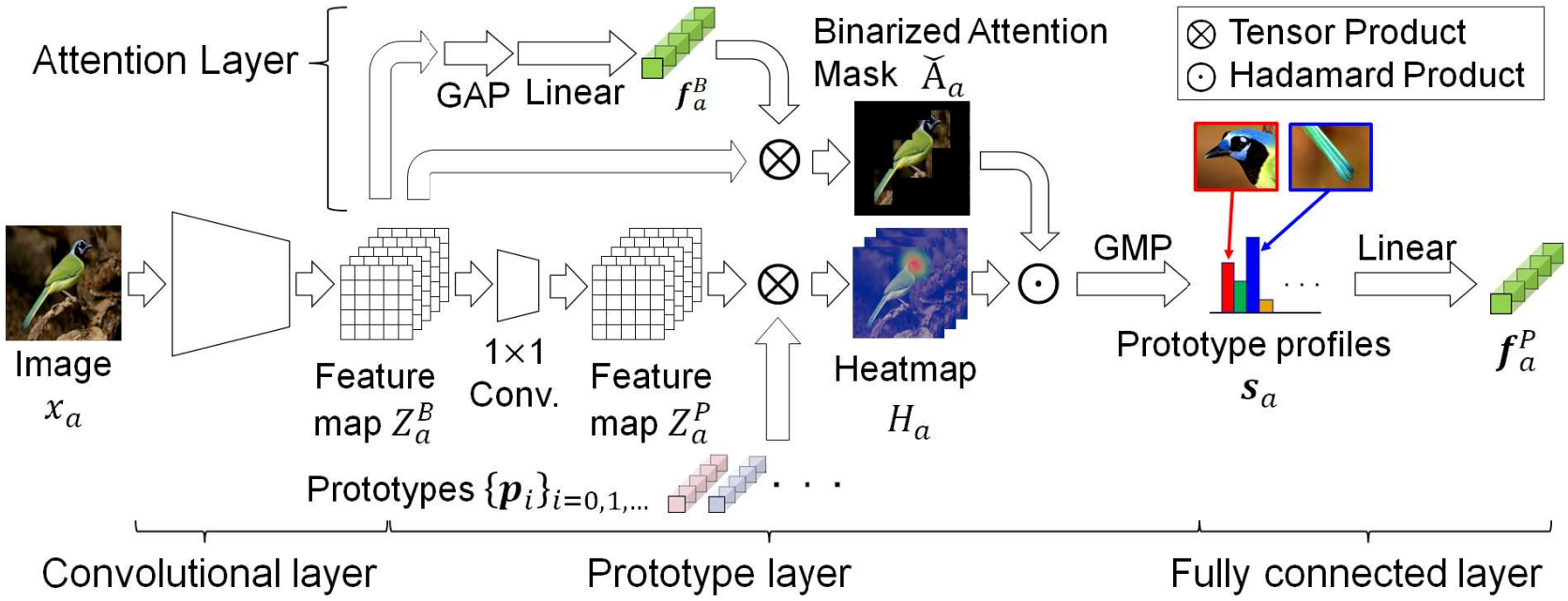

図 5.1: ProtoMetric のモデル構造．ProtoMetric は Convolutional layer，Attention layer，Prototype layer および Fully conected layer の 4 つのモジュールから構成される．

Algorithm 2 ProtoMetric の推論過程

Require: Input image x_a

Ensure: Feature vectors $\boldsymbol{f}_a^B$, $\boldsymbol{f}_a^P$

Ensure: Foreground mask $\check{\mathbf{A}}_a$

Ensure: Prototype profile $\boldsymbol{s}_a$ and prototype heatmap $\mathbf{H}_a$

// Convolutional layer

Transform image x_a into feature map $\mathbf{Z}_a^B$.

// Attention layer

Transform feature map $\mathbf{Z}_a^B$ into feature vector $\boldsymbol{f}_a^B$ with GAP and the linear layer.

Calculate the RAM $\mathbf{A}_a$ from $\mathbf{Z}_a^B$ and $\boldsymbol{f}_a^B$.

Obtain foreground mask $\check{\mathbf{A}}_a$ by binarizing $\mathbf{A}_a$.

// Prototype layer

Transform feature map $\mathbf{Z}_a^B$ into $\mathbf{Z}_a^P$ by 1×1 convolutional layer.

Calculate prototype heatmap $\mathbf{H}_a$ and prototype profile $\boldsymbol{s}_a$ following Eqs. 5.1 and 5.2, respectively.

// Fully connected layer

Transform prototype profile $\boldsymbol{s}_a$ into feature vector $\boldsymbol{f}_a^P$ with linear layer.

5.1 提案手法

本章では提案手法である ProtoMetric の説明を行う．以下では，ProtoMetric のモデル構造（5.1.1 章）および学習方法（5.1.2 章）について説明した後，ProtoMetric によるモデル推論の解釈方法について説明する（5.1.3 章）．

5.1.1 ProtoMetric のモデル構造

本章では ProtoMetric のモデル構造について説明する．図 5.1 に示すように ProtoMetric のモデル構造は Convolutional layer，Attention layer，Prototype layer，Fully connected layer により構成される．また提案手法のモデル推論のプロセスは Alg. 2 にまとめられる．以下では提案手法のモデル推論プロセスを詳細に説明した後 Prototype layer で用いる Multi-head trick について説明する．

■ モデル推論のプロセス

まず入力画像 x_a は Convolutional layer により特徴マップ $\mathbf{Z}_a^B$ に変換される．ここで下付き文字 a はデータインデックスであり，以下では特に断りのない限り下付き文字 $a, b, ...$ によりデータインデックスを表す．次に Attention layer は特徴マップ $\mathbf{Z}_a^B$ に Global Average Pooling (GAP) および全結合層を適用することで $\mathbf{Z}_a^B$ を特徴ベクトル $\boldsymbol{f}_a^B$ に変換する．変換後，提案手法では $\hat{\boldsymbol{f}}_a^B$ および $\mathbf{Z}_a^B$ を用いて，Ranking Activation Map（RAM）[60] により注視領域 $A_{a,hw}$ を算出する．ただし，$\hat{\boldsymbol{f}}_a^B$ は $\boldsymbol{f}_a^B$ の

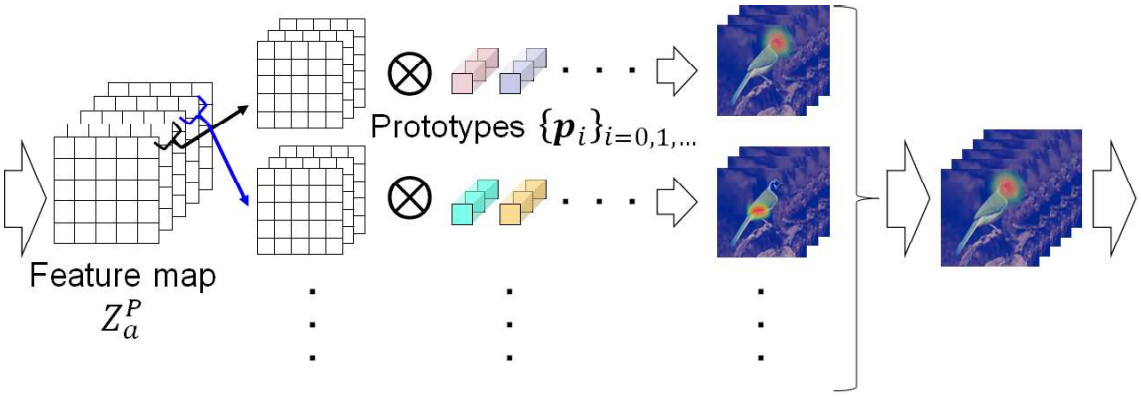

図 5.2: Multi-head trick を用いる場合における Prototype layer の構造.

L2 正規化結果であり，以下では断りの無い限りベクトル $\boldsymbol{v}$ の L2 正規化結果を $\hat{\boldsymbol{v}}$ により表す．さらに提案手法では RAM を降順に並べ累積和を取った値が閾値 t_{ram} 以下となる領域の値が 1，その他の領域の値が 0 となるように注視領域を二値化する．以下では二値化された注視領域を $\check{A}_{a,hw}$ と記し，$\check{A}_{a,hw}$ を Attention mask と呼称する．

続く Prototype layer では 1×1 畳み込み層により特徴マップ $\mathbf{Z}_a^B$ を $\mathbf{Z}_a^P$ に変換する．その後特徴マップ $\mathbf{Z}_a^P$ の各ピクセルに含まれる特徴ベクトルと Prototype $\{\boldsymbol{p}_i\}_{i=0,1,...}$ とのコサイン類似度を算出する．ただし，i は Prototype のインデックスであり，以下では特に断りのない限り，$i, j, ...$ により Prototype のインデックスを表す．これより入力画像 x_a に対する各 Prototype の類似度マップ $\mathbf{H}_a$ が算出される．すなわち類似度マップ $\mathbf{H}_a$ の位置 (h, w) における Prototype $\boldsymbol{p}_i$ に対応する成分 $H_{a,i,hw}$ は

$$H_{a,i,hw} = \max(0, \hat{\boldsymbol{z}}_{a,hw}^P \cdot \hat{\boldsymbol{p}}_i), \tag{5.1}$$

と定義される．ただし，$\boldsymbol{z}_{a,hw}^P$ は特徴マップ $\mathbf{Z}_a^P$ の位置 (h, w) に含まれる特徴ベクトルである．また以下では $\boldsymbol{z}_{a,hw}^P$ を画像パッチ特徴と呼称する．次に Prototype layer では $\mathbf{H}_a$ のうち Attention mask $\check{A}_{a,hw}$ が 1 となる領域において，画像内で最大の値を画像 x_a と各 Prototype との類似度 $\boldsymbol{s}_a$ として定義する．すなわち，$s_{a,i}$ を $\boldsymbol{s}_a$ のうち，Prototype $\boldsymbol{p}_i$ に対応する成分とすれば，

$$s_{a,i} = \max_{h,w} H_{a,i,hw} \check{A}_{a,hw}. \tag{5.2}$$

である．以下では $\boldsymbol{s}_a$ および $\mathbf{H}_a$ を Prototype プロファイルおよび Prototype ヒートマップと呼称する．最後に Prototype プロファイル $\boldsymbol{s}_a$ は Fully connected layer において全結合層により特徴ベクトル $\boldsymbol{f}_a^P$ に変換される．提案手法では $\hat{\boldsymbol{f}}_a^P$ により，画像間類似度を推論する．

■ Prototype layer における Multi-head trick

5.1.2 章で説明するように提案手法ではバッチ内サンプルの比較によりサンプル固有の Prototype を推定する．そのため，多くの Prototype 数を用いる場合には，一度のバッチ処理で抽出される Prototype

数の全 Prototype 数に対する割合は小さくなり，各サンプルへの適切な Prototype の割り当てが難しくなる．この課題に対処するため，提案手法では Prototype layer に Multi-head trick を適用する．具体的には図 5.2 に示すように，特徴マップ Z_a^P をチャンネル方向にヘッド数 H で分割し，各特徴マップに $\frac{P}{H}$ 個の Prototype を割り当て，Head 毎に後述する Cluster loss を算出する．これよりバッチサイズを小さく保ちつつ，大きな Prototype 数に対してもクラスと Prototype の関係を効率的に推定することが可能となる．本論文では特に断りのない限り，Head 数は 1 に設定される．

5.1.2 ProtoMetric の学習フレームワーク

先述したように ProtoMetric では事前定義された Prototype とクラスラベルの関係性を学習に利用しない．そのため，提案手法では sub-optimal な二段階の学習戦略を取ることなく end-to-end に学習を行うことが可能となる．具体的には，提案手法は Task loss，Auxiliary loss，Cluster loss，Suppression loss の 4 つの損失関数を最小化するように学習される．以下では各損失関数について詳細に説明する．学習の全プロセスは Alg. 3 にまとめられる．

■ Task Loss

Task loss L_{task} は画像分類および画像検索タスクを解くため，ProtoMetric の出力 $\boldsymbol{f}_a^P$ に課される損失関数である．提案手法では，詳細画像認識の実験設定では Proxy Anchor loss[67] を，画像検索の実験設定では Black-box モデルを教師モデルとした Relational Knowledge Distillation (RKD) loss[55] を Task loss として用いる．これは Prototype の学習のため，十分長く ProtoMetric を学習する必要がある一方で，画像検索タスクでは学習時とテスト時とで扱うクラスが異なるために過学習が発生しやすくなる課題に対処するためである．以下では各実験設定における Task loss を定式化する．また本章では明瞭性のため $\boldsymbol{f}_a^P$ を $\boldsymbol{f}_a$ と略記する．

詳細画像分類タスクにおける Task Loss　本実験設定では特に断りのない限り，Task loss として Proxy Anchor loss[67] を用いる．そこで本実験設定における Task loss L_{task} は以下のように定式化される：

$$L_{task} = \frac{1}{|\mathbb{Q}^+|} \sum_{q \in \mathbb{Q}^+} \log \left(1 + \sum_{a \in \mathbb{B}_q^+} e^{-\alpha(\hat{\boldsymbol{f}}_a \cdot \hat{\boldsymbol{g}}_q - \delta)} \right) + \frac{1}{|\mathbb{Q}|} \sum_{q \in \mathbb{Q}} \log \left(1 + \sum_{a \in \mathbb{B}_q^-} e^{\alpha(\hat{\boldsymbol{f}}_a \cdot \hat{\boldsymbol{g}}_q - \delta)} \right), \tag{5.3}$$

ここで，δ，α，$\mathbb{Q}$ および $\mathbb{Q}^+$ はそれぞれ，マージン，スケール因子，全てのプロキシインデックスの集合，ミニバッチ内に含まれるデータのクラスラベルに所属するプロキシインデックスの集合である．また，$\boldsymbol{g}_q$ はインデックス q に対応するプロキシであり，$\mathbb{B}_q^+$（$\mathbb{B}_q^-$）は q 番目のプロキシと同一の（異なる）クラスラベルをもつデータインデックスの集合である．

画像検索タスクにおける Task Loss　本実験設定では，RKD loss[55] を用いて事前学習された Black-box モデルを蒸留し，学習を行う．RKD loss は Distance-wise distillation loss L_D と Angle-wise distillation

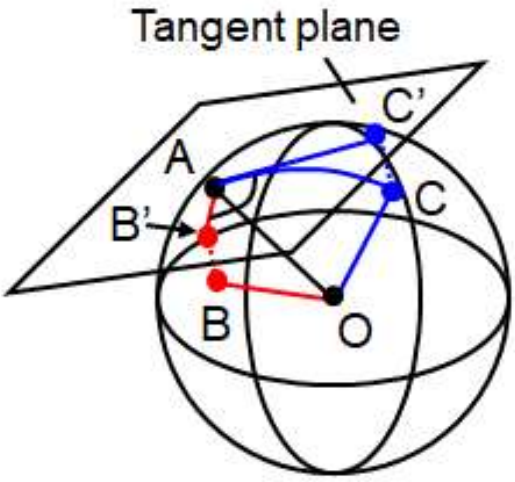

図 5.3: 原点 O を中心とする超球面上の三点 A，B，および C のなす角度．超球面上で定義される角度 $\angle BAC$ はユークリッド空間上の角度 $\angle B'AC'$ に等しい．ここで B' および C' は超球面上の点 B および C の点 A における超球面の接平面への射影である．

loss L_A の二つの損失関数から構成される．Park らによって提案された RKD loss[55] は Euclid 空間を対象に設計されているため，提案手法では超球面幾何を考慮するよう各損失関数を改変した．以下では各改変の詳細について説明する．

Distance-wise distillation loss L_D は教師モデルと生徒モデルより推論される画像間距離が，等しくなるよう課される損失関数である．超球面上の 2 点間の距離は 2 点を結ぶ大円の短い方の弧の長さとして定義される．そこで，x_a の教師モデル出力結果を $\boldsymbol{t}_a$ とすれば，Distance-wise distillation loss は

$$L_D = \sum_{a,b} l_\delta \left(\arccos(\hat{\boldsymbol{f}}_a \cdot \hat{\boldsymbol{f}}_b), \arccos(\hat{\boldsymbol{t}}_a \cdot \hat{\boldsymbol{t}}_b) \right), \tag{5.4}$$

と再定義される．ただし，l_δ は Smooth L1 loss であり，下式のように定義される．

$$l_\delta(x, y) = \begin{cases} \frac{1}{2}(x-y)^2 & \text{for } |x-y| \leq 1, \\ |x-y| - \frac{1}{2} & \text{otherwise.} \end{cases} \tag{5.5}$$

Angle-wise distillation loss L_A は教師モデルと生徒モデル各々より出力される，特徴ベクトル間のなす角度が等しくなるように課される損失関数である．図 5.3 に示すように，超球面上における 3 点 A, B, C のなす角 $\angle BAC$ は点 A と点 B, C の，点 A で超球面と接する接平面への射影 B', C' のなす角 $\angle B'AC'$ と等しい．よって，Angle-wise distillation loss は

$$L_A = \sum_{a,b,c} l_\delta \left(U(\boldsymbol{f}_b, \boldsymbol{f}_a) \cdot U(\boldsymbol{f}_c, \boldsymbol{f}_a), U(\boldsymbol{t}_b, \boldsymbol{t}_a) \cdot U(\boldsymbol{t}_c, \boldsymbol{t}_a) \right), \tag{5.6}$$

と再定義される．ただし，

$$U(\boldsymbol{u},\boldsymbol{v}) = \frac{V(\boldsymbol{u},\boldsymbol{v})}{\|V(\boldsymbol{u},\boldsymbol{v})\|}, \text{where } V(\boldsymbol{u},\boldsymbol{v}) = (\hat{\boldsymbol{u}}\cdot\hat{\boldsymbol{v}})\hat{\boldsymbol{v}} - \hat{\boldsymbol{u}}. \tag{5.7}$$

と定義した．

以上より，本実験設定における Task loss は下式のように定義される：

$$L_{task} = \lambda_D L_D + \lambda_A L_A, \tag{5.8}$$

ここで λ_D および λ_A は各損失関数の重みを調整するハイパーパラメータである．本論文では [55] に従い，各パラメータを 1.0 および 2.0 で固定した．

■ Auxiliary Loss

Auxirialy loss は Convolutional layer がより良い特徴抽出能力を獲得し，より良い Attention mask が Attention layer より出力されるようにするため，$\boldsymbol{f}_a^B$ に課される損失関数である．提案手法では Auxiliary loss として，Margin loss[33] を用いる．そこで，Auxiliary loss L_{aux} は以下のように定義される．

$$L_{aux} = \frac{1}{N}\sum_{a\in\mathbb{B}}([d_{ap} + m - \beta]_+ + [m - d_{an} + \beta]_+), \tag{5.9}$$

ここで m はマージン，β は学習可能なパラメータであり，a，p，n はそれぞれ Anchor，Positive，Negative サンプルのデータインデックスである．$[\cdot]_+$ は ReLU 関数を表しており，N は和の中で非ゼロの値をとる項の数である．また，d_{ap} および d_{an} はそれぞれ $\|\hat{\boldsymbol{f}}_a^B - \hat{\boldsymbol{f}}_p^B\|$，$\|\hat{\boldsymbol{f}}_a^B - \hat{\boldsymbol{f}}_n^B\|$ により定義される，Anchor，Positive サンプル間の距離および Anchor，Negative サンプル間の距離である．また画像検索の実験設定では過学習抑制のため正則化損失 [89] を追加した．

■ Cluster Loss

Cluster loss は各 Prototype が学習データ中のある画像パッチを代表するように，Prototype と画像パッチ特徴を近づけるため課される損失関数である．類似度学習では Prototype とクラスラベルの関係性を事前に定義することが出来ない．そのため，従来の ProtoPNet 派生手法と同様に Cluster loss を算出することは難しい．この課題に対し本論文では，ミニバッチ内のサンプルを比較することにより Prototype の各クラスへの帰属度を推定し，その推定に基づいて Cluster loss を算出する新規の手法を提案する．以下では，Cluster loss の算出方法について詳細に説明する．

図 5.4 に提案手法における Prototype 所属度の推定プロセスを示す．図 5.4 に示すように，提案手法ではまずバッチ内サンプルの比較によりサンプルに特有の Prototype を抽出する．Prototype $\boldsymbol{p}_i$ がサンプル x_a にどの程度含まれるかは，Prototype プロファイルの第 i 成分 $s_{a,i}$ により表現される．そこで x_a はバッチ内の他サンプル x_b に対し，$s_{a,i} - s_{b,i}$ が大きな値となる Prototype を含むと期待さ

れる．従って，各 x_b について $s_{a,i} - s_{b,i}$ が大きな値となる Prototype を抽出し，その和集合をとれば x_a に含まれる Prototype の集合を再構築できる．この抽出操作は Gumbel top-k trick[52] によるサンプリング操作：

$$E(x_a, \mathbf{p}_i) = \sum_{b \in \mathbb{B}/\{a\}} \Gamma_k \left(\frac{s_{a,i} - s_{b,i}}{\tau} \right), \tag{5.10}$$

により実行される．ただし，τ は温度パラメータであり，本論文では τ を 0.05 で固定した．また，Γ_k は下式で定義される Gumbel top-k 操作である．

$$\Gamma_k(s_i) = \mathbf{1} \left(i \in \arg\operatorname*{topk}_j \frac{\exp(s_j + \gamma_j)}{\sum_k \exp(s_k + \gamma_k)} \right), \tag{5.11}$$

ここで γ_* $(* \in \{j, k\})$ は標準 Gumbel 分布に従う確率変数であり $\mathbf{1}$ は括弧内の条件が真のときに 1，その他の時に 0 を返す指示関数である．また本論文では，Gumbel top-k 操作における k を 3 とした．得られたサンプリング結果を同一クラスラベルを持つサンプル間で平均化すれば各クラスに対する Prototype の所属度が得られる：

$$E(y, \mathbf{p}_i) = \frac{\sum_{a \in \mathbb{B}} \mathbf{1}(y_a = y) E(x_a, \boldsymbol{p}_i)}{\sum_{a \in \mathbb{B}} \mathbf{1}(y_a = y)}. \tag{5.12}$$

得られた各クラスに対する Prototype の所属度はクラスラベルに依らずある Prototype に偏る場合がある．そのため，サンプリング結果のバイアスを除去する必要がある．直感的には得られた所属度を表す行列 $\boldsymbol{E}$ について行方向の和が等しくなるよう正規化した後，列方向の和が等しくなるように正規化する方式が有効であるように思われる．ただし $E_{yi} \equiv \boldsymbol{E}[y, i] \equiv E(y, \boldsymbol{p}_i)$ であり，$\boldsymbol{E}[y, i]$ は行列 $\boldsymbol{E}$ の y 行 i 列の成分を表す．ここで $\boldsymbol{E}$ の行方向の正規化はバッチ内に含まれる全てのクラスについて Prototype 所属度を平均したとき各 Prototype の所属度が等しくなるようにする処理であり，列方向の正規化は各クラスに含まれる Prototype の総数が等しいようにする処理と捉えられる．しかし，上記の正規化方法はあるクラスがバッチ内に含まれるか否かを考慮しないため，学習にバイアスを導入する場合がある．そこで本論文ではクラスがバッチ内に含まれるか否かを考慮しながら，行方向および列方向の正規化を同時に行うバイアス除去方法を導入する．行方向及び列方向の同時正規化は与えられた行列 $\boldsymbol{E}$ に対し行方向の和が $\boldsymbol{u}$，列方向の和が $\boldsymbol{v}$ となるよう $\boldsymbol{E}$ を $\boldsymbol{E}'$ に変形する行列スケーリングの問題設定

$$\boldsymbol{E} \to \boldsymbol{E}' \text{ s.t. } \sum_i E_{ij} = u_j, \ \sum_j E_{ij} = v_i, \tag{5.13}$$

として定式化でき，Sinkhorn iteration[2, 25] により解が求まる．ここで各クラスに含まれる Prototype の総量が（1 に）等しいという仮定を満たすには全ての j に対し $v_j = 1$ とすればよい．そこで，クラスがバッチ内に含まれるか否かを考慮して $\boldsymbol{u}$ を決定すれば求めるバイアス除去を達成出来る．本論文ではこの目的のため，各クラスに対する Prototype の所属度を格納する外部メモリ $\boldsymbol{M}$ を利用する．より具体的には，バッチ内における Prototype 所属度の総和が，外部メモリ $\boldsymbol{M}$ に含まれる Prototype 所属度の，バッチ内に含まれるクラスに対する総和と一致するよう $\boldsymbol{u}$ を決定する．すなわち，

$$u_i = \frac{\sum_{y \in \mathbb{Y}} \boldsymbol{M}[y, i]}{|\mathbb{P}|}, \ v_j = 1, \tag{5.14}$$

とする．ただし，C，$\mathbb{Y}$，$\mathbb{P}$ はそれぞれ学習データセット全体に含まれるクラス数，バッチ内に含まれるクラスインデックスの集合，全ての Prototype インデックスの集合である．また $|\mathbb{P}|$ は $\mathbb{P}$ の cardinality である．外部メモリ $\boldsymbol{M}$ は全ての値が同じ値 $C/|\mathbb{P}|$ により初期化された後，バイアスの除去された Prototype 所属度を用いて下式に従い更新される．

$$\boldsymbol{M}[y,i] \leftarrow m \cdot \boldsymbol{M}[y,i] + (1-m) \cdot \boldsymbol{E}'[y,i], \tag{5.15}$$

ここで m はモメンタム係数であり，本論文では $m = 0.9$ とした．

Calculation of cluster loss　Prototype とクラスラベルの関係性が推定できれば，推定した関係性に基づき Cluster loss を算出出来る．ただし画像内の物体の見え方の違いを考慮するため，クラスに対する Prototype の所属度をサンプル毎の所属度に修正する．具体的にはサンプル x_a がどの程度 Prototype $\boldsymbol{p}_i$ で表現される画像特徴を含んでいるかに関する推定を，線形割り当て問題

$$T^{y*}_{a,i} = \arg\min_{T^{y}_{a,i}} \sum_{a\in\mathbb{B}, i\in\mathbb{P}} T^{y}_{a,i} C_{a,i} \text{ s.t. } \sum_{i\in\mathbb{B}} T^{y}_{a,i} = \mathbf{1}(y_a = y),\ \sum_{a\in\mathbb{P}} T^{y}_{a,i} = N_y \boldsymbol{E}'[y,i], \tag{5.16}$$

として定式化する．ただし，N_y はクラスラベルが y となるミニバッチ内のサンプル数であり，$C_{a,i} = \min_{\boldsymbol{z}\in\mathbf{Z}^P_a} \|\boldsymbol{z} - \boldsymbol{p}_i\|_2^2$ である．また 5.16 式の解は Sinkhorn Knopp Algorithm[11] により算出される．サンプル x_a が Prototype $\boldsymbol{p}_i$ を含むとき，すなわち $C_{a,i}$ が小さいときには解 $T^{y*}_{a,i}$ は大きな値を取り，x_a が $\boldsymbol{p}_i$ を含まないときには $T^{y*}_{a,i}$ は小さな値を取る．そこで 5.16 式の解 $T^{y*}_{a,i}$ はサンプル平均がクラス平均と一致する，各サンプルに対する Prototype の所属度と捉えることが出来る．よって，$T^{y*}_{a,i}$ により画像内の物体の見え方の違いを考慮した Cluster loss の算出が可能となる．また本論文においても ProtoPool と同様に Prototype と画像パッチ特徴の距離の最小値と平均値の差を最大化 [124] するよう Cluster loss を定義する．よって本論文における Cluster loss L_{clst} は下式のように定義される：

$$L_{clst} = \sum_{y\in\mathbb{Y}} \sum_{a\in\mathbb{B}, i\in\mathbb{P}} T^{y*}_{a,i} \hat{C}_{a,i},\ \text{where } \hat{C}_{a,i} = \min_{\boldsymbol{z}\in\mathbf{Z}^P_a} \|\boldsymbol{z} - \boldsymbol{p}_i\|_2^2 - \frac{\alpha}{|\mathbf{Z}^P_a|} \sum_{\boldsymbol{z}\in\mathbf{Z}^P_a} \|\boldsymbol{z} - \boldsymbol{p}_i\|_2^2. \tag{5.17}$$

■ Suppression Loss

ProtoMetric では，Prototype と画像パッチ特徴の距離の最小値と平均値の差を最大化する trick に加え，Attention layer により出力される Attention mask を利用して，背景領域を Prototype が学習することを防ぐ．この目的のため提案手法では以下の Suppression loss を導入する：

$$L_{supp} = \frac{1}{|\mathbb{B}|} \sum_{a\in\mathbb{B}} \frac{\sum_{i\in\mathbb{P},h,w} (1 - \check{A}_{a,hw}) H_{a,i,hw}}{\sum_{i\in\mathbb{P},h,w} (1 - \check{A}_{a,hw})}. \tag{5.18}$$

これより背景領域（Attention mask $\mathbf{A}_a = 0$ の領域）の画像パッチ特徴と Prototype を遠ざけ，背景領域を Prototype が学習することを防ぐことが出来る．

Algorithm 3 ProtoMeric の学習プロセスの全体

Require: Training datasets $\mathcal{T}$

Require: Number of classes C.

Require: Number of prototypes P.

Require: Model F parameterized by θ.

Initialize the external memory
$\boldsymbol{M}[i,j] \leftarrow \frac{P}{C}$ for all $0 < i < C, 0 < j < P$
Randomly initialize the prototypes $\{\mathbf{p}_i\}_{i=0,1,\ldots,P}$
for $epoch = 0$ to n_epochs **do**
  for mini-bath $\mathcal{B} = \{x_a, y_a\}_{a=0,1,\ldots} \in \mathcal{T}$ **do**
    // Encode the input images
    Take input image x_a into model F and obtain feature vectors $\boldsymbol{f}_a^B$ and $\boldsymbol{f}_a^P$, foreground mask $\check{\mathbf{A}}_a$, prototype profile $\boldsymbol{s}_a$, and prototype heatmap $\mathbf{H}_a$ following Algorithm 2.
    // Cluster loss
    Sample the prototypes following Eq. 5.10 and obtain the sampling results $E(x_a, \boldsymbol{p}_i)$.
    Modify $E(x_a, \boldsymbol{p}_i)$ to $E'(y_a, \boldsymbol{p}_i)$ with Sinkhorn algorithm so that Eq. 5.14 is satisfied.
    Update the external memory following Eq. 5.15.
    Solve the linear assignment problem defined in Eq. 5.17 and obtain the solution $T_{a,i}^{y*}$.
    Calculate the cluster loss L_{clst} following Eq. 5.17.
    // Other losses
    Calculate the task loss L_{task} with $\boldsymbol{f}_a^P$ and y_a.
    Calculate the auxiliary loss L_{aux} with $\boldsymbol{f}_a^B$ and y_a.
    Calculate the suppression loss L_{supp} with $\mathbf{H}_a$ and $\check{\mathbf{A}}_a$ following Eq. 5.18.
    Calculate the overall loss following Eq. 5.19.
    // Model update
    Update the model parameters θ and the prototypes on the basis of the gradient with respect to them.
  end for
end for
Conduct prototype projection following Eq. 5.20

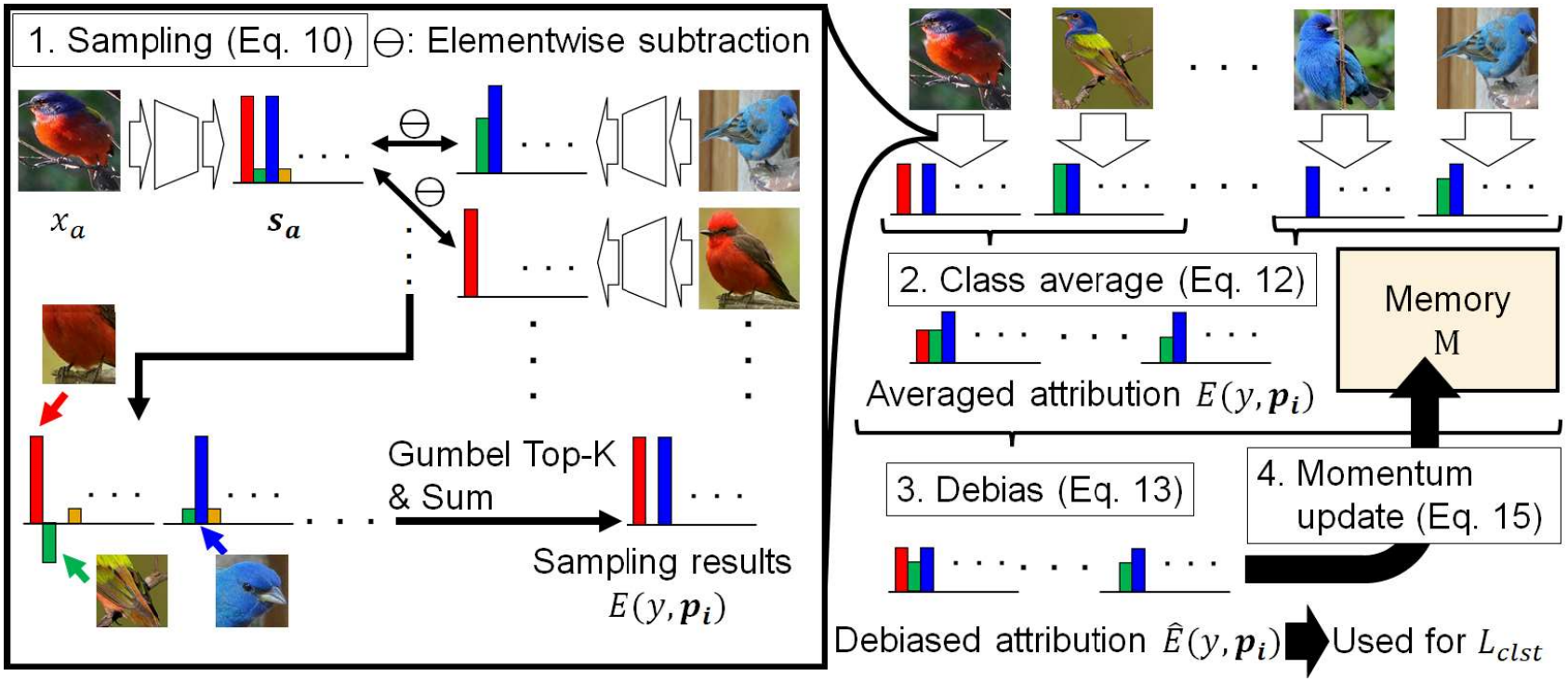

図 5.4: Prototype 所属度の推定プロセス.

■ **全体の損失関数と Prototype projection**

以上より ProtoMetric の学習で最小化する損失関数は

$$L_{total} = L_{task} + \lambda_{aux} L_{aux} + \lambda_{clst} L_{clst} + \lambda_{supp} L_{supp}, \tag{5.19}$$

となる．ただし λ_{aux}，λ_{clst}，λ_{supp} は各損失関数の重みを決定するハイパーパラメータである．また，提案手法では学習終了後，下式に従い各 Prototype を学習データ内で最も近い画像パッチ特徴に射影する（Prototype projection を実施する）．

$$\boldsymbol{p}_i \leftarrow \underset{\boldsymbol{z} \in \mathbf{Z}_a, a \in \mathbb{T}}{\arg\max}\, \hat{\boldsymbol{z}} \cdot \hat{\boldsymbol{p}}_i, \tag{5.20}$$

ただし，$\mathbb{T}$ は学習データセットに含まれる全てのデータインデックスの集合である．これより，各 Prototype が学習データ中のある画像パッチを代表することが保証され，case-based な推論による解釈可能性が実現される．

5.1.3 ProtoMetric の推論プロセスと推論根拠の解釈方法

本章では各実験設定における ProtoMetric の推論方法および推論の解釈方法について説明する．

詳細画像分類タスクにおける推論方法と解釈　詳細画像分類の実験設定ではまず学習データ内のサンプルより抽出される特徴ベクトルをクラス毎に平均化し，クラス重心を算出する．その後入力サンプルと最も大きいコサイン類似度を持つクラス重心に対応するクラスを予測クラスとする．すなわち，入力サンプル x_a に対する予測ラベル y_{pred} は，クラス y のクラス重心を $\boldsymbol{c}_y$ として，

$$y_{pred} = \underset{y}{\arg\max}\, \hat{\boldsymbol{f}}_a^B \cdot \hat{\boldsymbol{c}}_y. \tag{5.21}$$

となる．式 5.21 による推論は線形分類層による推論と等価となるため，従来の ProtoPNet と変わらない計算コストでクラス分類を行うことが可能となる．また，$W_{i,j}^{FC}$ を Fully connected layer における全結合層の重みとすれば，$\boldsymbol{f}_{a,i}^B = \sum_j W_{i,j}^{FC} s_{a,j}$ なので，

$$\hat{\boldsymbol{f}}_a^B \cdot \hat{\boldsymbol{c}}_y = \sum_{i,j} s_{a,j} \frac{W_{i,j}^{FC} \hat{c}_{y,i}}{\|\boldsymbol{f}_a^B\|} = \sum_j s_{a,j} C_j^{a,y}. \tag{5.22}$$

である．そこで，5.22 式右辺の各項 $s_{a,j} C_j^{a,y}$ の内，大きな値に対応する Prototype が画像 x_a がクラス y に分類される根拠となる．ただし $C_j^{a,y} = \sum_i \frac{W_{i,j}^{FC} \hat{c}_{y,i}}{\|\boldsymbol{f}_a^B\|}$ とした．

画像検索タスクにおける推論方法と解釈　5.1.1 節で述べたように，提案手法ではコサイン類似度を用いて画像間の類似度を推論する．入力画像 x_a および x_b 間の画像間類似度は 5.22 式と同様の議論から，

$$\hat{\boldsymbol{f}}_a^B \cdot \hat{\boldsymbol{f}}_b^B = \sum_{i,j,k} s_{a,j} \frac{W_{i,j}^{FC} W_{i,k}^{FC}}{\|\boldsymbol{f}_a^B\| \|\boldsymbol{f}_b^B\|} s_{b,k} = \sum_{j,k} s_{a,j} R_{j,k}^{a,b} s_{b,k}. \tag{5.23}$$

となる．そこで，5.23 式右辺の各項 $s_{a,j} R_{j,k}^{a,b} s_{b,k}$ のうち，大きな値に対応する Prototype が画像 x_a と x_b が類似する根拠となる．ただし，$R_{j,k}^{a,b} = \sum_i \frac{W_{i,j}^{FC} W_{i,k}^{FC}}{\|\boldsymbol{f}_a^B\| \|\boldsymbol{f}_b^B\|}$ とした．

5.2　実験

本論文では ProtoPNet 派生手法および画像検索タスクにおいて最も一般的に用いられる，CUB200-2011 データセット [8, 9] および Stanford Cars データセット [12] を用いて評価した．加えて Stanford Dogs[10] データセットを用いて Deformable ProtoPNet[116] との比較を実施した．本章ではまず，実装の詳細 (5.2.1 節) について説明した後，詳細画像分類問題における ProtoPNet 他手法と ProtoMetric の比較結果 (5.2.2 節) について説明する．その後提案手法の各モジュールに関する Ablation study の結果 (5.2.3 節) および詳細画像分類問題における ProtoMetric の定性的評価の結果 (5.2.4 節) について説明し，最後に ProtoMetric の画像検索タスクへの適用結果について説明する (5.2.5 節).

5.2.1　実験設定の詳細

提案手法は Roth らの実装 [73] を土台に実装した．ただし，従来手法 [45, 102, 124] に従い，CUB200-2011，Stanford Cars データセット内の画像は対象物体を囲む矩形領域により切り出した．また従来研究 [93, 116, 124] に従い，Resnet 50 は i-Naturalist2017[40] で事前学習されたモデルを，その他のモデルは Imagenet[7] で事前学習されたモデルを用いた．モデルバックボーンより出力される特徴マップ Z^B に 1×1 畳み込み層を適用することで得られる特徴マップ Z^P の次元数は，CUB200-2011，Stanford Cars および Stanford Dogs それぞれで 256，128 および 256 とした．同様に特徴マップ Z^B に GAP および全結合層を適用し得られる特徴量ベクトル $\mathbf{f}^B$ の次元数を，CUB200-2011，Stanford Cars および Stanford Dogs それぞれで 256，128 および 256 とした．また Fully connected layer から出力される特徴ベクトル $\mathbf{f}^P$ の次元数は 512 とした．詳細画像分類タスクにおいて Prototype 数は従来研究 [124, 93] に従い，CUB200-2011 データセットで 202 個，Stanford Cars データセットで 195 個とした．ただし画像検索タスクでは Prototype 数を 368 に設定し，Prototype layer における Head 数を 4 に設定した．また Stanford Dogs データセットでは Prototype 数および Head 数を 512 および 8 に設定した．学習における各損失関数の重みは詳細画像分類タスクでは $\lambda_{aux} = 1.0$，$\lambda_{clst} = 4.0$，$\lambda_{supp} = 0.1$ と設定し，画像検索タスクでは $\lambda_{aux} = 1.0$，$\lambda_{clst} = 0.8$，$\lambda_{supp} = 0.05$ と設定した．詳細画像分類タスクにおいて λ_{clst}，λ_{supp} に大きな値を設定する理由は Proxy Anchor loss の損失関数の

絶対値が他の損失関数と比較し大きな値となるためである．また Attention mask の閾値 t_{ram} は詳細画像分類タスクで 0.9，画像検索タスクで 0.99 に設定した．Proxy Anchor loss[67] や Margin loss[33], Multi-level distance regurarization (MDR) loss[89] におけるハイパーパラメータは論文値に従い設定した．本論文における Optimizer には Adam を採用し，モデルバックボーンの学習率は 1e-5，その他の層の学習率は 5e-4 とした．エポック数は画像検索タスクの実験設定では 150 とし，詳細画像分類タスクでは CUB200-2011 および Stanford Dogs で 150，Stanford Cars で 200 とした．ただし画像検索タスクでは学習前に Convolutional layer 以外の層を L_{task}，L_{aux} を用いて 20 エポックの間学習した．i-Naturalist2017 データセットで事前学習した Resnet50（iNR50）では十分良い特徴表現が獲得されており，長い学習時間を必要としない．そこで，モデルバックボーンとして iNR50 を用いる場合にはエポック数を 80 に設定した．いずれの実験設定においても Convolutional layer の学習開始直後の 10 エポックは L_{task} および L_{aux} のみで学習した．データ拡張は [93] に従い，RandomPerspective, ColorJitter，RandomHorizontalFlip，RandomAffine を用い，テスト時には画像サイズを 224×224 に Resize した．バッチサイズは 112 とし，バッチ内各クラスに 2 つのサンプルが含まれるようにした．本論文では 3 つの異なるシード (0, 1, 2) で学習を行い，学習終了時点のモデルを評価した結果の平均値を報告した．

5.2.2 詳細画像分類タスクへの適用

表 5.1，表 5.2 および表 5.3 に CUB200-2011，Stanford Cars および Stanford Dogs データセットにおける ProtoPNet 派生手法と ProtoMetric との比較結果を示す．ただし，ProtoMetric の Prototype 数を 512 とする場合には Prototype layer の Head 数を 4 とした．表 5.1 および表 5.2 より ProtoMetric は同一の Prototype 数の下で，全ての実験設定において ProtoPool を上回っていることが確認できる．これは従来の ProtoPNet 派生手法では sub-optimal な二段階の学習が必要であった一方で，ProtoMetric では提案する Cluster loss により end-to-end な学習を行うことが可能となったためと考えられる．また表 5.1，表 5.2 および表 5.3 の結果より ProtoMetric は，state-of-the-art 手法である Deformable ProtoPNet および Tesnet と比較し，全ての実験設定において少ない Prototype 数で同等以上の精度を達成していることが確認される．僅かに Tesnet と比較して精度劣化する場合がある理由としては，Tesnet は事前に定義された Prototype とクラスラベルの関係性を利用し，良い Prototype の特徴表現を獲得していることが挙げられる．クラスラベルと Prototype の関係性に依存しない，よりよい特徴表現の獲得は今後の課題である．

表 5.1: CUB200-2011 データセット [9] における提案手法と他の ProtoPNet 派生手法との比較結果．表には各モデルが必要とする Prototype 数（No. of proto.）と top-1 accuracy (Acc.) をまとめた．表内において，’iN’ は i-Naturalist 2017 データセット [40] で事前学習されたモデルを Convolutional layer に用いたことを表す．また表内では最も精度が高い手法を太字，二番目に精度が高い手法に下線を引き，提案手法を青字で表した．

Model	Arch.	No. of proto.	Acc.
ProtoMetric (Ours)	ResNet 34	202	81.4 %
ProtoPool[124]		202	80.3 %
ProtoPShare[97]		400	74.7 %
ProtoMetric (Ours)		512	82.6 %
ProtoPNet[45]		1655	78.6 %
TesNet[102]		2000	**82.8 %**
Def. ProtoPNet[116]		2000×2×2	81.1 %
ProtoMetric (Ours)	iN-ResNet 50	202	**87.7 %**
ProtoPool[124]		202	85.5 %
ProtoTree[93]		202	82.2 %
Def. ProtoPNet[116]		2000×2×2	85.6 %
ProtoMetric (Ours)	ResNet 152	202	**82.9 %**
ProtoPool[124]		202	81.5 %
ProtoPShare[97]		1000	73.6 %
ProtoPNet[45]		1734	79.2 %
TesNet[102]		2000	82.7 %
Def. ProtoPNet[116]		2000×2×2	82.0 %
ProtoMetric (Ours)	DenseNet 121	202	82.1 %
ProtoPool[124]		202	80.3 %
ProtoMetric (Ours)		512	84.7 %
ProtoPShare[97]		1000	76.5 %
ProtoPNet[45]		1734	80.2 %
TesNet[102]		2000	**84.8 %**
Def. ProtoPNet[116]		2000×2×2	82.6 %
ProtoMetric (Ours)	DenseNet 161	202	82.9 %
ProtoPool[124]		202	80.3 %
ProtoMetric (Ours)		512	84.5 %
ProtoPShare[97]		1000	76.5 %
ProtoPNet[45]		1734	80.1 %
TesNet[102]		2000	**84.6 %**
Def. ProtoPNet[116]		2000×2×2	83.3 %

表 5.2: Stanford Cars データセット [12] における提案手法と他の ProtoPNet 派生手法との比較結果. 表には各モデルが必要とする Prototype 数（No. of proto.）と top-1 accuracy (Acc.) をまとめた．また表内では最も精度が高い手法を太字，二番目に精度が高い手法に下線を引き，提案手法を青字で表した．

Model	Arch.	No. of proto.	Acc.
ProtoMetric (Ours)	ResNet 34	195	**91.4 %**
ProtoPool[124]		195	89.3 %
ProtoPShare[97]		480	86.4 %
ProtoPNet[45]		1960	88.8 %
TesNet[102]		1960	90.9 %
ProtoMetric (Ours)	ResNet 50	195	**92.0 %**
ProtoPool[124]		195	88.9 %
ProtoTree[93]		195	86.6 %
ProtoMetric (Ours)	DenseNet 121	195	91.0 %
ProtoPool[124]		195	86.4 %
ProtoMetric (Ours)		512	**92.5 %**
ProtoPShare[97]		980	84.8 %
ProtoPNet[45]		1960	87.7 %
TesNet[102]		1960	91.9 %

5.2.3 Ablation Study

本章ではまず Task loss L_{task} および Auxiliary loss L_{aux} についての Ablation study の結果について説明した後，Cluster loss L_{clst} における各コンポーネントの Ablation study の結果について説明する．

■ Task Loss および Auxiliary Loss に関する Ablation Study

Task loss および Auxiliary loss に関する Ablation Study の結果を表 5.4 に示す．結果より，Auxiliary loss を用いない場合には大幅な精度劣化が発生する事が確認される．また，Auxiliary loss として Proxy Anchor loss および Margin loss のいずれを用いても最終的な精度に大きく影響しないことが確認される．Auxiliary loss を導入しない場合における著しい精度劣化は，Attention layer における全結合層が学習されないために正確な Attention mask を獲得できないことが原因と考えられる．そこで Auxiliary loss は良い Attention mask を獲得するために必要であり，モデルの最終的な精度に大きく影響すると確認される．よって Auxiliary loss の有効性が確認されたと言える．また表 5.4 より，Task loss として Proxy Anchor loss を用いた場合には Margin Loss を用いた場合より高い精度を達成していることが確認される．この結果は Task loss は最終的なモデル性能に大きく影響し，精度向上には適切な損失関数の選択が必要であることを示唆している．そこで本論文では Task loss として Proxy Anchor loss を

表 5.3: Stanford Dogs[10] データセットにおける提案手法と他の ProtoPNet 派生手法との比較結果．表には各モデルが必要とする Prototype 数（No. of proto.）と top-1 accuracy (Acc.) をまとめた．また表内では最も精度が高い手法を太字，二番目に精度が高い手法に下線を引き，提案手法を青字で表した．

Model	Arch.	No. of proto.	Acc.
ProtoMetric (Ours)	VGG 19	512	**79.2 %**
ProtoPNet[45]		1200	73.6 %
Def. ProtoPNet[116]		1200×3×3	<u>77.9 %</u>
ProtoMetric (Ours)	ResNet 152	512	**87.7 %**
ProtoPNet[45]		1200	76.2 %
Def. ProtoPNet[116]		1200×3×3	<u>86.5 %</u>
ProtoMetric (Ours)	DenseNet 161	512	**85.5 %**
ProtoPNet[45]		1200	77.3 %
Def. ProtoPNet[116]		1200×3×3	<u>83.7 %</u>

採用することとした．また画像検索の実験設定において導入する正則化損失 [89] は Ranking ベースな手法を対象とする．そのため，本論文では Auxiliary loss に Margin loss を採用することとした．

■ Cluster Loss に関する Ablation Study

Cluster loss に関する Ablation Study の結果を表 5.5 にまとめる．表 5.5 の結果より，Cluster loss を適用しない場合 ('w/o Cluster loss') には，Prototype と最近傍画像パッチ特徴との平均コサイン類似度は低く，Prototype projection の前後で精度が大きく低下すると分かる．Prototype projection に伴う大幅な精度劣化は学習された Prototype が Prototype projection 前後で大きく変化したことを意味する．そのため，'w/o Cluster loss' では Prototype がある画像パッチを代表する特徴となるよう学習されたとは言えず，'This looks like that' フレームワークによる解釈可能性が成立していると言えない．また，各サンプルに対して最も類似度の高い Prototype と画像パッチ特徴とを近づける場合（'Take the most similar'）の結果は Cluster Loss を導入しない場合とほとんど変わらない．したがって'Take the most similar' においても'This looks like that' フレームワークによる解釈可能性が成立していないと言える．

一方 Prototype 所属度として，バッチ内サンプルの Prototype プロファイルを比較し，式 5.10 によるサンプリングの平均値を用いる場合（'Sampling (Eq. 5.10)'），最近傍の画像パッチ特徴との平均類似度は大きく向上し，Prototype projection 前後での精度差も小さくなる．これはバッチ内サンプルの違いに基づき Prototype 所属度を算出することの有効性を示唆している．加えて外部メモリを用いてデータセット全体を考慮したバイアス除去を行えば（'+ Debias with memory (Eq. 5.13)'）平均類似度はさらに増加し，Prototype projection 前後で精度差は無視できるほど小さくなる．これはサン

表 5.4: Task Loss（L_{task}）および Auxiliary Loss（L_{aux}）に関する Ablation Study の結果．Ablation Study は CUB200-2011[9] データセットにおいて異なる二つの Convolutional layer (R34 および iNR50) に対して実施した．表内で ResNet は'R' と略記される．また 'iN' は i-Naturalist 2017 データセット [40] において事前学習されたモデルであることを表す．

L_{task}	L_{aux}	R34	iNR50
Margin[33]	None	73.9 %	85.0 %
Margin[33]	Margin[33]	80.2 %	87.2 %
Margin[33]	Proxy-Anchor[67]	79.3 %	87.8 %
Proxy-Anchor[67]	None	79.0 %	87.0 %
Proxy-Anchor[67]	Margin[33]	81.1 %	87.7 %
Proxy-Anchor[67]	Proxy-Anchor[67]	**81.4 %**	**87.9 %**

表 5.5: Cluster Loss に関する Ablation Study の結果．本実験は i-Naturalist 2017 データセット [40] で事前学習された Resnet 50 を用いて CUB200-2011 データセット [9] において実施した．表内において'Avg. cossim.' は学習終了時における各 Prototype と学習データ内の画像パッチ特徴とのコサイン類似度の最大値の Prototype に関する平均値である．また，'Acc. before' および'Acc. after' は prototype projection 前後の top-1 accuracy [%] である．

Model	Acc. before	Acc. after	Avg. cossim
w/o Cluster loss	87.8 %	86.6 %	0.258
Take the most similar	87.7 %	85.9 %	0.293
Sampling (Eq. 5.10)	**88.3 %**	**87.9 %**	0.940
+ Debias with memory (Eq. 5.13)	87.7 %	87.6 %	0.980
+ Sample-aware modification (Eq. 5.17)	87.8 %	87.7 %	**0.983**

プリング結果の偏りを修正することで，全ての Prototype をある画像パッチ特徴と近づけることが可能となったためと考えられる．そこで，クラスと Prototype の関係性を事前定義することなしに'This looks like that' フレームワークによる解釈可能性を構築できたことが確認される．

さらに，Sample 毎に Prototype 所属度を修正すれば（'+ Sample-wise modification (Eq. 5.17)'），僅かに平均コサイン類似度及び精度が向上することが確認される．これは，画像内における物体の見え方の違いを考慮することで，より適切に Cluster loss を最適化出来るようになったためと考えられる．以上より L_{clst} における各コンポーネントの有効性が確認されたと言える．

5.2.4 定性的評価

図 5.5 および図 5.6 に CUB200-2011 および Stanford Cars データセットにおける提案手法のクラス分類予測に対する説明例を示す．図 5.5 および図 5.6 内の各行は 5.22 式の和における各項 $s_{a,j}C_j^{a,y}$ を

図 5.5: 詳細画像分類タスクにおいて CUB200-2011[9] データセットを用いた際の ProtoMetric の推論根拠の説明例．本実験では Convolutional layer として i-Naturalist 2017 データセット [40] で事前学習した Resnet 50 を採用した．‘Similarity’ および ‘weight’ 以下の数値は入力画像と Prototype との類似度及び 5.22 式で定義された $C_i^{y,a}$ を表す．

降順にソートした結果である．また，各行内の画像において矩形領域で囲まれた領域は，入力画像内で最も Prototype と類似する領域および Prototype である．ただし矩形領域の算出は従来手法 [45] に従った．Sec. 5.1.3 で説明したように，5.22 式の和において，大きな値をとる $s_{a,j}C_j^{a,y}$ に対応する Prototype が ProtoMetric のクラス予測に対する推論根拠となる．そこで，図 5.5 の例では，入力画像が’Green tailed towhee’ に分類された最大の理由は赤い頭頂部であり，このためクラススコアが 0.300 だけ増加したことが読み取れる．同様にして第 2，第 3 の推論根拠についても同様の解釈を行うことで定量的にクラス分類の推論根拠を解釈することが可能となる．以上より ProtoMetric は ProtoPNet 派生手法と同等の定量的なクラス予測の推論根拠を説明できることが確認される．

5.2.5　画像検索タスクへの適用

本章では画像検索タスクに ProtoMetric を適用した結果を説明する．5.1.2 節でも説明したように，画像検索タスクでは過学習に対処するため，学習済みの black-box モデルを教師モデルとして蒸留することで ProtoMetric を学習する．表 5.6 に教師モデルおよび蒸留した ProtoMetric の精度を示す．結

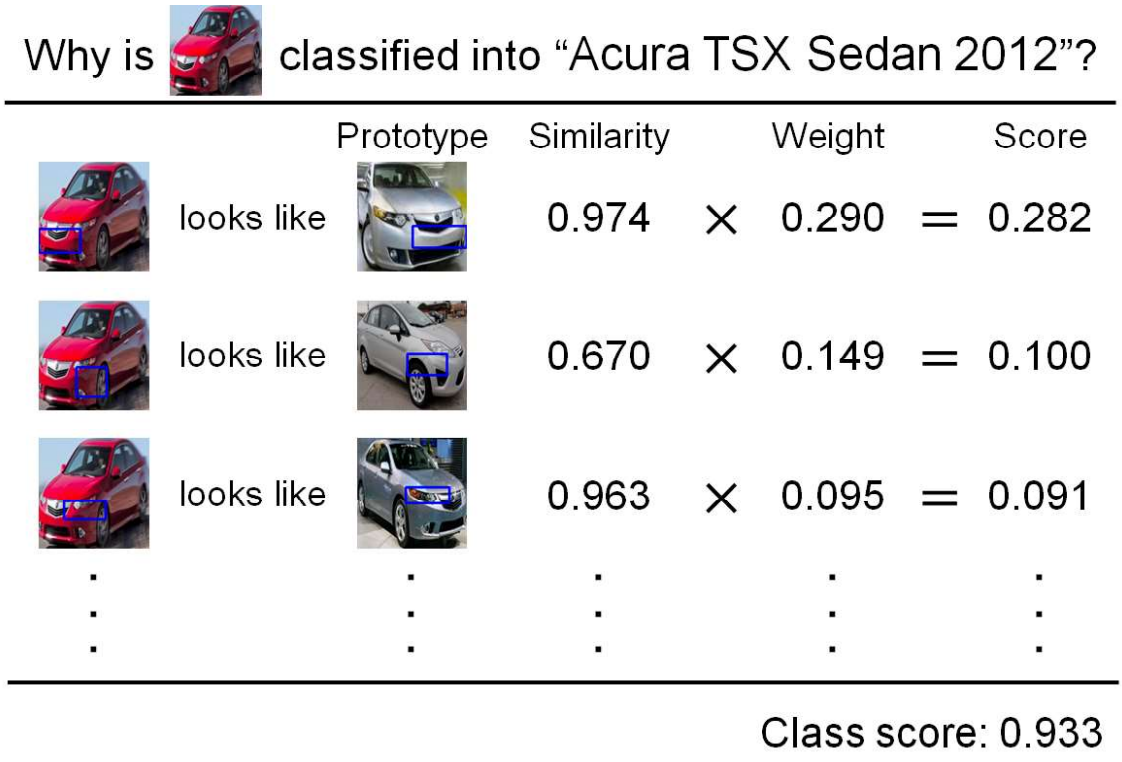

図 5.6: 詳細画像分類タスクにおいて Stanford Cars[12] データセットを用いた際の ProtoMetric の推論根拠の説明例．本実験では Convolutional layer として Imagenet[7] で事前学習した Resnet 50 を採用した．‘Similarity’ および ‘weight’ 以下の数値は入力画像と Prototype との類似度及び 5.22 式で定義された $C_i^{y,a}$ を表す．

果より，学習終了時点において ProtoMetric は蒸留元となる教師モデルとほとんど同等の精度を達成している事が確認できる．そこで，black-box モデルを蒸留する学習戦略により，画像検索タスクにおいて過学習することなく ProtoMetric を学習できることが確認される．本実験の目的は画像検索タスクに ProtoMetric を適用できると確認する事であり，他の深層距離学習手法と比較する事ではない点に注意されたい．他の画像検索手法との公平な比較 [72] は今後の課題である．

図 5.7 および図 5.8 に CUB200-2011 および Stanford Cars データセットにおける，画像検索結果に対する推論根拠の説明例を示す．図 5.7 および図 5.8 において，左端上の画像はクエリ画像であり，左端下の画像はクエリ画像と最も画像間類似度の大きいテストデータ内の画像（ギャラリ画像）である．また図内左端’Similarity’ 直下の数字はクエリ画像とギャラリ画像間の類似度を表している．図内右側の各行は 5.23 式の和における各項 $s_{a,j}R_{j,k}^{a,b}s_{b,k}$ を降順にソートした結果であり，各行内の画像内の矩形で囲まれた領域は画像内で最も Prototype と類似する領域および Prototype を表している．5.1.3 節で説明したように，5.23 式の和において，大きな値をとる $s_{a,j}R_{j,k}^{a,b}s_{b,k}$ に対応する Prototype のペアが画像間類似度に対する推論根拠となる．よって，図 5.7 の例では両画像においてある種のパターンを持つ羽根が存在することが二枚の画像が似ている最大の理由であり，このため 0.078 だけ画

表 5.6: 画像検索タスクにおける ProtoMetric の定量的評価．表内において ‘Teacher’ は知識蒸留における教師モデルとして用いた Black-box モデルである．表内の実験では Imagenet[7] で事前学習された Resnet 50 を Convolutional layer として採用し，CUB200-2011[9] および Stanford Cars[12] データセットを用いて実験を実施した．また，本実験では評価指標として Rank-1，Rank-2，Rank-4 accuracy（R1，R2，R4）に加え normalized mutual information（NMI）を採用した．

Model	CUB200-2011[9]				Stanford Cars[12]			
	R1	R2	R4	NMI	R1	R2	R4	NMI
Teacher[67]	70.5 %	80.3 %	87.3 %	0.727	90.6 %	94.2 %	96.6 %	0.766
ProtoMetric	70.2 %	79.7 %	86.7 %	0.715	90.3 %	94.3 %	96.5 %	0.763

像間類似度が増加したことが読み取れる．第二，第三の推論根拠についても同様にすれば定量的に画像間類似度の推論根拠を解釈することが可能となる．以上より ProtoMetric により，画像検索タスクにおける画像間類似度の’This looks like that’ フレームワークに基づく解釈が可能となることが確認される．

5.3 まとめ

本章では ProtoPNet を類似度学習に適用可能となるよう拡張した ProtoMetric を提案した．詳細画像認識タスクでは ProtoMetric は少ない Prototype 数で ProtoPNet 派生手法の state-of-the-art 手法と同等の精度を達成した．さらに本論文では case-study を通して ProtoMetric が画像検索タスクに適用可能であることを実験的に示した．筆者の知る限り，画像検索タスクにおいて’This looks like that’ フレームワークを構築した研究は本研究が初である．

本研究の Limitation には，画像検索タスクでは事前に訓練された教師モデルが必要となるために，学習の計算コストが大きくなる課題が挙げられる．教師モデルの蒸留は画像検索タスクにおける過学習を防ぎつつ，十分長い学習 iteration 数により Prototype を十分ある画像パッチに近づけるようにするため必要となる．画像検索タスクにおける過学習は未解決の open question であり，事前に学習された教師モデルを必要としない学習フレームワークの構築は今後の課題である．

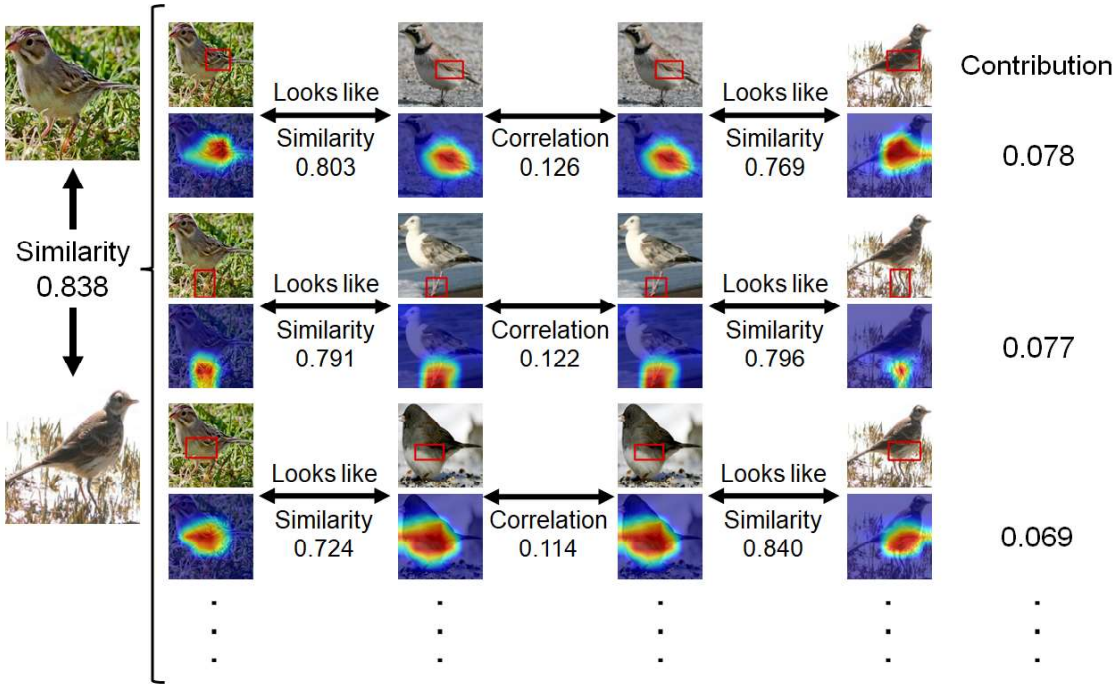

図 5.7: CUB200-2011[9] データセットを用いた際の画像検索タスクにおける ProtoMetric の推論根拠の説明例．本実験では Convolutional layer として Imagenet[7] で事前学習された Resnet 50 を採用した．‘Similarity’，‘Looks like’ および ‘Correlation’ 以下の数値は入力画像間のコサイン類似度，入力画像と Prototype との類似度，および 5.23 式で定義された Prototype 間の関連度 $K_{i,j}^{a,b}$ を表す．

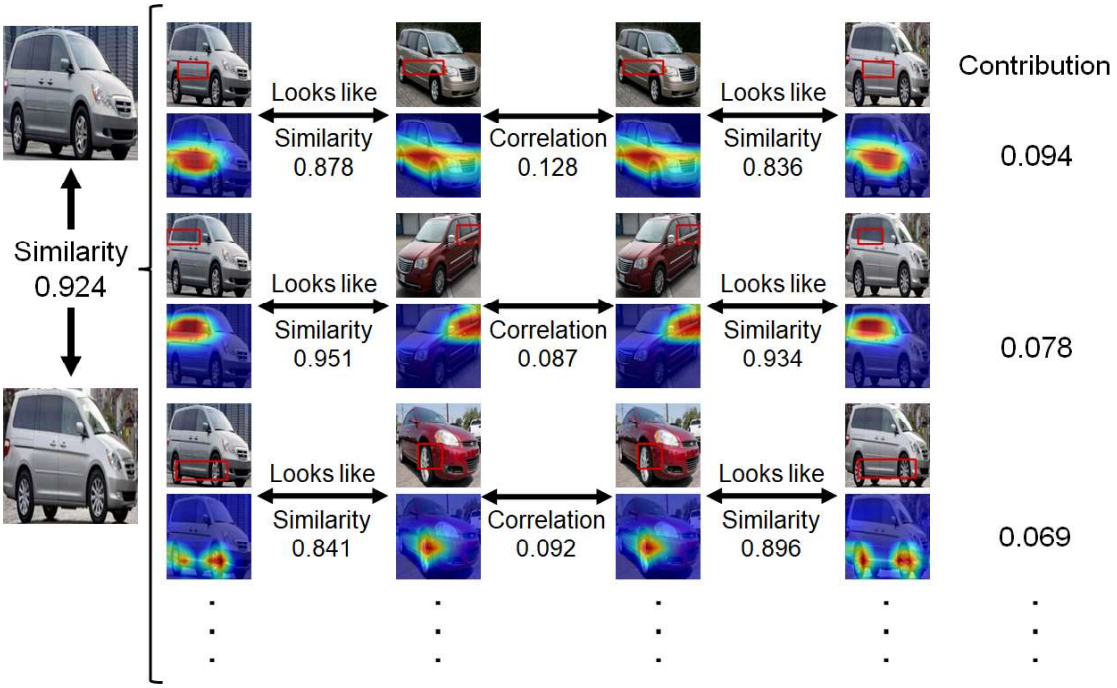

図 5.8: Stanford Cars[12] データセットを用いた際の画像検索タスクにおける ProtoMetric の推論根拠の説明例．本実験では Convolutional layer として Imagenet[7] で事前学習された Resnet 50 を採用した．‘Similarity’，‘Looks like’ および ‘Correlation’ 以下の数値は入力画像間のコサイン類似度，入力画像と Prototype との類似度，および 5.23 式で定義された Prototype 間の関連度 $K_{i,j}^{a,b}$ を表す．

第6章

結論と今後の展望

本論文では深層距離学習を実応用する上で課題となる二つの課題，ドメインシフトに伴う精度の劣化と解釈可能性の欠如の解決に取り組んだ．本論文では第一の課題に対し，局所的特徴と大域的特徴を考慮した新規の教師なしドメイン適応手法を提案した．また第二の課題に対しては本質的に解釈可能なモデルである ProtoPNet を類似度学習に適用可能となるよう拡張した ProtoMetric を提案した．以下に，本論文の結論と今後の展望を述べる．

6.1 結論

各章のまとめは以下のとおりである．2 章では本論文と密接に関連する深層距離学習，特に人物再同定における教師なしドメイン適応技術と深層学習の解釈可能性についてそれぞれ詳細に説明した．本論文では人物再同定における教師なしドメイン適応手法をソースドメインのデータをターゲットドメインのスタイルへ変換する生成画像を用いるアプローチと，クラスタリングによりターゲットドメインのデータへ疑似ラベルを付与する疑似ラベルを用いるアプローチに分類し，各アプローチを概説した．特に後者に関しては本論文と関連の深い大域的な特徴と局所的な特徴を用いた SSG[48] と ABMT[113] について詳細に説明した．また深層学習における解釈可能性の研究に関しては学習済みモデルを解析する Post-hoc な手法と本質的に解釈可能なモデルを構築する Ante-hoc な手法に大別し説明を行った．前者については更に Saliency-based な手法，Concept-based な手法，Surrogate Model を構築する手法に分類し，各手法の利点および欠点を取り上げながら各手法を体系的にまとめた．特に Post-hoc な手法では推論過程と説明が関連しないため，推論と何ら関係のない説明が生成される場合がある等，忠実性に課題を抱えていることを説明した．また Ante-hoc な手法については出力される説明の種類に応じて Attention-based な手法および Concept-based な手法に分類した．特に Concept-based の Ante-hoc な手法に関しては活発に研究のなされている Concept Bottleneck Model (CBM) [68] および Prototypical Part Network (ProtoPNet) [45] の派生手法を取り上げ，それらの利点および欠点について論じた．

3 章では人物再同定を対象として，局所的特徴と大域的特徴の両方を活用する教師なしドメイン適応手法を提案した．人物再同定タスクでは服の色などの画像内の大きな領域を占める大域的な特徴に加えて，時計や靴などの画像内の局所的な領域にのみ含まれる画像特徴が重要となる．しかし従来の教師なしドメイン適応手法では，GAP 出力により抽出される，画像の大域的な特徴のみを用いて学習が行われており，局所的な特徴が考慮されていない課題があった．この課題に対し，本論文で

は GAP 出力に加え，画像内の局所的特徴を抽出する特性の有る GMP 出力により抽出された各特徴を組み合わせて教師なしドメイン適応を行う手法を提案した．より具体的には GAP および GMP 出力をそれぞれ個別にクラスタリングして得られる疑似ラベルに加え，それらの積集合セットを用いて学習を行う手法を提案した．これより，提案手法は複数の公開された人物再同定データセットにおいて従来手法と比較して高い精度を達成できることを確認した．

4 章では ‘Gray-box’ モデルの中で特に注目を集めている ProtoPNet による ‘This looks like that’ フレームワークの従来手法を説明した．本論文では従来の ProtoPNet 派生手法をクラスに固有の Prototype を学習する手法とクラス間で共通した Prototype を学習する手法に分類した．前者の手法は学習の結果得られる Prototype の表現を改善することにより大幅な性能改善が達成されてきたことを説明し，後者の手法ではそれぞれ個別の効果的な学習フレームワークによりクラス間で共通した Prototype を学習することで，大幅なメモリ効率の改善が達成されることを説明した．一方で，これらの ProtoPNet 派生手法には学習データとテストデータで異なるクラスラベルをもつサンプルを対象とする類似度学習に適用することが難しい課題があることが確認された．

5 章では類似度学習に適用可能な ProtoPNet の拡張手法である ProtoMetric を提案した．ProtoPNet を類似度学習に適用することが難しい課題に対し，本論文ではサンプル間の比較によりクラスラベルと Prototype の関係性を推定しながら，その推定に基づき画像パッチ特徴と Prototype を近づける新規の Cluster Loss を提案した．詳細画像認識タスクにおける評価実験では，ProtoMetric は少ない Prototype 数で従来の ProtoPNet 派生手法における state-of-the-art と同等の精度を達成することを確認した．また，case-study を通して従来の ProtoPNet 派生手法を適用することが難しい画像検索タスクにおいても ProtoMetric により ‘This looks like that’ フレームワークに基づく解釈可能性が実現されることを確認した．一方で画像検索タスクにおける ProtoMetric の学習には学習済みの良い Black-box モデルを必要とする課題が残されている．これは画像検索タスクでは過学習が課題となる一方，Prototype の学習には十分長い学習を必要とすることから，本論文では Black-box モデルの蒸留を目的関数とすることで過学習を回避する戦略を採用したためである．画像検索タスクにおいていかに過学習を抑制するかは深層距離学習における open-question であり，今後の課題である．

6.2 展望

本論文では深層距離学習における二つの課題，ドメインシフトによる性能低下と解釈可能性の欠如に取り組んだ．本研究で提案した ProtoMetric は従来の ProtoPNet 派生手法と異なり，学習データとテストデータでクラスの異なる画像検索タスクに適用することが出来る．そこで ProtoMetric と本論文で提案した教師なしドメイン適応手法を組み合わせることにより，適用先となるドメインにおいて教師ラベルを作成することなく解釈可能な深層距離学習モデルが構築出来ると期待される．ProtoMetric は筆者の知る限り，クラスラベル以上の教示を必要としない，唯一の解釈可能な深層類似度学習モデルである．従って学習データとテストデータにおけるクラスラベルが異なる，open-set な画像認識タスクへの更なる ‘This looks like that’ フレームワークの応用が期待される．加えて，従来研究が十分なされていない，画像検索タスクにおけるモデル推論過程の理解やモデルデバッグに関する研究

が期待される.

謝　辞

本論文はグローリー株式会社研究開発センターおよび中部大学工学部機械知覚＆ロボティクスグループにおいて，多くの方々のご協力のもと行った研究成果をまとめたものである．

研究を進める上での貴重なアドバイスから論文の添削まで終始懇切丁寧なご指導とご助言をを頂きました中部大学工学部ロボット理工学科藤吉弘亘教授，同大学工学部情報工学科山下隆義教授，同大学 AI 数理データサイエンスセンター平川翼講師に深謝いたします．また他機関の優秀な研究者とのディスカッションの機会や国際会議における発表等，博士課程学生として多くの貴重な機会を与えてくださいましたことに改めて感謝申し上げます．ご多忙にも関わらず副査を快く引き受けて頂き，有益なご助言を賜りました中部大学工学部ロボット理工学研究科梅崎太造教授，九州大学大学院システム情報科学研究院情報知能工学部門内田誠一教授に謹んで感謝申し上げます．中部大学藤吉研究室宮腰あゆみさんには事務の面で多くのご助力を頂きました．また，土曜日のディスカッションを通して多くの知見を与えてくださった機械知覚＆ロボティクスグループの皆様に感謝申し上げます．

中部大学へ社会人博士として派遣していただき，本研究を長年にわたり強力に支援して下さったグローリー株式会社研究開発センターの亀山博史研究開発センター長に深謝いたします．また画像認識の研究を始めるきっかけならびに研究を進める上での良きアドバイスを数多くくださった大西弘之博士に深謝いたします．直接の上司として本研究を支援して下さった新技術創発部の福本一夫元部長，黍原文雄部長，安達和隆グループマネージャー，米澤亨研究ラボラトリー長に感謝申し上げます．事業化推進部の野村則夫元部長，森藤健部長，中嶋康博グループマネージャーには技術をどのように事業へとつなげていくかという観点で多くの御助言を頂きました．また，社会人博士として研究を進める上でご協力頂きました研究開発センターの皆さまに感謝申し上げます．

最後に本研究の場ならびに機会を与えてくださり，さらに本論文をまとめる機会を与えてくださった三和元純社長に深謝いたします．

参考文献

[1] L. S. Shapley, “Notes on the n-person game—ii:,” in *The value of an n-person game*, Santa Monica, CA, USA: RAND Corporation, 1951. [Online]. Available: https://www.rand.org/pubs/research_memoranda/RM0670.html.

[2] R. Sinkhorn, “Diagonal equivalence to matrices with prescribed row and column sums,” *The American Mathematical Monthly*, vol. 74, no. 4, pp. 402—405, 1967.

[3] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proc. Int. Conf. KnowledgeDiscovery and Data Mining (KDD)*. 1996, pp.226–231.

[4] A. Strehl and J. Ghosh, “Cluster ensembles – a knowledge reuse framework for combining multiple partitions,” in Journal of Machine Learning Research, vol.3, pp.583–617, 2002.

[5] D. Gray, S. Brennan, and H. Tao, “Evaluating appearance model for recognition, reacquision, and tracking,” in *IEEE Int. Workshop on Perform. Eval. of Tracking and Surveillance*, 2007

[6] A. Rosenberg and J. Hirschberg, “V-measure: a conditional entropy-based external cluster evaluation measure,” in *Proc. Joint Conf. Empirical Methods in Natural Lang. Process. and Comput. Natural Lang. Learn.*, 2007 , pp.410-420.

[7] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. “Imagenet: a large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.

[8] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, S. Diego, and P. Perona, “Caltech-ucsd birds 200,” *California Institute of Technology*, CNS-TR-201, 2010.

[9] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The caltech-ucsd birds-200-2011 dataset,” Comput. Neural Syst. Tech. Rep., 2011.

[10] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, “Novel dataset for fine-grained image categorization: Stanford dogs,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop on fine-grained vis. categorization (FGVC)*, vol. 2, no. 1, 2011.

[11] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transportation distances,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2013, pp. 2292-–2300.

[12] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3D object representations for fine-grained categorization,” in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2013, pp. 554—561.

[13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 770–778.

[14] P. Kontschieder, M. Fiterau, A. Criminisi, and S. R. Bulo. “Deep neural decision forests,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1467–1475.

[15] A. Mahendran and A. Vedaldi. “Understanding deep image representations by inverting them,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 5188–5196.

[16] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1116–1124.

[17] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. “Machine bias.” ProPublica; 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[18] I. Goodfellow, Y. Bengio, and A. Courville, “Deep learning,” MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org. Accessed: Feb 15, 2023.

[19] M. T. Ribeiro, S. Singh, C. Guestrin. “Why should I trust you?: Explaining the predictions of any classifier” in *Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135—1144.

[20] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in *Proc. Eur. Conf. Comput. Vis. Workshop (ECCVW)*, 2016, pp. 17–35.

[21] K. Sohn, “Improved deep metric learning with multi-class n-pair loss objective,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1857—1865.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826.

[23] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2921–2929.

[24] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, ”Scalable person re-identification: A benchmark,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1116–1124.

[25] J. Altschuler, J. Weed, and P. Rigollet, "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp.1961—1971.

[26] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. "Network dissection: Quantifying interpretability of deep visual representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6541–6549.

[27] W. Chen, X. Chen, J. Zhang, and K. Huang. "Beyond triplet loss: A deep quadruplet network for person reidentification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1320–1329.

[28] C. Olah, A. Mordvintsev, and L. Schubert. "Feature Visualization" Distill, 2017. [Online]. Available: https://distill.pub/2017/feature-visualization

[29] A. Hermans, L. Beyer, and B. Leibe. "In defense of the triplet loss for person reidentification," arXiv preprint arXiv:1703.07737, 2017.

[30] S. M. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 4765–4774.

[31] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh, "No fuss distance metric learning using proxies," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 360–368.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.

[33] C. Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2840—2848.

[34] C. Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl. "Sampling matters in deep embedding learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2840–2848.

[35] H. Zheng, J. Fu, T. Mei, and J. Luo "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 5209–5217.

[36] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Reranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1318–1327.

[37] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 7786—7795.

[38] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. “Sanity checks for saliency maps,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 9525—9536.

[39] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, “Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 994–1003.

[40] G. V. Horn, O. M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. “The inaturalist species classification and detection dataset,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8769–8778.

[41] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres. “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav),” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 2673–2682.

[42] V. Petsiuk, A. Das, and K. Saenko. “Rise: Randomized input sampling for explanation of black-box models.” in *Proc. British Mach. Vis. Conf. (BMVC)*, 2018, pp. 1-17.

[43] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu. “CosFace: Large Margin Cosine Loss for Deep Face Recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 5265–5274.

[44] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person transfer gan to bridge domain gap for person reidentification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 79–88.

[45] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, “This looks like that: Deep learning for interpretable image recognition,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 8928–8939.

[46] J. Deng, J. Guo, N. Xue and S. Zafeiriou. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4685–4694.

[47] Y. Deng, X. Lin, R. Li, and R. Ji, “Multi-scale gem pooling with n-pair center Loss for fine-grained image search,” in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2019, pp. 1000–1005.

[48] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. S. Huang, “Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person reidentification,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 6112–6121.

[49] H. Fukui, T. Hirakawa, T. Yamashita, H. Fujiyoshi “Attention branch network: Learning of attention mechanism for visual explanation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 10705–10714.

[50] A. Ghorbani, J. Wexler, J. Zou, and B. Kim "Towards automatic concept-based explanations," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 9273–9282.

[51] Y. Huang, Q. Wu, J. Xu, and Y. Zhong, "SBSGAN: Suppression of inter-domain background shift for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 9527–9536.

[52] W. Kool, H. van Hoof, and M. Welling, "Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3499–3508.

[53] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019, pp. 1487-1495.

[54] Y. Ming, P. Xu, H. Qu, and L. Ren "Prosenet: Interpretable and steerable sequence learning via prototypes," in *Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2019.

[55] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3967–3976.

[56] Q. Qian, L. Shang, B. Sun, J. Hu, H. Li, and R. Jin, "Softtriple loss: Deep metric learning without triplet sampling," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6450–6458.

[57] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[58] X. Sun and L. Zheng. "Dissecting person re-identification from the viewpoint of viewpoint," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 608–617.

[59] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, 2019, pp. 5022–5030.

[60] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 1389–1398.

[61] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification. " in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 598-607.

[62] H. Chang, D. Zhao, C. H. Wu, L. Li, N. Si, and R. He. "Visualization of spatial matching features during deep person reidentification," in Journal of Ambient Intelligence and Humanized Computing, 2020.

[63] G. Chen, Y. Lu, J. Lu, and J. Zhou, "Deep credible metric learning for unsupervised domain adaptation person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 643–659.

[64] Y. Ge, D. Chen, and H. Li, "Mutual mean teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," in *Int. Conf. Learn. Representation (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=rJlnOhVYPS

[65] Y. Ge, F. Zhu, D. Chen, R. Zhao, and H. Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-ID," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 11309—11321.

[66] Z. Ji, X. Zou, X. Lin, X. Liu, T. Huang, and S. Wu, "An attention-driven two-stage clustering method for unsupervised person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 20–36.

[67] S. Kim, D. Kim, M. Cho, and S. Kwak, "Proxy anchor loss for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3238–3247.

[68] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 5338–5348.

[69] J. Li and S. Zhang, "Joint visual and temporal consistency for unsupervised domain adaptive person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 483–499.

[70] M. Li, K. Kuang, Q. Zhu, X. Chen, Q. Guo, and F. Wu. "Ib-m: A flexible framework to align an interpretable model and a black-box model," in *Proc. Int. Conf. Bioinf. Biomed. (BIBM)*, 2020, pp. 643–649.

[71] D. Mekhazni, A. Bhuiyan, G. Ekladious, and E. Granger, "Unsupervised domain adaptation in the dissimilarity space for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 159–174.

[72] K. Musgrave, S. Belongie, and S. Lim "A metric learning reality check," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 681–699.

[73] K. Roth, T. Milbich, S. Sinha, P. Gupta, B. Ommer, and J. P. Cohen, "Revisiting training strategies and generalization performance in deep metric learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 8242–8252.

[74] C. Rudin, C. Wang, and B. Coker. "The age of secrecy and unfairness in recidivism prediction." Harvard Data Science Review, vol. 2, no. 1, 2020. https://doi.org/10.1162/99608f92.6ed64b30.

[75] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10981–10990.

[76] D. Wang and S. Zhang, "Unsupervised person re-identification via multi-label classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10981–10990.

[77] Y. Wang, S. Liao, and L. Shao. "Surpassing real-world source training data: Random 3d characters for generalizable person re-identification," in *28th ACM Int. Conf. Multimedia (ACMMM)*, 2020, pp.3422—3430.

[78] Z. Yunpeng, Y. Qixiang, L. Shijian, J. Mengxi, J. Rongrong, and T. Yonghong, "Multiple expert brainstorming for domain adaptive person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 594–611.

[79] F. Zhao, S. Liao, G. Xie, J. Zhao, K. Zhang, and L. Shao, "Unsupervised domain adaptation with noise resistible mutual-training for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 526–544.

[80] Y. Zou, X. Yang, Z. Yu, B. V. K. V. Kumar, and J. Kautz, "Joint disentangling and adaptation for cross-domain person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 87–104.

[81] A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, and C. Rudin. "A case-based interpretable deep learning model for classification of mass lesions in digital mammography," in *Nature Mach. Intell.*, vol 3, pp. 1061–1070, 2021.

[82] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond, "Joint Generative and Contrastive Learning for Unsupervised Person Re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 2004–2013.

[83] X. Chen, X. Liu, W. Liu, X. P. Zhang, Y. Zhang, T. Mei "Explainable person re-identification with attribute-guided metric distillation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 11813–11822.

[84] H. Feng, M. Chen, J. Hu, D. Shen, H. Liu, and D. Cai, "Complementary pseudo labels for unsupervised domain adaptation on person re-identification," in IEEE Trans. Image Process. , vol.30, pp.2898–2907, 2021.

[85] D. Fu, D. Chen, J. Bao, H. Yang, L. Yuan, L. Zhang, H. Li, and D. Chen. “Unsupervised pre-training for person re-identification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 14750–14759.

[86] Y. Ge, Y. Xiao, Z. Xu, M. Zheng, S. Karanam, T. Chen, L. Itti, and Z Wu. “A peek into the reasoning of neural networks: interpreting with structural visual concepts,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 2195–2204.

[87] T. Isobe, D. Li, L. Tian, W. Chen, Y. Shan, and S. Wang, “Towards discriminative representation learning for unsupervised person re-identification,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 8526–8536.

[88] E. Kim, S. Kim, M. Seo, and S. Yoon. “Xprotonet: Diagnosis in chest radiography with global and local explanations,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 15719–15728.

[89] Y. Kim and W. Park, “Multi-level distance regularization for deep metric learning,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 1827–1835.

[90] K. Kobs, M. Steininger, A. Dulny, and A. Hotho, “Do different deep metric learning losses lead to similar learned features?,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10644–10654.

[91] S. Kraft, K. Broelemann, A. Theissler, and G. Kasneci. “Sparrow: Semantically coherent prototypes for image classification,” in *Proc. British Mach. Vis. Conf. (BMVC)*, 2021, pp. 1-12.

[92] M. Mitsuhara, H. Fukui, Y. Sakashita, T. Ogata, T. Hirakawa, T. Yamashita, and H. Fujiyoshi “Embedding human knowledge into deep neural network via attention map” in *Proc. Int. Joint Conf. Comput. Vis. Imag. Comput. Graph. Theory and Appl. (VISSIGRAPP)*, 2021, pp. 626-636.

[93] M. Nauta, R. van Bree, and C. Seifert, “Neural prototype trees for interpretable fine-grained image recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 14933–14943.

[94] M. Nauta, A. Jutte, J. Provoost, and C. Seifert “This looks like that, because... Explaining prototypes for interpretable image recognition,” in *Proc. Joint Eur. Conf. Mach. Learn. and Knowl. Discovery in Databases (ECML PKDD)*, 2021, pp. 441-–456.

[95] J. Ni, Z. Chen, W. Cheng, B. Zong, D. Song, Y. Liu, X. Zhang, H. Chen. “Interpreting convolutional sequence model by learning local prototypes with adaptation regularization,” in *Proc. ACM Int. Conf. Information and Knowl. Manage. (CIKM)*, 2021, pp. 1366–1375.

[96] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[97] D. Rymarczyk, Ł. Struski, J. Tabor, and B. Zieliński, "Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification," in *Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2021, pp. 1420—1430.

[98] S. Santurkar, D. Tsipras, M. Elango, D. Bau, A. Torralba, and A. Madry. "Editing a classifier by rewriting its prediction rules," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 1–15.

[99] W. Shen, Q. Ren, D. Liu, Q. Zhang. "Interpreting representation quality of dnns for 3d point cloud processing," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 8857–8870.

[100] G. Singh and K. C. Yow. "These do not look like those: An interpretable deep learning model for image recognition" in *IEEE Access*, vol. 9, pp. 41482-41493, 2021, doi: 10.1109/ACCESS.2021.3064838.

[101] L. Trinh, M. Tsang, S. Rambhatla, and Y. Liu. "Interpretable and trustworthy deepfake detection via dynamic prototypes," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2021, pp. 1973–1983.

[102] J. Wang, H. Liu, X. Wang, and L. Jing, "Interpretable image recognition by constructing transparent embedding space," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 895–904.

[103] R. Wang, X. Wang, and D. I. Inouye, "Shapley Explanation Networks," in *Int. Conf. Learn. Representation (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=vsU0efpivw

[104] F. Zhang, M. Li, G. Zhai, and Y. Liu "Multi-branch and multi-scale attention learning for fine-grained visual categorization" in *Proc. Int. Conf. MultiMedia Modeling (MMM)*, 2021, pp.136-147.

[105] T. Zhang, L. Xie, L. Wei, Z. Zhuang, Y. Zhang, B. Li, Q. Tian. "Unrealperson: An adaptive pipeline towards costless person re-identification" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 11506–11515.

[106] Y. Zhang, A. Khakzar, Y. Li, A. Farshad, S. T. Kim, and N. Navab, "Fine-grained neural network explanation by identifying input features with predictive information," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 20040–20051.

[107] W. Zhao, Y. Rao, Z. Wang, J. Lu, J. Zhou. "Towards interpretable deep metric learning with structural matching," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 9887–9896.

[108] K. Zheng, C. Lan, W. Zeng, Z. Zhang, and Z. Zha, "Exploiting sample uncertainty for domain adaptive person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, pp. 3538–3546.

[109] S. Zhu, T. Yang, and C. Chen. "Visual explanation for deep metric learning," in *IEEE Trans. on Image Process. (TIP)*, vol. 30, pp. 7593-7607, 2021, doi: 10.1109/TIP.2021.3107214.

[110] P. Barbiero, G. Ciravegna, F. Giannini, P. Lió, M. Gori, and S. Melacci, "Entropy-based Logic Explanations of Neural Networks," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2022, pp. 6046–6054.

[111] M. Böhle and M. Fritz, and B. Schiele, "B-cos networks: Alignment is all we need for interpretability," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10329–10338.

[112] C. H. Chang, R. Caruana, and A. Goldenberg, "Node-gam: Neural generalized additive model for interpretable deep learning," in *Int. Conf. Learn. Representation (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=g8NJR6fCCl8

[113] H. Chen, B. Lagadec, and F. Bremond, "Enhancing diversity in teacher-student networks via asymmetric branches for unsupervised person re-identification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2021, pp. 1–10.

[114] K. Chen, W. Chen, T. He, R. Du, F. Wang, X. Sun, Y. Guo, G. Ding "Tagperson: A target-aware generation pipeline for person re-identification," in *30th ACM Int. Conf. Multimedia (ACMMM)*, 2022, pp. 560—571.

[115] E. Dai and S. Wang. "Towards prototype-based self-explainable graph neural network," arXiv preprint arXiv:2210.01974, 2022.

[116] J. Donnelly, A. J. Barnett, and C. Chen, "Deformable protopnet: an interpretable image classifier using deformable prototypes," in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, 2022, pages 10265–10275.

[117] D. Fu, D. Chen, H. Yang, J. Bao, L. Yuan, L. Zhang, H. Li, F. Wen, D. Chen "Large-scale pre-training for person re-identification with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 2476–2486.

[118] M. Hamilton, S. Lundberg, S. Fu, L. Zhang, and W. T. Freeman, "Axiomatic explanations for visual search, retrieval, and similarity learning," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=TqNsv1TuCX9

[119] T. Jing, H. Xia, R. Tian, H. Ding, X. Luo, J. Domeyer, R. Sherony, and Z. Ding. "Inaction: Interpretable action decision making for autonomous driving," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 370–387.

[120] S. S. Y. Kim, N. Meister, V. V. Ramaswamy, R. Fong, and O. Russakovsky. “Hive: Evaluating the human interpretability of visual explanations,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 280–298.

[121] B. A. Plummer, M. I. Vasileva, V. Petsiuk, K. Saenko, and D. Forsyth . “Why do these match? Explaining the behavior of image similarity models,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 652–669.

[122] R. Ragodos, T. Wang, Q. Lin, and X. Zhou, “Protox: Explaining a reinforcement learning agent via prototyping,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022. [Online]. Available: https://openreview.net/forum?id=nyBJcnhjAoy

[123] D. Rymarczyk, A. Pardyl, J. Kraus, A. Kaczyńska, M. Skomorowski, and B. Zieliński. “Protomil: Multiple instance learning with prototypical parts for whole-slide image classification,” in *Proc. Joint Eur. Conf. Mach. Learn. and Knowl. Discovery in Databases (ECML PKDD)*, 2022, pp. 421-–436.

[124] D. Rymarczyk, Ł. Struski, M. Górszczak, K. Lewandowska, J. Tabor, and B. Zieliński, “Protopool: interpretable image classification with differentiable prototypes assignment,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 351–368.

[125] Y. Wang, X. Liang, S. Liao. “Cloning outfits from real-world images to 3d characters for generalizable person re-identification” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 4900–4909.

[126] Z. Yang, M. Bastan, X. Zhu, D. Gray, and D. Samaras, “Hierarchical Proxy-Based Loss for Deep Metric Learning,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 1859–1868.

[127] Z. Zhang, Q. Liu, H. Wang, C. Lu, and C. Lee, “ProtGNN: Towards self-explaining graph neural networks,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2022, pp.9127–9135.

[128] S. Azzolin, A. Longa, P. Barbiero, P. Liò, and A. Passerini. “Global explainability of gnns via logic combination of learned concepts,” in *Int. Conf. Learn. Representation (ICLR)*, 2023. [Online]. Available: https://openreview.net/forum?id=OTbRTIY4YS.

[129] H. Ayoobi, N. Potyka, and F. Toni. “Protoargnet: Interpretable image classification with super-prototypes and argumentation,” arXiv preprint arXiv:2311.15438, 2023.

[130] A. Bontempelli, S. Teso, K. Tentori, F. Giunchiglia, and A. Passerini, “Concept-level debugging of part-prototype networks,” in *Int. Conf. Learn. Representation (ICLR)*, 2023. [Online]. Available: https://openreview.net/forum?id=oiwXWPDTyNk.

[131] M. Dani, I. Rio-Torto, S. Alaniz, Z. Akata. “Devil: Decoding vision features into language,” in *Proc. German Conf. Pattern Recognit. (GCPR)*, 2023.

[132] A. Dravid, Y. Gandelsman, A. A. Efros, and A Shocher. “Rosetta neurons: Mining the common units in a model zoo,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 1934–1943.

[133] J. Gao, X. Ma, and C. Xu. “Learning transferable conceptual prototypes for interpretable unsupervised domain adaptation”, arXiv preprint arXiv:2310.08071, 2023.

[134] S. Gautam, M. M.-C. Höhne, S. Hansen, R. Jenssen, and M. Kampffmeyer. “This looks more like that: Enhancing self-explaining models by prototypical relevance propagation,” in *Pattern Recognition*, 136:109172, 2023.

[135] S. Gulshad, T. Long, N. van Noord. “Hierarchical explanations for video action recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit. Workshop (CVPRW)*, 2023, pp. 3703–3708.

[136] R. Hesse, S. Schaub-Meyer, and S. Roth. “Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 3981–3991.

[137] Q. Huang, M. Xue, W. Huang, H. Zhang, J. Song, Y. Jing, and M. Song. “Evaluation and improvement of interpretability for self-explainable part-prototype networks,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 2011–2020.

[138] N. Kalibhat; S. Bhardwaj; C. B. Bruss, H. Firooz, M. Sanjabi, and S. Feizi. “Identifying interpretable subspaces in image representations,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 15623–15638.

[139] E. M. Kenny, M. Tucker, and J. Shah. “Towards interpretable deep reinforcement learning with human-friendly prototypes,” in *Int. Conf. on Learn. Representation (ICLR)*, 2023. [Online]. Available: https://openreview.net/forum?id=hWwY_Jq0xsN

[140] S. Kim, J. Oh, S. Lee, S. Yu, J. Do, T. Taghavi. “Grounding counterfactual explanation of image classifiers to textual concept space,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 10942–10950.

[141] C. Ma, B. Zhao, C. Chen, and C. Rudin. “This looks like those: Illuminating prototypical concepts using multiple visualizations,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023. [Online]. Available: https://openreview.net/forum?id=dCAk9VlegR

[142] M. Nauta, J. Schlötterer, M. Keulen, and C. Seifert. “Pip-net: Patch-based intuitive prototypes for interpretable image classification” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 2744–2753.

[143] R. Netzorg, J. Li, and B. Yu "Improving prototypical part networks with reward reweighing, reselection, and retraining," arXiv preprint arXiv:2307.03887, 2023.

[144] T. Oikarinen, S. Das, L. M. Nguyen, and T. W. Weng. "Label-free concept bottleneck models," in *Int. Conf. on Learn. Representation (ICLR)*, 2023. [Online]. Available: https://openreview.net/forum?id=FlCg47MNvBA

[145] T. Oikarinen, T. W. Weng. "Clip-dissect: automatic description of neuron representations in deep vision networks," in *Int. Conf. Learn. Representation (ICLR)*, 2023. [Online]. Available: https://openreview.net/forum?id=iPWiwWHc1V.

[146] V. V. Ramaswamy, S. S. Y. Kim, R. Fong, and O. Russakovsky. "Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 10932–10941.

[147] D. Rymarczyk, D. Dobrowolski, and T. Danel. "Progrest: Prototypical graph regression soft trees for molecular property prediction," in *Proc. SIAM Int. Conf. Data Mining (SDM)*, 2023, pp. 379–387.

[148] D. Rymarczyk, J. van de Weijer, B. Zieliński, and B. Twardowski. "Icicle: Interpretable class incremental continual learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 1887–1898.

[149] M. Sacha, D. Rymarczyk, Ł. Struski, J. Tabor, and B. Zieliński, "Protoseg: Interpretable semantic segmentation with prototypical parts," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2023, pp. 1481–1492.

[150] D. Srivastava, T. Oikarinen, and T. Weng. "Corrupting neuron explanations of deep visual features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 1877–1886.

[151] C. Wang, Y. Liu, Y. Chen, F. Liu, Y. Tian, D. J. McCarthy, H. Frazer, and G. Carneiro. "Learning support and trivial prototypes for interpretable image classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 2062–2072.

[152] C. Wang, Y. Chen, F. Liu, D. J. McCarthy, H. Frazer, and G. Carneiro. "Mixture of gaussian-distributed prototypes with generative modelling for interpretable image classification," arXiv preprint arXiv:2312.00092, 2023.

[153] X. Xu, Z. Wang, C. Deng, H. Yuan, and S. Ji. "Towards improved and interpretable deep metric learning via attentive grouping," in *IEEE Trans. Patt. Anal. and Mach. Intell. (TPAMI)*, vol. 45, no. 1, pp. 1189-1200, 1 Jan. 2023, doi: 10.1109/TPAMI.2022.3152495.

[154] Y. Yang, A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch, and M. Yatskar, "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 19187–19197.

[155] Y. Yu, S. Buchanan, D. Pai, T. Chu, Z. Wu, S. Tong, B. D. Haeffele, Y. Ma "White-box transformers via sparse rate reduction," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023. [Online]. Available: https://openreview.net/forum?id=THfl8hdVxH#

[156] Y. Yu, T. Chu, S. Tong, Z. Wu, D. Pai, S. Buchanan, and Y. Ma "Emergence of segmentation with minimalistic white-box transformers," arXiv preprint arXiv:2308.16271, 2023.

[157] M. Yuksekgonul, M. Wang, J. Zou "Post-hoc concept bottleneck models," in *Int. Conf. on Learn. Representation (ICLR)*, 2023. [Online]. Available: https://openreview.net/forum?id=nA5AZ8CEyow

研究業績一覧

学術論文

[1] 鵜飼　祐生, 藤吉　弘亘, “人物再同定における教師なしドメイン適応への大域・局所特徴の利用”, 電子情報通信学会論文誌, vol. J106-D, no. 3, pp. 195-206, 2023.

[2] Yuki Ukai, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi, “Toward prototypical part interpretable similarity learning with protometric,” in *IEEE Access*, vol. 11, pp. 62986-62997, 2023, doi: 10.1109/ACCESS.2023.3287638.

国際会議発表論文(査読あり)

[1] Yuki Ukai, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi, “This looks like it rather than that: Protoknn for similarity-based classifier,” in *The Eleventh International Conference on Learning Representation (ICLR)*, 2023.

付 録A

最適輸送問題とSinkhorn-Knoppアルゴリズム

本章では Cuturi によって提案された最適輸送問題の高速な近似解法である Sinkhorn-Knopp 法 [11] について説明する．最適輸送問題は下式のように定式化される．

$$\min_{T_{ij}} \sum_{i,j} C_{ij} T_{ij} \ s.t. \ \sum_j T_{ij} = a_i, \sum_i T_{ij} = b_j. \tag{A.1}$$

式 A.1 は非凸最適化問題であり，解が一つに定まらない．この問題に対し，Cuturi はエントロピー正則化を導入する事により，凸最適化問題となるように A.1 式を変換した．すなわち，エントロピー正則化の係数を λ として，

$$\min_{T_{ij}} \sum_{i,j} C_{ij} T_{ij} + \lambda T_{ij} \log T_{ij} \ s.t. \ \sum_j T_{ij} = a_i, \sum_i T_{ij} = b_j. \tag{A.2}$$

とした．A.2 式は Lagrange の未定乗数法を用いることにより，

$$\min_{T_{ij},\alpha_i,\beta_j} \sum_{i,j} C_{ij} T_{ij} + \lambda T_{ij} \log T_{ij} + \alpha_i * \left(a_i - \sum_j T_{ij} \right) + \beta_j * \left(b_j - \sum_i T_{ij} \right). \tag{A.3}$$

と書き換えることが出来る．そこで停留条件（すなわち Karush-Kuhn-Tucker (KKT) 条件）により，

$$\begin{aligned} & C_{ij} + \lambda \log T_{ij} + \lambda - \alpha_i - \beta_j = 0 \\ \Leftrightarrow & T_{ij} = \exp\left(\frac{\alpha_i}{\lambda} + \frac{1}{2}\right) \exp\left(-\frac{C_{ij}}{\lambda}\right) \exp\left(\frac{\beta_j}{\lambda} + \frac{1}{2}\right) \equiv A_i \exp\left(-\frac{C_{ij}}{\lambda}\right) B_j \end{aligned} \tag{A.4}$$

を得る．よってエントロピー正則化つきの最適輸送問題の解 T^*_{ij} は最適輸送問題の条件

$$\sum_j T_{ij} = a_i, \sum_i T_{ij} = b_j. \tag{A.5}$$

を満たすように Lagrange の未定乗数 A_i および B_j を決定することで得られる．特に A_i および B_j は適当な値で初期化した後，以下の繰り返しアルゴリズム（Sinkhorn Iteration）により求めることが

出来ると知られている：

$$
\begin{aligned}
A_i^{(t+1)} &\leftarrow \frac{a_i}{\sum_j \exp\left(-\frac{C_{ij}}{\lambda}\right) B_j^{(t)}} \\
B_j^{(t+1)} &\leftarrow \frac{b_j}{\sum_i A_i^{(t+1)} \exp\left(-\frac{C_{ij}}{\lambda}\right)}
\end{aligned}
\tag{A.6}
$$

付 録B

Prototreeにおける葉ノード更新則の導出

本章では任意の $\boldsymbol{\pi}$ に対して

$$\pi^*_{l,y} = \frac{1}{Z_l} \sum_{i \in \mathbb{B}} \frac{\mathbf{1}(y = y_i)\pi_{l,y}\mu_l(x_i)}{\sum_{m \in \mathbb{L}} \pi_{m,y}\mu_m(x_i)} \tag{B.1}$$

とすれば,

$$R(\boldsymbol{\pi}; \mathbb{B}) \geq R(\boldsymbol{\pi}^*; \mathbb{B}) \tag{B.2}$$

となることを証明する．ただし,

$$P_T(y|x, \boldsymbol{\pi}) = \sum_{l \in \mathbb{L}} \pi_{l,y}\mu_l(x), R(\boldsymbol{\pi}; \mathbb{B}) = -\frac{1}{|\mathbb{B}|} \sum_{i \in \mathbb{B}} P_T(y_i|x_i, \boldsymbol{\pi}) \tag{B.3}$$

であり，本章での記号の定義は 4 章で定義したものに準拠する．また以下では

$$\phi(\boldsymbol{\pi}, \bar{\boldsymbol{\pi}}) = R(\bar{\boldsymbol{\pi}}; \mathbb{B}) - \frac{1}{|\mathbb{B}|} \sum_{i \in \mathbb{B}} \sum_{l \in \mathbb{L}} \xi_l(\bar{\boldsymbol{\pi}}; x_i, y_i) \log \frac{\pi_{l,y_i}}{\bar{\pi}_{l,y_i}} \tag{B.4}$$

と定義する．ただし,

$$\xi_l(\boldsymbol{\pi}; x, y) = \frac{\pi_{l,y}\mu_l(x)}{P_T(y|x, \boldsymbol{\pi})} \tag{B.5}$$

である．ここで,

$$\phi(\boldsymbol{\pi}, \boldsymbol{\pi}) = R(\boldsymbol{\pi}; \mathbb{B}), \pi^*_{l,y} = \frac{1}{Z_l} \sum_{i \in \mathbb{B}} \mathbf{1}(y = y_i)\xi_l(\boldsymbol{\pi}; x_i, y_i) \tag{B.6}$$

となることに注意されたい．さて，B.1 式の証明には以下の 2 つの式が成立する事を証明すれば十分である．

$$\phi(\boldsymbol{\pi}, \bar{\boldsymbol{\pi}}) \geq R(\boldsymbol{\pi}; \mathbb{B}) \tag{B.7}$$

$$\forall \boldsymbol{\pi}, \bar{\boldsymbol{\pi}}\ \phi(\bar{\boldsymbol{\pi}}, \boldsymbol{\pi}) > \phi(\boldsymbol{\pi}^*, \boldsymbol{\pi}) \tag{B.8}$$

なぜなら，B.7 式および B.8 式を認めれば，任意の $\boldsymbol{\pi}$ について

$$R(\boldsymbol{\pi}; \mathbb{B}) = \phi(\boldsymbol{\pi}, \boldsymbol{\pi}) > \phi(\boldsymbol{\pi}^*, \boldsymbol{\pi}) \geq R(\boldsymbol{\pi}^*; \mathbb{B}) \tag{B.9}$$

となるためである．以下では，B.7 式および B.8 式を証明する．

B.7 式の証明　B.7 式は Jensen の不等式を用いることで簡単に証明することができる．ここで Jensen の不等式は $\lambda_i \geq 0$，$\sum_i \lambda_i = 1$ を満たす定数 λ_i および凸関数 f について成立する以下の不等式の事である．

$$\sum_i \lambda_i f(x_i) \geq f\left(\sum_i \lambda_i x_i\right) \tag{B.10}$$

今，$-\log(\cdot)$ が凸関数，$\sum_{l\in\mathbb{L}} \xi_l(\bar{\boldsymbol{\pi}}; x, y) = 1$ となることに注意すれば，

$$\begin{aligned} -\sum_{l\in\mathbb{L}} \xi_l(\bar{\boldsymbol{\pi}}; x, y) \log\left(\frac{\pi_{l,y}\mu_l(x)}{\xi_l(\bar{\boldsymbol{\pi}}; x, y)}\right) &\geq -\log\left(\sum_{l\in\mathbb{L}} \xi_l(\bar{\boldsymbol{\pi}}; x, y)\frac{\pi_{l,y}\mu_l}{\xi_l(\bar{\boldsymbol{\pi}}; x, y)}\right) \\ &= -\log\left(\sum_{l\in\mathbb{L}} \pi_{l,y}\mu_l\right) \end{aligned} \tag{B.11}$$

が成立するとわかるので，

$$\begin{aligned} R(\pi; \mathbb{B}) &= -\frac{1}{|\mathbb{B}|}\sum_{i\in\mathbb{B}} \log\left(\sum_{l\in\mathbb{L}} \pi_{l,y_i}\mu_l\right) \\ &\leq -\frac{1}{|\mathbb{B}|}\sum_{i\in\mathbb{B}}\sum_{l\in\mathbb{L}} \xi_l(\bar{\boldsymbol{\pi}}; x_i, y_i) \log\left(\frac{\pi_{l,y_i}\mu_l}{\xi_l(\bar{\boldsymbol{\pi}}; x_i, y_i)}\right) \\ &= -\frac{1}{|\mathbb{B}|}\sum_{i\in\mathbb{B}}\sum_{l\in\mathbb{L}} \xi_l(\bar{\boldsymbol{\pi}}; x_i, y_i)\left(\log(P_T(y_i|x_i, \boldsymbol{\pi})) + \log\left(\frac{\pi_{l,y_i}}{\bar{\pi}_{l,y_i}}\right)\right) \\ &= R(\bar{\boldsymbol{\pi}}; \mathbb{B}) - \frac{1}{|\mathbb{B}|}\sum_{i\in\mathbb{B}}\sum_{l\in\mathbb{L}} \xi_l(\bar{\boldsymbol{\pi}}; x_i, y_i) \log\frac{\pi_{l,y_i}}{\bar{\pi}_{l,y_i}} \\ &= \phi(\boldsymbol{\pi}, \bar{\boldsymbol{\pi}}) \end{aligned} \tag{B.12}$$

を得る．よって B.7 式は示された．

B.8 式の証明　B.8 式は単純な計算により証明出来る．実際，

$$\begin{aligned} &\phi(\boldsymbol{\pi}^*, \boldsymbol{\pi}) - \phi(\bar{\boldsymbol{\pi}}, \boldsymbol{\pi}) \\ &= R(\boldsymbol{\pi}; \mathbb{B}) - \frac{1}{|\mathbb{B}|}\sum_{i\in\mathbb{B}}\sum_{l\in\mathbb{L}} \xi_l(\boldsymbol{\pi}; x_i, y_i) \log\frac{\pi^*_{l,y_i}}{\pi_{l,y_i}} - R(\boldsymbol{\pi}; \mathbb{B}) - \frac{1}{|\mathbb{B}|}\sum_{i\in\mathbb{B}}\sum_{l\in\mathbb{L}} \xi_l(\boldsymbol{\pi}; x_i, y_i) \log\frac{\pi_{l,y_i}}{\bar{\pi}_{l,y_i}} \\ &= -\frac{1}{|\mathbb{B}|}\sum_{i\in\mathbb{B}}\sum_{l\in\mathbb{L}} \xi_l(\boldsymbol{\pi}; x_i, y_i) \log\frac{\pi^*_{l,y_i}}{\bar{\pi}_{l,y_i}} \\ &= -\frac{1}{|\mathbb{B}|}\sum_{i\in\mathbb{B}}\sum_{l\in\mathbb{L}}\sum_{y'\in\mathbb{Y}} \mathbf{1}(y' = y_i)\xi_l(\boldsymbol{\pi}; x_i, y_i) \log\frac{\pi^*_{l,y'}}{\bar{\pi}_{l,y'}} \\ &= -\frac{1}{|\mathbb{B}|}\sum_{y'\in\mathbb{Y}}\sum_{l\in\mathbb{L}} Z_l\pi^*_{l,y'} \log\frac{\pi^*_{l,y'}}{\bar{\pi}_{l,y'}} \quad (\because \text{Eq. B.6}) \\ &= -\frac{1}{|\mathbb{B}|}\sum_{l\in\mathbb{L}} Z_l KL(\boldsymbol{\pi}^*_l, \bar{\boldsymbol{\pi}}_l) \leq 0 \end{aligned} \tag{B.13}$$

より B.8 式も示された．ただし，$KL(\boldsymbol{\pi}^*_l, \bar{\boldsymbol{\pi}}_l) \geq 0$ は分布 $\boldsymbol{\pi}^*_l$ および $\bar{\boldsymbol{\pi}}_l$ 間の Kullback-Leibler divergence である．