大規模視覚言語モデルの少数データ適応
～学習効率と性能の両立を目指して～

2026年3月

加藤 直樹

学位論文 博士（工学）

大規模視覚言語モデルの少数データ適応
～学習効率と性能の両立を目指して～

2026年3月

慶應義塾大学大学院理工学研究科

加藤 直樹

論文要旨

深層学習では大規模な訓練データセットを用いてモデルを学習することで高精度な認識が可能であると知られている．例えば，画像認識分野においてはImageNetのような百万枚を超える規模のデータセットを用いることで，従来手法を大幅に上回る性能を達成している．しかし，大量のデータへのアノテーションは多大な労力を要し，専門知識を持つアノテーターの確保や品質管理などのコストが問題となっている．さらに，医用画像解析や希少動物の識別といった専門的なタスクによってはデータ自体の収集も困難であり，プライバシーや倫理的制約により大規模データセットの構築が現実的でない場合も多い．したがって，少数のラベル付きデータで高精度な認識を実現する手法が強く求められている．

近年，CLIPに代表される大規模視覚言語モデルが広く注目を集めている．これらのモデルは，インターネット上の膨大な画像とテキストのペアデータから学習することで，豊富な視覚的，言語的表現を獲得し，優れた汎化性能を示すことから，データの限られるタスクへの適用においても高い有効性が期待される．本論文では，このような大規模データで事前学習された視覚言語モデルを限られたデータを用いて下流タスクへ効率的に適応させることで少数データ画像認識を実現する手法について検討する．その際，事前学習モデルが獲得した汎用的な知識と下流タスクのドメイン知識を適切に組み合わせることで，多様な下流タスクにおける認識性能の改善を図る．また，計算効率性と実用性を兼ね備えた手法を開発することにより，リソース制約のある実応用シーンに適した少数データ学習の実現を目指す．

第1章では，既存の少数データ画像認識手法の問題を提起し，本研究の研究目的と貢献を示す．第2章では，少数データ画像認識の研究動向を概観し，その流れの中での本研究の位置付けを明らかにする．第3章では，学習データにおけるクラスごとのプロトタイプ表現に基づきアダプターを構築するProto-Adapterを提案する．事

前学習モデルのパラメータを変更することなく，学習不要かつ高速に適用可能な枠組みで高性能な少数データ画像認識を実現できることを示す．さらに，アダプターを距離学習の枠組みを用いてファインチューニングすることで，認識性能をさらに改善できることを実証する．第 4 章では，大規模視覚言語モデルを少数データに適応させる簡潔で効率的なベースライン手法として，線形識別器の残差学習に基づく適応手法を提案する．事前学習モデルとアダプターの推論結果に対する適切な重み付けが下流タスクによって異なることを明らかにし，この重み付けを適切に調整することにより，多様な下流タスクに対する平均性能がより複雑な既存手法を上回ることを実証する．第 5 章では，本論文の成果を総括し，今後の研究の展望について論じる．

Abstract

Deep learning achieves high recognition accuracy by training models on large-scale datasets. In image recognition, for instance, datasets with millions of images such as ImageNet have enabled performance that far exceeds conventional methods. However, collecting and annotating such data requires substantial effort, and in specialized domains such as medical imaging or rare-species identification, even data collection itself is challenging. Consequently, there is strong demand for methods that achieve high accuracy using only a small number of labeled examples.

Recently, large-scale vision-language models such as CLIP have attracted considerable attention. By learning from massive image–text pairs on the internet, these models acquire rich visual and linguistic representations and exhibit strong generalization, making them promising for tasks with limited data. This thesis investigates few-shot image recognition methods that efficiently adapt such pretrained vision-language models to downstream tasks using only limited labeled data. The aim is to improve recognition on diverse tasks by appropriately combining the general knowledge stored in the pretrained model with domain-specific information, while maintaining both computational efficiency and practical applicability under real-world resource constraints.

Chapter 1 raises problems with existing few-shot image recognition methods and presents the research objectives and contributions of this study. Chapter 2 reviews research trends in few-shot image recognition and clarifies the position of this research within that context. Chapter 3 proposes Proto-Adapter, which constructs adapters based on prototype representations for each class in the training data. We demonstrate that high-performance few-shot image recognition can be achieved in a training-free framework without modifying the parameters of the pretrained model. Furthermore, we show that

recognition performance can be further improved by fine-tuning the adapter using a metric learning framework. Chapter 4 presents a simple and efficient baseline for adapting large vision–language models to few-shot data, based on residual learning of linear classifiers. We show that the optimal weighting between the predictions of the pretrained model and those of the adapter varies across downstream tasks, and that appropriately adjusting this weighting yields average performance superior to that of more complex existing methods. Chapter 5 summarizes the achievements of this thesis and discusses future research prospects.

目 次

図 目 次

表 目 次

第1章　序論

本章では，本研究の背景と動機について述べる．まず，少数データ画像認識の基本概念と課題について説明し，続いて本研究の目的と主要な貢献について詳述する．最後に，本論文全体の構成を示す．

1.1　少数データ画像認識とその課題

深層学習では大規模な訓練データセットを用いてモデルを学習することにより高精度な認識が可能となることが知られている．例えば，画像認識分野においては多様なカテゴリの画像群で構成される ImageNet [1] のような百万枚を超える規模の画像データセットを用いることで，従来手法を大幅に上回る性能を達成している [2, 3, 4]．しかし，大量のデータへのアノテーションは多大な労力を要し，専門知識を持つアノテーターの確保や品質管理などのコストが問題となっている．さらに，図 1.1 に示すような，医用画像解析 [5, 6] や産業検査 [7, 8]，希少動物の識別 [9] といった専門的なタスクによってはデータ自体の収集も困難であり，プライバシーや倫理的制約により大規模データセットの構築が現実的でない場合も多い．したがって，少数のラベル付きデータで高精度な認識を実現する手法が強く求められている．

少数データ学習は，機械学習において限られた数の訓練サンプルから効果的に学習を行う手法の総称である [10, 11]．少数データ学習では各クラスにつき 1 個から数十個と，通常の教師あり学習と比べて少数の学習事例のみを用いて新たなタスクを解くことを目指す．高精度な少数データ学習技術が実現されることで，多くの実用

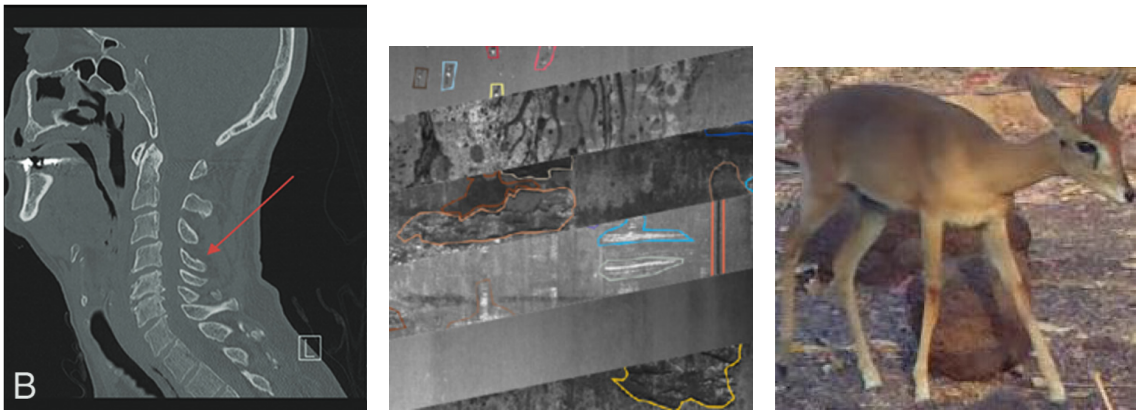

図 1.1: 少数データ画像認識の応用先の例．左：医用画像解析 [5]，中央：産業検査 [7]，右：希少動物の識別 [9].

上の利点が得られる．例えば，医療画像診断において，限られた症例データからでも高精度な診断モデルを構築できるようになり，診断精度の向上と医療の質の改善が期待される．また，製造業における品質検査では，不良品の発生頻度が低い場合でも効果的な検査システムを構築でき，生産効率の向上に寄与する．このような利点により，従来は大量のデータ収集が困難で機械学習の適用が制限されていた分野においても，機械学習技術の恩恵を受けることが可能となる．

少数データの代表的なアプローチとして，データ拡張，メタ学習，転移学習が存在する．データ拡張は，訓練データに対して，回転，スケーリング，ノイズ付加などといった，画像の意味情報を保った変換を適用することで，実質的な訓練データ数を増加させる手法である．少数データ学習に限らず，深層学習モデルの性能を改善させる技術として広く活用される．メタ学習は，少数データ学習の古典的なアプローチであり，複数の少数データタスクから得られた経験を活用して新しいタスクに素早く適応することで,「学習方法そのものを学習する」ことを目指す研究トピックである．Model-Agnostic Meta-Learning (MAML) [12] や Prototypical Networks [13] などが代表的である．しかし，単純な転移学習手法がメタ学習手法の性能を上回ることが確認されており，その有効性や実用性に対して疑問が呈されている [14]．転

移学習は，あるタスクやドメインで学習済みのモデルが持っている知識を，別のタスクやドメインに再利用して性能を上げる手法であり，下流タスクのデータ量に関わらず広く利用される．大規模データセットで事前学習されたモデルを下流タスクでファインチューニングする手法 [15, 16] や，固定した事前学習モデルを特徴抽出器として後段の識別に活用する手法 [17, 18] などが存在する．

近年の少数データ画像認識研究において，転移学習アプローチの有効性が広く実証されている．中でも，Contrastive Language-Image Pre-training（CLIP）[19] に代表される大規模視覚言語モデルは，インターネット上の膨大な画像，テキストペアデータでの事前学習により，豊富な視覚的，言語的表現を獲得している．これらのモデルは優れた汎化性能を示すため，少数データ画像認識分野においても活用が進んでいる．実際に，下流タスクの少数データを活用した適応手法により，対象ドメインにおける認識性能の顕著な改善が複数の研究において報告されている [20, 21, 22]．代表的な研究事例として，プロンプト調整手法である Context Optimization (CoOp) [20] は，CLIP において手動で設計しているテキストプロンプトを学習可能なベクトルに置き換え，少数の学習サンプルでそのパラメーターを最適化する．しかし，学習データ数が非常に少ないとき，適応により逆に認識性能が低下してしまう場合があることが問題である．また，適応の際に画像エンコーダーへの誤差逆伝播が必要であり，計算コストが比較的高いことも課題であると言える．CLIP-Adapter [21] は，画像エンコーダーとテキストエンコーダーの上層に残差接続のアダプター [23] を追加することで CLIP を下流タスクに適応させる．しかし，CLIP において大規模なデータから獲得された画像特徴量とテキスト分類器の整合性が崩れ，認識性能を損なってしまうことが懸念される．Tip-Adapter [22] は，少数データの画像特徴量及びラベルから成るキーバリューキャッシュで初期化したアダプターを CLIP の上層に追加する．アダプターの重みを調整することで，少数データでの分類性能を大きく向上させることができる．しかし，アダプターのサイズは学習サンプル数に依存して変化

するため，データ数の変動する実運用シーンにおいて利用しづらいことが課題である．またいずれの手法においても，多様な下流タスクに対する汎用的な適応性能には改善の余地が存在する．特に，飛行機 [24] やテクスチャ[25] の識別などのドメイン固有の知識を必要とするタスクへの認識性能が相対的に低下しやすい傾向にあり，その改善が必要である．

1.2 研究目的と貢献

本研究の目的は，少数データ画像認識における既存手法の限界を克服し，より効率的で実用性の高い学習手法を開発することである．具体的には，大規模データで事前学習された視覚言語モデルを少数の学習データで下流タスクへ効率的に適応させることで，高性能な少数データ画像認識を実現する手法の構築を目指す．前節で述べた既存手法の課題を踏まえ，本研究では以下 4 点の課題に取り組む．

- **効率的な適応手法の構築：**大規模事前学習モデルを少数データの下流タスクに効率的に適応させ，非常に少数の学習サンプルに対しても適用可能な手法を構築する．
- **計算効率性の向上：**計算コストの低い学習枠組みを構築し，リソースに制約がある状況でも高性能な少数データ学習を実現させる．
- **デプロイの容易性：**モデル構造が学習データ数に対して不変であり，使用可能なデータ数が変動する実運用シーンにおいても容易に利用可能な手法を構築する．
- **汎用性の確保：**多様なドメインの下流タスクに対して汎用的に適用可能で，ドメイン固有の知識を必要とするタスクでも安定した性能を発揮する手法を構築する．

これらの課題解決により，データ収集が困難な実世界のタスクに対して実用的な画像認識システムの構築を可能にすることを目指す．特に，リソース制約下でも高性能な少数データ学習を実現することで，従来は機械学習の適用が困難であった分野への技術普及を促進する．

1.2.1 プロトタイプ適応器による学習不要かつ効率的な大規模視覚言語モデルの少数データ適応

この研究では，クラスごとのプロトタイプ表現に基づきアダプターを構築する Proto-Adapter を提案し，学習不要な枠組みで高性能な少数データ画像認識を実現できることを示す．

本研究の主要な貢献は以下の通りである．

- 学習サンプル数に関わらず一定のアダプターサイズを維持する，CLIP のための新しい学習不要適応手法を提案する．これは各クラスの学習サンプルの特徴量を集約してアダプターの重みを構築することで実現される．
- 出力ロジットにおけるクラス間の距離に対して制約を導入する Additive Angular Margin Penalty を用いたファインチューニングにより，提案手法の性能をさらに向上させることができることを示す．
- ImageNet と他の 10 種類の画像認識データセットにおいて少数データ分類の性能を評価し，提案手法の既存の CLIP 適応手法に対する優位性を実証する．

1.2.2 線形識別器の残差学習による大規模視覚言語モデルの少数データ適応

この研究では，大規模視覚言語モデルの少数データ適応を簡素な枠組みで効率的に行うベースラインである，線形識別器の残差学習による少数データ適応手法を提案する．

本研究の主要な貢献は以下の通りである．

- 線形識別器の残差学習による大規模視覚言語モデルの少数データ適応手法を提案する．本手法は CLIP の頑健な特徴表現を維持しつつ，新たなドメインの知識を少数の学習データから効率的に導入することを可能とする．また本手法は事前学習された視覚言語モデルに対して最小限の変更で適用可能であり，少数データ適応におけるベースライン手法となることを意図している．
- 多様な画像認識ベンチマークでの性能評価により，提案手法がより複雑な比較手法と比べ，多様なドメインの下流タスクに対する汎用的な認識性能に優れることを実証する．

1.3 本論文の構成

第 1 章では，少数データ画像認識の問題設定を述べるとともに，既存研究の課題を提起した．また，本研究の研究目的と貢献を示した．第 2 章では，少数データ画像認識の研究動向を概観し，その流れの中での本研究の位置付けを明らかにする．第 3 章では，クラスごとのプロトタイプ表現に基づきアダプターを構築する Proto-Adapter を提案し，学習不要な枠組みで高性能な少数データ画像認識を実現できることを示す．この内容は，文献 [26] で発表したものである．第 4 章では，大規模視覚言語モ

デルの少数データ適応を簡素な枠組みで効率的に行うベースラインである，線形識別器の残差学習による少数データ適応手法を提案する．第 5 章では，本論文を総括するとともに，今後の展望について議論する．

第2章　関連研究

深層学習が発展した現在においても，大規模な学習画像の収集が困難な領域では，少数データ画像分類が依然として重要な研究課題である．本章では，少数データ学習に関する関連研究を概観する．まず，少数データ画像分類の主要なアプローチであるメタ学習，データ拡張，転移学習について，それぞれの概要と代表的な先行研究を説明する．次に，近年注目を集めている大規模視覚言語モデルを活用した少数データ画像認識手法について，視覚言語事前学習の基礎から CoOp，CLIP-Adapter，Tip-Adapter などの具体的な手法まで幅広く紹介する．最後に，これらの手法が抱える課題を整理し，本研究との関係をまとめる．

2.1　メタ学習

少数データ学習の古典的なアプローチであるメタ学習は，学習方法自体の学習を図る枠組みであり，限られた学習事例を用いて新たなタスクへ素早く適応することを目的とする．一般的な深層学習の枠組みでは単一のデータセットを用いてモデルのパラメーターを最適化するのに対し，メタ学習では多数の少数データタスクを使用し，タスク間で共通する学習の仕方を獲得しようとする点が異なる．

最初期の手法である Siamese Network [27] は，2 つの入力画像を同一の特徴抽出ネットワークに通し，得られた埋め込み表現間の距離が同一クラスなら小さく，異なるクラスなら大きくなるように学習することで，距離尺度を獲得する手法である．学習には，画像対に対して同一クラス，異なるクラスのラベルを与え，コントラスト

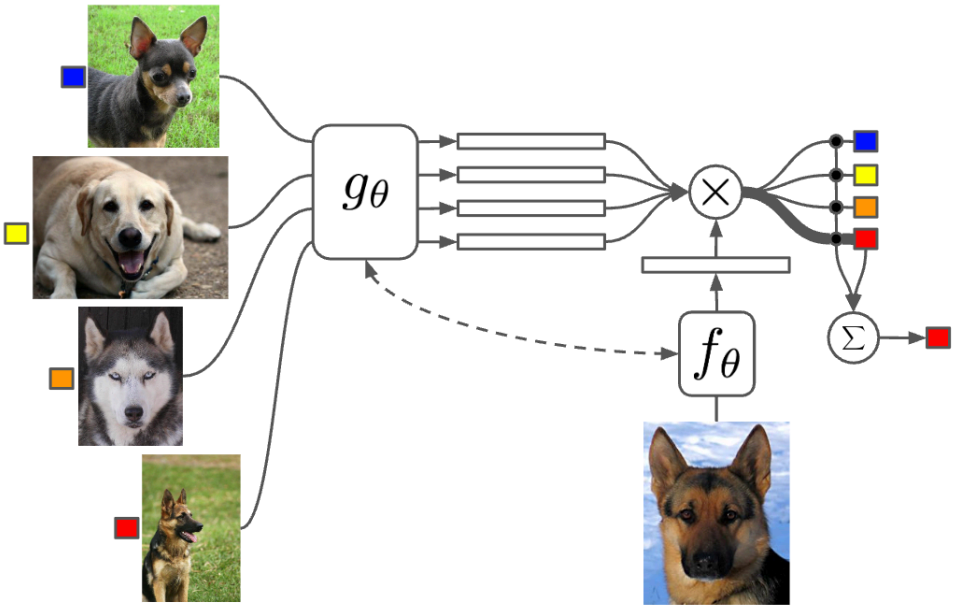

図 2.1: Matching Networks の枠組み.

損失により埋め込み空間を整形する枠組みが用いられる．推論時には，クエリ画像と各クラスの参照画像との距離を計算し，最も近い参照画像のクラスを出力することで 1-shot 認識を実現する．このように，クラスごとの分類器を直接学習するのではなく，クラスをまたいで再利用可能な類似度計算を学習する点が特徴であり，少数データ環境での認識を可能にした代表的な距離学習ベースの先駆的手法として位置づけられる．

図 2.1 に示す Matching Networks [28] は，少数のサポート集合とクエリ集合から成るエピソードを繰り返し学習することで，テスト時の少数データ設定と整合した学習を行う手法である．ここでサポート集合とは，各エピソードにおいて各クラスにつき数枚のラベル付き例から構成され，新規タスクに対する参照として用いられる少数の学習データである．各エピソードでは，サポート集合の各例をメモリとして保持し，クエリ画像の埋め込み表現とサポート例の埋め込み表現の類似度に基づく注意機構により，クエリのクラス確率を最近傍的に推定する．この枠組みにより，

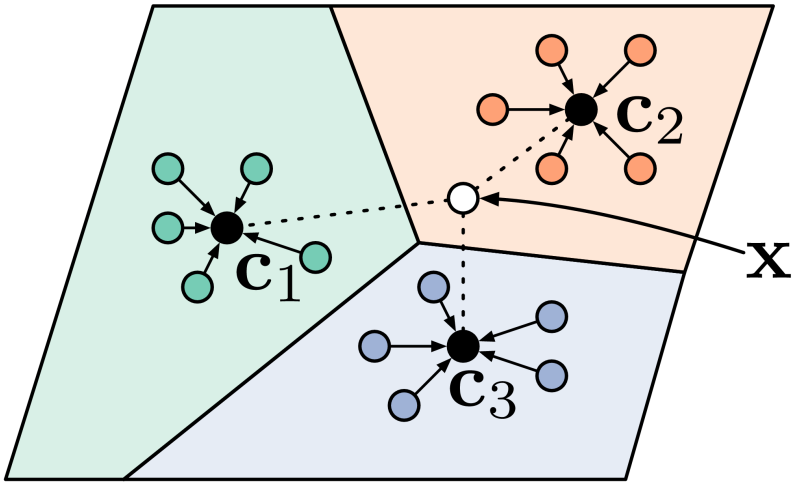

図 2.2: Prototypical Networks の枠組み．c は各クラスのプロトタイプを，x はクエリ埋め込みを表す．

クラスごとの分類器を明示的に学習せずとも，新規クラスに対してサポート例を参照して予測でき，汎化性能の向上が報告されている．

一方，Prototypical Networks [13] は，各クラスのサポート例の埋め込みを平均して得られるクラス代表ベクトルであるプロトタイプを導入し，クエリ埋め込みと各プロトタイプとの距離に基づいて分類する．これにより，サポート集合全体を参照する Matching Networks よりも推論が簡潔になる．エピソード学習を通して埋め込み空間におけるクラス間の距離構造を学習することで，少数データ条件下での安定した性能が得られる．

図 2.3 に示す最適化ベースのアプローチである Model-Agnostic Meta-Learning (MAML) [12] は，タスクごとに少数ステップの勾配更新（内側ループ）を行った後の性能が高くなるように，更新前の初期パラメータを学習する（外側ループ）手法である．具体的には，各エピソードでサポート集合を用いてモデルのパラメータ θ を θ' へと数ステップ更新し，その θ' でクエリ集合の損失を評価して，その損失が小さくなるように θ 自体を更新する．このように少数ステップのパラメータ更新で素早く適応できる初期値を獲得することで，分類器の構造に依存せず様々なモデルに適用可能な

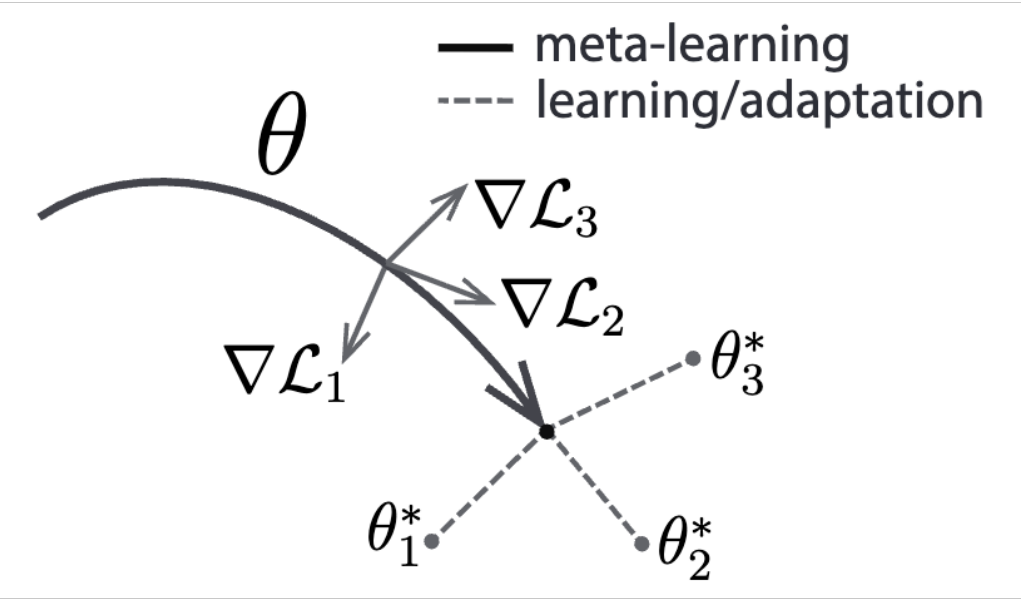

図 2.3: MAML の枠組み．θ はモデルのパラメータを，L は各種少数データタスクにおける損失関数を，θ^* は目的タスクで学習後のモデルパラメータを表す．

点が特徴である．

しかし，メタ学習手法の隆盛の一方で，大規模な事前学習モデルを単純に微調整する転移学習的アプローチが，標準的なベンチマーク設定においてメタ学習手法を上回る性能を示すことが確認されている [29]．さらに，多数のエピソードを想定したデータ分割や同一ドメイン内でのタスク生成といった，メタ学習で一般的に用いられる評価プロトコルが，現実の少数データ問題の条件と必ずしも一致しない可能性が指摘されており [30]，メタ学習の実用上の優位性や適用条件について再検証が進められている．

2.2 データ拡張

データ拡張は，元データのラベルを保ったまま画像に種々の変換を施し，訓練時に観測されるデータ分布の多様性を人工的に増やすことで，汎化性能の向上と過学

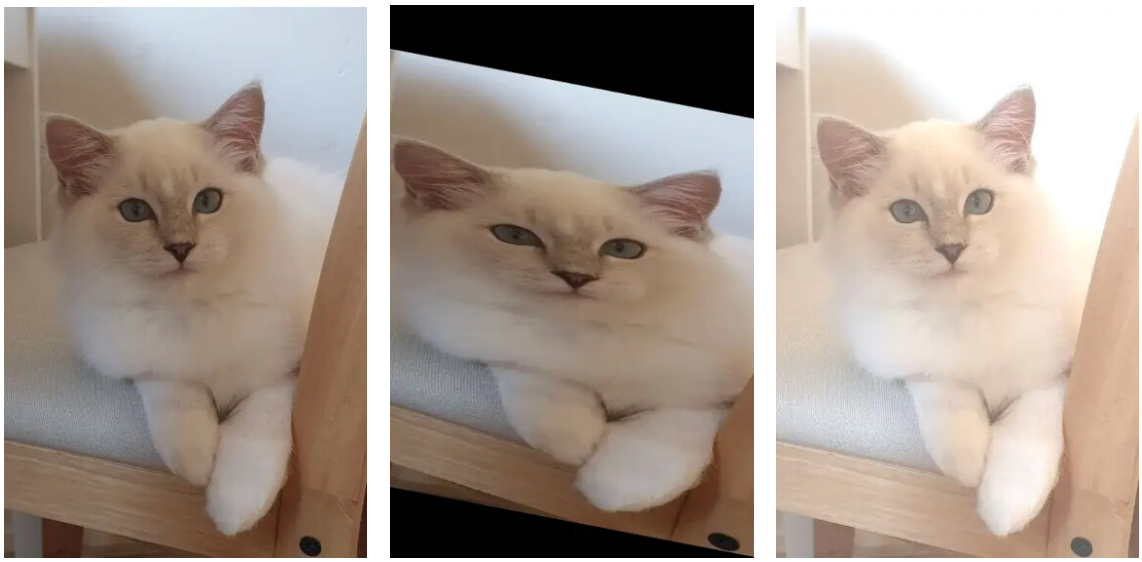

図 2.4: 代表的なデータ拡張の例．左：元画像，中央：幾何学的変換，右：色変換．

習の抑制を図る古典的かつ有効な技術である．特に少数データ設定では，学習サンプルが限られることでモデルが訓練画像の見た目（背景や撮影条件など）に過度に適合しやすいため，データ拡張による入力揺らぎの付与が重要となる．図 2.4 に示すように，回転・平行移動・反転・切り出しなどの幾何学変換や，明るさ・彩度・コントラストの変更といった色変換は，クラスの意味を保ちつつ見え方のみを変化させることで，撮影条件の違いに頑健な特徴表現の学習を促す．さらに，図 2.5 に示す Mixup [31] や CutMix [32] に代表される画像混合型の手法は，二枚の画像を線形結合または領域置換により合成し，混合比率に応じてラベルも混合して教師信号とする．これにより決定境界が滑らかになる方向に学習が誘導され，強い正則化効果とロバスト性向上が得られることが知られている．AutoAugment [33] や RandAugment [34] は，どの変換をどの強度で適用するかという方策を自動的に探索・最適化する枠組みであり，ImageNet [1] などのベンチマークにおいて性能向上を示してきた．特に RandAugment は探索を簡略化し，少数のハイパーパラメータで強力な拡張を実現する点で実用性が高い．また近年では，敵対的生成ネットワークや拡散モデルを用い

図 2.5: 画像の混合に基づくデータ拡張の例．左：Mixup，右：CutMix．これらのデータ拡張では，画像の混合比率に基づきそれらのラベルを混合して教師ラベルとして使用する．

て合成画像を生成し，品質やラベル整合性をフィルタリングした上で学習データに追加する生成ベースのデータ拡張の有効性も確認されている [35, 36]．合成データは単純な変換では得られない外観変化を導入できる一方，生成物の品質低下やラベルノイズが性能を損なう場合もあるため，選別やプロンプト設計などが重要となる．

2.3 転移学習

大規模なデータで事前学習したモデルのパラメータを異なるタスクへ転用する転移学習は，図 2.6 に示す，2014 年に Oquab らがその有効性を実証して以来，少数データタスクに対する深層学習モデル構築の標準的なアプローチとなった [37]．この枠組みでは，一般にソースタスク（例：ImageNet）の大量データで学習した特徴抽出器を初期値として利用し，目的タスクでは分類層のみを学習する，あるいは上位層を中心に微調整することで，少ない教師データでも高い精度を得ることを狙う．少数

データ環境では一から学習すると過学習しやすいのに対し，事前学習済み表現を活用することで学習すべきパラメータ数や探索空間を抑えられる点が利点である．近年では，自己教師あり事前学習により獲得された特徴表現の転移性能が教師あり事前学習を上回ることが複数の研究で報告されている [38, 39, 40]．自己教師あり学習ではラベルに依存せず，画像の再構成や対照学習などの事前課題を通じて汎用的な表現を獲得できるため，目的タスクのラベル空間やドメインがソースタスクと異なる場合でも頑健に機能しやすい．そのため，少数データ下流タスクにおいても，特徴抽出器の固定（線形プロービング）や限定的な微調整といった軽量な適応で良好な性能が得られることが多い．また，CLIP [19] や ALIGN [41] などの視覚言語基盤モデルは，インターネットから収集した画像とテキストの組から成る大規模な訓練データセットを用いて，画像とテキストを共通の埋め込み空間に写像するよう学習する．このような事前学習により，クラス名や説明文をテキストとして与えるだけで分類器（テキスト特徴）を構成できるため，追加学習なしのゼロショット分類，および少数のラベル付き例を用いた少数データ分類の双方を実現できる．特に，言語を介してクラス概念を指定できる点は，従来の閉集合分類器に比べて高い柔軟性を与える．さらに，CLIP の少数データ認識性能を向上させるため，テキストプロンプト中の文脈トークンを学習して下流タスクに適したプロンプトを獲得する CoOp [20] や，少数の追加パラメータで特徴を補正するアダプターを導入する CLIP-Adapter [21]，学習データの特徴量をキャッシュとして利用して予測を補強する Tip-Adapter [22] が提案されている．これらは，CLIP の事前学習済み表現をできるだけ保持しつつ，少数データでも過学習を抑えながら目的タスクへ適応させることを目的としている．これらの大規模視覚言語モデルを活用した少数データ画像認識手法について次節で説明する．

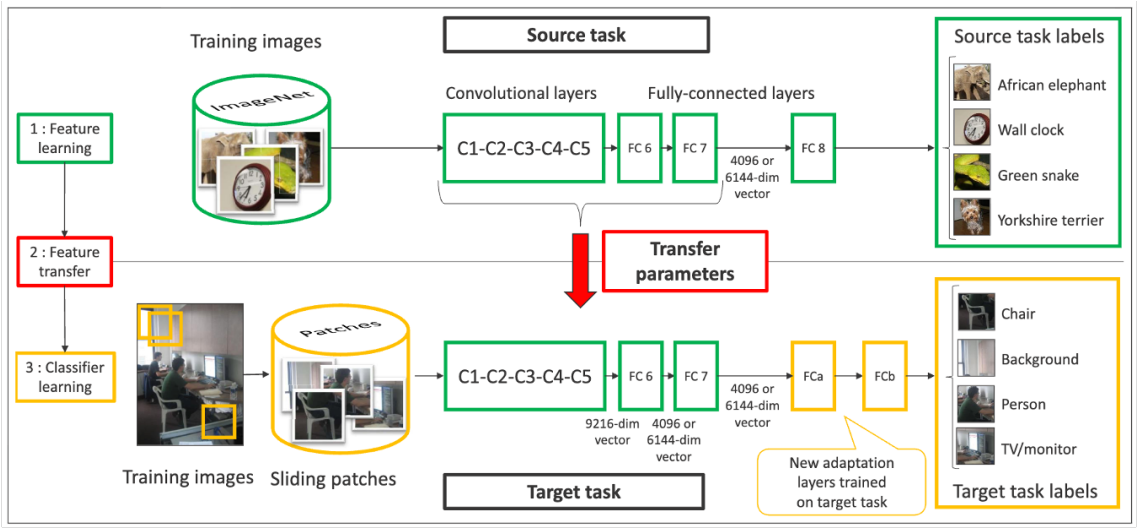

図 2.6: Oquab らの転移学習手法の枠組み．ソースタスクで学習した特徴抽出層のパラメータを目的タスクに流用し，その上層に新たに付け加えた適応層のパラメータのみを目的タスクで学習する．

2.4 大規模視覚言語モデルの活用

大規模視覚言語モデルは，インターネットから収集された大量の画像とテキストの組を用いた事前学習により構築されたモデルである．その汎化性能の高さから，画像分類に限らず，物体検出や領域分割，画像検索などの様々な視覚タスクへ応用されている．少数データ画像分類へ適用する研究も存在し，実用的な条件設定において高い性能を発揮することが確認されている．

2.4.1 CLIP

Contrastive Language-Image Pre-training (CLIP) [19] は，代表的な大規模視覚言語モデルであり，インターネットから収集された大量の画像とテキストの組のデータから両モダリティ間の対応関係を学習する．図 2.7 に CLIP の枠組みを示す．

CLIP は画像エンコーダーとテキストエンコーダーの 2 つで構成される．画像エン

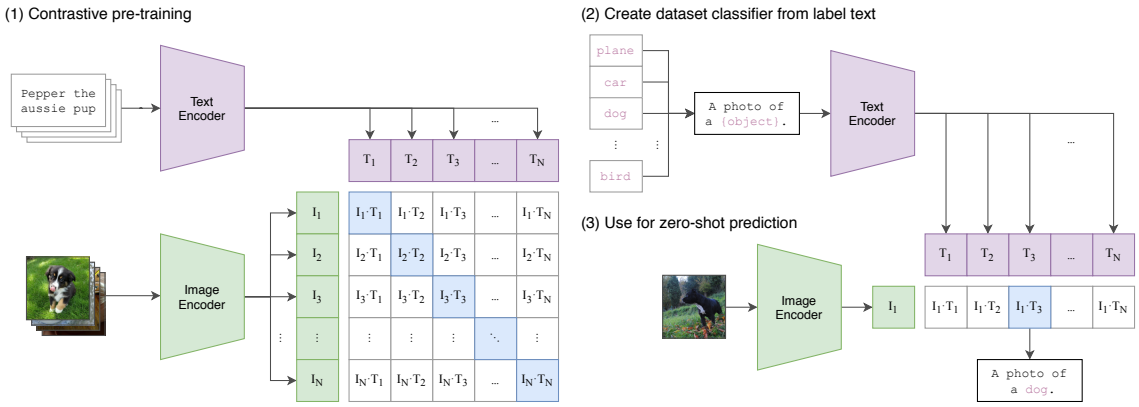

図 2.7: CLIP の事前学習と推論の枠組み.

コーダーは入力画像を D 次元の特徴ベクトルに変換し，テキストエンコーダーは入力テキストのトークン列を同じ D 次元の埋め込み空間に写像する．両エンコーダーは，画像とテキストの埋め込みベクトルが共通の表現空間で意味的に整合するよう共同で訓練される．

CLIP の学習では対照学習が用いられる [42, 43]．具体的には，学習バッチ内の正例ペア（対応する画像とテキスト）の埋め込みベクトル間のコサイン類似度を最大化し，同時に負例ペア（対応しない画像とテキスト）の類似度を最小化するよう InfoNCE 損失 [43] を用いて両エンコーダーのパラメーターを学習する．この学習により，意味的に関連する画像とテキストが埋め込み空間内で近い位置に配置される．CLIP は多様な視覚概念を獲得するために，インターネットから収集された約 4 億組の画像とテキストの組で構成される WIT-400M データセットで学習されている．

CLIP の特筆すべき特徴は，事前学習時に見たことのない新しいクラスに対するゼロショット推論が可能なことである．これは画像エンコーダーで抽出した画像特徴量と，テキストエンコーダーで生成したテキスト分類器との類似度計算により実現される．

推論時には，各クラスに対して「a photo of a [CLASS].」形式のテキストプロンプトを作成する．ここで [CLASS] は「cat」「dog」などの具体的なクラス名で置き換え

られる．N クラス分類問題において，テキスト分類器は以下のように構築される．

$$W_{\text{text}} = [w_1, w_2, \ldots, w_N] \in \mathbb{R}^{D \times N} \tag{2.1}$$

ここで w_i は i 番目のクラスに対するプロンプトをテキストエンコーダーに入力して得られる特徴ベクトルである．

画像 x に対する各クラスの予測ロジットは，画像特徴量 f とテキスト分類器の内積により以下のように計算される：

$$\text{logits} = W_{\text{text}}^T f / \tau \tag{2.2}$$

ここで τ は学習可能な温度パラメーターである．

従来の画像認識モデルが閉集合のクラスでの分類に限定されるのに対し，CLIP は自然言語による柔軟なクラス記述を通じて開集合の視覚概念認識を実現している．この特性により，事前学習時に明示的に学習していないクラスに対してもゼロショット認識が可能となり，実用的な応用において高い汎化性能を示している．

CLIP の画像エンコーダーから抽出された特徴量を用いて下流タスクの少数データで線形分類器を学習する線形プロービング手法についても検証が行われている．しかし，学習サンプル数が極めて限られた設定においては，線形プロービングの性能がゼロショット推論を下回るという課題が明らかになっている [19]．これは線形プロービングではテキストエンコーダーを使用しないため，CLIP が事前学習で獲得した汎用的な表現を十分に活用できないためと考えられる．このような背景から，CLIP の事前学習済み表現を保持しつつ，少数データ環境においても効果的に下流タスクへ適応させる手法の開発が重要な研究課題となっている．

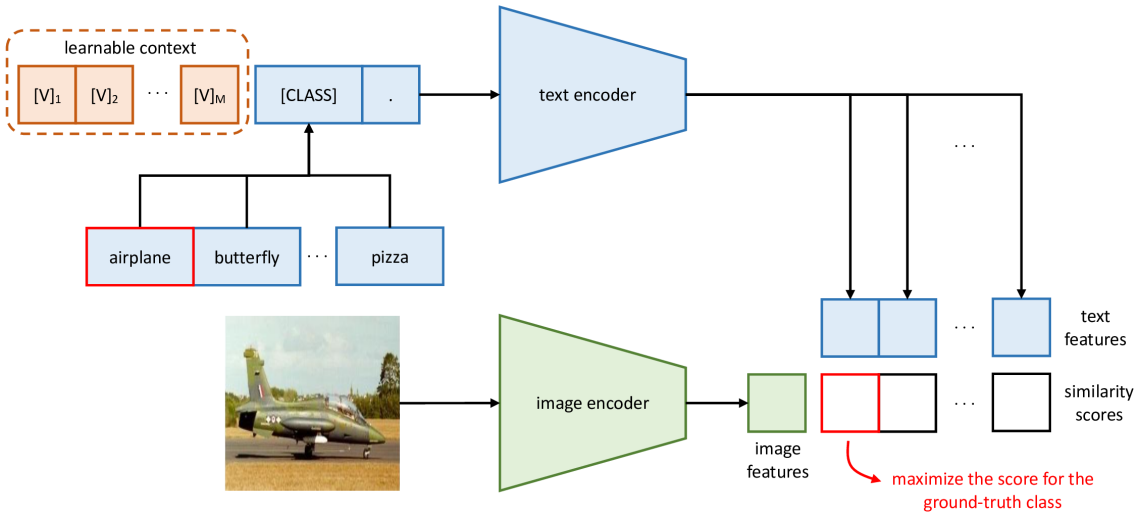

図 2.8: CoOp の枠組み.

2.4.2 CoOp

Context Optimization (CoOp) [20] は，CLIP の事前学習済みモデルを固定したまま，テキストプロンプトにおけるコンテキストワード部分のみを学習可能なベクトルとして最適化する手法である．図 2.8 に CoOp の枠組みを示す．従来の CLIP では「a photo of a [CLASS].」のような，人手で設計された固定のプロンプトテンプレートを使用していたが，CoOp では学習可能なコンテキストベクトルを導入することで，各下流タスクに適応したプロンプトを自動的に学習することができる．

CoOp では，プロンプトを学習可能なコンテキストベクトル $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M$ とクラス名 c_i の組み合わせで表現する．具体的には，以下の形式でプロンプトを構築する：

$$t_i = [\mathbf{v}_1][\mathbf{v}_2]\cdots[\mathbf{v}_M][c_i] \tag{2.3}$$

ここで M はコンテキストの長さ，$[\cdot]$ はトークンの連結を表す．コンテキストベクトル $\mathbf{v}_j \in \mathbb{R}^D$ は，単語埋め込みと同じ次元数の学習可能なパラメーターである．

CoOp の学習では，画像エンコーダーとテキストエンコーダーの重みを固定し，コ

ンテキストベクトルのみを更新する．損失関数には標準的な交差エントロピー損失を使用する．

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log p_i \tag{2.4}$$

ここで y_i は正解ラベル，p_i は予測確率である．予測確率は画像特徴量と学習されたプロンプトから生成されるテキスト特徴量のコサイン類似度に対し，ソフトマックス関数をかけることで計算される．

CoOp は少数のパラメーターのみを学習するため，少数データ環境において過学習の抑制を図っている．また，事前学習済みのエンコーダーを固定することで，CLIP が獲得した汎用的な表現を保持しながら下流タスクへの適応が可能となる．11 個のデータセットでの評価において，CoOp は手動で設計されたプロンプトを用いるゼロショット性能を大幅に上回ることが報告されている [20].

一方で，CoOp は学習されたコンテキストが人間にとって解釈困難であること，また新しいクラスに対する汎化性能において課題があることが指摘されている．これらの問題に対処するため，CoCoOp [44] や KgCoOp [45] などの拡張手法が提案されている．

2.4.3 CLIP-Adapter

CLIP-Adapter [21] は，CLIP の事前学習済みモデルの上層にアダプター層を追加することで，少数データでの効率的な適応を図る手法である．図 2.9 に CLIP-Adapter の枠組みを示す.

CLIP-Adapter では，CLIP の画像エンコーダーおよびテキストエンコーダーの上層それぞれに，2 層の多層パーセプトロンから成るアダプター層を付加する．CLIP の画像特徴量およびテキスト特徴量はそれぞれのアダプターによって処理された後，残差接続によって元の特徴量に付加される．これにより CLIP の汎用的な特徴表現

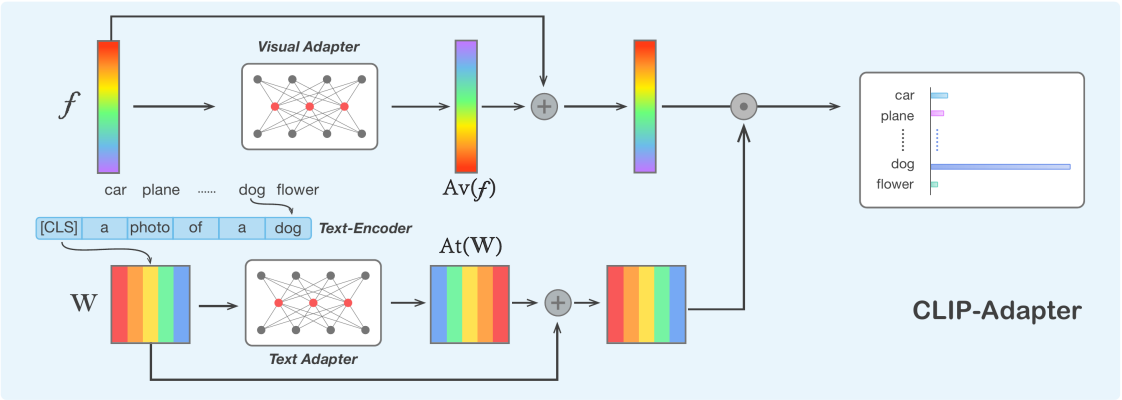

図 2.9: CLIP-Adapter の枠組み.

を崩さない適応を図っている．CLIP-Adapter の最終的な予測は，適応された画像特徴量およびテキスト特徴量に対して，CLIP と同様にコサイン類似度算出することで得られる．

学習時には，CLIP の画像エンコーダーとテキストエンコーダーの重みを固定し，アダプター層のパラメーターのみを学習する．損失関数には標準的な交差エントロピー損失を使用する．エンコーダーへの誤差逆伝播が必要ないため，CoOp と比べて効率的な学習が可能である．

一方で，CLIP-Adapter は適応により画像特徴量とテキスト特徴量を変換させるため，CLIP が大規模な事前学習を通して獲得した，それら特徴量間の整合性を崩してしまうことが懸念される．

2.4.4 Tip-Adapter

Tip-Adapter [22] は，CLIP の事前学習済みモデルに対して，少数データから構築したキャッシュモデルを組み合わせることで，効率的な少数データ適応を実現する手法である．図 2.10 に Tip-Adapter の枠組みを示す．

Tip-Adapter は，学習データの特徴量とラベルを直接キャッシュとして扱いアダプ

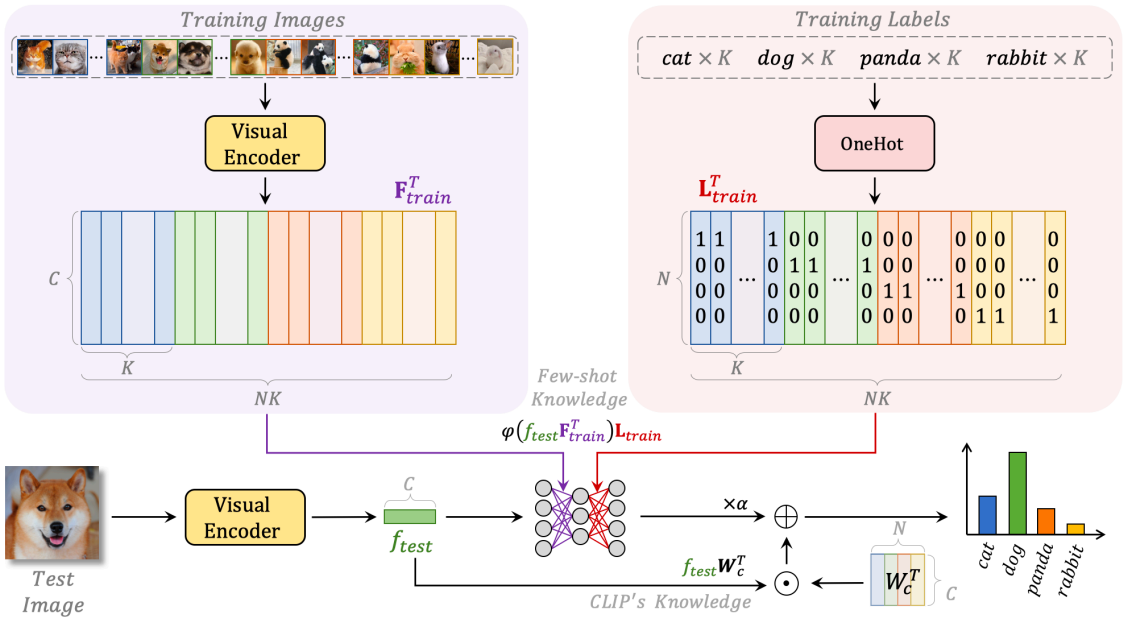

図 2.10: Tip-Adapter の枠組み.

ター層を構築することで，効率的な少数データ適応を図る手法である．この手法は，誤差逆伝播法を要としない学習不要版と，追加のアダプター学習を行う学習版の 2 種類のバリエーションを提供する．学習不要版では，学習データの特徴量とラベルそれぞれを 2 層のアダプターのパラメーターとして使用する．これにより，学習を行うことなく少数データ画像認識が可能となる．学習を行う場合は，CLIP のエンコーダーのパラメーターを凍結した状態で，アダプターのパラメーターのみを微調整する．これにより，認識性能をさらに改善することができる．また，Tip-Adapter は CLIP-Adapter と異なり CLIP 自体のゼロショット推論の推論経路を保持しているため，特徴量の整合性が崩れてしまう懸念を解消している．

一方で，キャッシュサイズが学習データ数に依存するため，データ数に応じてモデル構造が変化する．また，データ数が多い場合にメモリ使用量が増加する問題がある．

2.5 既存手法の課題

大規模視覚言語モデルの少数データ適応手法により，現実的な条件設定において比較的優れた認識性能の実現が可能となった．しかし，特にドメイン特有の知識を必要とするタスクに対する認識性能は芳しくなく，改善が必要とされる．本研究では，単一の軽量なアダプターを用いる簡素な枠組みを採用しながら，多様な下流タスクに対する汎用的な認識性能の向上を実現し，視覚言語モデルの少数データ適応における新たなベースラインを提案する．

第3章　プロトタイプ適応器による学習不要かつ効率的な大規模視覚言語モデルの少数データ適応

3.1　導入

少数データ画像分類は，限られた数のラベル付きサンプルで訓練されたモデルを用いて，未見の画像を分類することを目的とする．このタスクは，医用画像診断や外観検査など，大量のデータを収集することが困難な分野において重要である．少数データ画像分類に対応する主なアプローチとして，データ拡張 [46, 31]，転移学習 [47, 48, 30] およびメタ学習 [12, 13] が存在する．これらの中でも，メタ学習は少数データ学習の効率を向上させるために発展してきた古典的な研究分野である．メタ学習は，タスク集合からタスクをサンプリングして少数データ学習のシナリオをシミュレートすることによって，ソース（メタ学習）データセットからデータ効率的な学習器を作成し，この特殊化された学習器を目的（メタテスト）セットに適用することに焦点を当てている．多くの手法 [49] は，目的セットとクラスは異なるがドメインが類似したメタ訓練セットを利用し，少数ショット訓練データよりも大きなサンプルサイズを提供している．例えば，miniImageNet [28] のメタ訓練，テストセットは ImageNet-1K [1] における異なるクラス集合で構成され，CIFAR-FS [50] は CIFAR-100 [51] を分割することによって形成される．メタ訓練セットを用いること

でメタ学習や事前学習などを行うことにより，少数ショット訓練データだけでは十分に得られないドメイン固有の知識を提供し，モデルの性能向上に役立てている．しかし，実応用においてはそのようなデータ集合が得られる可能性は限られていることが多く，目的クラスの少数ショット訓練セットのみに依存する少数ショット学習手法が必要とされている．そのため，推定対象とするクラスの少数データのみを目的タスクに関するデータとして用いた少数データ学習手法が必要とされている．

一方で，強力な転移性能を持った事前学習モデルである Contrastive Vision-Language Pre-training (CLIP) [19] が近年コンピュータビジョン分野で注目を集めている．CLIP は，大規模な画像-テキストペアに対する対照事前学習を通じて，任意のカテゴリのデータに対してゼロショット推論を可能にする．少数ショットデータを用いた線形識別により，下流タスクにおける分類性能をさらに向上できることが示されている．CLIP の少数データでの転移性能は，メタ学習セットのような目的タスクに関する追加のデータを必要としないという点で，実応用において優れていると言える．近年，CLIP の少数ショット認識性能をさらに向上させることを目的とした複数の研究が公開されている．CLIP-Adapter [21] は，画像エンコーダーとテキストエンコーダーの上層に残差接続のアダプター [23] を組み込むことで，CLIP を下流タスクに適応させる．このアプローチでは，事前学習されたエンコーダーの重みを固定し，下流タスクの少数学習データを用いてアダプターの重みのみを学習する．Context Optimization (CoOp) [20] は，CLIP において手動で設計されているテキストエンコーダーの入力プロンプトを学習可能な分散表現に置き換え，少数データを用いた最適化により改善する．一方，Tip-Adapter [22] は，キーバリューキャッシュモデルを用いてアダプターを構築することで，確率的勾配降下法による学習を行うことなく CLIP を下流タスクに適応することを可能にする．さらにファインチューニングが可能な場合，アダプターの重みをファインチューニングすることで少数データ分類性能を大幅に向上させることができる．しかし，アダプターのサイズが学習サンプル数に比例して

増大する枠組みであるため，学習データ数が増加したときにモデルのサイズが大きくなりすぎてしまう問題がある．例えば，新たな学習データが継続的に追加されるシステムでは，データの蓄積に伴ってキャッシュが肥大化し，メモリ消費の増大や推論の遅延を招くため，計算資源の限られたエッジデバイスやリアルタイム性が求められるシステムに組み込みづらいことが課題である．

本研究では，Tip-Adapter の学習不要な性質を引き継ぎつつ，その課題を克服した新たな CLIP の適応手法である Proto-Adapter を提案する．Proto-Adapter は CLIP に対して単一の線形層から成るアダプターを付加する．アダプターは，少数学習データにおける各クラスの特徴量を集約したプロトタイプベクトルを用いて初期化される．このアプローチにより，学習サンプル数に対してアダプターのサイズが一定に保たれる．

図 3.1 は，異なる少数データ設定における Tip-Adapter と Proto-Adapter の性能比較を示している．驚くべきことに，Proto-Adapter は簡素かつ軽量な構造を持つのにも関わらず，Tip-Adapter の性能を上回ることを我々は確認した．さらに，顔認識において広く使用される深層距離学習手法である Additive Angular Margin Penalty [52] を用いてアダプターの重みをファインチューニングすることで，少数データ性能をさらに向上させることを提案する．このペナルティの導入により，少数のデータを用いた学習でも識別性の高い決定境界を持つモデルが得られることを我々は期待している．提案手法の有効性を実証するため，ImageNet および他の 10 種類のデータセットにおいて少数データ画像分類の包括的な実験を行う．

要約すると，本研究の貢献は以下の通りである．

- 学習サンプル数に関わらず一定のアダプターサイズを維持する，CLIP のための新しい学習不要適応手法を提案する．これは各クラスの学習サンプルの特徴量を集約してアダプターの重みを構築することで実現される．
- 提案手法の性能は，出力ロジットにおけるクラス間の距離に対して制約を導入

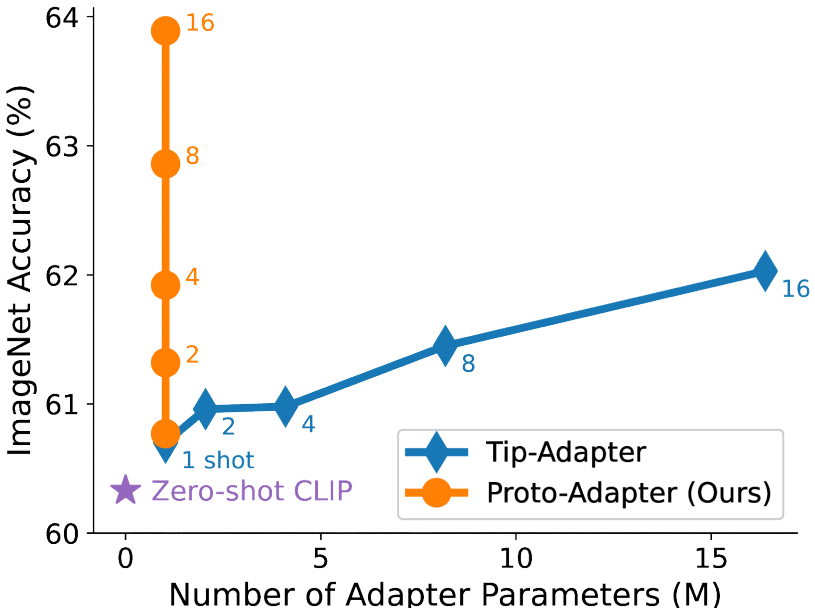

図 3.1: クラスごとの学習サンプル数を変化させたときの，ImageNet におけるパラメータ数と正解率の関係．両手法とも確率的勾配降下法による学習を行うことなく CLIP [19] を下流タスクに適応することができる．我々の Proto-Adapter は Tip-Adapter [22] を一貫して上回る性能を示しており，学習データ量に関わらず一定サイズかつ少数のアダプターパラメータを持つにも関わらずこの結果を達成している．

する Additive Angular Margin Penalty を用いてファインチューニングすることにより，さらに向上させることができる．

- ImageNet と他の 10 種類の画像認識データセットにおいて少数データ分類の性能を評価し，提案手法の既存の CLIP 適応手法に対する優位性を実証する．

3.2 手法

本節では，まず既存の少数データ適応手法である Tip-Adapter について簡潔に振り返る．その後，Tip-Adapter の課題点を解決している提案手法の詳細を述べる．

3.2.1 Tip-Adapter

Tip-Adapter [22] は，CLIP-Adapter [21] を学習不要かつノンパラメトリックに拡張した少数データ適応手法である．CLIP-Adapter に従い，重みの固定された事前学習済みの CLIP モデル [19] に軽量な 2 層の多層パーセプトロンを付加し，各入力画像に対して適応された，特徴量の残差成分を予測する．さらに，Tip-Adapter は少数データの学習セットからキーバリューキャッシュモデルを構築し，確率的勾配降下法での学習を行わずにノンパラメトリックな方法で，キャッシュをアダプターの多層パーセプトロンの重みに変換する．さらに，ファインチューニングが可能な場合，これらの重みをアダプターの初期値とし，アダプターのパラメータをファインチューニングすることで，認識性能を大きく改善することができる．

事前学習済みの CLIP モデルと少数データ分類のための N-class K-shot の訓練セットが与えられたとき，N 個のカテゴリのそれぞれに K 個のラベル付き画像があり，これらは $I_{N,K}$ として表され，そのラベルは $L_{N,K}$ と表される．Tip-Adapter はキーバリューキャッシュモデルを構築し，それに基づきアダプター層の重みを構築する．具体的には，全ての NK 個の学習サンプルに対して，CLIP を利用して少数データ学習画像から D 次元の L2 正規化された画像特徴量を抽出し，キー $F_{\text{train}} \in \mathbb{R}^{D\times NK}$ を下式のように作成する．

$$F_{\text{train}} = \text{CLIP}_{vis}(I_{N,K}) \tag{3.1}$$

ここで CLIP_{vis} は CLIP 画像エンコーダーを表す．値 $L_{\text{train}} \in \mathbb{R}^{N\times NK}$ は，少数学習データのラベル $L_{N,K}$ をワンホット符号化で変換することで得られる．

キャッシュモデルを構築した後，アダプターの重みはキャッシュされたキーとバリューの組を用いて構築される．そのため，クエリ画像に対するアダプターのロジッ

トは次のように作成される．

$$p_{\text{train}} = L_{\text{train}}\varphi(F_{\text{train}}^{T} f) \tag{3.2}$$

ここで $\varphi(x) = \exp(-\beta(1-x))$ はハイパーパラメータ β を持つ MLP の活性化関数を表し，$f \in \mathbb{R}^{D}$ は CLIP 画像エンコーダーによって抽出されたクエリ画像の L2 正規化された画像特徴量である．β は類似度の鋭さを制御する温度であり，値が大きいほど近い特徴量の寄与が強調される．

最終的な予測は，アダプターのロジット p_{train} と CLIP のゼロショットロジットの線形結合により得られる．

$$\text{logits} = \alpha p_{\text{train}} + W_{\text{text}}^{T} f \tag{3.3}$$

ここで α は混合係数であり，$W_{\text{text}} \in \mathbb{R}^{D\times N}$ は CLIP のテキスト分類器の重みを表す．ゼロショット CLIP に従い，W_{text} は各カテゴリ名を事前定義されたプロンプトテンプレートに当てはめ，それらを CLIP の事前学習済みテキストエンコーダーで符号化することによって構築される．

3.2.2 提案手法

我々は，Tip-Adapter [22] の学習不要な性質を保持しつつ，より簡素かつ軽量なアーキテクチャのアダプターを持つ Proto-Adapter を提案する．Proto-Adapter の全体的なパイプラインを図 3.2 に示す．各クラスのプロトタイプ表現を用いてアダプターの重みを構築することにより，Proto-Adapter におけるアダプターのサイズは小さく，学習サンプル数に対して不変である．さらに，Additive Angular Margin Penalty を適用してのファインチューニングにより，少数データ学習設定における認識性能をさらに向上させることができる．以下に提案手法の詳細を述べる．

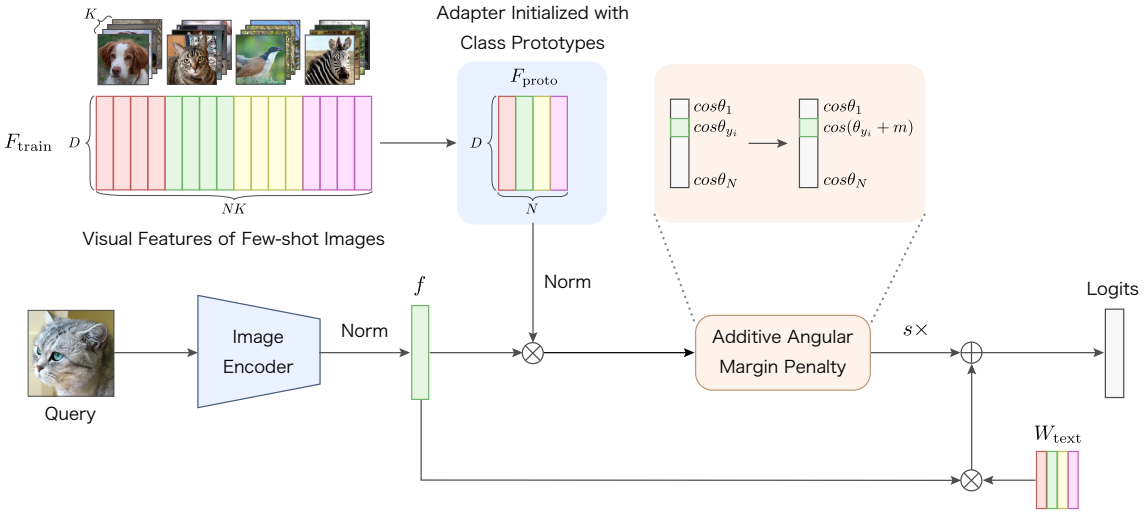

図 3.2: 提案する Proto-Adapter の概要．少数の学習画像から得られた視覚特徴量を集約して各クラスのプロトタイプベクトルが構築される．これらのベクトルは CLIP の画像エンコーダーの上部に配置されたアダプター層の重みとして機能する．さらなるファインチューニングが可能な場合，Additive Angular Margin Penalty [52] を用いてアダプター層をファインチューニングすることを提案する．このアプローチは，限られた学習データであっても，より識別性の高い予測を実現することを目的としている．

プロトタイプ表現に基づくアダプター

Proto-Adapter は，CLIP の画像エンコーダーの上部に単一層のアダプターのみを追加する．アダプターの重み $F_{\text{proto}} \in \mathbb{R}^{D \times N}$ は，各クラス $n \in \{1, \ldots, N\}$ の D 次元のプロトタイプベクトル $c_n \in \mathbb{R}^D$ を連結することによって作成される．

$$F_{\text{proto}} = [c_1, c_2, \cdots, c_N] \tag{3.4}$$

各プロトタイプベクトルは下式のように，CLIP の画像エンコーダーによって抽出された，あるクラスに属する学習画像全体の画像特徴量を平均化することにより作

成される．

$$c_n = \frac{1}{K} \sum_{k}^{K} \mathrm{CLIP}_{vis}(I_{n,k}) \tag{3.5}$$

ここで $I_{n,k}$ はクラス n の学習画像である．我々は，アダプターの重みを適切に正規化することがモデルの性能を大幅に改善することを発見した．具体的には，まずチャネル方向に特徴次元単位で L2 正規化を行い，続いてクラスごとの特徴ベクトル単位で L2 正規化を適用する．後者の特徴ベクトル単位の正規化は視覚言語モデルの推論において一般的に用いられるが，チャネル方向の正規化を併用することで，各チャネルの寄与を均一化し，事前学習で獲得された特徴表現をより有効に活用することを意図している．この正規化の有効性については，第 3.3.4 節のアブレーション実験で検証する．

視覚特徴量 f を持つクエリ画像が与えられたとき，アダプターのロジットは下式のように視覚特徴量とアダプターの重みの行列積により計算される．

$$p_{\text{proto}} = F_{\text{proto}}^{T} f \tag{3.6}$$

最終的な予測ロジットはアダプターのロジットと CLIP ロジットの線形結合として次のように得られる．

$$\text{logits} = \alpha p_{\text{proto}} + W_{\text{text}}^{T} f. \tag{3.7}$$

図 3.3 に Tip-Adapter と Proto-Adapter のアーキテクチャ比較を示す．提案する Proto-Adapter は，学習データにおけるクラスごとの画像特徴量をプロトタイプ表現に集約することにより，確率的勾配降下法による学習を行うことなくテストデータに対する推論が可能である．Proto-Adapter は Tip-Adapter において使用されている活性化関数や追加のハイパーパラメータを必要としない単一層のアダプターで構成されているため，事前学習済みの視覚言語モデルに対する最小限の変更で適用が可能である．

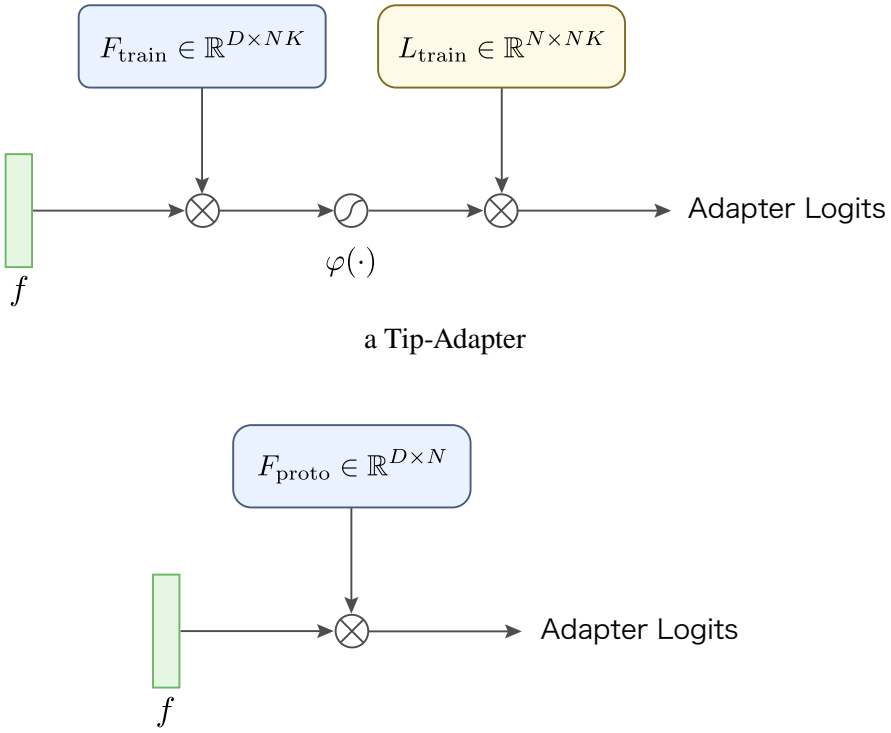

図 3.3: Tip-Adapter と Proto-Adapter のアーキテクチャの比較．(**a**) Tip-Adapter は 2 層の線形層と 1 つの活性化関数から構成されるアダプターを持ち，各線形層のサイズは学習データ数に比例して増大する．(**b**) 対照的に，Proto-Adapter は単一の固定サイズの線形層からなるアダプターのみで構成される．

Additive Angular Margin Penalty を用いたファインチューニング

Tip-Adapter と同様に，提案手法は少数の学習データを用いてアダプターの重みをファインチューニングすることにより，認識性能を向上させることが可能である．しかし，限られた数の事例を用いた学習で得られる決定境界は，多様なテストデータを正確に分類するのに十分な頑健性を持たない可能性があることを我々は懸念している．この問題に対処するため，顔認識において広く利用される距離学習技術を用いて，限られたデータであっても高い識別性を持つモデルを実現するファインチューニングのアプローチを導入する．具体的には，ArcFace [52] で提案された Additive

Angular Margin Penalty を用いてアダプターの重み $W = F_{\text{proto}}$ をファインチューニングする．

y_i 番目のクラスに属する i 番目のサンプルの正規化された視覚特徴量 f_i が与えられたとき，j 番目のクラスに対するロジットは $W_j^T f_i = \cos\theta_j$ として表すことができる．ここで $W_j \in \mathbb{R}^D$ はアダプターの重み W の j 番目の列を表し，θ_j は L2 正規化された重み W_j と特徴量 f_i の間の角度である．ArcFace に従い，f_i と W_{y_i} の間に Additive Angular Margin Penalty m を加えることで，正例クラス y_i に対するロジットを次のように取得する．

$$\text{logit}_{y_i} = \cos(\theta_{y_i} + m) \tag{3.8}$$

その後，得られたロジットに対してクロスエントロピー損失を適用し，アダプターの重み W をファインチューニングする．Additive Angular Margin Penalty を適用することにより，正例と負例の特徴空間での分離度が向上するようアダプターの重みが修正される．Additive Angular Margin Penalty は学習時のみに適用されるため，このアプローチによって推論時の枠組みが変わることはなく，推論時の計算コストを増加させることなく導入可能である．

3.3 評価実験

この節では，Proto-Adapter に関する包括的な評価実験を複数の画像認識ベンチマークで行い，その結果を報告する．

3.3.1 実験設定

データセット

提案手法を CLIP [19] で使用されている 11 種類の公開画像分類データセットで評価する．使用したデータセットは ImageNet [1]，StanfordCars [53]，UCF101 [54]，Caltech101 [55]，Flowers102 [56]，SUN397 [57]，DTD [25]，EuroSAT [58]，FGVCAircraft [24]，OxfordPets [59]，および Food101 [60] である．それぞれのデータセットの統計量を表 A.1 に，画像例を図 A.1 に示す．これらのデータセットは，一般的な物体，シーン，行動の分類から，より詳細なカテゴリ，さらにはテクスチャや衛星画像の識別などの特殊なタスクを含む，様々な視覚タスクを網羅している．

実装の詳細

Tip-Adapter と同様に，提案手法である Proto-Adapter には 2 つの方式が存在する．1 つは学習不要な方式で，もう 1 つは追加のファインチューニングを行う方式である．各方式は第 3.2.2 節で説明した方法に従って実装されている．本節では，これらの方式をそれぞれ Proto-Adapter および Proto-Adapter-F と呼称する．CLIP で使用されている少数データ学習の評価プロトコルに従い，1，2，4，8，16 ショットそれぞれでモデルを学習し，テストセット全体で評価を行う．CLIP のバックボーンについては，比較手法と一貫させるため，画像エンコーダーとして ResNet-50 [4] を，テキストエンコーダーとしてトランスフォーマー [61] を使用する．CLIP のテキストプロンプトを作成するためにプロンプトアンサンブル [19] を使用する．これは複数のテンプレートを CLIP のテキストエンコーダーに入力し，テキスト特徴量を平均化することにより成される．各データセットのテンプレートとして，Tip-Adapter で使用されているものと同じものを使用する．プロンプトテンプレートの一覧を表 B.1 に示す．Additive Angular Margin Penalty を用いてアダプターの重みをファインチュー

ニングする際は，最適化手法である Adam [62] を使用し，バッチサイズ 256 で 20 エポックにわたって Proto-Adapter を学習する．初期学習率を 4×10^{-4} とし，コサイン減衰学習率スケジュールに従って 4×10^{-5} まで減少させる．Angular Margin Penalty$m$ や α などの他のハイパーパラメータは，各データセットの検証セットを用いて調整する．Tip-Adapter と同様に，データ拡張としてランダムクロップ，リサイズ，およびランダム水平反転を行う．

3.3.2 ImageNet での性能比較

まず，画像分類の代表的なデータセットの一つである ImageNet [1] において，Proto-Adapter の性能を CLIP に基づく他の適応手法と比較する．比較手法は Zero-shot CLIP [19], Linear-probe CLIP [19]，CoOp [20]，CLIP-Adapter [21]，および Tip-Adapter [22] である．Zero-shot CLIP は下流データセットでの追加学習を必要としない．代わりに，テスト画像の画像特徴量と手動設計プロンプトのテキスト特徴量の類似度に基づいてゼロショット分類を行う．Linear-probe CLIP は元の画像エンコーダーを固定し，その上に接続されたロジスティック回帰による分類器のみを少数の学習データで学習する．CoOp はテキストプロンプトにおける context words を学習することで，プロンプトエンジニアリングを自動化する．我々は CoOp の中で最も性能の良い方式である統一コンテキスト方式を使用する．この方式は全てのクラスで同じコンテキストを共有する．CLIP-Adapter は，重みの固定された画像およびテキストエンコーダーの上部に追加された残差接続の特徴アダプターにより構成される，シンプルなアーキテクチャを持つ．Tip-Adapter は CLIP-Adapter と類似したアーキテクチャを持つが，画像ブランチのみにアダプターを付加し，少数学習セットの画像特徴量とラベルを用いて 2 層のアダプターを初期化する．これらの手法は全て，ResNet-50 [4] を画像エンコーダーとして用いた事前学習済み CLIP [19] に基づいている．Proto-Adapter と同様に，Zero-shot CLIP，CLIP-Adapter，および Tip-Adapter には 7 つのテンプレー

表 3.1: ImageNet における異なる少数データ設定での性能比較結果．すべての比較手法は，ResNet-50 を画像エンコーダーとして用いる事前学習済み CLIP に基づいている．‘FT’ はファインチューニングを表す．我々の Proto-Adapter は 1-shot 設定を除くすべての設定において他のすべての手法の性能を上回っている．

Method		**Shot**					
Models	**FT**	**0**	**1**	**2**	**4**	**8**	**16**
Zero-shot CLIP [19]		60.33	-	-	-	-	-
Tip-Adapter [22]		-	60.70	60.96	60.98	61.45	62.03
Proto-Adapter		-	**60.77**	**61.32**	**61.92**	**62.87**	**63.89**
Linear-probe CLIP [19]	✓	-	22.17	31.90	41.20	49.52	56.13
CoOp [20]	✓	-	47.62	50.88	56.22	59.93	62.95
CLIP-Adapter [21]	✓	-	61.20	61.52	61.84	62.68	63.59
Tip-Adapter-F [22]	✓	-	**61.32**	61.69	62.52	64.00	65.51
Proto-Adapter-F	✓	-	61.08	**62.05**	**63.05**	**64.49**	**66.17**

トを用いたプロンプトアンサンブルが使用されている．

表 3.1 に評価結果を示す．学習不要な CLIP 適応手法である Tip-Adapter と比較したとき，Proto-Adapter はより軽量なアダプター設計であるにも関わらず，すべての少数ショット設定においてそれを上回る性能を示している．性能の差は学習サンプル数が多いほど大きくなっており，1-shot 設定では+0.07 ポイントであるのに対して，16-shot 設定では 1.86 ポイントと大きな値となっている．このことは，Proto-Adapter が複数の学習サンプルの特徴量を効率的に軽量なアダプターに集約できていることを示している．また驚くべきことに，Proto-Adapter は学習不要な性質を持つにも関わらず，ファインチューニングを必要とする CLIP-Adapter と同等の性能を示しており，提案手法の有効性の高さを示している．さらに，提案手法においてファインチューニングを行った場合，Proto-Adapter-F は 1-shot 設定を除くすべての少数データ設定において，比較した手法の中で最も優れた性能を達成している．

3.3.3 複数のデータセットでの性能比較

表 3.2 は，ImageNet [1]，StanfordCars [53]，UCF101 [54]，Caltech101 [55]，Flowers102 [56]，SUN397 [57]，DTD [25]，EuroSAT [58]，FGVCAircraft [24]，OxfordPets [59]，および Food101 [60] の 11 種類のデータセットにおける提案手法の性能を示している．学習を必要としない CLIP の適応手法である Proto-Adapter や Tip-Adapter は，CLIP の平均精度を大幅に上回っている．特に改善幅が大きいのは FGVCAircraft や DTD，EuroSAT などの粒度の細かいカテゴリから構成されるデータセットである．この結果から，これらの CLIP 適応手法が少数の学習データの特徴量を有効に活用することにより，CLIP に対して効率的に新たなドメインの知識を付加することができていると言える．FGVCAircraft を除く全てのデータセットにおいて，Proto-Adapter は Tip-Adapter の性能を凌駕している．11 種類のデータセットにおける平均正解率は 72.69%であり，Tip-Adapter を+2.37 ポイントという大きな改善幅で上回っている．さらに，Proto-Adapter-F はより軽量なアダプターを持つのにも関わらず，Tip-Adapter-F と同程度の性能を示している．Proto-Adapter-F が Tip-Adapter-F より劣る FGVCAircraft や DTD などのデータセットは，飛行機やテクスチャの分類などの特殊な認識タスクとなっており，そのような特殊なデータセットに対する転移性能の改善が課題の一つであると言える．それでもやはり，これらの全体的な結果は Proto-Adapter が様々な下流タスクに対して軽量かつ効率的なアダプターを効率的に獲得できることを示していると言える．

3.3.4 アブレーション実験

本節では，Proto-Adapter に関する複数のアブレーション実験を行う．すべての実験は ImageNet で実施され，特に明言しない限り 16-shot 設定を採用する．

表 3.2: 16-shot 設定における 11 種類の画像分類ベンチマークでの性能比較結果．上段は学習不要な手法を示し，下段はファインチューニングを行う手法を示している．

	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101
Zero-shot CLIP [19]	60.32	85.92	85.83	55.74	66.02	77.32
Tip-Adapter [22]	62.01	90.43	88.14	66.77	89.89	77.83
Proto-Adapter	**63.89**	**91.85**	**88.55**	**70.35**	**93.06**	**78.73**
Tip-Adapter-F [22]	65.51	92.90	89.48	**75.49**	94.19	79.44
Proto-Adapter-F	**66.17**	**92.90**	**89.56**	75.00	**95.09**	**79.52**

	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	Avg
Zero-shot CLIP [19]	17.10	58.52	42.20	37.52	61.35	58.89
Tip-Adapter [22]	**29.76**	66.85	60.93	70.54	70.58	70.32
Proto-Adapter	27.21	**69.05**	**64.89**	**75.52**	**76.50**	**72.69**
Tip-Adapter-F [22]	**34.92**	71.43	**67.20**	**84.83**	78.54	**75.81**
Proto-Adapter-F	33.00	**71.82**	66.55	83.27	**78.56**	75.59

表 3.3: 異なる少数データ設定におけるプロトタイプベクトルの正規化の効果.

Normalization		Shot		
Channel-Wise	**Class-Wise**	**1**	**4**	**16**
		60.40	60.61	61.84
✓		60.65	60.96	61.75
	✓	60.66	61.34	62.70
✓	✓	**60.77**	**61.92**	**63.89**

プロトタイプベクトルの正規化

我々はまず，アダプターの重みを正規化することの影響を検証する．アダプターの重み $F_{\text{proto}} \in \mathbb{R}^{N\times D}$ の異なる軸に対して L2 正規化を適用した評価結果を表 3.3 に示す．すべての適用方法において，L2 正規化による性能改善が確認された．最も大幅な改善は，まずチャネル方向（第 1 軸）に正規化を適用し，続いてクラス方向（第 2 軸）に正規化を適用した場合に観察された．この 2 段階の正規化は，チャネル方向で各次元のスケール差を揃えて特定次元の過度な寄与を抑えつつ，クラス方向で各プロトタイプのノルムを統一することで，CLIP の埋め込みが前提とするコサイン類似度の幾何に整合させる効果がある．性能向上は学習データ数が多いほど大きく，最も学習データ量の多い 16-shot 設定において 2.05 ポイントの改善が確認された．これは，サンプル数が増えるほどプロトタイプの方向が安定し，ノルム差の影響が相対的に大きくなるため，正規化の効果がより顕著に現れたと考えられる．また，標準化などの他の正規化手法も検証したが，L2 正規化ほどの性能改善は見られなかった．標準化は平均と分散に依存して特徴ベクトルの方向を変えるため，方向情報を重視する CLIP の埋め込み空間とは必ずしも整合しない可能性がある．これらの結果は，提案手法においてクラスプロトタイプの正規化が極めて重要であることを示している．

表 3.4: Additive Angular Margin Penalty のマージンパラメータ m の効果．‘FT‘はファインチューニングを指す．

Shot	**1**	**2**	**4**	**8**	**16**
w/o FT	60.77	61.32	61.92	62.87	63.89
$m = 0$	60.99	61.66	62.40	63.79	65.56
$m = 0.1$	60.93	61.77	62.72	64.35	65.96
$m = 0.2$	60.86	61.97	62.99	**64.49**	**66.17**
$m = 0.3$	61.02	**62.05**	**63.05**	64.15	65.95
$m = 0.4$	**61.08**	62.01	62.82	63.95	65.54

Angular Margin Penalty *m*

次に，Angular Margin Penalty m がモデルの性能に与える効果を検証する．マージンパラメータが大きいほど，クラスの分離度がより大きくなるようモデルに制約が課される．表 3.4 に示す結果から，すべての少数データ設定において，Angular Margin Penalty を導入することにより認識性能が向上することが確認された．注目すべきことに，学習データ量が限られている場合，相対的に大きなマージンの有効性が大きくなっている．この結果は，学習データ量の限られる場面において，クラス間の分離を向上させることを目的とした距離学習が有効であることを示唆している．

画像エンコーダー

最後に，CLIP の画像エンコーダーが我々の手法に与える影響を調査する．表 3.5 は，様々なサイズの ResNet [4] と Vision Transformer (ViT) [61] を画像エンコーダーとして使用した性能比較結果を示している．我々の手法は，エンコーダーの種類やサイズに関わらず，一貫して比較手法を上回る性能を示している．特筆すべきことに，モデルのサイズが最も大きな ViT/16 を使用した場合において，我々の Proto-Adapter-F は Zero-shot CLIP を 5.5 ポイント上回っている．この結果は，Proto-Adapter が限ら

表 3.5: 16 ショット設定において様々な画像エンコーダーを用いたときの各手法の正解率（%）．Proto-Adapter は，画像エンコーダーに関わらず，一貫して比較手法の性能を上回っている．

Models	**RN50**	**RN101**	**ViT/32**	**ViT/16**
Zero-shot CLIP [19]	60.33	62.53	63.80	68.73
Tip-Adapter [22]	62.03	64.78	65.61	70.75
Proto-Adapter	**63.89**	**66.81**	**67.07**	**72.25**
CoOp [20]	62.95	66.60	66.85	71.92
CLIP-Adapter [21]	63.59	65.39	66.19	71.13
Tip-Adapter-F [22]	65.51	68.56	68.65	73.69
Proto-Adapter-F	**66.17**	**69.12**	**69.27**	**74.23**

れた量のラベル付きデータで構成される下流タスクに対して，大規模事前学習モデルに効果的に適応可能であることを実証している．

3.4 議論

我々が提案する Proto-Adapter は，汎用的な物体認識データセットである ImageNet において，ほぼ全ての実験設定において他の比較手法よりも優れた認識性能を示した．また，11 種類の画像分類データセットにおける平均性能は，ファインチューニングなしの設定で Tip-Adapter を大幅に上回り，ファインチューニングありの設定ではこれと同等の性能を示している．Proto-Adapter が Tip-Adapter と比べて軽量かつ一貫したサイズのアダプターを持つことを考えると，これらの結果は特筆すべきである．

性能向上の鍵は，プロトタイプベクトルの正規化にあることが確認された．この実験結果は，限られた学習データで視覚言語モデルを下流タスクに適応させる場合，アダプターのサイズよりもその適切な構築が重要であることを示唆している．

さらに，アダプターのファインチューニング時に Additive Angular Margin Penalty

を適用することで性能が向上することが確認された．この結果は，限られたデータでの学習において距離学習を導入することで，モデルの識別境界を改善できるという我々の仮説を支持している．

Proto-Adapter は汎用性の面においても優れている．画像エンコーダーに関わらず，CLIP モデルの少数ショット性能を大幅に向上させることができる．さらに，我々が提案したパイプラインは単一層のアダプターを追加するだけで，他の視覚言語モデルにも適用可能である．

さらに，Proto-Adapter は学習データ量に関わらずアダプターのサイズが一定であるとともに，保存すべき追加パラメータが少ないため，実利用シナリオにおいても展開が容易である．例えば，Tip-Adapter のようにサンプル数の増加に伴ってキャッシュが肥大化することがないため，継続的に学習データが追加される環境でも，メモリ消費や推論遅延を抑えたまま運用できる．

提案したアダプターを導入したモデルは，新しいクラスや学習データが追加された際に容易に更新することができる．ファインチューニングが不要な場合，式 (3.4) と (3.5) に基づき，確率的勾配降下法による学習を行うことなくアダプターを構築することができる．ファインチューニングを行う場合でも，計算コストの高い画像およびテキストエンコーダーへの誤差逆伝播は不要で，アダプター層の重みのみを更新するだけで済むため，極めて高速な学習が可能である．例えば，ImageNet の 16-shot 設定での学習は，単一の NVIDIA GeForce RTX 3090 GPU を用いて 8 分で完了することが確認された．

Proto-Adapter は既存の CLIP 適応手法と比較して優れた性能を示しているが，ドメイン特有の知識を必要とする特殊なタスク（例：FGVCAircraft や DTD）における性能には依然として改善の余地がある．また，外れ値がモデルの性能に悪影響を及ぼす可能性が懸念される．本研究では簡素な枠組みを維持するためこれらの問題に明示的に対処していないが，学習データが持つバイアスや外れ値への対応が今後の

研究課題として残されている．

3.5 本章のまとめ

本研究では，CLIP を下流タスクに対して限られた学習データを用いて適応させる新たな手法である Proto-Adapter を提案した．提案手法は，クラスごとのプロトタイプベクトルをアダプターの重みとして利用することにより，先行研究である Tip-Adapter における課題となっている，学習データの量に依存してアダプターのサイズが増大する問題を解決した．軽量かつ簡素なモデル構造であることにも関わらず，我々の手法は多くの画像分類ベンチマークにおいて Tip-Adapter を上回る性能を達成した．さらに，一般的に距離学習において用いられる技術である Additive Angular Margin Penalty をアダプターのファインチューニング時に適用することで，モデルの性能をさらに向上させることができることを実証した．本研究には様々な方向性への発展が可能である．今後の研究として，より効果的なプロトタイプベクトルの構築方法の探究，少数事例から多数事例シナリオまで学習データ量に関わらず効果的に適用可能な枠組みの構築，および我々のプロトタイプ表現に基づくアダプターの領域分割などのより高度なタスクへの拡張を構想している．我々は，この研究がそのような様々な方向性において，今後の研究の基礎となることを期待している．

第4章　線形識別器の残差学習による大規模視覚言語モデルの少数データ適応

4.1　導入

自動運転や様々な視覚タスクの自動化などに活用される画像認識技術は，近年の深層学習の発展に伴い急速に進歩している．深層学習では教師ラベルが付与された大規模なデータを用いて多層のニューラルネットワークを学習することで，多様なタスクを高い精度で解くことが可能である [2, 63, 64]．しかし，大量のデータへラベルを付与するアノテーション作業は多大な人的労力を要するとともに，医用画像解析や外観検査など，適用先によってはデータの収集自体が難しく，大規模な訓練データセットを構築することが困難な場合がある．

小規模なデータセットを用いた機械学習のアプローチとして，少数データ学習が存在する．少数データ学習では，僅かなラベル付きデータを用いて学習したモデルにより，未知のデータを正確に識別することを目的とする [10, 11]．少数データ学習による画像分類へのアプローチとして，データ拡張 [46, 31]，転移学習 [47, 48, 30]，メタ学習 [12, 13] などが挙げられる．それらの中でも，大規模視覚言語モデルの一種である Contrastive Vision-Language Pre-training (CLIP) [19] を下流タスクの少数のラベル付きデータで調整するアプローチ [20, 21, 22] の有効性が確認されている．CLIP

はインターネットから収集した画像とテキストのペアから成る大規模なデータセットで事前学習されており，視覚及び言語に関する頑健な特徴表現を獲得している．未知のクラスを含む下流タスクに対してゼロショットで推論することが可能であるが，下流タスクの少数データを用いた適応により，目的とするドメインに対する認識性能を改善できることが複数の研究で確認されている．代表的な研究事例として，プロンプト調整手法である Context Optimization (CoOp) [20] は，CLIP において手動で設計しているテキストプロンプトを学習可能なベクトルに置き換え，少数の学習サンプルでそのパラメーターを最適化する．CLIP-Adapter [21] は，画像エンコーダーとテキストエンコーダーの上層に残差接続のアダプター [23] を追加することで CLIP を下流タスクに適応させる．しかし，CLIP において大規模なデータから獲得された画像特徴量とテキスト分類器の整合性が崩れ，汎化性能を損なってしまうことが懸念される．Tip-Adapter [22] は，少数データの画像特徴量及びラベルから成るキーバリューキャッシュで初期化したアダプターを CLIP の上層に追加する．アダプターの重みを調整することで，少数データでの分類性能を大きく向上させることができる．しかし，アダプターのサイズは学習サンプル数に依存して変化するため，データ数の変動する実運用シーンにおいて利用しづらいことが課題である．またいずれの手法においても，多様な下流タスクに対する汎用的な適応性能には改善の余地が存在する．特に，飛行機 [24] やテクスチャ[25] の識別などのドメイン固有の知識を必要とするタスクへの認識性能が相対的に低下しやすい傾向にあり，その改善が必要である．

本研究では多様なドメインの下流タスクに対する汎用的な認識性能の改善を目的とし，線形識別器の残差学習による大規模視覚言語モデルの少数データ適応手法である Residual-Adapter を提案する．提案手法は CLIP の画像エンコーダーの上層に線形識別器をアダプターとして挿入し，CLIP の予測ロジットと線形識別器の予測ロジットの線形結合により画像の識別を行う．学習時は各種エンコーダーのパラメー

ターを凍結し，推論時と同一の推論経路を用いて線形識別器のパラメーターのみを確率的勾配降下法を用いて学習する．このような，事前学習されたエンコーダーのパラメーターを一切変更することなく，CLIP の予測値と教師ラベルの残差成分を学習する枠組みにより，視覚言語事前学習により獲得された頑健な特徴表現に基づく推論性能を保持しつつ，新たなドメインの知識を効率的にモデルに取り入れることが可能である．また，提案手法は画像エンコーダーの上層に追加した線形識別器のみを学習するため，学習時に CLIP の各エンコーダーへの誤差の逆伝播が必要なく，非常に高速な学習が可能である．我々は 11 種類の画像認識データセットを用いて提案手法の有効性を検証する実験を行った．提案手法は複数の少数データ設定において比較手法を上回る平均正解率を達成し，簡素な枠組みを持ちながら多様なドメインの下流タスクに対して高い汎用性を発揮することが確認された．

4.2 手法

4.2.1 視覚言語モデル

本研究において少数データ適応を図る視覚言語モデルである CLIP [19] について簡単に説明する．なお，提案手法は CLIP に限らず，ALIGN [41] や SigLIP [65, 66] などの CLIP と同様の枠組みを持つ任意の視覚言語モデルへ適用可能である．

CLIP は画像エンコーダーと言語エンコーダーの 2 種類のエンコーダーで構成される． 画像エンコーダーは入力画像に対して特徴抽出を行い，低次元な特徴空間へと写像する．一方，テキストエンコーダーは入力文字列（トークン列）を画像特徴量と次元数の等しい特徴空間へと埋め込む．これらのエンコーダーは，画像とテキストそれぞれの埋め込み空間を整合させるよう訓練される．具体的には，画像とテキストのペアから成るバッチに対し，一致するペアの埋め込みベクトル同士のコサイン類似度を最大化し，他の全ての不一致ペアのコサイン類似度を最小化するよう，対

照損失を目的関数に学習を行う [42, 43]．多様な視覚概念を学習するために，CLIP はインターネットから収集された 4 億の画像，テキストペアからなる大規模な訓練データセットで学習されている．

ゼロショット推論

CLIP は画像とテキストの説明が一致するかどうかを予測するように事前学習されているため，任意のクラスをゼロショットで認識することが可能である．これは，画像エンコーダーによって抽出された画像特徴量を，テキストエンコーダーによって作成されたテキスト分類器と比較することで実現される．具体的には，テキストエンコーダーは対象となるクラスを指定する「a photo of a [CLASS].」などの形式のプロンプトを入力文字列として受け取る．ここでクラストークンは「cat」「dog」などの具体的なクラス名で置き換える．D を特徴ベクトルの次元数，N をクラス数とするとき，テキスト分類器は $W_{\text{text}} = [w_1, w_2, ..., w_N] \in \mathbb{R}^{D \times N}$ のように，テキストエンコーダーにプロンプトを入力して出力された各クラスの特徴ベクトル w_i を並べて作成される行列である． CLIP における各クラスの予測ロジットは式（4.1）のように，テキスト分類器と画像特徴量 f の行列積に基づき算出される．

$$\text{logits} = W_{\text{text}}^T f / \tau \tag{4.1}$$

ここで τ はエンコーダーのパラメーターとともに学習される温度パラメーターである．

分類器を直接学習することで閉集合の視覚概念を学習するアプローチと比較して，CLIP はテキストエンコーダーを用いた視覚言語事前学習を通じて開集合の視覚概念の学習を図っており，陽に学習していないクラスをゼロショットで認識することを可能にしている．

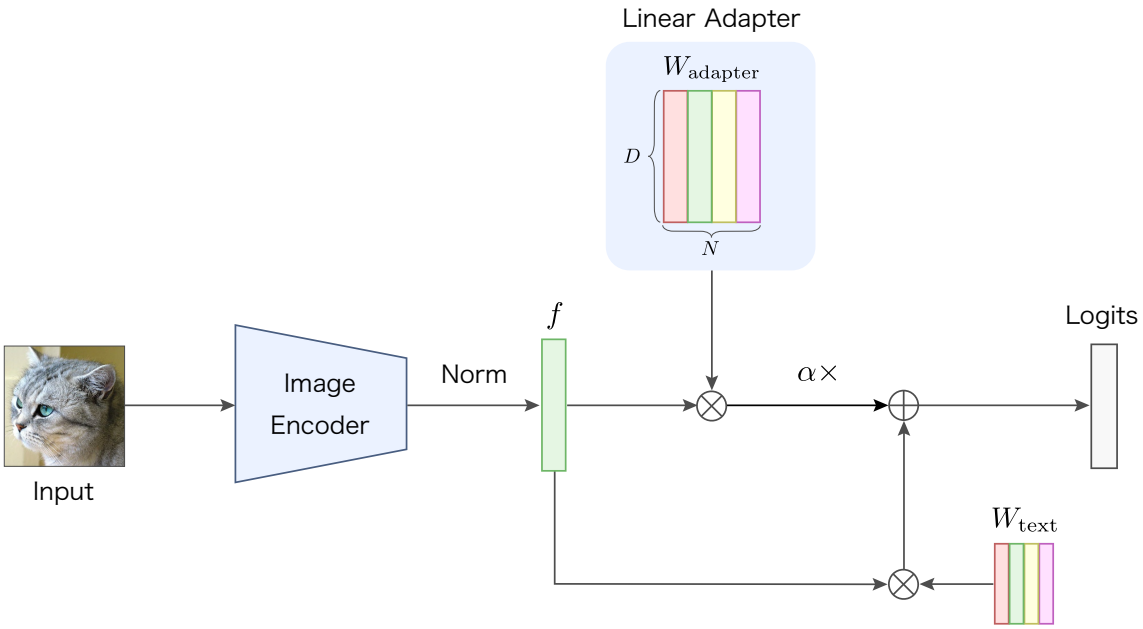

図 4.1: Residual-Adapter の全体構成図．本手法では CLIP の画像エンコーダーの上層に線形識別器をアダプターとして挿入する．CLIP のロジットと線形識別器のロジットの線形和により最終的な推定結果を出力する．少数データでの学習時は，CLIP の各種エンコーダーを凍結し，アダプターのパラメーターのみを確率的勾配降下法で最適化する．

4.2.2　線形識別器の残差学習

CLIP のような視覚言語モデルはゼロショット推論が可能な機構を持つものの，その認識性能には改善の余地が存在する．特に，ドメイン固有の知識を要するタスクに対する認識性能が低下しやすいことが課題である．視覚言語モデルを学習データの少ない下流タスクに対して高精度に適応させるために，我々は線形識別器の残差学習による視覚言語モデルの少数データ適応手法である Residual-Adapter を提案する．

モデル構成

Residual-Adapter の枠組みを図 4.1 に示す．我々は，CLIP の画像エンコーダーの上層に線形識別器 $W_{\text{adapter}} \in \mathbb{R}^{D\times N}$ を新たに追加する．この線形識別器は少数データ学

習を通じて下流タスクの知識を取り入れるアダプターとして機能する．提案手法における予測ロジットは式（4.2）のように，線形識別器による予測ロジットと CLIP の予測ロジットの線形結合により算出される．

$$\text{logits} = \alpha W_{\text{adapter}}^{T} f + W_{\text{text}}^{T} f / \tau \tag{4.2}$$

ここで，α は最終的な予測に対して線形識別器の予測をどれだけ重視するかを決定するスケーリングパラメーターである．我々は実験的に，適応先タスクに適したスケーリングパラメーターを設定することが認識性能に大きな影響を与えることを発見した．

提案手法の学習時は，視覚言語モデルの各種エンコーダーのパラメーターを固定し，追加した線形識別器のパラメーターのみを下流タスクの少数データを用いて学習する．一般的な分類タスクの学習に用いられるクロスエントロピー損失を目的関数とし，確率的勾配降下法を用いて線形識別器のパラメーターを学習する．その際，線形識別器の重みは 0 で初期化する．これにより，学習初期の予測ロジットは CLIP の予測ロジットと一貫し，CLIP の予測性能を初期値として学習を開始することができる．この初期化により，学習の初期段階から一定の性能を確保することで学習の安定化を図っている．本手法はモデル上層の線形識別器のみを学習するため，エンコーダーへの誤差の逆伝播を必要とせず，非常に高速な学習が可能である．例として，単一の NVIDIA GeForce RTX 3090 GPU を使用した場合，画像エンコーダーに ResNet-50 [4] を用いた際の ImageNet データセットにおける 16-shot 設定の学習は約 6 分で完了する．

このようなモデル構造及び学習の枠組みにより，Residual-Adapter は CLIP の予測ロジットと正解ラベルの残差成分を学習している．既存手法と Residual-Adapter のアーキテクチャ比較を図 4.2 に示す．Residual-Adapter は CLIP-Adapter [21] とは異なり CLIP の推論経路を保持しているため，視覚言語事前学習によって獲得された頑

健な特徴表現および特徴空間の整合性を損なうことなく，下流タスクに特有のドメイン知識を少数データから効率的に獲得することが可能である．また，提案手法は Tip-Adapter [22] において多層パーセプトロンで構築されているアダプターを線形識別器に置き換え簡素化した構造を持ちながら，多くの場合においてそれを上回る性能を発揮することを後述の実験にて示す．下流タスクごとでの適切なスケーリングパラメーターの設定により多様なタスクに適応できることが，提案手法の優位性に寄与していると考えられる．

線形識別器の統合

提案手法で導入する線形識別器は CLIP のテキスト分類器と同一サイズの行列であるため，推論時はこれらの重みを統合することが可能である．重みを統合する場合の予測ロジットの算出方法は式 (4.3) で表される．

$$\begin{aligned} \text{logits} &= (\alpha W_{\text{adapter}} + W_{\text{text}}/\tau)^T f \\ &= W_{\text{merged}}^T f \end{aligned} \tag{4.3}$$

重みの統合により，提案手法は CLIP と同一のモデル構造となる． この統合された重みは，CLIP の持つ汎用的な言語表現に，少数データから新たに得られるドメイン知識を融合させたものであると解釈できる．

4.3 評価実験

本章では，多様なベンチマークを用いて少数データ画像分類に関する包括的な評価実験を行い，提案手法の有効性を検証する．

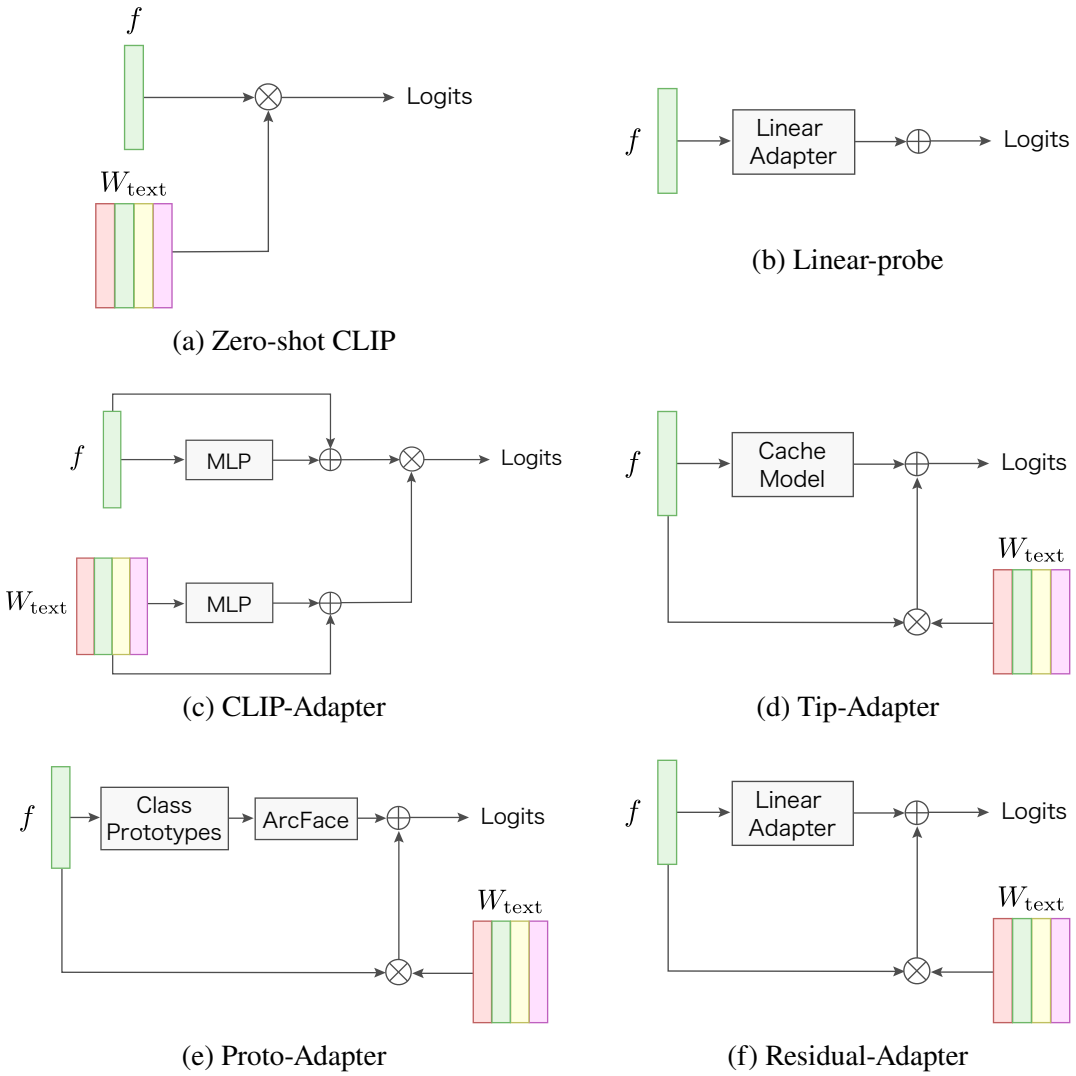

図 4.2: CLIP 適応手法のアーキテクチャ比較．f は CLIP の画像特徴量を，W_{text} は CLIP のテキスト分類器を表す．(a) Zero-shot CLIP：画像特徴量とテキスト分類器の行列積により予測ロジットを算出する．(b) Linear-probe：画像特徴量に対して線形識別器を適用する．(c) CLIP-Adapter：画像特徴量とテキスト分類器それぞれを残差接続付きの多層パーセプトロンを用いて下流タスクへ適応させる．(d) Tip-Adapter：学習データから構築したキーバリューキャッシュモデルと CLIP の予測ロジットを線形結合する．必要に応じてキャッシュモデルを微調整する．(e) Proto-Adapter：学習データから構築したクラスプロトタイプと CLIP の予測ロジットを線形結合する．必要に応じて，ArcFace を用いてクラスプロトタイプを微調整する．(f) Residual-Adapter：線形識別器と CLIP の予測ロジットを線形結合する簡素な適応の枠組みを持つ．

4.3.1 実験設定

評価方法

公開データセットを用いて画像識別の評価実験を行う．使用するデータセットは，ImageNet [1]，StanfordCars [53]，UCF101 [54], Caltech101[55]，Flowers102 [56], SUN397 [57]，DTD [25]，EuroSAT [58]，FGVCAircraft [24]，OxfordPets [59]，Food101 [60] の 11 種類である．それぞれのデータセットの統計量を表 A.1 に，画像例を図 A.1 に示す．これらのデータセットは，一般的な物体やシーン，行動の分類に加え，詳細なカテゴリーの分類や，テクスチャや衛星画像の認識といった特殊なタスクを含み，画像識別に関する包括的なベンチマークを構成している．学習データ設定として，N 個の各予測対象クラスに K 枚のラベル付き画像が含まれている K-shot N-class 設定を用いる． CLIP [19] で用いられている少数データ学習の評価プロトコルに従い，学習セットにおける 1，2，4，8，16-shot いずれかの設定でモデルを学習し，テストセット全体で評価を行う．

比較手法

評価実験において，既存の CLIP に基づく少数データ適応手法と提案手法である Residual-Adapter の性能を比較する．比較手法は，Zero-shot CLIP [19]，CoOp [20], CLIP-Adapter [21]，Tip-Adapter-F [22] 及び Proto-Adapter [26] である．Zero-shot CLIP は，画像とテキストのペアから成る大規模なデータセットで事前学習された視覚言語モデルである．このモデルは，テスト画像から抽出した画像特徴量と各クラスのテキスト特徴量との間のコサイン類似度に基づき，追加の学習を必要とせずにゼロショットでの分類を行う．CoOp は，CLIP におけるプロンプトエンジニアリングを自動化することを目的とし，テキストプロンプト内のコンテキストワードを少数データを用いて学習する手法である．本実験では，全クラスで共通のコンテキストを共有す

る統一コンテキスト方式の CoOp を採用している．これは，CoOp の各バリエーションの中で最も高い性能を示すことが知られている．CLIP-Adapter は，事前学習済みの画像及びテキストエンコーダーを固定したまま，それらの上層に残差接続を持つアダプター層を追加することで，特徴量を下流タスクへと適応させる．Tip-Adapter は，少数の学習データから構築したキーバリューキャッシュモデルを用いて，CLIP の特徴量を下流タスクに適応させる手法である．確率的勾配降下法による学習を必要とせずに適応が可能であるが，キャッシュモデルをファインチューニングした Tip-Adapter-F ではさらなる性能向上を実現できる．Proto-Adapter は，各クラスのプロトタイプ表現に基づいて固定サイズのアダプターを初期化する．Tip-Adapter と同様に学習を必要としない適応が可能であるが，Angular Margin Penalty [52] を課したファインチューニングを行うことで，より高性能な適応を達成している．これらの手法は全て，ResNet-50 [4] を画像エンコーダーとする事前学習済み CLIP [19] に基づいている．提案手法と同様に，Zero-shot CLIP，CLIP-Adapter，Tip-Adapter，Proto-Adapter では後述するプロンプトアンサンブルを使用している．

4.3.2 実装の詳細

提案手法は CLIP における任意のバックボーン及びテキストプロンプトを利用可能である．CLIP のバックボーンには比較手法と同様に，画像エンコーダーに ResNet-50 [4]，テキストエンコーダーに Transformer [67] を用いる．テキスト分類器の作成にはプロンプトアンサンブル [19] を用いる．これは，複数のプロンプトテンプレートを CLIP のテキストエンコーダーに入力し，それらのテキスト特徴量を平均化する手法である．我々は各データセットに対して Tip-Adapter [22] と同一のテンプレートを使用する．各データセットに対して使用するプロンプトテンプレートの一覧を表 B.1 に示す．

少数データを用いたアダプターの学習では，$\beta_1 = \beta_2 = 0.9$ とする AdamW [68, 62]

表 4.1: 各種少数データ設定における平均正解率．全ての比較手法は ResNet-50 を画像エンコーダーとする事前学習済みの CLIP に基づいている．

Method	**Shot**				
	1	2	4	8	16
Zero-shot CLIP [19]			58.89		
CoOp [20]	59.59	62.32	66.77	69.89	73.42
CLIP-Adapter [21]	62.67	65.55	68.61	71.40	74.44
Tip-Adapter-F [22]	**64.60**	**66.65**	69.67	72.45	75.83
Linear-probe (Our impl.)	39.78	50.47	61.27	67.33	73.52
Residual-Adapter	63.49	65.96	**69.80**	**72.82**	**76.44**

を最適化手法として利用し，バッチサイズ 256 で 20 エポックの学習を行う．学習率は最初の 1 エポックで 0 から 4e-4 まで線形のウォームアップを行い，その後はコサインスケジュールに従い減衰させる．τ は 200 とし，スケーリングパラメーター α は検証セットを用いてデータセットごとに調整する．データ拡張にはランダムクロップ，リサイズ及びランダムな水平反転を使用する．

4.3.3 提案手法の有効性の検証

提案手法は，CLIP による推論経路と線形識別器による推論経路の 2 つの推論経路を持つ．提案手法から線形識別器の推論経路を取り除くと CLIP と等価となり，CLIP による推論経路を取り除くと線形識別器のみによる分類となる．本節では，提案手法の性能を CLIP 及び線形識別器と比較することで，提案した残差学習アプローチの有効性を検証する．各種少数データ設定での 11 種のデータセットにおける平均正解率を表 4.1 に示す．表中の Linear-probe は，提案手法から CLIP による推論経路を取り除き，線形識別器のみで学習及び推論を行った結果を表す．線形識別器単体の場合，4-shot 以上の学習データ数において CLIP の性能を上回るものの，それ未満の

表 4.2: 11 種類の画像分類ベンチマークにおける 16-shot 設定での性能比較結果. 提案手法が最も高い平均性能を示している.

	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101
Zero-shot CLIP [19]	60.32	85.92	85.83	55.74	66.02	77.32
CoOp [20]	62.95	91.83	87.01	73.36	94.51	74.67
CLIP-Adapter [21]	63.59	92.49	87.84	74.01	93.90	78.25
Tip-Adapter-F [22]	65.51	92.86	**89.70**	75.74	94.80	79.43
Proto-Adapter-F [26]	**66.17**	92.90	89.56	75.00	95.09	**79.52**
Linear-probe (Our impl.)	59.38	92.49	70.99	73.56	95.01	73.49
Residual-Adapter	65.20	**93.35**	89.59	**76.51**	**95.17**	79.19

	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	Avg
Zero-shot CLIP [19]	17.10	58.52	42.20	37.52	61.35	58.89
CoOp [20]	31.26	69.26	63.58	83.53	75.71	73.42
CLIP-Adapter [21]	32.10	69.55	65.96	84.43	76.76	74.44
Tip-Adapter-F [22]	35.55	71.47	66.55	84.54	78.03	75.83
Proto-Adapter-F [26]	33.00	**71.82**	66.55	83.27	78.56	75.59
Linear-probe (Our impl.)	37.26	67.97	**67.85**	85.12	76.63	73.52
Residual-Adapter	**37.29**	70.73	**67.85**	**85.89**	**80.02**	**76.44**

データ数では CLIP に劣る結果となった．特に 1-shot 設定では大幅な性能低下が見られた．これは，線形識別器はテキストエンコーダーによるテキスト分類器を利用しないため，事前学習で獲得された知識を十分に活用できないことが要因と考えられる．一方，提案手法は全ての学習データ数において CLIP を上回る性能を示しており，1-shot 設定においてすでに CLIP の平均正解率を 4.60 ポイントと大幅に上回っている．学習データの増加に伴い更なる性能の向上が見られ，16-shot 設定では CLIP の性能を 17.55 ポイント上回る 76.44%の平均正解率を達成している．これらの結果は，提案手法が CLIP の頑健なゼロショット推論能力と線形識別器による適応を効果的に組み合わせることで，極めて少数の学習データからでも有効な適応を実現していることを示している．

4.3.4 既存手法との性能比較

提案手法の性能を CLIP に基づく既存の少数データ適応手法と比較する．表 4.1 に示す結果から，提案手法は全ての少数データ設定において CoOp 及び CLIP-Adapter の性能を上回っていることが分かる．これは，提案手法が CLIP の推論経路を維持することで特徴空間の整合性を保った適応を行うためであると考えられる．さらに，既存手法の中で最も高い性能を示す Tip-Adapter を 5 つ中 3 つの少数データ設定で凌駕している．提案手法は下流タスクごとに適切なスケーリングパラメーターを設定することで多様なタスクへの適応性能を高めているため，Tip-Adapter と比べて簡潔なモデル構造や学習の枠組みを持ちつつも，同等以上の性能を達成できていると考えられる．16-shot 設定でのデータセットごとでの各手法の性能比較結果を表 4.2 に示す．データセットによって最も性能の高い手法が異なる結果となった．例えば，一般物体認識タスクである ImageNet では Proto-Adapter が最良の性能を示している一方，ドメイン固有の知識が求められる FGVCAircraft では比較手法の中で相対的に性能が低く，手法によって得意及び不得意な領域が存在することが示唆される．一

表 4.3: 各種画像エンコーダーを用いたときの平均正解率．16-shot 設定．

Models	RN50	RN101	ViT-B/32	ViT-B/16
Zero-shot CLIP [19]	58.89	59.68	62.18	65.53
Linear-probe (Our impl.)	73.52	75.98	75.72	79.96
Residual-Adapter	**76.44**	**77.52**	**77.79**	**81.44**

方，提案手法はデータセットに依存することなく常に良好な性能を示している．11 のデータセット中 7 つで最も優れた性能を達成し，最も高い平均正解率を達成している．提案手法は比較手法と比べて簡素な枠組みながら高い平均性能を示しており，広範なドメインに対して汎用性の高い手法であるといえる．

4.3.5 構成要素の比較実験

本節では，提案手法の構成要素に関する比較実験を行う．全ての実験は 16-shot 設定を採用し，11 種類の画像分類ベンチマークにおける平均正解率を報告する．

画像エンコーダー

CLIP の画像エンコーダーが提案手法に与える影響を調査するため，様々なサイズの ResNet [4] 及び Vision Transformer (ViT) [61] を画像エンコーダーとして用いた評価検証を行う．表 4.3 に示す結果から，提案手法はエンコーダーの種類やサイズによらず，一貫して CLIP や線形識別器単体の性能を上回ることが確認された．エンコーダーの規模が大きくなるにつれて CLIP のゼロショット性能は向上するが，提案手法と CLIP の性能差はエンコーダーのサイズによらず概ね一定に保たれている．最も計算量の多い ViT-B/16 を用いた場合においても，提案手法は CLIP の性能を 15.91 ポイント改善しており，提案した少数データ適応手法が大規模なエンコーダーに対

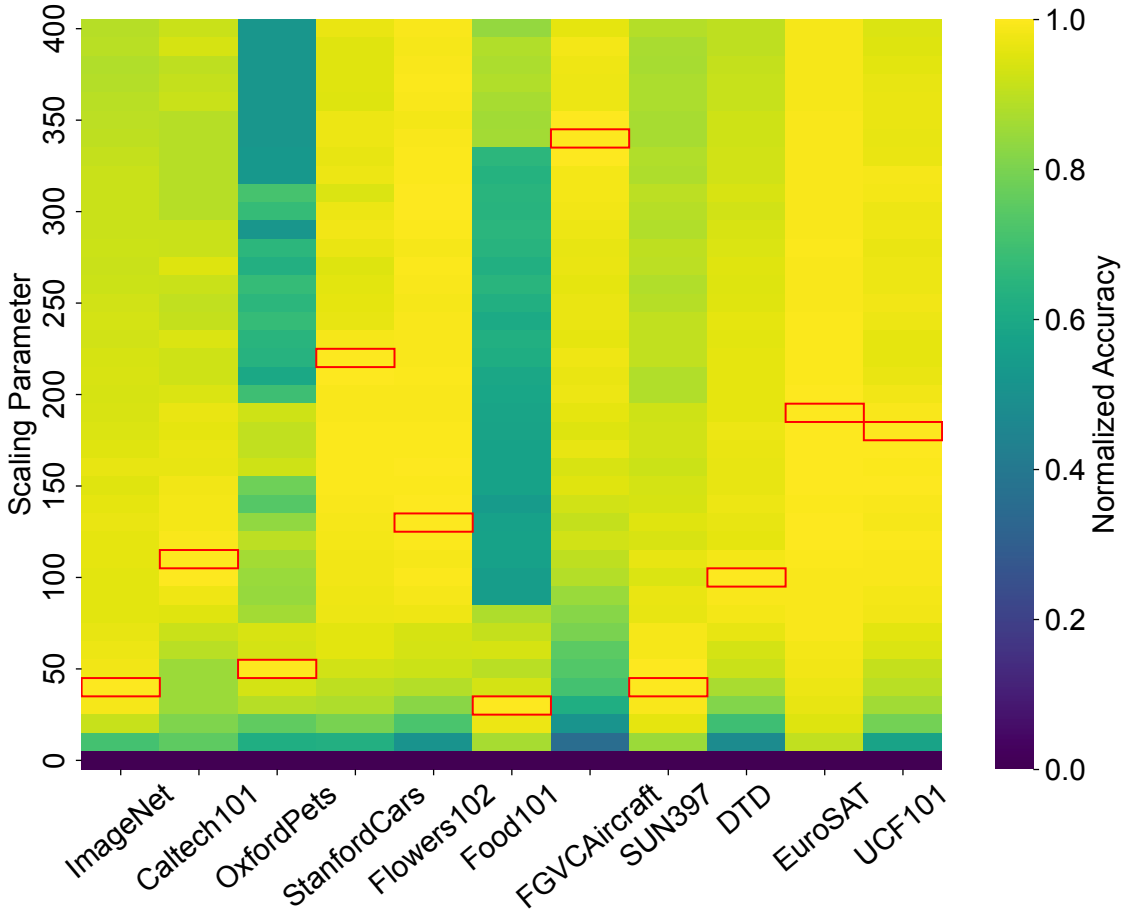

図 4.3: スケーリングパラメーターによる性能変化．データセットごとに正解率の上限を 1，下限を 0 に正規化して表示．各データセットにおいて最良の結果を赤枠で囲んでいる．

しても効果的に機能することが確認された．これらの結果は，提案手法がエンコーダーの表現力に依存することなく，CLIP に下流タスクのドメイン知識を適切に付加できることを示唆している．

スケーリングパラメーター

スケーリングパラメーターは，CLIP とアダプターそれぞれの予測ロジットの混合比率を制御するパラメーターである．スケーリングパラメーターが小さいほど，

CLIP が事前学習で獲得した知識を重視した推論が行われ，パラメーターが大きいほど，下流タスクから得られた知識が重視される．図 4.3 に示すデータセットごとでのスケーリングパラメーターと正解率の関係を表すヒートマップから，データセットによって最適なスケーリングパラメーターが異なることが分かる．例えば，ImageNet や Food101 では 50 未満の相対的に小さな値で最高性能を示すのに対し，StanfordCars や FGVCAircraft では 200 以上の大きな値が最適となっている．これは，データセットの性質によって CLIP の事前知識と下流タスクの知識の最適な組み合わせ比率が異なることを示唆している．また図 4.4 に示すように，CLIP に対する提案手法の性能改善比率が大きいデータセットほど，大きなスケーリングパラメーターが最適となる傾向が観察された．特に性能改善比率の大きな FGVCAircraft と EuroSAT はそれぞれ航空機及び衛星画像の分類タスクであり，これらの認識にはタスク特有の知識が重要となる．これらの結果は，CLIP が事前学習で獲得していない知識が重要なタスクにおいて，下流タスクから学習した知識を重視した推論を行うことが効果的であることを示唆している．このことから，十分な検証データを確保することが難しい場合は，事前学習タスクと下流タスクの相違度が大きいほど大きなスケーリングパラメーターを設定することが有効であると考えられる．

4.4 本章のまとめ

本研究では，CLIP の少数データ適応において多様なドメインの下流タスクへの汎用性の改善を目的とし，線形識別器を用いた残差学習に基づく手法である Residual-Adapter を提案した．提案手法は，CLIP の予測ロジットと正解ラベルの残差成分を線形識別器のアダプターで学習することにより，CLIP の頑健な特徴表現を維持しつつ，新たなドメインの知識を少数の学習データから効率的に導入することを可能とした．実験結果から，提案手法は複数の少数データ設定において既存の手法を上回

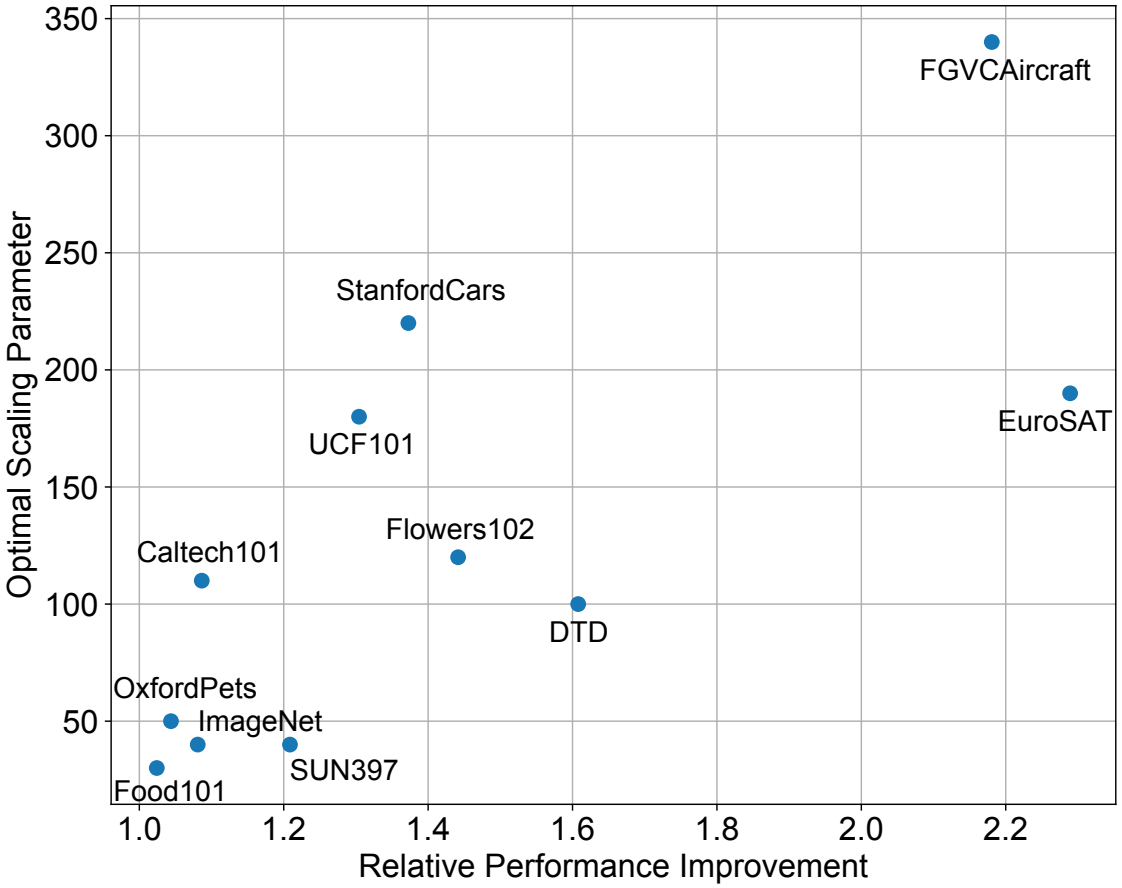

図 4.4: CLIP に対する提案手法の相対的な性能改善比率と最適なスケーリングパラメーターの関係．これらには相関関係が見られる．

る性能を発揮することが確認された．CLIP に対して最小限の変更のみを加えた提案手法がより複雑な構造を持つ既存手法の性能を上回ったことは，提案手法の汎用的な性能の高さを示唆している．本手法はデータ数によらず一貫した構造のアダプターを持つとともに高速な学習が可能であるため，計算資源に制約のある実利用の場において特に有用であるといえる．我々は提案手法が，大規模視覚言語モデルを任意の少数データタスクへ適応させる上での優れたベースラインとなることを期待している．残課題として，事前学習タスクと下流タスクの類似度を測定し，その結果に基づいてスケーリングパラメーターを適切に設定する方法を確立する必要があ

る．本研究では検証セットを用いてスケーリングパラメーターを調整したが，実運用においては十分な検証データを確保できない場合も想定されるため，タスク間の相違度に基づく最適なスケーリングパラメーターの推定や自動調整の枠組みが重要である．事前学習タスクと下流タスクの類似度指標として，データセットが持つ画像情報や言語情報，クラス情報に基づく指標を用いることや，事前学習モデルを下流タスクに適用した際の予測確信度に基づく事前知識の適合度を利用することが候補として挙げられる．更なる展望として，他の大規模視覚言語モデルとの組み合わせによる性能向上の可能性を探ることや，異なるモダリティへの応用可能性の検討が挙げられる．また，提案手法の理論的な解析や，学習事例の効果的な選択方法の探求についても取り組みの余地が残されている．

第5章　結論

本章では，本研究で得られた成果をまとめ，研究の意義と限界を整理し，今後の展望について述べることで本論文を総括する．

5.1　本研究のまとめ

本論文では，大規模視覚言語モデルを活用した少数データ画像認識の効果的な適応手法を探求した．従来の深層学習では大量のラベル付きデータが必要であり，データ収集やアノテーションのコストが課題となっていた．特に医療画像解析や希少動物の識別といった専門的なタスクでは，プライバシーや倫理的制約により大規模データセットの構築が困難である．このような背景から，少数のラベル付きデータで高精度な認識を実現する手法が強く求められている．

本研究では，Contrastive Language-Image Pre-training（CLIP）に代表される大規模視覚言語モデルの優れた汎化性能に着目し，これらのモデルを下流タスクに対して少数データで効率的に適応させる手法を提案した．具体的には，事前学習モデルに対する変更を最小限に抑えながら，軽量なアダプター機構や残差学習を通じて多様な下流タスクにおける認識性能の向上を図った．また，多様なデータセットを用いた広範囲な実験により，提案手法の有効性と汎用性を実証した．

5.1.1 Proto-Adapter

第 3 章では，クラスごとのプロトタイプ表現に基づきアダプターを構築する Proto-Adapter を提案した．この手法は確率的勾配降下法を用いた学習が不要な枠組みを持つため，非常に高速な適用が可能でありながら，高性能な少数データ画像認識を実現できることを示した．さらに，深層距離学習に用いられる Additive Angular Margin Penalty を用いてアダプターをファインチューニングすることで，認識性能をさらに改善できることを示した．Proto-Adapter は学習データの数が変動してもアダプターのサイズが不変であるため，使用できるデータ数が変動しうる実用的なシナリオにおいて利用しやすい手法であると言える．またファインチューニングを行う場合，行わない場合どちらにおいても高速な適用が可能であり，計算リソースや使用できるデータの量に制約のある場面において，目的とするタスクへの適応を効率的に行うことができる手法と言える．

5.1.2 Residual-Adapter

第 4 章では，大規模視覚言語モデルの少数データ適応を簡素な枠組みで効率的に行うベースライン手法として，線形識別器の残差学習による適応手法である Residual-Adapter を提案した．この手法では，CLIP の画像エンコーダー上層に線形識別器を追加し，事前学習パラメータを凍結して線形識別器のみを残差学習により下流タスクへ適応させる．この手法により，視覚言語事前学習で獲得された頑健な特徴表現を保持しつつ，新たなドメインの知識を効率的にモデルに取り入れることが可能である．また，下流タスクによって線形識別器の予測ロジットの最適な重みが大きく異なることを確認し，その適切な設定により簡素な枠組みながら優れた認識性能を実現できることを示した．本手法は Proto-Adapter と同様に高速な学習が可能であるとともに，既存手法と比べて非常に簡素な枠組みを持つにも関わらず多様なドメイ

ンに対する汎用的な適応性能に優れることを確認した．

5.2 課題と展望

5.2.1 本研究の限界

本研究で提案した手法には，いくつかの限界が存在する．第一に，事前学習モデルへの依存性の問題がある．提案手法は事前学習モデルの性能に大きく依存しており，事前学習モデルが苦手とするドメインでは性能向上が限定的である場合がある．第二に，極めて少数のデータ設定における性能に関する課題がある．各クラス 1 から 2 サンプルという極めて少ないデータ設定では，提案手法の優位性が十分に発揮されない場合がある．第三に，専門的なドメインへの汎化性能の問題がある．提案手法により改善が見られたものの，さらなる性能改善の余地が存在する．

5.2.2 研究展望

本研究の成果と課題を踏まえ，今後の研究展望として以下の方向性が考えられる．少数データ画像分類のさらなる性能改善のためには，汎化性能に優れる大規模な事前学習モデルの使用は前提となると考えられる．その中で，専門的なドメインへの汎化は依然として困難な問題であり，この解決に向けていくつかのアプローチが考えられる．第一に，より大規模な言語モデルの活用により，事前学習で獲得される特徴表現の質を改善させることが期待される．第二に，下流タスクに適した事前学習モデルの選定や，複数の事前学習モデルの組み合わせによる性能向上が有望である．第三に，学習事例の効果的な選択方法の探求も重要であり，これは能動学習の領域と密接に関連し，少数データ学習との相乗効果が期待される．また，ラベルなしデータの活用による半教師あり学習のアプローチも，アノテーションコストの削

減に大いに役立つと考えられる．次に，学習データ量に関わらず汎用的に利用可能な枠組みの構築も重要な研究方向である．本研究では少数データ設定を主眼としたが，より幅広いデータ量の設定で効果的に機能する手法の開発が求められる．さらに，他のタスクや異なるモダリティへの拡張も重要な展望である．領域分割や物体検出といった他のタスクへの応用や，動画，テキスト，音声，3 次元データなどの異なるモダリティへの応用可能性を検討することで，提案手法の適用範囲を大幅に拡張できると考えられる．最後に，少数データ設定における汎化性能の理論的解析も重要な研究課題である．提案手法がなぜ効果的なのか，どのような条件下で最適な性能を発揮するのかを理論的に明らかにすることで，より確実で予測可能な手法の開発が可能になると期待される．

5.3 おわりに

本論文では，大規模視覚言語モデルを活用した少数データ画像認識の効率的な適応手法について研究した．クラスごとのプロトタイプ表現を用いることによる学習不要な適応手法である Proto-Adapter と，線形識別器の残差学習による効率的な適応手法である Residual-Adapter の 2 つの手法を提案し，いずれも従来手法と比較して優れた認識性能および実用上の優位性を持つことを実証した．これらの成果により，少数データ画像認識の分野において，大規模視覚言語モデルを活用した効率的な適応手法の基盤を構築することができたと言える．深層学習の発展により，大量のデータを用いた高精度な認識システムが実現されている一方で，データ収集やアノテーションのコストは依然として大きな課題である．本研究で提案した手法は，この課題に対する実用的な解決策を提供するものであり，深層学習技術の社会実装をより身近なものにする重要な一歩である．少数データ学習の分野は急速に発展しており，大規模視覚言語モデルなどの新たな技術の登場により，さらなる可能性が広がって

いる．本研究の成果が，今後の研究発展の基盤となり，実社会における様々な課題解決に活用されることを期待している．最後に，限られたデータから最大限の価値を引き出すという少数データ学習の本質は，効率的で持続可能な機械学習技術の社会実装に向けた重要な研究分野であり続けると考えられる．本研究がその発展に少しでも寄与できれば幸いである．

謝辞

本研究は，著者が慶應義塾大学大学院理工学研究科後期博士課程在学中に，同大学理工学部青木義満教授のご指導のもとで行われました．本論文の執筆にあたり，多くの方々よりご指導ならびにご協力を賜りました．ここに深く感謝申し上げます．

はじめに，本論文の主査であり指導教員である青木義満教授に，心より感謝申し上げます．学部生の頃から博士課程に至るまで長年にわたり，多大なるご指導とご支援を賜りました．研究に関するご指導に加え，研究に専念できる環境を整えていただいたことで，充実した研究活動を行うことができました．また，博士課程への進学に際しては，学業と仕事の両立に関し多大なるご配慮とご協力を賜りましたこと，重ねて御礼申し上げます．青木研究室の一員として受け入れていただいたことは，私の人生における大きな転機の一つでした．研究室で学ばせていただいたことは現在の仕事にも繋がっており，今後の人生においても大きな糧になると確信しております．

そして，本研究の副査を快くお引き受けいただきました池原雅章教授，杉本麻樹教授，村田真悟准教授に深く感謝申し上げます．学位審査を通して多くの貴重なご助言とご指摘をいただき，本論文をより充実した内容とすることができました．

また，共同研究先である株式会社明電舎の皆様に深く感謝申し上げます．研究題目の策定から研究の推進に至るまで，数多くの貴重なご助言と多大なるご協力を賜りました．

加えて，青木研究室の皆様に深く感謝いたします．先輩方からのご指導に加え，皆様からのご助言ならびに議論を通して，多くのことを学ばせていただきました．共

に切磋琢磨する仲間がいたことが，困難な研究課題に取り組む上で大きな支えとなりました．素晴らしいメンバーに恵まれた研究室で，充実した研究生活を送ることができたことを大変嬉しく思います．

最後に，家族に心より感謝いたします．両親には修士課程までの間，多大な経済的支援をしていただきました．家族が私の決断を温かく見守り，協力してくれたことが，研究を続けていく上で心強い支えとなりました．

改めて皆様に深甚なる謝意を表します．

参考文献

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1026–1034, 2012.

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[5] JE Small, P Osler, AB Paul, and M Kunst. Ct cervical spine fracture detection using a convolutional neural network. *American Journal of Neuroradiology*, 42(7):1341–1347, 2021.

[6] Yun-Woo Chang, Jung Kyu Ryu, Jin Kyung An, Nami Choi, Young Mi Park, Kyung Hee Ko, and Kyunghwa Han. Artificial intelligence for breast cancer screening in mammography (ai-stream): preliminary analysis of a prospective multicenter cohort study. *Nature Communications*, 16(1):2248, 2025.

[7] Severstal: Steel defect detection. https://www.kaggle.com/competitions/severstal-steel-defect-detection. Accessed: 2026-02-08.

[8] Weidong Zhao, Feng Chen, Hancheng Huang, Dan Li, and Wei Cheng. A new steel defect detection algorithm based on deep learning. *Computational Intelligence and Neuroscience*, 2021(1):5592878, 2021.

[9] Haoyu Chen, Stacy Lindshield, Papa Ibnou Ndiaye, Yaya Hamady Ndiaye, Jill D Pruetz, and Amy R Reibman. Applying few-shot learning for in-the-wild camera-trap species classification. *AI*, 4(3):574–597, 2023.

[10] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3):1–34, 2020.

[11] Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, 55(13s):1–40, 2023.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning*, pages 1126–1135, 2017.

[13] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4080–4090, 2017.

[14] Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *Proceedings of the International Conference on Learning Representations*, 2020.

[15] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[16] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[17] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the International Conference on Machine Learning*, pages 647–655. PMLR, 2014.

[18] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 806–813, 2014.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021.

[20] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[21] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

[22] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

[23] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *Proceedings of the International Conference on Machine Learning*, pages 2790–2799, 2019.

[24] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[25] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

[26] Naoki Kato, Yoshiki Nota, and Yoshimitsu Aoki. Proto-adapter: Efficient training-free clip-adapter for few-shot image classification. *Sensors*, 24(11):3624, 2024.

[27] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *Proceedings of the International Conference on Machine Learning Workshop*, pages 1–30, 2015.

[28] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3637–3645, 2016.

[29] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9068–9077, 2022.

[30] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Proceedings of the European Conference on Computer Vision*, pages 266–282, 2020.

[31] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[32] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[33] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[34] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, pages 702–703, 2020.

[35] Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific reports*, 9(1):16884, 2019.

[36] Brandon Trabucco, Kyle Doherty, Max A Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *Proceedings of the International Conference on Learning Representations*, pages 1–15, 2024.

[37] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1717–1724, 2014.

[38] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, pages 1597–1607, 2020.

[39] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[40] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[41] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-

language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*, pages 4904–4916, 2021.

[42] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pages 1857–1865, 2016.

[43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[44] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[45] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023.

[46] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[47] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.

[48] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

[49] Han Lin, Guangxing Han, Jiawei Ma, Shiyuan Huang, Xudong Lin, and Shih-Fu Chang. Supervised masked knowledge distillation for few-shot transformers. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19649–19659, 2023.

[50] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

[51] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

[52] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[53] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision workshop*, pages 554–561, 2013.

[54] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[55] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004.

[56] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008.

[57] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010.

[58] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[59] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012.

[60] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision*, pages 446–461, 2014.

[61] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[62] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[63] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[64] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.

[65] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[66] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.

[68] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

付 録A　データセット

3.3 節における Proto-Adapter の評価実験および，4.3 節における提案手法の評価実験では 11 種類のデータセットを少数データ画像分類の評価実験に使用した．それぞれのデータセットの統計量を表 A.1 に，画像例を図 A.1 に示す.

ImageNet [1] は 1,000 クラス，約 130 万枚の訓練画像からなる大規模な画像分類データセットである．日常的な物体，動物，乗り物など多様なカテゴリを含み，画像分類タスクの評価用途に広く利用される．Caltech101 [55] は，101 クラスのオブジェクトカテゴリと 1 つの背景クラスからなる物体認識データセットである．動物，乗り物，日用品など多様な物体が含まれる．OxfordPets [59] は，37 種類の犬と猫の品種を含むペット分類データセットである．各品種について約 200 枚の画像が含まれる．StanfordCars [53] は，196 クラスの車種を含む細粒度分類データセットである．メーカー，モデル，製造年などにより詳細に分類されている．Flowers102 [56] は，イギリスで一般的な 102 種類の花の細粒度分類データセットである．各クラスは 40 から 258 枚の画像で構成される．Food101 [60] は，101 種類の食べ物カテゴリを含むデータセットである．各クラスには 1,000 枚の画像が含まれ，実世界のノイズを多く含む．FGVCAircraft [24] は，100 種類の航空機モデルを含む細粒度分類データセットである．Boeing 737 や Airbus A380 などの機種が含まれる．SUN397 [57] は，397 種類のシーンカテゴリを含む大規模なシーン認識データセットである．屋内外の多様な環境が含まれる．DTD（Describable Textures Dataset） [25] は，47 種類のテクスチャカテゴリを含むデータセットである．各カテゴリは質感を表す形容詞（例：striped，dotted）で記述される．EuroSAT [58] は，10 クラスの土地利用・土

地被覆分類のための衛星画像データセットである．住宅地，森林，農地などが含まれる．UCF101 [54] は，101 種類の人間の行動を含むデータセットである．スポーツや日常動作などのカテゴリが含まれる．

表 A.1: 各データセットのクラス数およびデータ数.

Dataset	**Classes**	**Train Size**	**Test Size**
ImageNet	1,000	1,281,167	50,000
Caltech101	102	3,060	6,085
OxfordPets	37	3,680	3,669
StanfordCars	196	8,144	8,041
Flowers102	102	2,040	6,149
Food101	101	75,750	25,250
FGVCAircraft	100	6,667	3,333
SUN397	397	19,850	19,850
DTD	47	3,760	1,880
EuroSAT	10	10,000	5,000
UCF101	101	9,537	1,794

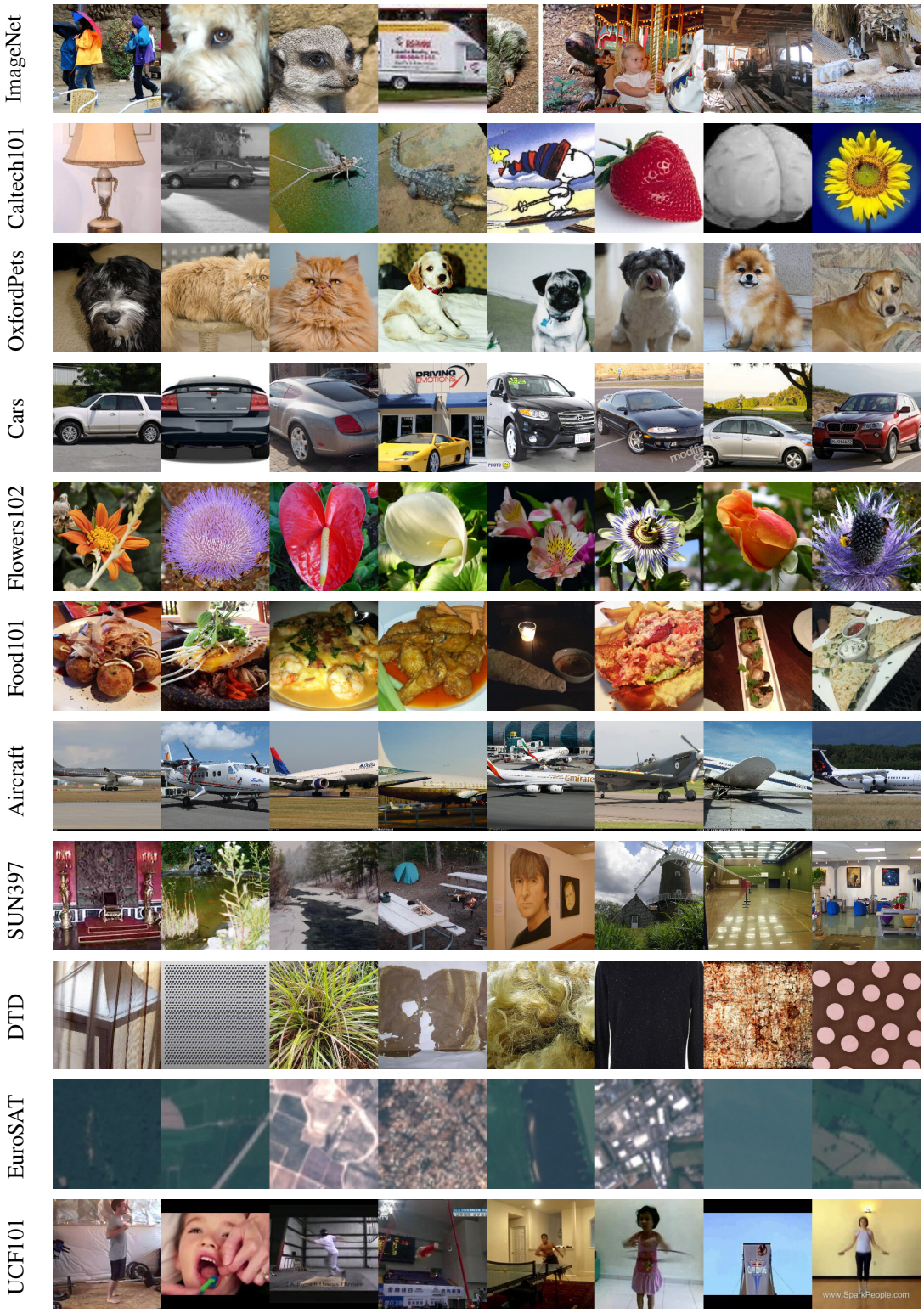

図 A.1: 各データセットの画像例.

付 録B　プロンプトテンプレート

3.3 節における Proto-Adapter の評価実験および，4.3 節における提案手法の評価実験において，各データセットに対して使用したプロンプトテンプレートの一覧を表 B.1 に示す．ImageNet [1] では 7 つのテンプレートを用いたプロンプトアンサンブルを，その他データセットでは単一のプロンプトテンプレートを使用した．

表 B.1: 各データセットに対して使用したプロンプトテンプレート．

Dataset	**Prompt Templates**
ImageNet	”itap of a [CLASS].”
	”a bad photo of the [CLASS].”
	”a origami [CLASS].”
	”a photo of the large [CLASS].”
	”a [CLASS] in a video game.”
	”art of the [CLASS].”
	”a photo of the small [CLASS].”
Caltech101	”a photo of a [CLASS].”
OxfordPets	”a photo of a [CLASS], a type of pet.”
StanfordCars	”a photo of a [CLASS].”
Flowers102	”a photo of a [CLASS], a type of flower.”
Food101	”a photo of [CLASS], a type of food.”
FGVCAircraft	”a photo of a [CLASS], a type of aircraft.”
SUN397	”a photo of a [CLASS].”
DTD	”[CLASS] texture.”
EuroSAT	”a centered satellite photo of [CLASS].”
UCF101	”a photo of a person doing [CLASS].”