JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 8, pp. 17-30, August 2026
https://doi.org/10.9708/jksci.2026.31.08.017

Rethinking PEFT Evaluation: A Class-Balanced Protocol Across Data Scales for Fair Comparison

Du-Hwan Hur*, Dong-Hyun Kim**, Seung-Hwan Bae***

*Ph.D. candidate, Vision & Learning Lab, Inha University, Incheon, Korea
**M. S. candidate, Vision & Learning Lab, Inha University, Incheon, Korea
***Associate Professor, Vision & Learning Lab, Dept. of Electrical and Computer Engineering, Inha University, Incheon, Korea

[Abstract]

In this paper, we propose a class-balanced evaluation protocol across data scales based on the VTAB-1k benchmark to analyze Parameter-Efficient Fine-Tuning (PEFT) under the extreme data scarcity and class imbalance of real-world domains such as medical and satellite imagery. The pipeline independently controls data scale (10%–100%) and class balance, benchmarking Adapter, LoRA, and Visual Prompt Tuning (VPT) on 19 tasks over a frozen ViT backbone, together with full fine-tuning and linear-probing baselines and a recent variant, DoRA. A clear performance crossover is observed: VPT, strong in specialized and structured domains, dominates below the 50% scale, whereas LoRA, strong in natural domains, prevails as data increases. Class balancing trades a small amount of overall Top-1 accuracy for higher balanced accuracy in low-data regimes, and its effect vanishes at larger scales as the balanced subset converges to the original distribution. These controlled results demonstrate that no single PEFT method is universally optimal, and the resulting guidelines support the practical deployment of PEFT in data-limited applications such as medical imaging, remote sensing, and industrial inspection.

▸ **Key words**: Parameter-Efficient Fine-Tuning, Low-Data Learning, Visual Prompt Tuning, Low-Rank Adaptation, Adapter

• First Author: Du-Hwan Hur, Corresponding Author: Seung-Hwan Bae
*Du-Hwan Hur (gjenghks@inha.edu), Vision & Learning Lab, Inha University
**Dong-Hyun Kim (ddongpal0730@inha.edu), Vision & Learning Lab, Inha University
***Seung-Hwan Bae (shbae@inha.ac.kr), Vision & Learning Lab, Dept. of Electrical and Computer Engineering, Inha University
• Received: 2026. 06. 22, Revised: 2026. 07. 19, Accepted: 2026. 07. 21.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

[요 약]

본 논문은 데이터 양(10%~100%)과 클래스 균형 여부를 독립적으로 통제하는 데이터 규모 별 클래스 균형 평가 프로토콜을 VTAB-1k 벤치마크에서 제안한다. 최근 대규모 사전학습 모델을 적은 비용으로 특정 데이터셋에 적응시키는 파라미터 효율적 미세조정(PEFT)이 주목받고 있으나, 의료·위성 영상과 같이 데이터가 극히 제한되고 클래스가 불균형한 현실적 환경에서의 학습 및 예측양상은 충분히 규명되지 않았다. 제안하는 프로토콜은 동결된 ViT 백본 위에서 주요 PEFT 기법(Adapter, LoRA, VPT)을 자연·특수·구조적 도메인의 19개 데이터셋에 대해 동일 조건에서 벤치마크하고, 완전 미세조정 및 linear probing baseline과 최신 변형 기법 DoRA를 함께 비교한다. 실험 결과, 데이터가 50%보다 적을 때는 특수·구조적 도메인에 강한 VPT가, 그보다 많을 때는 자연 도메인에 강한 LoRA가 우세하여 데이터 양에 따라 유리한 기법이 달라졌다. 또한 클래스 수를 맞춘 균형 학습은 데이터가 적을 때 전체 Top-1 정확도를 조금 낮추는 대신 모든 클래스를 고르게 평가하는 balanced accuracy를 높였고, 데이터가 많아지면 균형 데이터가 원본과 거의 같아져 그 효과가 사라졌다. 이처럼 모든 상황에서 가장 좋은 단일 PEFT 기법은 존재하지 않으며, 본 연구를 통해 도메인 특성·데이터 규모·클래스 분포에 따른 실용적 기법 선택 지침을 제시하여 의료·원격탐사·산업 검사 등 데이터가 제한된 다양한 응용 분야에서 PEFT의 효율적 적용에 기여할 수 있다.

▸ **주제어**: 파라미터 효율적 미세조정, 소규모 데이터 학습, 비주얼 프롬프트 튜닝, 저순위 적응, 어댑터

I. Introduction

최근 컴퓨터 비전 분야는 큰 패러다임 전환을 맞이하고 있다[1]. 기존 파운데이션 모델인 Vision Transformer (ViT)[2]를 넘어, 수억에서 수십억 개에 이르는 파라미터를 갖는 대규모 기반 모델(foundation model)이 등장하면서 다양한 데이터셋에서 뛰어난 성능을 달성하고 있다[3,4,5]. 그러나 파라미터 규모가 수십억 단위로 증가함에 따라 모델 전체를 미세조정하는 완전 미세조정(full fine-tuning)은 막대한 연산 자원을 요구하게 된다. 또한 클래스가 불균형한 long-tail 환경에서는 완전 미세조정이 사전학습으로 확보한 특징 표현을 왜곡하여 오히려 소수(tail) 클래스의 성능을 저하시킬 수 있음이 보고되었다[6]. 이를 해결하기 위해 파라미터 효율적 미세조정(Parameter-Efficient Fine-Tuning, PEFT) 기법이 제안되었다[7,8]. 대표적으로 어댑터(Adapter)[9]는 트랜스포머 블록 사이에 병목 계층을 삽입하고, 저순위 적응(Low-Rank Adaptation, LoRA)[10]은 가중치 갱신에 저순위 행렬을 주입하며, 비주얼 프롬프트 튜닝(Visual Prompt Tuning, VPT)[11]은 입력 공간에 학습 가능한 프롬프트 토큰을 추가한다. 이들은 소수의 파라미터만 갱신하여 전이학습 효율을 높인다.

대규모 사전학습 모델이 소량의 데이터만으로 새로운 태스크에 적응할 수 있음이 알려지면서[12] PEFT도 소규모 데이터 환경에서 효과적일 것으로 기대되어 왔으나, 그

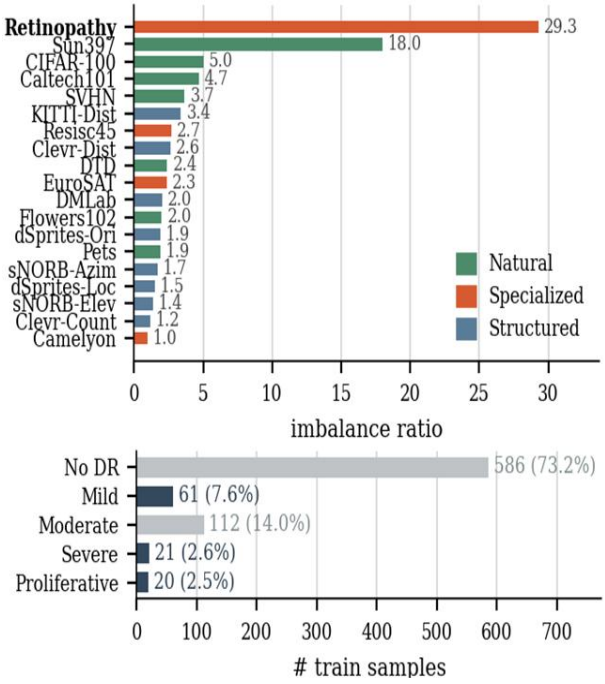

Fig. 1. Class imbalance in the VTAB-1k benchmark. (Top) Class-imbalance ratio across all 19 tasks, by domain, (Bottom) Retinopathy per-class distribution

성능은 충분히 검증되지 않았다. 의료 영상이나 특수 센서 기반 영상과 같이 실제 응용에서는 엄격한 데이터 제약과 상당한 도메인 격차가 존재하며, 실제 환경에서 기존 PEFT의 보고된 장점이 실제로 유지되는지는 불분명하다. 예를 들어 VTAB-1k[13]에 포함된 당뇨망막병증(Diabetic Retinopathy) 데이터셋은 Fig. 1과 같이 정상(No DR) 클

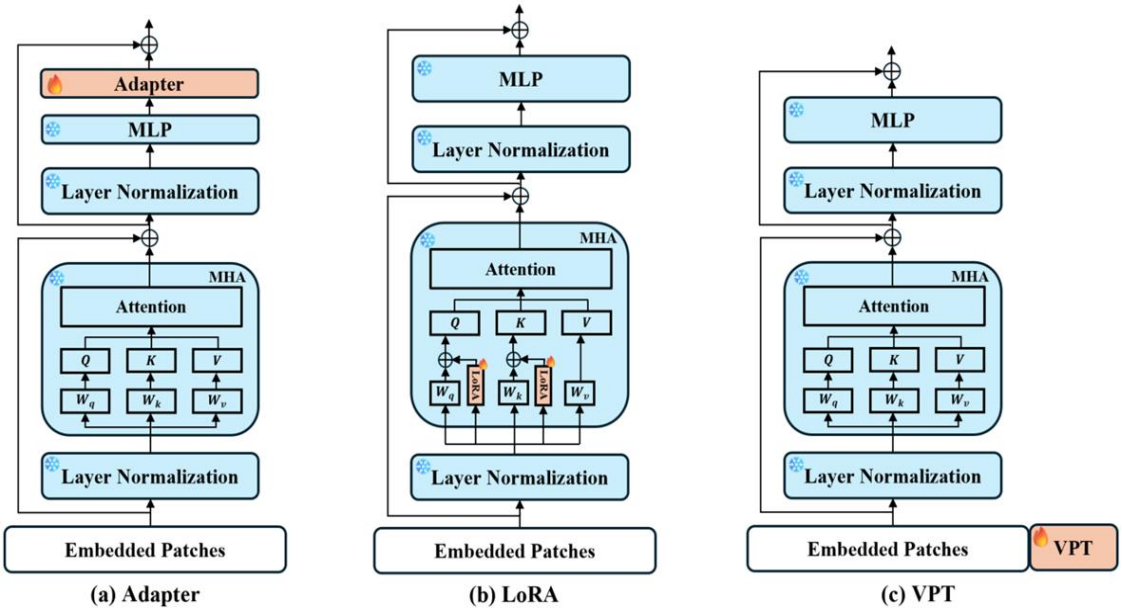

Fig. 2. Architectural comparison of PEFT methods: (a) Adapter with bottleneck modules, (b) LoRA with low-rank updates, (c) VPT with trainable prompt tokens.

래스가 전체의 73.2%를 차지하는 극심한 클래스 불균형을 나타낸다. 이러한 불균형은 특정 데이터셋에 국한된 예외가 아니다. Fig. 1과 같이 VTAB-1k의 19개 데이터셋 대다수에서 가장 많은 클래스와 가장 적은 클래스의 표본 수가 큰 차이를 보이며, 이는 자연·특수 도메인에서 특히 두드러진다. 따라서 이 문제는 개별 데이터셋 차원이 아니라 VTAB-1k 전반의 차원에서 다루어져야 하며, 클래스 분포를 통제하지 않은 평가로는 이를 해소할 수 없다.

이에 본 논문은 PEFT 기법 별로 클래스 균형에 따른 영향력과 데이터 비율 별 비교 및 분석을 수행한다. 각 데이터셋에 대해 클래스별로 동일한 수의 표본을 추출하여 원본 크기의 10%에서 100%까지 학습 집합을 재구성함으로써, 클래스 불균형 없이 다양한 데이터 가용 조건에서 PEFT 기법을 공정하게 비교한다. 이 설정 하에서 Adapter, LoRA, VPT를 자연·특수·구조적 도메인에 걸친 19개 비전 데이터셋에서 벤치마크하였다. 분석 결과, 데이터가 많아질수록 Adapter와 LoRA의 성능이 향상되는 반면, VPT는 위성 영상이나 의료 진단과 같은 특수·구조적 데이터셋에서 극단적으로 데이터가 적은 환경에 더 강건하였다.

본 논문의 주요 기여는 다음과 같다. (1) 공정한 PEFT 평가를 위한 데이터 규모 별 클래스 균형 VTAB-1k 평가 프로토콜을 제안한다. (2) 3개의 도메인, 19개의 데이터셋에서 Adapter, LoRA, VPT를 벤치마크하고, 최신 변형 기법 DoRA[14]까지 비교를 확장한다. (3) 데이터가 적은 구간의 VPT와 데이터가 많은 구간의 LoRA가 교차하는 성능 전환점을 규명하고, 도메인 유형과 데이터 규모에 따라 어떤 기법이 우수한지를 종합 분석한다. (4) 클래스 균형화의 이점이 데이터가 부족한 구간에 집중되고 데이터가 늘수록 사라짐을 정량적으로 확인하고, 이를 바탕으로 도메인 특성·데이터 규모·클래스 분포에 따른 PEFT 기법 선택 지침을 제시한다.

II. Related Works

PEFT는 연산 자원이나 레이블 데이터가 제한된 환경에서 완전 미세조정의 실용적 대안으로 활발히 연구되어 왔다[7,8,14]. 이 기법들은 전체 파라미터 중 일부만을 수정하여 대규모 사전학습 모델을 하위 데이터셋에 적응시킴으로써 학습 비용과 메모리 사용량을 크게 줄인다. 선행 연구[15]는 비전 모델을 위한 PEFT 기법을 다음 세 범주로 분류한다.

- **추가 기반 튜닝(Addition-based Tuning):** 사전학습 백본 가중치를 변경하지 않고 병목 계층이나 프롬프트 토큰 형태의 학습 가능한 모듈을 도입하는 방식으로, Adapter[9]와 VPT[11]가 대표적이다.
- **부분 기반 튜닝(Partial-based Tuning):** 내부 파라미터의 일부만 선택적으로 갱신하는 방식으로, 특정 파라미터를 직접 수정하는 BitFit[17]이나 학습 가능한 저순위 행렬을 추가·통합하는 LoRA[10]가 이에 해당한다.
- **통합 기반 튜닝(Unified-based Tuning):** 여러 튜닝 전략을 결합하고 구조 탐색으로 구성을 최적화하는 혼합

프레임워크로, NOAH[18] 등이 있다.

한편 기존 VTAB-1k 기반 평가는 세 가지 공통된 한계를 지닌다. 첫째, VPT[11], NOAH[18] 및 대규모 벤치마크 연구[15]를 포함한 대부분의 선행 평가는 각 데이터셋의 학습 표본 전체를 사용하는 단일 데이터 지점에서 단일 성능만을 보고하므로, 최적의 기법이 데이터 규모에 따라 어떻게 달라지는지 관측할 수 없다. 둘째, VTAB-1k에 내재한 클래스 불균형 문제를 고려하지 않아, 보고된 성능 차이가 기법의 우열 때문인지, 데이터 클래스 분포가 치우친 탓인지 구분할 수 없다. 이는 클래스 불균형이 다수 클래스로의 편향을 유발해 소수 클래스 성능을 저하시킨다는 점이 널리 보고된[19] 것에 비추어 볼 때, 공정 비교의 실질적 장애가 된다. 셋째, 그 결과 어떤 PEFT 기법이 우수한가라는 질문에 데이터 규모나 클래스 분포와 무관하게 정답이 하나로 정해져 있다고 암묵적으로 가정하게 된다. 이러한 한계는 데이터 중심(data-centric) 관점의 최근 흐름과도 맞닿아 있다. 도메인 격차, 레이블 노이즈, 롱테일 분포 등 데이터 품질 요인이 전이 성능을 좌우한다는 논의가 활발해지고 있으며, 평가 벤치마크가 이러한 요인을 통제하지 않으면 기법 간 비교가 왜곡될 수 있다는 인식이 확산되고 있다. 본 연구는 이들 요인 가운데 명시적 통제가 가능하고 그 영향이 가장 직접적인 두 축, 즉 데이터 규모와 클래스 분포에 초점을 맞춘다. 도메인 격차는 VTAB-1k의 세 도메인 구분을 분석 축으로 활용해 간접적으로 다루며, 레이블 노이즈 등 나머지 품질 축으로의 확장은 본 프로토콜의 향후 과제 중 하나다.

본 연구의 프로토콜은 이 두 요인(데이터 규모, 클래스 분포)을 독립적으로 통제함으로써, 기존 설정에서는 구조적으로 관측 불가능했던 두 현상을 정량적으로 드러낸다. 첫째, 클래스 균형 설정에서 기존 평가에 해당하는 100% 지점에서는 LoRA가 최고 성능(71.60%)이지만, 10% 지점에서는 VPT(41.63%)로 뒤바뀌어, 데이터 규모에 따라 우세 기법 자체가 역전된다. 이 역전은 특히 사전학습과 이질적인 구조적 도메인에서 두드러진다. 즉 단일 지점 평가는 데이터가 적은구간의 기법 특성을 원리적으로 포착하지 못한다. 둘째, 클래스 균형화 설정의 효과는 10% 지점에서 최대 +5.2%p(Adapter)에 달하지만 100% 지점에서는 +0.05%p로 줄어들어, 이러한 이점이 데이터가 부족한 구간에 집중되고 데이터가 늘수록 약해짐을 보여준다. 이 두 정량적 발견이 본 프로토콜이 기존 평가 방식과 구별되는 학술적 근거이며, 본 연구는 이를 통해 조건(도메인·규모·분포)에 따른 기법 선택 지침이라는, 단일 지점 평가로는 도출할 수 없는 결론을 제공한다.

각 범주 내에서도 다양한 변형이 제안되어 왔다. Adapter 계열에서는 MLP 계층에 병렬 병목 구조를 도입하여 확장성과 효율을 개선한 AdaptFormer[20] 등이 제안되었고, LoRA 계열에서는 가중치 분해 방식을 정교화한 DoRA 등이, VPT 계열에서는 프롬프트의 표현력과 효율을 높인 E²VPT[21]과 같은 후속 연구가 이어졌다. 이처럼 PEFT 연구는 빠르게 확장되고 있으나 대부분 충분한 데이터를 가정한 표준 설정에서 평가되어, 극단적인 소규모 데이터 조건에서의 학습 및 예측양상은 여전히 불분명하다. 본 연구는 이러한 변형들의 기반이 되는 세 가지 핵심 기법을 통제된 조건에서 분석함으로써, 향후 다양한 변형 기법으로 확장 가능한 진단 기준을 제공한다.

본 연구에서는 위 첫 두 범주에 해당하면서 구조적 통합 지점이 서로 다른 Adapter, LoRA, VPT에 집중한다. 세 기법은 각각 중간 계층, 어텐션 가중치, 입력 토큰을 수정하여 통합 지점이 명확히 구분되고, 모두 오픈소스로 널리 채택되어 구조 재설계 없이 적용 가능하며, 이러한 구조적 차이 덕분에 데이터 규모·도메인에 따른 트레이드오프를 의미 있게 비교할 수 있다. 반면 NOAH[18]와 같은 통합 기법이나 BitFit 같은 경량 대안은 혼합적 특성이나 상이한 가정으로 인해 공정한 독립 비교가 어려워 본 연구 범위에서 제외하였다. 비교 대상인 세 기법의 구조는 Fig. 2와 같으며, 각 기법의 동작 원리를 이어서 상세히 기술한다.

1. Adapter

Adapter[9]는 작은 모듈을 삽입하여 모델을 새로운 데이터셋에 적응시키는 기법으로, 사전학습 가중치는 동결(frozen)된 채 새로 추가된 모듈의 파라미터만 학습된다. Fig. 2-(a)와 같이 모듈은 트랜스포머 블록 내 MLP 하위 계층 직후에 삽입된다. 은닉 상태 $h \in \mathbb{R}^d$가 주어지면, 하향 투영 행렬 $W_{down} \in \mathbb{R}^{d \times m}$로 차원을 압축하고, 비선형 활성화 함수(예: GELU[16])를 적용한 뒤, 상향 투영 행렬 $W_{up} \in \mathbb{R}^{m \times d}$로 원래 차원을 복원하여 잔차 연결로 더한다. 이 과정은 다음과 같다.

$$Adapter(h) = (f(h W_{down} + b_{down})) W_{up} + b_{up} \quad (1)$$

$$h' = h + Adapter(h) \quad (2)$$

여기서 $b_{down} \in \mathbb{R}^m, b_{up} \in \mathbb{R}^d$는 각 투영 계층의 편향 벡터이다.

2. Low-Rank Adaptation (LoRA)

LoRA[10]는 가중치 갱신량 ΔW가 매우 낮은 내재적 순위(intrinsic rank)를 갖는다는 가설에 기반하여, ΔW를 두 저순위 행렬 $B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}$의 곱으로 분해한다.

$$\Delta W = BA \tag{3}$$

여기서 $r \ll min(d,k)$이다. 전이학습 시 원본 가중치 W_0는 동결되고 A, B만 학습되며, 입력 x에 대한 출력은 다음과 같다.

$$h' = xW_0 + x\Delta W = xW_0 + xBA \tag{4}$$

원 논문[10]은 Query·Value 행렬에 적용하였지만, 본 연구는 VTAB-1k 상의 대표 비교 연구인 NOAH[18]의 설정을 따라 Fig. 2-(b)와 같이 Query·Key 가중치 행렬에 적용한다. 추론 시 두 행렬을 단일 행렬로 병합($W = W_0 + BA$)할 수 있어 추가적인 추론 지연이 없다는 장점이 있다.

3. Visual Prompt Tuning (VPT)

VPT[11]는 입력 수준에 소수의 학습 가능한 프롬프트를 추가하여 모델을 적응시킨다. 사전학습 백본 전체를 동결한 채 입력 시퀀스만 수정한다는 점에서 가장 극단적인 PEFT 중 하나이다(Fig. 2-(c)). p개의 프롬프트 토큰 $P \in \mathbb{R}^{p \times d}$를 원본 패치 임베딩 $E_{patches} \in \mathbb{R}^{N \times d}$ 앞에 추가하여, $[CLS]$를 포함한 최종 입력 시퀀스 E'을 구성한다.

$$E' = [[CLS]; P; E_{patches}] \tag{5}$$

ViT 백본은 동결되고 프롬프트 P만 갱신되며, 모델은 학습된 프롬프트를 새로운 데이터셋 수행을 위한 추가 맥락으로 활용하도록 학습된다.

III. Methodology

1. Low-data Evaluation Protocol

기존 PEFT 평가는 대부분 모든 학습 표본을 사용하는 단일 데이터 설정에서 단일 성능만을 보고하고, 데이터셋 고유의 클래스 불균형을 통제하지 않은 채 어떤 기법이 우수한지를 결정해 왔다. 이러한 평가는 어떤 기법이 우수한

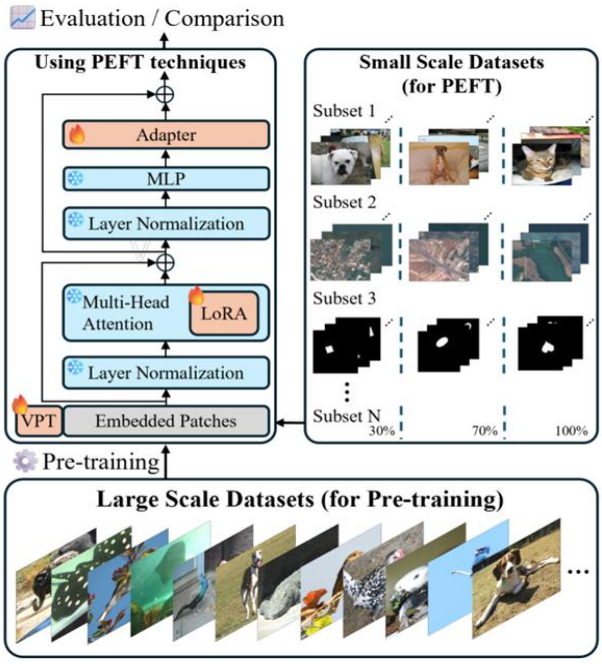

Fig. 3. Overview of the low-data evaluation protocol.

지가 데이터 규모나 분포와 무관하게 하나로 정해져 있다고 암묵적으로 가정하고 있다. 그러나 의료 영상이나 위성 영상과 같이 데이터가 극히 제한되고 분포가 편향된 현실적 환경에서는 이 가정이 성립하지 않으며, 데이터 규모와 클래스 분포라는 두 요인이 뒤섞여 공정한 비교를 어렵게 한다. 따라서 PEFT 기법을 현실적 제약 하에서 올바르게 평가하려면, 이 두 요인을 독립된 축으로 분리하여 통제하는 평가 설계가 필요하다.

이를 위해 본 논문은 Fig. 3과 같은 데이터 규모 별 클래스 균형 평가 프로토콜을 제안한다. 이 프로토콜은 (1) 대규모 데이터셋에서 사전학습된 ViT-B/16[2]을 동결된 백본으로 사용하고, (2) VTAB-1k[13]를 기반으로 클래스별 균일 추출을 통해 10%~100% 규모의 데이터 규모 별 클래스 균형 부분집합을 구성하며, (3) Adapter, LoRA, VPT를 동결 백본에 독립적으로 적용해 지정 파라미터만 갱신하고, (4) 모든 데이터 규모·도메인에서 Top-1 정확도를 측정하는 네 단계로 구성된다. 나아가 완전한 비교를 위해 원본(불균형)과 클래스 균형 설정 모두에서 실험을 수행한다. 또한 세 대표 기법에 더해 완전 미세조정과 linear probing을 성능의 상·하한 기준선으로, LoRA의 최신 변형인 DoRA를 변형 기법의 사례로 동일 조건에서 함께 평가하며, 클래스 불균형의 영향을 다각도로 확인하기 위해 대표 데이터셋에 대해 Top-1 정확도뿐만 아니라 balanced accuracy와 macro-F1을 평가 지표로 함께 측정한다. 이처럼 데이터 규모와 클래스 균형을 독립적으

로 통제하는 본 프로토콜은, 단일 설정 평가가 포착할 수 없었던 기법 간 성능 교차(crossover)와 클래스 균형화 효과의 데이터 의존성을 드러내며, 바로 이 점이 본 평가 설계가 필요하고 기존 분석과 차별되는 이유이다.

2. Dataset Construction

VTAB-1k[13]는 자연·특수·구조적 세 도메인의 19개 데이터셋으로 구성되며, 세 도메인은 사전학습 표현과의 관계에 따라 다음과 같이 구분된다. 자연(Natural) 도메인은 일반 카메라로 촬영한 일상 사물·장면 영상으로, 대상 개념이 ImageNet 계열 사전학습 데이터와 시각적으로 중첩되어 사전학습 표현을 그대로 활용하기에 유리하다(Caltech101, CIFAR-100, DTD, Flowers102, Pets, SVHN, Sun397). 특수(Specialized) 도메인은 위성 센서나 의료 장비 등 비일상적 장비로 취득한 영상으로, 영상의 저수준 통계와 촬영 조건이 자연 영상과 크게 달라 도메인 격차가 존재한다(EuroSAT, Resisc45, Patch Camelyon, Retinopathy). 구조적(Structured) 도메인은 합성 장면에서 물체의 개수·거리·자세 등 기하학적 속성을 예측하는 데이터셋으로, 대상의 의미적 범주가 아니라 장면의 구성 규칙을 판별해야 하므로 사전학습이 제공하는 의미 중심 표현만으로는 해결이 어렵다(Clevr-Count, Clevr-Dist, DMLab, KITTI-Dist, dSprites-Loc, dSprites-Ori, sNORB-Azim, sNORB-Elev). 본 논문은 이 세 도메인이 사전학습 표현과 맺는 관계가 서로 다르다는 점에 착안하여, 도메인을 PEFT 기법 선택의 명시적 분석 축으로 삼는다. 각 데이터셋은 본래 1,000개 표본(학습 800, 검증 200)을 포함한다. 다수의 데이터셋이 클래스 불균형을 보이는데, 이는 각 표본의 영향력이 큰 데이터가 적은 환경에서 공정한 비교를 저해한다. 이를 완화하기 위해 클래스별로 동일한 수의 표본을 균일 추출하여 클래스 균형 부분집합을 구성하고, 전체 표본 수를 조절해 10%~100%의 규모로 변화시키되 각 규모에서 클래스 분포의 균등성을 최대한 유지하였다. 본 논문에서 데이터 규모별 적응이란 클래스당 표본 수를 고정하는 전통적 k-shot 설정이 아니라, VTAB-1k의 학습 표본(800개)을 10%~100%로 부분 추출하여 데이터 양을 통제하는 적응 설정을 지칭한다. 예컨대 10% 설정의 학습 표본은 총 80개로, 데이터셋의 클래스 수에 따라 클래스당 표본이 수십 개에서 수 개 이하 수준까지 감소하여 기존 소규모 데이터 연구의 표본 규모와 유사하거나 그보다 극단적인 조건에 해당한다.

클래스 균형 부분집합은 다음 절차로 구성한다. 각 데이터셋에서 클래스별 가용 표본 수를 집계한 뒤, 목표 데이터 비율(10%~100%)에 해당하는 전체 표본 수를 클래스에 균등하게 배분하되, 가용 표본이 소진된 클래스는 건너뛰고 남은 클래스에서 계속 추출하여 목표 표본 수를 채운다. 이때 소수 클래스는 가용 표본 전체가 선택되고 그 부족분은 다수 클래스에서 보충되므로, 균형화의 정도는 데이터 비율이 낮을수록 완전하고 비율이 커질수록 원본 분포에 점진적으로 수렴하며, 100%에서는 원본 집합과 동일해진다. 이 구조적 성질은 이후 관측되는 균형화 효과의 데이터 규모 의존성을 해석하는 데 핵심적이다. 아울러 동일한 비율에서 원본(불균형) 부분집합도 함께 구성한다. 원본 부분집합은 데이터셋 고유의 클래스 분포를 유지한 채 목표 비율만큼 무작위 추출한 것으로, 균형 부분집합과 데이터 비율은 같되 클래스 분포만 다르다. 이 원본·균형 쌍(pair) 구성은 단순한 통제 장치가 아니라, 클래스 균형화가 효과적인가라는 미해결 질문을 정면으로 다루기 위한 설계다. 기존 연구는 균형화를 당연한 전처리 관행으로 가정하거나 아예 고려하지 않았으나, 본 이원 구성은 균형화의 효과를 데이터 규모별로 분리 측정할 수 있게 하여, 데이터 부족과 클래스 분포라는 두 요인이 각 PEFT 기법에 미치는 영향을 독립적으로 관찰하는 데 핵심적인 역할을 한다.

IV. Experiments

1. Implementation Details

모든 실험은 단일 NVIDIA A6000 GPU에서 PyTorch 환경에서 수행한다. 배치 크기는 표본 수가 배치보다 작아질 수 있는 극단적으로 데이터가 적은 환경에서 학습 불안정을 방지하기 위해, 통상적인 64 대신 8로 설정하였다. 선형 스케일링 규칙(Linear Scaling Rule)에 따라 배치를 8분의 1로 줄였으므로 학습률도 통상값을 8로 나누어 동일하게 축소하였다. 모든 실험 조건은 공정하고 재현 가능한 평가를 위해 일관되게 통제하였으며, 주 평가 지표로 Top-1 정확도를 사용한다. 다만 Top-1 정확도는 표본 수가 많은 다수 클래스의 영향을 크게 받으므로, 클래스 불균형을 다루는 본 연구에서는 추가적인 평가 지표를 함께 사용한다. balanced accuracy(BAcc)는 클래스별 재현율(recall)을 클래스에 대해 단순 평균한 값으로, 모든 클래

Table 1. Hyperparameter settings for all experiments.

Type	Setting	Value
Common	Backbone	ViT-B/16 (ImageNet-21k)
	Optimizer / LR schedule	AdamW / Cosine decay
	Learning rate	1.25×10^{-4}
	Weight decay	1×10^{-4}
	Batch size / Epochs	8 / 100
Adapter	Bottleneck dim (m) / Activation function	8 / GELU
LoRA	Rank (r)	8
VPT	Prompt length (p)	100

스를 동일한 비중으로 취급하여 소수 클래스의 검출 능력을 반영한다. macro-F1은 클래스별 정밀도(precision)와 재현율의 조화평균인 F1 점수를 클래스에 대해 단순 평균한 값으로, 소수 클래스를 놓치지 않는 능력뿐 아니라 다른 클래스를 소수 클래스로 잘못 예측하는 오탐까지 함께 반영한다. 따라서 세 지표를 병기하면 다수 클래스 편향(Top-1 대비 BAcc), 그리고 재현율과 정밀도의 균형(BAcc 대비macro-F1)을 각각 분리하여 관찰할 수 있다. 또한 10%부터 100%까지 10개 데이터 비율 전반에서 동일한 경향이 일관되게 관찰되어, 본 연구가 제시하는 기법 간 성능 역전과 균형화 효과가 특정 설정에 국한되지 않는 강건한 결론임을 뒷받침한다.

각 기법의 삽입 위치와 세부 하이퍼파라미터는 NOAH의 표준 구성을 따랐으며, Table 1에서 확인할 수 있다. 구체적으로 Adapter는 각 트랜스포머 블록의 MLP 이후에 삽입하고(m=8), LoRA는 어텐션의 Query·Key 투영 행렬 W_q, W_k에 rank r=8로 적용하며, VPT는 모든 층에 프롬프트를 삽입하는 Deep 방식으로 프롬프트 길이 p=100을 사용한다(파라미터 정합 실험에서는 p=25). DoRA는 LoRA와 동일한 rank와 적용 행렬을 사용하여 가중치 분해 방식의 차이만이 비교되도록 하였다. 학습률은 통상값 0.001을 선형 스케일링 규칙에 따라 8로 나눈 1.25×10^{-4}로 설정하였고, 모든 기법에 동일하게 적용한다. 모든 부분집합 추출과 학습에는 고정 랜덤 시드를 사용하였고, 각 데이터 비율의 부분집합은 고정된 단일 시드로 추출하여 세 기법에 공통 적용함으로써 기법 간 비교의 공정성을 확보하였다. 통계 검증에서는 시드를 3개로 바꾸어 실험을 반복하였으며, 각 시드는 부분집합 추출과 학습을 함께 재수행하고, 그 결과를 평균과 표준편차로 제시한다.

2. Overall Performance Analysis

먼저 데이터 규모와 클래스 분포가 전체 성능에 미치는 영향을 분석한다. 세 기법의 전체 평균 Top-1 정확도를 Table 2에 나타내었으며, 이로부터 PEFT 성능이 데이터 규모와 클래스 분포에 따라 복합적인 양상을 보임을 확인할 수 있다. 아울러 각 데이터 비율에서 최저 성능 기법 대비 상대적 향상률(relative improvement)을 표기하여 기법 간 격차의 상대적 크기를 함께 제시한다. 데이터가 극히 제한된 10~40%인 구간에서는 VPT가 가장 우수하였다. 균형 설정에서 VPT는 10%의 데이터만으로 41.63%를 기록하여 최고 성능을 보였고, LoRA(40.85%)와 Adapter(40.44%)가 근소한 차이로 뒤를 이었다. 이는 내부 가중치를 직접 수정하지 않고 입력 수준의 프롬프트만으로 사전학습 표현을 활용하는 VPT 고유의 방식에서 비롯된다. 주목할 점은 클래스 균형 설정에서 이 격차가 크게 줄어든다는 사실이다. 이는 균형화의 효과가 데이터가 적은 구간에서 LoRA, Adapter의 초기 학습 안정성에 선택적으로 작용함을 보여준다. 다만 현재 결과는 19개 데이터셋 평균 Top-1 정확도에 기반한 것이다. 클래스 불균형이 극심한 개별 데이터셋에서는 균형화가 Top-1 정확도와 balanced accuracy을 교환하는 상이한 양상이 나타나며, 이는 IV.4 절에서 추가적인 평가 지표를 통해 분석한다. 그러나 데이터 비율이 40%를 넘으면서 기법 간 성

Table 2. Average Top-1 accuracy (%) for the entire dataset.

Type	Method	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Imbalance	Adapter	35.22	47.31	54.18	58.79	62.71	64.51	66.48	67.78	68.70	70.10
	LoRA	36.01	49.87	56.63	61.55	**64.57**	**66.62**	**68.31**	**69.62**	**70.65**	**71.80**
	VPT	**41.38**	**51.80**	**57.30**	**61.75**	63.61	66.60	67.85	69.23	70.18	70.35
	Average	37.54	49.66	56.04	60.70	63.63	65.91	67.55	68.88	69.84	70.75
Balance	Adapter	40.44	49.48	55.75	60.14	63.23	65.75	67.56	68.59	69.30	70.15
	LoRA	40.85	51.57	58.64	63.09	**65.92**	**67.68**	**69.23**	**70.14**	**70.95**	**71.60**
	VPT	**41.63**	**53.95**	**59.71**	**63.41**	65.29	65.26	68.53	69.68	69.58	70.53
	Average	40.97 (+3.43)	51.67 (+2.01)	58.03 (+1.99)	62.21 (+1.51)	64.81 (+1.18)	66.23 (+0.32)	68.44 (+0.89)	69.47 (+0.59)	69.94 (+0.10)	70.76 (+0.01)

Table 3. PEFT performance by domain, data size, and balance condition (Top-1 accuracy, %).

Domain	Type	Method	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Natural	Imbalance	Adapter	31.30	47.58	56.05	62.50	67.70	71.52	73.76	75.50	76.55	78.16
		LoRA	**31.64**	**50.14**	**58.44**	**64.61**	**69.17**	**72.22**	**74.29**	**76.24**	**77.58**	**78.87**
		VPT	28.52	43.60	54.69	59.90	65.02	67.99	69.77	71.85	73.01	74.00
	Balance	Adapter	45.04	55.00	60.62	65.64	69.78	73.52	75.33	76.64	77.60	78.22
		LoRA	**45.43**	**57.26**	**64.56**	**68.56**	**72.27**	**75.03**	**76.46**	**77.59**	**78.07**	**78.58**
		VPT	30.65	50.58	59.94	64.15	67.05	69.59	71.74	72.60	73.54	73.78
Specialized	Imbalance	Adapter	62.05	71.34	74.61	78.32	79.53	80.68	81.44	81.95	82.43	83.18
		LoRA	62.89	72.01	76.01	79.13	80.22	81.75	82.21	82.77	83.18	83.70
		VPT	**69.73**	**75.92**	**78.58**	**80.70**	**81.53**	**82.45**	**83.05**	**83.21**	**84.02**	**83.96**
	Balance	Adapter	62.35	68.21	71.85	78.15	80.41	81.53	81.76	82.54	82.55	83.49
		LoRA	61.51	68.34	74.40	79.80	81.36	81.84	82.36	83.07	83.69	83.56
		VPT	**72.08**	**75.46**	**79.08**	**80.89**	**81.52**	**82.16**	**83.43**	**83.82**	**84.21**	**84.21**
Structured	Imbalance	Adapter	25.22	35.06	42.32	45.78	49.93	50.28	52.63	53.95	54.96	56.51
		LoRA	26.40	38.56	45.36	50.09	52.73	54.17	56.12	57.26	58.31	59.66
		VPT	**38.45**	**46.92**	**48.94**	**53.90**	**53.43**	**57.46**	**58.57**	**59.95**	**60.79**	**60.35**
	Balance	Adapter	25.47	35.29	43.43	46.33	48.92	51.07	53.66	54.56	55.41	56.42
		LoRA	26.52	38.21	45.58	49.96	52.64	**54.18**	56.35	57.15	58.35	59.50
		VPT	**36.01**	**46.14**	**49.84**	**54.02**	**55.63**	53.04	**58.27**	**60.07**	**58.80**	**60.85**

능이 역전되었다. VPT가 40%까지 근소한 우위를 유지했으나 50% 구간부터 LoRA가 역전하였고, 50%→100% 구간에서 LoRA는 불균형 설정 7.23%p(64.57→71.80), 균형 설정 5.68%p(65.92→71.60)로 가장 큰 향상을 보였다. 이는 LoRA가 MHA 블록의 Query·Key 행렬에 직접 작용하여 어텐션의 표현력을 점진적으로 정제하기 때문이다. 이처럼 데이터가 적은(10%~40%) 구간의 VPT와 데이터가 많은(70%~100%) 구간의 LoRA가 교차하는 명확한 성능 전환점은 데이터 규모를 연속적으로 통제할 때 비로소 드러난다.

종합하면 Table 2는 두 가지 핵심 사실을 보여준다. 첫째, 최적의 PEFT 기법은 고정되어 있지 않고 데이터 규모에 따라 달라진다. 둘째, 클래스 균형화의 효과는 데이터가 적은 구간에 집중되며, 데이터가 늘어날수록 균형 부분집합이 원본 분포에 수렴하면서 그 효과가 사라진다. 이러한 균형화 효과의 데이터 의존성은 이어지는 IV.3 절에서 더욱 뚜렷하게 나타난다.

아울러 비교의 기준점을 제공하기 위해 Table 4와 같이 대표 데이터셋(SVHN·Retinopathy·Clevr-Count)에서 완전 미세조정(full fine-tuning)과 linear probing 베이스라인을 동일 프로토콜로 평가하였다. 완전 미세조정은 자연 도메인의 SVHN에서는 100% 기준 90.4%로 PEFT 대비 약 6%p 높았으나, 구조적 도메인의 Clevr-Count에서는 45.1%로 LoRA(82.7%)의 절반 수준에 그쳤다. 특히 Retinopathy에서는 Top-1 정확도가 73.7%로 가장 높았음에도 balanced accuracy는 28.0%에 불과하여, 8,600만 개의 전체 파라미터가 소수의 불균형 표본에 적합되는 과정에서 다수 클래스로 붕괴함을 보여준다. 한편 분류기만 학습하는 linear probing은 SVHN(33.4%)과 Clevr-Count(31.6%)에서 Top-1 정확도가 가장 낮지만, Retinopathy에서는 Top-1 정확도가 73.6%로 높게 나타났음에도 balanced accuracy는 약 20%에 머물러 완전 미세조정과 마찬가지로 다수 클래스로 붕괴한 상태임을 보여준다. 요컨대 완전 미세조정의 우위는 도메인 정합성이 높고 Top-1 정확도로 평가하는 조건에 한정되며, 그 밖에 소규모 데이터 환경에서는 상한(완전 미세조정)과 하한(linear probing) 사이에서 PEFT가 정확도-비용 균형상 더 유리하다.

3. Domain-Specific Analysis

IV. 2절은 전반적인 경향만 보여주므로 데이터셋 및 도메인별로 분해하여 Table 3에 제시한다. Table 3은 본 연구의 핵심 발견, 즉 PEFT 기법이 도메인 특성에 따라 서로 다른 방식으로 사전학습 표현을 활용한다는 사실을 가장 직접적으로 뒷받침한다. LoRA와 Adapter는 사전학습 백본이 이미 보유한 표현을 내부 가중치 수준에서 점진적으로 조정하는 반면, VPT는 백본 표현을 그대로 유지한 채 입력 수준의 프롬프트만으로 데이터셋에 맞는 표현을 새로 구성한다. 이

Table 4. Reference baselines under the imbalanced setting (mean±std over three seeds): full fine-tuning as an upper-capacity reference and linear probing as a lower-capacity reference.

Domain	Method	Metric	10%	50%	100%
SVHN	Full FT	Top-1	46.7±8.4	85.9±1.3	90.4±0.7
		BAcc	41.5±5.9	84.1±2.0	89.3±1.0
	Linear Prob	Top-1	22.5±1.0	28.5±0.6	33.4±0.1
		BAcc	14.7±1.0	19.9±0.7	26.3±0.2
Retinopathy	Full FT	Top-1	70.8±2.1	72.4±1.3	73.7±0.7
		BAcc	24.5±1.7	27.8±1.5	28.0±1.1
	Linear Prob	Top-1	73.6±0.0	73.6±0.0	73.6±0.0
		BAcc	20.0±0.0	20.3±0.2	21.1±0.0
Clevr-Count	Full FT	Top-1	28.0±1.5	39.0±0.9	45.1±0.4
		BAcc	27.9±1.4	38.9±0.9	45.0±0.3
	Linear Prob	Top-1	19.7±4.1	29.3±0.4	31.6±0.0
		BAcc	19.7±4.0	29.1±0.6	31.4±0.0

는 도메인별로 분해할 때 비로소 드러난다. 사전학습과 유사한 자연(Natural) 도메인에서는 LoRA·Adapter가 뚜렷이 우세하였다. 불균형 100%에서 LoRA(78.87%)와 Adapter(78.16%)가 최고 성능을 기록한 반면 VPT는 가장 낮았다. 자연 도메인은 사전학습 데이터셋(ImageNet-21k)과 시각 개념을 공유하여 ViT 특징이 이미 효과적이므로, 강력한 표현이 확립된 경우에는 내부 표현을 직접 정제하는 LoRA·Adapter가 입력에 프롬프트만 더하는 VPT보다 효과적이다. 반면 사전학습과 이질적인 특수(Specialized)·구조적(Structured) 도메인에서는 VPT가 우세하였다. 불균형 10%에서 VPT는 특수 도메인에서 69.73%로 LoRA(62.89%)·Adapter(62.05%)를 6%p 이상, 구조적 도메인에서 38.45%로 LoRA(26.40%)·Adapter(25.22%)를 12%p 이상 앞섰으며, 이 우위는 전 구간에서 유지되었다. 사전학습 표현이 부합하지 않는 영역에서는 특징을 수정하기보다 입력 프롬프트로 모델을 데이터셋에 맞게 조정하는 VPT가 효과적이기 때문이며, 특히 새로운 기하학적 규칙 학습이 필요한 구조적 데이터셋에서 내부 가중치 일부만 바꾸는 LoRA·Adapter의 한계가 두드러진다.

클래스 균형화의 효과 또한 도메인에 따라 갈렸다. 자연 도메인에서는 데이터가 적은 구간의 이득이 가장 커 10% 지점에서 LoRA가 31.6%에서 45.4%로 13.8%p 상승하였다. 특수 도메인에서는 기법별로 상이했으며, 구조적 도메인에서는 이득이 대체로 ±1%p 이내로 미미하고 VPT는 여러 구간에서 오히려 하락하여(예: 60% 지점 57.5%→53.0%), 개수·거리 등 데이터셋 정보가 레이블 분포 자체에 반영된 구조적 데이터셋에서는 인위적 균등화가 유효 정보를 왜곡할 수 있음을 시사한다. 이러한 균형화 역효과는 균형 부분집합을 만드는 방식에서 비롯된다. 클래스마다 표본 수를 똑같이 맞추려면 다수 클래스의 표본 일부를 버려야 하는데, 이때 두 가지 비용이 생긴다. 첫째, 같은 데이터 비율이라도 실제로 학습에 쓰는 정보량이 줄어든다. 둘째, 테스트 데이터는 여전히 원본의 불균형 분포를 따르므로, 균형하게 학습한 모델과 불균형한 테스트 분포 사이에 클래스 사전 확률의 불일치(prior shift)가 생긴다. 데이터가 극히 적을 때는 소수 클래스의 최소 표본을 확보하고 배치 구성을 안정화하는 이득이 이 비용보다 크다. 그러나 데이터가 늘수록 이 이득은 포화되는 반면 정보 손실과 분포 왜곡의 비용은 계속 쌓여, 결국 균형화의 순효과가 음(−)으로 바뀐다. 구조적 도메인에서 역효과가 특히 큰 것도 같은 이유다. 개수·거리처럼 연속적인 값을 이산화한 구조적 데이터셋에서는 레이블 분포 자체가 데이터셋의 기하학적 규칙을 담은 정보인데, 억지로 클래스 수를 맞추면 이 정보가 그만큼 더 많이 훼손되기 때문이다. 이 두 비용은 균형화가 부분적으로만 일어나는 중간 구간에서 가장 크게 나타난다. 실제로 Table 3의 구조적 도메인에서 VPT는 60% 지점에서 균형화 이후 57.5%에서 53.0%로 4.4%p 하락하였다. 반면 100% 지점에서는 균형 추출이 원본 분포에 수렴하여 그 차이가 ±0.5%p 이내로 사라진다(예: Adapter 56.5%→56.4%). 요컨대 균형화는 어디서나 유효한 전처리가 아니라, 데이터 규모와 도메인 특성에 따라 효과가 달라지는 조건부 설계 선택이다.

4. Statistical Reliability and Generalization Analysis

관측된 성능 차이가 무작위 변동의 산물이 아님을 검증하기 위해 두 층위의 통계 분석을 수행하였다. 먼저 도메

Table 5. Top-1 accuracy, balanced accuracy (BAcc), and macro-F1 (mean±std, 3 seeds) on three representative datasets, including DoRA as a recent variant. Bold: BAcc on Retinopathy at 10%, where imbalanced training collapses toward the majority class

Dataset	Method	Metric	Imbalance			Balance		
			10%	50%	100%	10%	50%	100%
SVHN (Natural)	Adapter	Top-1	22.2±2.3	61.7±2.1	78.8±1.3	18.4±2.0	62.4±0.8	78.9±0.2
		BAcc	17.1±1.9	56.1±1.1	75.4±1.6	19.0±1.1	61.4±1.0	75.5±0.4
		macro-F1	14.9±2.3	55.2±1.6	75.6±1.7	17.1±1.8	59.8±1.0	75.6±0.4
	LoRA	Top-1	26.6±4.3	70.5±0.9	82.0±0.8	21.8±1.4	72.6±1.7	81.9±0.5
		BAcc	20.1±2.7	66.1±0.8	79.5±1.1	21.9±1.4	71.5±1.3	79.3±0.7
		macro-F1	17.2±3.4	65.9±1.1	79.8±1.0	19.9±1.4	70.5±1.3	79.7±0.7
	VPT	Top-1	33.4±11.5	74.5±4.6	84.8±1.5	28.7±6.7	77.0±5.0	85.2±2.4
		BAcc	27.4±8.0	70.8±6.2	82.7±1.6	28.0±6.6	76.0±5.0	83.3±2.9
		macro-F1	26.5±7.9	71.1±5.7	83.1±1.6	26.3±5.8	75.3±5.1	83.6±2.7
	DoRA	Top-1	26.5±3.8	71.3±0.6	82.0±0.5	21.7±1.8	72.2±1.1	81.8±0.3
		BAcc	20.1±2.5	66.6±0.5	79.3±0.7	21.8±1.4	71.0±0.8	79.4±0.2
		macro-F1	17.0±3.2	66.5±0.5	79.8±0.7	19.8±1.7	70.2±0.7	79.8±0.2
Retinopathy (Specialized)	Adapter	Top-1	66.4±3.1	70.0±2.0	70.9±1.2	32.8±4.9	61.1±4.9	70.8±0.8
		BAcc	24.7±1.1	31.4±2.1	37.1±0.8	36.8±2.6	38.9±0.3	36.8±0.4
		macro-F1	23.4±2.3	32.5±2.3	39.6±0.5	23.0±1.0	38.1±0.7	39.2±0.6
	LoRA	Top-1	68.3±1.6	68.6±2.3	69.1±0.6	30.4±3.0	57.2±1.8	69.1±0.9
		BAcc	23.4±0.7	34.4±0.9	37.1±0.1	38.0±1.7	39.7±0.2	37.2±1.4
		macro-F1	22.6±1.6	35.8±1.6	38.9±0.2	21.9±0.5	38.0±0.3	39.1±0.6
	VPT	Top-1	64.3±7.5	63.8±0.5	64.4±0.5	30.8±1.0	45.7±7.4	65.2±1.6
		BAcc	27.6±3.3	31.3±3.0	32.8±0.4	39.0±0.9	37.3±2.0	32.2±1.1
		macro-F1	26.9±3.0	32.8±3.0	34.6±0.7	24.6±1.1	34.1±0.6	34.3±1.0
	DoRA	Top-1	68.0±2.3	68.2±1.8	68.4±1.7	31.0±2.6	54.0±1.4	68.9±0.5
		BAcc	23.2±0.5	34.9±1.5	36.8±0.6	37.9±1.2	39.9±0.4	36.5±0.7
		macro-F1	22.4±1.3	36.0±2.3	38.8±0.3	22.4±0.9	37.7±0.5	38.4±0.8
Clevr-Count (Structured)	Adapter	Top-1	37.7±0.7	74.2±1.2	82.1±0.3	36.4±5.4	76.8±0.3	81.6±0.4
		BAcc	37.6±0.7	74.1±1.2	82.0±0.3	36.3±5.4	76.8±0.3	81.6±0.4
		macro-F1	34.0±1.1	73.8±1.1	82.0±0.3	35.2±4.6	76.6±0.2	81.5±0.4
	LoRA	Top-1	36.7±0.9	77.8±0.3	82.7±0.3	37.5±1.0	77.6±1.1	82.8±0.4
		BAcc	36.6±1.0	77.7±0.3	82.7±0.3	37.5±1.0	77.5±1.1	82.7±0.4
		macro-F1	33.3±1.2	77.8±0.4	82.8±0.3	36.4±0.7	77.6±1.1	82.8±0.4
	VPT	Top-1	46.1±1.5	70.5±1.3	77.8±0.4	44.9±2.0	71.9±1.9	78.4±1.0
		BAcc	46.0±1.5	70.4±1.3	77.8±0.4	44.9±2.0	71.8±1.9	78.3±1.0
		macro-F1	45.6±1.7	70.3±1.2	77.7±0.4	44.2±2.2	71.6±2.0	78.3±1.0
	DoRA	Top-1	36.9±1.4	77.8±0.8	82.5±0.3	37.5±1.4	77.2±1.3	82.3±0.2
		BAcc	36.8±1.5	77.8±0.7	82.4±0.3	37.4±1.3	77.1±1.3	82.3±0.2
		macro-F1	33.2±1.5	77.8±0.8	82.5±0.3	36.4±1.2	77.2±1.4	82.4±0.2

인별 대표 데이터셋인 SVHN(자연), Retinopathy(특수), Clevr-Count(구조)에 대해, 부분집합은 고정한 채 학습 시드만 달리하여 3회 반복 학습하고 Top-1 정확도와 balanced accuracy, macro-F1의 평균과 표준편차를 Table 5에 제시한다. Adapter와 LoRA의 표준편차는 전 조건에서 0.1~5.4%p 수준으로, 도메인 분석에서 보고한 기법 간 격차보다 작아 해당 결론이 시드 변동에 강건함을 보여준다. 흥미롭게도 VPT는 극단적으로 데이터가 적은 (10%) 구간에서 표준편차가 최대 ±11.5%p에 달해, 프롬프트 초기화에 따른 학습 불안정성이 세 기법 중 가장 크

Table 6. Comparison of training times (mm:ss) by PEFT methods. Bold indicates the fastest method within each column.

Domain	Method	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Natural	Adapter	**14:32**	**15:44**	16:46	18:03	19:22	19:53	21:04	22:14	23:12	**23:29**
	LoRA	15:10	**15:44**	**16:32**	**17:25**	**18:38**	**19:18**	**20:48**	**21:58**	**22:43**	24:12
	VPT	16:22	17:53	18:34	20:06	21:30	22:24	22:58	24:54	25:05	26:57
Specialized	Adapter	**18:49**	**19:34**	21:18	22:49	23:24	24:50	25:10	26:19	27:37	28:02
	LoRA	19:45	19:52	**21:05**	**21:39**	**23:04**	**24:20**	**24:49**	**25:34**	**27:00**	**27:55**
	VPT	21:46	23:35	24:14	24:47	26:55	28:07	29:10	30:03	30:44	32:08
Structured	Adapter	**20:30**	22:09	23:28	24:41	25:18	27:15	28:18	28:28	29:35	31:35
	LoRA	21:06	**21:59**	**23:19**	**24:03**	**25:16**	**26:01**	**27:05**	**28:26**	**29:32**	**31:15**
	VPT	24:34	25:52	27:20	28:53	30:16	31:06	32:38	33:30	34:12	35:25

다는 새로운 관측을 제공한다. 이는 VPT의 소규모 데이터 우위가 평균적으로는 유효하되 개별 시행의 변동 폭이 크므로, 실무 적용 시 복수 시드 검증이 권장됨을 시사한다. 한편 100% 지점의 균형화 전후 격차는 세 기법·세 데이터셋 전체에서 -0.4~+0.8%p로 시드 표준편차 이내였는데, 이는 위 데이터 구조에서 기술한 바와 같이 고비율 구간에서 균형 추출이 원본 분포에 수렴하기 때문으로, 균형화의 효과가 데이터가 부족할수록 커지고 데이터가 늘수록 사라진다는 본문의 해석을 절차적 수준에서 뒷받침한다.

Table 5는 평가 지표의 선택이 결론에 미치는 영향도 드러낸다. 클래스 불균형이 극심한 Retinopathy의 10% 지점에서 불균형 설정은 세 기법 모두 Top-1 정확도 64~68%로 높아 보이지만 balanced accuracy는 23~28%로 우연 수준(20%)에 근접하여, 모델이 다수 클래스(No DR)로 붕괴한 상태임을 보여준다. 반대로 균형 학습은 Top-1 정확도가 30% 내외로 낮아지는 대신 balanced accuracy가 37~39%로 상승한다. 즉 데이터가 적은 구간에서 균형화로 Top-1 정확도가 하락하는 것은 실제 성능 저하가 아니라, 다수 클래스로 쏠려 부풀려졌던 Top-1 정확도가 공정한 balanced accuracy로 교정된 결과다. 이는 클래스 균형화가 평가 지표에 의존하며, 오분류 비용이 클래스마다 다른 의료 진단과 같은 응용에서는 Top-1 정확도 기준의 결론을 그대로 적용해서는 안 됨을 시사한다. 한편 macro-F1을 함께 보면 균형화의 효과가 보다 정확하게 규정된다. 같은 Retinopathy 10% 지점에서 균형 설정은 balanced accuracy를 크게 끌어올리지만(LoRA 23.4% → 38.0%) macro-F1은 개선되지 않는다(22.6% → 21.9%). balanced accuracy가 재현율만을 반영하는 반면 macro-F1은 정밀도를 함께 반영하므로, 이 대비는 균형화가 소수 클래스의 검출 성능을 높이는 대신 오탐을 늘리는 트레이드오프임을 보여준다. 즉 균형화는 소수 클래스 검출을 절대적으로 높이는 것이 아니라, 미검출(false negative)과 오탐(false positive) 사이의 균형점을 옮긴다. 따라서 균형화는 미검출 비용이 오탐 비용보다 큰 의료 스크리닝과 같은 응용에서 유효하며, 정밀도가 함께 요구되는 응용에서는 그 이득이 제한적이다. 반면 불균형이 완만한 SVHN에서는 균형화가 macro-F1까지 개선하여(Adapter 10% 기준 14.9% → 17.1%), 이러한 재현율-정밀도 교환이 불균형이 극심한 데이터셋에 국한된 현상임을 확인할 수 있다. 아울러 이 결과는 Table 2가 보여준 평균적 균형화 이득과 모순되지 않는다. 19개 데이터셋 평균의 이득은 불균형이 완만한 다수의 데이터셋에서 비롯되는 반면, Retinopathy와 같이 불균형이 극심한 데이터셋에서는 지표 간 교환이 지배적으로 나타나기 때문이다. 반면 클래스 분포가 본래 균등에 가까운 Clevr-Count에서는 두 지표가 사실상 일치하여(격차 0.1%p 내외), 지표 간 괴리 자체가 불균형의 진단 지표로 기능함을 확인할 수 있다. 이상의 시드 반복 분석은 도메인 분석(Table 3)에서 보고한 기법 간 순위와 균형화 효과의 방향이 대표 데이터셋 수준에서 재현됨을 보여주며, 주요 결론이 특정 시드나 단일 시행에 의존하지 않음을 뒷받침한다.

마지막으로, 제안 프로토콜이 특정 기법에 종속되지 않음을 확인하기 위해 LoRA의 최신 변형인 DoRA를 Table 5와 같이 구조적 수정 없이 동일 프로토콜에 적용하였다. DoRA는 가중치 갱신을 크기(magnitude)와 방향(direction)으로 분해하여 LoRA의 표현력을 개선한 기법이다. 실험 결과 DoRA는 본문이 보고한 핵심 경향을 모두 동일하게 재현하였다. 데이터 규모에 따른 성능 향상(SVHN 불균형 Top-1 정확도 26.5%→82.0%), 100% 지점에서의 균형화 효과 소멸(Top-1 정확도 불균형 설정 82.0% vs 균형 설정 81.8%), 그리고 소규모 데이터 불균

형 구간에서의 다수 클래스 붕괴(Retinopathy 10%: Top-1 정확도 68.0%, BAcc 23.2%)와 균형화에 의한 balanced accuracy 회복(23.2%→37.9%)이 그것으로, 본 프로토콜의 분석 축이 최신 변형 기법에도 그대로 적용됨을 보여준다. 주목할 점은 DoRA가 본 소규모 데이터 설정의 사실상 전 구간에서 LoRA를 넘어서지 못하고 대등한 수준에 머물렀다는 사실이다(예: Top-1 정확도 SVHN 100% 82.0% vs 81.8%, Clevr-Count 100% 82.5% vs 82.3%). 이는 충분한 데이터를 전제한 표준 벤치마크에서 보고된 개선이 극단적으로 데이터가 적은 환경으로 자동 이전되지 않음을 보여주는 결과로, 최신 기법의 우수성이 데이터가 적은 환경까지 자동으로 이어지지는 않으므로 이러한 환경에서는 개별적인 검증이 필요하다는 본 연구의 문제의식을 재확인한다.

5. Efficiency Analysis

연산 효율성 또한 실제 응용에서의 중요한 고려 사항이다. 각 기법의 학습 가능 파라미터 수는 Adapter 0.16M, LoRA 0.29M, VPT 1.07M으로, VPT가 가장 많고 Adapter가 가장 적다. 이 차이는 학습 시간에 직접 반영되어, Table 6을 보면, 측정된 학습 시간은 약 15~35분 범위에서 파라미터가 가장 많은 VPT가 모든 도메인에서 가장 길었고 Adapter가 가장 빨랐다. 반면 Table 7과 같이, 영상 당 평균 추론 시간은 세 기법 모두 8~10ms로 차이가 미미하여, 모든 기법이 실시간 서비스 환경에 적용 가능한 수준임을 보여준다. 나아가 파라미터 규모 차이가 기법 간 비교를 왜곡할 가능성을 검증하기 위해, LoRA(약 0.29M)와 파라미터 규모를 일치시킨 VPT(p=25, 약 0.27M)를 19개 데이터셋 전체에서 추가 평가하였다. 100% 데이터에서 VPT(p=25)는 70.5%로 LoRA(71.8%)와 대등한 성능을 유지하면서도 데이터가 적은 (10%) 구간의 우위(37.2% vs 36.0%)를 유지하였고, 100% 조건에서 19개 중 11개 데이터셋에서는 오히려 p=25가 p=100보다 높은 성능을 보였다(예: Pets에서 10% 기준 +8.5%p, 100% 기준 +2.2%p). 이는 파라미터 수가 성능의 1차적 결정 요인이 아니며, 데이터가 적은 구간의 기법 간 격차가 모델 용량 차이에서 비롯된 것이 아님을 보여준다. 이러한 prompt length 민감도는 원 VPT 문헌[11]의 보고와도 일치한다.

종합하면, LoRA는 Adapter 대비 약 1.8배 많은 파라미터로 자연 도메인과 전체 평균에서 확실한 우위를 확보하며 이 우위는 데이터가 많아질수록 두드러지므로, 데이터가 풍부할수록 효율이 극대화되는 균형 잡힌 전략이다. 반면 VPT는 상대적으로 많은 파라미터를 쓰지만 극단적으로 데이터가 적은 경우, 특히 특수·구조적 도메인에서 뛰어난 초기 성능을 달성한다. 다만 데이터가 늘어날수록 고정 프롬프트의 표현력이 한계에 도달해 향상이 둔화된다. 이를 바탕으로 다음 지침을 제시한다. 최소 자원으로 만족스러운 성능을 원할 때는 Adapter, 데이터가 풍부하여 확장 가능한 향상을 원할 때는 LoRA, 데이터가 매우 적거나 특수 도메인에서 강력한 초기 성능을 신속히 확보하려 할 때는 VPT가 가장 효과적이다.

Table 7. PEFT inference time comparison

Adapter	LoRA	VPT
9.58	9.28	8.42

V. Conclusions

본 연구는 VTAB-1k 상에 데이터 규모와 클래스 균형을 명시적으로 통제하는 체계적인 데이터 규모 별 클래스 균형 평가 프로토콜을 구축하고, 이를 통해 Adapter, LoRA, VPT를 동일 조건에서 심층 비교하였다. 기존 연구가 완전 데이터라는 단일 지점이나 불균형이 통제되지 않은 조건에서 PEFT를 평가한 것과 달리, 본 연구는 데이터 규모·도메인·클래스 분포를 동시에 통제함으로써 기법 간 성능 역전과 균형화 효과를 정량적으로 규명하였다는 점에서 차별화된다. 분석 결과 LoRA와 Adapter는 사전학습 표현을 내부 가중치 수준에서 조정하여 자연 도메인에서, VPT는 입력 프롬프트로 데이터셋에 맞는 표현을 새로 구성하여 특수·구조적 도메인에서 강점을 보였다. 또한 데이터 규모에 따라 동적으로 변화하는 명확한 트레이드오프(데이터가 적을 때의 VPT, 많을 때의 LoRA)와, 클래스 균형화 설정의 효과가 데이터 규모와 평가 지표에 조건부임을 확인하였다. 균형화 설정은 데이터가 적은 구간에서 Top-1 정확도를 다소 낮추는 대신 모든 클래스를 고르게 평가하는 balanced accuracy를 높이는 교환 관계를 보이나, 오탐이 함께 증가하여 정밀도를 반영하는 macro-F1은 개선되지 않아 이는 성능의 절대적 향상이 아니라 미검출과 오탐 사이의 균형점을 옮기는 것으로 해석된다.

학술적으로 본 연구는 단일 기법의 우열을 넘어 PEFT의 작동 원리를 기법별 적응 방식, 즉 내부 가중치를 조정하는 방식과 입력 프롬프트로 데이터셋에 맞는 표현을 구성하는 방식의 차이라는 관점에서 해석하는 틀을 제시하며, 실용적으로는 도메인 특성·데이터 규모·클래스 분포에

따른 구체적 기법 선택 지침을 제공한다. 예컨대 판독 레이블 확보 비용이 커 표본이 수십 장 수준에 그치고 정상 소견이 대부분을 차지하는 의료 영상 진단에서는, 소규모 데이터·특수 도메인에 강건한 VPT를 우선 적용하고 클래스 균형 추출을 병행하는 구성이 유리하다. 반면 레이블링이 비교적 용이하여 표본을 충분히 확보할 수 있는 원격탐사 영상 분류에서는 데이터 증가에 따른 향상 폭이 가장 큰 LoRA가 적합하다. 산업 검사와 같이 결함 유형이 지속적으로 추가되어 신규 데이터셋마다 신속한 초기 성능 확보가 요구되는 환경에서는, 백본을 공유한 채 프롬프트만 교체할 수 있는 VPT가 배포 효율 측면에서 이점을 가진다. 한편 본 연구는 세 대표 기법과 영상 분류에 초점을 맞추었으나, 제안한 평가 프로토콜은 다양한 PEFT 변형 기법을 동일 조건에서 평가할 수 있는 확장 가능한 틀을 제공한다. 실제로 본 연구는 이 틀을 LoRA의 최신 변형인 DoRA에 구조적 수정 없이 적용하여, 프로토콜의 기법 비종속성과 함께 표준 벤치마크의 개선이 소규모 데이터 환경으로 자동 이전되지 않음을 확인하였다. 이러한 범용적 설계는 같은 방식으로 AdaptFormer, E²VPT 등 Adapter·VPT 계열의 변형에도 수정 없이 적용될 수 있으며, 이들에 대한 검증은 본 프로토콜의 확장성을 추가로 입증할 후속 과제다. 향후 객체 검출·영상 분할 등 다양한 하위 데이터셋으로 분석을 확장하고, 데이터 규모에 따라 기법을 동적으로 전환하는 전략을 탐구하고자 한다.

ACKNOWLEDGEMENT

This work was supported in part by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No.RS-2022- NR071978) and funded by the Ministry of Education (No.RS-2022-NR070869); supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No.RS-2022-II220448: Deep Total Recall, 10%, and IITP-2026-RS-2026-25544527: Leading Generative AI Human Resources Development, RS-2026-25548560: Artificial Intelligence Innovation Graduate School (Inha University)).

REFERENCES

[1] J. Zhang, J. Huang, S. Jin, and S. Lu, “Vision-language models for vision tasks: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., Vol. 46, No. 8, pp. 5625-5644, Aug. 2024. DOI: 10.1109/TPAMI.2024.3369699

[2] A. Dosovitskiy, et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, May 2021.

[3] M. Oquab, et al., “DINOv2: Learning robust visual features without supervision,” Trans. Mach. Learn. Res. (TMLR), Jan. 2024. DOI: 10.48550/arXiv.2304.07193

[4] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, “Scaling vision transformers,” Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1204-1213, New Orleans, USA, Jun. 2022. DOI: 10.1109/CVPR52688.2022.01179

[5] A. Kirillov, et al., “Segment anything,” Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 3992-4003, Paris, France, Oct. 2023. DOI: 10.1109/ICCV51070.2023.00371

[6] J.-X. Shi, T. Wei, Z. Zhou, J.-J. Shao, X.-Y. Han, and Y.-F. Li, “Long-tail learning with foundation model: Heavy fine-tuning hurts,” Proc. Int. Conf. Mach. Learn. (ICML), pp. 45014-45039, Vienna, Austria, Jul. 2024. DOI: 10.48550/arXiv.2309.10019

[7] V. Lialin, V. Deshpande, and A. Rumshisky, “Scaling down to scale up: A guide to parameter-efficient fine-tuning,” arXiv:2303.15647, Mar. 2023. DOI: 10.48550/arXiv.2303.15647

[8] J. He, et al., “Towards a unified view of parameter-efficient transfer learning,” Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, Apr. 2022. DOI: 10.48550/arXiv.2110.04366

[9] N. Houlsby, et al., “Parameter-efficient transfer learning for NLP,” Proc. Int. Conf. Mach. Learn. (ICML), pp. 2790-2799, Long Beach, USA, Jun. 2019. DOI: 10.48550/arXiv.1902.00751

[10] E. J. Hu, et al., “LoRA: Low-rank adaptation of large language models,” Proc. Int. Conf. Learn. Represent. (ICLR), Virtual, Apr. 2022. DOI: 10.48550/arXiv.2106.09685

[11] M. Jia, et al., “Visual prompt tuning,” Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 709-727, Tel Aviv, Israel, Oct. 2022. DOI: 10.1007/978-3-031-19827-4_41

[12] M. Tsimpoukelli, et al., “Multimodal few-shot learning with frozen language models,” Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 200-212, Virtual, Dec. 2021. DOI: 10.48550/arXiv.2106.13884

[13] X. Zhai, et al., “A large-scale study of representation learning with the visual task adaptation benchmark,” arXiv:1910.04867, Oct. 2019. DOI: 10.48550/arXiv.1910.04867

[14] S. Liu, et al., “DoRA: Weight-decomposed low-rank adaptation,” Proc. Int. Conf. Mach. Learn. (ICML), pp. 32100-32121, Vienna, Austria, Jul. 2024. DOI: 10.48550/arXiv.2402.09353

[15] Y. Xin, et al., “Parameter-efficient fine-tuning for pre-trained

vision models: A survey and benchmark," arXiv:2402.02242, Feb. 2024. DOI: 10.48550/arXiv.2402.02242

[16] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," arXiv:1606.08415, Jun. 2016. DOI: 10.48550/arXiv.1606.08415

[17] E. B. Zaken, Y. Goldberg, and S. Ravfogel, "BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL), pp. 1-9, Dublin, Ireland, May 2022. DOI: 10.18653/v1/2022.acl-short.1

[18] Y. Zhang, K. Zhou, and Z. Liu, "Neural prompt search," IEEE Trans. Pattern Anal. Mach. Intell., Vol. 47, No. 7, pp. 5268-5280, Jul. 2025. DOI: 10.1109/TPAMI.2024.3435939

[19] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," IEEE Trans. Pattern Anal. Mach. Intell., Vol. 45, No. 9, pp. 10795-10816, Sep. 2023. DOI: 10.1109/TPAMI.2023.3268118

[20] S. Chen, et al., "AdaptFormer: Adapting vision transformers for scalable visual recognition," Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 16664-16678, New Orleans, USA, Dec. 2022. DOI: 10.48550/arXiv.2205.13535

[21] C. Han, et al., "E²VPT: An effective and efficient approach for visual prompt tuning," Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 17491-17502, Paris, France, Oct. 2023. DOI: 10.1109/ICCV51070.2023.01604

Authors

Du-Hwan Hur received the B.S. degree in Computer Engineering from Hanbat National University in 2021, and is currently pursuing the M.S. - Ph.D. integrated course degree with the Department of Electrical Computer Engineering at Inha University, Korea. His current research interests are object detection, model compression, efficient learning, and on-device AI.

Dong-Hyun Kim received the B.S. degree in Computer Science and Engineering from Inha University in 2025, and is currently pursuing an M.S. - Ph.D. integrated course degree with the Department of Electrical and Computer Engineering at Inha University. His current research interests are object detection, efficient learning, and continual learning.

Seung-Hwan Bae received the BS degree in information and communication engineering from Chungbuk National University, in 2009 and the MS and PhD degrees in information and communications from the Gwangju Institute of Science and Technology (GIST), in 2010 and 2015, respectively. He was a senior researcher at Electronics and Telecommunications Research Institute (ETRI) in Korea from 2015 to 2017. He was an assistant professor in the Department of Computer Science and Engineering at Incheon National University, Korea from 2017 to 2020. He is currently an Associate Professor with the Department of Electrical and Computer Engineering at Inha University, His research interests include object tracking, object detection, generative model learning, continual learning, on-device ML, etc.