레터논문 (Letter Paper)
방송공학회논문지 제31권 제1호, 2026년 1월 (JBE Vol.31, No.1, January 2026)
https://doi.org/10.5909/JBE.2026.31.1.185
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

CLS Token 기반 LoRA 미세튜닝을 통한 합성-실사 도메인 적응 및 시맨틱 표현 분석

용 윤 정[a)‡]

CLS-Path LoRA Fine-Tuning for Synthetic-to-Real Domain Adaptation and Semantic Alignment Analysis

Yunjeong Yong[a)‡]

요 약

합성 데이터는 비전 모델 학습의 데이터 부족을 해결할 대안이나, 실사 도메인과의 시맨틱 및 분포 격차로 인해 적용에 한계가 있다. 본 논문은 Vision Transformer(ViT)의 CLS 토큰 경로에 집중하여 이 격차를 해소하는 Residual LoRA와 대조 정렬을 결합한 새로운 PEFT 구조를 제안한다. DINOv2 백본을 동결하여 기하학적 특징을 보존하고, 출력단에 Residual LoRA, Cosine Classifier, Contrastive Alignment를 결합해 특징 공간에서의 사후 보정(Post-hoc Refinement)을 수행한다. 특히 클래스당 10장의 실사 데이터(10-shot)만을 사용하는 극한 환경에서, 제안 기법은 전체 파라미터의 약 1% 수준 학습으로 합성 데이터 단독 학습 대비 8.7%의 정확도 향상을 달성했다. 또한 t-SNE 분석을 통해 시맨틱 공간에서의 효과적인 도메인 정렬을 확인하였다.

Abstract

Synthetic data addresses data scarcity in vision models but suffers from semantic gaps when applied to real-world scenarios. This paper introduces a novel parameter-efficient fine-tuning (PEFT) architecture targeting the CLS token pathway to mitigate this gap. We freeze the DINOv2 backbone to preserve geometric representations and perform post-hoc refinement using a Residual LoRA module, cosine classifier, and contrastive alignment head. Under an extreme 10-shot real-data setting, our method updates about 1% of parameters yet achieves an 8.7% accuracy improvement over synthetic-only training. t-SNE analysis confirms effective semantic alignment between synthetic and real domains.

Keyword : Domain Adaptation, Synthetic Data, Vision Transformer, CLS Token, Residual LoRA

a) 씨이랩 AI/ML팀(AI/ML Team, Xiilab)
‡ Corresponding Author : 용윤정(Yunjeong Yong)
E-mail: y.yong@xiilab.com
Tel: +82-2-2039-3145
ORCID: https://orcid.org/0009-0001-3164-8569

※ This work was supported by the Technology Innovation Development Program for SMEs (Market Expansion Type) funded by the Ministry of SMEs and Startups (MSS), Republic of Korea, and managed by the Korea Technology and Information Promotion Agency for SMEs (TIPA) (RS-2024-00470370).
· Manuscript December 4, 2025; Revised January 8, 2026; Accepted January 9, 2026.

Ⅰ. 서 론

대규모 합성 데이터를 활용한 비전 모델 학습은 비용을 절감할 수 있으나, 텍스처·조명 차이에 따른 도메인 갭(Domain Gap)으로 실환경 성능 저하가 발생한다[1,2]. 이를 완화하기 위한 대규모 백본(DINOv2[3] 등) 미세튜닝은 과적합·파국적 망각 위험과 높은 계산 비용을 동반하며, LoRA[4] 등 PEFT 또한 주로 Transformer 내부(QKV/FFN) 연산에 개입하는 방식이 일반적이다. 본 연구는 ViT[5]에서 전역 표현을 담당하는 CLS 토큰에 착안하여, 백본을 완전 동결한 상태에서 CLS 특징 공간에만 최소 파라미터를 주입하는 사후 보정(Post-hoc Refinement) 기반 구조로 합성-실사 시맨틱 불일치를 완화한다. 구체적으로 Residual LoRA와 Cosine classifier[6], Contrastive Alignment[7]를 결합해 표현 공간 잔차 보정과 시맨틱 정렬을 동시에 수행한다. 제안 기법은 10-shot 실사 환경에서 전체 파라미터의 약 1% 미만(0.26M)만 학습하면서도 합성 단독 학습 대비 +8.70%p(70.17%→78.87%) 성능 향상을 달성하였다. 주요 기여는 다음과 같다.

CLS 경로 특화 Residual LoRA: 백본 동결 상태에서 CLS 출력에 저랭크 잔차 보정을 적용하여 도메인 불일치를 완화한다.

안정적 정렬 구조: Cosine Classifier와 Contrastive Alignment를 결합해 도메인 정렬과 클래스 응집도를 강화한다.

효율성 입증: 전체 파라미터의 1% 미만만 학습하여 47개 클래스 분류에서 유의미한 성능 향상을 입증하였다.

Ⅱ. 제안 방법

제안하는 모델의 전체 구조는 그림 1과 같으며, 시맨틱 불일치를 CLS 특징 공간의 사후 보정으로 해결한다.

1. CLS 경로 특화 Residual LoRA

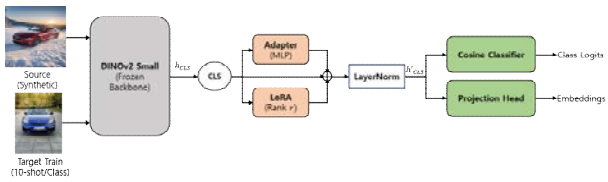

그림 1. CLS-Residual LoRA 기반 모델 구조
Fig. 1. Overview of the proposed CLS-Residual LoRA architecture.

본 논문의 핵심은 LoRA의 적용 위치를 Transformer 내부(QKV/FFN) 가중치(W_q, W_v)가 아닌 최종 출력단 특징 공간으로 재정의한 사후 보정 전략에 있다. 클래스당 실사 데이터가 10장에 불과하여 데이터 확보가 극도로 제한된 환경에서는, 모델 내부 가중치를 직접 수정할 경우 사전 학습된 범용적 기하학적 특징이 훼손되거나 소량의 데이터에만 편향되는 과적합 문제가 발생하기 쉽다.

이를 해결하기 위해 DINOv2 백본을 완전 동결하여 특징 추출 능력을 보존하면서, 식 (1)과 같이 전역 시맨틱이 응축된 h_{CLS}에 Adapter와 LoRA 출력을 잔차로 더해 합성-실사 간 분포 차이를 정교하게 보정한다.

2. 강건한 정렬 전략: Cosine & Contrastive

도메인 적응 과정에서의 안정성을 극대화하기 위해, 본 연구는 스케일 불변성과 시맨틱 응집을 동시에 확보하는 전략을 도입하였다. 첫째, 특징 벡터의 크기에 민감한 일반적인 선형 분류기 대신 Cosine Classifier를 채택하여, 벡터의 방향성에 기반한 분류를 수행함으로써 도메인 간 스케일 차이에 강건하게 대응한다. 둘째, 별도의 Projection Head를 통해 Contrastive Alignment Loss를 적용하여, 동일 클래스에 속한 합성 데이터와 실사 데이터가 임베딩 공간에서 서로 밀집하도록 강제한다. 이는 단순한 분류 경계 학습을 넘어, 두 도메인의 시맨틱 분포를 근본적으로 일치시키는 역할을 한다.

Ⅲ. 실험 및 결과

1. 데이터셋 및 실험 환경

$$h'_{CLS} = LayerNorm(h_{CLS} + F_{Adapt}(h_{CLS}) + F_{LoRA}(h_{CLS})) \quad (1)$$

표 1. CLS-Residual LoRA 구성 요소별 Ablation Study 결과
Table 1. Ablation Study Results for Each Component of the Proposed CLS-Residual LoRA Method

Category	Classifier	Res. Lora	Adapter	Contrastive	Target Train Data (10)	Valid Accuracy (%)
Baseline (Source Only)	Linear	-	-	-	-	70.17
Cosine Baseline	Cosine	-	-	-	√	78.10
LoRA Only (Linear)	Linear	√	-	-	√	77.77
LoRA + Cosine	Cosine	√	-	-	√	77.61
LoRA + Cosine + Contrastive	Cosine	√	-	√	√	77.61
CLS-Residual LoRA	Cosine	√	√	√	√	78.87

47개 클래스에 대해 합성(Source, 약 19만 장)에서 실사(Target)로의 적응을 평가하였다. 실사 데이터는 클래스당 10장만 학습(10-shot)하고, 별도 실사 4,278장으로 평가했으며, 학습/평가 간 중복을 제거해 데이터 누수를 방지하였다. Source는 SD3.5[8]로 조명·배경·시점 변화를 포함해 생성했고, Target은 웹 수집 후 라벨 정합 검수 및 노이즈 제거를 거쳐 구축하였다. 백본은 DINOv2-Small을 사용했으며 학습 파라미터는 1.1%(0.26M)이다.

2. Ablation Study

제안 기법의 각 구성 요소가 성능에 미치는 영향을 표 1에서 분석하였다. 합성 데이터만 사용한 Baseline은 70.17%였으나, 클래스당 10장의 실사 데이터를 활용한 Cosine Baseline은 78.10%로 가장 큰 성능 향상을 보였다. 또한 Residual LoRA, Adapter, Contrastive Alignment를 결합한 CLS-Residual LoRA는 78.87%로 최고 성능을 기록하여, 백본 수정 없이 CLS 특징 공간의 잔차 보정만으로도 효과적인 합성 실사 도메인 적응이 가능함을 확인하였다. 특히 LoRA를 Transformer 내부가 아닌 CLS 출력 특징 공간의 사후 보정으로 적용함으로써, 사전학습 표현을 보존하면서 도메인 시맨틱 갭을 효율적으로 완화한다.

3. 기존 기법 대비 성능 비교

클래스당 실사 10장만을 사용하는 제한된 도메인 전이 설정에서 제안 기법과 기존 미세튜닝 및 PEFT 기법의 성능을 비교하였다. 그 결과는 표 2에 정리하였다. Baseline (Source Only)은 70.17%로 도메인 갭으로 인한 성능 저하를 확인하였고, Linear Probe(75.42%) 및 Cosine Baseline (78.10%)은 백본 동결 상태에서도 타겟 데이터로 성능을 개선하였다. Standard LoRA(77.95%)와 Houlsby Adapter (77.63%)는 백본 일부를 업데이트했으나, 제안 기법은 백본을 완전 동결한 채 1.1%(0.26M)만 학습하고도 78.87%를 달성하였다. 이는 Full Fine-tuning(79.24%) 대비 0.37%p 차이에 불과하면서도 학습 파라미터를 약 91배 절감한 결과로, 극소량 실사 데이터 환경에서 효율적인 대안임을 보여준다.

4. CLS 토큰 기반 표현의 타당성 검증

CLS 토큰만으로 도메인 적응이 충분하다는 가정을 검증하기 위해 동일 조건에서 표현 방식만 변경하여 비교 실험을 수행하였으며 결과는 표 3에 정리하였다. CLS-only는 79.34%로 가장 높은 정확도를 기록했고, Patch-only (76.45%) 및 Full Patch(77.89%)는 오히려 낮은 성능을 보였다. 이는 제한된 데이터 환경에서 patch 기반 표현이 텍스

표 2. 기존 기법 대비 성능 비교
Table 2. Performance comparison with conventional baselines under the synthetic-to-real 10-shot setting

Method	Head	Backbone Update	Trainable Parameters (%)	Valid Accuracy (%)
Baseline (Source Only)	Linear	Frozen	-	70.17
Linear Probe[5, 7]	Linear	Frozen	<0.1	75.42
Cosine Baseline[6]	Cosine	Frozen	<0.1	78.10
Standard LoRA (QKV)[4]	Linear	Partial	1.8	77.95
Adapter (Houlsby)[9]	Linear	Partial	3.2	77.63
Full Fine-tuning[3]	Linear	Full	100	79.24
CLS-Residual LoRA (Ours)	Cosine	Frozen	1.1(0.26M)	78.87

표 3. CLS 토큰 기반 표현의 타당성 검증(합성→실사 10-shot)
Table 3. Validation of CLS-token assumption under the synthetic-to-real 10-shot setting

Method	Representation	Valid Accuracy (%)	Trainable Parameters (%)	GFLOPs
Patch-only	Mean pooling of patch tokens	76.45	0.28	1.84
CLS+Patch	Concat (CLS, pooled patch)	78.92	0.52	2.31
Full Patch	All patch tokens + attention pooling	77.89	2.14	18.42
CLS-only (Ours)	CLS token	79.34	0.26	0.12

처·조명 변화에 민감해 도메인 편향 및 과적합을 유발할 수 있음을 시사하며, CLS-only가 성능과 효율을 동시에 만족하는 안정적인 표현임을 정량적으로 뒷받침한다.

5. 시맨틱 정렬 분석

그림 2는 제안 기법 적용 후 47클래스 전체에 대해 산출된 CLS 토큰 임베딩의 t-SNE 결과를 보여주며, 합성 - 실사 도메인 간 시맨틱 정렬 효과를 정성적으로 확인할 수 있다. 동일 클래스 내에서 두 도메인의 샘플들이 밀집하며 혼합된 분포를 형성하는 것을 확인할 수 있으며, 이는 제안된 Residual LoRA 기반 정렬 구조가 도메인 간 시맨틱 갭을 효과적으로 감소시켰음을 시각적으로 보여준다. 또한 정렬 이후에도 클래스 간 분리도가 유지되는 것을 확인하여, 단순 혼합이 아닌 시맨틱 구조 보존 하의 정렬이 이루어졌음을 확인하였다.

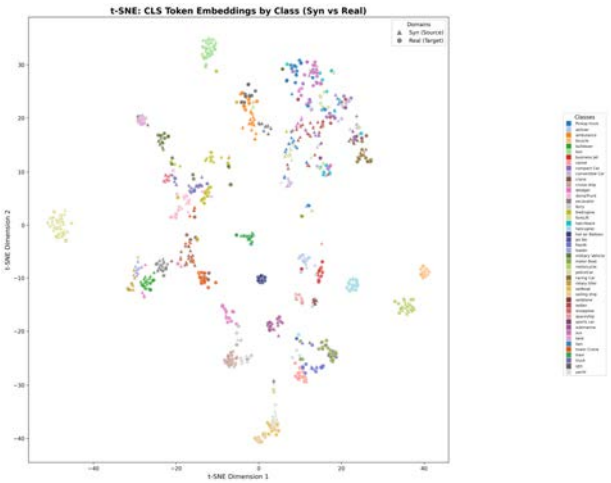

그림 2. CLS 토큰 임베딩의 t-SNE 시각화
Fig. 2. t-SNE visualization of CLS token embeddings (combined by class and domain)

Ⅳ. 결 론

본 논문은 ViT의 CLS 토큰 경로에 특화된 Residual LoRA 기반 PEFT를 제안하여 합성-실사 도메인 갭을 특징 공간 사후 보정으로 완화하였다. DINOv2 백본을 동결한 채 약 1% 수준의 파라미터만 학습하여 10-shot 실사 환경에서도 성능 향상과 시맨틱 정렬을 확인하였다. 향후 다양한 도메인 전이 데이터셋으로 일반성을 추가 검증하고, 객체 검출 및 의미 분할 등 다양한 다운스트림 태스크로 확장하여 실환경 인지 문제에서의 범용성을 실증할 예정이다.

참 고 문 헌 (References)

[1] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial Discriminative Domain Adaptation,” Proc. CVPR, pp. 7167-7176, 2017.
doi: https://doi.org/10.1109/CVPR.2017.316

[2] Y. Ganin et al., “Domain-Adversarial Training of Neural Networks,” J. Mach. Learn. Res., vol. 17, no. 59, pp. 1–35, 2016, https://jmlr.org/papers/v17/15-239.html

[3] M. Oquab et al., “DINOv2: Learning Robust Visual Features without Supervision,” Trans. Mach. Learn. Res. (TMLR), 2023, https://openreview.net/forum?id=a68SUt6zFt

[4] E. J. Hu et al., “LoRA: Low-Rank Adaptation for Large Language Models,” ICLR, 2022, https://openreview.net/forum?id=nuztlsLxeJb

[5] A. Dosovitskiy et al., “An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale,” ICLR, 2021, https://openreview.net/forum?id=YicbFdNTTy

[6] F. Wang et al., “Cosine Classifiers Are Great Domain Generalizers,” arXiv:2201.07185, 2022, https://arxiv.org/abs/2201.07185
doi: https://doi.org/10.48550/arXiv.2201.07185

[7] T. Chen et al., “A Simple Framework for Contrastive Learning of Visual Representations,” Proc. ICML, pp. 1599-1608, 2020, https://proceedings.mlr.press/v119/chen20j.html

[8] S. Xie et al., “A Survey on Diffusion Models in Vision,” IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI), 2023. https://ieeexplore.ieee.org/document/10081412
doi: https://doi.org/10.1109/TPAMI.2023.3261908

[9] N. Houlsby et al., “Parameter-Efficient Transfer Learning for NLP,” Proc. ICML, pp. 2790-2799, 2019, https://proceedings.mlr.press/v97/houlsby19a.html