CLIP meets DINO for Tuning Zero-Shot Classifier using Unlabeled Image Collections

Mohamed Fazli[1], Rufael Fekadu[1], Jameel Hassan[1][2], Mustansar Fiaz[3], Alham Fikri[1], Hisham Cholakkal[1]

[1] MBZUAI [2] John Hopkins University [3] IBM Research

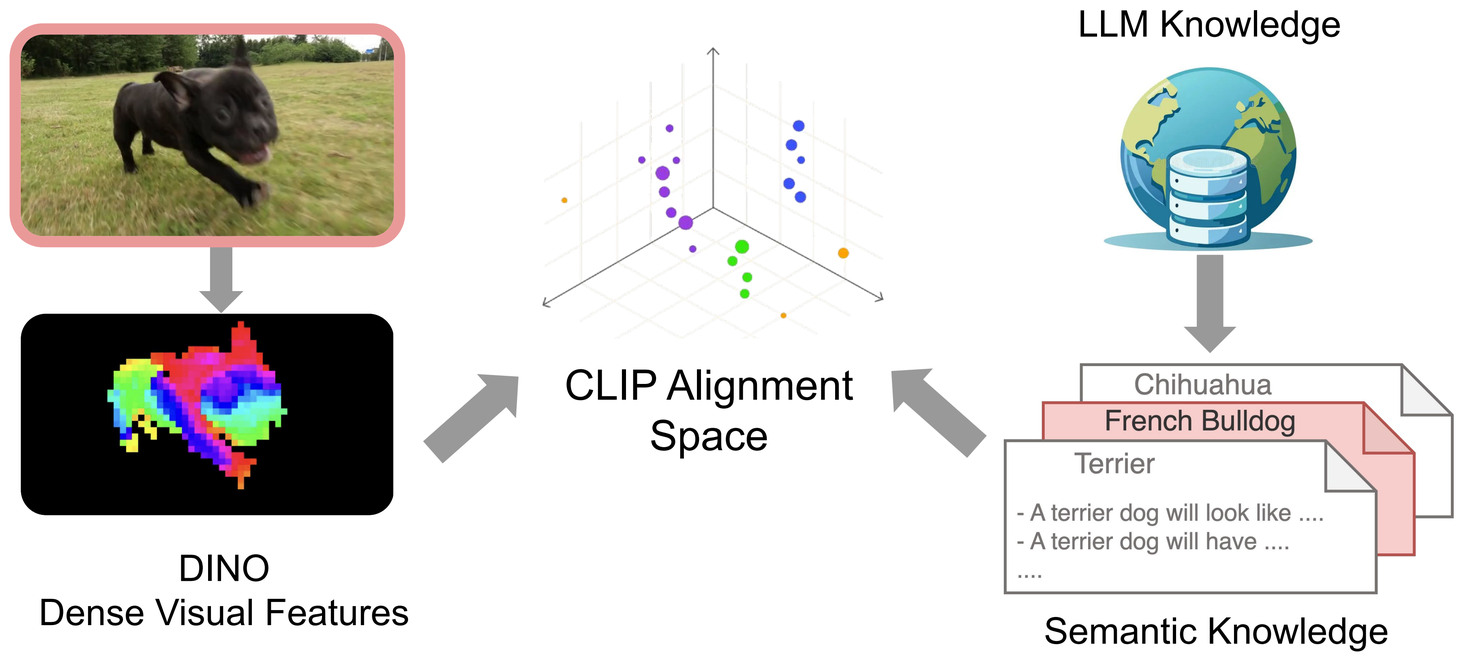
Introduction

Motivation

- Zero-shot CLIP features are sub-optimal.
- DINO visuals are rich but SSL methods usually require labeled linear probes; labels are costly.

Research Question

Can we efficiently fine-tune CLIP using unlabeled images and class-specific visual descriptions from LLMs to achieve label-free, dataset-adaptive classification?



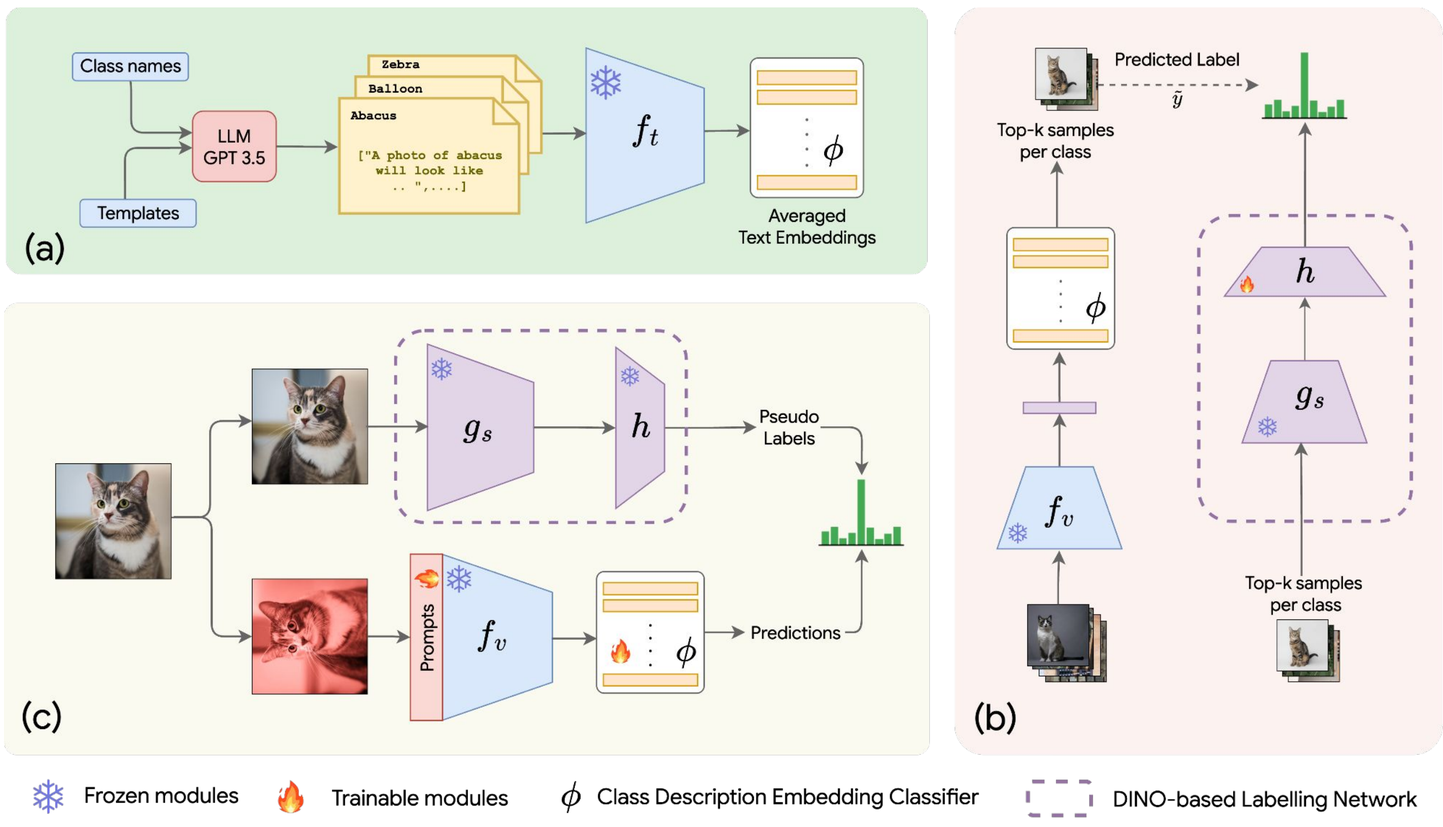
Methodology

We propose **NoLA** (No Labels Attached), a lightweight, label-free prompt-tuning framework that adapts CLIP using only unlabeled images. Our approach unfolds in three steps:

A. The class-specific visual descriptions is used to distil rich textual priors into CLIP's text space.

B. The SSL DINO is aligned with CLIP using top-k confident predictions per class.

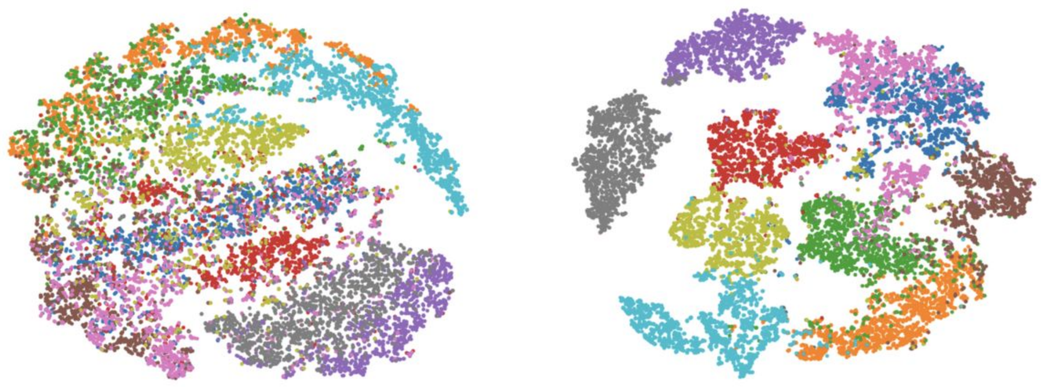
C. The aligned DINO backbone is used to generate confident pseudo labels to finetune the CLIP vision encoder via prompt tuning.



Quantitative Results

	ImageNet	EuroSAT	Caltech	Pets	UCF101	DTD	Flowers	SUN397	RESISC45	CFAR10	CFAR100	Average
Few-Shot Methods												
CoOp (5-Shot)	61.3	71.8	93.2	-	74.3	41.1	85.8	67.3	-	86.6	63.2	-
CoOp (10-Shot)	62.3	81.6	94.6	-	77.2	65.8	92.1	69.0	-	88.5	66.6	-
Label-Free Methods												
CLIP	61.9	40.6	90.5	85.0	61.0	42.9	66.6	60.8	49.8	88.8	64.2	64.7
CuPL	63.4	62.2	90.6	87.2	63.9	48.0	71.5	66.0	61.9	89.2	65.8	70.0
MetaPrompt	**65.0**	55.6	92.9	88.1	67.9	50.8	73.9	**67.0**	64.0	89.9	66.3	71.0
LaFTer	64.2	**73.9**	93.3	82.7	**68.2**	46.1	71.0	64.5	**68.3**	**95.8**	**74.6**	**73.0**
ProText	64.9	51.4	**93.4**	**89.0**	66.4	50.7	**74.2**	**66.8**	57.4	89.5	66.1	70.0
WaffleCLIP	63.5	46.7	**94.8**	88.1	65.8	**51.0**	68.7	65.6	63.4	90.9	67.2	69.6
NoLA (Ours)	**65.4**	**73.5**	**94.8**	**89.3**	**68.3**	**56.1**	**82.7**	**67.0**	**75.4**	**94.9**	**75.6**	**76.6**

Qualitative Results



Base CLIP Embeddings

NoLA Adapted CLIP Embeddings

Ablations

	EuroSAT	Caltech	Pets	Flowers	SUN397	CIFAR100	Average
AdaptCLIPZS	81.8	95.3	93.7	81.3	72.1	73.9	83.0
LaFTer	72.1	94.3	81.5	65.9	65.9	76.3	76.0
NoLA (Ours)	**79.1**	**95.3**	**91.7**	**84.3**	**69.3**	**77.5**	**82.9**

SSL	EuroSAT	Caltech101	Pets	Flowers	SUN397	CIFAR100	Average
CLIP ZS	42.3	93.3	55.1	67.5	62.5	68.4	64.9
MAE	**74.1**	93.0	85.0	76.8	63.0	67.7	76.6
BeiT	72.2	91.8	85.2	78.0	64.1	68.2	74.9
DINO (NoLA)	73.5	**94.8**	**89.3**	**82.7**	**67.0**	**75.6**	**80.5**

Trainable components	Prompts	CDE	Avg. Top-1 Acc. (%)
Setting 1	✗	✗	77.9
Setting 2	✓	✗	78.2
Setting 3	✗	✓	79.6
NoLA	✓	✓	**80.5**

	Avg. Top-1 Acc. (%)
CLIP zero-shot	67.9
(+) CDE classifier	72.0
(+) DL network	73.9
(+) DINO-assisted Prompt Learning	**80.5**

Conclusion

- NoLA delivers label-free, parameter-efficient adaptation using unlabeled images and class-specific visual descriptions.
- A scalable, cost-effective path for adapting VLMs when annotations are scarce.

References

[1] 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 ICML, PMLR, 2021.

[2] 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 ICCV, 2021.

[3] Muhammad Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Horst Possegger, Mateusz Kozinski, Rogerio Feris, and Horst Bischof. ***Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections***. Advances in NeurIPS, 2024

[4] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. ***What does a platypus look like? generating customized prompts for zero-shot image classification.*** In Proceedings of the IEEE/CVF ICCV, 2023.