UNIC: Universal Classification Models via Multi-teacher Distillation

Mert Bulent SARIYILDIZ Philippe WEINZAEPFEL Thomas LUCAS Diane LARLUS Yannis KALANTIDIS

NAVER LABS
Europe

https://github.com/naver/unic

Goal

Learn a **single** encoder from **strong** and **complementary** teachers

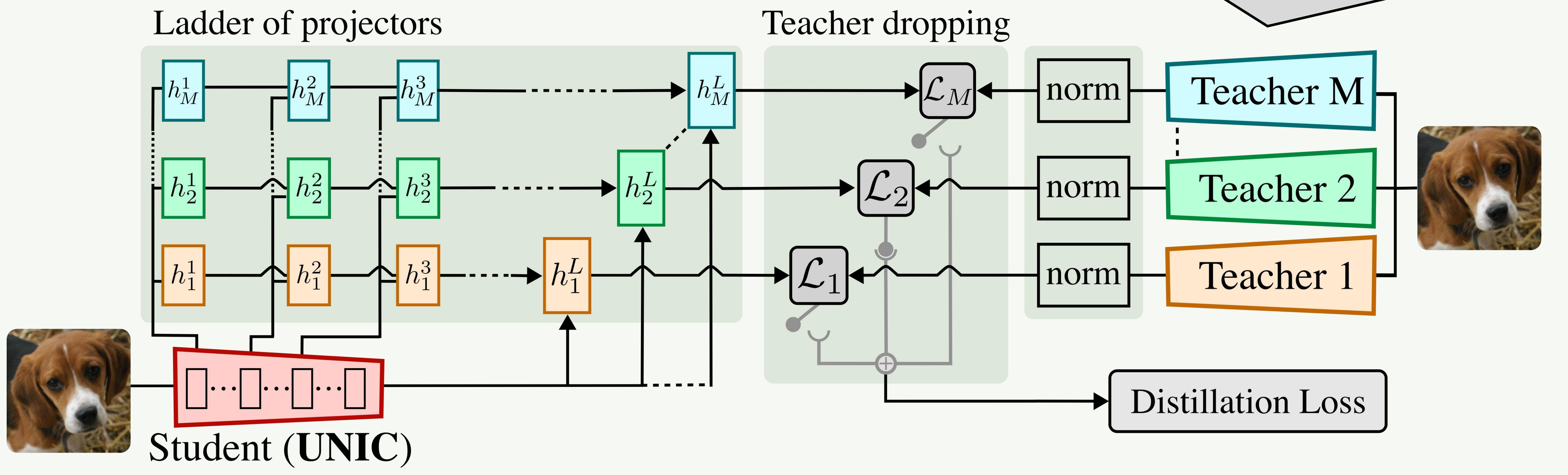
Approach

Multi-teacher distillation

3 components

- Feature normalization
- Ladder of projectors
- Teacher dropping regularization

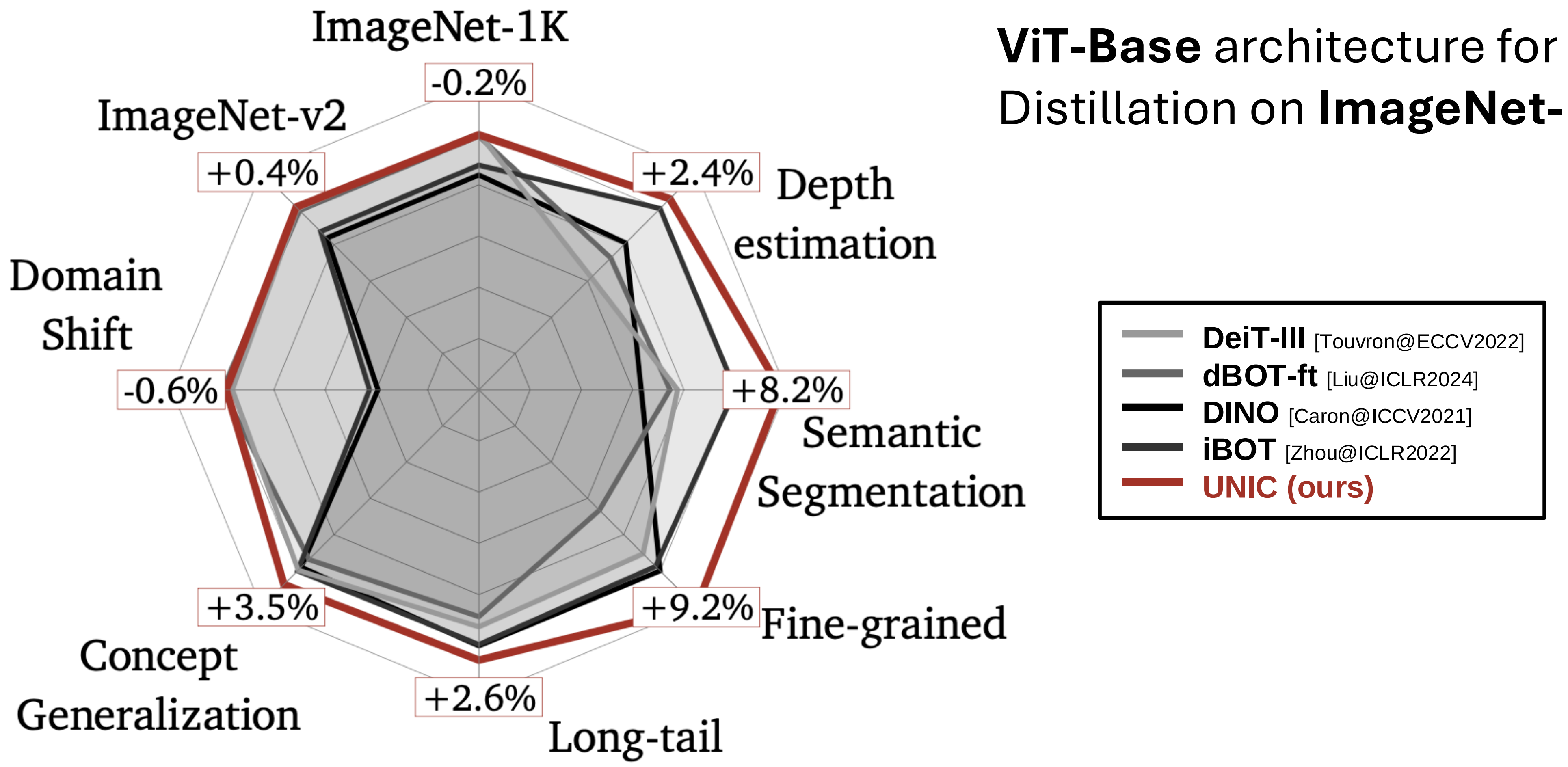
Diverse training objectives		**Complementary strengths**
Supervised learning	→	Classifying training concepts
Self-supervised learning	→	Transfer learning
Masked image modeling	→	Dense prediction



Universal encoder: A **single model** that combines the teacher's **strenghts**

Teachers trained on ImageNet-1K

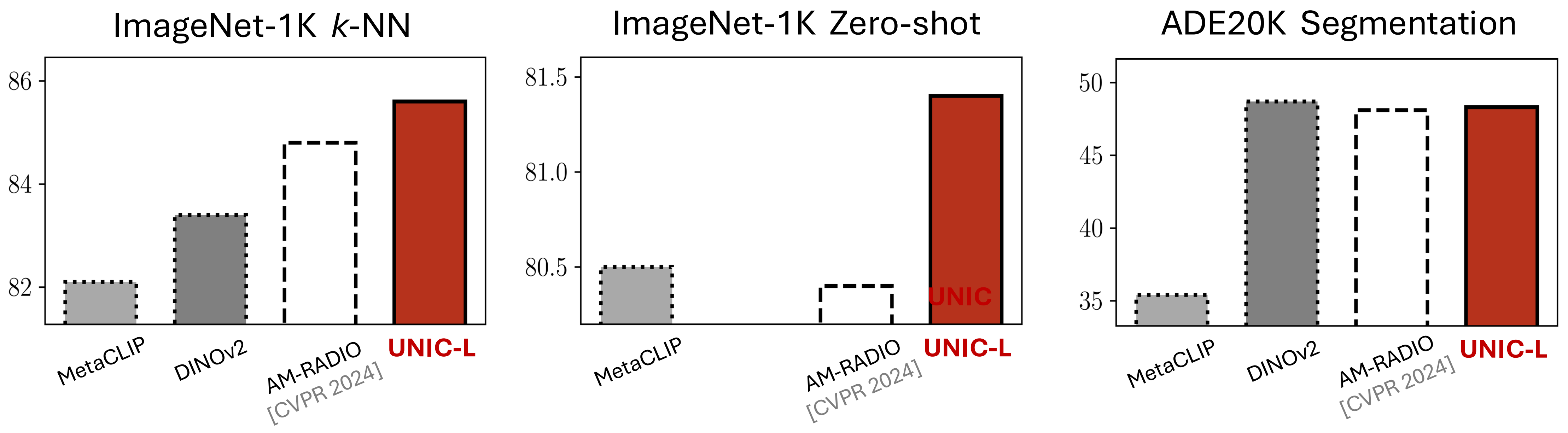
ViT-Base architecture for all
Distillation on **ImageNet-1K**



✓ **UNIC** is **on par or better** than **the best teacher** for each task

Comparison to the state of the art

Teachers: **MetaCLIP-Huge** and **DINOv2-Giant**, Student: **ViT-Large**
Distillation on **ImageNet-1K**



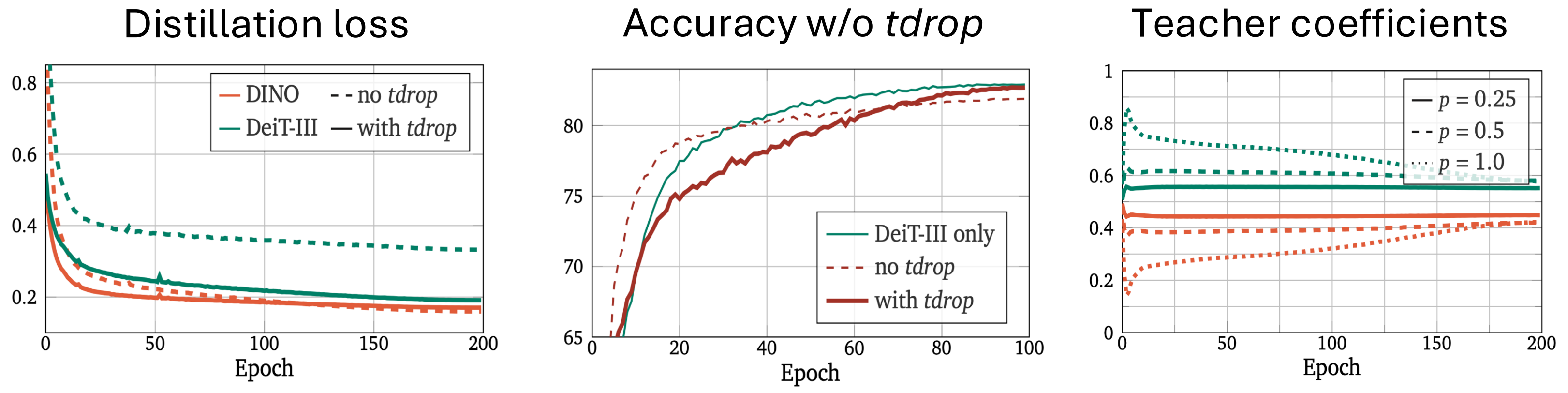
✓ **UNIC-L outperforms the best teacher** in the vast majority of cases

Impact of components

Feature normalization (*norm*) Ladder of projectors (*LP*) Teacher dropping (*tdrop*)

Model	norm	LP	*tdrop*	IN-val top-1 (↑)	Transfer top-1 (↑)	Segmentation mIoU (↑)	Depth RMSE (↓)
Teacher models							
DINO				77.7	72.4	30.4	0.570
DeiT-III				83.6	68.5	32.3	0.589
best teacher				83.6	72.4	32.3	0.570
Multi-teacher distillation (DINO & DeiT-III teachers)							
basic setup				78.7	73.1	33.9	0.560
	✓			81.4	73.8	36.1	0.558
UNIC	✓	✓		82.7	74.2	37.4	0.546
	✓	✓	✓	83.2	73.5	37.3	0.547

Teacher dropping analysis



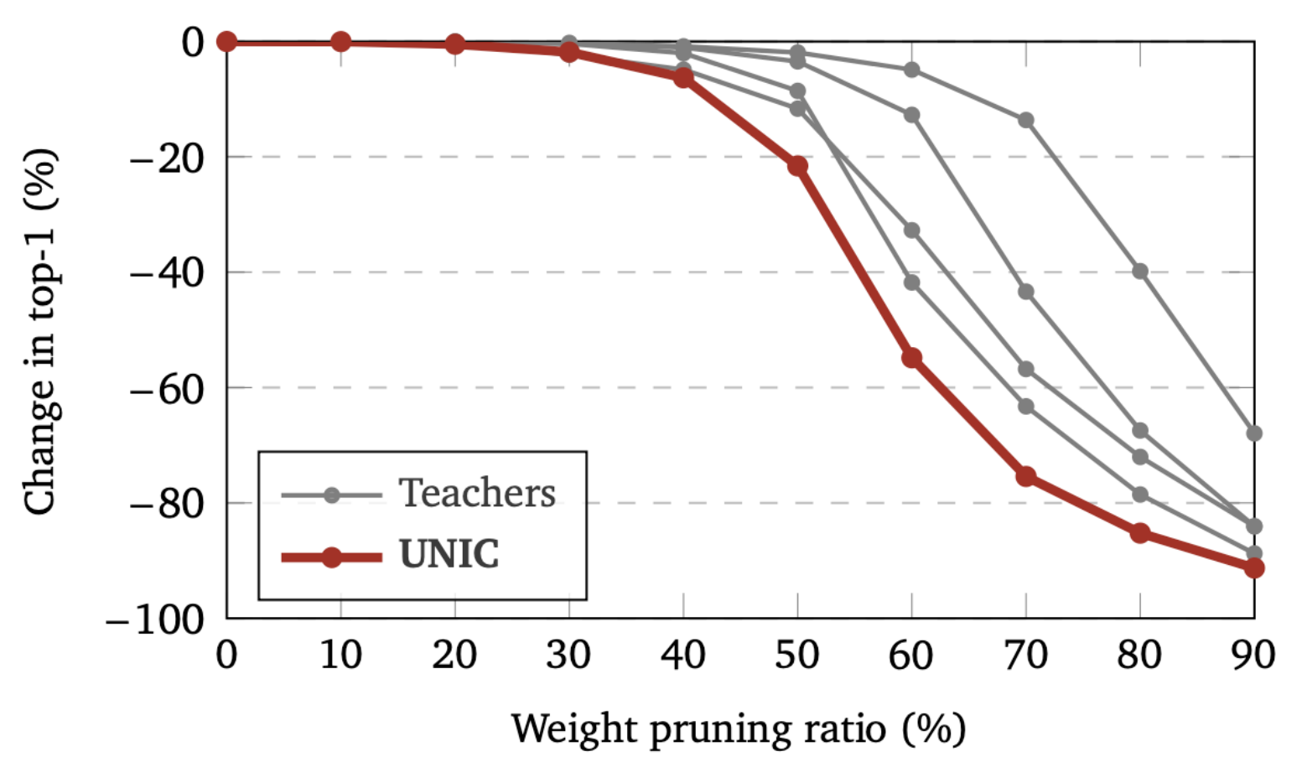
- ✓ Distillation loss is balanced across teachers
- ✓ More discriminative features are learned, albeit slower
- ✓ Teacher importance is dynamically determined

✓ Teacher dropping is effective at **learning all teachers equally well**

Model weight and feature space utilization analysis

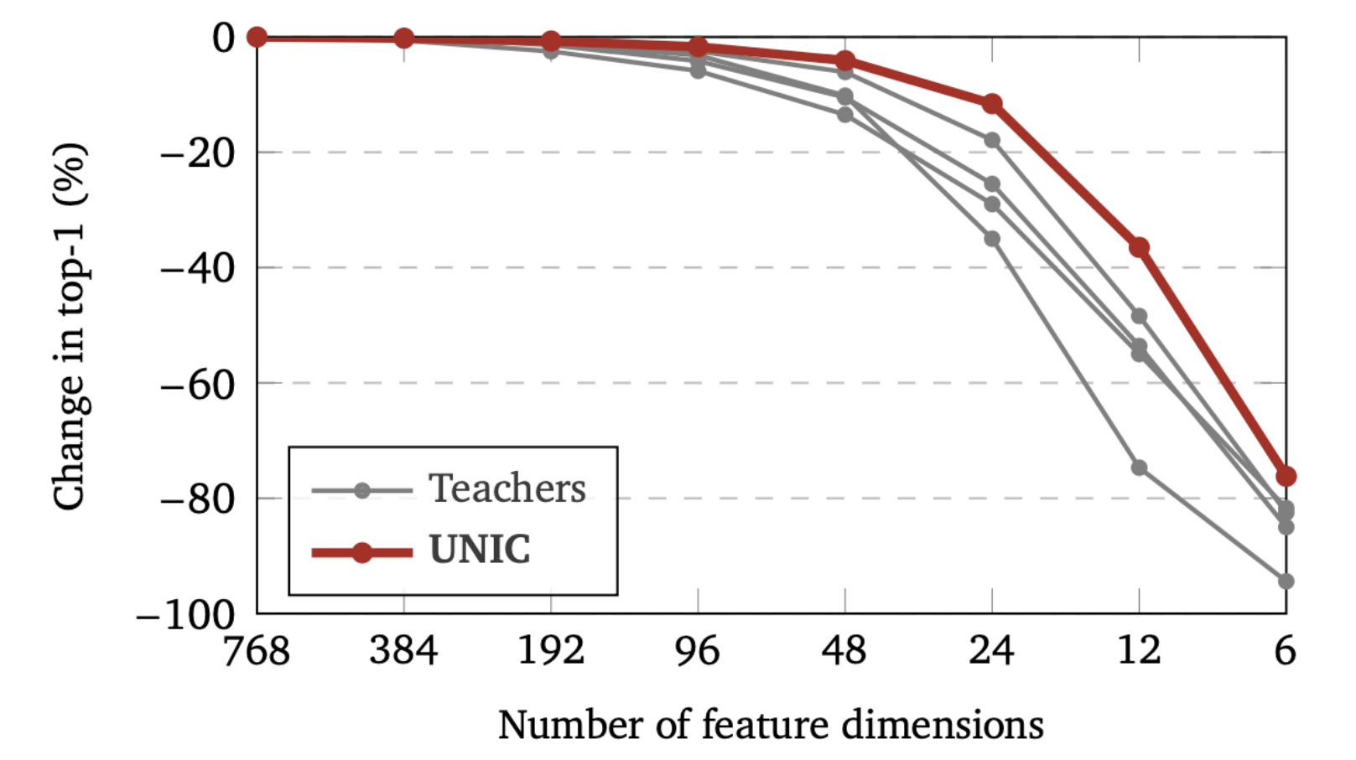
Weight utilization study

- Prune model weights
- Extract ImageNet-1K features
- Train a linear classifier



Feature space utilization study

- Extract ImageNet-1K features
- Apply PCA
- Train a linear classifier



✓ **UNIC** encoders **utilize model weights more effectively**

✓ **UNIC** features are **more resilient** to dimensionality reduction