KOO-FU CLIP: CLOSED-FORM ADAPTATION OF VISION–LANGUAGE MODELS VIA FUKUNAGA–KOONTZ LINEAR DISCRIMINANT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models such as CLIP provide general-purpose representations, but their embeddings are not explicitly optimized for class separation or compactness in downstream supervised tasks. We propose Koo-Fu CLIP, a supervised adaptation method based on **Fu**kunaga–**Koo**ntz Linear Discriminant Analysis. The method combines regularized within-class whitening with a rotation that orders projection directions by between-class scatter. The resulting closed-form transformation supports multiple embedding sizes through truncation, without iterative optimization.

Experiments with CLIP, SigLIP2, and DINOv3 on seven datasets show improvements over the original embeddings in prototype and nearest-neighbor classification across most evaluated settings. Dimensionality reduction often further improves nearest-neighbor accuracy while reducing storage requirements and distance-computation costs. For example, on ImageNet-1k, Koo-Fu CLIP raises CLIP 15-nearest-neighbor accuracy from 79.42% to 82.18% while reducing embedding dimensionality from 768 to 192. Although learned linear projections generally achieve higher accuracy, Koo-Fu CLIP obtains all output dimensionalities from a single fitted transformation.

1 INTRODUCTION

Vision-language models (VLMs) such as Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021) and Sigmoid Loss for Language-Image Pre-Training (SigLIP) (Zhai et al., 2023; Tschannen et al., 2025) learn image representations that support classification and retrieval. Vision-only models such as DINOv3 learn transferable representations without textual supervision (Siméoni et al., 2026). However, these embeddings are not necessarily optimized for class separation or compactness in a particular downstream task.

We propose **Koo-Fu CLIP**, a supervised adaptation of frozen foundation-model embeddings based on Fukunaga-Koontz Linear Discriminant Analysis (LDA). The method combines regularized within-class whitening with a rotation that orders directions by between-class scatter in the whitened space. For a fixed regularization parameter, the projection is computed in closed form from labeled training data, and its ordered directions allow dimensionality reduction without refitting.

Related approaches use Gaussian discriminant analysis for CLIP classification (Wang et al., 2024) and regularized discriminant projections for out-of-distribution detection (Chen et al., 2023); we study compact embeddings for visual classification. Like Matryoshka Representation Learning (Kusupati et al., 2022), Koo-Fu CLIP supports multiple embedding sizes, but derives ordered directions in closed form rather than training nested representations.

We evaluate Koo-Fu CLIP on CLIP, SigLIP2, and DINOv3 embeddings using nearest visual prototype (NVP) and k-nearest neighbors (k-NN) classification across seven datasets. The best reported configurations improve prototype accuracy in all 21 model–dataset combinations and full-dimensional 15-NN accuracy in 20 of 21 combinations. Dimensionality reduction further improves 15-NN accuracy in 19 of the 21 configurations with reported dimensionality results. Smaller embeddings reduce reference-database storage and distance-computation costs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

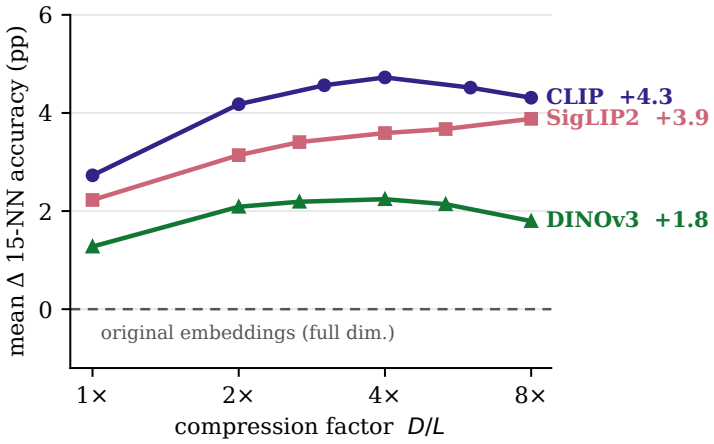

Figure 1: Mean change in 15-NN accuracy over the original embeddings across ImageNet-1k, ReImageNet, ImageNet-21k, ObjectNet, and FungiTastic-Mini. Each curve uses the source and regularization settings listed for each dataset in Tab. 2, held fixed as dimensions decrease. Compression factors use each encoder's actual embedding dimensions; the largest factor is eight for all three encoders.

Linear projections trained with contrastive loss, alone or combined with cross-entropy, generally achieve higher 15-NN accuracy on most comparison datasets, while Koo-Fu CLIP generally performs better on ImageNet-21k. These baselines require separate training for each output dimensionality, whereas Koo-Fu CLIP supports truncation of a single fitted transformation.

Our contributions are:

- A closed-form adaptation of frozen foundation-model embeddings using regularized Fukunaga-Koontz discriminative projections.
- Dimensionality reduction through ordered projection directions, supporting multiple embedding sizes from a single fitted transformation.
- An evaluation of prototype and nearest-neighbor classification across three encoders and seven datasets, including transformation transfer, regularization, and comparisons with learned linear projections.

2 Related work

Foundation models and visual classification. Vision-language models such as CLIP (Radford et al., 2021) learn aligned image and text representations for zero-shot recognition and transfer learning. SigLIP and SigLIP2 (Zhai et al., 2023; Tschannen et al., 2025) extend this approach using a sigmoid objective, while DINO models learn visual representations without text supervision (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2026). Our study examines supervised transformation and compression of these representations for nearest-prototype and nearest-neighbor classification.

CLIP adaptation. Linear probing learns a classifier on frozen features, while CoOp, CoCoOp, and inverse prompting optimize textual prompts (Zhou et al., 2022b;a; Parisot et al., 2023). Tip-Adapter constructs a feature cache without iterative training; its fine-tuned variant further optimizes the cache (Zhang et al., 2022). Feature adaptation methods modify the visual representation: CLIP-Adapter learns residual feature adapters (Gao et al., 2024), and ProLIP fine-tunes the visual projection matrix with regularization toward its pretrained weights (Fahes et al., 2026). CLIP-SVD fine-tunes the singular values of weight matrices in the image and text encoders while keeping their singular vectors fixed (Koleilat et al., 2026). Our method instead estimates a linear transformation of extracted embeddings in closed form from labeled-data statistics.

Compact CLIP representations. Matryoshka Representation Learning trains nested embeddings whose prefixes support different representation sizes (Kusupati et al., 2022). Adaptive Prior Refinement (APE) selects existing CLIP feature channels (Zhu et al., 2023), while InDiReCT learns a compact mapping from text embeddings for image retrieval (Kobs et al., 2023). Our method instead combines coordinates using scatter estimates from labeled images, requiring no text encoder and supporting truncation without retraining the backbone.

Discriminative projections and metric learning. Discriminative linear projections have been used to improve local descriptor matching and reduce dimensionality (Mikolajczyk & Matas, 2007; Cai et al., 2008). For deep image retrieval, Radenović et al. (2019) combine discriminative whitening and truncation using pair-based supervision, a close methodological predecessor to our class-based construction. Mensink et al. (2012) learn low-dimensional metrics for nearest-class-mean and nearest-neighbor classification on ImageNet. Their setting relates to our evaluation, but their optimized objectives differ from our scatter-based eigendecomposition.

Discriminant analysis of CLIP features. WDiscOOD (Chen et al., 2023) uses within-class whitening and regularized LDA on deep features, including CLIP features, for out-of-distribution detection. Our study shares this discriminative construction but evaluates prototype and nearest-neighbor classification under dimensionality reduction. Wang et al. (2024) estimate a CLIP classifier from class means and a shared covariance using Gaussian discriminant analysis, then combine its predictions with a textual classifier. Our contribution is the systematic evaluation of classical discriminative projections as compact representations of frozen foundation-model features.

3 METHOD

This section first summarizes the theoretical background of the Fukunaga-Koontz transform and then explains how it can be applied to CLIP and other pretrained image encoders.

3.1 PRELIMINARIES

Linear Discriminant Analysis Linear Discriminant Analysis (LDA) is a supervised statistical method used to reduce the dimensionality of a feature space while preserving its discriminative capabilities (Fisher, 1936). Its advantages include simplicity and efficiency, since applying the fitted projection does not require access to the training data, making it applicable even for learning from datasets with millions of data points.

LDA finds a linear mapping from the original feature space to a lower-dimensional subspace in which the class-specific mean vectors are maximally separated, while the variance (scatter) of samples within each class is minimized. Formally, given a labeled dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with samples $\mathbf{x}_i \in \mathbb{R}^D$ and class labels $y_i \in \{1, \dots, K\}$, let

$$\boldsymbol{\mu} = \frac{1}{N} \sum_i \mathbf{x}_i, \qquad \boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i:y_i=k} \mathbf{x}_i \tag{1}$$

be the global data mean vector and the mean vector of class k, respectively, where N_k is the number of samples in class k,

$$\mathbf{S}_k = \sum_{i:y_i=k} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \tag{2}$$

the conditional scatter matrix of class k, and

$$\mathbf{S}_w = \sum_{k=1}^{K} \mathbf{S}_k, \qquad \mathbf{S}_b = \sum_{k=1}^{K} N_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^\top, \tag{3}$$

the within-class and between-class scatter matrices. LDA aims to find a projection matrix $\mathbf{W} \in \mathbb{R}^{D \times L}$ such that the objective

$$\frac{\det(\mathbf{W}^\top \mathbf{S}_b \mathbf{W})}{\det(\mathbf{W}^\top \mathbf{S}_w \mathbf{W})} \tag{4}$$

is maximal. The solution can be found in closed form by solving the generalized eigenvalue problem

$$\mathbf{S}_b \mathbf{v} = \lambda \mathbf{S}_w \mathbf{v}. \tag{5}$$

Finally, to perform the mapping to a lower-dimensional space, the eigenvalues λ_j are sorted in descending order ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$) and the ultimate projection matrix, defining a mapping into an L-dimensional subspace, is formed by the top L eigenvectors $\mathbf{v}_j$ corresponding to the highest eigenvalues:

$$\mathbf{W}_L = [\mathbf{v}_1, \ldots, \mathbf{v}_L]. \tag{6}$$

Since $\text{rank}(\mathbf{S}_b) \leq \min(D, K-1)$, at most $\min(D, K-1)$ directions have nonzero discriminative eigenvalues (Fukunaga, 1990).

Fukunaga-Koontz Transform

The Fukunaga-Koontz transform is a linear feature transformation based on simultaneous diagonalization of two scatter matrices (Fukunaga & Koontz, 1970; Fukunaga, 1990). Applied to the within-class and between-class scatter matrices, this framework provides a constructive view of the LDA projection through a *whitening transformation* followed by a *discriminative rotation*. Whitening normalizes the pooled within-class scatter, and the rotation orders directions by their between-class scatter in the whitened space. For positive-definite within-class scatter, these two steps solve the LDA generalized eigenvalue problem.

The whitening begins with the eigendecomposition of the within-class scatter matrix:

$$\mathbf{S}_w = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top. \tag{7}$$

From this, we compute the inverse square root matrix

$$\mathbf{S}_w^{-1/2} = \mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{V}^\top, \tag{8}$$

where $\mathbf{\Lambda}^{-1/2}$ is the diagonal matrix of the inverse square roots of the eigenvalues ($\sqrt{\frac{1}{\lambda_j}}$). Satisfying

$$\mathbf{S}_w^{-1/2}\mathbf{S}_w\mathbf{S}_w^{-1/2} = \mathbf{I}, \tag{9}$$

the matrix $\mathbf{S}_w^{-1/2}$ makes the pooled within-class scatter isotropic.

Computing the inverse square root requires positive eigenvalues, but floating-point errors can introduce small negative eigenvalues into the empirical within-class scatter. Following prior work (Friedman, 1989; Ledoit & Wolf, 2004; Chen et al., 2023; Wang et al., 2024), we apply regularization, which, as we demonstrate, can also improve classification accuracy:

$$\mathbf{S}'_w = \mathbf{S}_w + \lambda\mathbf{I}, \tag{10}$$

$$\mathbf{Z} := (\mathbf{S}'_w)^{-1/2}, \tag{11}$$

with λ becoming the only hyperparameter of the proposed method.

For convenience, we define the whitening function $f(\cdot)$ as the projection of a centered vector onto the sphered space:

$$f(\mathbf{x}) = \mathbf{Z}\mathbf{x}. \tag{12}$$

Then, the between-class scatter matrix (Eq. (3)) is computed in the whitened space, which is defined as the range of the matrix $\mathbf{S}_w^{-1/2}$. This requires mapping the global and per-class means to the whitened space:

$$\mathbf{S}'_b = \sum_{k=1}^{K} N_k f(\boldsymbol{\mu}_k - \boldsymbol{\mu}) f(\boldsymbol{\mu}_k - \boldsymbol{\mu})^\top. \tag{13}$$

Next, we perform the eigendecomposition $\mathbf{S}'_b = \mathbf{U}\mathbf{\Gamma}\mathbf{U}^\top$, where the columns of $\mathbf{U}$ represent the directions of maximum class separability. Similarly to LDA, dimensionality reduction is performed by sorting the eigenvalues of the whitened between-class scatter matrix $\mathbf{S}'_b$ ($\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_D$) and truncating the rotation matrix $\mathbf{U}$ to keep only the eigenvectors corresponding to the top L eigenvalues:

$$\mathbf{U}_L = [\mathbf{u}_1, \ldots, \mathbf{u}_L]. \tag{14}$$

The ultimate projection matrix $\mathbf{T} \in \mathbb{R}^{L \times D}$ is obtained by composing the truncated rotation with the whitening transformation:

$$\mathbf{T} = \mathbf{U}_L^\top \mathbf{Z}. \tag{15}$$

This projection maps the original features into an optimized L-dimensional discriminative subspace.

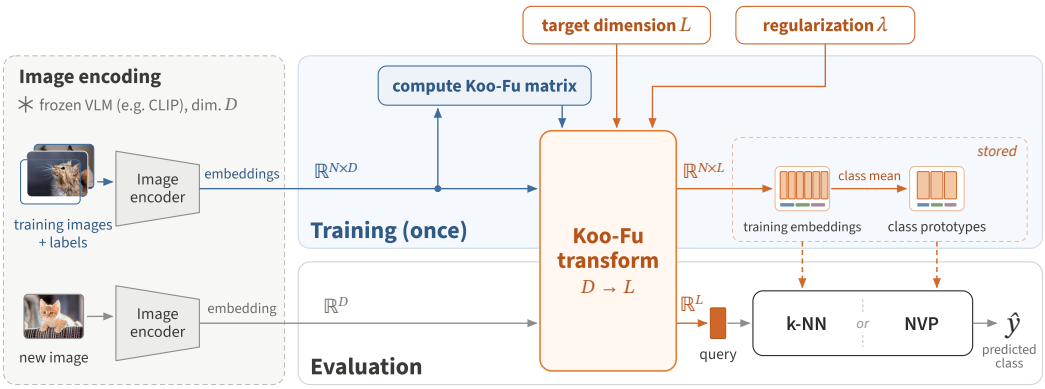

Figure 2: Overview of Koo-Fu CLIP. A frozen image encoder produces D-dimensional embeddings. Labeled training embeddings determine the projection $\mathbf{T}$, which maps both training and query embeddings to L dimensions. Classification uses the transformed training embeddings for nearest-neighbor prediction or their class means for nearest-prototype prediction.

3.2 Koo-Fu CLIP

Koo-Fu CLIP adapts pretrained image embeddings to a downstream classification task using the projection derived in Sec. 3.1. Figure 2 illustrates the training and inference pipeline.

Let e denote an image encoder and $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ the labeled training embeddings, where $\mathbf{x}_i = e(I_i) \in \mathbb{R}^D$ for training image I_i. We write $c(\mathbf{z}; \mathcal{A})$ for a general classifier that predicts a class from an embedding $\mathbf{z}$ using a labeled reference set $\mathcal{A}$ in the same feature space. For example, c may use nearest neighbors or class prototypes. For a query image I, classification with the original embeddings is therefore

$$\hat{y}_{\text{raw}} = c(e(I); \mathcal{D}). \tag{16}$$

To construct our representation, we estimate $\mathbf{S}_w$ and $\mathbf{S}_b$ from $\mathcal{D}$ and compute the regularized whitening matrix $\mathbf{Z} = (\mathbf{S}_w + \lambda \mathbf{I})^{-1/2}$. We then obtain $\mathbf{U}_L$ from the whitened between-class scatter in Eq. (13) and form $\mathbf{T} = \mathbf{U}_L^\top \mathbf{Z}$ as in Eq. (15). Applying this projection to the training embeddings gives

$$\mathcal{D}_{\mathbf{T}} = \{(\mathbf{T}\mathbf{x}_i, y_i)\}_{i=1}^N. \tag{17}$$

The same classifier then operates on the transformed query and reference embeddings:

$$\hat{y}_{\text{ours}} = c(\mathbf{T}e(I); \mathcal{D}_{\mathbf{T}}). \tag{18}$$

Thus, Koo-Fu CLIP changes the representation used by the classifier while retaining the pretrained image encoder.

4 Experiments

4.1 Setup

Models

We evaluate three pretrained image encoders across all tasks.
OpenAI's CLIP ViT-L/14 (Radford et al., 2021) produces 768-dimensional embeddings.
SigLIP2 (`hf-hub:timm/ViT-L-16-SigLIP2-256`) (Tschannen et al., 2025) and DINOv3 (`hf-hub:timm/vit_large_patch16_dinov3.lvd1689m`) (Siméoni et al., 2026) both produce 1024-dimensional embeddings. We use OpenCLIP (Ilharco et al., 2021) and PyTorch Image Models (timm) (Wightman, 2019) for the pretrained image encoders.

Datasets

We evaluate on seven datasets covering general object recognition, robustness to distribution shifts, and fine-grained fungal recognition.

ImageNet-1k (Russakovsky et al., 2015) contains over 1.28 million training images and 50,000 validation images across 1,000 object classes defined by WordNet (Miller, 1995). We use the standard training and validation splits.
ReImageNet (Kisel et al., 2025; Volkov et al., 2026) is a subset of the ImageNet-1k validation split with corrected labels. We use the August 2026 release, which excludes 2,016 images without reliable labels, leaving 47,984 validation images. The standard ImageNet-1k training split supplies the data for dataset-specific transformation fitting, prototypes, and reference neighbors. This exclusion and its more permissive metric (see below) may inflate performance.
ImageNet-21k (Ridnik et al., 2021) contains 13,153,480 images across 19,167 WordNet classes in the Winter 2021 release used here. It is the larger dataset from which ImageNet-1k was sampled.

ImageNet-A (Hendrycks et al., 2021b) contains approximately 7,500 natural, unedited images across 200 classes, selected to elicit high-confidence misclassifications from standard vision models. It evaluates robustness to natural adversarial examples.
ImageNet-R (Hendrycks et al., 2021a) contains approximately 30,000 artistic renditions of 200 ImageNet classes, including paintings, cartoons, and sculptures. It evaluates generalization across visual styles.
ObjectNet (Barbu et al., 2019) contains 50,273 images across 313 object classes, including 113 shared with ImageNet-1k. Controlled variations in backgrounds, rotations, and viewpoints test robustness to atypical object contexts.

FungiTastic-Mini (Picek et al., 2025) provides 46,842 training images across 215 classes and 9,412 validation images covering 193 of these classes in our evaluation. Derived from expert-verified fungal records, it evaluates fine-grained recognition.

Classifiers

k-nearest neighbors (*k*-NN) classification assigns the most common label among the k nearest training embeddings, retrieved using FAISS (Douze et al., 2026). Nearest visual prototype (NVP) classification assigns the label of the nearest class-mean embedding, requiring storage of only one vector per class. Before storage, all embeddings are normalized, and vector similarity is measured using cosine similarity.

Metrics

Our primary metric is *accuracy*, that is, the fraction of evaluated samples classified correctly. For ReImageNet, we adopt the ReaL accuracy (Beyer et al., 2020; Kisel et al., 2025): a prediction is correct if its class matches any accepted ground-truth label.

Evaluation

For each encoder and dataset, we learn the proposed transformation using the available training data and select the regularization parameter λ using associated validation data. We use the same regularization search grid for all encoders, datasets, classifiers, and transformation sources:

$$\lambda \in \{0, 50, 100, 150, 250, 500, 750, 1000, 1500, 2000, 5000\}.$$

For ImageNet-A, ImageNet-R, and ObjectNet, we construct approximately 50/50 training and validation splits while ensuring that all classes are represented appropriately. For ImageNet-21k, we randomly select 10% of the samples in each class, capped at 50 per class, for validation, resulting in approximately a 95:5 training-to-validation ratio. For ReImageNet, we use the standard ImageNet-1k training split and the reduced validation set with corrected labels. Splits are not changed across experiments. We consider dataset-specific transformations, learned on the training split of the evaluation dataset, and transferred transformations, learned on ImageNet-1k or ImageNet-21k. In both settings, we evaluate on the validation split of the target dataset.

We report the best evaluated transformation settings separately for the prototype-based classifier (NVP) and for *k*-NN in Tabs. 3 and 4. For the dimensionality sweeps, we hold the source and λ listed in Tabs. 1 and 2 fixed. We fix k=15 based on preliminary experiments, ranging from 1 to 20, which produced results close to the optimum for most of the datasets.

For both NVP and 15-NN, we evaluate the following output dimensionalities L, including the original embedding size:

$$\begin{aligned} \text{CLIP:} \quad & \{768, 640, 512, 384, 256, 192, 128, 96, 76, 64, 32\}, \\ \text{SigLIP2, DINOv3:} \quad & \{1024, 768, 640, 512, 384, 256, 192, 128, 102, 80, 40\}. \end{aligned}$$

Table 1: NVP accuracy (%) with Koo-Fu adaptation and compression. Training sources: IN-1k = ImageNet-1k, IN-21k = ImageNet-21k, same = evaluation dataset (ReImageNet uses ImageNet-1k training data). Original is unadapted; Full uses D; Reduced reports the best accuracy at $L < D$ (smallest L for ties at reported precision). Δ: Reduced minus Full (percentage points). Final columns: minimum evaluated dimensions matching or exceeding Full/Original accuracy. Each sweep fixes the source and λ.

Model	Dataset	Source (λ)	Original	Full	Reduced @ L	Δ	$\geq$Full	$\geq$Orig.
CLIP (768D)	ImageNet-1k	IN-1k (100)	75.11	79.07	79.10 @ 512	+0.03	384	76
	ReImageNet	IN-1k (50)	79.66	83.81	83.86 @ 512	+0.05	512	76
	ImageNet-21k	IN-21k (2000)	32.03	34.32	34.32 @ 640	0.00	640	192
	ImageNet-A	same (0)	67.71	73.75	73.83 @ 256	+0.08	192	64
	ImageNet-R	same (0)	84.88	89.29	89.41 @ 192	+0.12	192	64
	ObjectNet	same (0)	59.97	65.61	65.65 @ 512	+0.04	384	64
	FungiTastic-Mini	same (0)	33.16	55.25	55.40 @ 384	+0.15	192	32
SigLIP2 (1024D)	ImageNet-1k	IN-1k (150)	81.61	84.11	84.10 @ 768	−0.01	1024	102
	ReImageNet	IN-1k (100)	86.07	88.63	88.71 @ 384	+0.08	384	80
	ImageNet-21k	IN-21k (1500)	37.24	39.19	39.07 @ 768	−0.12	1024	256
	ImageNet-A	IN-1k (150)	73.27	78.62	79.87 @ 384	+1.25	192	80
	ImageNet-R	same (50)	93.24	95.17	95.19 @ 256	+0.02	256	80
	ObjectNet	same (50)	69.87	74.00	74.10 @ 384	+0.10	256	80
	FungiTastic-Mini	same (0)	42.96	58.64	58.86 @ 102	+0.22	102	40
DINOv3 (1024D)	ImageNet-1k	IN-1k (50)	82.73	83.44	83.35 @ 768	−0.09	1024	256
	ReImageNet	IN-1k (50)	87.62	88.30	88.24 @ 768	−0.06	1024	256
	ImageNet-21k	IN-21k (100)	36.54	36.88	36.48 @ 768	−0.40	1024	1024
	ImageNet-A	same (50)	84.46	85.81	86.15 @ 128	+0.34	128	102
	ImageNet-R	same (150)	92.67	93.27	93.27 @ 256	0.00	256	192
	ObjectNet	same (50)	71.74	74.32	74.47 @ 384	+0.15	256	128
	FungiTastic-Mini	same (0)	45.35	70.27	70.37 @ 192	+0.10	192	40

We also evaluate learned linear projections using the same training data and 15-NN protocol, comparing methods at the subset of output dimensionalities reported in Tab. 5.

Learned linear baselines

As an alternative to the scatter-based construction of $\mathbf{T}$ in Eq. (15), we directly optimize a bias-free linear projection for each output dimensionality. The first baseline uses the supervised contrastive loss SupCon (Khosla et al., 2020) with temperature 0.1. The second adds a linear classification layer with softmax and jointly minimizes SupCon and cross-entropy with equal weights. Both baselines normalize the projected embeddings, and only these representations are used for 15-NN evaluation. All baseline models are randomly initialized and optimized with stochastic gradient descent for 120 epochs, using a batch size of 4096 and a learning rate of 1.0 for the first 100 epochs and 0.1 for the final 20 epochs. We perform one training run per baseline configuration and loss function, without repeated random initializations.

4.2 Results

NVP classification

Full-dimensional Koo-Fu adaptation improves NVP accuracy in all 21 model–dataset combinations in Tab. 1. The largest gains occur on FungiTastic-Mini: 22.09, 15.68, and 24.92 percentage points for CLIP, SigLIP2, and DINOv3, respectively. Table 3 in Sec. A reports the complete comparison of transformation training sources and regularization settings. Dataset-specific transformations perform best on ImageNet-R, ObjectNet, and FungiTastic-Mini, while transferred transformations are preferable in some ImageNet-A settings. Regularization can mitigate accuracy losses: on ImageNet-A, it raises the accuracy of the dataset-specific DINOv3 transformation from 83.11% to 85.81%. However, all three models achieve their best FungiTastic-Mini prototype accuracy without regularization.

Table 2: 15-NN accuracy (%) with Koo-Fu adaptation and compression. Column definitions and tie-breaking follow Tab. 1. The source and λ are fixed within each sweep.

Model	Dataset	Source (λ)	Original	Full	Reduced @ L	Δ	$\geq$Full	$\geq$Orig.
CLIP (768D)	ImageNet-1k	IN-1k (500)	79.42	81.05	82.18 @ 192	+1.13	96	64
	ReImageNet	IN-1k (150)	82.41	84.24	85.75 @ 192	+1.51	96	64
	ImageNet-21k	IN-1k (500)	38.71	39.69	40.20 @ 384	+0.51	256	192
	ImageNet-A	IN-1k (500)	60.17	65.10	70.45 @ 192	+5.35	64	64
	ImageNet-R	same (50)	76.68	83.20	87.73 @ 128	+4.53	64	32
	ObjectNet	same (50)	56.08	60.51	64.00 @ 128	+3.49	64	32
	FungiTastic-Mini	same (50)	39.24	44.01	51.49 @ 32	+7.48	32	32
SigLIP2 (1024D)	ImageNet-1k	IN-1k (500)	84.10	85.12	85.89 @ 256	+0.77	102	80
	ReImageNet	IN-1k (500)	86.84	88.11	88.84 @ 192	+0.73	80	80
	ImageNet-21k	IN-1k (1000)	42.95	43.35	43.91 @ 384	+0.56	192	192
	ImageNet-A	IN-1k (500)	65.20	71.33	75.43 @ 256	+4.10	80	40
	ImageNet-R	same (50)	84.91	92.51	94.65 @ 192	+2.14	40	40
	ObjectNet	same (50)	65.97	70.24	73.24 @ 128	+3.00	80	40
	FungiTastic-Mini	same (50)	46.69	50.86	59.31 @ 40	+8.45	40	40
DINOv3 (1024D)	ImageNet-1k	IN-1k (50)	85.21	85.28	85.89 @ 384	+0.61	192	128
	ReImageNet	IN-1k (50)	88.29	88.45	89.07 @ 384	+0.62	128	128
	ImageNet-21k	IN-21k (50)	43.66	43.64	43.40 @ 768	−0.24	1024	n/a
	ImageNet-A	IN-1k (0)	79.39	81.76	81.08 @ 768	−0.68	1024	512
	ImageNet-R	same (50)	90.34	91.98	93.62 @ 192	+1.64	128	80
	ObjectNet	same (50)	72.00	73.18	75.14 @ 256	+1.96	102	80
	FungiTastic-Mini	same (50)	61.79	66.78	71.20 @ 102	+4.42	40	40

Dimensionality reduction provides limited additional NVP accuracy gains in most settings. In 17 of the 21 combinations in Tab. 1, the best reported reduced accuracy differs from Full by at most 0.20 percentage points. The largest increase is for SigLIP2 on ImageNet-A, from 78.62% to 79.87% at 384 dimensions. Compact representations can nevertheless retain full-dimensional accuracy: SigLIP2 on FungiTastic-Mini reaches 58.86% at 102 dimensions, compared with 58.64% at 1024 dimensions. For DINOv3 on ImageNet-1k, ReImageNet, and ImageNet-21k, every evaluated reduction lowers NVP accuracy; on ImageNet-21k, even the best reduced result (36.48%) falls below the original representation (36.54%).

k-NN classification

At full dimensionality, Koo-Fu improves 15-NN accuracy in 20 of 21 model–dataset combinations (Tab. 4 in Sec. A). The exception is DINOv3 on ImageNet-21k, with 43.64% accuracy versus 43.66% for the original embeddings. Regularization preferences differ between classifiers: on FungiTastic-Mini, unregularized transformations reduce 15-NN accuracy for every model, whereas $\lambda = 50$ yields gains of 4.77, 4.17, and 4.99 percentage points for CLIP, SigLIP2, and DINOv3.

Table 2 reports the transformation configurations used for dimensionality reduction alongside their accuracies. Dimensionality reduction further improves accuracy in 19 of the 21 combinations in Tab. 2. For DINOv3 on ImageNet-A and ImageNet-21k, the best reduced accuracies are 81.08% and 43.40% at 768 dimensions, below their respective full-dimensional accuracies of 81.76% and 43.64%. For CLIP on ImageNet-1k, reducing embeddings from 768 to 192 dimensions raises accuracy from 81.05% to 82.18%, while 64 dimensions retain at least the original accuracy of 79.42%. On FungiTastic-Mini, the best reduced representations reach 51.49%, 59.31%, and 71.20% for CLIP, SigLIP2, and DINOv3 at 32, 40, and 102 dimensions, respectively. Performance improvements on FungiTastic-Mini likely reflect its strong similarity to related categories, in which a substantial fraction of the variability present in the original embedding space may be non-informative for class discrimination. Koo-Fu suppresses such variation while retaining discriminative information and leveraging it further.

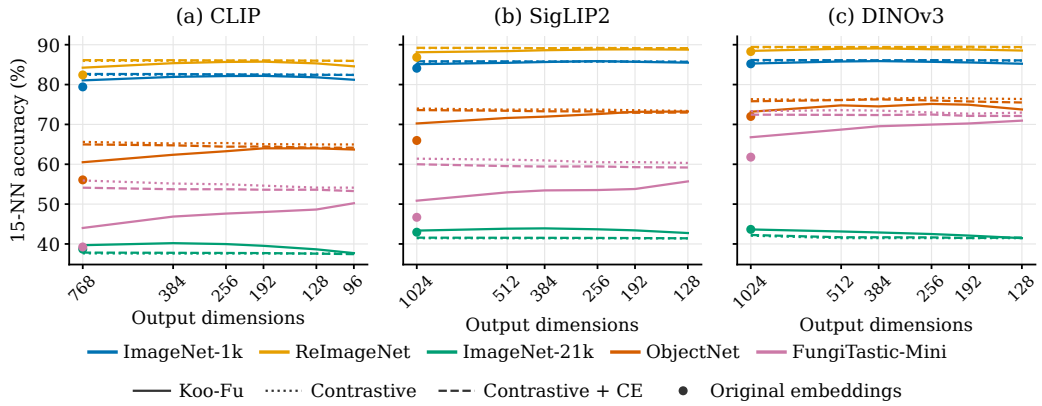

Figure 3: 15-NN accuracy versus output dimensionality for CLIP, SigLIP2, and DINOv3. Colors identify datasets; solid, dotted, and dashed curves denote Koo-Fu, contrastive linear projections, and linear projections trained with contrastive and cross-entropy (CE) losses, respectively. Circles indicate the accuracy of the original embeddings. Table 5 in the appendix reports the underlying numerical results.

Comparison with learned linear projections

Figure 3 shows that learned linear projections generally achieve higher 15-NN accuracy on ImageNet-1k, ReImageNet, ObjectNet, and FungiTastic-Mini, although dimensionality reduction often narrows the gap. For example, SigLIP2 on ImageNet-1k reaches 85.89% with Koo-Fu at 256 dimensions, compared with 85.79% and 85.81% for the two baselines. On ImageNet-21k, Koo-Fu outperforms both baselines at every evaluated dimensionality for CLIP and SigLIP2 and at all but the smallest dimensionality for DINOv3. Adding cross-entropy provides no consistent advantage over contrastive training alone. Although these results from a single run do not capture variability across random seeds or the impact of hyperparameter tuning, we note that the resulting curves remain stable across all tested combinations.

For a fixed λ, Koo-Fu supports all output dimensionalities through one closed-form fit, whereas the learned baselines require separate training for each size. Only λ requires selection through validation, and no optimizer hyperparameters need to be tuned.

5 Conclusion

We introduced Koo-Fu CLIP, a supervised method for adapting and compressing frozen image embeddings through regularized Fukunaga-Koontz discriminant analysis. For a fixed regularization parameter, a single closed-form transformation provides ordered projection directions that support multiple embedding sizes without refitting.

Across three encoders and seven datasets, the best evaluated transformations improve full-dimensional nearest-prototype accuracy in all 21 model–dataset combinations and 15-NN accuracy in 20 of 21. Dimensionality reduction further improves 15-NN accuracy in 19 combinations, whereas its additional gains for nearest-prototype classification are generally small. These results show that compact representations can improve classification accuracy, although the preferred transformation source, regularization, and embedding size depend on the task and classifier.

Learned linear projections generally achieve higher accuracy on most comparison datasets, while Koo-Fu CLIP generally performs better on ImageNet-21k. Its practical advantage is obtaining multiple representation sizes without iterative training or separate fitting for each size.

In its current form, Koo-Fu CLIP is not applicable to zero-shot classification: it requires labeled visual prototypes or reference samples and does not explicitly preserve image–text alignment. Extending the transformation to support textual class representations is a direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on research methodology and experiments, implementing some methods, and interpreting results. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, propose or refine hypotheses, or clean or reformat datasets. Formulating and proving mathematical claims, assistance with translation, writing of proofs and qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for drafting parts of the research paper, brainstorming and identifying relevant literature. LLM-generated code was verified by comparing its outputs to another implementation, and the generated interpretations were subjected to human review. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Andrei Barbu, David Mayo, Julian Alver, Alexander Luo, Christopher Katz, Dan Katz, Josef Baruch, Devon Atwood, Luya Cobos, Moshe Krymolowski, et al. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with ImageNet? *arXiv preprint arXiv:2006.07159*, 2020.

Hongping Cai, Krystian Mikolajczyk, and Jiří Matas. Learning linear discriminant projections for dimensionality reduction of image descriptors. In *Proceedings of the British Machine Vision Conference*, pp. 51.1–51.10. BMVA Press, 2008. doi: 10.5244/C.22.51.

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*, pp. 9650–9660, 2021.

Yiye Chen, Yunzhi Lin, Ruinian Xu, and Patricio A Vela. WDiscOOD: Out-of-distribution detection via whitened linear discriminant analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5298–5307, 2023.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library. *IEEE Transactions on Big Data*, 12(2):346–361, 2026. doi: 10.1109/TBDATA.2025.3618474.

Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. CLIP's visual embedding projector is a few-shot cornucopia. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 1936.

Jerome H Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 1989.

Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

Keinosuke Fukunaga and Warren LG Koontz. Application of the Karhunen-Loève expansion to feature selection and ordering. *IEEE Transactions on Computers*, C-19(4):311–318, 1970.

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. doi: 10.1007/s11263-023-01891-x.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. URL https://doi.org/10.5281/zenodo.5143773.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673. Curran Associates, Inc., 2020.

Nikita Kisel, Illia Volkov, Kateřina Hanzelková, Klára Janoušková, and Jiří Matas. Flaws of ImageNet, computer vision's favorite dataset. In *The Fourth Blogpost Track at ICLR 2025*, 2025. URL https://openreview.net/forum?id=c6igqDpJCC.

Konstantin Kobs, Michael Steininger, and Andreas Hotho. InDiReCT: Language-guided zero-shot deep metric learning for images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

Taha Koleilat, Hassan Rivaz, and Yiming Xiao. CLIP-SVD: Efficient and interpretable vision–language adaptation via singular values. *Transactions on Machine Learning Research*, 2026.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, et al. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30233–30249, 2022.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 2004.

Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *European Conference on Computer Vision*, 2012. doi: 10.1007/978-3-642-33709-3_35.

Krystian Mikolajczyk and Jiří Matas. Improving descriptors for fast tree matching by optimal linear projection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2007. doi: 10.1109/ICCV.2007.4408871.

George A Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Sarah Parisot, Yongxin Yang, and Steven McDonagh. Learning to name classes for vision and language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23477–23486, 2023.

Lukas Picek, Klara Janouskova, Vojtech Cermak, and Jiri Matas. FungiTastic: A multi-modal dataset and benchmark for image categorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2025.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2019. doi: 10.1109/TPAMI.2018.2846566.

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 *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. ImageNet-21K pretraining for the masses. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Illia Volkov, Nikita Kisel, Tetiana Mishkina, Klara Janouskova, and Jiri Matas. Doomed to reannotate, forever: The ImageNet story. *arXiv preprint arXiv:2608.13783*, 2026.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free CLIP-based adaptation. In *International Conference on Learning Representations*, 2024.

Ross Wightman. PyTorch image models, 2019. URL `https://github.com/rwightman/pytorch-image-models`.

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*, pp. 11975–11986, 2023.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *European Conference on Computer Vision*, pp. 493–510, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19832-8. doi: 10.1007/978-3-031-19833-5_29.

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*, 2022a.

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, July 2022b. ISSN 1573-1405. doi: 10.1007/s11263-022-01653-1.

Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot CLIP with adaptive prior refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2605–2615, 2023.