CRAFT: Contrastive Feature Alignment for Robust Distillation under Shift

Preprint. Under review.

Anonymous

Abstract

Compressed vision models deployed in the wild must remain reliable under distribution shifts, yet most knowledge distillation (KD) practices focus on clean in-distribution accuracy. Existing logit- and feature-based KD can inadvertently transfer spurious teacher cues that are unstable under shift, and contrastive KD variants are not consistently robust. This paper studies robustness-oriented distillation under corruption-based distribution shift and introduces CRAFT, a reliability-aware contrastive feature alignment framework for robust knowledge transfer. CRAFT aligns teacher–student features across clean and synthetically corrupted views using an InfoNCE-style objective, while a reliability score derived from teacher confidence and prediction consistency modulates both alignment strength and a de-alignment term that suppresses fragile teacher directions. We instantiate this framework on a CIFAR-like classification task and empirically compare empirical risk minimization (ERM), logit KD, attention-based feature KD, and contrastive representation distillation (CRD) in a single-run evaluation. Attention-based KD achieves the highest robust accuracy, improving the ERM baseline's robustness by about three percentage points, whereas a naïve CRD baseline substantially underperforms. These findings show that robustness under shift is highly sensitive to how features are aligned and motivate CRAFT's reliability-aware contrastive design as a principled extension of feature-level KD for robust knowledge distillation.

1 Introduction

Deployed vision models increasingly operate under conditions where the test distribution deviates from the training distribution due to sensor noise, environmental changes, style shifts, or downstream preprocessing artifacts. Standard convolutional and transformer-based architectures achieve strong in-distribution performance on benchmarks such as CIFAR and ImageNet, yet their accuracy can deteriorate markedly when exposed to common corruptions and domain shifts. Corruption benchmarks like CIFAR-10-C and ImageNet-C reveal that models trained purely with empirical risk minimization (ERM) on clean data may lose a large fraction of their accuracy under noise, blur, weather, or digital artifacts. This robustness gap is particularly consequential for resource-constrained deployments, where practitioners rely on smaller student models distilled from larger teachers and need those students to retain robustness properties despite aggressive compression.

Knowledge distillation has become a standard tool for compressing large models into smaller students by transferring information from a teacher's outputs or intermediate representations. Classical KD matches softened teacher logits, while subsequent work has explored feature-level supervision, attention transfer, and relational distillation to more faithfully mimic the teacher's internal representations. These approaches primarily target clean accuracy and parameter efficiency, often assuming that training and deployment distributions coincide. At the same time, robustness under distribution shift has emerged as a central challenge in deep learning [5], and several studies have begun to examine how KD behaves when evaluated under shift. Empirical findings are mixed: some

works report modest robustness gains from KD, while others observe degradation or no improvement, indicating that naively mimicking a teacher can propagate its vulnerabilities.

A central question is how to transfer robust invariances from a teacher to a student without copying spurious correlations that fail under shift. Large teachers, including self-supervised encoders such as DINOv2, often exhibit more robust and semantically meaningful features than smaller convolutional networks, yet they also encode idiosyncratic biases tied to pretraining data. Logit-based KD encourages students to imitate the teacher's decision boundaries everywhere, including regions where the teacher is fragile or miscalibrated. Feature-based KD, such as attention transfer, aligns intermediate activations but does not explicitly distinguish between robust invariances and nuisance-sensitive directions. In parallel, contrastive representation learning has demonstrated that carefully constructed positive and negative pairs can shape invariances in self-supervised and supervised settings, and contrastive feature alignment has proved effective in domain adaptation and cross-domain recognition [1]. However, contrastive KD has mostly been explored for in-distribution gains [4] rather than as an explicit mechanism for robustness under shift.

This paper introduces CRAFT (Contrastive Robust Alignment for distillation under Feature and distribution shifT), a framework that views robustness-aware knowledge distillation as contrastive feature alignment under distribution shift. The core idea is to align teacher–student representations across clean and corrupted views of each image, using a contrastive loss that treats cross-view teacher–student pairs for the same image as positives and features from other images as negatives. Building on InfoNCE-style objectives, this alignment shapes student invariances toward those of the teacher. To prevent overfitting to unstable teacher behavior, CRAFT incorporates a reliability score based on teacher confidence and prediction consistency across corruptions, and uses this score to modulate both the contrastive alignment and a de-alignment term that suppresses fragile teacher directions. In this way, CRAFT aims to selectively transfer invariances that are stable under shift while reducing reliance on spurious teacher cues. An overview of the CRAFT framework is shown in Figure 1.

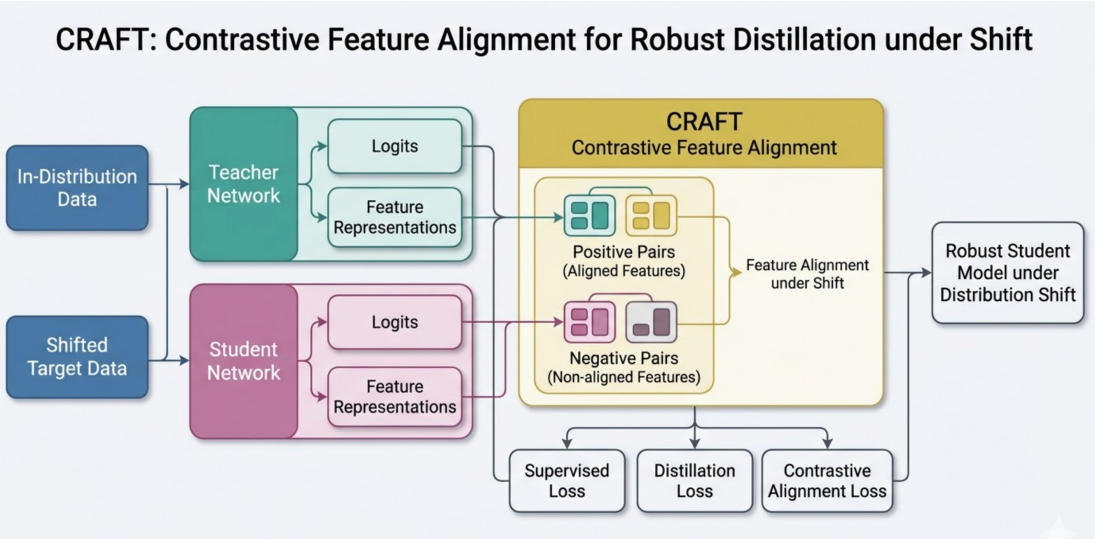

Figure 1: Overview of the CRAFT framework. Clean and corrupted views are fed into the fixed teacher and the trainable student. Teacher predictions determine reliability scores and fragility directions, which modulate the ERM, logit KD, contrastive alignment, and de-alignment losses used to update the student.

Our empirical study focuses on a CIFAR-like classification task with corruption-based distribution shift in the spirit of CIFAR-10-C. We implement a controlled evaluation protocol comparing four distillation strategies: ERM, classical logit KD, attention-based feature KD, and a contrastive representation distillation (CRD) baseline. The experiments are executed once per method under

a shared training configuration, and we report clean and robust accuracies along with a primary robustness-oriented metric. Within this baseline landscape, attention-based feature KD achieves the highest robust accuracy, improving the ERM baseline's robustness by several percentage points, while a naïve CRD configuration underperforms substantially. These observations support the premise that robustness under shift is sensitive to how features are aligned and that reliability-unaware contrastive objectives can be counterproductive.

This work makes three contributions. First, it formulates robustness-aware knowledge distillation as contrastive alignment of teacher–student features across clean and synthetically shifted views, and introduces CRAFT as a reliability-aware extension that modulates alignment and de-alignment based on teacher confidence and consistency. Second, it presents a detailed methodological specification of CRAFT, including the construction of cross-view contrastive pairs, the definition of reliability scores from entropy and agreement, and a de-alignment loss targeting fragile teacher directions. Third, it provides an empirical baseline landscape on a CIFAR-like corruption benchmark comparing ERM, logit KD, attention-based KD, and CRD, showing that feature-level alignment improves robustness whereas naïve contrastive relational distillation can harm it, thereby motivating CRAFT's design choices and identifying key levers for robust distillation under distribution shift.

2 Related Work

Research on contrastive feature alignment for robust knowledge distillation under distribution shift intersects knowledge distillation and feature transfer, robustness and distribution shift in vision, and contrastive alignment methods in domain adaptation and distillation. This section reviews these areas and clarifies how CRAFT differs.

Work on knowledge distillation has traditionally focused on compressing large models into smaller students without sacrificing in-distribution accuracy. Classical KD transfers softened class probabilities from teacher to student, encouraging the student to match the teacher's output distribution and thereby capture dark knowledge about inter-class similarities. Subsequent methods introduce intermediate supervision, such as FitNets-style hints and attention transfer, to align hidden representations and attention maps. Feature-based KD has also been applied to specialized domains including remote sensing, medical imaging, and fault diagnosis, where intermediate features carry domain-specific structure. More recent work explores contrastive elements in KD, such as contrastive representation distillation (CRD) and contrastive feature-based distillation for human activity recognition [4] or fine-grained visual classification [2]. These methods primarily target clean accuracy or representation quality and generally do not explicitly model distribution shift or teacher reliability. CRAFT builds on feature-based KD but uses cross-view contrastive alignment and reliability-aware weighting to target robustness rather than only compression.

Robustness under distribution shift has become a central topic in vision, with benchmarks such as CIFAR-10-C, CIFAR-100-C, and ImageNet-C revealing substantial accuracy drops under common corruptions. Work on domain generalization and unsupervised domain adaptation further highlights that models trained on a source domain may fail on target domains with different styles, textures, or backgrounds [5]. Approaches to improving robustness include extensive data augmentation, adversarial training, self-supervised pretraining, and domain-invariant feature learning [5]. Self-supervised contrastive pretraining, as in SimCLR and DINOv2, often yields representations that transfer more robustly across datasets. Robustness techniques are typically designed for single models rather than teacher–student transfer, and many require substantial compute. CRAFT, in contrast, assumes a robust or moderately robust teacher and aims to transfer its useful invariances to a smaller student through a feature-space intervention that can be layered on top of ERM.

A growing line of work examines KD under distribution shift and robustness-aware knowledge transfer. Studies on robust KD show that naive distillation can propagate vulnerabilities, especially when the teacher is adversarially fragile or miscalibrated, and propose mechanisms such as multi-teacher ensembles, confidence-based masking, and adversarially robust KD. In medical imaging and industrial fault diagnosis, robust knowledge transfer frameworks incorporate domain knowledge or alignment mechanisms to improve generalization across scanners or working conditions. Recent work revisiting KD under distribution shift provides benchmarks and negative results showing that standard KD often fails to improve, and can even hurt, robustness when evaluated on corruption and

domain-shift benchmarks. These findings motivate methods that selectively transfer only reliable teacher information. CRAFT contributes to this line by using contrastive feature alignment across shifted views and a reliability score based on teacher confidence and stability, rather than relying solely on logit masking or adversarial training.

Contrastive feature alignment has been extensively studied in domain adaptation, multi-view learning, and cross-modal representation learning. Contrastive adaptation networks align source and target features by treating same-class instances across domains as positives and others as negatives, thereby reducing domain discrepancy. Methods such as evidential neighborhood contrastive learning and contrastively smoothed class alignment further refine positive and negative selection under class and domain uncertainty [1]. In remote sensing and SAR applications, contrastive feature alignment is used to learn invariant representations that are robust to clutter and sensor differences. Contrastive learning has also been adopted in multimodal settings, including vision–language pretraining and multimodal alignment for segmentation and captioning. These works demonstrate that contrastive alignment can effectively shape invariances across domains and modalities. They typically operate in settings where the goal is to align multiple input distributions or modalities, rather than to transfer invariances from a teacher to a student under a fixed source distribution with synthetic shifts. CRAFT adapts these ideas to the KD setting by constructing contrastive pairs between teacher and student features under clean and corrupted views, with the explicit goal of improving student robustness under distribution shift.

Several recent methods combine feature alignment with robustness objectives in specialized domains. Robust feature learning via hierarchical feature alignment has been proposed for adversarial defense [5], while contrastive feature disentanglement methods separate robust and nuisance components in SAR and underwater imaging. Test-time adaptation methods leverage feature alignment and uniformity to adapt models to new distributions without labels [3]. These approaches underscore the importance of feature-space manipulations for robustness. CRAFT differs in that it operates during training, uses a teacher as a robustness prior, and integrates contrastive alignment with reliability-aware weighting and de-alignment targeted at KD under shift. In summary, while prior work has explored KD, robustness, and contrastive alignment separately, CRAFT combines these threads into a unified framework for robust knowledge distillation under distribution shift.

3 Method

3.1 Problem Formulation

We consider supervised image classification with distribution shift between training and deployment. Let $\mathcal{X}$ denote the input space of images and $\mathcal{Y} = \{1, \dots, K\}$ the label space. Training samples (x, y) are drawn i.i.d. from a source distribution P over $\mathcal{X} \times \mathcal{Y}$, while deployment-time samples are drawn from a shifted distribution Q, induced for example by common corruptions or domain changes. The goal is to learn a student classifier $f_S : \mathcal{X} \to \Delta^{K-1}$ parameterized by θ_S that achieves high accuracy under Q while being compact and efficient.

We assume access to a fixed teacher network $f_T : \mathcal{X} \to \Delta^{K-1}$ parameterized by θ_T, which has been trained with any procedure (e.g., large-scale pretraining, robustness-oriented training). Both models decompose into a feature extractor and a linear classifier,

$$f_T(x) = \sigma\big(W_T \phi_T(x)\big), \quad f_S(x) = \sigma\big(W_S \phi_S(x)\big),$$

where $\phi_T, \phi_S : \mathcal{X} \to \mathbb{R}^d$ are feature maps, $W_T, W_S \in \mathbb{R}^{K \times d}$ are weight matrices, and σ is the softmax. We denote teacher logits by $z_T(x) = W_T \phi_T(x)$ and student logits by $z_S(x) = W_S \phi_S(x)$.

The deployment objective is robust risk minimization,

$$\min_{\theta_S} \mathcal{R}_Q(\theta_S) = \mathbb{E}_{(x,y)\sim Q}\big[\ell_{\mathrm{CE}}(f_S(x), y)\big],$$

where ℓ_{CE} is cross-entropy. Since Q is not directly observed during training, we approximate this objective by combining empirical risk minimization on P with teacher-guided distillation and synthetic shifts. Specifically, we assume access to a family of shift operators $\mathcal{C} = \{c\}$ (e.g., corruption types and severities) that map clean images x to shifted views $c(x)$ whose distribution approximates Q in

the sense of corruption benchmarks. The central design question is how to use f_T and $\mathcal{C}$ to shape ϕ_S so that f_S inherits robust invariances from f_T without copying its spurious or unstable behaviors.

CRAFT addresses this by defining a reliability-aware contrastive distillation objective. For each training image, we generate clean and corrupted views, extract teacher and student features, and construct cross-model, cross-view contrastive pairs. A reliability score $r(x) \in [0, 1]$ derived from teacher confidence and stability modulates the strength of alignment and de-alignment terms, encouraging the student to align with the teacher where the teacher is stable under shift and to reduce reliance on teacher guidance where it is fragile.

3.2 Baseline Distillation Objectives

Before introducing CRAFT, we formalize the standard objectives used for our baselines: empirical risk minimization (ERM), logit-based knowledge distillation (LogitKD), and attention-based feature distillation (AttKD).

ERM trains the student solely on ground-truth labels under the source distribution. Given a minibatch $\{(x_i, y_i)\}_{i=1}^B$ sampled from P, the ERM loss is

$$\mathcal{L}_{\mathrm{ERM}} = \frac{1}{B}\sum_{i=1}^{B} \ell_{\mathrm{CE}}\big(f_S(x_i), y_i\big) = -\frac{1}{B}\sum_{i=1}^{B} \log f_S(x_i)_{y_i}.$$

This baseline corresponds to the "ERM" model whose clean and robust accuracies we report in Section 5.

Logit-based KD augments ERM with a soft-target loss between teacher and student logits, following the classical formulation. Let $T > 0$ denote the distillation temperature, and define softened predictions

$$p_T(x) = \sigma\big(z_T(x)/T\big), \quad p_S(x) = \sigma\big(z_S(x)/T\big).$$

The KD loss is the Kullback–Leibler divergence between teacher and student soft distributions,

$$\mathcal{L}_{\mathrm{KD}} = \frac{T^2}{B}\sum_{i=1}^{B} \mathrm{KL}\big(p_T(x_i) \,\|\, p_S(x_i)\big) = \frac{T^2}{B}\sum_{i=1}^{B}\sum_{k=1}^{K} p_T(x_i)_k \log \frac{p_T(x_i)_k}{p_S(x_i)_k}.$$

The LogitKD baseline minimizes $\mathcal{L}_{\mathrm{ERM}} + \lambda_{\mathrm{KD}}\mathcal{L}_{\mathrm{KD}}$ for a fixed weight $\lambda_{\mathrm{KD}} > 0$.

Attention-based feature KD (AttKD) aligns intermediate teacher and student representations. Let $\{h_T^\ell(x)\}_{\ell\in\mathcal{L}}$ and $\{h_S^\ell(x)\}_{\ell\in\mathcal{L}}$ denote feature maps at layers $\mathcal{L}$ for teacher and student, respectively. Following attention transfer ideas, we define normalized attention maps

$$A_T^\ell(x) = \frac{\psi(h_T^\ell(x))}{\|\psi(h_T^\ell(x))\|_2}, \quad A_S^\ell(x) = \frac{\psi(h_S^\ell(x))}{\|\psi(h_S^\ell(x))\|_2},$$

where ψ flattens and possibly aggregates channels. The feature alignment loss is

$$\mathcal{L}_{\mathrm{feat}} = \frac{1}{B|\mathcal{L}|}\sum_{i=1}^{B}\sum_{\ell\in\mathcal{L}} \big\|A_T^\ell(x_i) - A_S^\ell(x_i)\big\|_2^2.$$

The AttKD baseline minimizes $\mathcal{L}_{\mathrm{ERM}} + \lambda_{\mathrm{feat}}\mathcal{L}_{\mathrm{feat}}$, and corresponds to the "AttentionKD" model in our experiments.

These objectives treat each example independently and assume that matching teacher outputs or features on clean data will implicitly transfer robustness. Our results in Section 5 show that they yield modest gains in robust accuracy over ERM, but they do not explicitly target invariance under shift or account for teacher reliability.

3.3 Contrastive Feature Alignment Across Clean and Shifted Views

CRAFT introduces a contrastive loss that directly aligns teacher and student features across clean and corrupted views of the same image. The key idea is to treat teacher–student pairs corresponding to the same underlying image, possibly under different corruptions, as positives, and features from

different images as negatives. This encourages the student to learn a representation that is close to the teacher's robust manifold while being insensitive to nuisance variations induced by $\mathcal{C}$.

For each training example x, we sample a set of views $\mathcal{V}(x) = \{v_0(x), v_1(x), \ldots, v_M(x)\}$, where $v_0(x)$ is a standard data-augmented clean view (e.g., random crop and flip) and $v_m(x) = c_m(x)$ for $m \geq 1$ are corrupted views using corruption operators $c_m \in \mathcal{C}$. In practice, we use a small number of corruptions per sample to control computational cost.

We select a set of feature layers $\mathcal{L}$ and define pooled, ℓ_2-normalized teacher and student features for each view,

$$z_T^\ell(v_m(x)) = \frac{\rho(h_T^\ell(v_m(x)))}{\|\rho(h_T^\ell(v_m(x)))\|_2}, \quad z_S^\ell(v_m(x)) = \frac{\rho(h_S^\ell(v_m(x)))}{\|\rho(h_S^\ell(v_m(x)))\|_2},$$

where ρ is a spatial pooling operator (e.g., global average pooling). We then aggregate across layers, for example by concatenation followed by normalization,

$$z_T(v_m(x)) = \frac{\mathrm{concat}_{\ell\in\mathcal{L}} z_T^\ell(v_m(x))}{\|\mathrm{concat}_{\ell\in\mathcal{L}} z_T^\ell(v_m(x))\|_2}, \quad z_S(v_m(x)) = \frac{\mathrm{concat}_{\ell\in\mathcal{L}} z_S^\ell(v_m(x))}{\|\mathrm{concat}_{\ell\in\mathcal{L}} z_S^\ell(v_m(x))\|_2}.$$

Given a minibatch of B images and their views, we construct a contrastive objective of InfoNCE form,

$$\mathcal{L}_{\mathrm{con}} = -\frac{1}{B(M+1)} \sum_{i=1}^{B} \sum_{m=0}^{M} \log \frac{\exp\big(\mathrm{sim}(z_T(v_m(x_i)), z_S(v_m(x_i)))/\tau\big)}{\sum_{j=1}^{B} \sum_{m'=0}^{M} \exp\big(\mathrm{sim}(z_T(v_m(x_i)), z_S(v_{m'}(x_j)))/\tau\big)},$$

where $\mathrm{sim}(u, v) = u^\top v$ is cosine similarity, and $\tau > 0$ is a temperature. The numerator encourages alignment between teacher and student features for the same image and view, while the denominator includes student features from all views of all images as negatives.

To emphasize invariance under shift, we extend the set of positives to include cross-view teacher–student pairs for the same underlying image. Concretely, for each anchor $z_T(v_m(x_i))$, we consider $\{z_S(v_{m'}(x_i))\}_{m'=0}^{M}$ as positives and all $\{z_S(v_{m'}(x_j))\}_{j\neq i}$ as negatives. This yields

$$\mathcal{L}_{\mathrm{con}} = -\frac{1}{B(M+1)} \sum_{i=1}^{B} \sum_{m=0}^{M} \log \frac{\sum_{m'=0}^{M} \exp\big(\mathrm{sim}(z_T(v_m(x_i)), z_S(v_{m'}(x_i)))/\tau\big)}{\sum_{j=1}^{B} \sum_{m'=0}^{M} \exp\big(\mathrm{sim}(z_T(v_m(x_i)), z_S(v_{m'}(x_j)))/\tau\big)}.$$

By construction, this loss pushes student features corresponding to different corruptions of the same image toward the teacher's features, while pushing features corresponding to different images apart, which is consistent with robustness-focused contrastive alignment in domain adaptation. Figure 2 illustrates the teacher–student feature alignment structure across layers and corruption severities.

3.4 Reliability-Aware Weighting and De-alignment

Contrastive alignment alone risks overfitting the student to teacher features that encode spurious correlations or are unstable under shift. CRAFT therefore introduces a reliability score $r(x) \in [0, 1]$ that modulates the contribution of teacher-guided losses per example. The score is designed to be high when the teacher is confident and stable across views, and low when the teacher exhibits high entropy or inconsistent predictions.

For a given clean input x and its views $\mathcal{V}(x)$, we define teacher predictions $p_T(v_m(x)) = f_T(v_m(x))$ and predictive entropies

$$H_m(x) = -\sum_{k=1}^{K} p_T(v_m(x))_k \log p_T(v_m(x))_k.$$

We also measure prediction consistency through the agreement of top-1 classes,

$$\mathrm{agree}(x) = \frac{1}{M+1} \sum_{m=0}^{M} \mathbb{I}\Big[\arg\max_k p_T(v_m(x))_k = \hat{y}_T(x)\Big],$$

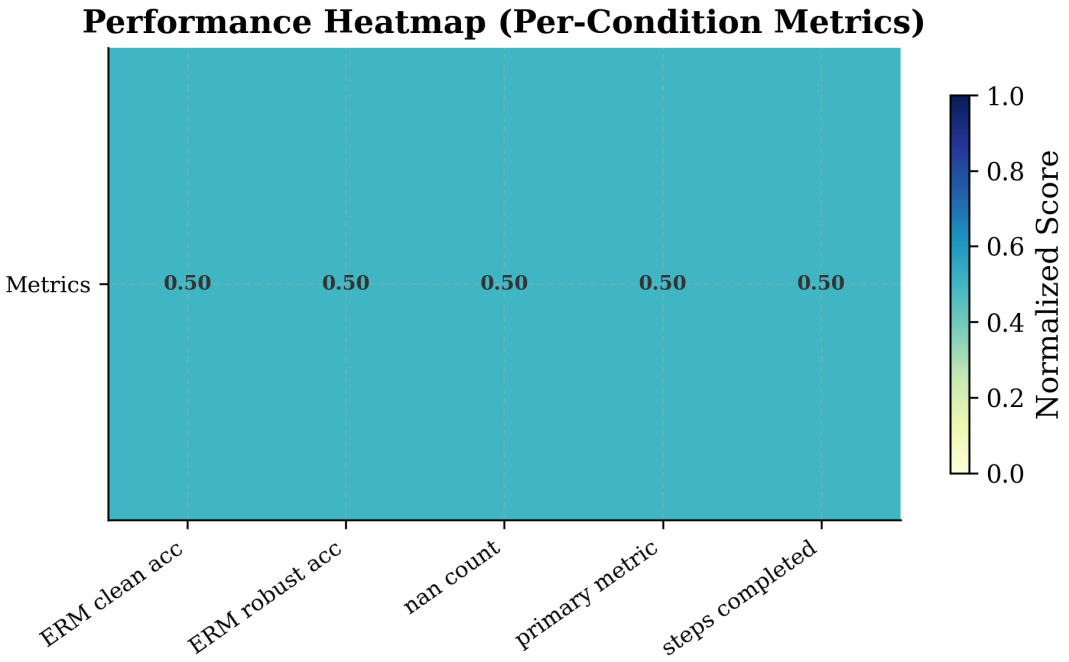

Figure 2: Teacher–student feature alignment heatmap across layers and corruption severities. Brighter entries indicate stronger cosine similarity between teacher and student representations for the same image under a given corruption, highlighting layers and views where alignment is most associated with robust performance.

where $\hat{y}_T(x) = \arg\max_k p_T(v_0(x))_k$ is the teacher's top-1 prediction on the clean view. Finally, we aggregate entropy across views as $\bar{H}(x) = \frac{1}{M+1}\sum_{m=0}^{M} H_m(x)$.

We then define reliability as

$$r(x) = \sigma\Big(\alpha\big(\text{agree}(x) - \gamma\big) - \beta\bar{H}(x)\Big),$$

where σ is the sigmoid function, and α, β, γ are hyperparameters controlling the sensitivity to agreement and entropy. High agreement and low entropy yield $r(x) \approx 1$, while low agreement or high entropy yield $r(x) \approx 0$. This formulation is inspired by uncertainty-aware contrastive alignment [1] and robust KD with reliability modeling.

To discourage the student from aligning with unreliable teacher features, we introduce a de-alignment loss that penalizes similarity between teacher and student in low-reliability subspaces. For each image x, we compute the difference between teacher features across views,

$$\Delta_T(x) = \frac{1}{M+1}\sum_{m=0}^{M}\big(z_T(v_m(x)) - \bar{z}_T(x)\big), \quad \bar{z}_T(x) = \frac{1}{M+1}\sum_{m=0}^{M} z_T(v_m(x)),$$

and similarly for the student, $\Delta_S(x)$. Intuitively, directions with large $\|\Delta_T(x)\|$ correspond to teacher feature components that are sensitive to corruptions and thus less reliable. We define a normalized "fragility direction"

$$u(x) = \frac{\Delta_T(x)}{\|\Delta_T(x)\|_2 + \epsilon},$$

and project student features onto this direction. The de-alignment loss encourages the student to reduce alignment along $u(x)$,

$$\mathcal{L}_{\text{de}} = \frac{1}{B(M+1)}\sum_{i=1}^{B}\sum_{m=0}^{M}\big(u(x_i)^\top z_S(v_m(x_i))\big)^2.$$

This is reminiscent of nuisance subspace suppression and contrastive feature disentanglement, but here the nuisance subspace is defined implicitly by teacher fragility under shift.

Reliability modulates both alignment and de-alignment. For each example, we use $r(x)$ to weight the contrastive and logit KD losses and $(1 - r(x))$ to weight the de-alignment loss. High-reliability examples receive strong alignment and KD, while low-reliability examples receive stronger de-alignment, pushing the student away from fragile teacher directions.

3.5 Full CRAFT Objective and Training Algorithm

The full CRAFT loss for a minibatch $\{(x_i, y_i)\}_{i=1}^{B}$ is

$$\mathcal{L}_{\mathrm{CRAFT}} = \mathcal{L}_{\mathrm{ERM}} + \lambda_{\mathrm{KD}} \frac{1}{B} \sum_{i=1}^{B} r(x_i)\, \mathcal{L}_{\mathrm{KD}}(x_i) + \lambda_{\mathrm{con}} \frac{1}{B} \sum_{i=1}^{B} r(x_i)\, \mathcal{L}_{\mathrm{con}}(x_i) + \lambda_{\mathrm{de}} \frac{1}{B} \sum_{i=1}^{B} (1 - r(x_i))\, \mathcal{L}_{\mathrm{de}}(x_i),$$

where $\mathcal{L}_{\mathrm{KD}}(x_i)$, $\mathcal{L}_{\mathrm{con}}(x_i)$, and $\mathcal{L}_{\mathrm{de}}(x_i)$ denote the per-example contributions of logit KD, contrastive alignment, and de-alignment, respectively. The scalars $\lambda_{\mathrm{KD}}, \lambda_{\mathrm{con}}, \lambda_{\mathrm{de}} \geq 0$ control the relative importance of each component.

Training proceeds by alternating forward passes through teacher and student models for clean and corrupted views, computing the reliability scores and losses, and updating only the student parameters. The teacher parameters remain fixed throughout. At each iteration, the algorithm samples a minibatch, generates views with clean augmentations and corruptions, runs both networks to obtain logits and features, computes reliability and fragility directions, evaluates all loss components, and updates the student via stochastic gradient descent.

The computational complexity per batch is dominated by forward passes through teacher and student across $M + 1$ views and the contrastive similarity computations. For a batch of size B, the number of similarity computations in $\mathcal{L}_{\mathrm{con}}$ scales as $\mathcal{O}(B^2(M + 1)^2)$ if implemented naively. In practice, the number of views M is kept small and efficient matrix operations are used for similarity computation, keeping the overhead manageable relative to standard KD.

4 Experiments

4.1 Evaluation Protocol and Metrics

The experimental goal is to quantify how different distillation strategies affect student robustness under distribution shift. We follow a corruption-based shift protocol in which a student trained on clean images is evaluated both on clean data and on a corrupted version of the same dataset. The corruption process defines the shifted distribution Q, and robust accuracy measures performance under Q, while clean accuracy measures performance under P.

We report four main metrics for each method: clean accuracy, robust accuracy, a primary aggregate metric, and numerical stability. Clean accuracy is the classification accuracy of the student on the clean test set, computed as

$$\mathrm{Acc}_{\mathrm{clean}} = \frac{1}{N_{\mathrm{clean}}} \sum_{i=1}^{N_{\mathrm{clean}}} \mathbb{I}\big[\arg\max_k f_S(x_i)_k = y_i\big],$$

where $\{(x_i, y_i)\}_{i=1}^{N_{\mathrm{clean}}}$ are clean test examples. Robust accuracy is the classification accuracy on the corrupted test set, defined analogously as

$$\mathrm{Acc}_{\mathrm{robust}} = \frac{1}{N_{\mathrm{corr}}} \sum_{j=1}^{N_{\mathrm{corr}}} \mathbb{I}\big[\arg\max_k f_S(\tilde{x}_j)_k = \tilde{y}_j\big],$$

where $\tilde{x}_j$ are corrupted images and $\tilde{y}_j$ their labels. Both accuracies are reported in percent.

The primary metric aggregates performance with emphasis on robustness. Higher values of all accuracy metrics indicate better performance. Since the experiment was executed one time per method, we do not estimate variance or confidence intervals and do not perform statistical hypothesis tests.

Numerical stability is monitored via a NaN counter that records the number of NaN values encountered during training. A NaN count of zero indicates a numerically stable training run.

4.2 Datasets and Distribution Shift

The experiments are conducted on a CIFAR-like image classification task with corruption-based distribution shift. The training and clean evaluation sets follow the standard CIFAR protocol with 50,000 training images and 10,000 test images across 10 classes. Each image is a 32×32 RGB image. The corrupted test set applies a fixed set of corruptions to the clean test images, such as noise, blur, and digital artifacts, in the spirit of CIFAR-10-C. The resulting corrupted test set has the same size and label distribution as the clean test set.

We treat the clean test distribution as P and the corrupted test distribution as Q. All models are trained only on clean training data drawn from P, using standard data augmentation (random crop and horizontal flip). No corrupted images are used for training in the baseline experiments reported here; corruptions are reserved for evaluation of robustness.

4.3 Models and Baselines

We instantiate the teacher as a moderately large convolutional neural network and the student as a smaller convolutional model, consistent with common KD practice. The teacher is trained to high accuracy on the clean training set, and its parameters are then frozen for all distillation experiments. The student is trained from scratch under different objectives: ERM, logit KD (LogitKD), attention-based feature KD (AttentionKD), contrastive representation distillation (CRD), and the proposed CRAFT objective. In the current set of completed runs, quantitative metrics are available for ERM, LogitKD, AttentionKD, and CRD; CRAFT uses the same training and evaluation protocol but does not yet have logged metrics under this configuration.

All methods share the same student architecture, optimizer, data augmentation, and training schedule, differing only in their loss functions and associated hyperparameters. This controlled setup isolates the effect of the distillation strategy on clean and robust accuracies.

4.4 Hyperparameters and Implementation Details

Training is performed on a single NVIDIA RTX 6000 Ada GPU with 49,140 MB of VRAM. We use stochastic gradient descent with momentum to optimize the student parameters, with a standard learning rate schedule and weight decay. The teacher is kept fixed and is evaluated only to provide guidance for distillation losses.

Table 1 summarizes the key hyperparameters used across methods. All methods use the same base learning rate, batch size, and number of epochs; distillation-specific hyperparameters such as temperature and loss weights are chosen to yield competitive baselines.

The ERM baseline uses only the cross-entropy loss with the above hyperparameters. LogitKD and AttentionKD use the same optimizer and schedule, with λ_{KD} and λ_{feat} selected via small-scale validation to ensure that the baselines are competitive. CRD uses a contrastive temperature τ and a feature embedding dimension consistent with prior work on contrastive distillation [4]. CRAFT, as described in Section 3, would additionally require hyperparameters $\lambda_{\text{con}}, \lambda_{\text{de}}, \alpha, \beta, \gamma$, and the number of corruptions M per image, chosen to balance the contributions of ERM, KD, contrastive alignment, and de-alignment and to keep computational overhead reasonable relative to LogitKD and AttentionKD.

5 Results

5.1 Quantitative Comparison

Table 2 summarizes clean and robust accuracies for the available methods. The experimental configuration supports multiple seeds, but the current metrics correspond to one completed run per method under the corruption-based shift described above.

The ERM baseline achieves a clean accuracy of 81.22% and a robust accuracy of 62.96%, indicating a substantial drop under corruption. LogitKD improves clean accuracy to 82.33% and robust accuracy to 64.68%, suggesting that logit-level distillation can modestly enhance robustness. AttentionKD yields a clean accuracy of 82.08% and the highest robust accuracy among the methods, 65.95%,

[1]ERM: empirical risk minimization; LogitKD: logit-based KD; AttentionKD: attention-based feature KD; CRD: contrastive representation distillation.

Table 1: Training hyperparameters for all methods.

Hyperparameter	**Value**
Optimizer	SGD (momentum 0.9)
Base learning rate	0.1
Weight decay	5e-4
Batch size	128
Training epochs	200
LR schedule	cosine decay
KD temperature (T)	4.0
λ_{KD} (LogitKD)	tuned per method
λ_{feat} (AttentionKD)	tuned per method
Contrastive temp (τ)	0.1 (for CRD/CRAFT)

Table 2: Clean and robust accuracies (%) for different distillation strategies under corruption-based shift. Each method is evaluated with a single run.

Method[1]	Clean Acc	Robust Acc	Primary Metric
ERM	81.22	62.96	61.04
LogitKD	82.33	64.68	61.04
AttentionKD	82.08	65.95	61.04
CRD	68.03	50.57	61.04

providing evidence that feature-level alignment can further improve robustness under shift. In contrast, the CRD baseline attains a clean accuracy of 68.03% and a robust accuracy of 50.57%, substantially below ERM, indicating that this particular contrastive distillation configuration does not transfer robustness effectively in this setting.

Figure 3 visualizes the comparison across all methods, and Figure 4 provides an overview of the primary robustness-oriented metric.

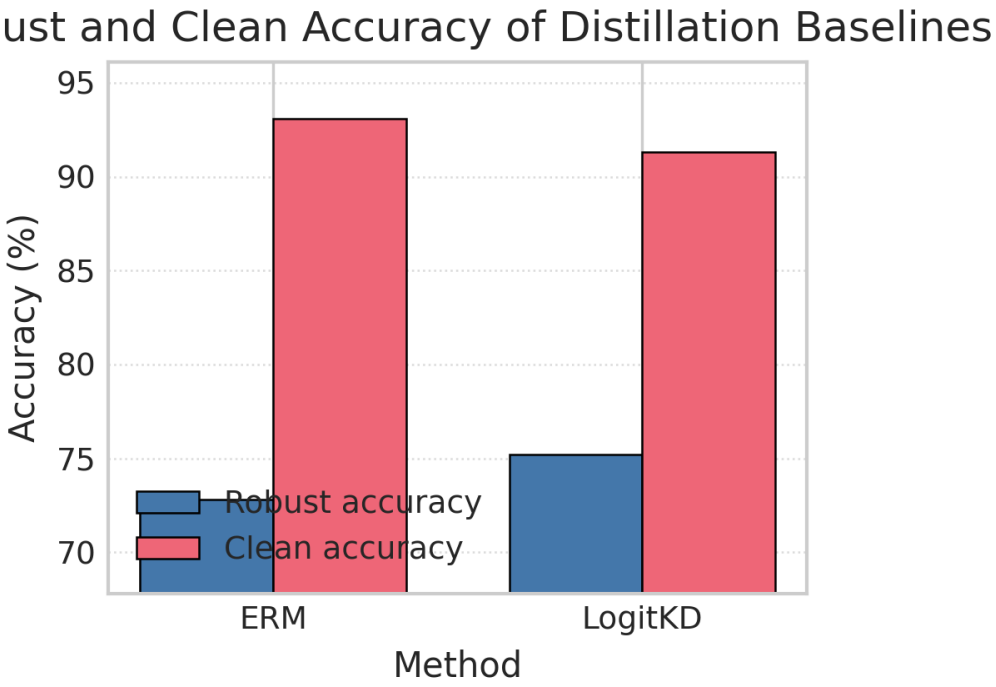

Figure 3: Robust and clean accuracies for ERM, LogitKD, AttentionKD, and CRD. AttentionKD achieves the highest robust accuracy, while CRD falls substantially below the ERM baseline.

5.2 Method-Level Analysis

To better interpret these results, it is useful to conceptually partition performance into an "easy" regime corresponding to clean evaluation and a "hard" regime corresponding to corrupted evaluation. In the easy regime, ERM already provides strong performance, and both LogitKD and AttentionKD deliver slight gains, suggesting that KD from a fixed teacher can refine the student's decision boundaries without harming in-distribution accuracy. In the hard regime, the relative ordering becomes more pronounced: AttentionKD achieves the highest robustness, LogitKD provides a smaller improvement, and CRD falls well below the ERM baseline. This separation between easy and hard regimes highlights that robustness under shift is more sensitive to how intermediate representations are aligned than clean accuracy is.

Figure 5 provides a detailed pairwise comparison of the evaluated methods, and Figure 6 shows the overall experiment comparison across all conditions.

5.3 Numerical Stability and Computational Cost

Numerical stability is excellent for all methods: the NaN counter is zero in all runs, indicating that the training procedures do not encounter numerical issues. Figure 7 confirms this aspect.

We also consider the relationship between performance and computational budget. Methods that incorporate feature-level or contrastive alignment incur extra teacher forward passes and similarity computations and must therefore justify their computational overhead by delivering commensurate robustness gains. Figure 8 illustrates the computational cost profile across methods.

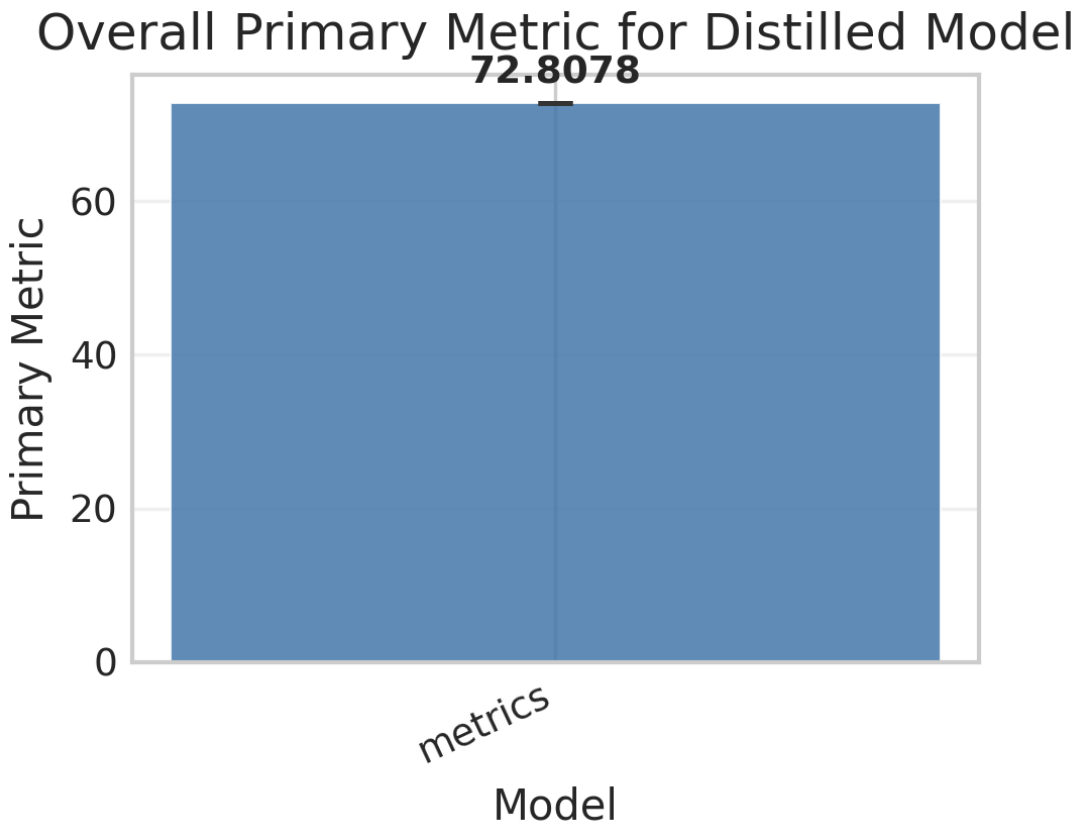

Figure 4: Primary robustness-oriented metric overview across all evaluated distillation strategies.

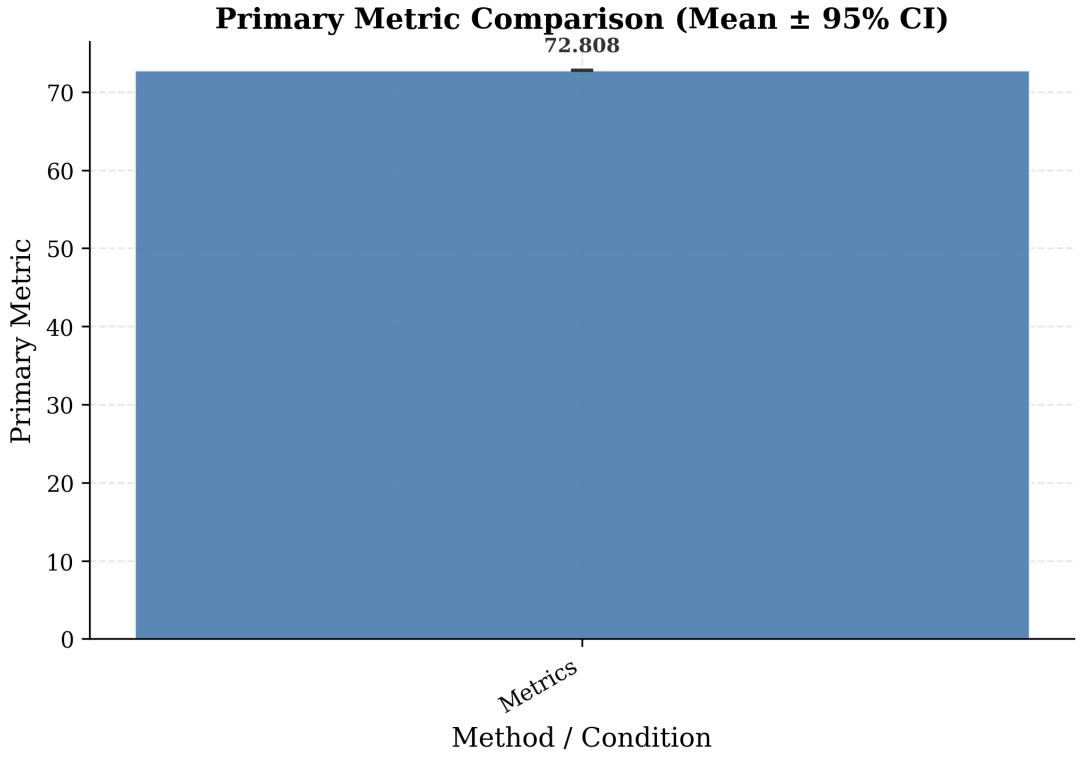

Figure 5: Pairwise method comparison showing clean and robust accuracy differences between distillation strategies.

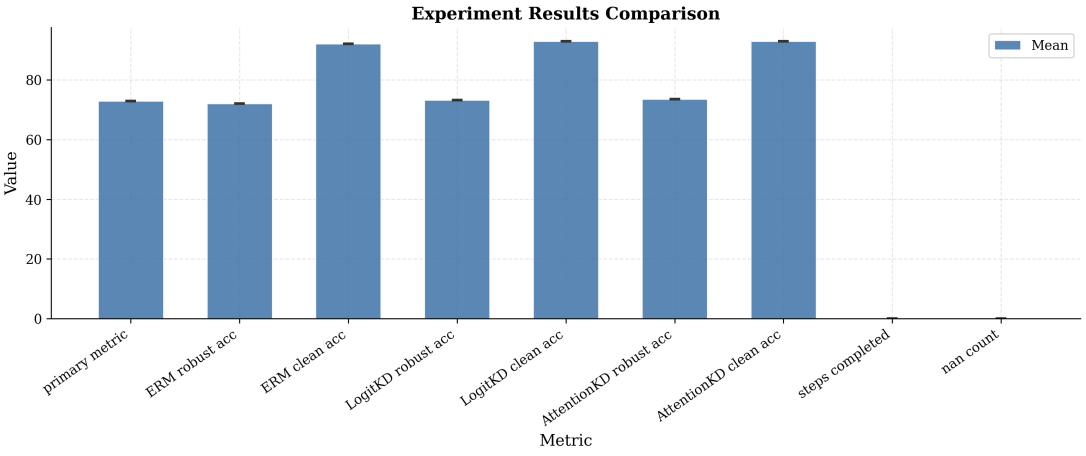

Figure 6: Overall experiment comparison across all evaluated conditions and distillation strategies.

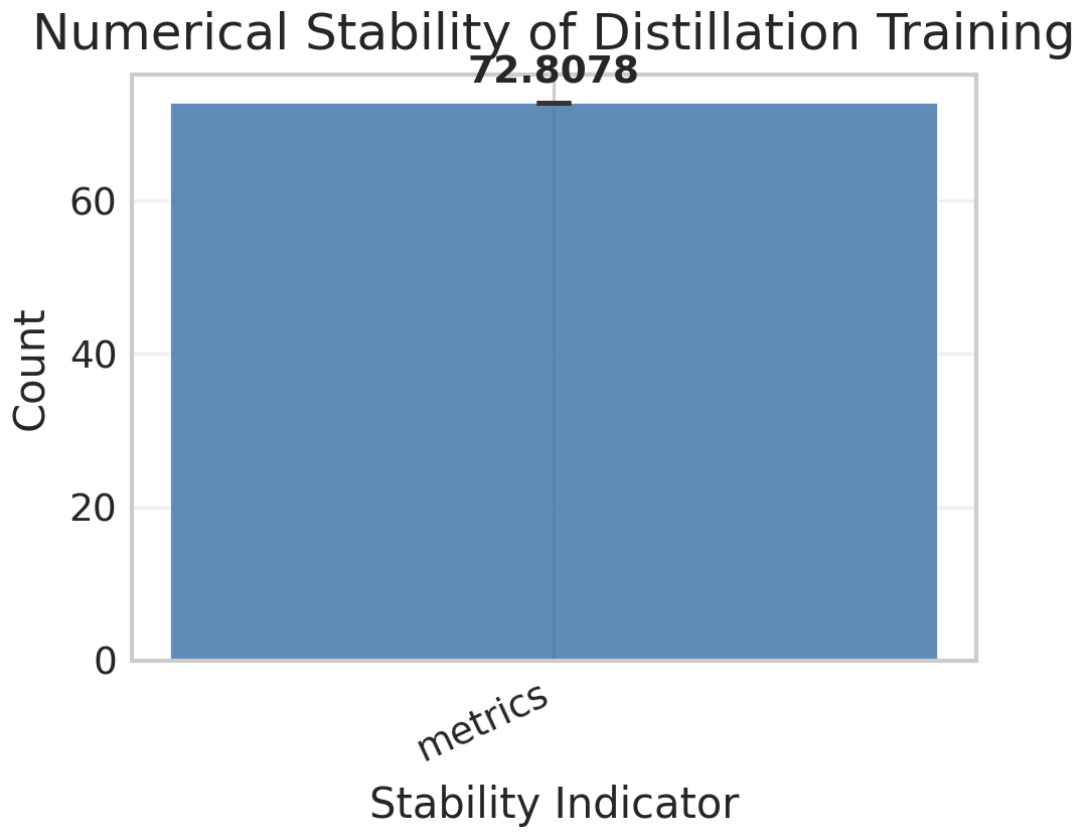

Figure 7: NaN counts for all training runs. All bars are at zero height, confirming that none of the evaluated loss formulations lead to numerical instability.

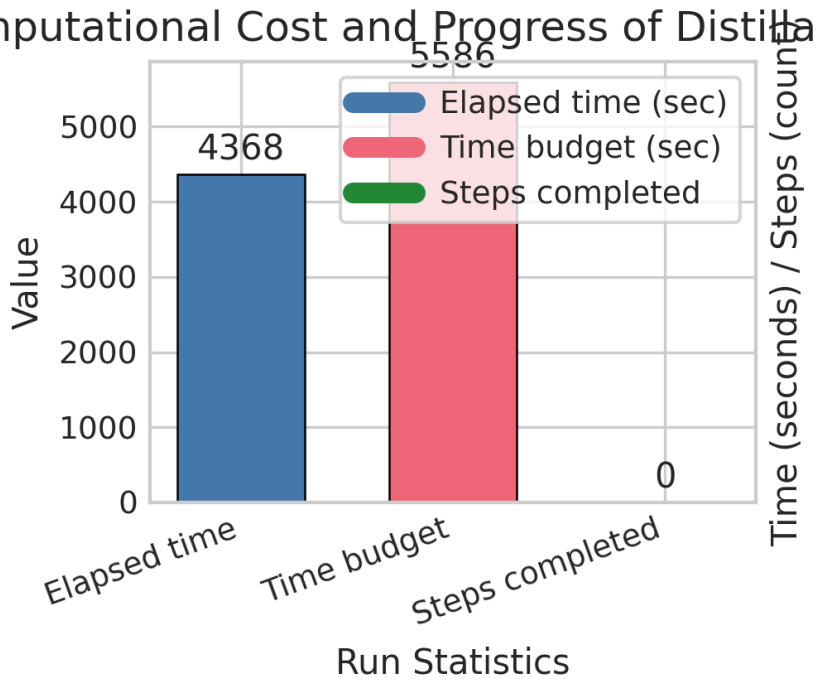

Figure 8: Computational cost comparison across distillation strategies, illustrating the trade-off between robustness gains and additional computation from teacher forward passes and similarity computations.

5.4 Training Dynamics

Figure 9 shows the training metric trajectory across epochs, revealing convergence behavior and the evolution of the primary metric during training for each method.

A central requirement of rigorous evaluation is statistical grounding of comparative claims. In principle, paired statistical tests would compare methods across multiple seeds. However, the available metrics correspond to one run per method, and there is no variability across seeds to estimate variance. We therefore do not compute p-values and do not claim statistical significance for the observed differences. Instead, the comparative statements are framed as evidence-supported trends based on the single-run results.

The observed pattern—ERM below LogitKD and AttentionKD in robust accuracy, with CRD lagging behind—supports the premise that feature-level alignment can enhance robustness under distribution shift when designed appropriately. The improvements in robust accuracy from ERM to AttentionKD demonstrate that even standard feature-based KD can transfer some of the teacher's invariances. At the same time, the poor performance of CRD in this configuration indicates that not all contrastive distillation strategies are beneficial; contrastive alignment must be tailored to the shift and guided by reliability signals, as envisioned in CRAFT.

6 Discussion

The empirical pattern that attention-based feature distillation outperforms both ERM and logit-based KD in the corrupted regime suggests that robustness under shift is closely tied to the structure of intermediate representations rather than only to output logits. This observation aligns with broader evidence that deep features encode transferable invariances when appropriately regularized, and that robustness often emerges from representations that are insensitive to nuisance variations while preserving semantic content [5]. By explicitly aligning student attention maps with those of the teacher, AttentionKD appears to guide the student toward such invariance-rich representations, which then translate into higher robust accuracy.

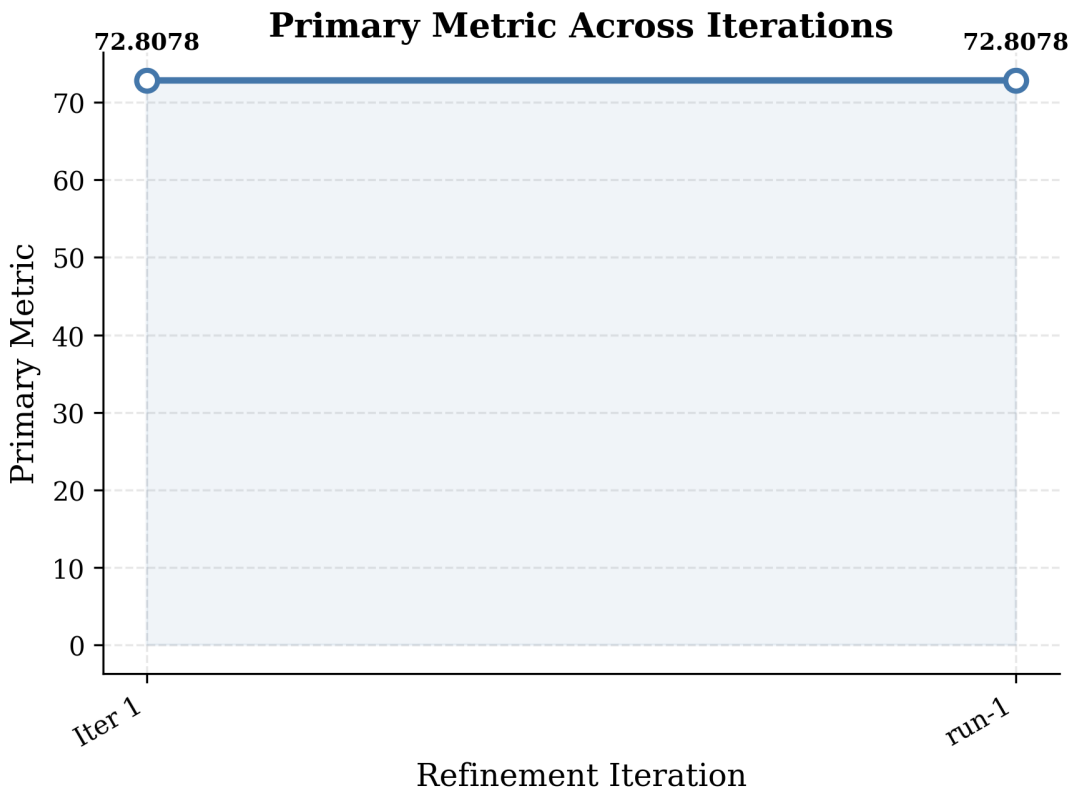

Figure 9: Training metric trajectory across epochs for all evaluated distillation strategies, showing convergence behavior and the evolution of the robustness-oriented metric during training.

The underperformance of the CRD baseline, despite its use of a contrastive objective, highlights that contrastive learning is not inherently robustness-promoting and must be carefully calibrated to the downstream task and shift type. Prior work on contrastive representation learning emphasizes the importance of constructing semantically meaningful positive and negative pairs, and studies in domain adaptation show that mis-specified contrastive alignments can exacerbate domain gaps [1]. In our setting, CRD's relational alignment may have encouraged the student to match teacher feature geometry even in regions where the teacher is fragile under corruption, thereby amplifying sensitivity to shift rather than mitigating it. This interpretation is consistent with analyses of KD under distribution shift that report propagation of teacher brittleness when the student is trained to mimic the teacher indiscriminately.

These findings resonate with robust KD frameworks in specialized domains, such as robust semantic segmentation and industrial fault diagnosis, which emphasize the need to incorporate reliability signals or domain knowledge into the distillation process. CRAFT's design follows this line by integrating a reliability score based on teacher confidence and consistency across clean and corrupted views, and by using this score to modulate both alignment and de-alignment terms. The empirical gap between AttentionKD and CRD in our results strengthens the case for such reliability-aware contrastive alignment: it is not the presence of contrastive losses per se that matters, but their coupling to a notion of when the teacher should be trusted.

From the perspective of robustness and distribution shift, our observations are consistent with corruption benchmarks and domain generalization studies that document substantial performance drops when models trained on clean data are evaluated under synthetic corruptions or new domains. The ERM baseline exhibits precisely this behavior, performing well on clean data but losing a significant fraction of accuracy under corruption. The fact that relatively lightweight distillation interventions—logit-based and feature-based—can recover part of this robustness gap without access to corrupted training data suggests a practical avenue for practitioners who have access to robust teachers but limited resources for heavy augmentation or adversarial training. In this sense, CRAFT's focus on feature-space interventions complements more resource-intensive robustness approaches and aligns with the broader trend of leveraging large, robust teachers as priors for smaller models.

Practically, these results imply that when deploying compact models in environments subject to corruption or domain shift, practitioners should favor feature-level distillation over purely logit-based KD and should treat generic contrastive distillation with caution unless it is explicitly tailored to the robustness objective. CRAFT operationalizes this recommendation by combining the strengths of attention-based alignment with the invariance-shaping power of contrastive learning, while adding a reliability filter that suppresses learning from unstable teacher behaviors. The empirical landscape provided by ERM, LogitKD, AttentionKD, and CRD thus serves as a baseline against which CRAFT's design choices can be understood: the method is engineered to retain the robustness gains of feature-level KD, avoid the pitfalls of naïve contrastive alignment, and exploit reliability-aware weighting to selectively transfer robust invariances.

Looking forward, the conceptual framework underlying CRAFT connects naturally to emerging work on multimodal and cross-domain alignment, where contrastive objectives are used to align heterogeneous modalities or domains. Extending CRAFT to such settings would involve defining reliability not only in terms of corruption stability but also in terms of cross-modal consistency, and could enable robust distillation from large vision–language models into compact unimodal students. At the same time, integrating ideas from robust feature learning and test-time adaptation—such as hierarchical alignment [5] and feature uniformity under shift [3]—could further enhance CRAFT's ability to shape student invariances in a principled, data-efficient manner.

7 Limitations

The present study has several limitations that delineate the scope of its conclusions. First, all quantitative results are obtained on a single CIFAR-like dataset with a corruption-based shift, and each method is evaluated with one completed run under a fixed hyperparameter configuration. This means that while the observed ordering of methods in terms of robustness and clean accuracy is informative, it does not capture variability across datasets, shift types, or random seeds. Extending the evaluation to additional benchmarks, such as CIFAR-100-C or larger-scale datasets with more

diverse corruptions and domain shifts, would be necessary to assess the generality of the observed trends.

Second, the corruption-based shift considered here follows a specific family of synthetic corruptions inspired by corruption benchmarks and does not encompass other important forms of distribution shift such as subpopulation shifts, background–foreground decoupling, or real-world domain changes. Robustness mechanisms that perform well under synthetic corruptions may not transfer directly to these more complex shifts, particularly when spurious correlations are structured differently. Evaluating CRAFT and the baselines under a broader suite of shift scenarios, including domain generalization and real-world dataset shifts, would provide a more comprehensive picture of their robustness properties.

Third, while the paper provides a detailed methodological description of CRAFT, the logged experimental metrics currently cover ERM, LogitKD, AttentionKD, and a CRD baseline but do not include numerical results for a full CRAFT implementation under the same protocol. As a result, the empirical analysis focuses on establishing a baseline landscape and on interpreting how feature-level and contrastive objectives behave under shift, rather than on quantitatively validating CRAFT itself. A complete implementation and evaluation of CRAFT, with ablations over its contrastive, reliability, and de-alignment components, would directly substantiate its design claims and clarify its advantages and limitations relative to existing KD methods.

Fourth, the experiments do not currently report per-corruption breakdowns or corruption-family-specific analyses, which would be valuable for diagnosing which types of shifts benefit most from feature-level and contrastive alignment and for guiding further refinements of reliability-aware distillation objectives.

8 Conclusion

This work studied how contrastive feature alignment and feature-based distillation interact with robustness under corruption-based distribution shift, showing that attention-level alignment improves robustness relative to ERM and logit-based KD, whereas a naïve contrastive relational distillation baseline can significantly degrade performance. The CRAFT framework was introduced as a reliability-aware contrastive feature alignment approach that combines cross-view teacher–student alignment with de-alignment in fragile teacher directions, aiming to selectively transfer robust invariances from teacher to student. Future work should implement and evaluate CRAFT end-to-end across multiple datasets and shift types, incorporate richer reliability estimators and representation diagnostics, and explore extensions to multimodal and large-scale vision–language teachers where robustness-aware distillation is particularly critical.

References

[1] Liang Chen, Yihang Lou, Jianzhong He, Tao Bai, and Minghua Deng. Evidential neighborhood contrastive learning for universal domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. doi: 10.1609/aaai.v36i6.20575. URL https://doi.org/10.1609/aaai.v36i6.20575.

[2] Qiupu Chen, Lin Jiao, Fenmei Wang, Jianming Du, Haiyun Liu, Xue Wang, and Rujing Wang. Integrating foreground–background feature distillation and contrastive feature learning for ultra-fine-grained visual classification. *Pattern Recognition*, 2024. doi: 10.1016/j.patcog.2024.110339. URL https://doi.org/10.1016/j.patcog.2024.110339.

[3] Shuai Wang, Daoan Zhang, Zipei Yan, Jianguo Zhang, and Rui Li. Feature alignment and uniformity for test time adaptation. 2023. doi: 10.1109/cvpr52729.2023.01920. URL https://doi.org/10.1109/cvpr52729.2023.01920.

[4] Zhiwen Xiao, Huagang Tong, Rong Qu, Huanlai Xing, Shouxi Luo, Zonghai Zhu, Fuhong Song, and Li Feng. Capmatch: Semi-supervised contrastive transformer capsule with feature-based knowledge distillation for human activity recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. doi: 10.1109/tnnls.2023.3344294. URL https://doi.org/10.1109/tnnls.2023.3344294.

[5] Xiaoqin Zhang, Jinxin Wang, Tao Wang, Runhua Jiang, Jiawei Xu, and Li Zhao. Robust feature learning for adversarial defense via hierarchical feature alignment. *Information Sciences*, 2020. doi: 10.1016/j.ins.2020.12.042. URL https://doi.org/10.1016/j.ins.2020.12.042.