Anti-Shortcut Distillation via Temporal Negative Knowledge Transfer

Syed Muhammad Raza
iamraza1998@gmail.com
Omer Tariq
omer.tariq@neubility.co.kr
Jeongbae Son
thswjdqo92@neubility.co.kr

Perception AI,
Neubility Inc.,
Seoul, South Korea

Abstract

Knowledge distillation (KD) trains a compact student by attracting it towards a converged teacher. It is silent about which directions the teacher itself learned to suppress: repulsive and bias-aware objectives exist, but none exploits the teacher's own trajectory to identify what the student should avoid. We observe that the missing signal is already encoded in the teacher's optimization trajectory: features that an early-stage teacher emphasizes but that a converged teacher attenuates are precisely the shortcut directions worth pushing the student away from. We instantiate this observation as **A**nti-**S**hortcut **D**istillation (ASD), a push–pull KD framework that treats the converged teacher T_{final} as a positive semantic anchor and an early-checkpoint teacher T_{early} as a temporal negative reference. ASD couples two losses: a temporal contrastive loss ($\mathcal{L}_{\text{TC}}$) that places the early-teacher feature as a same-sample negative against in-batch and memory-bank final-teacher features in an InfoNCE objective; and a shortcut suppression loss ($\mathcal{L}_{\text{SS}}$) that penalizes student projection onto the top eigenvectors of $\mathbb{E}[\Delta h \Delta h^{\top}]$, the uncentered second-moment matrix of early-to-final feature displacements. Across 13 teacher–student pairs on CIFAR-100, ImageNet-100, and TinyImageNet, ASD attains the highest clean top-1 accuracy on more than 10 pairs and outperforms standard KD on 12. On CIFAR-100-C corruption robustness, ASD obtains the lowest mean Corruption Error (86.1 mCE) on the most challenging cross-architecture pair (WRN-40-2→ShuffleNet-V2). Mechanistic diagnostics confirm the intended geometry: the ASD student is systematically anti-aligned with the shortcut direction, while its projection onto the robust subspace is substantially larger (0.45 vs. 0.12). ASD github repository

1 Introduction

Knowledge distillation (KD) [11, 16] compresses a high-capacity teacher into a compact student by transferring supervisory structure richer than one-hot labels. Recent methods have expanded the types of knowledge transferred during distillation: FitNets [27] regresses intermediate features, AT [35] matches spatial attention, RKD [24] and PKT [25] preserve sample geometry, CRD [30] adds contrastive transfer, DKD [37] decouples logits, and ReviewKD [3] aggregates feature hierarchies, but all share one geometric premise: distillation is treated as a unilateral attractive force. The student is pulled toward the teacher, but the objective is silent about which representational directions the student should avoid. Repulsive

© 2026. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

sample-level negatives [30], checkpoint-based positive targets [32], and bias-aware distillation [20, 22] partially address this asymmetry, but none uses the teacher itself to identify harmful representational directions.

This asymmetry matters because teacher representations are themselves products of a temporal optimization path. Networks tend to exploit local textures, colour statistics, and background cues early in training before consolidating shape-biased, semantically stable features [1, 9, 10]. A converged teacher records the endpoint of this process, not the shortcut-prone directions traversed and attenuated along the way. A compact student trained under the same data distribution may therefore recover the easiest spurious correlations available, weakening robustness under corruption and shift [13, 28, 29]. The teacher's trajectory thus carries a second form of knowledge: not only features worth imitating, but also directions worth rejecting.

We introduce Anti-Shortcut Distillation (ASD) to convert this latent trajectory signal into explicit push–pull supervision (fig. 1). ASD treats the converged teacher T_{final} as a positive semantic anchor and an early-checkpoint teacher T_{early} as a temporal negative reference. For an input x, the displacement $\Delta h(x) = h_{\text{early}}(x) - h_{\text{final}}(x)$ is a sample-dependent proxy for the representation components emphasized by the early state and attenuated by convergence. This sharply separates ASD from prior trajectory-aware distillation [8, 17, 23, 32, 36]: intermediate checkpoints are not additional teachers to imitate, but evidence of representational directions from which the student should be separated.

ASD realizes this temporal signal through two complementary objectives. A temporal contrastive loss $\mathcal{L}_{\text{TC}}$ places the early-teacher feature as a same-sample negative against in-batch and memory-queue final-teacher features in an InfoNCE [26, 31, 33] objective, so the student is pulled toward T_{final} and pushed away from T_{early} for the same input. A shortcut-suppression loss $\mathcal{L}_{\text{SS}}$ penalizes the student's projection onto the leading eigenvectors of the uncentered second-moment matrix of early-to-final feature differences, constraining the representation at the subspace level rather than per sample. Both auxiliary terms are linearly warmed up and combined with cross-entropy and logit distillation; a trainable projector handles mismatched feature dimensions and is discarded at inference.

Contributions. First, we propose a signed KD framework that treats T_{final} as a positive semantic anchor and T_{early} as a temporal negative reference, combining final-teacher imitation with explicit shortcut avoidance through $\mathcal{L}_{\text{TC}}$. Second, we identify $\Delta h = h_{\text{early}} - h_{\text{final}}$ as a label-free shortcut indicator and suppress student projection onto the leading eigenvectors of $\mathbb{E}[\Delta h \Delta h^\top]$ through $\mathcal{L}_{\text{SS}}$. Third, we provide a compact analysis showing that, under a two-subspace decomposition of teacher features, the top eigenspace of $\mathbb{E}[\Delta h \Delta h^\top]$ recovers the shortcut subspace, with a Davis–Kahan [6, 34] perturbation bound on the empirical (theorem 4). Finally, across 13 teacher–student pairs on CIFAR-100, ImageNet-100, and TinyImageNet, ASD attains the best clean accuracy on 10 pairs and surpasses KD on 12; mechanistic diagnostics show ASD students are anti-aligned with Δh ($\cos(h_S, \Delta h) = -0.36$ vs. 0.00 for KD) while increasing robust-subspace projection (0.45 vs. 0.12).

2 Related Work

Response-based KD [11, 16] aligns softened logits; feature/attention methods [27, 35] align hidden activations; relational methods [24, 25] preserve sample geometry; and recent objectives [3, 30, 37] broaden transfer through contrastive, decoupled, or hierarchical alignment. All share an attractive premise. ASD instead uses the teacher to define what should be suppressed, not only imitated. TAKD [23], CkptKD [32], snapshot distillation [17], Born-Again

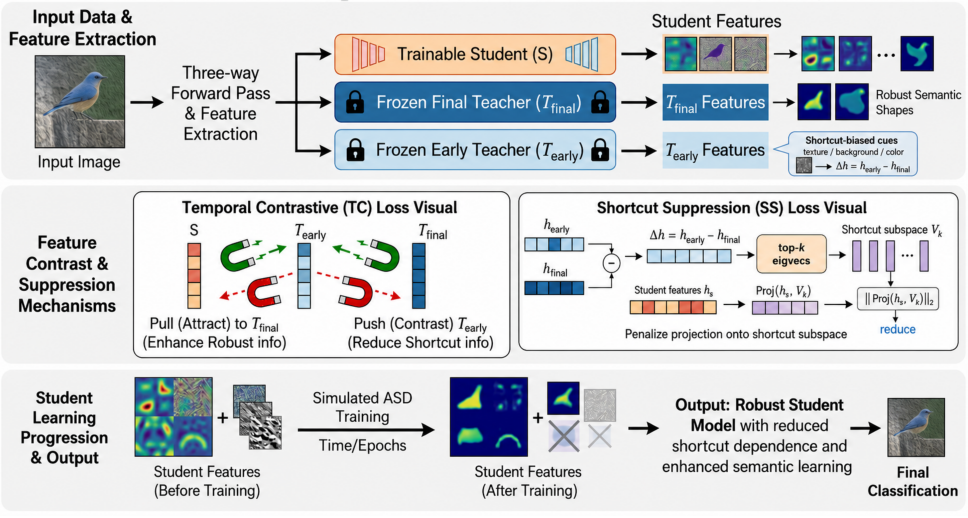


Figure 1: Anti-Shortcut Distillation (ASD) trains a student using two frozen teacher checkpoints: a converged teacher T_{final} providing robust semantic targets and an early teacher T_{early} exposing shortcut-biased cues. $\mathcal{L}_{\text{TC}}$ attracts the student toward T_{final} while repelling it from T_{early}; $\mathcal{L}_{\text{SS}}$ penalizes projection onto the principal shortcut subspace defined by $\Delta h = h_{\text{early}} - h_{\text{final}}$.

Networks [8], and BYOT [36] all recognise that intermediate teacher states contain useful supervision and treat them as positive targets. ASD inverts the sign: T_{early} is a temporal negative that exposes shortcut-prone directions the student should be pushed away from. Lukasik *et al.* [22] show distillation amplifies teacher errors on subgroups; Lee & Lee [20] debias by transplanting the teacher's last layer. Both diagnose when transfer is harmful at the label/head level; ASD operationalizes the harmful directions, extracting them label-free from the early-to-final displacement. Networks fit simple, dominant patterns before more robust structure [1, 29]; in image classification this manifests as texture bias [9, 10], weakening robustness under shift [13, 28]. ASD targets this failure mode at distillation time using the teacher's own early-to-final feature change as a label-free shortcut indicator. SimCLR [4], MoCo [12], and memory-bank methods [33] learn invariances by separating positives from negatives; supervised contrastive learning [18] adds label structure; CRD [30] brings contrast into KD with sample-level negatives. ASD moves the contrastive axis to teacher time: positives and negatives are distinct optimization states of the same teacher.

3 Method

ASD is a training-time distillation framework built around a three-way forward pass: the trainable student, the frozen converged teacher, and a frozen early teacher checkpoint. The final teacher provides the semantic target to imitate, whereas the early teacher exposes shortcut-biased directions that should not be transferred to the student. This section describes how ASD converts this teacher trajectory into two coupled mechanisms: temporal contrastive alignment (section 3.3) and eigenvector-based shortcut suppression (section 3.4).

3.1 Problem Setup

Let $f_T : \mathcal{X} \to \mathbb{R}^C$ be a teacher classifier over C classes, and let $\phi_T(x) \in \mathbb{R}^{d_T}$ denote its penultimate feature for input x. From the teacher training trajectory we retain two frozen snapshots: the final teacher T_{final} and an early checkpoint T_{early}. The student f_S has the same output dimension C, but may use a different architecture with pre-logit feature $\phi_S(x) \in \mathbb{R}^{d_S}$. We define

$$h_{\text{final}}(x) = \phi_{T_{\text{final}}}(x), \qquad h_{\text{early}}(x) = \phi_{T_{\text{early}}}(x). \tag{1}$$

When $d_S \neq d_T$, a trainable projector $g : \mathbb{R}^{d_S} \to \mathbb{R}^{d_T}$ maps student features into the teacher feature space; when $d_S = d_T$, g is the identity. The student feature used by ASD is therefore

$$h_S(x) = g(\phi_S(x)). \tag{2}$$

Both teacher networks remain frozen throughout student training; only the student and, when needed, the projector are optimized. The projector is discarded at inference, so the deployed model is the student alone.

3.2 Temporal Shortcut Signal

Standard KD observes only the final teacher state. ASD instead uses the teacher's temporal change as supervision. Early teacher features tend to retain low-level, shortcut-prone information that the converged teacher later reduces or reorganizes into more stable semantic representations. For each input, ASD measures this change through the early-to-final displacement

$$\Delta h(x) = h_{\text{early}}(x) - h_{\text{final}}(x). \tag{3}$$

This vector is not treated as noise. It is the central anti-shortcut signal: components emphasized by T_{early} but attenuated by T_{final} become candidate directions to suppress during student learning. For a mini-batch $\mathcal{B}$, ASD forms the uncentered second-moment matrix

$$\widehat{\mathbf{C}}_{\text{sc}} = \frac{1}{|\mathcal{B}|} \sum_{x_i \in \mathcal{B}} \Delta h(x_i)\, \Delta h(x_i)^\top \in \mathbb{R}^{d_T \times d_T}. \tag{4}$$

The matrix is deliberately uncentered: the mean displacement may itself encode a dominant shortcut direction, and centering would remove that signal. Let $\mathbf{U}_K = [u_1, \ldots, u_K]$ denote the top-K orthonormal eigenvectors of $\widehat{\mathbf{C}}_{\text{sc}}$. These eigenvectors span the shortcut subspace used by ASD in the current training step. Although $\widehat{\mathbf{C}}_{\text{sc}}$ is rank-deficient at typical batch sizes ($B \ll d_T$), $\mathcal{L}_{\text{SS}}$ uses only the $K{=}4 \ll B$ dimensional subspace $\text{span}(\mathbf{U}_K)$ and is invariant to rotations within it; accuracy and robustness are insensitive to $B \in \{32, \ldots, 256\}$, while subspace agreement improves with B at the rate predicted by section 4.3 (Supplementary Sec. A.4.5). The same temporal displacement now serves two roles: it defines a negative teacher state for contrastive learning (section 3.3) and a geometric subspace for shortcut suppression (section 3.4).

3.3 Temporal Contrastive Loss

The Temporal Contrastive Loss makes the student's representation closer to the final teacher than to the early teacher for the same input. For a mini-batch $\{x_i\}_{i=1}^B$, let z_S^i, z_F^i, and z_E^i be the ℓ_2-normalized versions of $h_S(x_i)$, $h_{\text{final}}(x_i)$, and $h_{\text{early}}(x_i)$, respectively. ASD treats z_F^i as the positive target for z_S^i and z_E^i as the temporal negative. Other final-teacher features inside the

mini-batch and a FIFO memory queue $\mathcal{M} = \{q_m\}_{m=1}^{M}$ of normalized T_{final} features provide additional negatives. The loss for sample x_i is

$$\mathcal{L}_{\text{TC}}^{(i)} = -\log \frac{\exp\left(z_S^i \cdot z_F^i / \tau_c\right)}{D_i}, \tag{5}$$

where the denominator D_i assembles one positive and three types of negatives,

$$\begin{aligned} D_i = & \underbrace{\exp\left(z_S^i \cdot z_F^i / \tau_c\right)}_{\text{positive}} + \underbrace{\exp\left(z_S^i \cdot z_E^i / \tau_c\right)}_{\text{temporal negative}} \\ & + \underbrace{\sum_{j \neq i} \exp\left(z_S^i \cdot z_F^j / \tau_c\right)}_{\text{in-batch final-teacher}} + \underbrace{\sum_{m=1}^{M} \exp\left(z_S^i \cdot q_m / \tau_c\right)}_{\text{memory-bank final-teacher}}. \end{aligned} \tag{6}$$

The positive term pulls the student toward the converged teacher; the same-sample temporal negative makes the early shortcut state explicitly competitive; the in-batch and memory-bank terms prevent the contrastive problem from degenerating into a binary comparison and provide a large negative set required for an informative InfoNCE bound [26, 31]. Replacing the temporal negative with random in-batch negatives recovers a CRD-style objective; the resulting ablation (section 6.3) isolates the contribution of T_{early}.

3.4 Shortcut Suppression Loss

Temporal contrast reduces similarity to T_{early}, but it does not directly constrain the subspace in which student features may lie. ASD therefore adds a geometric suppression term. Given the shortcut eigenspace $\mathbf{U}_K$, the normalized student feature is projected onto this subspace and penalized when the projection magnitude exceeds a margin $\varepsilon > 0$:

$$\mathcal{L}_{\text{SS}} = \frac{1}{B} \sum_{i=1}^{B} \max\left(0, \left\| \bar{h}_S(x_i)^\top \mathbf{U}_K \right\|_2 - \varepsilon\right), \qquad \text{where} \quad \bar{h}_S(x_i) = \frac{h_S(x_i)}{\|h_S(x_i)\|_2}. \tag{7}$$

Here $\left\|\bar{h}_S^\top \mathbf{U}_K\right\|_2 \in [0,1]$ is the (unsigned) magnitude of the student feature's projection onto $\text{span}(\mathbf{U}_K)$; the hinge is inactive below ε and grows linearly above it. We use $\varepsilon = 0.1$ throughout. We stress that this projection margin is distinct from the signed-cosine threshold $\varepsilon_{\text{diag}} = -0.1$ used only as a diagnostic in section 6.4; the former constrains a subspace projection during training, the latter measures signed alignment with a single direction at evaluation. While $\mathcal{L}_{\text{TC}}$ supplies a relative push away from the early teacher at the sample level, $\mathcal{L}_{\text{SS}}$ imposes an explicit eigenspace constraint at the batch level. Together they prevent the student from matching the final teacher while silently recovering the same shortcuts from data: $\mathcal{L}_{\text{TC}}$ alone cannot forbid the student from re-discovering shortcut directions, and $\mathcal{L}_{\text{SS}}$ alone carries no per-sample temporal signal.

3.5 Projector for Heterogeneous Architectures

For heterogeneous teacher–student pairs ($d_S \neq d_T$), direct feature comparison is undefined. We use the trainable projector

$$g = \text{Linear}(d_S, d_T) \to \text{BN}(d_T) \to \text{ReLU} \to \text{Linear}(d_T, d_T). \tag{8}$$

The projector is optimized jointly with the student and used only to compute feature-level ASD losses; it is removed at inference, keeping the deployed model architecture-unchanged.

3.6 Training Objective

ASD combines task supervision, final-teacher logit distillation, temporal contrast, and shortcut-subspace suppression:

$$\mathcal{L}_{\mathrm{ASD}} = \mathcal{L}_{\mathrm{CE}} + \alpha_{\mathrm{kd}}\mathcal{L}_{\mathrm{KD}} + \beta(t)\,\alpha_{\mathrm{tc}}\mathcal{L}_{\mathrm{TC}} + \gamma(t)\,\alpha_{\mathrm{ss}}\mathcal{L}_{\mathrm{SS}}. \tag{9}$$

Here $\mathcal{L}_{\mathrm{CE}}$ is cross-entropy and $\mathcal{L}_{\mathrm{KD}}$ is temperature-scaled logit distillation [16] from T_{final}. Because the student representation is unstable early in training, ASD warms up the two auxiliary terms linearly:

$$\beta(t) = \gamma(t) = \min\left(\frac{t}{T_{\mathrm{warmup}}}, 1\right). \tag{10}$$

The resulting objective is a push–pull distillation process: the student is pulled toward the final teacher's semantic representation while being pushed away from shortcut directions exposed by the early teacher.

4 Theoretical Analysis

We give a compact analysis of ASD at the population level and defer extended proofs to the supplementary material. The aim is not to make hard generalization guarantees, but to give a precise meaning to the two losses and to the role of Δh as a shortcut indicator. We state one information-theoretic interpretation of $\mathcal{L}_{\mathrm{TC}}$ (theorem 1), one geometric reading of $\mathcal{L}_{\mathrm{SS}}$ (theorem 2), and one structural theorem (theorem 4) that justifies Δh as a shortcut indicator.

4.1 Information-Theoretic Interpretation of $\mathcal{L}_{\mathrm{TC}}$

The InfoNCE objective lower-bounds the mutual information between the two endpoints of the positive pair, with a $\log N$ saturation term depending on the number of marginal negatives [26, 31]. We give the version of this bound that applies to $\mathcal{L}_{\mathrm{TC}}$, with the temporal negative correctly handled.

Proposition 1 (InfoNCE alignment with T_{final})**.** *Let $\widetilde{\mathcal{L}_{\mathrm{TC}}}$ denote the variant of $\mathcal{L}_{\mathrm{TC}}$ in which the same-sample temporal-negative term $\exp(z_S^i \cdot z_E^i/\tau_c)$ is removed from the denominator of eq. (6), leaving only the positive and the $|\mathcal{N}_F| = (B-1)+M$ in-batch and memory-bank* final-teacher *negatives, drawn i.i.d. from the marginal $p_{h_{\mathrm{final}}}$. Then*

$$I(h_S; h_{\mathrm{final}}) \geq \log\big(|\mathcal{N}_F|+1\big) - \mathbb{E}[\widetilde{\mathcal{L}_{\mathrm{TC}}}]. \tag{11}$$

Furthermore, $\mathcal{L}_{\mathrm{TC}} \geq \widetilde{\mathcal{L}_{\mathrm{TC}}}$ pointwise, so $\mathbb{E}[\widetilde{\mathcal{L}_{\mathrm{TC}}}] \leq \mathbb{E}[\mathcal{L}_{\mathrm{TC}}]$ and the same bound is preserved when $\mathcal{L}_{\mathrm{TC}}$ replaces $\widetilde{\mathcal{L}_{\mathrm{TC}}}$.

Proof sketch. Equation (11) is the standard InfoNCE bound [26, 31] applied with positive pair $(h_S, h_{\mathrm{final}})$ and negatives drawn i.i.d. from $p_{h_{\mathrm{final}}}$. Adding an extra positive term to the denominator of a softmax can only decrease the log-probability of the positive class, so $\mathcal{L}_{\mathrm{TC}} \geq \widetilde{\mathcal{L}_{\mathrm{TC}}}$. The full proof and a discussion of bound tightness are in the supplementary. □

The proposition gives the temporal negative an operational reading: it does not change the MI bound on T_{final} (the bound is controlled by the marginal T_{final} negatives), but it adds an explicit penalty on similarity to T_{early} *at no cost* to the bound. This is precisely how ASD differs from CRD [30]: the same-sample temporal negative changes the geometry of the contrastive decision, not merely the count of negatives.

4.2 Geometric Interpretation of $\mathcal{L}_{\text{SS}}$

Proposition 2 (Angular margin from shortcut suppression)**.** *Let* $\theta(x) \in [0, \frac{\pi}{2}]$ *be the angle between* $\bar{h}_S(x)$ *and its orthogonal projection onto* $\text{span}(\mathbf{U}_K)$*, i.e. the angle to the nearest point of the subspace. Since* $\bar{h}_S(x)$ *is unit-norm and* $\mathbf{U}_K$ *has orthonormal columns,* $\left\|\bar{h}_S(x)^\top \mathbf{U}_K\right\|_2 = \cos\theta(x)$ *exactly. Consequently, if* $\mathcal{L}_{\text{SS}} = 0$ *on a sample* x *then* $\cos\theta(x) \leq \varepsilon$*, equivalently* $\theta(x) \geq \arccos(\varepsilon)$*. With* $\varepsilon = 0.1$ *this enforces* $\theta(x) \geq 84.3°$*.*

The proposition gives $\mathcal{L}_{\text{SS}}$ a direct geometric reading: when it vanishes, the student feature is separated from *every* direction in the shortcut subspace by at least $\arccos(\varepsilon)$, a strictly stronger constraint than the relative push provided by $\mathcal{L}_{\text{TC}}$ (section 6.1).

4.3 When Does Δh Identify the Shortcut Subspace?

We formalize when the top eigenspace of $\mathbb{E}[\Delta h \Delta h^\top]$ identifies a meaningful shortcut subspace through a two-subspace decomposition of teacher features.

Assumption 3 (Two-subspace decomposition)**.** Teacher features at any state $t \in \{\text{early}, \text{final}\}$ admit the decomposition

$$h_t(x) \;=\; r_t(x) + s_t(x), \qquad r_t(x) \in \mathcal{R},\;\; s_t(x) \in \mathcal{S}, \tag{12}$$

where $\mathcal{R}$ and $\mathcal{S}$ are orthogonal complementary subspaces of $\mathbb{R}^{d_T}$.

Theorem 4 (Shortcut subspace identification)**.** *Under assumption 3, let* $G_r = \mathbb{E}[(r_{\text{early}} - r_{\text{final}})$
$(r_{\text{early}} - r_{\text{final}})^\top]$ *and* $G_s = \mathbb{E}[(s_{\text{early}} - s_{\text{final}})(s_{\text{early}} - s_{\text{final}})^\top]$*. If the shortcut component dominates,*

$$\lambda_K(G_s) \;>\; \|G_r\|_2\,, \tag{13}$$

then the top-K eigenspace of $\mathbb{E}[\Delta h \Delta h^\top]$ *coincides with the top-K eigenspace of* G_s*, which lies entirely in* $\mathcal{S}$*. Consequently, minimizing* $\mathcal{L}_{\text{SS}}$ *reduces student projection onto* $\mathcal{S}$*.*

Proof sketch. Under $\mathcal{R} \perp \mathcal{S}$, the cross terms in $\mathbb{E}[\Delta h \Delta h^\top]$ vanish, so $\mathbb{E}[\Delta h \Delta h^\top] = G_r \oplus G_s$ is block-diagonal in the basis $\mathcal{R} \cup \mathcal{S}$. Eigenvectors of a block-diagonal symmetric matrix are contained in the individual blocks; eq. (13) forces the top-K eigenvalues to come from G_s, so the top-K eigenspace lies in $\mathcal{S}$. For the empirical $\widehat{\mathbf{C}}_{\text{sc}}$, Davis–Kahan sin-$\Theta$ [5, 34] bounds the deviation from the population subspace by $\left\|\mathbf{C}_{\text{sc}} - \widehat{\mathbf{C}}_{\text{sc}}\right\|_2 / (\lambda_K(G_s) - \|G_r\|_2)$. The full proof is in the supplementary. □

Interpretation. Equation (13) states that between early and final teacher states the dominant change in feature variance occurs in the shortcut subspace, not the robust subspace. This is the empirical regime that motivates ASD, and we verify it directly: the principal angle between the shortcut subspace and the T_{final}-PCA subspace is 65.3° against 38.6° to the T_{early}-PCA subspace (section 6.4). Across the teacher trajectory (Supplementary Fig. A.6), shortcut energy in $\text{span}(\mathbf{U}_K)$ peaks at T_{early} (0.25) and drops by ∼49% at convergence (0.13), while robust T_{final}-PCA energy remains stable (∼0.11–0.13), supporting the premise of Δh.

4.4 Robustness under Shortcut-Aligned Shifts

If $\mathcal{L}_{\text{SS}}$ is small, the student is confined to a region whose projection onto $\text{span}(\mathbf{U}_K)$ is at most ε. For a shift that perturbs features within $\text{span}(\mathbf{U}_K)$ and a locally-linear classifier head, the

first-order change in classifier output is bounded by this projection. ASD therefore exhibits reduced first-order sensitivity to shortcut-aligned shifts compared with a KD student of larger shortcut projection. The linearity and shift-support conditions are restrictive; we treat this as a design hypothesis and evaluate it empirically in sections 5.6 and 6.4.

5 Experiments

5.1 Experimental Setup

Datasets and shortcut benchmarks. CIFAR-100 [19] provides the main controlled distillation setting (50K/10K train/test, 100 classes). CIFAR-100-C [13] contains 15 corruption types across 5 severity levels and is used only for robustness evaluation. ImageNet-100 (a 100-class subset of ImageNet [6]) evaluates ASD at larger image resolution and data scale. TinyImageNet stresses the method with 200 classes, 100K/10K images, and 64×64 inputs. Corruptions alone are not shortcut benchmarks, so we additionally evaluate the trained ImageNet-100 students (ResNet-50→MobileNet-V2), without retraining, on *Stylized-ImageNet-100* (AdaIN [9], full 100-class val set), *ImageNet-R* [14] (19 overlapping classes), and *ImageNet-A* [15] (15-class overlap, 680 images); protocol and per-seed results in Supplementary Sec. A.3.9.

Baselines. We compare with CE-only, KD [16], FitNets [27], AT [35], DKD [37], CRD [30], and CkptKD [32], covering logit transfer, feature regression, attention alignment, decoupled logits, contrastive representation distillation, and checkpoint reuse. All methods share the optimizer, schedule, augmentation, and training budget.

Training details. CIFAR-100: 240 epochs, SGD (momentum 0.9, weight decay 5×10^{-4}), initial LR 0.05, cosine schedule, batch 64. ImageNet-100: 100 epochs, LR 0.1, batch 256. TinyImageNet: 200 epochs. T_{early} is the 15%-of-training checkpoint on every dataset. ASD hyperparameters: $\alpha_{\text{kd}} = 1.0$, $\alpha_{\text{tc}} = 0.8$, $\alpha_{\text{ss}} = 1.0$, $\tau_{\text{kd}} = 4.0$, $\tau_c = 0.07$, $\varepsilon = 0.1$, $K = 4$, $M = 4096$, $T_{\text{warmup}} = 20$ epochs. Robustness is mean Corruption Error (mCE) on CIFAR-100-C [13] across all 15 corruption types at severity 5. Results are mean±std over 3 seeds unless noted.

5.2 Same-Family CIFAR-100 Results

Teacher Student	ResNet-34 ResNet-18	WRN-40-2 WRN-16-2	ResNet-32×4 ResNet-8×4	WRN-40-2 WRN-40-1	VGG-13 VGG-8
Teacher	79.91	76.49	79.72	76.49	76.15
Student	78.79	73.33	71.90	71.49	71.10
KD [16]	80.22 ± 0.19 (↓)	74.74 ± 0.17 (↓)	73.32 ± 0.11 (↓)	73.38 ± 0.47 (↓)	73.68 ± 0.35 (↓)
FitNets [27]	80.32 ± 0.17 (↑)	74.94 ± 0.24 (↑)	73.55 ± 0.04 (↑)	73.54 ± 0.18 (↑)	73.91 ± 0.11 (↑)
AT [35]	80.30 ± 0.12 (↑)	74.99 ± 0.15 (↑)	**73.78 ± 0.15** (↑)	73.79 ± 0.16 (↑)	71.61 ± 0.12 (↓)
DKD [37]	80.04 ± 0.36 (↓)	75.08 ± 0.14 (↑)	73.69 ± 0.32 (↑)	73.63 ± 0.07 (↑)	73.99 ± 0.05 (↑)
CRD [30]	80.71 ± 0.34 (↑)	74.76 ± 0.12 (↑)	73.15 ± 0.12 (↓)	73.63 ± 0.33 (↑)	74.04 ± 0.18 (↑)
CkptKD [32]	73.94 ± 0.13 (↓)	71.31 ± 0.06 (↓)	72.17 ± 0.15 (↓)	70.65 ± 0.15 (↓)	70.48 ± 0.12 (↓)
ASD (ours)	**80.50 ± 0.07** (↑)	**75.87 ± 0.29** (↑)	73.13 ± 0.12 (↓)	**74.16 ± 0.21** (↑)	**74.26 ± 0.13** (↑)

Table 1: Top-1 accuracy (%) on CIFAR-100 for same-family teacher–student pairs. P1–P5 denote the five teacher–student pairs shown in the header. ↑ / ↓ indicates better/worse performance than KD. Grey bold marks the best result per column. Results are mean ± std over 3 seeds.

Same-family pairs isolate capacity compression under shared architectural bias (table 1). ASD improves over KD on 4 of 5 pairs (P1, P2, P4, P5) and obtains the best accuracy on P2 (75.87%), P4 (74.16%) and P5 (74.26%). On P1, CRD edges ASD by 0.21 pp on clean accuracy, but with nearly 5× larger seed variance (±0.34 vs. ±0.07); we will see in section 5.6 that ASD's robustness on P1 is comparable to CRD's while preserving lower seed-level variability.

The single same-family failure is P3 (ResNet-32×4→ResNet-8×4), where ASD reaches 73.13%, slightly below KD (73.32%) and AT (73.78%). This is the most aggressively compressed student in our sweep. We discuss the capacity boundary that this case represents in section 5.7.

5.3 Cross-Architecture CIFAR-100 Results

Teacher Student	WRN-40-2 ShuffleNet-V2	WRN-40-2 ShuffleNet-V1	ResNet-32×4 ShuffleNet-V1	ResNet-32×4 ShuffleNet-V2	VGG-13 MobileNet-V2
Teacher Student	76.49 63.26	76.49 71.46	79.72 71.46	79.72 63.26	76.15 64.12
CE only	63.26 ± 0.27	71.46 ± 0.38	71.46 ± 0.38	63.26 ± 0.27	64.12 ± 0.31
KD [16]	66.72 ± 0.32 (↓)	73.79 ± 0.26 (↓)	73.09 ± 0.38 (↓)	66.17 ± 0.18 (↓)	65.86 ± 0.28 (↓)
FitNets [27]	67.07 ± 0.15 (↑)	74.27 ± 0.04 (↑)	73.20 ± 0.50 (↑)	66.07 ± 0.36 (↓)	66.18 ± 0.22 (↑)
AT [35]	67.66 ± 0.19 (↑)	73.97 ± 0.21 (↑)	73.25 ± 0.28 (↑)	66.28 ± 0.24 (↑)	66.38 ± 0.19 (↑)
DKD [37]	66.84 ± 0.31 (↑)	73.63 ± 0.39 (↓)	72.33 ± 0.33 (↓)	66.42 ± 0.28 (↑)	66.02 ± 0.24 (↑)
CRD [30]	67.81 ± 0.42 (↑)	74.34 ± 0.47 (↑)	73.60 ± 0.16 (↑)	66.98 ± 0.36 (↑)	66.64 ± 0.31 (↑)
CkptKD [32]	67.27 ± 0.03 (↑)	72.66 ± 0.20 (↓)	73.39 ± 0.01 (↑)	66.52 ± 0.08 (↑)	66.08 ± 0.14 (↑)
ASD (ours)	**68.05 ± 0.14** (↑)	**74.94 ± 0.09** (↑)	**73.98 ± 0.27** (↑)	**67.18 ± 0.19** (↑)	**67.02 ± 0.18** (↑)

Table 2: Top-1 accuracy (%) on CIFAR-100 for cross-architecture teacher–student pairs. P6–P10 denote the five pairs shown in the header. ↑/↓ indicates better/worse performance than KD. Grey bold marks the best result per column. Results are mean ± std over 3 seeds.

ASD attains the best accuracy on *all* five heterogeneous pairs (table 2), with the strongest absolute gain over KD on WRN-40-2→ShuffleNet-V2 (+1.33 pp). This is the regime where push–pull supervision matters most: when teacher and student have different representational bases, purely attractive transfer leaves the student free to find architecture-specific shortcuts that reproduce teacher predictions without recovering the teacher's robust geometry. ASD removes this ambiguity by explicitly designating directions the student should not occupy. ASD also exceeds CRD on every cross-architecture pair, including the difficult WRN-40-2→ShuffleNet-V2 case where it improves over CRD by 0.24 pp; sample-level contrastive negatives alone are not sufficient in this regime.

5.4 Beyond CIFAR-100

Table 3 tests whether the temporal shortcut signal remains useful at higher resolution and class count. ASD obtains the best accuracy on all 3 pairs (83.24% on P11, 46.00% on P12, 85.48% on P13). The TinyImageNet result has notably larger seed variance (±1.45) than the CIFAR-100 results, reflecting a less stable optimization landscape on this dataset. The mean is still highest, and ASD wins on 3/3 seeds against CRD with paired RNG initialization (per-seed margins +0.62, +1.14, +1.18 pp; see supplementary for the seed-level analysis).

5.5 Semantic Segmentation on ADE20K

To test whether the temporal shortcut signal transfers beyond classification, we evaluate ASD on dense prediction. Both ASD losses act on penultimate features, so they extend to

Teacher → Student Dataset	ResNet-50 → MobileNet-V2 ImageNet-100	ResNet-34 → ResNet-18 TinyImageNet	ResNet-50 → ResNet-18 ImageNet-100
Teacher	85.30	43.14	85.30
Student	80.21	42.98	83.85
CE only	80.21 ± 0.07	42.98 ± 0.46	83.85 ± 0.15
KD [16]	82.43 ± 0.41 (↓)	44.36 ± 1.23 (↓)	84.72 ± 0.35 (↓)
FitNets [27]	82.51 ± 0.30 (↑)	44.58 ± 0.83 (↑)	84.88 ± 0.22 (↑)
AT [35]	82.90 ± 0.30 (↑)	44.42 ± 0.91 (↑)	84.92 ± 0.26 (↑)
DKD [47]	82.54 ± 0.18 (↑)	45.44 ± 1.11† (↑)	84.85 ± 0.18 (↑)
CRD [30]	83.01 ± 0.18 (↑)	45.02 ± 1.08 (↑)	85.24 ± 0.19 (↑)
CkptKD [53]	82.70 ± 0.30 (↑)	44.75 ± 1.02 (↑)	84.80 ± 0.28 (↑)
ASD (ours)	**83.24 ± 0.13** (↑)	**46.00 ± 1.45** (↑)	**85.48 ± 0.15** (↑)

Table 3: Top-1 accuracy (%) on ImageNet-100 and TinyImageNet. P11–P13 denote the teacher–student pairs shown in the header. ↑/↓ indicates above/below KD. Grey bold marks the best result per column. Results are mean ± std over 3 seeds.

Teacher	Student	From Scratch		KD [16]		Af-DCD [7]		UFD-KD [21]	**ASD (ours)**	
		T. mIoU	S. mIoU	pixAcc	mIoU	pixAcc	mIoU	mIoU	pixAcc	mIoU
ResNet-101	ResNet-18	42.70	33.91	76.32	34.88	77.43	36.21	36.89	78.92	**38.87**
ResNet-101	MobileNetV2	42.70	34.32	76.74	34.92	77.28	35.97	36.24	78.65	**37.91**

Table 4: ADE20K semantic segmentation with DeepLabv3. Results are reported as pixAcc/mIoU (%) using 512×512 inputs and 40k iterations. From-Scratch reports teacher/student mIoU; KD, Af-DCD [7], and UFD-KD [21] follow [21]. Bold denotes best mIoU per row.

segmentation without modification: we apply $\mathcal{L}_{\text{TC}}$ and $\mathcal{L}_{\text{SS}}$ to the backbone feature map of a DeepLabv3 [2] network, treating each spatial location as a sample when forming the temporal displacement $\Delta h = h_{\text{early}} - h_{\text{final}}$ and the uncentered shortcut covariance. We follow the protocol of [21]: a ResNet-101 teacher distilled into ResNet-18 and MobileNetV2 students, 512×512 inputs, 40k iterations, SGD with momentum 0.9 and learning rate 0.02, reporting pixel accuracy and mean IoU on the ADE20K [38] validation set.

Table 4 reports the results. ASD improves both students over the KD baseline, the dense contrastive method Af-DCD [7], and the recent frequency-decoupling method UFD-KD [21], reaching 38.87 mIoU for ResNet-18 (+1.98 over UFD-KD, +3.99 over the from-scratch student) and 37.91 mIoU for MobileNetV2 (+1.67 over UFD-KD). The gain is consistent with the mechanistic picture from classification: dense prediction is especially exposed to texture and local-statistic shortcuts, so suppressing the early-teacher displacement subspace and contrasting against the early teacher state transfers the same shape-biased, shortcut-resistant structure to the per-pixel features. ASD requires no segmentation-specific component beyond pooling the two frozen teacher checkpoints' feature maps, which keeps the added cost in line with the classification setting. t-SNE and qualitative mask visualizations (Supplementary Section A.10) corroborate these gains: the ASD student forms tighter, better-separated per-class clusters and recovers sharper object boundaries than the KD student.

5.6 Corruption Robustness

Robustness (table 5) tests whether reduced shortcut projection translates into reduced sensitivity to corrupted local statistics. ASD obtains the lowest mCE on P4 (91.7), P5 (85.2), and the most challenging cross-architecture pair P6 (86.1). These are pairs where ASD also

Teacher Student	ResNet-34 ResNet-18	WRN-40-2 WRN-16-2	ResNet-32×4 ResNet-8×4	WRN-40-2 WRN-40-1	VGG-13 VGG-8	WRN-40-2 ShuffleNet-V2
CE only	87.0 ± 0.3 (↑)	95.9 ± 0.2 (↑)	99.6 ± 0.4 (↑)	93.7 ± 0.4 (↑)	87.0 ± 0.3 (↑)	88.5 ± 0.4 (↑)
KD [16]	86.2 ± 0.4 (↑)	93.7 ± 0.4 (↑)	98.1 ± 0.3 (↑)	92.2 ± 0.6 (↑)	85.6 ± 0.3 (↑)	86.9 ± 0.4 (↑)
FitNets [27]	86.0 ± 0.4 (↓)	93.6 ± 0.1 (↓)	98.0 ± 0.3 (↓)	92.1 ± 0.3 (↓)	85.5 ± 0.2 (↓)	86.8 ± 0.3 (↓)
AT [35]	85.7 ± 0.6 (↓)	93.6 ± 0.5 (↓)	97.8 ± 0.5 (↓)	92.0 ± 0.4 (↓)	85.5 ± 0.3 (↓)	87.4 ± 0.2 (↑)
DKD [37]	85.9 ± 0.9 (↓)	**93.4 ± 0.3** (↓)	97.5 ± 0.2 (↓)	92.4 ± 0.3 (↑)	85.5 ± 0.1 (↓)	87.4 ± 0.5 (↑)
CRD [30]	**85.2 ± 0.6** (↓)	93.8 ± 0.3 (↑)	98.2 ± 0.1 (↑)	92.7 ± 0.6 (↑)	85.9 ± 0.5 (↑)	86.7 ± 0.3 (↓)
CkptKD [34]	90.9 ± 0.1 (↑)	99.1 ± 0.2 (↑)	**96.6 ± 0.2** (↓)	98.7 ± 0.2 (↑)	90.0 ± 0.2 (↑)	92.2 ± 0.1 (↑)
ASD (ours)	85.6 ± 0.4 (↓)	94.2 ± 0.4 (↑)	97.9 ± 0.2 (↓)	**91.7 ± 0.4** (↓)	**85.2 ± 0.3** (↓)	**86.1 ± 0.4** (↓)

Table 5: Mean Corruption Error (mCE, %) on CIFAR-100-C at severity 5. Lower is better. ↓/↑ indicates improvement/degradation vs. KD. Grey bold marks the lowest mCE per column. Results are mean ± std over 3 seeds.

performs strongly on clean accuracy, suggesting that shortcut suppression does not simply trade accuracy for robustness; when the student has sufficient capacity, it redirects the representation toward features that improve both.

P1: narrower trade-off. On P1, CRD has the lowest mCE (85.2); ASD reaches 85.6, still improving over KD (86.2). The two reduce error differently: CRD strengthens instance-level geometry, ASD penalizes temporal shortcut directions.

Structural failure of CkptKD. CkptKD degrades robustness severely on five of six pairs (P1: 86.2 → 90.9; P2: 93.7 → 99.1; P6: 86.9 → 92.2). The reason is structural: CkptKD treats an early checkpoint as an *attractive* teacher, codifying its texture- and background-biased features. ASD uses the same source in the opposite direction and obtains the opposite effect.

Honest negative: P2. On WRN-40-2→WRN-16-2, ASD reaches 94.2 mCE, above KD (93.7) and DKD (93.4): the dominant error mode here is class-boundary calibration rather than shortcut reliance, where DKD's logit separation is more effective (per-category decomposition in Supplementary Sec. A.9.2). We report this without exclusion.

5.7 The P3 Capacity Boundary

P3 (ResNet-32×4→ResNet-8×4) is the only same-family pair where ASD trails KD on clean accuracy, and the only pair where the negative-source ablation (section 6.3) prefers in-batch negatives. ResNet-8×4 has 6× fewer parameters than its teacher, and for such aggressive compression some low-level cues function as necessary discriminative scaffolding rather than dispensable shortcuts. This is the practical boundary of ASD's applicability: shortcut suppression is most reliable when the student can *replace* shortcut cues with semantic structure. Consistently, CkptKD which encourages retention of early-teacher features has its only robustness improvement on P3, (Supplementary Sec. A.9.1).

6 Ablation Studies

6.1 Loss Components

Table 6 decomposes ASD into temporal contrast and shortcut suppression on P1. Removing $\mathcal{L}_{\text{SS}}$ (leaving $\mathcal{L}_{\text{TC}}$ only) gives 80.19% clean accuracy with 85.8 mCE; removing $\mathcal{L}_{\text{TC}}$ (leaving $\mathcal{L}_{\text{SS}}$ only) gives 80.15% and 85.9 mCE; the full model reaches 80.50% and 85.6 mCE. The pattern supports the intended division of labour: $\mathcal{L}_{\text{TC}}$ mainly improves discriminative transfer by making the final teacher more probable than the early teacher in representation

Method	Clean Acc (%) ↑	mCE (%) ↓
KD [16]	80.22 ± 0.19 (↓)	86.2 ± 0.4 (↑)
ASD w/o $\mathcal{L}_{SS}$	80.19 ± 0.07 (↓)	85.8 ± 0.7 (↓)
ASD w/o $\mathcal{L}_{TC}$	80.15 ± 0.19 (↓)	85.9 ± 0.4 (↓)
ASD (ours)	**80.50 ± 0.07** (↑)	**85.6 ± 0.4** (↓)

Table 6: Loss ablation on P1. Removing $\mathcal{L}_{SS}$ keeps only temporal contrast, while removing $\mathcal{L}_{TC}$ keeps only shortcut suppression.

T_{early} epoch	% training	Clean Acc (%)
12	5%	80.25
24	10%	80.26
36	15%	**80.50 ± 0.07** ← default
48	20%	80.16
72	30%	80.56

Table 7: Sensitivity of ASD to the early checkpoint epoch T_{early} on P1. Epoch 36 is adopted as the default setting.

space, while $\mathcal{L}_{SS}$ acts more directly on robustness by constraining the student's projection onto shortcut eigendirections. Either component alone improves part of the objective, but neither fully reconstructs the push–pull geometry. Their combination is not redundant: one term defines the temporal contrast, the other enforces a geometric exclusion region.

6.2 T_{early} Sensitivity

Table 7 shows ASD is stable across the 10%–30% early-to-middle range: epoch 36 (fixed 15% default) gives 80.50 ± 0.07, epoch 72 gives 80.56%, and only epoch 12 is weaker. This stability makes the default practical: T_{early} is a zero-cost byproduct of training T_{final} (Supplementary Sec. A.8), and if only T_{final} is available, a 10-epoch high-LR fine-tune gives an early-like reference (80.18% on P1), close to KD (80.22%) and the true early checkpoint (80.27%; Supplementary Sec. A.4.6).

6.3 Negative Source for $\mathcal{L}_{\text{TC}}$

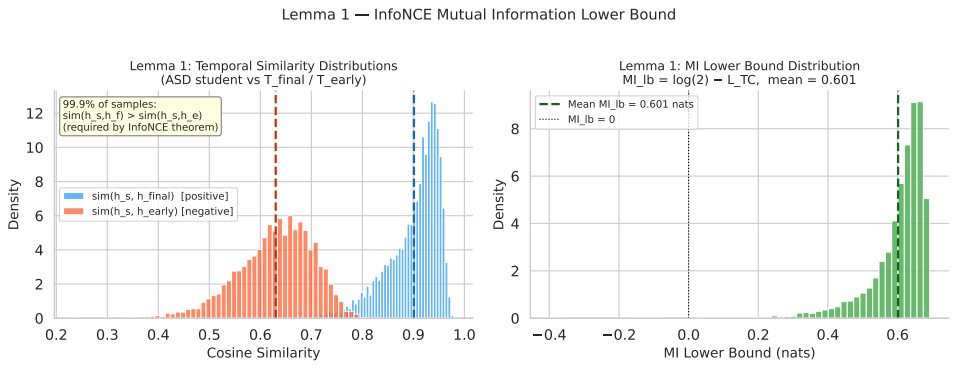
Negative Source	Objective	ResNet-34 ResNet-18	WRN-40-2 WRN-16-2	ResNet-32×4 ResNet-8×4	WRN-40-2 ShuffleNet-V2	VGG-13 VGG-8
In-batch ($i \neq j$)	CE + KD	80.22 ± 0.19	74.74 ± 0.17	**73.32 ± 0.11**	66.72 ± 0.32	73.68 ± 0.35
T_{early} (ours)	ASD	**80.50 ± 0.07**	**75.87 ± 0.29**	73.13 ± 0.12	**68.05 ± 0.14**	**74.26 ± 0.13**

Table 8: Negative-source ablation for $\mathcal{L}_{\text{TC}}$ on five CIFAR-100 pairs. T_{early} provides structured temporal negatives and improves 4/5 pairs; P3 reflects low shortcut reliance. Grey bold marks the best result per column. Results are mean ± std over 3 seeds.

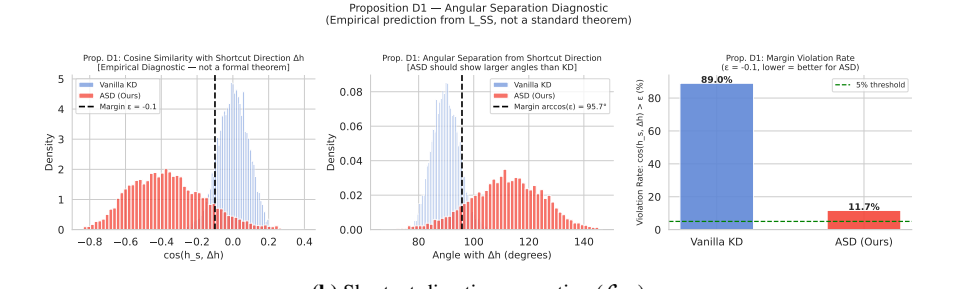
Table 8 most directly defends the novelty claim against CRD. Replacing the same-sample temporal negative with random in-batch negatives reduces $\mathcal{L}_{\text{TC}}$ to a CRD-style objective and removes the teacher trajectory signal entirely. On 4 of 5 pairs, using T_{early} as the structured negative outperforms in-batch negatives, with the largest gain on the cross-architecture pair WRN-40-2→ShuffleNet-V2 (+1.33 pp). The exception is again P3, consistent with the capacity boundary discussed in section 5.7. The pattern supports the interpretation that cross-architecture students benefit most from knowing not only what to imitate, but also which teacher state to avoid.

6.4 Mechanistic Validation

We verify that the learned student actually follows the geometry the method prescribes. All diagnostics use P1 (ResNet-34→ResNet-18 on CIFAR-100); full procedures, bootstrap stability, and the texture-shift protocol are in Supplementary Sec. A.3. Figure 2 and fig. 3 consolidate the four core diagnostics. The mechanism is not pair-specific: Supplementary Sec. A.3.7 repeats the diagnostics on P5, P3, and P12, in addition to P1. ASD reproduces



(a) Temporal alignment ($\mathcal{L}_{\text{TC}}$)



(b) Shortcut-direction separation ($\mathcal{L}_{\text{SS}}$)

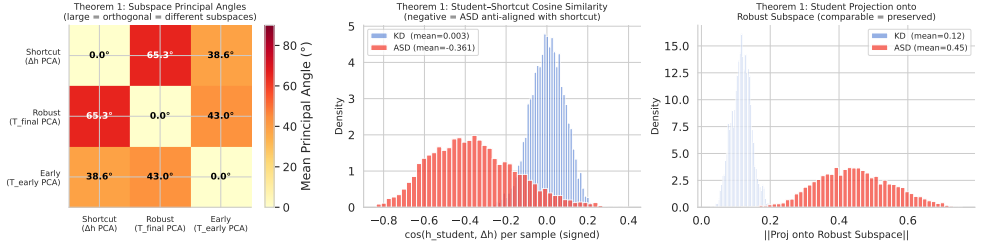
Figure 2: *Alignment and shortcut-direction diagnostics on P1.* ASD aligns the student with T_{final} over T_{early} for 99.9% of samples with an InfoNCE bound of 0.601 nats, and produces a negative shortcut cosine compared with KD (-0.361 vs. 0.003).

the same signature on every pair: higher T_{final} than T_{early} alignment (e.g., $0.90/0.61$ on P1 and $0.52/0.34$ on P12) and reduced shortcut magnitude (-16 to -28%), with CIFAR-100-C gains concentrated on texture/local-statistic corruptions.

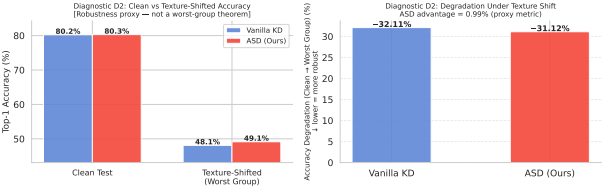
Temporal alignment ($\mathcal{L}_{\text{TC}}$ diagnostic, fig. 2a): At convergence, 99.9% of samples satisfy $\text{sim}(h_S, h_{\text{final}}) > \text{sim}(h_S, h_{\text{early}})$, confirming that $\mathcal{L}_{\text{TC}}$ specifically moves the student closer to the converged teacher than to the early teacher. The binary-diagnostic InfoNCE lower bound is 0.601 nats, close to its hard binary ceiling of $\log 2 \approx 0.693$ nats; we read this as a directional indicator that the student systematically prefers T_{final}, not as an absolute mutual-information claim. A non-bound, full-set diagnostic is in Supplementary Sec. A.3.6.

Anti-alignment with the shortcut direction ($\mathcal{L}_{\text{SS}}$ diagnostic, fig. 2b): The signed cosine $\cos(h_S, \Delta h)$ has mean -0.361 for ASD versus 0.003 for KD. Under the diagnostic threshold $\varepsilon_{\text{diag}} = -0.1$, 89.0% of KD samples fall on the shortcut-aligned side, against 11.7% for ASD. ASD does not merely fail to align with Δh it is *anti*-aligned; the representational signature predicted by $\mathcal{L}_{\text{TC}}$.

Subspace geometry (theorem 4 diagnostic, fig. 3a): The principal angle between the shortcut subspace $\text{span}(\mathbf{U}_K)$ and the T_{final}-PCA robust subspace is 65.3°, against 38.6° to the T_{early}-PCA subspace. The shortcut subspace lies substantially closer to the early teacher than to the final teacher, consistent with theorem 4 and eq. (13). The ASD student's projection onto the robust subspace has mean magnitude 0.45 versus 0.12 for KD: ASD does not



(a) Subspace geometry (theorem 4)



(b) Texture-shift robustness

Figure 3: *Subspace and robustness diagnostics on P1.* ASD separates shortcut directions, improves robust-subspace projection over KD (0.45 vs. 0.12), and slightly reduces texture-shift degradation (-31.12% vs. -32.11%).

merely preserve the robust subspace, it re-allocates capacity toward it, explaining why clean accuracy and robustness improve together.

Texture-shift diagnostic (fig. 3b): Under a stylized-input perturbation following [9], ASD obtains 80.3% clean and 49.1% accuracy (degradation -31.12%), versus $80.2\%/48.1\%$ for KD (-32.11%). The gap is modest but directionally consistent: reduced shortcut projection reduces sensitivity to texture perturbation without eliminating it.

Subspace stability: The shortcut covariance has a small eigengap (0.059 absolute, 0.029 relative), so individual eigenvectors are not stably oriented; however, bootstrap resampling gives an inter-resample principal angle of $32.6° \pm 1.6°$ the *region* of shortcut activation is reproducible even though the eigenbasis is not. As $\mathcal{L}_{\text{SS}}$ depends on subspace projection, not eigenvector orientation, this is sufficient (Supplementary Sec. A.7).

7 Conclusion

We presented Anti-Shortcut Distillation, which converts a teacher's training trajectory into both attractive and repulsive supervision. ASD treats T_{final} as a positive anchor and T_{early} as a temporal negative; $\mathcal{L}_{\text{TC}}$ and $\mathcal{L}_{\text{SS}}$ align the student with converged teacher features while suppressing projection onto shortcut-prone directions. Across 13 teacher–student pairs, ASD attains the best clean accuracy on 10 and improves over KD on 12, obtains the lowest mCE on the hardest cross-architecture pair (86.1), and is mechanistically anti-aligned with the shortcut direction ($\cos(h_S, \Delta h) = -0.36$ vs. 0.00 for KD) while tripling robust-subspace projection (0.45 vs. 0.12).

Limitations and future work: ASD stores one extra frozen checkpoint and computes a per-batch eigendecomposition, a small marginal cost. P3 highlights a capacity boundary: severely compressed students may still need low-level cues for discrimination. A transformer-student result (DeiT-Tiny, Supplementary Sec. A.4.7) suggests the signal is not CNN-specific; large-model and self-supervised validation remain open alongside a tighter $\mathcal{L}_{\text{TC}}$–$\mathcal{L}_{\text{SS}}$ characterization.

Acknowledgment

This work was supported by Institute of Information and communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00443391, Korea-Japan Joint International Research on Deep Learning-based Autonomous Mobility Control Technology for Autonomous Mobile Robots).

References

[1] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *ICML*, 2017.

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[3] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, 2021.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[5] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[7] Jiawei Fan, Chao Li, Xiaolong Liu, Meina Song, and Anbang Yao. Augmentation-free dense contrastive knowledge distillation for efficient semantic segmentation. In *NeurIPS*, 2023.

[8] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, 2018.

[9] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

[10] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zeiler, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[11] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[13] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

[14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021.

[15] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[17] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get m for free. In *ICLR*, 2017.

[18] Prannay Khosla, Yonglong Tian, Tan Wang, Yandong Guo, Yang Song, Dilip Krishnan, Ting Chen, Phillip Isola, Huiwen Zhao, Hervé Jégou, et al. Supervised contrastive learning. In *NeurIPS*, 2020.

[19] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

[20] Jiwoon Lee and Jaeho Lee. Debiased distillation by transplanting the last layer. *arXiv preprint arXiv:2302.11187*, 2023.

[21] Sihan Lu, Yang Zheng, Jie Liu, and Zhenghao Xi. UFD-KD: Unified frequency decoupled knowledge distillation. In *BMVC*, 2025.

[22] Michal Lukasik, Srinadh Bhojanapalli, Aditya Krishna Menon, and Sanjiv Kumar. Teacher's pet: Understanding and mitigating biases in distillation. *arXiv preprint arXiv:2106.10494*, 2021.

[23] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, 2020.

[24] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019.

[25] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, 2018.

[26] Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *ICML*, 2019.

[27] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *ICLR*, 2015.

[28] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

[29] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *NeurIPS*, 2020.

[30] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020.

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *NeurIPS Deep Learning Symposium*, 2018.

[32] Chaofei Wang, Qisen Yang, Rui Huang, Shiji Song, and Gao Huang. Efficient knowledge distillation from model checkpoints. In *NeurIPS*, 2022.

[33] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[34] Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

[35] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

[36] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*, 2019.

[37] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, 2022.

[38] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017.