DiD It in 87 Minutes: A Label-Free Softmax-to-Linear Adaptation of Vision Transformers for Object Detection

Anonymous authors
Paper under double-blind review

Abstract

While linear attention is a compelling mechanism for high-resolution object detection due to its reduced cost for global token mixing, converting the Softmax-attention ViT backbone of a trained detector into a linear-attention one is not a trivial drop-in replacement. Directly swapping the attention operator leads to severe performance degradation, and generic label-free distillation, though effective for classification, often fails on detection tasks. We argue that the central challenge is *detector-interface preservation*: the converted backbone must reproduce the exact feature tensors expected by the fixed downstream detector, rather than merely imitating internal Softmax hidden states. To address this, we introduce Detector-Interface Distillation (DiD), a label-free conversion method that exclusively trains the linear-attention backbone by aligning detector-facing interface tensors with those of a frozen Softmax teacher. On DOTA-v1.5, DiD substantially outperforms established baselines and matches supervised, fully trained linear models. Adaptation completes in roughly 87 minutes on 4 GPUs, and the linearized backbone cuts inference latency by $\sim$62% and peak memory by $\sim$49%. We hope our findings offer the community a simple, label-free route to reusing trained Softmax detectors as efficient linear ones, and encourage interface-aware objectives in future architecture-conversion work.

1 Introduction

Vision Transformers (ViTs) have become a dominant backbone family in modern computer vision. Their strong performance is driven by scalable self-attention (Zhai et al., 2022; Dehghani et al., 2023; Fang et al., 2023), large-scale pretraining (Singh et al., 2023; Yang et al., 2025; Wang et al., 2025), and effective transfer to downstream tasks such as image classification (Dosovitskiy et al., 2020), semantic segmentation (Zhang et al., 2022a; Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2025), and object detection (Fang et al., 2021; Zhang et al., 2021; Li et al., 2022; Chen et al., 2022). In detection, ViT-based systems such as ViTDet (Li et al., 2022) and transformer detectors such as YOLOS (Fang et al., 2021) show that pure or nearly pure transformer backbones can provide strong representations when paired with an appropriate detector head or feature pyramid. However, this success comes with a fundamental limitation: standard Softmax attention scales quadratically with the number of image tokens (Keles et al., 2023). This cost becomes especially problematic for dense prediction, where high-resolution inputs are often required to localize small objects (Wang et al., 2021; Liu et al., 2021; Yang et al., 2021; Yuan et al., 2021; Ranftl et al., 2021), as in application scenarios like remote sensing (Xia et al., 2018) and autonomous driving (Yu et al., 2020).

To address this cost, a large body of work has explored more efficient ViT designs. One line of research reduces attention complexity through locality or hierarchy (Wang et al., 2021; Liu et al., 2021) that restricts self-attention to local windows or progressively aggregated regions, substantially improving efficiency while maintaining strong dense-prediction performance. These hierarchical backbones have become highly practical for detection and have been adopted in strong detector families (Zhang et al., 2022b; Zong et al., 2023; Liu et al., 2024; Ren et al., 2024). Nevertheless, hierarchy and locality also impose architectural constraints: long-range interactions become indirect (Yang et al., 2021; Liu et al., 2022; Ding et al., 2022), and the model no longer preserves the same global token-mixing behavior as a standard Softmax-attention ViT.

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

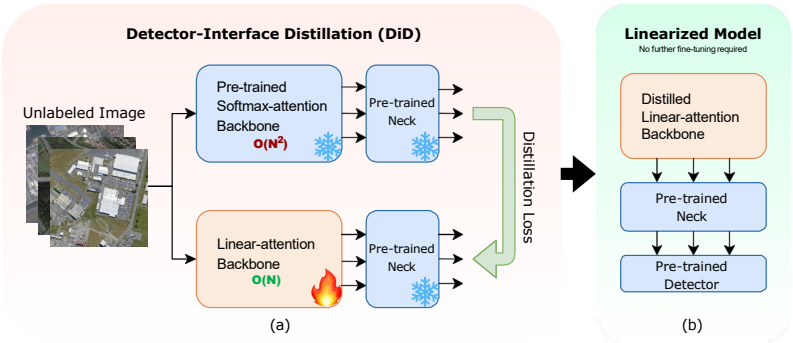


Figure 1: **Overview of the label-free Softmax-to-linear detector conversion setting.** (a) Starting from a trained detector with a Softmax-attention ViT backbone, we construct a student by replacing the backbone attention operator with linear attention. During conversion, the Softmax teacher and the linear student receive the same unlabeled images; only the student backbone is updated, while the downstream detector remains fixed. Snowflake symbols indicate frozen modules, while the fire symbol indicates the trainable module. (b) After conversion, the distilled linear backbone is used directly in the detector without additional fine-tuning.

Another line of work targets the quadratic attention operator itself (Katharopoulos et al., 2020; Han et al., 2024; Zhang et al., 2024; Zheng, 2025), inspired by earlier work in NLP (Kitaev et al., 2020; Choromanski et al., 2020; Qin et al., 2022). Linear-attention methods replace the Softmax kernel with feature-map-based formulations whose complexity scales linearly with sequence length. Early linear-attention variants often trailed quadratic Softmax attention in accuracy (Katharopoulos et al., 2020), but more recent approaches (Han et al., 2024; Zhang et al., 2024; Zheng, 2025) have narrowed this gap and, in some settings, surpassed quadratic attention. These results suggest that linear attention can be competitive under suitable architectures and training regimes. However, adopting linear attention in an existing detector typically demands pre-training the linear backbone from scratch and then fine-tuning the entire detector to recover accuracy, both of which require substantial compute in addition to labeled data.

This creates a practical conversion problem: **can a quadratic-attention backbone in an existing object detector be converted into a linear-attention backbone without labels and without retraining the entire detector?**

In this paper, we study this problem in the setting of lightweight object-detection conversion. Rather than retraining the full detector, we replace the quadratic attention modules in the backbone of a trained ViT detector with linear attention and perform a short label-free adaptation stage that trains only the backbone, as illustrated in Fig. 1. The difficulty is that the converted backbone must remain compatible with a detector pipeline that was trained around the original Softmax feature distribution.

We show that this is not merely a generic representation-transfer problem. For classification, preserving global semantic behavior is often sufficient (Tung & Mori, 2019; Tian et al., 2019), but object detection is more restrictive because downstream performance depends on the deployed multi-scale interface tensors consumed by the detector (Guo et al., 2021; Golizadeh et al., 2025). Prior detector distillation works suggest that supervision at intermediate detector features can be particularly effective for object detection (Guo et al., 2021; Golizadeh et al., 2025). We build on this general intuition by preserving the detector-facing feature contract required by the fixed downstream detector after the attention operator is changed. Based on this observation, we propose *Detector-Interface Distillation* (DiD), a label-free conversion method that supervises detector-facing interface tensors rather than the full sequence of hidden states. Across our experiments, this simple detector-aware objective substantially improves Softmax-to-linear transfer for detection and yields a practical adaptation procedure that completes in roughly 87 minutes on 4 GPUs.

Our contributions are as follows:

1. We identify detector-interface mismatch as a key failure mode in Softmax-to-linear attention conversion for object detection. Direct attention replacement collapses performance, and generic label-free distillation remains insufficient though it preserves well for classification.
2. We propose Detector-Interface Distillation (DiD), a simple label-free conversion method that aligns detector-facing interface tensors while keeping the downstream detector fixed.

3. We provide a controlled study showing that detector-interface supervision substantially outperforms direct copy, generic self-supervised distillation, and hidden-state distillation, and is competitive with full supervised training of the linear detector.
4. We quantify the practical payoff of conversion, showing that the resulting ReLU backbone substantially reduces latency and memory while the conservative 40k-step adaptation completes in about 87 minutes on 4 GPUs.

2 Related Work

Efficient Attention for Vision. The quadratic complexity of Softmax attention has motivated several families of efficient visual architectures. Local and hierarchical transformers restrict attention to windows or progressively aggregated regions (Wang et al., 2021; Liu et al., 2021; Yang et al., 2021; Ding et al., 2022), while sparse designs reduce the number of token interactions. These approaches are highly effective for dense prediction, but long-range information is propagated indirectly and the resulting architecture differs substantially from a standard global-attention ViT. Linear-attention methods instead modify the attention operator itself (Katharopoulos et al., 2020; Qin et al., 2022; Han et al., 2024; Zhang et al., 2024; Zheng, 2025): replacing Softmax with a feature-map formulation and reordering the matrix multiplication reduces the attention complexity from quadratic to linear in sequence length. These designs have narrowed the accuracy gap to Softmax attention, making linear attention a promising option for high-resolution vision. More recent attempts reuse pre-trained Softmax weights to build linear-attention ViTs (Li et al., 2026a;b), but still rely on multi-stage adaptation or full retraining. Our work is complementary: rather than designing or training a new linear-attention architecture from scratch, we study how to convert the backbone of an already trained detector without retraining the detection pipeline.

Label-Free Model Conversion. Knowledge distillation transfers behavior from a frozen teacher to a student using teacher-derived supervision (Hinton et al., 2015; Gou et al., 2021). In label-free settings, common objectives match logits, global representations, hidden states, attention outputs, or attention statistics. Softmax-to-linear conversion methods similarly treat operator replacement as a distillation problem, matching hidden states (Goldstein et al., 2025), attention behavior (Zhang et al., 2024), or low-rank approximations of attention maps (Zhang et al., 2025). Related operator-conversion pipelines have also been explored for language models (Kasai et al., 2021; Wang et al., 2024). Most existing conversion methods assume that any remaining mismatch can be repaired through subsequent supervised fine-tuning. Our setting removes this recovery stage: the converted backbone aims to be compatible with a frozen detector, without detection labels. This changes the objective from approximate operator imitation to preservation of the deployed task interface.

Knowledge Distillation for Object Detection. Detection distillation has shown that intermediate feature supervision can be more effective than output-level imitation because detectors depend on structured spatial features and strongly imbalanced foreground-background signals (Guo et al., 2021; Yang et al., 2022; Golizadeh et al., 2025). Conventional detector distillation, however, generally transfers a larger teacher into a smaller student, often with detection labels and while training the detector pathway. DiD addresses a different setting: teacher and student have comparable capacity, the attention operator is changed post hoc, the downstream detector remains fixed, and conversion uses only unlabeled images. We build on the general importance of detector features but target the specific interface compatibility required by architectural conversion. Our setting is also related to attention-transfer studies (Li et al., 2024; Qin et al., 2026), which transfer the attention behavior of a pre-trained ViT to a student model; DiD instead preserves the detector-facing interface after the attention operator itself is replaced.

3 Problem Formulation and Diagnostic Study

3.1 Label-Free Softmax-to-Linear Conversion

For query, key, and value projections Q, K, and V, standard Softmax attention is

$$\text{Attention}_{\text{Softmax}}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{1}$$

Table 1: **Cross-task Softmax-to-ReLU transfer.** Each task reports its native score followed by the percentage retained relative to its own Softmax reference. Generic distillation nearly restores classification accuracy but recovers only half of the DOTA-v1.5 detector mAP.

Method	Classification				Detection	
	ImageNet		CIFAR-100		DOTA-v1.5	
	Acc.	Retained (%)	Acc.	Retained (%)	mAP	Retained (%)
Softmax reference	79.83	100.0	78.09	100.0	33.5	100.0
Direct copy	5.08	6.4	3.54	4.5	1.7	5.1
Generic distillation	77.83	97.5	75.92	97.2	18.1	54.0

which explicitly forms a token-to-token affinity matrix. Linear attention replaces this operation with an associative feature-map form,

$$\mathrm{Attention}_{\mathrm{linear}}(Q, K, V) = \phi(Q)\left(\phi(K)^\top V\right), \tag{2}$$

where $\phi(\cdot)$ is a non-negative feature map. This reordering avoids explicitly materializing the full affinity matrix and improves scaling with sequence length. However, it also changes the backbone function: parameters trained jointly with Softmax attention are not guaranteed to preserve the teacher behavior after the operator is replaced.

Let B_t denote the trained Softmax backbone and B_s the linear-attention student initialized from all compatible teacher parameters. We decompose the trained downstream detector pathway as

$$G = H \circ N, \tag{3}$$

where N denotes the neck or feature-pyramid transformation and H denotes the subsequent prediction modules. During conversion, B_t, N, and H remain frozen, and only B_s is optimized using unlabeled images $x \sim \mathcal{D}$:

$$\theta_s^\star = \arg\min_{\theta_s} \mathbb{E}_{x\sim\mathcal{D}}\left[\mathcal{L}_{\mathrm{conv}}\big(B_s(x), B_t(x); N, H\big)\right]. \tag{4}$$

The central question is therefore not merely how closely B_s imitates B_t, but which teacher-student discrepancy must be minimized to keep the downstream detector functional.

3.2 Why Is Generic Transfer Insufficient?

To test whether a classification-oriented transfer objective is sufficient for object detection, we compare direct copy and generic LightlyTrain-style global-feature distillation on ImageNet and CIFAR-100 classification, and on DOTA-v1.5 detection with RetinaNet. Because classification accuracy and detection mAP are not directly comparable, each task reports both its native metric and the percentage retained relative to its own Softmax reference.

Direct copy collapses across all three tasks (Table 1), confirming that the attention operator cannot be replaced without adaptation. Generic distillation, however, exhibits a clear task gap: it retains 97.5% and 97.2% of the Softmax-reference accuracy on ImageNet and CIFAR-100, respectively, but only 54.0% of the reference mAP on DOTA-v1.5. Preserving global semantic representations is therefore insufficient to maintain the structured, multi-scale features required by a fixed detector, motivating a supervision target tied to the deployed detection pathway.

3.3 Where Should Conversion Be Supervised?

The cross-task comparison suggests that preserving global semantic representations is insufficient for object detection. We therefore ask where supervision should be applied so that the converted backbone remains compatible with the fixed detector. Using RetinaNet on DOTA-v1.5, we conduct a controlled 20k-step conversion in which the reconstruction loss and optimization settings are fixed while only the supervised tensors are changed. We compare five candidate targets: attention outputs, hidden states from all transformer layers, hidden states from the backbone layers read by the neck, the final hidden state, and the multi-scale features produced after the frozen neck / FPN and passed to the detector heads. For a selected tensor set $\mathcal{V}$, we use the following loss:

$$\mathcal{L}_{\mathrm{place}} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \|v_s - v_t\|_F^2, \tag{5}$$

where v_t and v_s denote corresponding teacher and student tensors.

Algorithm 1 DiD PyTorch-style pseudocode.

```
# B_t: frozen Softmax teacher backbone
# B_s: trainable linear-attention student
# N_t, N_s: identical frozen neck / FPN copies
# H: frozen detector heads (unused in loss)
# beta: weights, beta[i] for interface m = i+1
B_s.params = B_t.params # copy compatible weights
for x in unlabeled_images: # no detection labels
    R_t = B_t.readout_features(x) # frozen path
    R_s = B_s.readout_features(x) # trainable

    Z_t = N_t(R_t) # tensors z_t^(1), ..., z_t^(M)
    Z_s = N_s(R_s) # tensors z_s^(1), ..., z_s^(M)

    # weighted detector-interface reconstruction
    loss = 0
    for i in range(M):
        loss += beta[i]*mse(Z_s[i], Z_t[i].detach())
    loss /= sum(beta)

    loss.backward() # update only student backbone
    update(B_s) # AdamW step

# deploy: drop teacher, reuse the frozen detector
converted_detector = H o N_s o B_s
```

As shown in Table 2, matching all transformer hidden states is the strongest detector-agnostic target, but it remains 5.5 mAP below supervision after the neck / FPN. Matching only the backbone layers supplied to the neck is also insufficient. Therefore, proximity to the detector does not by itself guarantee compatibility; the converted backbone must reproduce the transformed multi-scale features that the detector actually receives.

Table 2: **Effect of supervision placement for RetinaNet conversion on DOTA-v1.5.** Matching multi-scale features produced after the neck / FPN is substantially more effective than matching attention outputs or internal backbone states.

Supervision target	mAP	mAP_{50}	mAP_{75}
Attention outputs	7.8	17.4	5.6
Hidden states from all layers	25.2	44.7	24.6
Hidden states from readout layers	20.3	37.4	19.0
Final hidden state	12.6	24.1	11.7
Post-neck / FPN detector features	**30.7**	**54.1**	**30.1**

This experiment identifies the feature boundary between the frozen neck / FPN and the detector heads as the appropriate conversion target. We refer to this boundary as the *detector-facing interface* and formalize its preservation as Detector-Interface Distillation in the next section.

4 DiD: Detector-Interface Distillation

4.1 Overview

Detector-Interface Distillation (DiD) adapts a linear-attention student backbone to a detector that was originally trained with a Softmax-attention backbone. The frozen teacher and the student process the same unlabeled image, and their backbone readout features are passed through identical frozen neck / FPN pathways. DiD then aligns the multi-scale tensors produced at the output of the neck / FPN, while updating only the student backbone.

The key design is therefore not to reproduce the teacher's complete internal computation trace. Softmax and linear attention implement different token-mixing rules, so their intermediate hidden states need not follow identical trajectories. DiD instead preserves the feature boundary on which the fixed detector depends. Internal transformer states and final detector predictions are not used as distillation targets. Algorithm 1 summarizes this computation in pseudocode.

4.2 Detector-Facing Interface

Let B_t denote the frozen Softmax teacher backbone and B_s the trainable linear-attention student backbone. For an input image x, the two backbones expose the readout sets

$$\mathcal{R}_t(x) = \left\{ r_t^{(\ell)}(x) \right\}_{\ell \in \mathcal{I}}, \qquad \mathcal{R}_s(x) = \left\{ r_s^{(\ell)}(x) \right\}_{\ell \in \mathcal{I}}, \tag{6}$$

where $\mathcal{I}$ indexes the backbone layers read by the detector neck. The exact readout locations depend on the detector architecture. Let N denote the pretrained neck or feature-pyramid pathway. During

conversion, the teacher and student use identical frozen copies of N. Their detector-facing outputs are

$$\mathcal{Z}_t(x) = \left\{ z_t^{(m)}(x) \right\}_{m=1}^{M} = N\left(\mathcal{R}_t(x)\right), \tag{7}$$

and

$$\mathcal{Z}_s(x) = \left\{ z_s^{(m)}(x) \right\}_{m=1}^{M} = N\left(\mathcal{R}_s(x)\right). \tag{8}$$

We refer to $\mathcal{Z}_t(x)$ and $\mathcal{Z}_s(x)$ as the *detector-facing interfaces*: they are the multi-scale feature tensors presented by the frozen neck / FPN to the downstream detector. This definition separates the supervision target from the backbone readouts themselves. The readout tensors $\mathcal{R}(x)$ are inputs to the neck / FPN, whereas the interface tensors $\mathcal{Z}(x)$ are its outputs and constitute the feature contract consumed by the detector.

DiD supervises this post-neck / FPN interface rather than every internal hidden state. It also does not distill final detector predictions, because doing so introduces detector-specific prediction, matching, and assignment logic beyond the feature interface itself. In the RetinaNet setting used for the supervision-placement and objective-design studies, $M = 5$.

4.3 Weighted Interface Objective

DiD aligns corresponding teacher and student interface tensors using the weighted reconstruction objective:

$$\mathcal{L}_{\mathrm{DiD}}(x) = \frac{1}{\sum_{m=1}^{M} \beta_m} \sum_{m=1}^{M} \beta_m \left\| z_s^{(m)}(x) - z_t^{(m)}(x) \right\|_F^2, \tag{9}$$

where $\beta_m > 0$ controls the contribution of interface output m. For the five-interface RetinaNet setting, the default configuration uses $(\beta_1, \ldots, \beta_5) = (3.0,\ 2.5,\ 2.0,\ 1.5,\ 1.0)$, following the early-to-late output ordering used in the implementation. We use plain MSE because it directly preserves the absolute feature scale and channel responses expected by the frozen detector. The interface-weighting and loss-design comparisons are reported in Sec. 5.6.

4.4 Conversion and Deployment

Recall that N denotes the frozen neck / FPN and H denotes the frozen detector prediction modules that consume its outputs. Given an unlabeled conversion set $\mathcal{D}$, DiD optimizes only the student-backbone parameters:

$$\theta_s^{\star} = \arg\min_{\theta_s} \mathbb{E}_{x \sim \mathcal{D}} \left[\mathcal{L}_{\mathrm{DiD}}(x) \right]. \tag{10}$$

For each image, the teacher provides the detached targets $\mathcal{Z}_t(x)$ and the student produces $\mathcal{Z}_s(x)$. The teacher backbone, both neck / FPN pathways, and the detector prediction modules remain frozen. DiD uses no ground-truth boxes, class labels, detector-output targets, or pseudo-label assignment.

After conversion, the teacher branch is discarded. The converted detector becomes $f_s^{\star} = H \circ N \circ B_s^{\star}$, which reuses the original neck / FPN and prediction modules without additional fine-tuning.

5 Experiments

5.1 Experimental Setup

Datasets and evaluation. We evaluate label-free detector conversion on DOTA-v1.5 (Xia et al., 2018), a high-resolution aerial-detection benchmark where global token mixing is costly and small objects are common. Conversion uses the training images only as unlabeled inputs; detection annotations are never used to optimize the student. The converted detectors are evaluated using mAP, mAP_{50}, and mAP_{75}. A generalization study on Pascal VOC (Everingham et al., 2010), testing whether the conversion behavior transfers beyond aerial imagery, is provided in the Appendix.

Architectures. We use RetinaNet (Lin et al., 2017) and Faster R-CNN (Ren et al., 2015) with a ViT-S backbone. The student is initialized by copying all compatible parameters from the trained Softmax detector and replacing only the backbone attention operator. During conversion, the Softmax teacher, neck / FPN, and detector heads remain frozen, and only the linear-attention student backbone is optimized. DiD supervises the multi-scale tensors exposed by the frozen neck / FPN to the

Table 3: **Label-free Softmax-to-linear conversion on DOTA-v1.5 with ViT-S.** Full training is a supervised reference; all other adaptation methods use no detection labels.

Detector	Method	ReLU			ELU		
		mAP	mAP_{50}	mAP_{75}	mAP	mAP_{50}	mAP_{75}
RetinaNet	Softmax reference	33.5	58.1	33.2	33.5	58.1	33.2
	Full training	30.9	53.3	30.9	30.8	53.0	**31.3**
	Direct copy	1.7	3.8	1.2	1.3	2.9	0.9
	Generic distillation	18.1	35.2	16.6	9.6	19.2	5.3
	Softmax-generated pseudo-label distillation	14.8	33.6	10.1	19.6	38.2	17.2
	Hidden-state distillation	28.1	49.7	27.7	26.7	47.4	26.6
	DiD (Ours)	**31.1**	**54.5**	**31.0**	**31.1**	**54.4**	30.7
Faster R-CNN	Softmax reference	33.1	56.0	34.1	33.1	56.0	34.1
	Full training	28.9	50.0	29.1	28.6	50.0	28.6
	Direct copy	0.3	0.7	0.2	0.2	0.4	0.1
	Generic distillation	14.9	30.8	12.9	8.9	19.3	7.1
	Softmax-generated pseudo-label distillation	25.7	46.2	25.3	25.4	46.1	24.9
	Hidden-state distillation	27.1	48.1	26.9	27.2	48.1	26.6
	DiD (Ours)	**29.3**	**51.1**	**29.8**	**28.9**	**50.2**	**29.4**

downstream detector, using the early-heavy interface weighting and plain MSE defaults of Sec. 4. We evaluate two linear-attention feature maps: $\phi(x) = \mathrm{ReLU}(x)$ following Han et al. (2024), and $\phi(x) = \mathrm{ELU}(x) + 1$ following Katharopoulos et al. (2020).

Conversion protocol. Unless otherwise stated, all conversion methods use 40k iterations with AdamW, an initial learning rate of 10^{-3}, and learning-rate decay at 4k and 32k iterations. The default augmentation is random horizontal flipping. Design ablations use a reduced 20k-step schedule. All experiments are conducted on NVIDIA RTX A6000 GPUs. Each experiment is repeated three times with different seeds; standard deviations are reported in the Appendix.

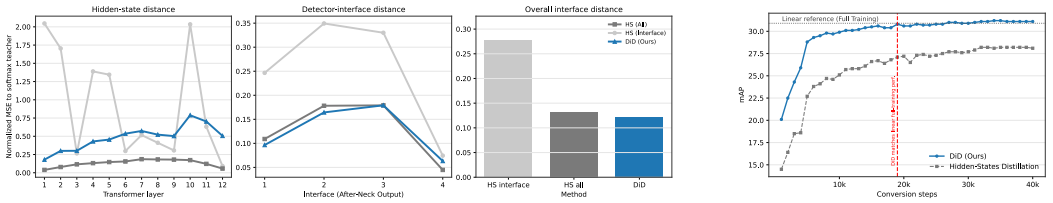
5.2 Baselines

We compare DiD with four label-free alternatives and one supervised reference. *Direct copy* replaces the attention operator and evaluates the copied parameters without adaptation, following the attention-weight transfer setting of RADLADS (Goldstein et al., 2025). *Generic distillation* uses LightlyTrain (Lightly AI, 2024) to align global teacher-student representations. *Softmax-generated pseudo-label distillation* uses AutoDistill (Contributors, 2023) to train the student from the frozen teacher's predicted detections. *Hidden-state distillation* follows RADLADS-style intermediate representation matching across transformer layers. *Full training* trains the corresponding linear-attention detector using detection labels and is included only as a supervised reference.

5.3 Main Results

Table 3 establishes three consistent trends: Firstly, direct copy produces near-zero mAP for both attention kernels and detector families, confirming that the attention operator cannot be replaced without adaptation. Secondly, generic representation matching and pseudo-label supervision recover only part of the lost detector performance. Hidden-state distillation is substantially stronger, showing that intermediate representations matter, but it remains below DiD in every setting. Thirdly, DiD is the strongest label-free method. For RetinaNet, it reaches 31.1 mAP with both ReLU and ELU, compared with 28.1 and 26.7 for hidden-state distillation. Its performance is also comparable to the supervised linear-attention references of 30.9 and 30.8 mAP. We interpret the small numerical differences around the supervised reference as comparable operating points rather than evidence that label-free conversion is superior to supervised training.

Faster R-CNN exhibits the same overall ranking. DiD reaches 29.3 mAP with ReLU and 28.9 with ELU, outperforming all tested label-free baselines and reaching performance comparable to the corresponding supervised linear-attention references. The consistency between RetinaNet, an anchor-based one-stage detector, and Faster R-CNN, a proposal-based two-stage detector, indicates that DiD is not tied to a particular downstream detection architecture.



(a) (b)

Figure 2: **Conversion behavior on RetinaNet.** (a) Hidden-state distillation more closely matches internal transformer states, while DiD more closely matches the detector-facing interface. (b) DiD consistently outperforms hidden-state distillation during conversion and reaches the supervised linear-training regime after approximately 20k-30k iterations.

5.4 Interface Alignment and Conversion Dynamics

To understand why DiD outperforms hidden-state distillation, we compare how the two objectives affect teacher-student alignment inside the backbone and at the detector-facing interface. Fig. 2(a) compares the normalized feature error $\mathrm{nMSE}(s,t) = \mathbb{E}\|s-t\|_2^2 / \mathbb{E}\|t\|_2^2$ at internal backbone states and detector-facing tensors.

All-layer hidden-state distillation achieves lower mismatch within the transformer, whereas DiD achieves lower mismatch at the detector-facing interface and higher detector accuracy. This result provides representational evidence for the internal-compensation interpretation: after the attention operator changes, close agreement at every intermediate state is not necessary, provided that the student resolves the discrepancy before its features are consumed by the frozen detector.

Fig. 2(b) shows that DiD maintains an advantage over hidden-state distillation throughout conversion and reaches the supervised linear-training regime after approximately 20k-30k steps. We retain 40k steps as default, while the trajectory shows that most practical recovery occurs earlier.

5.5 Inference and Conversion Efficiency

The converted model realizes the systems benefit that motivates linear attention. Fig. 3 compares backbone-only inference for Softmax attention and the ReLU-linearized ViT. The linearized backbone reduces average latency by approximately 62% and peak memory by approximately 49%, showing that DiD produces a deployable efficiency improvement rather than only a representational substitution.

Table 4: **RetinaNet ReLU conversion throughput at fixed global batch size 8.**

GPUs	Global batch	Iter/s	Wall time (min)	
			20k	40k
1	8	2.98	112	224
2	8	5.20	64	128
4	8	7.67	44	87

The adaptation stage is also short relative to full detector training. Table 4 reports conversion throughput at a fixed global batch size of 8: increasing the number of GPUs from one to four raises throughput from 2.98 to 7.67 iterations per second. The 20k-step practical regime takes approximately 44 minutes on four GPUs, while the conservative 40k-step default takes approximately 87 minutes.

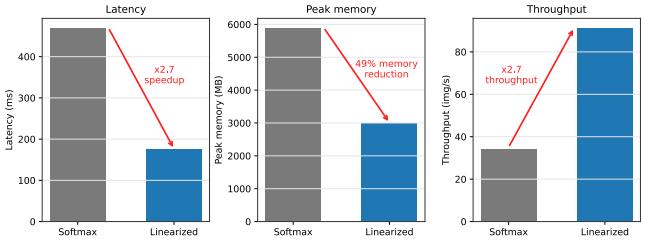


Figure 3: **Backbone-only inference comparison between the Softmax and ReLU-linearized ViT.** Linear attention substantially reduces latency and peak memory.

5.6 Ablation Studies

All ablations use RetinaNet with ReLU linear attention on DOTA-v1.5 and a reduced 20k-step conversion schedule unless otherwise stated.

Interface weighting. Table 5 compares uniform, late-heavy, and early-heavy weighting over the five detector-facing tensors. Early-heavy weighting performs best among the evaluated schemes, reaching 30.7 mAP compared with 29.8 for uniform weighting and 28.4 for late-heavy weighting.

This result shows that the five interface levels do not contribute equally under the tested conversion setting. We therefore adopt early-heavy weighting as the default without claiming that this ordering is universally optimal across detector architectures.

Table 5: **Effect of detector-interface weighting.**

Interface weighting	mAP	mAP_{50}	mAP_{75}
Uniform	29.8	52.1	29.4
Late-heavy	28.4	49.8	28.2
Early-heavy	**30.7**	**54.1**	**30.1**

Reconstruction loss. Table 6 compares MSE, L1, Smooth L1, and their cosine-augmented variants. Plain MSE performs best, while Smooth L1 remains close and cosine regularization provides no consistent benefit. This result confirms that DiD does not rely on a specialized reconstruction objective: a simple MSE loss is sufficient once the detector-facing interface has been selected as the conversion target.

Table 6: **Effect of the reconstruction loss.**

Loss	mAP	mAP_{50}	mAP_{75}
MSE	**30.7**	**54.1**	**30.1**
L1	30.0	52.5	29.3
Smooth L1	30.3	53.0	29.7
MSE + cosine	30.4	53.2	29.8
L1 + cosine	29.3	51.5	28.8
Smooth L1 + cosine	30.4	53.1	29.9

Augmentation severity. Light detection-style augmentation performs best in Table 7. Strong color and geometric perturbations reduce mAP, consistent with the purpose of conversion: DiD aligns the student to an already trained detector interface rather than learning invariances from scratch.

Table 7: **Effect of augmentation severity.**

Train augmentation	mAP	mAP_{50}	mAP_{75}
RandomHorizontalFlip	**30.7**	**54.1**	**30.1**
+ Medium ColorJitter + RandomAffine	30.0	52.2	29.8
+ Strong ColorJitter + RandomAffine	27.0	47.9	26.4

Source domain. Table 8 compares unlabeled conversion images from DOTA-v1.5, FAIR1M (Sun et al., 2022), and ImageNet (Russakovsky et al., 2015). In-domain DOTA images perform best, followed by the related aerial domain and then generic natural images. Because supervision comes from the teacher, conversion data is most useful when it exposes the student to the visual distribution that the deployed detector will serve.

Table 8: **Effect of the conversion-data domain.**

Conversion source	mAP	mAP_{50}	mAP_{75}
DOTA-v1.5	**30.7**	**54.1**	**30.1**
FAIR1M	25.1	43.5	25.0
ImageNet	20.2	36.6	19.7

Conversion budget. While Fig. 2(b) compares the optimization trajectories of DiD and hidden-state distillation, Fig. 4 focuses specifically on the effect of the DiD conversion budget with RetinaNet on DOTA-v1.5. Most of the final performance is recovered within 20k-40k iterations, after which the improvement becomes modest. This supports the 40k schedule as a conservative default while showing that a shorter 20k conversion already reaches the practical operating regime.

More ablations are provided in the Appendix.

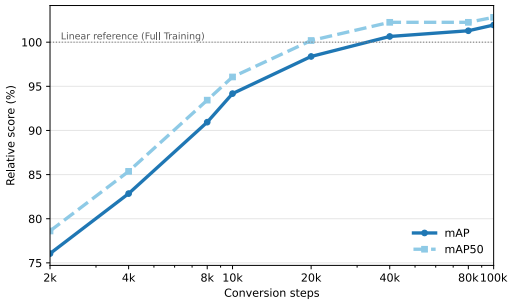


Figure 4: **Scaling behavior of RetinaNet DiD.** Scores are normalized to the linear full-training.

6 Conclusion

We presented Detector-Interface Distillation (DiD), a label-free method for converting trained Softmax-attention ViT detectors into efficient linear-attention ones. Our central finding is that successful conversion is governed by compatibility at the detector-facing interface: supervising the tensors actually consumed by the frozen detector lets the linear backbone compensate internally, where direct operator replacement and generic distillation fails. On DOTA-v1.5, DiD is the strongest label-free conversion method for both RetinaNet and Faster R-CNN, reaches an operating point comparable to fully supervised linear-attention training, and completes in about 87 minutes on 4 GPUs while substantially reducing inference latency and peak memory. We hope DiD offers a simple post-training route to efficient ViT detectors, and that interface-aware objectives extend beyond the ViT-S scale studied here to larger backbones, additional detector families, and broader conversion settings.

AI USE STATEMENT

We certify the use of AI to aid in the writing and polishing of this manuscript, and the use of AI to aid code implementations of involved experiments in this work. All AI-assisted content shown in this manuscript has been reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of AI.

ETHICS STATEMENT

This work aims to propose a label-free adaptation method to convert the Softmax-attention ViT backbone of a trained detector into a linear-attention one. This method relies on existing public pre-trained Softmax ViT weights, some of which are trained on large-scale curated or web-scale datasets and may inherit dataset biases. The proposed method for Softmax-to-linear conversion can make this adaptation more effective, but it does not remove biases or safety concerns inherited from the original Softmax weights.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, we commit to making our source code publicly available upon publication. The code will include our implementations and scripts to replicate the empirical results presented in this manuscript. Comprehensive experimental setups of our analysis, including details of datasets, evaluation protocols, architectures, and conversion protocols, are provided in Sec. 5.1. Additionally, all experiments are repeated over 3 random seeds to ensure statistical significance. Table A provides the statistical significance analysis of our obtained results.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 1

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 16

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 1

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020. 2

AutoDistill Contributors. Autodistill: Compositional foundation model pipelines. https://github.com/autodistill/autodistill, 2023. 7, 15

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, pp. 7480–7512. PMLR, 2023. 1

Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *ECCV*, pp. 74–92. Springer, 2022. 1, 3

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 6, 16

Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. In *NeurIPS*, pp. 26183–26197, 2021. 1

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, pp. 19358–19369, 2023. 1

Daniel Goldstein, Eric Alcaide, Janna Lu, and Eugene Cheah. Radlads: Rapid attention distillation to linear attention decoders at scale. *arXiv preprint arXiv:2505.03005*, 2025. 3, 7, 15

Mahdi Golizadeh, Nassibeh Golizadeh, Mohammad Ali Keyvanrad, and Hossein Shirazi. Architectural insights into knowledge distillation for object detection: A comprehensive review. *arXiv preprint arXiv:2508.03317*, 2025. 2, 3

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *IJCV*, 129(6):1789–1819, 2021. 3

Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *CVPR*, pp. 2154–2164, 2021. 2, 3

Dongchen Han, Yifan Pu, Zhuofan Xia, Yizeng Han, Xuran Pan, Xiu Li, Jiwen Lu, Shiji Song, and Gao Huang. Bridging the divide: Reconsidering softmax and linear attention. In *NeurIPS*, pp. 79221–79245, 2024. 2, 3, 7

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

Jungo Kasai, Hao Peng, Yizhe Zhang, Dani Yogatama, Gabriel Ilharco, Nikolaos Pappas, Yi Mao, Weizhu Chen, and Noah A Smith. Finetuning pretrained transformers into rnns. In *EMNLP*, pp. 10630–10643, 2021. 3

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, pp. 5156–5165. PMLR, 2020. 2, 3, 7

Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. On the computational complexity of self-attention. In *International conference on algorithmic learning theory*, pp. 597–619. PMLR, 2023. 1

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pp. 4015–4026, 2023. 1

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The Efficient Transformer. *arXiv e-prints*, art. arXiv:2001.04451, January 2020. doi: 10.48550/arXiv.2001.04451. 2

Alexander C Li, Yuandong Tian, Beidi Chen, Deepak Pathak, and Xinlei Chen. On the surprising effectiveness of attention transfer for vision transformers. In *NeurIPS*, pp. 113963–113990, 2024. 3

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, pp. 280–296. Springer, 2022. 1

Yifan Li, Seunghyun Yoon, Viet Dac Lai, Franck Dernoncourt, Jason Kuen, Yu Kong, and Trung Bui. Vit-adala: Adapting vision transformers with linear attention. *arXiv preprint arXiv:2603.16063*, 2026a. 3

Yining Li, Dongchen Han, Zeyu Liu, Hanyi Wang, Yulin Wang, and Gao Huang. Linearizing vision transformer with test-time training. *arXiv preprint arXiv:2605.02772*, 2026b. 3

Lightly AI. lightly-train: A framework for self-supervised learning in computer vision. https://github.com/lightly-ai/lightly-train, 2024. 7, 15

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pp. 2980–2988, 2017. 6

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, pp. 38–55. Springer, 2024. 1

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pp. 10012–10022, 2021. 1, 3

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, pp. 12009–12019, 2022. 1

Huaiyuan Qin, Muli Yang, Gabriel James Goenawan, Peng Hu, Chen Gong, Xi Peng, and Hongyuan Zhu. Attention transfer is not universally effective for vision transformers. *arXiv preprint arXiv:2605.07191*, 2026. 3

Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. *arXiv preprint arXiv:2202.08791*, 2022. 2, 3

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pp. 12179–12188, 2021. 1

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *ICLR*, pp. 28085–28128, 2025. 1

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 6

Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, et al. Dino-x: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*, 2024. 1

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 9

Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *ICCV*, 2023. 1

Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022. 9

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 2

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, pp. 1365–1374, 2019. 2

Junxiong Wang, Daniele Paliotta, Avner May, Alexander M Rush, and Tri Dao. The mamba in the llama: Distilling and accelerating hybrid models. In *NeurIPS*, pp. 62432–62457, 2024. 3

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, pp. 568–578, 2021. 1, 3

Xiao Wang, Ibrahim Alabdulmohsin, Daniel Salz, Zhe Li, Keran Rong, and Xiaohua Zhai. Scaling pre-training to one hundred billion data for vision language models. *arXiv preprint arXiv:2502.07617*, 2025. 1

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *CVPR*, pp. 3974–3983, 2018. 1, 6

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal self-attention for local-global interactions in vision transformers. *arXiv preprint arXiv:2107.00641*, 2021. 1, 3

Wenxuan Yang, Qingqv Wei, Chenxi Ma, Weimin Tan, and Bo Yan. Scaling laws for data-efficient visual transfer learning. In *ACM MM*, pp. 2968–2976, 2025. 1

Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *CVPR*, pp. 4643–4652, 2022. 3

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pp. 2636–2645, 2020. 1

Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. In *NeurIPS*, pp. 7281–7293, 2021. 1

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *CVPR*, pp. 12104–12113, 2022. 1

Bowen Zhang, Zhi Tian, Quan Tang, Xiangxiang Chu, Xiaolin Wei, Chunhua Shen, et al. Segvit: Semantic segmentation with plain vision transformers. In *NeurIPS*, pp. 4971–4982, 2022a. 1

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022b. 1

Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. The hedgehog & the porcupine: Expressive linear attentions with softmax mimicry. *arXiv preprint arXiv:2402.04347*, 2024. 2, 3

Michael Zhang, Simran Arora, Rahul Chalamala, Benjamin Spector, Alan Wu, Krithik Ramesh, Aaryan Singhal, and Christopher Ré. Lolcats: On low-rank linearizing of large language models. In *ICLR*, pp. 46200–46253, 2025. 3

Zixiao Zhang, Xiaoqiang Lu, Guojin Cao, Yuting Yang, Licheng Jiao, and Fang Liu. Vit-yolo: Transformer-based yolo for object detection. In *ICCV*, pp. 2799–2808, 2021. 1

Chuanyang Zheng. The linear attention resurrection in vision transformer. *arXiv preprint arXiv:2501.16182*, 2025. 2, 3

Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *ICCV*, pp. 6748–6758, 2023. 1

Contents

1 Introduction **1**

2 Related Work **3**

3 Problem Formulation and Diagnostic Study **3**

3.1 Label-Free Softmax-to-Linear Conversion 3

3.2 Why Is Generic Transfer Insufficient? 4

3.3 Where Should Conversion Be Supervised? 4

4 DiD: Detector-Interface Distillation **5**

4.1 Overview 5

4.2 Detector-Facing Interface 5

4.3 Weighted Interface Objective 6

4.4 Conversion and Deployment 6

5 Experiments **6**

5.1 Experimental Setup 6

5.2 Baselines 7

5.3 Main Results 7

5.4 Interface Alignment and Conversion Dynamics 8

5.5 Inference and Conversion Efficiency 8

5.6 Ablation Studies 8

6 Conclusion **9**

A Baseline Objectives **15**

B Variance Across Seeds **15**

C Ablation on Teacher Initialization **16**

D Generalization to Pascal VOC **16**

A Baseline Objectives

This section formalizes the optimization-based baseline objectives used in the main paper. Let θ_t and θ_s denote the teacher and student parameters, respectively.

Generic distillation. Our first optimization baseline adopts a generic teacher-student distillation objective on unlabeled images, instantiated with the LightlyTrain framework. The teacher is kept fixed, and the student is optimized as

$$\theta_s^\star = \arg\min_{\theta_s} \mathbb{E}_{x\sim\mathcal{D}} \left[\mathcal{L}_{\text{ssl}}(x; \theta_s, \theta_t)\right]. \tag{A}$$

Concretely, the loss matches global representations across two stochastic augmentations $x^{(1)}$ and $x^{(2)}$ of the same unlabeled image:

$$\mathcal{L}_{\text{ssl}} = \mathcal{C}_{\text{vec}}\left(g_s(x^{(1)}), g_t(x^{(2)})\right) + \mathcal{C}_{\text{vec}}\left(g_s(x^{(2)}), g_t(x^{(1)})\right), \tag{B}$$

where $g_t(\cdot)$ and $g_s(\cdot)$ denote the teacher and student global projection heads, and

$$\mathcal{C}_{\text{vec}}(a, b) = 1 - \frac{\langle a, b\rangle}{\|a\|_2 \, \|b\|_2}. \tag{C}$$

This baseline preserves high-level semantic consistency, but it does not explicitly constrain the detector-facing interface tensors consumed by the downstream detector. In practice, we use the LightlyTrain (Lightly AI, 2024) implementation and follow its recommended setting.

Softmax-generated pseudo-label distillation. The next baseline uses the frozen Softmax detector as a pseudo-label teacher. For each unlabeled image x, the teacher produces pseudo detections $\hat{y}_t(x)$, and the student is trained with the corresponding detection loss:

$$\theta_s^\star = \arg\min_{\theta_s} \mathbb{E}_{x\sim\mathcal{D}} \left[\mathcal{L}_{\text{det}}\left(f_s(x), \hat{y}_t(x)\right)\right]. \tag{D}$$

This baseline transfers detector supervision at the output level rather than through intermediate feature matching. It is more task-specific than generic self-supervised distillation, but it still does not directly enforce compatibility at the deployed detector-facing interface tensors. We implement this baseline with the AutoDistill (Contributors, 2023) framework.

Hidden-state distillation. A more direct representation-matching baseline supervises backbone hidden states across transformer layers, following the general strategy of RADLADS (Goldstein et al., 2025). Let $\mathcal{K}$ denote the set of supervised hidden-state layers, and let $h_t^{(\ell)}(x)$ and $h_s^{(\ell)}(x)$ denote the teacher and student hidden states at layer ℓ. Let $\alpha_\ell > 0$ be the weight assigned to layer $\ell \in \mathcal{K}$. In our implementation, α_ℓ may differ between detector-interface layers and non-interface layers, although we use equal weighting by default. The loss is

$$\mathcal{L}_{\text{hid}} = \frac{1}{\sum_{\ell\in\mathcal{K}} \alpha_\ell} \sum_{\ell\in\mathcal{K}} \alpha_\ell \Big(\|h_s^{(\ell)}(x) - h_t^{(\ell)}(x)\|_F^2 + \lambda_{\cos}\mathcal{C}_{\text{tok}}\left(h_s^{(\ell)}(x), h_t^{(\ell)}(x)\right)\Big), \tag{E}$$

where $\|\cdot\|_F$ is the Frobenius norm and

$$\mathcal{C}_{\text{tok}}(a, b) = \frac{1}{|\Omega(a)|} \sum_{u\in\Omega(a)} \left(1 - \frac{\langle a_u, b_u\rangle}{\|a_u\|_2 \, \|b_u\|_2}\right) \tag{F}$$

is a tokenwise cosine alignment loss, with $\Omega(a)$ denoting the set of token locations in the hidden-state tensor. This objective encourages the converted student to preserve the teacher's intermediate transformer behavior in a generic, detector-agnostic way.

These baselines test whether global semantic distillation, output-level pseudo-label supervision, or hidden-state imitation suffice for Softmax-to-linear conversion in object detection.

B Variance Across Seeds

Each experiment is repeated three times with different random seeds. Table A reports the mean and standard deviation over the three seeds for DiD with RetinaNet and Faster R-CNN under the ReLU and ELU kernels on DOTA-v1.5.

Table A: **Mean ± standard deviation over three seeds for DiD conversion on DOTA-v1.5 with a ViT-S backbone.**

Detector	Kernel	mAP	mAP_{50}	mAP_{75}
RetinaNet	ReLU	31.1 ± 0.1	54.5 ± 0.2	31.0 ± 0.1
	ELU	31.1 ± 0.1	54.4 ± 0.1	30.7 ± 0.1
Faster R-CNN	ReLU	29.3 ± 0.2	51.1 ± 0.2	29.8 ± 0.2
	ELU	28.9 ± 0.1	50.2 ± 0.1	29.4 ± 0.2

Table B: **Compatibility with different teacher initializations for RetinaNet ReLU conversion on DOTA-v1.5.**

Method	ImageNet			DINO			No pretraining		
	mAP	mAP_{50}	mAP_{75}	mAP	mAP_{50}	mAP_{75}	mAP	mAP_{50}	mAP_{75}
Softmax reference	33.5	58.1	33.2	31.4	53.8	31.5	25.9	46.5	25.2
Hidden-state distillation	28.1	49.7	27.7	28.7	49.7	28.5	26.0	46.4	25.3
DiD	**31.1**	**54.5**	**31.0**	**31.0**	**53.1**	**31.5**	**26.2**	**46.7**	**25.4**

Table C: **Label-free Softmax-to-linear conversion on Pascal VOC with ViT-S.** Full training is a supervised reference; all other adaptation methods use no detection labels.

Detector	Method	ReLU			ELU		
		mAP	mAP_{50}	mAP_{75}	mAP	mAP_{50}	mAP_{75}
RetinaNet	Softmax reference	66.6	90.3	75.2	66.6	90.3	75.2
	Full training	60.3	85.5	67.4	59.4	84.8	66.2
	Direct copy	3.2	7.0	1.8	2.1	5.0	1.0
	Generic distillation	46.5	75.8	49.2	38.0	67.1	37.2
	Softmax-generated pseudo-label distillation	54.0	82.2	58.4	55.6	83.5	60.7
	Hidden-state distillation	57.5	84.3	63.8	56.1	83.2	61.9
	DiD (Ours)	**60.8**	**86.2**	**68.0**	**59.7**	**85.2**	**66.6**
Faster R-CNN	Softmax reference	66.0	89.8	75.8	66.0	89.8	75.8
	Full training	57.6	83.4	64.5	56.7	82.7	63.1
	Direct copy	0.7	1.6	0.3	0.4	1.0	0.1
	Generic distillation	44.5	74.5	46.4	35.8	64.0	34.6
	Softmax-generated pseudo-label distillation	54.3	82.4	59.7	53.7	81.9	58.4
	Hidden-state distillation	55.4	82.6	61.3	55.0	82.1	60.7
	DiD (Ours)	**58.0**	**84.1**	**65.3**	**56.9**	**83.2**	**63.8**

C Ablation on Teacher Initialization

We test ImageNet initialization, DINO initialization (Caron et al., 2021), and detector training without backbone pretraining. Table B shows that DiD remains stronger than hidden-state distillation for all three teachers. The final converted accuracy nevertheless depends on the interface learned by the original detector: stronger Softmax teachers generally yield stronger converted models.

In the no-pretraining setting, DiD slightly exceeds the Softmax reference by 0.3 mAP. We treat this small difference as a setting-specific result rather than evidence that conversion generally improves the teacher. Overall, DiD is compatible with all tested teacher initializations, and its converted accuracy inherits the operating point of the learned detector interface.

D Generalization to Pascal VOC

Table C applies the same conversion protocol and baselines as the main DOTA-v1.5 study to Pascal VOC (Everingham et al., 2010) to test whether the conversion behavior transfers beyond aerial imagery.