RT-CO-DETR: Enhancing Real-Time Detectors via Multi-Level Knowledge Distillation from a Detector Teacher

Tran Hoang Nam
FPT University
Ho Chi Minh City, Vietnam
hnamt04@gmail.com

Tran Gia Hien
FPT University
Ho Chi Minh City, Vietnam
hientgse181821@fpt.edu.vn

Tran Hoang Phuc
FPT University
Ho Chi Minh City, Vietnam
phucthse183342@fpt.edu.vn

Bui Nhat Anh
FPT University
Ho Chi Minh City, Vietnam
nhatanh011204@outlook.com

***Abstract*—Knowledge distillation offers a promising avenue for enhancing real-time object detectors, yet its efficacy is often hindered by architectural mismatches. This paper investigates this challenge by first empirically demonstrating the limitations of distilling from foundation model backbones. Our initial experiments show that a DINOv3-trained ConvNeXt teacher provides only a marginal 6.00% mAP@.50-.95, while a Vision Transformer (ViT) teacher degrades performance to just 2.90% mAP on the TACO dataset, both falling drastically short of the baseline. Motivated by this critical finding, we introduce RT-CO-DETR, a novel framework that pivots to a detector-to-detector distillation strategy using an architecturally congruent teacher: Conditional DETR. Our core contribution is a holistic, multi-level distillation across features, classification logits, and bounding box regression. Evaluated on the challenging TACO dataset, RT-CO-DETR achieves 26.10% mAP, a significant 9.2% relative improvement over the baseline, while critically preserving the original inference speed and parameter count. Our work validates that a comprehensive, multi-level distillation from an architecturally similar teacher is a more practical and effective methodology for developing state-of-the-art, efficient object detectors.**

***Index Terms*—Object Detection, Knowledge Distillation, RT-DETR, Vision Transformer, DINOv3, Conditional DETR, Trash Detection, TACO Dataset.**

I. Introduction

Real-time object detection is a cornerstone of many practical computer vision systems, from autonomous driving to environmental monitoring. Models like the Real-Time DEtection TRansformer (RT-DETR) (Zhao et al., 2024) have emerged as strong contenders in this domain, offering an effective balance between the high accuracy of Transformer-based architectures and the computational efficiency required for deployment. However, a primary challenge remains: how to further enhance the performance of these models on complex, specialized datasets—such as the Trash Annotations in Context (TACO) dataset (Proença and Simões, 2020) with its highly varied and cluttered objects—without introducing significant inference latency.

The advent of large-scale, self-supervised foundation models, particularly Vision Transformers (ViTs) like DINOv3 (Siméoni et al., 2025), has opened new frontiers for performance enhancement. These models are renowned for producing rich semantic features, but their application is challenged by complex datasets like TACO, which features a wide variety of objects in cluttered scenes, as illustrated in Figure 1.

Fig. 1. Examples from the TACO dataset, showcasing the diversity of objects and challenging backgrounds, which motivate the need for robust detection models.

A prevailing trend to leverage these powerful backbones is through *architectural fusion*, where components of a foundation model are structurally integrated into a detector. Recent works have explored this path, proposing hybrid models that combine DINOv3 with architectures like YOLO (P et al., 2025) and RT-DETR itself (Huang et al., 2025). While these approaches show potential, they often require significant re-engineering and can compromise low-latency performance that makes models like RT-DETR practical.

An alternative, more deployment-friendly approach is *knowledge distillation* (KD) (Hinton et al., 2015), where a smaller "student" model learns from a larger "teacher" during training. However, the effectiveness of KD is not guaranteed. Our preliminary experiments explored a feature-level distillation from backbones trained with the DINOv3 self-supervised methodology. We tested two distinct teacher architectures: one

based on ConvNeXt (Liu et al., 2022) and another on a Vision Transformer. The results highlighted a critical challenge: while the ConvNeXt-based teacher provided a marginal performance of 6.00% mAP@.50-.95, the more architecturally dissimilar ViT-based teacher degraded performance to just 2.90% mAP, both starkly underperforming compared to the 23.90% mAP baseline. This provided a key insight: the distillation process is highly sensitive to the architectural mismatch between the teacher's backbone and the student's, limiting its potential.

Furthermore, evaluating any detection model requires a high-quality, reliably annotated dataset. Recognizing the scarcity of large-scale object detection datasets for litter, we initially attempted to create one by aggregating six public classification datasets and applying an automated labeling procedure using a pre-trained YOLOv8 model. This effort, however, revealed another critical pitfall: the generic detector produced bounding boxes with significant ***semantic mismatch***, leading to noisy and incorrect labels (e.g., assigning the 'Paper' label to a detected cigarette). This demonstrated that naive auto-labeling is an unreliable methodology for creating clean training data.

To isolate the effects of our distillation method from potential data quality issues, and to address the architectural mismatch challenge, we pivoted our approach. We selected the manually annotated TACO dataset (Proença and Simões, 2020) as a clean benchmark. In this paper, we introduce **RT-CO-DETR**, a model enhanced through a holistic, multi-level distillation from an architecturally congruent teacher. Instead of a generic classifier, we employ a more powerful object detector from the same family, **Co**nditional-DETR (Meng et al., 2023), as our teacher. Our key innovation lies in a comprehensive distillation strategy that transfers knowledge at all stages of the detection pipeline, inspired by seminal works in feature and prediction-level distillation:

- **Feature-level Loss:** Aligning student and teacher backbone features using Mean Squared Error, a technique pioneered by works like FitNets (Romero et al., 2015).
- **Classification Loss:** Transferring "dark knowledge" by minimizing the KL-Divergence between logits, as proposed in the original KD paper by (Hinton et al., 2015).
- **Regression Loss:** Supervising the student's bounding box predictions using the teacher's outputs via an L1 loss, a common practice in detector distillation (Wang et al., 2019).

This detector-to-detector, multi-level approach ensures a richer and more relevant knowledge transfer, effectively bridging the performance gap while critically **preserving the original, high-speed architecture of RT-DETR**. Furthermore, to ensure scientific validity, all experiments were conducted under a unified hyperparameter configuration, guaranteeing a fair comparison.

Our contributions are threefold:

1) We propose RT-CO-DETR, a model that significantly improves upon the RT-DETR baseline through a novel, multi-level distillation strategy with zero added inference cost.
2) We provide direct empirical evidence that a holistic, detector-to-detector distillation is superior to simpler, feature-only distillation from classifier backbones, especially when architectural mismatches and data quality issues are considered.
3) We validate that our framework is a practical and powerful method for creating state-of-the-art, real-time object detectors, offering a compelling alternative to complex architectural fusion.

Project code and Preprocessed data are publicly available at:
https://github.com/nam-htran/RT-CO-DETR
https://github.com/nam-htran/RT-DisDINOV3
https://www.kaggle.com/datasets/tranhoangnamk18hcm/dsp-pre-final

II. Related Work

In this section, we review prior work in three key areas relevant to our research: the evolution of real-time object detectors, the application of knowledge distillation to detection tasks, and the recent trend of leveraging large-scale foundation models.

A. Real-Time Object Detection

The field of object detection has seen a rapid evolution, traditionally dominated by CNN-based, single-stage detectors like the YOLO family (Redmon et al., 2016), which prioritized inference speed. A paradigm shift occurred with the introduction of DEtection TRansformers (DETR) (Carion et al., 2020), which framed object detection as a direct set prediction problem, eliminating the need for complex hand-designed components like anchor generation and non-maximum suppression. However, the original DETR suffered from slow convergence and high computational cost.

To bridge this gap, subsequent works like Conditional DETR (Meng et al., 2023) introduced mechanisms like conditional cross-attention to accelerate training convergence. More recently, RT-DETR (Zhao et al., 2024) presented a successful hybrid architecture, combining a CNN backbone for efficient feature extraction with a lightweight Transformer decoder. This design achieves a compelling balance between speed and accuracy, making it a strong baseline for real-time applications, and serves as the student model in our work. Our choice of Conditional-DETR as the teacher is motivated by its architectural similarity and superior performance within the same DETR family.

B. Knowledge Distillation in Object Detection

Knowledge Distillation (KD), first popularized by (Hinton et al., 2015), is a technique where a compact "student" model is trained to mimic the behavior of a larger, more powerful "teacher" model. The original formulation focused on transferring "dark knowledge" by matching the softened probability distributions of the classification logits using KL-Divergence. This concept was extended by works like FitNets (Romero et al., 2015), which introduced the idea of distilling

knowledge from intermediate feature maps, forcing the student to learn similar internal representations.

Applying KD to object detection is more complex as it involves both classification and localization tasks. Seminal works in this area (Wang et al., 2019) have shown that effective distillation requires transferring knowledge from all parts of the detector, including feature pyramids, classification heads, and bounding box regression heads. The application of KD to enhance RT-DETR, specifically, has also been explored. For instance, (Liang et al., 2025) proposed RT-DETR-FFD, a knowledge distillation-enhanced model for the specific task of fabric defect detection. Their work successfully improved the accuracy of a lightweight RT-DETR student without increasing computational overhead, further validating the potential of KD as an effective optimization strategy for the RT-DETR family.

The principle of mitigating architectural mismatch, which is central to our work, is also recognized in other advanced distillation strategies. (Cao et al., 2023) introduced Multi-Teacher Progressive Distillation to bridge the large "capacity gap" between powerful Transformer-based teachers and CNN-based students. By using a chain of intermediate teachers, they progressively transfer knowledge, highlighting the importance of addressing the teacher-student discrepancy. In contrast, our RT-CO-DETR framework presents a more direct and simpler alternative by selecting an architecturally congruent teacher from the outset, thereby avoiding the need for a complex multi-stage training pipeline. Our proposed RT-CO-DETR framework builds upon these principles by implementing a comprehensive, multi-level strategy that combines feature, classification, and regression distillation in a single, unified training process.

C. Foundation Models in Detection

The rise of large-scale, self-supervised foundation models like DINOv3 (Siméoni et al., 2025) has provided a new source of powerful, general-purpose visual representations. A common strategy to leverage these models is through *architectural fusion*, where the foundation model's backbone is directly integrated into a detector architecture (P et al., 2025; Huang et al., 2025). While promising, this approach can increase model complexity and latency. An alternative is to use these backbones as teachers in a feature-level distillation setup. However, as our preliminary experiments in the Introduction highlighted, this approach is vulnerable to the *architectural mismatch* between teacher and student, which can severely limit the effectiveness of the knowledge transfer. Our work pivots from this approach towards a detector-to-detector strategy to mitigate this very issue.

III. Methodology

Our research methodology is twofold. First, we conduct an ablation study to investigate the impact of *architectural mismatch* in knowledge distillation. Second, based on the insights from this study, we propose our primary framework, **RT-CO-DETR**, which employs a multi-level, detector-to-detector distillation strategy.

A. Ablation Study: Distillation from Foundation Models

To empirically test the effects of architectural mismatch, we first implemented a simpler, feature-level distillation framework, as detailed in Algorithm 2. In this setup, the student model's backbone was trained to mimic the intermediate feature representations of two different teacher backbones, both pre-trained using the **DINOv3** self-supervised methodology:

- **ConvNeXt Teacher:** A modern CNN architecture (Liu et al., 2022), representing a teacher with a relatively similar inductive bias to the student's CNN backbone.
- **ViT Teacher:** A pure Vision Transformer architecture, representing a teacher with a significantly different, attention-based inductive bias.

For both cases, distillation was performed *only* at the feature level, using projection layers and an MSE loss. This controlled experiment was designed to isolate the impact of the teacher's architecture on the final performance of the fine-tuned detector.

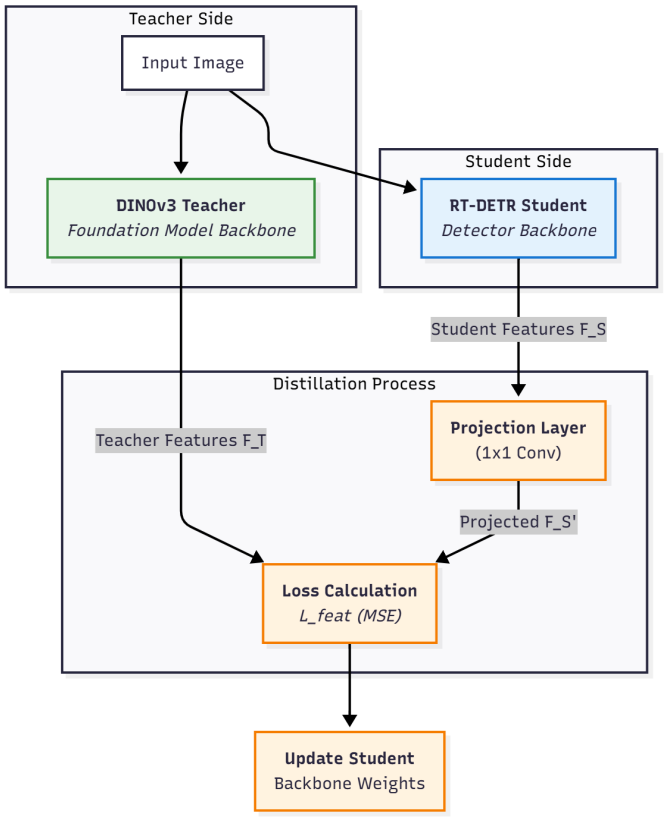


Fig. 2. The feature-only distillation framework used in our ablation study. The student backbone learns directly from a general-purpose DINOv3-trained teacher backbone. This setup highlights the challenge of architectural mismatch.

B. Proposed Framework: RT-CO-DETR

Our proposed framework, **RT-CO-DETR**, is designed to enhance the baseline RT-DETR model without altering its architecture or increasing inference cost. It uses a multi-level knowledge distillation process where a student model,

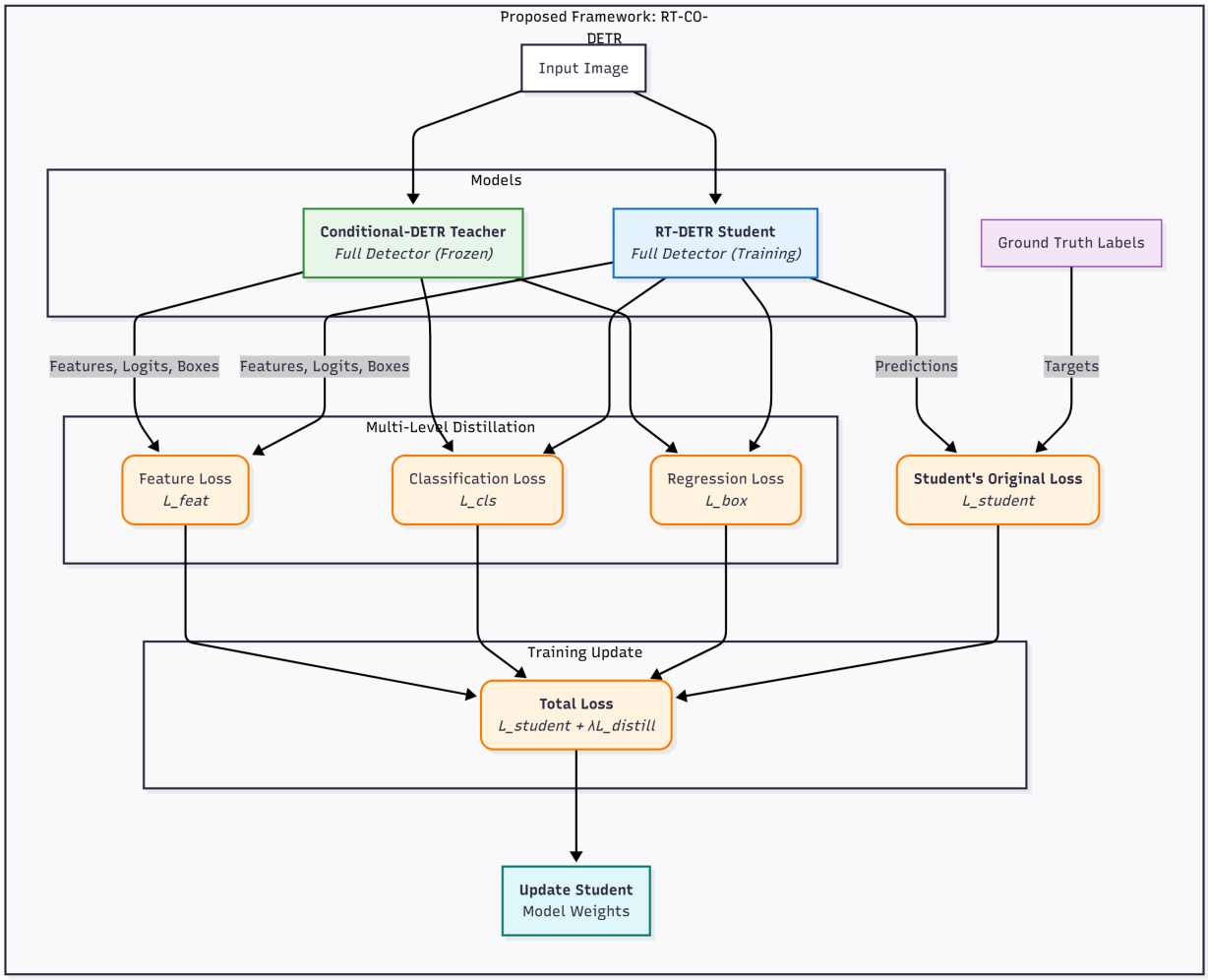


Fig. 3. The overall architecture of our proposed RT-CO-DETR framework. The student model (RT-DETR, right) learns from an architecturally congruent teacher (Conditional-DETR, left) via three distinct distillation losses. This detector-to-detector approach transfers knowledge at the feature-level (L_{feat}), classification-level (L_{cls}), and regression-level (L_{box}), ensuring a comprehensive knowledge transfer.

S (full RT-DETR), learns from a larger teacher model, T (full **Co**nditional-DETR). The overall training is guided by a composite loss function:

$$L_{\text{total}} = L_{\text{student}} + \lambda_{\text{feat}} L_{\text{feat}} + \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{box}} L_{\text{box}} \quad (1)$$

where L_{student} is the original loss of RT-DETR, and λ values are hyper-parameters balancing each distillation term. The framework is depicted in Figure 3, and the comprehensive training procedure is detailed in Algorithm 1.

1) Feature-level Distillation: To encourage the student's backbone to learn richer spatial representations, we align its intermediate feature maps with those of the teacher. Let F_i^S and F_i^T be the feature maps from the i-th stage. We use a 1x1 convolution ($Conv_{1x1}$) as a projection layer. The loss is calculated as the Mean Squared Error (MSE):

$$L_{\text{feat}} = \sum_{i \in \text{stages}} \left\| \text{Conv}_{1x1}(F_i^S) - \text{Interpolate}(F_i^T) \right\|_2^2 \quad (2)$$

2) Classification Distillation: To transfer the nuanced *"dark knowledge"* from the teacher's classification head, we use the Kullback-Leibler (KL) Divergence loss on the softened probability distributions, controlled by a temperature parameter τ. A critical challenge is the mismatch in the number of output classes. We address this by slicing the logits to a common dimension before applying the loss:

$$L_{\text{cls}} = \tau^2 \cdot \text{KLDiv}\left(\sigma(Z^S/\tau), \sigma(Z^T/\tau)\right) \quad (3)$$

where σ is the Softmax function.

3) Regression Distillation: To improve the student's localization accuracy, we directly supervise its bounding box predictions using the teacher's outputs. Let B^S and B^T be the predicted bounding box coordinates. The regression distillation loss is a simple L1 loss:

$$L_{\text{box}} = \left\| B^S - B^T \right\|_1 \quad (4)$$

IV. Experiments

A. Dataset and Metrics

All our experiments are conducted on the **Trash Annotations in Context (TACO)** dataset (Proença and Simões, 2020).

Algorithm 1 RT-CO-DETR Training Procedure

Input: Training dataloader $\mathcal{D}_{\text{train}}$, Student model S, Teacher model T, Projection layers P.
Hyperparameters: Epochs E, learning rate η, loss weights $\lambda_{\text{feat}}, \lambda_{\text{cls}}, \lambda_{\text{box}}$.
Initialize Optimizer for parameters of S and P with η.
Freeze parameters of teacher model T.
for epoch = 1 to E **do**
  for images, ground_truth in $\mathcal{D}_{\text{train}}$ **do**
    // Get teacher outputs in no_grad mode
    $F^T, Z^T, B^T \leftarrow T(\text{images})$
    // Get student outputs
    $F^S, Z^S, B^S \leftarrow S(\text{images})$
    // Calculate student's original loss
    $L_{\text{student}} \leftarrow \text{ComputeLoss}(Z^S, B^S, \text{ground\_truth})$
    // Calculate distillation losses
    $L_{\text{feat}} \leftarrow \text{MSELoss}(P(F^S), F^T)$
    $L_{\text{cls}} \leftarrow \text{KLDivLoss}(Z^S, Z^T, \tau)$
    $L_{\text{box}} \leftarrow \text{L1Loss}(B^S, B^T)$
    // Calculate total loss
    $L_{\text{total}} \leftarrow L_{\text{student}} + \lambda_{\text{feat}} L_{\text{feat}} + \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{box}} L_{\text{box}}$
    // Backpropagation
    Perform backward pass on L_{total}.
    Update parameters of S and P.
  end for
end for

Algorithm 2 Feature-level Distillation from Foundation Model (Ablation Study)

Input: Training images $\mathcal{D}_{\text{images}}$, Student backbone S_{bb}, Teacher backbone T_{bb}, Projection layers P.
Hyperparameters: Epochs E, learning rate η.
Initialize Optimizer for parameters of S_{bb} and P with η.
Freeze parameters of teacher backbone T_{bb}.
for epoch = 1 to E **do**
  for images in $\mathcal{D}_{\text{images}}$ **do**
    // Get teacher features in no_grad mode
    $F^T \leftarrow T_{\text{bb}}(\text{images})$
    // Get student features
    $F^S \leftarrow S_{\text{bb}}(\text{images})$
    // Calculate feature distillation loss
    $L_{\text{feat}} \leftarrow \text{MSELoss}(P(F^S), F^T)$
    // Backpropagation
    Perform backward pass on L_{feat}.
    Update parameters of S_{bb} and P.
  end for
end for

To ensure a standardized and robust evaluation benchmark, we first applied a comprehensive preprocessing pipeline to the original dataset. The initial 1,500 images were analyzed, revealing and removing a small number of perceptual duplicates, resulting in a clean set of 1,498 unique images. All unique images were then resized to a fixed resolution of 640x640 pixels, and their corresponding bounding box annotations were scaled accordingly. Finally, this cleaned and standardized dataset was randomly shuffled and split into training and validation sets using an 85/15 ratio, yielding 1,273 images for training and 225 for validation. This process mitigates potential biases from duplicate data and standardizes the input, ensuring that performance differences are attributable to the models themselves. For evaluation, we use standard COCO metrics, including mean Average Precision (*mAP@50-95*), *mAP@50*, and *mAP@75*. We also report inference *speed* (ms) and *model parameters* (M) to evaluate efficiency.

B. Experimental Setup

To validate our approach, we compare several models:

- **RT-DETR-L (Baseline):** The standard RT-DETR-L model fine-tuned on TACO.
- **YOLOv11-L (Baseline):** A strong baseline from the YOLO family for reference.
- **RT-DisDINOv3 (Ablation):** Our initial experiments as described in Section III-A, serving as an ablation study on architectural mismatch.
- **RT-CO-DETR (Proposed):** Our final model using multi-level distillation from a Conditional-DETR teacher.

To ensure a *fair comparison*, all models were trained under a unified hyperparameter configuration.

C. Implementation Details

Our framework is implemented using **PyTorch** and trained in a distributed manner on 2 NVIDIA T4 GPUs using 'torchrun'. For full reproducibility, our training used the *AdamW* optimizer with a base learning rate of 5×10^{-5} and a weight decay of 1×10^{-4}. All models were trained for 50 epochs with a batch size of 16 per GPU (total batch size 32) and utilized an input image size of 640x640. For the distillation process, the critical loss weights were set to $\lambda_{\text{feat}} = 1.0$, $\lambda_{\text{cls}} = 0.8$, and $\lambda_{\text{box}} = 1.2$, with a temperature of $\tau = 4.0$. **All final performance metrics, including inference speed, were measured on a single NVIDIA T4 GPU using a unified benchmark script, which is provided in our project repository.**

V. Results and Discussion

The quantitative results of all our experiments, evaluated on the TACO validation set, are summarized in Table I.

A. Performance Analysis

Our proposed method, **RT-CO-DETR**, clearly demonstrates the effectiveness of a holistic, detector-to-detector distillation strategy. As shown in Table I, it achieves an mAP@.50-.95 of **26.10%**, which represents a significant **relative improvement**

Fig. 4. Qualitative comparison on a challenging, cluttered image from the TACO dataset, evaluated at a high confidence threshold of 0.70. From left to right: RT-DETR-L (Baseline), our proposed RT-CO-DETR, and YOLOv11-L (as a reference baseline). Our model (center) demonstrates a superior ability to detect a significantly greater number of objects with high confidence compared to both baselines.

TABLE I
MAIN PERFORMANCE COMPARISON ON THE TACO VALIDATION SET. THE TABLE INCLUDES BASELINES, RESULTS FROM OUR ABLATION STUDY ON ARCHITECTURAL MISMATCH, AND OUR PROPOSED MODEL'S FINAL PERFORMANCE. ALL METRICS WERE GENERATED BY A UNIFIED BENCHMARK SCRIPT ON A SINGLE NVIDIA T4 GPU.

Model	mAP@.50-.95	mAP@.50	Speed (ms)	Params (M)	FLOPs (G)
Baselines					
RT-DETR-L (Baseline)	23.90%	30.00%	59.68	40.92	136.06
YOLOv11-L (Baseline)	25.66%	29.60%	30.87	25.36	87.53
Ablation Study: Feature-only Distillation from DINOv3					
RT-DisDINOv3 (ConvNeXt)	6.00%	8.20%	58.67	40.92	136.06
RT-DisDINOv3 (ViT)	2.90%	4.00%	58.12	40.92	136.06
Proposed Method					
RT-CO-DETR (Proposed)	**26.10%**	**31.00%**	**58.50**	**40.92**	**136.06**

of 9.2% over the 23.90% mAP of the standard RT-DETR-L baseline. This substantial gain is achieved with **zero added computational cost**, as both the parameter count (40.92 M) and FLOPs (136.06 G) remain identical to the baseline. Critically, the inference speed is also preserved (58.50 ms vs. 59.68 ms). Furthermore, our distilled model not only improves upon its own architectural family but also surpasses the strong **YOLOv11-L baseline** in accuracy (26.10% vs. 25.66% mAP), proving its competitiveness for applications where precision is paramount.

The results from our **ablation study** on DINOv3-based teachers provide crucial context. Distilling from a DINOv3-ViT teacher resulted in a dismal 2.90% mAP, and the architecturally more similar DINOv3-ConvNeXt teacher only yielded 6.00% mAP. Both results are drastically lower than the 23.90% mAP of the standard RT-DETR-L baseline and validate our central hypothesis: **architectural mismatch** between a general-purpose teacher and a specialized student detector severely hinders knowledge transfer. The superior performance of RT-CO-DETR confirms that a multi-level distillation from an architecturally congruent teacher is a far more effective approach.

B. Qualitative Analysis

The dramatic impact of our distillation framework is best illustrated through a qualitative comparison, shown in Figure 4. On a highly cluttered scene from the TACO dataset, the performance difference between the models is stark, especially when evaluated at a stringent confidence threshold of 0.70.

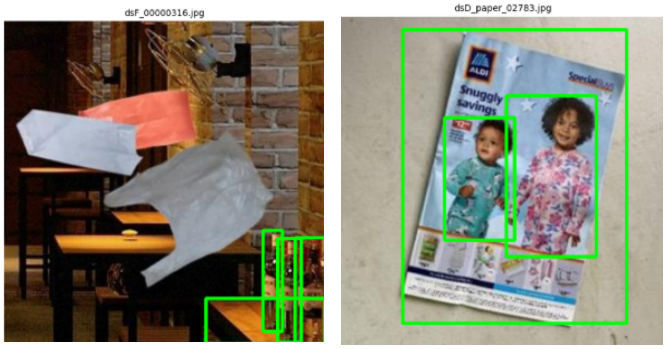
The **RT-DETR baseline** (left) struggles significantly at this high threshold, correctly identifying only a single plastic bottle while failing to recognize the majority of objects. The **YOLOv11-L baseline** (right) performs better by detecting several plastic items, but it still misses many objects that are clearly visible in the heap.

In stark contrast, even at this demanding confidence level, our proposed **RT-CO-DETR** model (center) demonstrates a transformative improvement. It successfully identifies a significantly larger number of objects, all with high confidence scores well above the 0.70 threshold. This visual evidence strongly suggests that our multi-level distillation process does not just marginally improve the model, but fundamentally enhances its ability to discern and localize a diverse range of objects with high certainty in complex, real-world scenarios. The student model has effectively learned the teacher's more sophisticated understanding of the scene.

C. On the Challenges of Automated Annotation

A significant aspect of our initial research was an attempt to construct a large-scale object detection dataset for litter. We aggregated several public image classification datasets and employed a pre-trained YOLOv8 model to automatically generate bounding boxes. However, this effort revealed a critical flaw we term **semantic mismatch**. The generic detector, pre-trained on COCO, proficiently identified common objects like 'person' or 'bottle', but our pipeline then incorrectly assigned the class label of the source image folder to these detections.

This resulted in absurd and detrimental labels, such as a bounding box around a person being labeled as 'Paper' or background bottles being labeled as 'Plastic', as illustrated in Figure 5.



(a) Source class: 'Plastic'. (b) Source class: 'Paper'.

Fig. 5. Examples of **semantic mismatch** from our naive auto-labeling pipeline. (a) In an image intended for the 'Plastic' class, the generic YOLOv8 model detects background bottles and incorrectly labels them as 'Plastic'. (b) In an image for the 'Paper' class, the model detects people and incorrectly labels them as 'Paper'.

This process created a dataset where bounding boxes were syntactically correct but *semantically wrong*, providing a noisy and counter-productive training signal that would effectively teach a model to associate the features of a person with the 'Paper' label. This finding underscores a broader challenge: the transition from image-level classification to object-level detection is non-trivial. Our experiment highlights the immense value of manually curated benchmarks like TACO and serves as a strong cautionary tale. It validates our decision to pivot to a clean, reliable dataset for **rigorous scientific evaluation**, ensuring our final results are attributable to our distillation method and not distorted by data artifacts. It also suggests that future efforts in low-cost dataset creation must focus on more sophisticated techniques, such as weakly-supervised or human-in-the-loop systems, to avoid this pitfall.

VI. Conclusion and Future Work

In this paper, we introduced **RT-CO-DETR**, a framework for enhancing real-time object detectors through a holistic, multi-level knowledge distillation process. We first established the limitations of naive approaches, highlighting the ineffectiveness of using architecturally dissimilar backbones and the unreliability of automated annotation for custom datasets. Our proposed solution addresses these issues by using a powerful, architecturally congruent teacher (Conditional-DETR) on a clean, manually annotated benchmark. By transferring knowledge across features, classification, and regression, our method significantly boosted the performance of RT-DETR on the TACO dataset, surpassing not only its own baseline but also a strong YOLO-family competitor in terms of accuracy, all without any additional inference cost.

Our work validates that a comprehensive, detector-to-detector distillation is a practical and powerful methodology. Building upon these findings, we suggest several promising directions for future research:

- **Semi-Supervised Learning with Pseudo-Labeling:** The large, aggregated classification dataset we initially built, despite being unsuitable for naive auto-labeling, presents a valuable resource. A future direction could be to use our trained RT-CO-DETR model to generate high-confidence pseudo-labels on this unlabeled (in terms of bounding boxes) data. A new student model could then be retrained on a combination of the original TACO data and this newly labeled set, potentially leading to further performance gains through a self-training loop.
- **Post-Distillation Efficiency Enhancement:** While RT-CO-DETR improves accuracy, a speed gap remains when compared to architectures like YOLO. Future work could investigate the application of model compression techniques, such as *quantization* or *pruning*, to the distilled model. This could help bridge the latency gap while preserving the accuracy benefits gained from distillation.
- **Exploring Advanced Distillation Techniques:** Our framework relies on classic distillation losses. An interesting avenue for exploration would be to employ more advanced, Transformer-specific distillation methods, such as *attention-based distillation*, to better transfer the relational knowledge captured within the attention maps of the teacher's Transformer layers.

References

Shengcao Cao, Mengtian Li, James Hays, Deva Ramanan, Yi-Xiong Wang, and Liang-Yan Gui. Learning lightweight object detectors via multi-teacher progressive distillation, 2023. URL https://arxiv.org/abs/2308.09105.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. URL https://arxiv.org/abs/2005.12872.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Shihua Huang, Yongjie Hou, Longfei Liu, Xuanlong Yu, and Xi Shen. Real-time object detection meets dinov3, 2025. URL https://arxiv.org/abs/2509.20787.

Gengliang Liang, Shijia Yu, and Shuguang Han. Rt-detr-ffd: A knowledge distillation-enhanced lightweight model for printed fabric defect detection. *Electronics*, 14:2789, 07 2025. doi: 10.3390/electronics14142789.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. URL https://arxiv.org/abs/2201.03545.

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence, 2023. URL https://arxiv.org/abs/2108.06152.

Malaisree P, Youwai S, Kitkobsin T, Janrungautai S, Amorndechaphon D, and Rojanavasu P. Dino-yolo: Self-supervised

pre-training for data-efficient object detection in civil engineering applications, 2025. URL https://arxiv.org/abs/2510.25140.

Pedro F Proença and Pedro Simões. Taco: Trash annotations in context for litter detection, 2020. URL https://arxiv.org/abs/2003.06975.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016. URL https://arxiv.org/abs/1506.02640.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015. URL https://arxiv.org/abs/1412.6550.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation, 2019. URL https://arxiv.org/abs/1906.03609.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection, 2024. URL https://arxiv.org/abs/2304.08069.