EFFICIENT MATCHING-FREE DISTILLATION FOR DETECTION TRANSFORMERS VIA ACTIVE SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Most existing knowledge distillation approaches for DETR-based detectors depend on query matching between teacher and student models, typically utilizing Hungarian matching algorithms, which are inefficient and time consuming. To mitigate these limitations, we propose an effective and efficient distillation framework that obviates the need for matching. Specifically, we introduce a novel active sampling and alignment strategy tailored for matching-free knowledge distillation. In our approach, the output from both the teacher and student models queries are regarded as representations of their corresponding output distributions. Then, with appropriate sampling points, we concurrently sample from both distributions and then enforce consistency between the sampled outcomes, thereby aligning the distribution between teacher and student. For the sampling procedure, we devise a simple but effective attention-based sampling module, complemented by a dedicated learning strategy for effective distribution sampling. Additionally, for the selection of sampling points during distillation, we propose a prior-guided point sampling approach that more accurately captures the teacher's output distribution, enhancing alignment with the student's distribution. Extensive experiments conducted across multiple datasets and baseline detectors validate that our method substantially enhances the performance of the student model. Compared to the DETR-Distill, our approach achieves superior performance while accelerating the distillation training process by 3.8 times. The code is available in the supplementary materials and will be publicly released upon acceptance of this paper.

1 INTRODUCTION

In recent years, DETR-based methods (Carion et al., 2020; Zhu et al., 2020; Dai et al., 2021; Zhang et al., 2022; Roh et al., 2021) have demonstrated considerable potential in tasks such as object detection. DETR innovatively introduces the Transformer (Vaswani et al., 2017; Dosovitskiy et al., 2020) architecture into object detection, eliminating the reliance on traditional anchor boxes and post-processing steps (Redmon et al., 2016; Girshick, 2015). This approach not only streamlines the object detection pipeline but also achieves performance comparable to, or even surpassing, conventional methods on benchmark datasets such as COCO (Lin et al., 2014).

Despite the remarkable performance of DETR-based detectors, their high computational cost poses significant challenges for deployment in real-time applications. Knowledge distillation (KD) (Hinton et al., 2015) emerges as an effective model compression technique, optimizing the training of a student model by minimizing the discrepancy between its outputs and those of a high-performing teacher model (e.g., through temperature-scaled softmax distributions). This process enables the transfer of knowledge from a large, high-performance teacher model to a smaller student model, thereby achieving a balance between performance and efficiency in the student model's detection results.

Recent studies have investigated the application of knowledge distillation (KD) techniques to DETR-based detection frameworks. Since DETR formulates object detection as a set prediction task by employing object queries to represent potential targets, the distillation process typically involves aligning query-level predictions between teacher and student models. However, this procedure is computationally expensive and time-consuming. Specifically, DETR often requires a large number of queries (e.g., 300), and the matching operation—commonly performed via bipartite matching al-

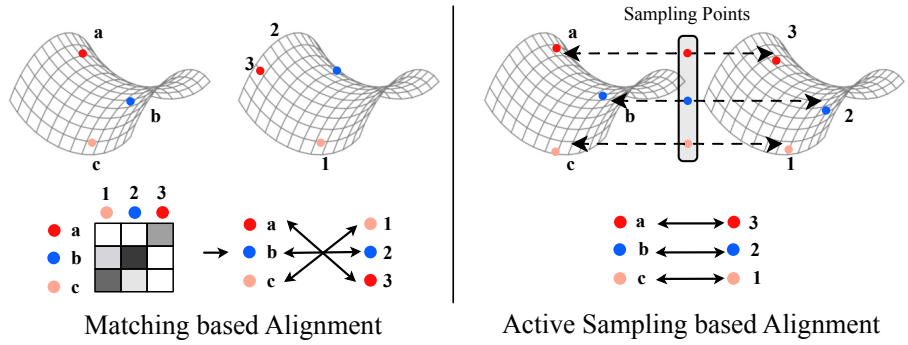

Figure 1: Comparison between matching-based methods and active sampling methods. Matching-based methods require the computation of a cost matrix between points, followed by the establishment of correspondences based on the matrix. These steps are often time-consuming, and the computational cost becomes even more pronounced when the number of points to be matched is large. In contrast, active sampling methods do not require the computation of a cost matrix, making them more convenient and efficient.

gorithms such as the Hungarian method on CPUs—introduces substantial overhead, thereby limiting training efficiency. For example, experiments with Deformable DETR (Zhu et al., 2020) demonstrate that the matching step alone accounts for a significant portion of the overall training time. To mitigate these limitations, several strategies have been proposed. DETRDistill (Chang et al., 2023), for instance, introduces a Query-prior Assignment mechanism that incorporates the teacher's queries as an additional prior for the student. This design encourages the student to generate predictions conditioned on the teacher's stable bipartite matching outcomes, thereby circumventing the need for redundant matching with the teacher's outputs. Empirical results suggest that this approach accelerates convergence and enhances the student's performance. Similarly, KD_DETR (Wang et al., 2024) adopts dedicated object queries that disentangle detection from distillation, ensuring consistent supervision between teacher and student models while avoiding explicit feature alignment through matching. Despite their merits, these methods remain limited. Relying directly on the teacher's bipartite matching outcomes may be suboptimal, as discrepancies in feature distributions between teacher and student can undermine the quality of transferred knowledge. Ablation studies in DETRDistill further indicate that the Query-prior Assignment strategy yields only marginal gains (e.g., +0.4 AP), suggesting that such straightforward alignment may be insufficient for effective optimization. In the case of KD_DETR, additional query points are required to enable the distillation process, which increases the model's complexity. Moreover, this approach may suffer from the issue of inconsistent predictions at identical query points, which can arise from feature discrepancies between the teacher and student models.

To address the limitations of existing knowledge distillation approaches for DETR-based detectors, we propose a novel framework that eliminates the reliance on inefficient and time-consuming Hungarian matching. Specifically, we introduce an active sampling and alignment method designed to achieve query distribution alignment between teacher and student models without explicit matching. The key idea is to approximate distribution alignment indirectly by enforcing consistency between sampled outcomes derived under a tailored sampling strategy. To this end, we treat the query outputs of both teacher and student models as representations of their respective prediction distributions and design a lightweight yet effective module that performs sampling directly from these outputs. By jointly sampling from both models and aligning the resulting samples, the student distribution is guided toward that of the teacher. Furthermore, during distillation, we propose a prior-based sampling scheme, which enhances the representativeness of selected sampling points and strengthens distribution alignment, ultimately improving the effectiveness of knowledge transfer. Extensive experimental results demonstrate that our method not only substantially reduces the complexity of the distillation process but also improves the performance of the distilled model, exhibiting significant advantages over previous approaches. We believe our work will bring insights into the related fields.

The main contributions of this paper can be summarized as follows:

- This study introduces a novel distillation method for DETR-based models, which leverages active sampling and alignment to perform knowledge transferring, eliminating costly matching operations and yielding substantial gains in efficiency and performance.
- We propose an effective sampling module specifically tailored to the prediction space of query-based detectors. The module projects the prediction outputs of individual queries into a high-dimensional representation space, wherein the query outputs can be more effectively sampled due to enhanced separability.
- During distillation, we further propose a prior-based sampling strategy that selects sampling points capable of better preserving the distributional characteristics of the teacher model, thereby leading to more effective alignment between teacher and student distributions.
- Extensive experiments conducted on multiple datasets and several popular DETR-based detectors validate the effectiveness of our proposed method. The results show that our approach achieves highly competitive performance and efficiency compared to existing methods.

2 Related Works

DETR based Detection

The Transformer architecture has demonstrated remarkable success in natural language processing tasks, prompting researchers to explore its application to vision tasks. The seminal work, Detection Transformer (DETR), introduced an end-to-end Transformer-based object detector that eliminates the need for post-processing steps. Unlike traditional object detection methods, DETR reformulates object detection as a set prediction problem, optimized via bipartite matching. However, DETR-based methods suffer from challenges such as slow convergence, which has spurred a series of subsequent improvements aimed at addressing these limitations. Deformable DETR (Zhu et al., 2020) enhances DETR by introducing a deformable attention module that generates reference points for query elements. Each reference point attends to a limited set of positions on the feature map, reducing interference from background noise and irrelevant regions. This focused attention mechanism accelerates convergence by prioritizing target regions. Conditional DETR (Meng et al., 2021) further improves convergence by incorporating additional prior information into the decoder's object queries. It decouples contextual and positional features within the queries and generates positional features based on spatial locations, thereby reducing the reliance on global content in cross-attention mechanisms and expediting optimization. Building upon Conditional DETR, DAB-DETR (Liu et al., 2022) integrates width and height information into positional features to model objects at varying scales more effectively. Anchor DETR employs predefined anchor points as initial query representations, providing a stable starting point that mitigates instability caused by random initialization during early training. By encoding anchor points into object queries through multiple patterns, Anchor DETR (Wang et al., 2022) enhances adaptability to complex scenes, further accelerating convergence. Additionally, it introduces row-column decoupled attention to reduce memory costs. DN-DETR (Li et al., 2022) proposes a denoising-based approach by introducing noisy query samples and training the model to predict denoised queries. This query denoising task accelerates training by enhancing optimization stability. Furthermore, DN-DETR incorporates a group-based one-to-many label assignment strategy to increase supervisory signals, thereby improving convergence speed. Similarly, H-DETR (Jia et al., 2023) enhances supervision by introducing multiple positive queries as decoder inputs, strengthening target supervision and further accelerating convergence.

Several recent studies have focused on improving the efficiency of Transformer-based object detection methods. Sparse DETR (Roh et al., 2021) enhances computational efficiency by sparsifying the encoder's computations, thereby reducing training costs and accelerating convergence. By leveraging sparse encoder outputs, the decoder can still effectively perform object queries and predictions while maintaining detection performance with minimal degradation. Efficient DETR (Yao et al., 2021) proposes the integration of a Region Proposal Network (RPN) to generate object queries, initializing them with dense priors. This approach improves convergence speed and, due to the higher quality of initial queries, allows for the use of fewer decoder layers, resulting in reduced computational complexity and faster inference. PnP DETR (Wang et al., 2021) introduces a Poll and Pool (PnP) sampling module designed to selectively extract critical information from feature maps before

feeding them into the Transformer encoder. This selective processing reduces unnecessary computations, enhancing overall efficiency. The DINO series (Zhang et al., 2022) builds upon these advancements by incorporating novel techniques and scaling up both model and dataset sizes to further unlock the potential of DETR-based architectures, achieving improved performance while maintaining computational feasibility.

Knowledge distillation

Given the focus of this paper on improving knowledge distillation algorithms for DETR-based object detection, this section reviews relevant works on knowledge distillation under the DETR framework. As mentioned above, unlike traditional CNN-based object detectors, which rely on spatial feature map alignment for distillation, DETR represents a paradigm shift in object detection by reformulating it as a set prediction problem using a Transformer encoder-decoder architecture. The learnable and permutation-invariant nature of object queries in DETR introduces a lack of consistent distillation points between teacher and student models, posing significant challenges for knowledge distillation in this context. KD-DETR (Zhang et al., 2022) addresses the alignment issue by introducing a set of non-learnable, shared "probing queries" that serve as consistent distillation points between the teacher and student models. By decoupling the detection and distillation tasks, KD-DETR establishes a general paradigm for DETR distillation. Similarly, DETRDistill (Chang et al., 2023)leverages DETR's native Hungarian matching algorithm to align predictions between teacher and student models. It first identifies optimal one-to-one correspondences between their prediction sets and then applies a response-based distillation loss to the matched pairs. Additionally, DETRDistill incorporates an object-aware feature distillation method to enhance the student's learning of object-centric features. OD-DETR (Wu et al., 2024) focuses on stabilizing and accelerating DETR's notoriously slow training through knowledge distillation. It employs an online distillation strategy where a teacher model, constructed via exponential moving average (EMA), guides the student. Specifically, OD-DETR transfers the teacher's learned query-to-ground-truth box matching relationships and even initial object queries to the student, significantly improving distillation efficiency without introducing additional parameters. CLoCKDistill (Lan & Tian, 2025) introduces a Query-to-Feature (Q2F) module that aligns teacher model queries with specific locations on the encoder's feature map, enabling effective distillation on the feature map to transfer precise positional and contextual knowledge. Knowledge Distillation via Query Selection (Liu et al., 2024) observes that many queries correspond to background regions, introducing noise if distilled directly. To mitigate this, it proposes distilling only queries matched to positive samples by the teacher, thus filtering low-quality supervision and improving both efficiency and accuracy. D3ETR (Chen et al., 2022) focuses on knowledge transfer within the decoder. It proposes multi-layer decoder distillation, requiring the student to mimic not only the teacher's final predictions but also the refinement process at each decoder step. To this end, D3ETR designs an attention-matrix-based alignment mechanism to address query misalignment across layers, significantly accelerating the student model's convergence. However, the aforementioned methods still exhibit certain limitations. For instance, some rely on additional queries or involve complex distillation strategies, which lack simplicity and efficiency, thereby constraining their practicality in real-world applications.

3 METHOD

3.1 PROBLEM FORMULATION

In the DETR-base object detection framework, the feature encoder processes the image input to produce a feature representation Z, while the decoder takes input Z and N learnable object queries $Q \in \mathbb{R}^{N_{\text{obj}} \times D}$, processing them through an M-layer Transformer network to generate the model's output:

$$Q' = \text{Decoder}(Q, Z), \quad Q' \in \mathbb{R}^{N_{\text{obj}} \times D}. \tag{1}$$

For each object query Q'_i, a feed-forward network (FFN) is applied to predicts object bounding boxes and class probabilities:

$$\hat{B}_i = \text{FFN}_b(Q'_i), \quad \hat{C}_i = \text{FFN}_c(Q'), \tag{2}$$

where $\hat{B} = \{\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_{N-1}\}$, $\hat{b}_i \in \mathbb{R}^{N \times 4}$ represents bounding box coordinates (center, width, height), and $\hat{C} = \{\hat{c}_0, \hat{c}_1, \ldots, \hat{c}_{N-1}\}$, $\hat{c}_i \in \mathbb{R}^{N \times (C+1)}$ denotes class probabilities. During training,

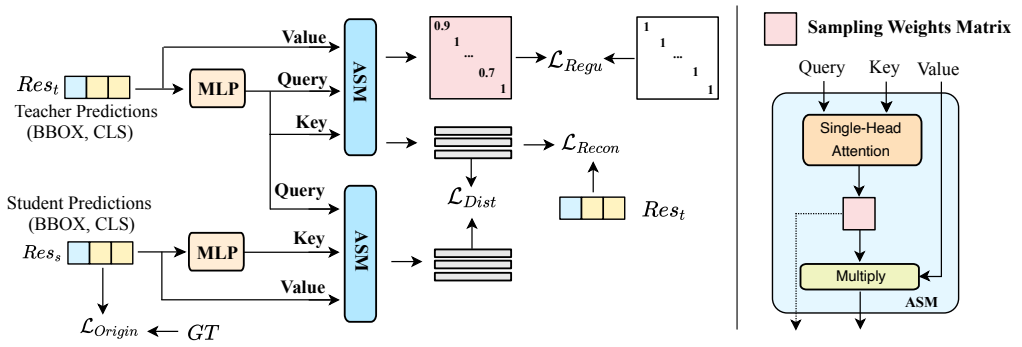

Figure 2: The flowchart of our proposed distillation framework based on active sampling alignment. For the teacher model predictions, including bounding boxes (BBOX) and classifications (CLS), are concatenated and subsequently processed through an MLP to generate queries and keys. These are then input into the ASM model, which outputs the single-head attention weights and the weighted sum of values. The attention weights are constrained to the diagonal of the matrix to ensure that each query has a high similarity score only with itself, allowing for precise sampling at specific locations. For distillation, the query from the teacher model is used as the sampling point, and sample results are generated from the student's outputs. A distillation loss is applied to align the sampling results from both the teacher's and student's predictions.

label assignment is formulated as the problem of minimizing the matching cost between model predictions and ground-truth (GT) annotations, yielding a bipartite matching through the Hungarian algorithm. The optimal matching is defined as:

$$\hat{\sigma} = \arg\min_{\sigma} \sum_{i=1}^{N} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma_i}), \tag{3}$$

where σ denotes a permutation of N elements and $\hat{\sigma}$ represents the optimal assignment. Each ground-truth instance is denoted as $y_i = (c_i, b_i)$, where c_i corresponds to the target class (which may be $\varnothing$) and b_i is the ground-truth bounding box. The pairwise matching cost $\mathcal{L}_{\text{match}}$ is defined as:

$$\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma_i}) = \mathcal{L}_{\text{cls}}(c_i, \hat{c}_{\sigma_i}) + \mathbf{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{bbox}}(b_i, \hat{b}_{\sigma_i}), \tag{4}$$

where $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{bbox}}$ denote the classification and bounding-box regression losses, respectively. Queries matched to ground-truth objects are treated as positive samples, while unmatched queries are supervised as negative samples. The overall detection loss is expressed as:

$$\mathcal{L}_{\text{det}}(y, \hat{y}_{\hat{\sigma}}) = \sum_{i=1}^{N} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma_i}), \tag{5}$$

In the context of knowledge distillation, where teacher and student models are involved. Conventional distillation methods typically require matching the teacher and student predictions before performing distillation, treating the teacher's outputs as pseudo-ground-truths. The distillation loss can be written as:

$$\mathcal{L}_{\text{distillation}} = \sum_{i=1}^{N} \mathcal{L}_{\text{match}}(\hat{y}_i^{tea}, \hat{y}_{\sigma_i}^{stu}), \tag{6}$$

The above fromulation establishes a relationship between the teacher's and student's predictions, guiding the student to approximate the teacher's outputs. However, this process has several limitations. First, the large number of queries in DETR models renders the matching process computationally expensive. Second, the teacher and student models may have different numbers of queries, making Hungarian matching ill-suited to handle such disparities, potentially leading to ignored or forced matches that reduce effective knowledge transfer. Finally, DETR's object queries are egocentric, initialized and optimized independently for each model, resulting in no fixed spatial or

semantic correspondence between teacher and student queries. While Hungarian matching attempts to establish one-to-one correspondences based on matching costs, it cannot ensure strict cross-model consistency, particularly for redundant negative queries.

To address these challenges, we propose a novel approach that leverages active sampling and alignment to eliminate the need for inefficient and time-consuming matching operations, which also demonstrates significant advantages in both efficiency and performance. For the teacher's output distribution Ω_{tea} and the student's output distribution Ω_{stu}, we select appropriate sampling points $P = \{p_0, p_1, ..., m-1\}$ in the prediction space and then sample from both distributions:

$$\hat{S}_{\text{tea}} = \textbf{Sampling}(\Omega_{\text{tea}}, P), \quad \hat{S}_{\text{stu}} = \textbf{Sampling}(\Omega_{\text{stu}}, P), \tag{7}$$

$\hat{S}_{\text{tea}}$ and $\hat{S}_{\text{stu}} \in \mathbb{R}^{m\times(4+c+1)}$, we then enforce consistency between the sampling results to align the student's output distribution with the teacher's:

$$\mathcal{L}_{\text{distillation}} = \mathcal{D}(\hat{S}_{\text{tea}}, \hat{S}_{\text{stu}}), \tag{8}$$

where $\mathcal{D}(,)$ is a distance measurement function to measure the distance between two inputs.

3.2 Active Sampling

Unlike traditional CNN-based models, which produce spatially continuous features $Out_{cnn} \in \mathbb{R}^{H\times W\times D}$, DETR-based methods generate outputs that depend on discrete, permutation-invariant queries. This property renders most traditional CNN sampling techniques, which rely on spatial correspondence, inapplicable. To overcome the challenge of aligning distributions over unordered outputs, we exploit the order-agnostic nature of the attention mechanism in Transformers to facilitate distribution alignment. We first provide a review of the attention mechanism computation in Transformers.

The attention mechanism operates on a set of vectors: queries $Q \in \mathbb{R}^{n\times d_k}$, keys $K \in \mathbb{R}^{m\times d_k}$, and values $V \in \mathbb{R}^{m\times d_v}$, where n and m denote the sequence lengths of queries and key-value pairs, respectively, and d_k and d_v represent the dimensionalities of the keys and values, respectively. These vectors are typically obtained through linear transformations of the input sequence. The attention scores are computed as the dot product between queries and keys, measuring their similarity. For scaled dot-product attention, the score is calculated as:

$$\text{score} = \frac{QK^T}{\sqrt{d_k}} \in \mathbb{R}^{n\times m}, \tag{9}$$

where $\sqrt{d_k}$ is a scaling factor to prevent large values in high dimensions. The attention scores are then normalized using the softmax function to obtain attention weights:

$$W = \text{softmax}\left(score\right). \tag{10}$$

Finally, the attention weights are used to compute a weighted sum of the value vectors, producing the final output:

$$\text{Attention}(Q, K, V) = W * V. \tag{11}$$

From the above attention output, we observe that when the weight distribution is highly concentrated, resembling a Dirac delta distribution:

$$w_i = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise}, \end{cases} \tag{12}$$

Then the attention procedure transforms into a sampling operation from the set of values $V = \{v_0, v_1, \ldots, v_{n-1}\}$. Notably, this sampling process is independent of the order of items in the set.

Building on the above observation, we propose employing an order-agnostic sampling mechanism to align the output distributions of teacher and student models in DETR-based detector knowledge distillation. Specifically, we interpret the query operation in attention as a sampling process over the

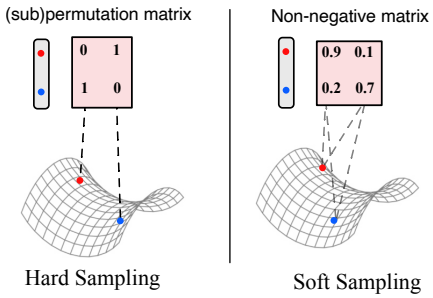

Figure 3: This figure provides a conceptual illustration contrasting Hard Sampling with Soft Sampling mechanism. On the left, Hard Sampling is depicted as a discrete selection process, utilizing a (sub)permutation matrix to enforce a strict one-to-one mapping where each input query is matched to a single, distinct point on the target manifold. In contrast, our Soft Sampling approach, shown on the right, transforms this into a differentiable and probabilistic process. It employs a non-negative matrix to synthesize a new representation as a weighted average of all available target points. This allows each input query to create a blended, "soft" sample that incorporates rich information from the entire set, rather than being limited to a single hard choice.

key space, where the query vector functions as the sampling point. Given the teacher model's output $\hat{Y}_{\text{tea}} = \{\hat{y}_0^{tea}, \hat{y}_1^{tea}, \ldots, \hat{y}_{N-1}^{tea}\}$ and the student model's output $\hat{Y}_{\text{stu}} = \{\hat{y}_0^{stu}, \hat{y}_1^{stu}, \ldots, \hat{y}_{N-1}^{stu}\}$, which correspond to the distributional Ω_{tea} and Ω_{stu}, respectively. Our objective is to sample the outputs from both the student and teacher at the same position p_i. Based on the above discussion, the sampling process can be reformulated as an attention operation. Without loss of generality, we illustrate the sampling process using the teacher model as an example:

$$w_i = \text{softmax}\left(\frac{p_i\, Q_{\text{tea}}^T}{\sqrt{d_k}}\right) \in \mathbb{R}^{n\times m}, \tag{13}$$

$$\text{Attention}(Q, K, V) = w_i * \hat{Y}_{\text{tea}}. \tag{14}$$

Notably, each sampling point p_i should lie within the same distributional space as the outputs of the model. In the following subsection, we will provide a detailed description of the selection of sampling points p_i. In practice, the weight distribution w_i does not necessarily conform to a Dirac delta distribution. Consequently, the sampling process becomes a weighted selection over the entire space, a procedure that we refer to as *soft sampling* in this work.

3.3 Key Component in Knowledge Distillation

With the module designed in the preceding subsection, we are able to actively sample the unordered outputs of both the teacher and student models. Knowledge distillation can then be performed according to Equations (7) and (8). In what follows, we present a detailed discussion of several critical factors that influence the distillation process, namely the sampling points selection, single-head attention, and distillation temperature setting in knowledge distillation.

Sampling Points Selection. Previous matching-based studies have demonstrated that the selection of positive and negative samples plays a crucial role in distillation. Analogously, in our active sampling and alignment approach, the choice of sampling points is equally critical. For a sampling point p_i to be effective, it must first lie within the output distribution space of the model to ensure the validity of the sampled outcomes. Furthermore, the distribution of the selected sampling points should closely approximate that of the teacher model, as this alignment enhances the representativeness of the sampled results with respect to the teacher's distribution. Lastly, the sampling points should be computationally feasible to obtain. Based by this analysis, in our experiments, we adopt the teacher model's query outputs directly as the sampling points.

Single-Head Attention. In standard Transformer architectures, input features are typically partitioned into m groups, with each group independently executing the attention operation described in Equations (9–11). The outputs of these groups are subsequently concatenated to produce the

final output, a procedure commonly referred to as multi-head attention. This mechanism enables the model to concurrently capture diverse aspects of the input sequence, such as semantic and syntactic information, thereby enhancing the flexibility and representational capacity of the attention operation. However, in the context of this work, where the objective is to select specific outputs at precise locations within the distribution space, the multi-head attention mechanism is suboptimal. Consequently, we employ a single-head attention mechanism to ensure that, during each sampling iteration, every sampling point corresponds to a unique output within the distribution space.

Distillation Temperature. As discussed in Section 3.2, the weight w in the sampling module tends to yield highly concentrated distributions. In practice, however, we observe that such concentration can impede distillation efficiency, since multiple student queries may predominantly align with a single teacher query. To address this issue, we introduce a temperature coefficient to soften the weights during the distillation process:

$$w_i = \text{Softmax}\left(\frac{p_i Q_{\text{tea}}^T}{\sqrt{d_k}}/T\right), \tag{15}$$

where T denotes the distillation temperature, which smooths the weight distribution, enabling more distributed query alignments and improving the distillation process.

3.4 Overall loss

The distillation training process is supervised by multiple loss, which are described in detail below.

Sampling Loss: For training the sampling module, we introduce a regularization loss on the sampling weights W, Specifically, for the sampling process $s_i = \text{sampling}(\Omega, \hat{y}_i)$, we aim to encourage the sampling weight w_i to approach 1 while all other weights approach 0. To this end, we design an entropy-based regularization loss to supervise the weight distribution:

$$\mathcal{L}_{sampling} = -\sum_{i=0}^{m-1} \log(W_{ii}). \tag{16}$$

The above loss is designed to ensure that the weights can be concentrated on specific areas in the distribution space during sampling.

Distillation Loss: The distillation involves losses for both bounding box (*bbox*) and classification (*cls*) predictions. For the bounding box distillation, we employ the L1 loss to enforce consistency between the student and teacher models. For classification prediction distillation, we employ the Kullback-Leibler (KL) divergence to align the student's class probabilities with those of the teacher:

$$\begin{aligned} \mathcal{L}_{distill} &= \mathcal{L}_{distill}^{bbox} + \mathcal{L}_{distill}^{cls} \\ &= \lambda_{bbox}\mathcal{L}_{L1}(S_{stu}^{bbox}, S_{tea}^{bbox}) + \lambda_{cls}\mathcal{L}_{kl}(S_{stu}^{cls}, S_{tea}^{cls}), \end{aligned} \tag{17}$$

where S_{stu}^{bbox} and S_{tea}^{bbox} represent the student and teacher bounding box predictions, S_{stu}^{cls} and S_{stu}^{cls} denote the teacher and student class probability distributions, and λ_{bbox} and λ_{cls} are the loss coefficient to balance the contribution $\mathcal{L}_{L1}(,)$ and $\mathcal{L}_{cls}(,)$.

Overall Loss: The overall loss function for the training process is a weighted combination of the DETR training loss and the distillation-specific losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{detr}} + \mathcal{L}_{\text{sampling}} + \mathcal{L}_{distill}. \tag{18}$$

4 Experiments

4.1 Setup

Datasets This study employs the demanding large-scale MS COCO benchmark, with train2017 (118K images) applied for training and val2017 (5K images) for validation. Evaluation is performed using the standard COCO-style metric, specifically average precision (mAP). The mAP is the average AP over 10 different IoU thresholds and across all 80 classes.

Table 1: Performance Comparison of Different Detectors and Settings on COCO Dataset.

Detector	Setting	Epoch	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
AdaMixer	Teacher	12	45.3	64.6	49.2	27.3	48.3	61.9
	Student	12	42.3	61.2	45.6	25.3	44.8	58.2
	FGD	12	40.7 (-1.6)	59.3	43.4	23.4	43.3	55.8
	MGD	12	42.3 (+0.0)	61.3	45.5	24.5	45.0	58.9
	FitNet	12	42.9 (+0.6)	61.7	46.2	24.7	45.8	59.4
	LD	12	41.4 (-0.7)	60.4	44.7	23.6	44.2	57.6
	DETRDistill	12	44.7 (+2.4)	62.9	48.2	26.7	47.6	61.0
	Ours	12	**45.2 (+2.9)**	**63.4**	**49.1**	**27.5**	**48.4**	**61.9**
Deformable DETR	Teacher	50	44.8	64.1	48.9	26.5	48.3	59.6
	Student	50	44.1	63.2	47.9	27.0	47.4	58.3
	FGD	50	44.1 (+0.0)	63.1	48.0	25.9	47.7	58.8
	MGD	50	44.0 (-0.1)	63.1	48.0	25.9	47.3	58.6
	FitNet	50	44.9 (+0.8)	64.3	48.9	27.2	48.4	59.6
	LD	50	43.7 (-0.4)	62.4	47.2	25.3	46.8	58.8
	DETRDistill	50	46.6 (+2.5)	65.6	50.7	28.5	50.0	60.4
	Ours	50	**47.0 (+3.0)**	**66.5**	**51.1**	**29.0**	**51.2**	**60.2**

Table 2: Per-Sample training latency comparison. * indicates the setting used to achieve the performance reported in the original work. † is the setting we used to achieve the performance reported in this work.

Method	Configuration	Latency (ms)↓
Base Student	-	193
Distill with DETRDistill	6 Layers*	335*(+142)
	5 Layers	322
	3 Layers	281
	1 Layer	236
Distill with Ours	2 Layers†	230†(+37)
	1 Layer	221

DETR Models This investigation evaluates three distinct DETR-based detection frameworks: Deformable DETR, and AdaMixer. These models were selected owing to their representative architectural designs and demonstrated superior performance. In the context of the ablation study, AdaMixer was adopted as the baseline for experimental analysis and parameter tuning, attributed to its facile training procedure and expedited convergence characteristics.

Implementation Details All models are trained on 4 NVIDIA V100 GPUs. Unless otherwise specified, we train the teacher model for 1× schedule (12 epochs) or 50 epochs using ResNet-101 as the backbone with Adam optimizer, and train the student model with the same learning schedule using ResNet-50 as the backbone, following each baseline's settings.

4.2 Main results

We conduct experiments on multiple baselines to assess the effectiveness of our proposed matching-free DETR distillation for object detection tasks, including FGD (Yang et al., 2022), MGD (Yang et al., 2022), FitNet (Romero et al., 2014), LD (Zheng et al., 2022), and DETRDistill (Chang et al., 2023). To ensure a fair comparison with previous work, we select two base detectors (AdaMixer and Deformable-DETR) and followed their official experiments setting unless otherwise specified. As detailed in Table 4, our method consistently enhances performance in all settings. In detail, with AdaMixer (Gao et al., 2022), our approach yields a 2.9 mAP gain, elevating the student's performance to nearly match the teacher model. When using Deformable-DETRZhu et al. (2020), we achieve a 3.0 mAP gain, surpassing all other listed distillation baselines.

Table 3: Performance Comparison of Different Distillation Methods on COCO Dataset

Multi-head attention (nums=8)				One-head attention			
AP	**AP_S**	**AP_M**	**AP_L**	**AP**	**AP_S**	**AP_M**	**AP_L**
46.2	28.4	50.0	59.8	47.0	29.0	51.2	60.2

Table 4: An ablation study on the effect of distillation temperature on final performance.

Distillations	**AP**	**AP_S**	**AP_M**	**AP_L**
Baseline	25.4	11.2	28.5	37.1
+ours (T=1)	26.3	9.7	28.9	40.3
+ours (T=2)	26.9	10.7	29.6	39.5
+ours (T=4)	28.5	11.8	31.8	41.3
+ours (T=8)	26.8	10.5	30.1	39.7

4.3 Computational Cost

In addition to the performance improvements detailed in our main results, we also evaluate the computational efficiency of our proposed method. A key aspect of a practical distillation framework is ensuring that training process does not incur significant computational overhead. To this end, we measure and compare the training latency of our approach against the DETRDistill applied in different number of layers. The results are presented in Table 2. The base student model training has an initial latency of 193 ms. As shown, the DETRDistill method introduces a considerable latency increase, which grows substantially from 236 ms with one layer to 335 ms with six layers. In contrast, our configuration adds a training latency of only 37 ms, making it approximately **3.8 times faster** than it in the official DETRDistill setting. Furthermore, our model shows better scalability, with only a minor increase to 230 ms for a two-layer setup. This analysis demonstrates that our matching-free distillation not only achieves the strong mAP performance reported in Table 4 but does so with a lower computational cost, making it a more efficient and practical solution.

4.4 Ablation study

We conducted an experimental analysis on the number of heads in the attention module of the ASM module. It can be observed that using the conventional multi-head attention mechanism actually leads to a performance decline. This is because, in our case, the attention operation is used for the sampling process, and multiple heads complicate the computation of the sampling weights, ultimately affecting performance.

Parameter Anaysis: We conducted an ablation study to investigate the effect of distillation temperature on final performance, with the results presented in Table 4. It can be observed that the distillation performance reaches its optimal point when the temperature coefficient is equal to 4. Additionally, distillation performance decreases on both sides of this value.

5 Conclusion

This work presents a novel, matching-free knowledge distillation strategy for DETR-based detectors, overcoming the significant computational bottleneck of query matching in existing methods. Our core contribution is a prior-driven, learnable sampling mechanism that implicitly aligns student and teacher queries, completely sidestepping the need for direct, costly matching. We believe this approach is a leap towards truly matching-free supervision for Detection Transformers. By eliminating this cumbersome step, our method not only accelerates the training process but also offers new insights to the community, paving the way for more efficient and practical deployment of state-of-the-art object detection models.

REFERENCES

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Jiahao Chang, Shuo Wang, Hai-Ming Xu, Zehui Chen, Chenhongyi Yang, and Feng Zhao. Detrdistill: A universal knowledge distillation framework for detr-families. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6898–6908, 2023.

Xiaokang Chen, Jiahui Chen, Yan Liu, and Gang Zeng. D˜3 etr: Decoder distillation for detection transformer. *arXiv preprint arXiv:2211.09768*, 2022.

Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2988–2997, 2021.

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.

Ziteng Gao, Limin Wang, Bing Han, and Sheng Guo. Adamixer: A fast-converging query-based object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5364–5373, 2022.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

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

Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19702–19712, 2023.

Qizhen Lan and Qing Tian. Clockdistill: Consistent location-and-context-aware knowledge distillation for detrs. *arXiv preprint arXiv:2502.10683*, 2025.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13619–13627, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.

Yi Liu, Luting Wang, Zongheng Tang, Yue Liao, Yifan Sun, Lijun Zhang, and Si Liu. Knowledge distillation via query selection for detection transformer. *arXiv preprint arXiv:2409.06443*, 2024.

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3651–3660, 2021.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse detr: Efficient end-to-end object detection with learnable sparsity. *arXiv preprint arXiv:2111.14330*, 2021.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. arxiv 2014. *arXiv preprint arXiv:1412.6550*, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Tao Wang, Li Yuan, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Pnp-detr: Towards efficient visual analysis with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4661–4670, 2021.

Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2567–2575, 2022.

Yu Wang, Xin Li, Shengzhao Weng, Gang Zhang, Haixiao Yue, Haocheng Feng, Junyu Han, and Errui Ding. Kd-detr: Knowledge distillation for detection transformer with consistent distillation points sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16016–16025, 2024.

Shengjian Wu, Li Sun, and Qingli Li. Od-detr: online distillation for stabilizing training of detection transformer. *arXiv preprint arXiv:2406.05791*, 2024.

Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4643–4652, 2022.

Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021.

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*, 2022.

Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Ming-Ming Cheng. Localization distillation for dense object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9407–9416, 2022.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

A APPENDIX

B APPENDIX

B.1 ANNOUNCEMENT FOR LLM TOOL USAGE IN THIS PAPER

We employed a large language model (Google's Gemini) as a general-purpose writing assistance tool during the final stages of manuscript preparation. The precise role of the LLM was confined to language enhancement, which included refining sentence structure, improving clarity, and checking for grammatical and typographical errors. All suggestions provided by the LLM were critically reviewed, and the authors made the final decisions on all textual modifications. We affirm that no part of the core research, including the ideation, methodology, and interpretation of results, was generated by the LLM. All authors have reviewed the final manuscript and assume complete responsibility for its content and scientific integrity.

B.2 REPRODUCIBILITY STATEMENT

To support the verification and extension of our research, we have made our source code available in the supplementary materials. The successful reproduction of our results can be guided by the "Setup" section within the main body of the paper.