TEF-GUIDED AND QUALITY-AWARE DISTILLED DETR FOR LANGUAGE-BASED AUDIO MOMENT RETRIEVAL

Technical Report

Yutao Xu, Liu Yang, Weixi Zheng

School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China
SheryTao@e.gzhu.edu.cn, yanlow2013@gzhu.edu.cn, 2112406058@e.gzhu.edu.cn

ABSTRACT

This technical report describes our system for Task 6 *language-based audio moment retrieval* of the DCASE 2026 challenge. The system builds upoin the official QD-DETR/AM-DETR baseline, which employs fixed CLAP audio and text features. The core design of the proposed system is a quality-aware distilled DETR framework for precise temporal grounding. It improves the baseline in three aspects, including query-conditioned audio representation, temporal-position encoding with temporal endpoint features, and localization-quality-aware candidate scoring with prediction-level consensus. Specifically, it integrates text-guided audio refinement, focal and IoU-aware quality objectives, boundary auxiliary supervision, span evidence adaptation, teacher distillation, recall-balanced checkpoint averaging, and weighted box fusion oriented to R1@0.7. Besides, since R1@0.7 is the primary competition target, our model selection prioritizes precise top-1 localization rather than broad recall alone. Experiment results show that, on the CASTELLA test set, the proposed system achieves improvements of 6.76 in R1@0.5, 5.79 in R1@0.7, and 3.47 in average mAP over the official baseline.

Index Terms— Audio moment retrieval, temporal grounding, DETR, CLAP, knowledge distillation, quality-aware ranking

1. INTRODUCTION

Language-based audio moment retrieval aims to identify the start and end timestamps of an audio segment that corresponds to a natural-language query [1]. Given a long audio recording and a free-form textual description, the system returns the temporal window in which the described acoustic event or scene occurs. This task is useful for audio archive search, multimedia indexing, surveillance audio analysis, human-computer interaction, and accessibility tools, where users need to locate short events in long recordings without manually listening to the entire file. Unlike clip-level audio-text retrieval, audio moment retrieval requires fine-grained temporal grounding and must rank candidate moments under strict temporal Intersection over Union (IoU) thresholds. Therefore, a successful system must jointly model semantic relevance, temporal boundaries, and the confidence of each candidate segment.

The audio setting introduces additional difficulties. Acoustic events may be short, repeated, overlapping, or semantically similar across different intervals, while long recordings often contain background sounds unrelated to the query. In addition, the same textual query may correspond to events with different durations or ambiguous acoustic contexts, making simple clip-level matching insufficient. DCASE 2026 Task 6 requires systems to retrieve moments from CASTELLA long audio using textual queries and to output ranked temporal windows. The official ranking emphasizes R1@0.7, so the top-1 predicted window must tightly overlap with the reference moment. The system must therefore estimate both query-event relevance and localization quality, especially for high-IoU predictions.

The DCASE 2026 Task 6 baseline provides fixed CLAP audio-text features [2] and a transformer-based temporal grounding pipeline related to DETR-style set prediction [3]. Our system retains this foundation while improving the downstream components most relevant to strict localization: TEF-guided token construction, text guided audio refinement, quality-aware scoring, boundary supervision, span evidence adaptation, distillation, checkpoint averaging, and R1@0.7-oriented prediction consensus. These components are designed to keep the robustness of the baseline feature representation while making the final ranking more consistent with precise temporal overlap. In particular, the system places stronger emphasis on candidates that are not only semantically matched to the query but also temporally compact and well aligned with the target event. This design also keeps the system compatible with the official feature format and evaluation protocol, so the improvements mainly come from localization-aware modeling and prediction ranking rather than from additional external audio features.

2. BASELINE SYSTEM

This section describes the feature format and prediction mechanism inherited from the official implementation. The long audio is divided into one-second clips, each represented by a 768-dimensional CLAP audio embedding. A temporal endpoint feature (TEF), consisting of normalized start and end positions, is appended to each audio clip feature. TEF provides explicit temporal position cues for language-conditioned localization, while the Transformer encoder supplies sequence-level positional modeling [4]. The text query is represented by a sequence of CLAP text features. Before being fed into the temporal grounding network, the audio and text features are projected to a hidden dimension of 256.

The DETR-based network consists of a text-to-audio cross-attention encoder, an audio temporal Transformer encoder, and a QD-DETR decoder with learnable temporal queries. Each decoder query predicts a candidate moment in normalized center-width format along with a foreground/background confidence score. Hungarian matching associates predictions with ground-truth moments [5], consistent with set-prediction detection frameworks [3]. The baseline training objective comprises a span L_1 loss, a generalized temporal IoU loss, a classification loss, auxiliary decoder losses, and a saliency loss. This design avoids hand-crafted proposal gen-

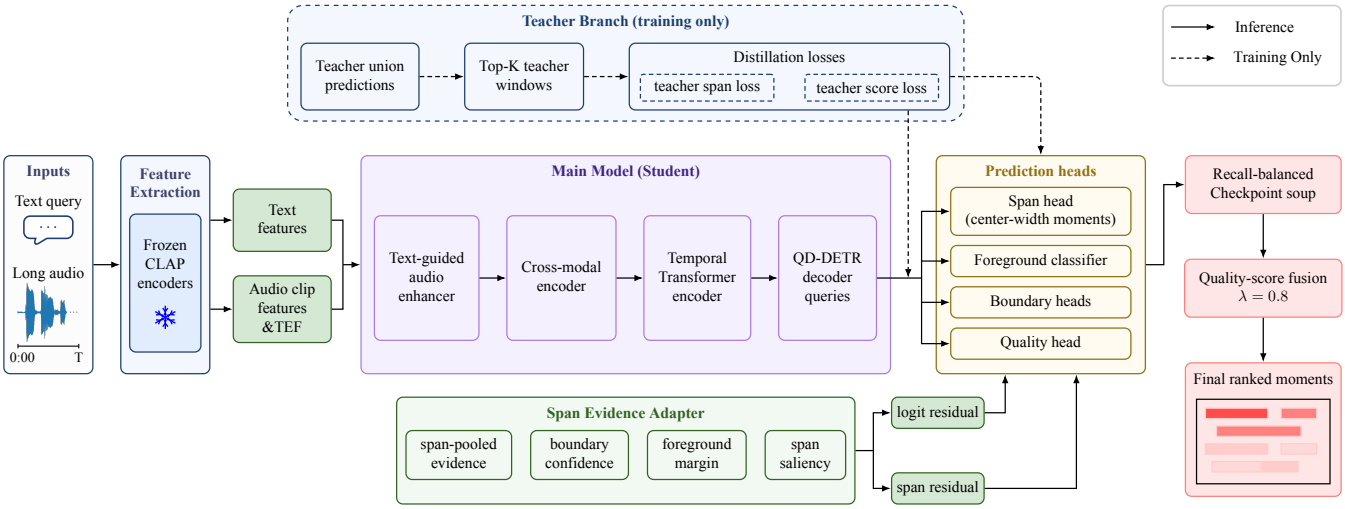

Figure 1: Overall diagram of the proposed TEF-guided and quality-aware distilled DETR system.

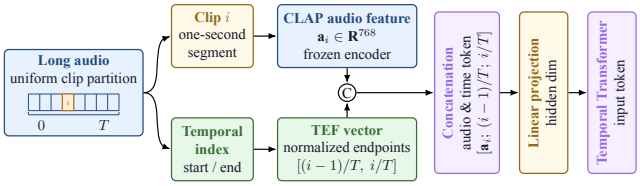

Figure 2: Illustration of TEF. For each one-second audio clip, normalized start and end positions are concatenated to the CLAP audio embedding before projection.

eration, but it places great importance on the quality of candidate scoring, as the final submission retains only a small number of ranked temporal windows per query.

3. PROPOSED SYSTEM

The proposed system keeps the baseline data pipeline and Transformer backbone, but modifies the representation learning, span calibration, and scoring components. Fig. 1 summarizes the overall diagram of the proposed pipeline. The system first constructs TEF-guided audio tokens, then applies query-conditioned temporal grounding with prediction heads and span evidence refinement, and finally performs teacher-guided calibration and prediction-level fusion to optimize the primary R1@0.7 metric. This process can be formulated as,

$$(\hat{\boldsymbol{y}}, \hat{\boldsymbol{c}}) = f_{\mathrm{DETR}}\left(g_{\mathrm{audio}}(\boldsymbol{z}_A, \boldsymbol{z}_T), \boldsymbol{z}_T\right), \tag{1}$$

where $\boldsymbol{z}_A$ and $\boldsymbol{z}_T$ denote projected audio and text features, $g_{\mathrm{audio}}()$ is the text-guided audio refinement module, and $f_{\mathrm{DETR}}()$ is the temporal DETR network. The outputs $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{c}}$ denote the predicted temporal moments and their confidence scores, respectively.

3.1. TEF-guided input token construction

The first stage of the proposed system is feature extraction from the long audio and text query. As illustrated in Fig. 2, each long recording is uniformly divided into one-second clips. For the i-th clip in an audio sequence of length T, the temporal endpoint feature (TEF) $\boldsymbol{t}_i$ is defined as the normalized start and end positions,

$$\boldsymbol{t}_i = [(i-1)/T,\ i/T],\ i = 1, 2, \cdots, T. \tag{2}$$

This two-dimensional vector is concatenated with the frozen CLAP audio embedding $\boldsymbol{a}_i$ to form a time-aware audio token $[\boldsymbol{a}_i; \boldsymbol{t}_i]$, which is then projected to the hidden dimension of the temporal grounding network. TEF supplies explicit position information before the Transformer encoder.

3.2. Text-guided audio enhancer

After TEF-guided token construction, the student branch applies a lightweight text-guided audio enhancer before the cross-modal Transformer. Audio tokens attend to text tokens, a pooled text representation generates FiLM-style modulation parameters, and a temporal convolution models local audio context. This FiLM-style conditioning follows the principle of feature-wise modulation conditioned on task information [6]. In our system, the modulation signal is derived from the text query, allowing the same audio clip sequence to be represented differently for different retrieval requests.

The text-guided audio enhancer is deliberately shallow. It adapts frozen CLAP features to the current query while preserving the original audio tokens through a residual path. The final configuration adopts a conservative single-scale temporal context, which yields the strongest test-side R1@0.7 performance.

3.3. Prediction heads and quality-aware losses

After the student temporal grounding model produces decoder queries, the prediction heads in Fig. 1 estimate foreground confidence, temporal boundaries, span coordinates, and localization quality. The original foreground/background classification objective is sensitive to the imbalance between a small number of matched foreground queries and many background queries. To reduce the influence of easy negatives, focal loss is employed [7]:

$$\mathcal{L}_{\mathrm{focal}} = -\alpha_t (1 - p_t)^{\gamma} \log p_t, \tag{3}$$

where p_t is the target-class probability, α_t is the class-balancing factor, and γ is the focusing parameter. In the experiments, $\alpha_t = 0.75$ and $\gamma = 2.0$.

In addition, an IoU-aware quality loss is added. For a matched prediction, the target quality is the temporal IoU between the predicted moment and the matched ground truth, following the localization-quality motivation of generalized IoU [8]. The ranking target is also related to quality-aware dense detection [9]. For unmatched queries, the target is zero. The foreground logit margin is trained with binary cross entropy (BCE) as,

$$\mathcal{L}_{\text{quality}} = \text{BCEWithLogits}(s_{\text{fg}} - s_{\text{bg}}, q), \tag{4}$$

where the scalar q is the detached IoU target, s_{fg} and s_{bg} are the foreground and background logits. This encourages confidence scores to reflect localization quality.

3.4. Boundary-aware auxiliary supervision

To improve boundary localization, we add start and end heads on top of the encoded audio memory. Ground-truth boundaries are represented by Gaussian soft targets, and the boundary loss is defined as,

$$\mathcal{L}_{\text{boundary}} = \frac{1}{2}\left(\text{BCE}(\hat{\boldsymbol{s}}_{\text{start}}, \boldsymbol{s}_{\text{start}}) + \text{BCE}(\hat{\boldsymbol{s}}_{\text{end}}, \boldsymbol{s}_{\text{end}})\right). \tag{5}$$

Here, $\hat{\boldsymbol{s}}_{\text{start}}$ and $\hat{\boldsymbol{s}}_{\text{end}}$ are predicted boundary scores, and $\boldsymbol{s}_{\text{start}}$ and $\boldsymbol{s}_{\text{end}}$ are Gaussian soft targets. This auxiliary task only regularizes the audio memory toward sharper temporal boundary information.

3.5. Quality-score calibration

Our system introduces an independent quality head to separate semantic foreground classification from localization quality estimation. The quality head employs decoder query features, boundary confidence, and span width to predict candidate quality. During inference, the final score is a fusion of foreground probability p_{fg} and quality score p_{quality}, i.e.,

$$\hat{c} = (1-\lambda)p_{\text{fg}} + \lambda p_{\text{quality}}, \tag{6}$$

where $\hat{c}$ is the calibrated score, p_{fg} is the foreground probability, p_{quality} is the quality score, and λ is the interpolation weight. The final score-calibrated system uses $\lambda = 0.8$.

3.6. Span evidence adapter

The quality head improves ranking but does not directly update span coordinates. A lightweight span evidence adapter is therefore added to pool evidence inside each predicted window and predict small residual updates to logits and normalized center-width spans:

$$(\Delta\boldsymbol{s}, \Delta\boldsymbol{y}) = h_{\text{span}}\left[\boldsymbol{q}, \boldsymbol{e}_{\text{span}}, \boldsymbol{q} \odot \boldsymbol{e}_{\text{span}}, b, w, m, r\right], \tag{7}$$

where $\boldsymbol{q}$ is the decoder query feature, $\boldsymbol{e}_{\text{span}}$ is span-pooled audio evidence, b is boundary confidence, w is span width, m is foreground margin, and r is span saliency. Moreover, $\Delta\boldsymbol{s}$ and $\Delta\boldsymbol{y}$ are residual corrections for logits and normalized center-width spans. The final layer is zero-initialized, so the adapter acts as checkpoint-compatible local refinement rather than replacing the DETR decoder.

3.7. Teacher distillation and checkpoint soup

To transfer complementary behavior into one model, a prediction-level teacher is built from the union of the original strong checkpoint and the span-evidence checkpoint. During student training, the top-K windows of the teacher are used as soft targets, and two auxiliary losses inspired by knowledge distillation are added [10],

$$\mathcal{L}_{\text{teacher}} = \lambda_{\text{span}}\mathcal{L}_{\text{span}} + \lambda_{\text{score}}\mathcal{L}_{\text{score}}, \tag{8}$$

in which $\mathcal{L}_{\text{span}}$ supervises student spans, $\mathcal{L}_{\text{score}}$ supervises ranking scores. The final configuration adopts $K = 3$, $\lambda_{\text{span}} = 2.0$, and $\lambda_{\text{score}} = 0.5$.

Conservative checkpoint averaging is then applied to stabilize the distilled student. The checkpoints selected by average mAP and mAP@0.5 are averaged with weights 0.8 and 0.2, following the intuition of model soups [11]. This distilled soup is further averaged with the original strong checkpoint using weights 0.7 and 0.3 to recover R1@0.5, producing the recall-balanced prediction set used by the final fusion stage.

3.8. R1@0.7-oriented prediction consensus

For the final submission, a lightweight prediction-level consensus stage is used. This stage fuses the recall-balanced prediction set with a distilled high-mAP anchor using weighted temporal box fusion, adapted from box-level ensembling in object detection [12]. The main prediction set receives weight 0.9 and the anchor receives weight 0.1. Candidate windows with temporal IoU above 0.85 are merged by score-weighted averaging of their boundaries. To emphasize robust high-IoU top-1 retrieval, small consensus bonuses are added for cross-source agreement and top-rank agreement:

$$s_{\text{fuse}} = \max_i s_i + \beta_n(n-1) + \beta_s(m-1) + \beta_r r, \tag{9}$$

where n is the number of clustered windows, m is the number of distinct prediction sources, and r is the average reciprocal rank inside the cluster. Moreover, s_i is the i-th candidate score, s_{fuse} is the fused score, and β_n, β_s, β_r weight count, source, and rank bonuses. The final run uses $\beta_n = 0.05$, $\beta_s = 0.02$, and $\beta_r = 0.02$.

This fusion uses only model predictions and confidence/rank information. Compared with selecting the strongest single checkpoint by R1@0.5 or mAP, the consensus stage slightly reduces broad recall but improves the main R1@0.7 metric and average mAP. This output is therefore used as the final submission candidate.

4. EXPERIMENTAL SETUP

Evaluation is performed on the CASTELLA validation and test splits using the official metric script. The metrics are Recall@1 at IoU thresholds 0.5 and 0.7, average mAP over thresholds from 0.5 to 0.95, mAP@0.5, and mAP@0.75. All metrics are reported to characterize both broad retrieval and high-precision localization, but R1@0.7 is used as the primary model-selection reference because it most directly measures precise top-1 moment retrieval.

Unless otherwise noted, the model uses hidden dimension 256, 2 encoder layers, 2 decoder layers, 8 attention heads, 10 temporal queries, batch size 32, AdamW optimization [13] with learning rate 10^{-4}, and 200 training epochs. The training schedule follows the

baseline configuration except for the additional objectives and fine-tuning stages described in Section 3. The final main loss is

$$\begin{aligned}\mathcal{L} =& \lambda_{\text{span}}\mathcal{L}_1 + \lambda_{\text{giou}}\mathcal{L}_{\text{gIoU}} + \lambda_{\text{focal}}\mathcal{L}_{\text{focal}} \\ &+ \lambda_{\text{quality}}\mathcal{L}_{\text{quality}} + \lambda_{\text{boundary}}\mathcal{L}_{\text{boundary}} \\ &+ \lambda_{\text{sal}}\mathcal{L}_{\text{saliency}} + \mathcal{L}_{\text{teacher}},\end{aligned} \tag{10}$$

where $\mathcal{L}_1$, $\mathcal{L}_{\text{gIoU}}$, and $\mathcal{L}_{\text{saliency}}$ denote span loss, temporal-IoU loss, and saliency loss. The corresponding weights to these objectives are 10, 1, 4, 1, 1, and 1 for span, gIoU, focal, quality, boundary, and saliency. The teacher terms are disabled in non-distillation runs.

5. RESULTS

Table 1 compares the final system with the official baseline on the CASTELLA test set. The primary R1@0.7 metric increases from 10.32 to 16.11, and all auxiliary metrics also improve. The gain is especially large for mAP@0.75, indicating that the proposed quality and boundary modeling improves high-IoU ranking rather than only broad candidate recall.

Table 2 summarizes the cumulative ablation results and the final fusion comparison, where "Bnd.", "Avg.", "Q-head", "Span", "Distill", "Bal-WBF", and "R1-WBF" denote boundary supervision, checkpoint averaging, quality-head calibration, span evidence adaptation, teacher distillation, balanced weighted box fusion, and R1@0.7-oriented weighted box fusion, respectively. The ablation study reveals that single-model modifications and prediction-level fusion contribute in different ways. Boundary supervision and quality-head calibration improve the ranking behavior of the baseline, whereas the span evidence adapter and distillation enhance mAP-oriented ranking. The final WBF configuration specifically increases the primary R1@0.7 metric from 15.74 to 16.11 while raising the average mAP from 12.43 to 12.58. Among all the normal-channel systems evaluated in this report, the final consensus WBF output achieves the best R1@0.7, average mAP, mAP@0.5, and mAP@0.75. Although it is slightly below the balanced prediction WBF on R1@0.5, the drop is modest compared to the gain on the primary R1@0.7 metric. Consequently, the submitted system is selected based on the R1@0.7 consensus WBF output.

Table 1: Comparison with the baseline on CASTELLA test set.

System	R1@0.5	R1@0.7	mAP	mAP@0.5	mAP@0.75
Baseline	23.16	10.32	9.11	20.34	6.96
Ours	**29.92**	**16.11**	**12.58**	**25.52**	**11.02**
Gain	+6.76	+5.79	+3.47	+5.18	+4.06

6. DISCUSSION

A trade-off is observed between broad recall and high-IoU top-1 localization. For instance, the balanced prediction fusion achieves the best R1@0.5, whereas the final R1@0.7-oriented WBF yields the best primary metric and the best average mAP. This trade-off is expected as a broader top candidate can increase overlap at an IoU threshold of 0.5 while reducing strict localization quality at IoU 0.7 or above. The final system therefore employs score calibration and consensus bonuses to favor candidates that are both semantically reliable and temporally precise.

Not every architectural extension generalized equally well to the test set. Some development variants improved validation metrics but led to a drop in test-side R1@0.7, suggesting that the current data regime favors conservative calibration, quality-aware ranking, and agreement-based prediction fusion over large backbone changes. Consequently, the final system adopts a checkpoint-compatible span evidence adapter, distillation from a prediction-level teacher, recall-balanced checkpoint averaging, and a compact consensus WBF stage, rather than a substantially larger model.

7. CONCLUSION

This report has presented a quality-aware distilled DETR system for DCASE 2026 Task 6. Starting from the official CLAP-based QD-DETR baseline, the system introduces text-guided audio refinement, TEF-based temporal-position encoding, focal classification, IoU-aware quality learning, boundary-aware auxiliary supervision, span evidence adaptation, teacher distillation, recall-balanced checkpoint averaging, separate quality-head calibration, and R1@0.7-oriented prediction consensus. On the CASTELLA test, the final normal-channel system improves R1@0.5 from 23.16 to 29.92, R1@0.7 from 10.32 to 16.11, and average mAP from 9.11 to 12.58. The final configuration is selected to prioritize precise top-1 localization rather than broad retrieval recall alone. Ablation results show that TEF-based temporal cues, localization-quality scoring, boundary supervision, and prediction consensus are complementary, namely, TEF stabilizes clip-level position information, the quality branch calibrates candidate confidence, and WBF favors temporally consistent windows across prediction sources.

Table 2: Ablation study and prediction-fusion comparison on the CASTELLA test set. The configuration columns indicate whether each component is used. R1@0.7 is the primary competition metric and is therefore used as the final selection criterion.

Type	Configuration							Test metrics				
	Bnd.	Avg.	Q-head	Span	Distill	Bal.-WBF	R1-WBF	R1@0.5	R1@0.7	mAP	mAP@0.5	mAP@0.75
student model	×	×	×	×	×	×	×	29.18	13.88	11.62	24.81	9.18
	√	×	×	×	×	×	×	28.51	15.00	11.73	24.19	9.87
	√	√	×	×	×	×	×	29.32	15.07	11.96	25.16	10.30
	√	√	√	×	×	×	×	29.77	15.07	11.98	25.22	10.30
	√	√	√	√	×	×	×	27.17	14.55	12.33	24.19	10.98
	√	√	√	√	√	×	×	28.21	15.22	12.36	24.54	10.88
prediction fusion	√	√	√	√	√	√	×	**30.14**	15.74	12.43	25.39	10.92
	√	√	√	√	√	√	√	29.92	**16.11**	**12.58**	**25.52**	**11.02**

8. REFERENCES

[1] H. Munakata, T. Nishimura, S. Nakada, and T. Komatsu, "Language-based audio moment retrieval," arXiv:2409.15672, 2025.

[2] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *Proc. ICASSP*, 2024.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. ECCV*, 2020.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017.

[5] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1–2, pp. 83–97, 1955.

[6] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, 2018.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. ICCV*, 2017.

[8] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. CVPR*, 2019.

[9] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. NeurIPS*, 2020.

[10] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531, 2015.

[11] M. Wortsman, G. Ilharco, S. Y. Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, and L. Schmidt, "Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time," in *Proc. ICML*, 2022.

[12] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*, vol. 107, 2021.

[13] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.