Efficient SwinDETR: Distillation Through Attention and Post Training Quantization

Sekwang Lee 20210807 **Chaerin Lee** [redacted] **Jihoon Lim** [redacted] **Myunglyul Lee** [redacted]

Abstract

In this work, we propose SwinDETR, which integrates a Swin Transformer backbone into the DETR framework to overcome the low-resolution feature limitations of CNN backbones. SwinDETR removes the class head from the Swin backbone and employs a fully connected layer to align its hierarchical feature maps with the embedding dimension of the DETR encoder, thereby enhancing detection performance across scales, especially for small objects. We further introduce trainable distillation queries inspired by the DeiT distillation token concept and a hard-label distillation scheme that uses teacher model predictions as pseudo-ground truths. Layer-wise cosine similarity analysis reveals that distillation and object queries remain highly similar, indicating limited representational separation and constraining distillation gains. Finally, we apply three post-training quantization methods—weight-only quantization, dynamic quantization, and SmoothQuant—to the trained SwinDETR model and validate their impacts on accuracy, inference speed, and model size.

1 Introduction

1.1 Motivation

DETR [1] is an end-to-end object detection model leveraging Transformers and Bipartite Matching, achieving strong performance without NMS or anchors. However, due to low-resolution feature maps produced by its CNN backbone (e.g., ResNet [2]), DETR underperforms Faster R-CNN [3] by 5.5 points on small object detection (AP_S). To address this, we propose SwinDETR, which replaces DETR's backbone with a Swin Transformer [4]. The Swin Transformer hierarchically merges patches to extract multi-resolution features and employs Shifted Window Self-Attention for inter-window information exchange. Additionally, we introduce a novel attention-based knowledge distillation method using Distillation Queries, and apply post-training quantization to improve the efficiency of the large-parameter SwinDETR model.

1.2 Related work

Multi-scale feature representation. To address CNN's low-resolution feature maps, FPN [5] integrates spatial and semantic information via bottom-up, top-down pathways and lateral connections, while Deformable DETR [6] uses multi-scale inputs and deformable attention to focus on key regions.

Distillation on Vision Transformer. DeiT [7] achieves strong performance on small datasets using a Distillation Token, inspiring the introduction of Distillation Queries in our DETR decoder. DearKD [8] injects CNN inductive bias into early ViT [9] layers to improve data efficiency and generalization.

Post-training quantization on Vision Transformer. CNN-centric quantization methods suffer larger accuracy drops on transformer models, leading to transformer-specific approaches like

SmoothQuant [10]. ViT-optimized quantization techniques such as FQ-ViT [11] and RepQ-ViT [12] have also been proposed.

2 Method

SwinDETR architecture. We design the SwinDETR architecture by first removing the classification head from the Swin Transformer backbone and inserting a fully connected layer. This layer projects the hierarchical feature maps extracted at each Swin stage into the embedding dimension required by the DETR encoder, ensuring proper alignment of feature spaces.

Distillation through attention on DETR-decoder. Drawing on the distillation-token concept from DeiT, we introduce a set of trainable distillation queries alongside the standard object queries as inputs to the DETR decoder. These distillation queries are initialized as learnable embedding vectors, identical in form to the object queries, and interact with all other embeddings via self-attention. At the final decoder layer, the distillation queries yield additional outputs that are trained to match the teacher model's predictions, thereby transferring knowledge through the network (see Figure 1).

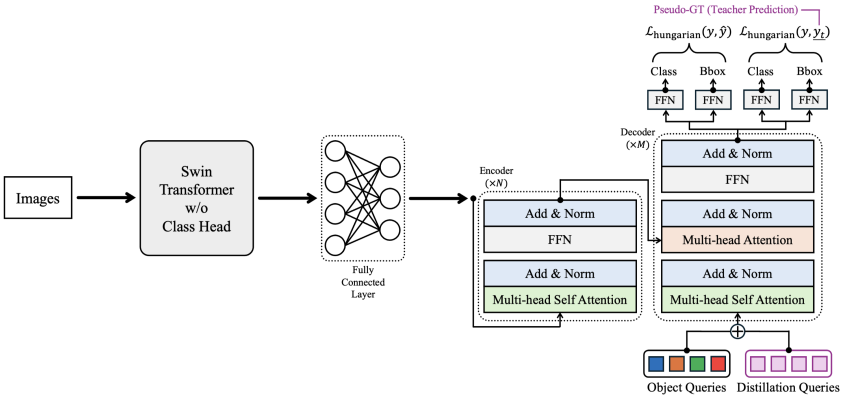

Figure 1: SwinDETR with distillation queries structure.

We also introduce a hard-label distillation technique that treats the teacher model's predictions as ground truth during the distillation process. This method affords a conceptually simple implementation for bounding box distillation. During training, the teacher's predicted object classes and bounding boxes for each input image are filtered by confidence score, and only those exceeding a threshold (0.5) are adopted as pseudo ground truth y_t. These pseudo labels are then matched one-to-one with the decoder outputs corresponding to the student model's distillation queries and optimized using the Hungarian loss employed in DETR. The loss function is defined as follows:

$$\mathcal{L}_{\text{distill}} = \frac{1}{2}\,\mathcal{L}_{\text{Hungarian}}(y, \hat{y}) + \frac{1}{2}\,\mathcal{L}_{\text{Hungarian}}(y, y_t) \tag{1}$$

$$\mathcal{L}_{\text{Hungarian}}(y, \hat{y}) = \sum_{i=1}^{N} \Big[\lambda_{\text{cls}} \big(-\log \hat{p}_{\hat{\sigma}(i)}(c_i) \big) + \mathbb{1}_{\{c_i \neq \varnothing\}} \Big(\lambda_{\text{giou}}\, \mathcal{L}_{\text{giou}} \big(b_{\sigma(i)}, \hat{b}_i \big) + \lambda_{\text{L1}} \big\| b_{\sigma(i)} - \hat{b}_i \big\|_1 \Big) \Big] \tag{2}$$

Post-training quantization on SwinDETR. We evaluate the effectiveness of LLM-based post-training quantization (PTQ) on SwinDETR—transformer-based vision model—by applying three PTQ schemes—weight-only quantization, dynamic quantization, and SmoothQuant—to a pre-trained SwinDETR model. Weight-only quantization converts only the model weights to low-bit representations, significantly reducing parameter memory while retaining FP32 activations to minimize accuracy degradation. Dynamic quantization automatically casts mainly nn.Linear operations to INT8 at runtime without any calibration, thus improving CPU inference speed. SmoothQuant employs

channel-wise scaling to flatten activation distributions and reduce quantization errors under low-bit arithmetic; for implementation simplicity, we apply SmoothQuant only to the weights.

3 Experiment

Experiment settings. We used the COCO 2017 dataset resized to 224×224 and trained both the proposed SwinDETR and the baseline DETR for 200 epochs with a batch size of 24 under identical hardware constraints. The Swin Transformer employed the Swin-B backbone, and Faster R-CNN (ResNet-50 + FPN) served as the teacher model for distillation. Distillation queries were set to 100—matching the number of object queries—while all other hyperparameters followed those of the original DETR. Post-training quantization was performed by converting model weights from 32-bit floating point to 8-bit integer format.

Comparing with DETR and exploring distillation. We compared the performance of SwinDETR and its distillation-augmented variant against the original DETR. SwinDETR achieved superior overall results, particularly improving AP_S, whereas the distilled SwinDETR saw declines in all metrics except AP_S. To verify whether distillation queries learn distinct representations from object queries, we measured their layer-wise cosine similarity after 100 and 200 training epochs. In both cases, the average similarity remained around 0.99, with only slight decreases across layers as training progressed. This high similarity indicates that distillation queries and object queries acquire nearly identical embeddings, which likely constrains the effectiveness of the distillation scheme (see Table 1).

Table 1: Detection performance of DETR, SwinDETR, and distilled SwinDETR.

Model	mAP (%)	AP_{50} (%)	AP_{75} (%)	AP_S (%)	AP_M (%)	AP_L (%)
DETR	3.7	9.3	1.1	0.1	1.3	8.8
SwinDETR (Ours)	**4.9**	**11.9**	**3.1**	**0.2**	**1.7**	**10.7**
SwinDETR+Distillation (Ours)	2.1	5.7	1.2	**0.2**	0.8	4.0

Exploring the effect of quantization. Among the evaluated PTQ methods for SwinDETR, weight-only quantization achieved the greatest compression ratio, while dynamic quantization delivered the highest CPU inference speed. Accuracy measurements revealed slight decreases in both mAP and AP_L for weight-only and dynamic quantization, whereas SmoothQuant improved both metrics. We attribute this to the regularization effect of quantization, which enhanced generalization performance more than any precision loss (see Table 2).

Table 2: Effect of post-training quantization methods on SwinDETR accuracy, inference speed, and model weight size.

PTQ method	mAP (%)	AP_S (%)	AP_M (%)	AP_L (%)	CPU Eval time	weight size (MB)
None	4.9	**0.2**	**1.7**	10.7	6m 43s	486.55
Weight-only	4.8	**0.2**	**1.7**	10.1	6m 39s	**122.10**
Dynamic	4.7	**0.2**	1.5	9.9	**5m 6s**	137.52
SmoothQuant	**5.1**	0.1	**1.7**	**11.0**	6m 14s	122.74

4 Conclusion

We introduced SwinDETR to enhance DETR's feature representation across diverse object scales by integrating a Swin Transformer backbone. Our distillation approach, however, showed limited gains due to class and distillation queries learning similar embeddings. Future work will explore constraints—such as enforcing orthogonality between query sets—or adopt soft distillation losses that minimize the KL divergence between teacher and student bounding-box predictions. Also, the three PTQ techniques applied to SwinDETR exhibit complementary strengths. Going forward, developing a unified quantization strategy that combines weight-only, dynamic, and SmoothQuant methods may yield further improvements in model efficiency and accuracy.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *The European Conference on Computer Vision (ECCV)*, 2020.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[5] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2017.

[6] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *International Conference on Learning Representations (ICLR)*, 2021.

[7] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *International Conference on Machine Learning (ICML)*, 2021.

[8] Yao Chen, Xinshi Yang, Hang Liu, Peng Yin, Zichuan Liu, Aoyang Zhou, Boxin Gao, Zhoujun Lin, and Xiaodong Liu. Dearkd: Data-efficient early exiting for knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[10] Zhewei Xiao, Zijian Wei, Yue Wu, Ji Lin, Yonggan Zhou, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. *International Conference on Machine Learning (ICML)*, 2023.

[11] Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. Fq-vit: Post-training quantization for fully quantized vision transformer. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

[12] Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.