Small Object Detection on Resource-Constrained Edge Devices: A Literature Review

Knowledge Distillation for Object Detection on Edge Devices

Knowledge distillation (KD) has become a popular model compression technique to deploy high-accuracy detectors on edge hardware with limited compute. In KD, a large **teacher** model's knowledge is transferred to a smaller **student** model, boosting the student's accuracy without increasing complexity [1]. Recent studies show KD can significantly reduce resource requirements while maintaining accuracy on edge devices [2]. For example, Setyanto *et al.* (2025) compared KD methods for object detection on edge and found that distilling a YOLOv5 teacher into a tiny YOLOv5 student yielded a compact model with minimal accuracy drop [3] [4].

Several works focus specifically on **small object detection** with KD. Zhu *et al.* (CVPR 2023) propose **ScaleKD**, a scale-aware distillation framework for small objects [5]. They introduce a *scale-decoupled feature distillation* module to let the student explicitly mimic the teacher's features at small-object scales, and a *cross-scale assistant* to handle noisy teacher predictions [6] [7]. ScaleKD boosted small object detection performance on COCO and VisDrone datasets while keeping inference efficient [8] [9]. In another recent work, Zhou *et al.* (2024) present **KDSMALL**, a lightweight small-object detector distilled from a larger model [10]. KDSMALL uses a multi-scale feature extractor and attention mechanisms, combined with KD, to improve detection of tiny objects [11]. The authors report significant accuracy gains over vanilla YOLO on small objects, and real-time performance when deployed on edge GPUs (e.g. NVIDIA Jetson TX2) [12] [13]. Similarly, an approach called **BiKD-YOLO** (2024) integrates *BiFormer* attention and knowledge distillation to enhance YOLO for UAV-based small object detection [14] [15]. These studies demonstrate that KD is a powerful tool for compressing detectors for edge deployment, and specialized distillation techniques (scale-aware feature transfer, attention-based distillation, etc.) can substantially improve small object accuracy.

Synthetic Data Augmentation for Small Object Detection

A key challenge for small object detectors is data scarcity – small instances are relatively rare and hard to capture in datasets [16]. To address this, researchers have explored **synthetic data augmentation**, including using generative adversarial networks (GANs) to create or enhance training examples. *Perceptual GAN* (Li *et al.*, 2017) was a pioneering work that used a GAN to super-resolve small objects' feature representations, effectively "hallucinating" large-object detail for tiny instances to aid detection [17] [18]. By generating super-resolved representations for small traffic signs and pedestrians, their detector achieved state-of-the-art results on those small objects [17] [19].

More recently, Bosquet *et al.* (2023) proposed a **full augmentation pipeline** combining GAN-based generation with copy-paste placement and blending. Their pipeline uses a *Downsampling GAN (DS-GAN)* to synthesize realistic small objects from larger object crops [20] [21]. The synthetic objects are then segmented, inpainted into new backgrounds, and blended for realism [22] [21]. This approach improved

detection performance by up to **11.9% AP** (small objects, UAVDT dataset) compared to no augmentation [23]. Notably, they found traditional resizing of objects creates unrealistic artifacts, whereas DS-GAN produces more authentic low-resolution appearances [24] [25]. Other studies have similarly used GANs: Stachoń and Pietroń (2022) augment a detector's training set on Pascal VOC with GAN-generated samples, reporting that GAN-based augmentation outperformed basic geometric augmentations for small objects [26] [27]. These works show that GANs can increase the **diversity and quantity** of small-object training data, mitigating the lack of samples and improving detector accuracy.

Aside from GANs, simpler synthetic augmentation techniques have also proven effective. **Copy-paste augmentation** is a straightforward method where small objects are cut from images and pasted multiple times into new locations. Kisantal *et al.* (2019) demonstrated that oversampling images containing small instances and copy-pasting the small objects to increase their count led to a **7–10% relative improvement** in small-object AP on COCO [28]. Unlike GANs, this does not create new object appearances, but it amplifies underrepresented examples and can be easier to implement. In practice, a combination of techniques may be used: for example, one might generate synthetic small objects with a GAN and also apply copy-paste or context simulation to cover a range of scenarios [22] [21]. Overall, literature suggests that **augmenting small object data**, whether via GANs or other synthetic means, is crucial to boost detection performance – often yielding substantial gains with relatively low cost.

Anchor-Free Detection and Small Objects

Modern object detectors have trended toward **anchor-free** designs, which can be beneficial for small objects. Traditional anchor-based detectors (e.g. Faster R-CNN, YOLOv3) rely on predefined anchor box sizes. Small objects often fail to match these anchor boxes well – for instance, an analysis of Mask R-CNN on COCO found that predicted anchors rarely overlap small ground-truth objects sufficiently [16]. This mismatch leads to few positive samples during training, hurting small-object recall. Anchor-free detectors avoid this issue by predicting object locations directly (e.g. via keypoints or center locations) instead of classifying numerous anchor boxes. Approaches like CenterNet and FCOS, which predict objects as point centers with associated size, inherently handle arbitrary object scales without needing carefully tuned anchors. This can improve small object detection since even tiny objects can activate a detection point on high-resolution feature maps, rather than relying on an anchor of appropriate size.

Many small-object-specific models adopt multi-scale feature fusion (e.g. FPN) and anchor-free outputs. For example, Zhu *et al.* (2020) introduced an anchor-free detector for tiny faces, arguing that removing anchors simplifies the search space for small instances and reduces false negatives [16]. In the small object literature, even anchor-based frameworks often optimize anchor configurations or incorporate anchor-free layers for the smallest scales. In summary, while anchor-free vs. anchor-based is an ongoing debate, it's widely accepted that **anchor design must be handled carefully for small targets**, and anchor-free methods offer a more flexible solution by eliminating the rigid grid of anchor boxes that small objects can slip through.

Adaptive Model Compression Under Real-Time Constraints

Another frontier is **adaptive model compression** – dynamically adjusting a model's complexity at runtime based on current resource constraints (like available power, thermal headroom, or frame-rate requirements). Traditional compression (pruning, quantization, KD) is done offline, yielding a fixed model.

However, edge devices experience varying conditions (battery levels, CPU/GPU load, thermal throttling) that may demand the model to speed up or slow down. Recent research addresses this via *runtime-switchable networks*. For example, **Once-For-All (OFA)** networks are trained to support multiple sub-models (e.g. different widths or depths) that can be selected on the fly [29]. Cai *et al.* (2020) showed that a single OFA model can be deployed with different resource footprints (e.g. a smaller sub-network when power is limited) without retraining [29]. Similarly, **slimmable networks** allow the model to "thin" or "widen" dynamically, trading accuracy for efficiency as needed.

Beyond multi-configuration models, some systems explicitly monitor and adapt to device status. **DNNShifter** (Eccles *et al.*, 2024) is a runtime DNN management system that pre-trains a *portfolio* of pruned model variants and **switches** between them based on operating conditions [30] [31]. It addresses the need to maintain performance under fluctuating workloads or thermal limits: if the device heats up or the workload spikes, a lighter model variant is loaded to keep inference real-time and avoid throttling [30] [32]. The model-switching overhead is kept low (tens of milliseconds) so that adaptation is seamless [33]. While these solutions are not tailored to knowledge distillation per se, they often *use* KD or related compression to obtain the multiple efficient models. The concept of **adaptive distillation** – e.g. adjusting distillation intensity or teacher guidance during runtime – is still emerging, but one could imagine a system that gradually distills a model further when thermal constraints tighten.

In summary, adaptive approaches ensure that edge vision systems meet real-time constraints **at all times**. They complement static compression by adding a control layer: the model can dial up or down its complexity to balance accuracy and efficiency. This is especially relevant for battery-powered or thermally constrained platforms like drones and mobile GPUs. A framework that unifies KD-based compression with runtime adaptability would be quite novel, as most prior KD works train a fixed student, and most adaptive systems focus on switching between pre-compressed models rather than on-the-fly knowledge transfer.

Edge-Optimized Small Object Detectors and Deployment

Finally, numerous studies have engineered **edge-friendly detection models** aimed at devices like the NVIDIA Jetson series (TX2, Xavier NX, etc.). A common theme is using lightweight architectures (MobileNet, Tiny YOLO, etc.) augmented with tricks to preserve small object accuracy. For instance, Yu *et al.* (2023) developed **SFHG-YOLO**, based on YOLOv5s, to detect tiny pineapple buds on UAVs [34]. They introduced specialized modules to enhance fine-grained features and reported real-time performance on a Jetson Xavier NX, even with the challenging small targets [35]. Other works like **EL-YOLO** (2024) and **RTS-Net** (2024) incorporate attention mechanisms (e.g. BiFormer or transformer layers) to boost small object features, while keeping the model size small enough for NX deployment [36] [37]. These optimized models, when paired with TensorRT acceleration or mixed-precision (FP16) on Jetson, can achieve inference speeds suitable for live applications (often 20–30+ FPS for VGA or 720p input) [38].

It's noteworthy that many small-object edge detectors target **drone and mobile vision** scenarios. Drones equipped with Jetson modules require detectors that are both **fast** and **sensitive** to small items (e.g. distant vehicles, humans, or signs). By pruning networks, using compact backbones, and focusing on multi-scale feature fusion, researchers have shown it's possible to get decent accuracy on tiny objects within the tight compute budgets. For example, Hua *et al.* (2024) in BiKD-YOLO report improved precision on aerial small objects compared to standard YOLOv5, thanks to KD from a larger teacher and an efficient attention module [14] [39]. Such results underscore that edge-optimized models can tackle small object detection, but

often **each component must be carefully designed** – from data augmentation, to architecture, to compression – to compensate for the limited computation.

Novelty of the Proposed *SynKD* Framework

Given this landscape, we can contextualize the proposed **SynKD** approach. SynKD appears to combine multiple state-of-the-art strategies: knowledge distillation for model compression, synthetic data (possibly GAN-generated) for augmenting small objects, and an anchor-free detection architecture – all targeted for edge deployment (e.g. Jetson Xavier NX). Our literature review finds **individual precedents** for each of these components, but very few (if any) frameworks that integrate all in a unified solution. For example, prior works distilled large models into small ones for better small-object accuracy [8] [11], and others generated synthetic small examples to improve training [20] [28], but a system that employs *both* KD and extensive synthetic augmentation is novel. Moreover, SynKD's anchor-free paradigm aligns with modern best practices, as it avoids anchor tuning issues for tiny objects – an aspect supported by recent detector designs but not unique on its own. The inclusion of an **adaptive distillation** element (adjusting model complexity at runtime based on thermal/power constraints) would be an especially novel contribution. While dynamic model scaling is known in the systems community [30] [33], applying it in conjunction with on-device knowledge distillation (e.g. periodically refining the student model or switching distilled models on-the-fly) goes beyond current literature.

In summary, similar ideas to SynKD exist in pieces throughout the last 7–8 years, but **the particular combination** proposed is new. No prior study was found that implements a full small-object detection pipeline with knowledge distillation, GAN-based data augmentation, anchor-free detection, and adaptive model management for edge devices all together. Thus, SynKD stands to be a **novel framework**, leveraging proven techniques from past work in a synergistic way. By uniting these components, SynKD can push the boundaries of what's achievable for small object detection on resource-constrained edge devices, marking a clear step beyond existing implementations in both scope and capability.

References: The review above synthesizes findings from diverse sources, including peer-reviewed conference papers and journals, with citations indicating key supporting works (e.g. CVPR, IEEE Access, Pattern Recognition, etc.). These illustrate the evolution of small object detection techniques and how the proposed approach builds upon and innovates beyond the state of the art [5] [20] [30].

---

[1] [2] **Overview of the KeepEdge edge intelligence framework. | Download Scientific Diagram**
https://www.researchgate.net/figure/Overview-of-the-KeepEdge-edge-intelligence-framework_fig2_359126497

[3] **Knowledge Distillation for Object Detection Based on Mutual ...**
https://www.researchgate.net/publication/354411749_Knowledge_Distillation_for_Object_Detection_Based_on_Mutual_Information

[4] **In Kee Kim - DBLP**
https://dblp.org/pid/96/69

[5] [6] [7] [8] [9] **CVPR 2023 Open Access Repository**
https://openaccess.thecvf.com/content/CVPR2023/html/Zhu_ScaleKD_Distilling_Scale-Aware_Knowledge_in_Small_Object_Detector_CVPR_2023_paper.html

10 11 13 **Complex Scene Understanding and Object Detection Algorithm Assisted by Artificial Intelligence | Academic Journal of Science and Technology**
https://drpress.org/ojs/index.php/ajst/article/view/25804

12 **Real-time small object detection on embedded hardware for 360 ...**
https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13207/132070E/Real-time-small-object-detection-on-embedded-hardware-for-360/10.1117/12.3031503.full

14 **BiKD-Yolo: A Small Object Detection Algorithm from UAV ...**
https://ieeexplore.ieee.org/document/10702331/

15 39 **BiKD-Yolo: A Small Object Detection Algorithm from UAV ...**
https://www.semanticscholar.org/paper/BiKD-Yolo%3A-A-Small-Object-Detection-Algorithm-from-Hua-Cai/b67e967b15e3488ca34fef3082e4a6972e84b090

16 28 **[1902.07296] Augmentation for small object detection**
https://arxiv.org/abs/1902.07296

17 18 19 **Perceptual Generative Adversarial Networks for Small Object Detection**
https://openaccess.thecvf.com/content_cvpr_2017/papers/Li_Perceptual_Generative_Adversarial_CVPR_2017_paper.pdf

20 21 22 23 24 25 **A full data augmentation pipeline for small object detection based on generative adversarial networks**
https://www.micc.unifi.it/seidenari/publication/pr-22/pr-22.pdf

26 27 **[2208.13591] Chosen methods of improving small object recognition with weak recognizable features**
https://arxiv.org/abs/2208.13591

29 **DNNShifter: An efficient DNN pruning system for edge computing**
https://ouci.dntb.gov.ua/en/works/4YwVQ3q9/

30 31 32 33 **Dnnshifter: An Efficient Dnn Pruning System for Edge Computing by Bailey Jack Eccles, Philip Rodgers, Peter Kilpatrick, Ivor Spence, Blesson Varghese :: SSRN**
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4482180

34 35 **SFHG-YOLO: A Simple Real-Time Small-Object-Detection Method ...**
https://www.mdpi.com/1424-8220/23/22/9242

36 **A multi-scale small object detection algorithm SMA-YOLO for UAV ...**
https://www.nature.com/articles/s41598-025-92344-7

37 **BiKD-Yolo: A Small Object Detection Algorithm from UAV ...**
https://www.researchgate.net/publication/384757127_BiKD-Yolo_A_Small_Object_Detection_Algorithm_from_UAV_Perspective_Based_on_BiFormer_Attention_and_Knowledge_Distillation

38 **EL-YOLO: An efficient and lightweight low-altitude aerial objects ...**
https://www.sciencedirect.com/science/article/abs/pii/S0957417424017159