INTISARI

Estimasi kerumunan secara *real-time* pada perangkat *edge* dengan sumber daya terbatas merupakan tantangan krusial dalam aplikasi keamanan dan manajemen publik. Model deteksi berakurasi tinggi yang ada saat ini seringkali memerlukan komputasi mahal sehingga tidak praktis untuk implementasi di lapangan. Untuk mengatasi kendala ini, penelitian ini menyajikan optimisasi menyeluruh pada kerangka kerja P2PNet untuk deteksi kepala yang efisien. Melalui strategi optimisasi dengan substitusi *backbone* VGG dengan arsitektur ringan (MobileNetV2, ShuffleNetV2, ResNet18) dan penerapan kuantisasi statis pasca-pelatihan (INT8) berhasil dicapai peningkatan kecepatan inferensi lebih dari 50 kali lipat serta pengurangan ukuran model hingga 15 kali lipat. Sebagai inovasi utama, penelitian ini memperkenalkan *Hungarian Matching Scheduler*, sebuah strategi pelatihan korektif untuk mengatasi penurunan akurasi lokalisasi yang merupakan efek samping umum dari proses kompresi. *Hungarian Matching Scheduler* secara dinamis memperketat batasan jarak pencocokan selama pelatihan, memaksa model untuk mempelajari representasi spasial yang lebih presisi. Hasil evaluasi pada *dataset benchmark* ShanghaiTech dan UCF-QNRF menunjukkan bahwa model yang dioptimalkan, khususnya yang dilatih menggunakan *Hungarian Matching Scheduler*, mampu memberikan keseimbangan (*trade-off*) terbaik antara akurasi dan efisiensi komputasi. Model yang dihasilkan mampu mencapai kecepatan inferensi melebihi 15 FPS pada CPU peningkatan drastis dari *baseline* di bawah 1 FPS sambil mempertahankan akurasi deteksi. Pencapaian ini membuktikan kelayakan implementasi analisis kerumunan yang akurat secara *real-time* pada perangkat *edge* untuk aplikasi praktis seperti sistem pemantauan cerdas dan manajemen ruang publik.

Kata kunci : *Crowd Counting*, Deteksi Objek, Kompresi, *Lightweight*, Kuantisasi

ABSTRACT

Real-time crowd estimation on resource-constrained edge devices is a critical task for public safety and management, yet existing high-accuracy models are often too computationally expensive for practical deployment. This paper addresses this challenge by optimizing the P2PNet framework for efficient head detection. This research propose a dual-pronged optimization strategy that combines architectural modification by replacing the heavy VGG backbone with lightweight alternatives like MobileNetV2 and ShuffleNetV2 and post-training static quantization to INT8 precision. This approach achieves a more than 50-fold increase in inference speed and a 15-fold reduction in model size. To counteract the localization errors introduced by these compression techniques, a training strategy, the Hungarian Matching Scheduler, is introduced. It dynamically adjusts matching constraints to stabilize and refine model learning. Extensive experiments on standard benchmarks (ShanghaiTech, UCF-QNRF) demonstrate that the optimized models, particularly when trained with Hungarian Matching Scheduler, offer a superior trade-off between speed, size, and accuracy. For instance, the developed lightweight models achieve inference speeds exceeding 15 FPS on a CPU, a significant improvement over the baseline's sub-1 FPS, while maintaining competitive counting and localization performance, thus enabling practical real-time crowd analysis on edge devices for practical applications such as intelligent monitoring systems and public space management.

Keywords : Crowd Counting, Compression, Lightweight, Object Detection, Quantization