Real-Time Localized Image Enhancement via YOLOv8 and Lightweight Super Resolution Model

Huihao Xing, Youran Geng, Yuhao Zhang

April 6, 2024

Abstract

This project proposes a real-time framework for enhancing image clarity in localized regions by integrating YOLOv8 for object detection and a lightweight super-resolution (SR) model. Targeting applications such as video conferencing and live streaming, the system dynamically identifies regions of interest (e.g., human subjects, animals, plants, cars, etc.) and applies adaptive SR enhancement while maintaining a processing speed of 30+ FPS on consumer-grade GPUs, and other potential devices, such as smartphones or iPad. Our approach balances computational efficiency and perceptual quality through model optimization and/or hardware acceleration.

Introduction

Modern visual communication systems demand real-time image enhancement to improve user experience in low-bandwidth or low-resolution scenarios. Existing solutions often either process entire frames (wasting computation on irrelevant regions) or rely on heavy SR models incompatible with real-time constraints. We aim to compare the resource consumption between enhancing only the segmented areas versus processing the entire image. We proposed a solution based on the following steps:

1. **Segmentation**: YOLOv8 detects and segments subjects from entire frame.
2. **Enhancement**: A pruned ESRGAN model(tentative) super-resolves target regions (2x - 4x upscaling) while preserving background details.
3. **Optimization(tentative)**: TensorRT quantization and CUDA-accelerated fusion ensure end-to-end latency < 700ms.

This approach ideally enables resource-efficient enhancement focused on the subjects, achieving both speed and quality in general.

Related Work

Object Detection YOLO variants [1] dominate real-time detection, with YOLOv8 offering improved segmentation masks over previous versions. Mask R-CNN [2] provides higher accuracy but is 5x slower.

Super Resolution ESRGAN [3] achieves photorealistic results but requires 200ms per 512x512 patch. Lightweight models like FSRCNN [4] sacrifice quality for speed (10ms per patch).

Real-Time Systems NVIDIA DeepStream [5] demonstrates GPU-accelerated pipelines, while MobileSR [6] optimizes SR for edge devices. None combine adaptive detection and enhancement.

Proposed Work

Our architecture has three stages(two scheduled stages and one tentative stage):

1. Adaptive Region Selection YOLOv8 processes 1920*1080 inputs at 30 FPS, outputting segmentation section.

2. Multi-Scale SR Enhancement A modified ESRGAN model processes segmented regions. The model selects constant upscale factors of $2x$, which suggests that the enhanced area will be 4 times of original pixels numbers.

3. Hybrid Deployment(tentative) Using TensorRT, we quantize YOLOv8 to INT8 (2x speedup) and the SR model to FP16. CUDA kernels accelerate mask fusion to avoid CPU-GPU transfers.

Datasets

This project will be used a variety of datasets. Some interesting datasets from Kaggle:

1. Kaggle - Human Segmentation Dataset - Supervise.ly
2. Kaggle - Human Segmentation MADS
3. Kaggle - Human Segmentation - TikTok Dances

The datasets above capture a wide range of real-life scenarios, including video clips (especially of TikTok dances), the MADS dataset featuring martial arts, dances and sports

movements, as well as static images such as individual portraits, family photos, and scenic shots with people in various poses. Specifically, the human subjects may be facing the camera directly or performing dynamic actions, providing rich variations in pose, background, and interaction. Therefore, we believe the aggregated dataset is representative of what most TikTok contents would be.

Exploratory Data Analysis of the Aggregated Dataset

Each of the first three human segmentation datasets is relatively clean. Each of them contain a `df.csv` specifying the corresponding collage, images, and masks, and the corresponding images and masks share the same file name (under different directories). We do not need collages as they are just combinations of images and masks.

The (black-and-white) mask photos depict the areas with human body. In our project, the YOLO model will first convert the masks to boundaries (as a collection of pixel points) and learn the associations between the photo and the boundaries. For prediction step, we will first predict a boundary from YOLO, then convert the boundary to a mask. Finally, we put the mask on the original photo to filter out the human segmentation we want for super-resolution.

Evaluation

Quantitative Metrics:

- **Segmentation Accuracy**
  - Mean Average Precision (mAP): This metric evaluates the overall detection performance in various classes by computing the average precision and recall. A higher mAP indicates that the detected bounding boxes closely match the ground truth across different scenes.
  - Intersection of Union (IoU): IoU measures the overlap between the predicted bounding boxes and the ground truth boxes. A higher IoU reflects more precise localization, which is critical for accurately defining the regions that the super-resolution module will enhance.
  - Precision and Recall: These metrics quantify the detector's ability to correctly identify objects (precision) and capture all relevant objects (recall). Balancing precision and recall is essential to minimize false positives and false negatives, ensuring that the enhancement process focuses on the correct regions.
- **Speed (Real-Time Performance)**
  - FPS ($\geq$30 FPS on 1080p input)

- End-to-end latency (<700ms, including detection + enhancement + fusion)

Qualitative Evaluation:

- User study (5+ participants rating enhanced videos on a 1-5 scale)
- Visual comparison (generate GIFs of before/after enhancement)

Timeline

Total Duration: March 3, 2024 - May 1, 2024 (9 weeks)

Week	Milestone	Deliverables
Week 1-2	Data Preparation & Baseline Testing	• Clean and annotate 3 datasets • Implement YOLOv8-INT8 quantization benchmark
Week 3-4	Model Development	• Train lightweight ESRGAN • Implement CUDA-accelerated mask fusion module
Week 5	System Integration	• End-to-end pipeline prototype (FPS $\geq$ 20) • Validate CUDA optimization results
Week 6	Performance Optimization	• Achieve 30+ FPS with TensorRT deployment • Analyze quantization-sensitive layers
Week 7	Evaluation Phase	• Complete quantitative metrics comparison table • Collect subjective evaluation survey data
Week 8	Final Debugging	• System robustness testing (low-light/occlusion scenarios) • Write technical documentation
Week 9	Project Delivery	• Submit full code/model weights • Generate visual comparison report

Conclusion

In this project, we propose a real-time framework for localized image enhancement by integrating YOLOv8 for object detection and a lightweight super-resolution (SR) model. Our approach addresses the limitations of existing solutions by dynamically identifying

regions of interest (e.g., human subjects) and applying adaptive SR enhancement while maintaining a processing speed of 30+ FPS on consumer-grade GPUs. By leveraging model optimization techniques such as TensorRT quantization and CUDA-accelerated fusion, we achieve a balance between computational efficiency and perceptual quality.

Key contributions of our work include:

- A hybrid pipeline combining YOLOv8 for segmentation and a pruned ESRGAN model for super-resolution, enabling resource-efficient enhancement focused on human subjects.
- A deployment strategy using TensorRT and CUDA to ensure end-to-end latency of less than 700ms, making the system suitable for real-time applications such as video conferencing and live streaming.
- Comprehensive evaluation metrics, including FPS, End-to-end latency, and user study to validate the system's performance and image quality.

We anticipate that this framework will significantly improve user experience in low-bandwidth or low-resolution scenarios, offering a practical solution for real-time image enhancement. Future work may explore extending the system to handle more complex scenes, integrating additional optimization techniques, or deploying the framework on edge devices for broader applicability.

By achieving the proposed milestones within the 10-week timeline, we aim to deliver a robust, efficient, and high-quality solution for real-time localized image enhancement.

Code Availability

The implementation of our real-time enhancement pipeline, including YOLOv8 quantization, lightweight ESRGAN training, and CUDA-accelerated fusion modules, is publicly available under the MIT License:

- **GitHub Repository**: Link to Repo

The repository contains:

- Pre-trained models (YOLOv8-seg) and relevant files for segmentation purpose
- Benchmarking scripts for an initial run
- Docker support for reproducible deployment

References

[1] Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. *You Only Look Once: Unified, Real-Time Object Detection*. IEEE, 2016.

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick. *Mask R-CNN*. IEEE, 2017.

[3] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, Xiaoou Tang. *ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks*. Computer Vision Foundation, 2018.

[4] Chao Dong, Chen Change Loy, Xiaoou Tang. *Accelerating the Super-Resolution Convolutional Neural Network*. ECCV: European Conference on Computer Vision, 2016.

[5] NVIDIA Corporation. *NVIDIA DeepStream SDK*. NVIDIA Developer, N/A.

[6] Xindong Zhang, Hui Zeng, and Lei Zhang. *Edge-oriented Convolution Block for Real-time Super Resolution on Mobile Devices*. ACM Multimedia, 2021.