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

Huihao Xing, Youran Geng, Yuhao Zhang

Feburary 26, 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 subject) and applies adaptive SR enhancement while maintaining a processing speed of 30+ FPS on consumer-grade GPUs. 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 address these limitations by proposing a hybrid pipeline:

1. **Localization**: YOLOv8 detects and segments human subjects.
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 enables resource-efficient enhancement focused on human subject, 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 schedule 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
4. Kaggle - TikTok Dataset with depth prediction, model included

5. Kaggle - Human Dataset with YOLO labels, this may contain more various images.

ChatGPT proposed more recommendations from Wikipedia, which may be useful. We Will refer to them later if we need some extra datasets.

Evaluation

Quantitative Metrics:

- **Speed (Real-Time Performance)**
  - FPS ($\geq$30 FPS on 1080p input)
  - End-to-end latency ($<$700ms, including detection + enhancement + fusion)
- **Comparative Experiments**
  - Baseline 1: Full-frame enhancement using ESRGAN
  - Baseline 2: YOLOv8 + FSRCNN
  - SOTA methods: Compare with latest CVPR/ICCV real-time SR papers

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-INT
Week 3-4	Model Development	Train lightweight ESRGAN Implement CUDA-accelerat
Week 5	System Integration	End-to-end pipeline prototype (FPS $\geq$ 20) Validate CU
Week 6	Performance Optimization	Achieve 30+ FPS with TensorRT deployment Analyze q
Week 7	Evaluation Phase	Complete quantitative metrics comparison table Collect
Week 8	Final Debugging	System robustness testing (low-light/occlusion scenarios
Week 9	Project Delivery	Submit full code/model weights Generate visual compar

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 localization 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.

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.