A Hybrid Approach of EDVR and YOLOv8 for Super-Resolution

Abul Monsur Mannan

Abstract

Advances in super-resolution (SR) and object detection have often been used largely in parallel within the domain of computer vision. While 'Enhanced Deformable Video Restoration' (EDVR) excels in generating high-resolution images, models like YOLOv8 are optimized for detecting and segmenting objects within those images. This research seeks to synergize these specialized capabilities into a unified framework while adopting a 'qualitative methodology' approach. The study fuses the aligned feature maps produced by EDVR with YOLOv8-seg's contextual layers on a standard Set5 dataset utilizing a PyTorch-based implementation. The hypothesis posits that the fusion of features from both models will result in high-quality super-resolution outputs, augmented with precise object detection and segmentation. The proposed model has applications across various domains, including surveillance, medical imaging, and autonomous vehicles.

Introduction

In the context of computer vision, Super-Resolution (SR) is a foundational technique for enhancing image and video frame quality. Notable progress has been made by models like Enhanced Deformable Video Restoration (EDVR) in achieving Super Resolution. However, the integration of object detection and segmentation into SR remains an active and exploratory field. Models such as YOLOv8-seg [4] have shown strong segmentation abilities, particularly for capturing contextual information in images.

A research gap exists where EDVR excels in video super-resolution and YOLOv8-seg specializes in object detection and segmentation. There's a lack of methodologies that combine their strengths to comprehensively enhance super-resolution quality in images and video frames.

The research hypothesis suggests that fusing feature maps from YOLOv8-seg and EDVR will generate high-quality image and video frames with improved perceptual quality and accurate object detection masks. This hypothesis draws inspiration from works like Fast-SAM [26], SEEM, and Improved EDVR.

The study's significance lies in its potential contribution to computer vision, particularly in tasks requiring super-resolution alongside object detection and segmentation. This impact spans domains such as surveillance, medical imaging, and autonomous driving systems. The study bridges a gap in the current understanding, opening avenues for advanced applications in various sectors.

Literature Review

Researchers have introduced diverse learning models to tackle the super-resolution challenge, each with unique strengths and limitations. Interpolation, a widely used technique for resolution enhancement, is countered by methods like sparse coding and dictionary learning to map low-resolution images to high-resolution space. Patch-based methods like self-model and non-local means enhance textures and details but might introduce blocking artefacts. Deep learning models, particularly Convolutional Neural Networks (CNNs), such as SRCNN, VDSR, and EDSR, have excelled in capturing nonlinear LR-HR relationships, offering promising results in super-resolution. GAN-based models like SRGAN and video-specific models like EDVR show potential for high-quality outputs.

In the realm of video super-resolution (VSR), techniques using deep neural networks have advanced significantly. Spatial and temporal information play pivotal roles, with models like DUF and EDVR incorporating optical flow and temporal dependencies to enhance quality. Spatial information encompasses pixel values, gradients, and structures, while temporal information addresses changes over time in sequences. Techniques like patch-based methods, CNNs, sparse coding, and motion compensation extract spatial details, while optical flow estimation and temporal super-resolution networks capture temporal changes.

Comparing super-resolution methods reveals their strengths and weaknesses. Bicubic interpolation results in blurry images, while deep learning models like CNNs capture complex relationships but require substantial resources. Proper dataset selection and augmentation are essential for robust performance. Two models, EDVR and YOLOv8-seg, are chosen for research. EDVR excels in video super-resolution, leveraging temporal coherence and motion compensation, yet faces complexity challenges. YOLOv8-seg combines object detection and semantic segmentation for robust performance but may be impacted by detection and segmentation inaccuracies. Both models guide the research towards effective super-resolution solutions.

Problem Statement

The focus of this survey is twofold. First, a basic survey of SR models and the selection of base models for the second step of this study.

The second part is to develop image resolution models that can effectively improve the quality and analysis of images, in computer vision applications.

Research Objective

The primary aim of the study is to conduct a survey on learning models used for achieving super-resolution (SR) in images. The study is motivated to explore the advancements and techniques in learning models that can enhance image resolution and quality. The study will examine the techniques employed in these models, how the contributors evaluate their performance and identify their limitations. Through this analysis, the aim is to gain insights into the strengths, weaknesses, and areas for improvement of models.

In addition, the research aims to propose, implement, and evaluate novel image resolution model. The study will also explore techniques and existing models that can enhance image quality through SR. The focus will be on developing efficient models, with improved generalization and robustness capabilities. As well as, the study intends to undertake an investigation into the possibility of boosting video SR performance by using two models which are possibly YOLOv8-seg and EDVR.

This study shall, therefore, involve the extraction and integration of the different video frames with the aim of discovering the way these models can complement each other and can be used for video SR.

Research Methodology

The research methodology is composed of a comprehensive investigation of techniques to enhance image and video resolutions. The focus of the study is on both classical algorithms and deep learning-based methods.

a) Methodology breakdown

In this study, we are following a qualitative research approach. The first part goes with the investigation of the methods, SR models and techniques followed by the authors of YOLOv8 and EDVR to fulfil our research objective. Following this, we examine the architecture of SEEM model from (Lu et al.) and Improved-EDVR from Huang et al. to materialize our research objective. These papers and their mentioned architecture and models are peer-reviewed in several other papers and practically implemented in real-life applications.

The second part consists of an in-depth examination of the methods and techniques the authors followed in those models to achieve video/image SR.

In the last and third part, we provide a detailed explanation of the algorithms, frameworks, and tools we utilize for implementing and testing the our proposed models. The model creation, implementation and results will come in the following section of this paper.

b) About EDVR and YOLOv8

This study focuses on evaluating YOLOv8 and EDVR, contributing to the existing knowledge of image/video super-resolution. The EDVR architecture comprises components such as Pyramid, Cascading, and Deformable (PCD) alignment module for large motion handling, Temporal and Spatial Attention (TSA) fusion module for feature emphasis, Feature extraction network, Enhanced Deformable Convolutional Network (EDCN) for restoration, and modules for Reconstruction and PreDeblur. The latter mitigates blur in video sequences by learning complex mappings. YOLOv8-seg utilizes CSPDarknet53 backbone with Cross-Stage Partial (CSP) connections for feature extraction. Beginning with convolutional and pooling layers, CSPDarknet53 captures low-level to abstract features, enabling effective representation. Extracted features serve object detection and semantic segmentation tasks, achieving a balance between fine details and global context through well-designed architecture and connections.

c) Implementation background

To facilitate our work we have inspired by and chosen two models for achieving a better image/video SR for our study. These models are SAM-guided Ed refinement Module (SEEM) by (Lu et al., 2023) and Improved EDVR by (Huang & Chen, 2022). We follow their method of work and tools to work with EDVR and YOLOv80seg pretrain models.

- *SEEM:*

SEEM's plug-in module boosts the EDVR's performance and quality by using the semantic information from SAM to improve the alignment and fusion of multiple frames. According to the paper, SEEM can enhance both the foreground and background regions of the video frames, resulting in more accurate and realistic super-resolution outputs. For example, in section 4.2 of the paper, the authors show some visual comparisons between EDVR and EDVR+SEEM on the REDS dataset. They demonstrate that SEEM can better handle large motions, occlusions, and complex textures, such as the moving car, the occluded person, and the brick wall.

They also provide some quantitative results in their paper, where they report that SEEM can improve the PSNR and SSIM metrics of EDVR by 0.14dB and 0.003, respectively, on the REDS dataset.

Moreover, SEEM can also reduce the number of parameters and FLOPs of EDVR by 11.6% and 10.8%, respectively, without sacrificing performance. Therefore, SEEM's plug-in module can boost EDVR's performance and quality by utilizing a more robust and semantic-aware prior for VSR.

- *IMPROVED EDVR:*

The authors identify two main challenges in VSR: accuracy and the need for high-speed, possibly real-time, models.

To address these challenges, the authors propose several improvements – 1) a preprocessing module, 2) A temporal 3D, 3) a convolutional fusion module, and 4) a new reconstruction block. Additionally, the authors employ multiple programmatic methods to accelerate both the model training and inference processes, making the model more practical for real-world applications.

We assume the proposed model not only outperforms the baseline EDVR model in terms of PSNR and SSIM but also maintains a similar level of parameter count. This makes the model both robust and efficient, reducing timing cost and memory consumption while delivering extraordinary performance.

Research Methodology

Our module hypothesis:

To formulate a model, we intend to use the YOLOv8-seg model's feature maps and fuse those features with the EDVR's aligned feature. We use the following approach to achieve our goal:

First, we aim to collect all the feature maps from the image/video frame produced by the YOLOv8-seg model.

This will give us an idea of where the objects are located in each frame through feature masks. This gives us the internal information of the feature maps like shape and size.

Next, we can use EDVR to super-resolve each frame of the video. This will give us high-quality image frames with improved perceptual quality. Tweaking EDVR architecture will be required here.

Then, we can fuse the features extracted from YOLOv8-seg with the aligned features from EDVR to generate high-quality image or video frames with very good image segmentation.

We hypothesise that the above approach allows us to generate high-quality image/video frames with object detection masks. However, it is important to note that this approach may require significant computational resources due to the complexity of both models.

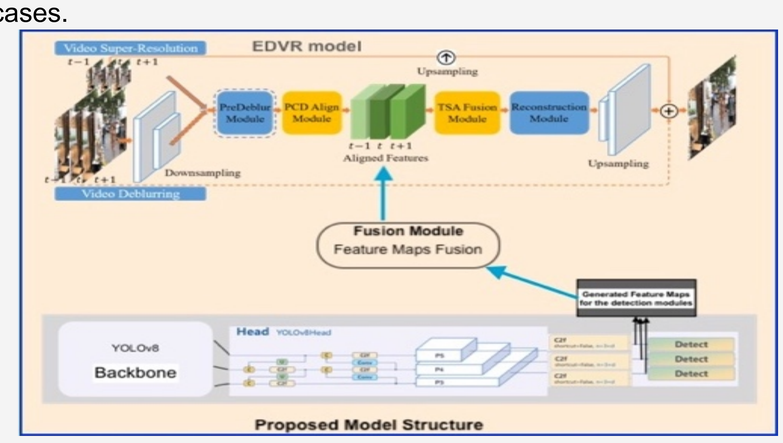
Experiment, Results and Discussion

- *Fusing two models' feature maps*

The proposed method involves gathering feature maps from the YOLOv8-seg model using a pretrained model and Pytorch. The EDVR model is then used to fuse the feature maps from the YOLOv8-seg model with the PCD module's outputs for feature alignment with neighboring features.

Two fusion modules are introduced: Attention Fusion Module and Stack Fusion Module. These modules perform attention-based fusion of feature maps, with attention multiplication being more suitable for tasks requiring spatial focus and concatenation for tasks requiring richer feature representations.

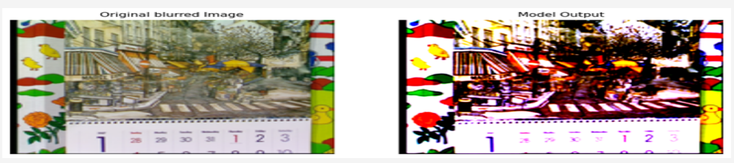
Both modules are initialized at runtime based on the shape of the input tensors, offering flexibility but may not be optimal for all use-cases.



The final model structure consists of two modules: Attention Fusion Module, which multiplies the input tensor with attention maps, and Stack Fusion Module, which concatenates the input tensor with attention maps and applies a 1x1 convolution.

- *Results and discussion*

The study compared the quality of a model for super resolution problem using two computer vision models.



Despite limited resources, the model produced good quality outputs, matching or even better than the ground truth and blurred LR images.

However, due to unavailability of required GPUs and time constraints, the model could not be fully trained and tested on a full dataset. The PSNR test showed lower results compared to the original model. The architecture uses two distinct computer vision models for super resolution, but combining them on large image frame data requires additional computational power and resources.

Conclusion

The study proposes a hybrid super-resolution model that combines YOLOv8-seg and EDVR for computer vision tasks. The model extracts feature maps from YOLOv8-seg and aligns them with EDVR features to produce high-quality image frames with enhanced object detection masks.

Despite high computational requirements, the model showed promise. The study was constrained by limited computational resources and time, but the model's potential to advance super-resolution methodologies is evident. Further research is needed to fine-tune the model and evaluate its performance under different conditions and metrics.

Acknowledgement

I would like to thank my supervisor Dr Yuhang Ye for providing invaluable support and help in developing and writing this research work. I am grateful to my supervisor for generously giving me directions and suggestions for my paper revisions and the direction of the work and paper submissions. I must also offer my thanks and appreciation to Dr Enda Fallon for all of his support and help.