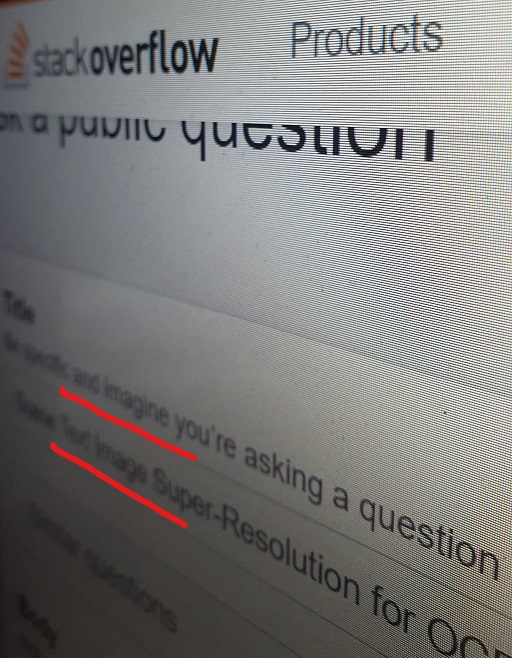
<https://stackoverflow.com/questions/64808986/scene-text-image-super-resolution-for-ocr>

Vấn đề trong bài toán OCR trên là ảnh đầu vào bị nhiễu (do chụp rung, do hiệu ứng chuyển động, do vị trí góc nên không lấy nét được)



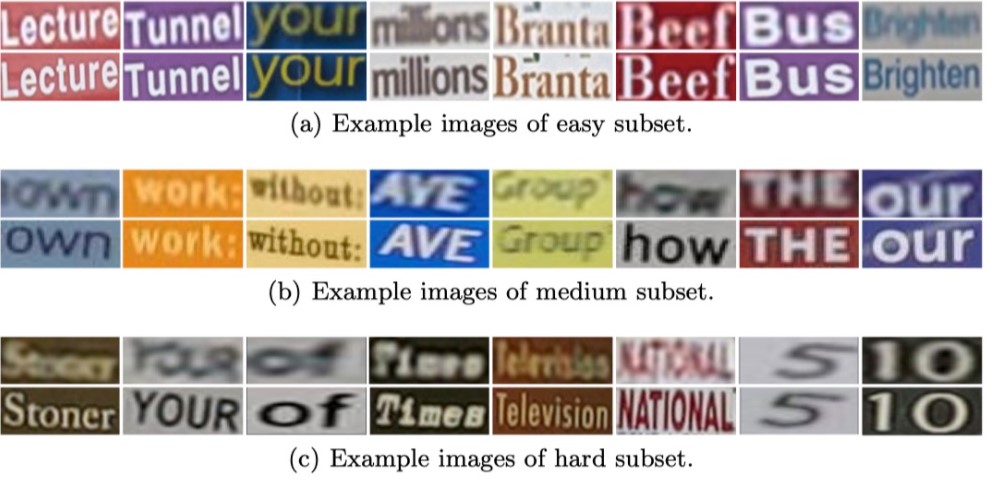
Tham khảo nhiều phương pháp (tesseract, opencv) nhưng không cải thiện được

Thử nghiệm các model liên quan đến SR (Super Resolution)

1. TSRN (Text Super-Resolution Network)

Không xử lý được đối với ảnh chữ màu xanh ở trên

Example của TSRN, em thấy áp dụng model này để train cho bài toán OCR của mình sẽ ra kết quả tốt



Link: [\*1609.04802 (arxiv.org)](https://arxiv.org/pdf/1609.04802)

1. Neural enhance

Không xử lý được, kể cả lấy sample trong link cũng không cho ra kết quả tốt như trong đó

1. ISR (Image Segmentation and Recognition)

Không áp dụng được

Một số vấn đề tiếp theo của bài viết nói về deblur, không phù hợp với vấn đề hiện tại của dự án.

Git hub: [twtygqyy/pytorch-SRResNet: pytorch implementation for Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network arXiv:1609.04802 (github.com)](https://github.com/twtygqyy/pytorch-SRResNet)

Paper: [\*1609.04802 (arxiv.org)](https://arxiv.org/pdf/1609.04802)

Bài viết thảo luận về sự phát triển của kỹ thuật siêu phân giải ảnh đơn (SISR), nêu bật các phương pháp và tiến bộ khác nhau trong lĩnh vực này.

Các phương pháp ban đầu như lọc tuyến tính, bicubic hoặc Lanczos tuy nhanh nhưng lại đơn giản hóa vấn đề quá mức, dẫn đến kết cấu quá mịn.

Các kỹ thuật và phương pháp bảo toàn cạnh tập trung vào việc thiết lập các ánh xạ phức tạp giữa các ảnh có độ phân giải thấp và cao đã được đề xuất.

Các phương pháp dựa trên cặp ví dụ và cách tiếp cận dựa trên học tập sử dụng dữ liệu huấn luyện đã cho thấy những cải tiến đáng kể.

Những tiến bộ gần đây bao gồm các phương pháp dựa trên dự đoán, kỹ thuật cảm biến nén và phương pháp nhúng vùng lân cận.

Sự xuất hiện của các thuật toán dựa trên mạng nơ ron tích chập (CNN) đã cho thấy hiệu suất vượt trội trong SISR.

Giới thiệu SRGAN, một mạng đối thủ tổng hợp (GAN) dành cho hình ảnh có độ phân giải siêu cao, đặt ra một trạng thái mới về nghệ thuật cho SISR ảnh thực tế với hệ số nâng cấp cao (4x).

SRGAN sử dụng chức năng mất cảm nhận, kết hợp mất đối thủ và mất nội dung, để tạo ra kết cấu hình ảnh chân thực từ các hình ảnh được lấy mẫu xuống.

Kiến trúc mạng bao gồm ResNet sâu được tối ưu hóa cho MSE và mức suy hao được tính toán trên các bản đồ đặc trưng của mạng VGG để cải thiện chất lượng cảm nhận.

**Method**

**SISR (Single Image Super Resolution)**

Mục đích là ước tính hình ảnh có độ phân giải cao từ đầu vào có độ phân giải thấp bằng cách sử dụng hàm tạo G được huấn luyện trên cặp hình ảnh LR và HR

Hàm tạo là mạng dư sâu được tối ưu hóa cho MSE, với perceptual loss function kết hợp adversarial loss và content loss

**GAN**

SRGAN sử dụng khung GAN cho hình ảnh có độ phân giải siêu cao, cho phép suy luận các hình ảnh tự nhiên giống như ảnh thực tế cho các yếu tố nâng cấp tỷ lệ 4×.

Sự mất mát đối nghịch hướng giải pháp tới các đặc điểm hình ảnh tự nhiên, trong khi sự mất mát nội dung tập trung vào sự tương đồng về cảm nhận hơn là sự tương tự về không gian pixel.

**Perceptual loss function**

Perceptual loss function là sự kết hợp có trọng số của một số thành phần mất mát để nắm bắt các đặc điểm mong muốn riêng biệt của hình ảnh siêu phân giải.

Nó bao gồm adversarial loss do mạng phân biệt đối xử gây ra và content loss được thúc đẩy bởi sự tương đồng về nhận thức, nâng cao độ trung thực của kết cấu được phục hồi.

(Chưa quá hiểu các công thức trong bài báo)

***Tóm lại:***

INPUT

Đầu vào của SRGAN là hình ảnh có độ phân giải thấp (ảnh LR) đã được downsampled từ hình ảnh có độ phân giải cao (ảnh HR).

Những hình ảnh LR này được sử dụng làm đầu vào cho hàm tạo G, nhằm mục đích tạo ra một phiên bản siêu phân giải của hình ảnh LR với độ chi tiết và độ trung thực được nâng cao.

OUTPUT

Đầu ra của SRGAN là hình ảnh có độ phân giải cao (hình ảnh SR) được tạo ra bởi hàm sinh G, đây là mạng dư sâu được tối ưu hóa cho MSE với chức năng mất nhận thức.

Hình ảnh SR dự kiến sẽ hấp dẫn về mặt hình ảnh, mang tính chân thực và chứa các chi tiết kết cấu mịn hơn so với hình ảnh LR đầu vào, đạt được mức tăng đáng kể về chất lượng cảm nhận theo thử nghiệm MOS.

**Scene Text Image Super-Resolution in the Wild**

**Medium**: [OCR with Deep Learning in PyTorch (Low Resolution OCR in the Wild) | by Mohamed Hasan | Medium](https://eng-mhasan.medium.com/ocr-with-deep-learning-in-pytorch-95a347e2ef29)

**Github**: [WenjiaWang0312/TextZoom: A super-resolution dataset of paired LR-HR scene text images (github.com)](https://github.com/WenjiaWang0312/TextZoom/tree/master)

**Paper**: [2005.03341 (arxiv.org)](https://arxiv.org/pdf/2005.03341)

* Super Resolution giúp tạo ra output là hr image, với input là các ảnh lr tương ứng.
* Một số phương pháp cũ sử dụng filter như bilinear, bicubic với output hr image (mỗi pixel hr tạo ra bằng cách interpolating (nội suy) các màu pixel lân cận)
* Với inputh là các lr img, mục tiêu của SR sẽ là tối thiểu hóa khoảng cách giữa output là hr img so với ảnh gốc.

Đối với dataset, ta có thể thu thập/tạo theo 2 cách:

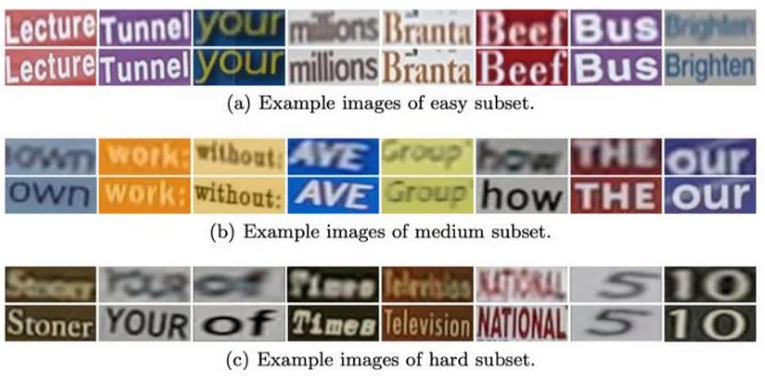
* Thu thập các ảnh hr, sau đó biến thành ảnh lr
* Cách 2 là thu thập các cặp ảnh lr-hr bằng cách chụp và điều chỉnh tiêu cự camera

Trong bài viết trên, ta có dataset TextZoom được thu thập theo cách 2 và model TSRN để dự đoán các lr img

Figure dưới so sánh ảnh tổng hợp & ảnh chụp và ảnh hr



Dataset TextZoom được chia thành 3 bộ con (easy subset – medium subset – hard subset)

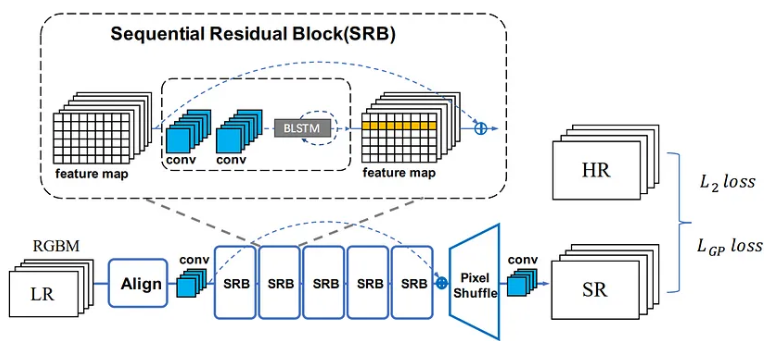


**Model TSRN**

**TSRN** là mô hình end to end, dùng để tái tạo lại text từ các lr img.

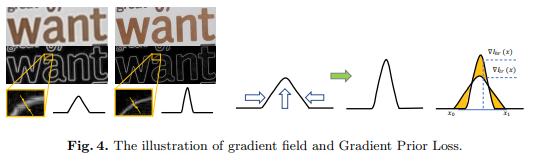
Input là lr img – Output là hr img

Dưới đây là pipeline



Giải thích pipeline:

* LR Input = RGB + binary mask (binary mask = mean gray scale of image)
* Thêm 1 central alignment module phía trước network
* Tiếp tục sử dụng CNN để trích xuất các đặc trưng cơ bản từ image
* Sau đó, thay thế các block cơ bản thành khối SRB (Sequential Residual Block) xếp chồng lên nhau → trích xuất các đặc trưng sâu hơn theo 1 cách tuần tự
* SR Ouput được generate ra sẽ được supervised bởi 2 hàm loss:
  + Gradient Prior Loss (Lgp): 1 hàm dùng để tăng cường biên hình dạng của ký tự.



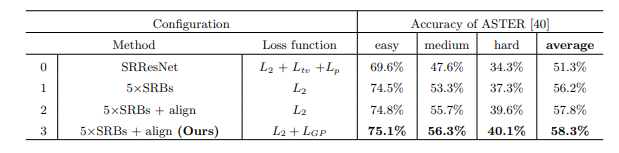


* + Hàm L2, để tính khoảng cách giữa SR output và HR image

**Ablation Study on TSRN**

Theo Pipeline:

* Việc thêm 5 khối SRB xếp chồng lên nhau đã giúp tăng accuracy lên 9% so với SRResnet.
* Align module giúp tăng accuracy lên 1,5%
* Hàm Lgp cũng giúp tăng accuracy lên 0.5%, mặc dù tăng không đáng kể, nhưng kết quả trực quan vẫn tốt hơn

****

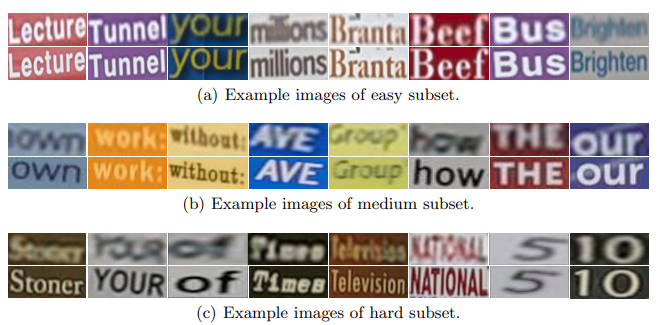
Ảnh dưới đây so sánh trực quan accuracy của ASTER, những chữ đỏ là nhận dạng sai



Thông qua các thử nghiệm, TSRN có kết quả vượt trội so với các phương pháp SR khác



**Thử nghiệm trên Textzoom**

****