Two-Stage Seamless Text Erasing on Real-World Scene Images

Benjamin Conrad[1] Pei-I Chen[2]
[1] University of Amsterdam [2] Jumio AI Labs

Introduction

- **Text erasing** is the task of removing all words, numbers and characters found in an image and filling in these pixels in a realistic fashion.

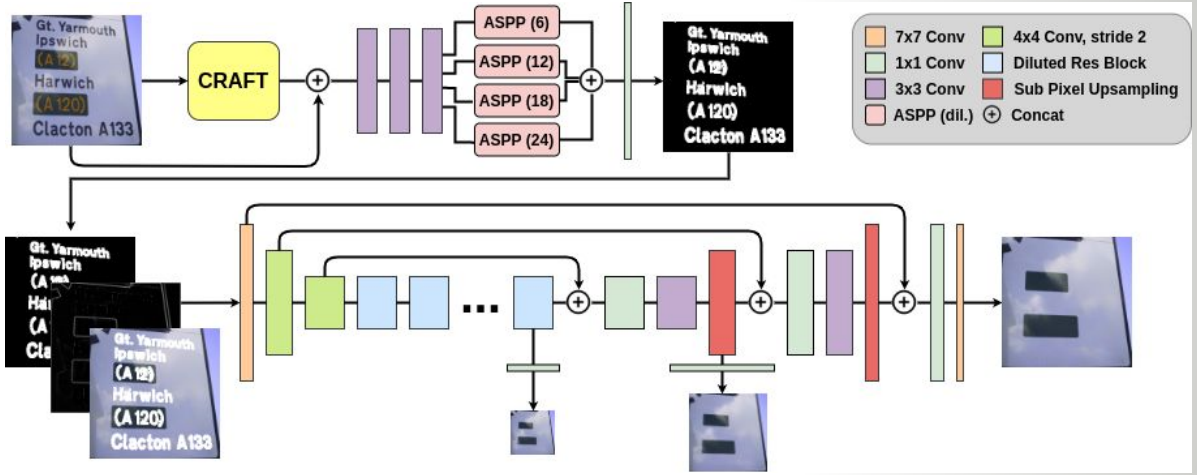
- **Use cases:** Removing sensitive information (license plate numbers), text swapping, dataset creation.

- Previous approaches **struggle on real-world examples**:
  - Fails to remove all text
  - Noticeable artifacts
  - Lack of fine details

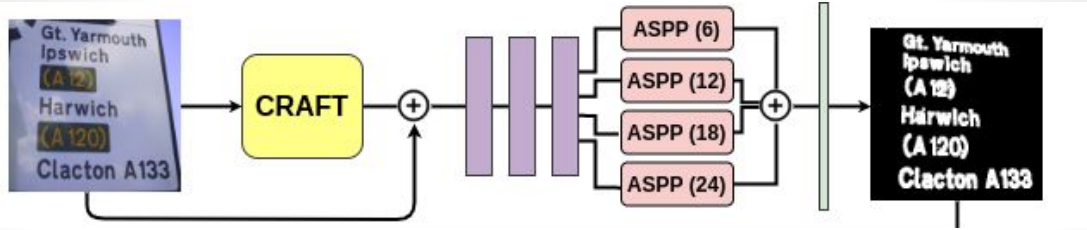
Two-Stage Text Erasing Pipeline

Stage 1: Text Mask Generator

Stage 2: Inpainting Model



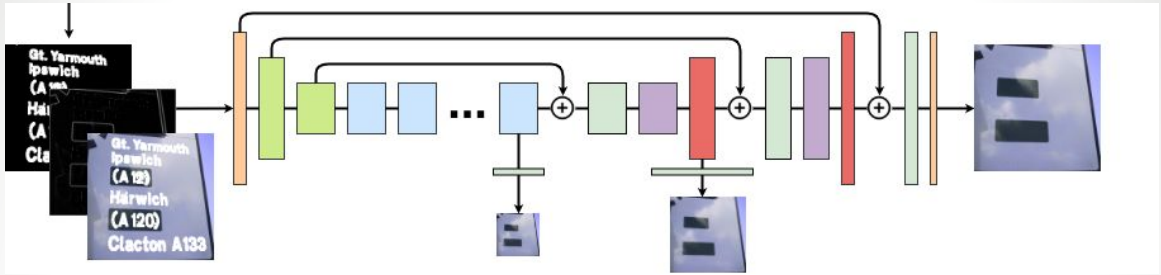
Stage 1: Text Mask Generator



- Image passed through CRAFT[1] text detector to extract per character saliency information.
- Detector output is further refined using an ASPP segmentation head to generate a binary text segmentation mask.
- To ensure that text is entirely covered, the model is trained using the Tversky Loss[2].
  - α=0.1 β=0.9 □ **Penalize false negatives more than false positives**.

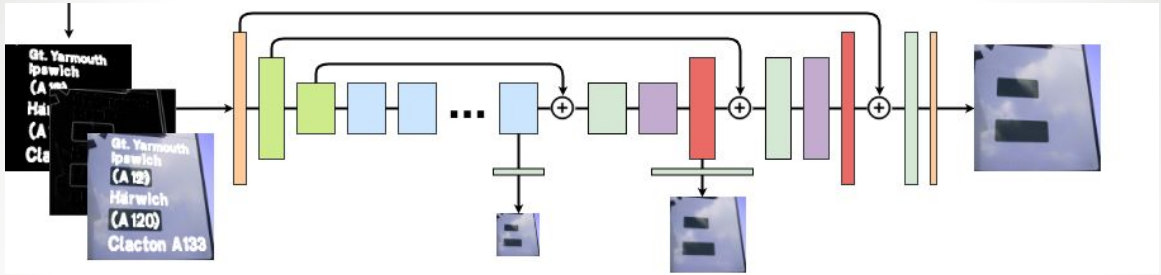
$$L_{tvky} = \frac{TP}{TP + \alpha FP + \beta FN}$$

Stage 2: Inpainting Model



- The masked image, image gradients and generated mask are passed through the inpainting model to produce a text-free version of the original image.
- Architecture builds off EdgeConnect's[2] image completion network.
  - Includes additional improvements such as **skip connections, sub-pixel upsampling[3] and multiscale generation.**

Stage 2: Inpainting Model



- **Multiscale Gradient Reconstruction Loss:**
  - Enforces model to produce **sharp edges** as well as **smooth surfaces**.
  - Low computational cost

$$L_{gr} = \frac{\sum_i \|\nabla I_{pred_i} - \nabla I_i\|_2^2}{S_i}$$

- **Other Loss Functions:**
  - **Full resolution**: L1, gradient, style, perceptual, total variation, adversarial.
  - **All resolutions**: L1, gradient

Training

- Text mask generator trained on **real-world images** from ICDAR 2013 and TotalText.
- Inpainting model trained on **synthetic images** from SynthText.

- Each stage is **trained separately** and combined during evaluation to create the text erasing pipeline

Synthetic Evaluation

- Evaluated on SCUT synthetic text erasing dataset.
- Our method **matches the state of the art** while not being optimized for synthetic data.

Method	PSNR†	SSIM†	MAE*	Reported PSNR†	Reported SSIM†
EnsNet [2]*	31.18	91.12	0.018	37.36	96.44
MTRNet [3]	30.56	90.14	0.021	29.71	94.43
MTRNet++ [4]	**33.43**	93.10	0.015	34.55	**98.45**
WS-TE (ResNet-50) [5]*	30.73	93.43	0.016	37.44	93.69
WS-TE (ResNet-152) [5]	-	-	-	**37.46**	93.64
Ours	32.97	**94.90**	**0.013**	32.97	94.90

Table 3. Quantitative results on the SCUT dataset, including our re-calculated values along with each method's reported values.
*Values are from re-implementations. †Higher is better.*Lower is better.

Input EnsNet[4] MTRNet[5] MTRNet++[6] Zdenek et al.[7] Ours

Real-World Evaluation

- Human perceptual study conducted with images from ICDAR 2013 test set.
- Our method is **significantly prefered** over previous state of the art on real-world images.

Method	# Images	% Votes
WS-TE (ResNet-50) [5]	12	18%
Ours	**213**	**82%**
Tie	8	-

Table 2. Results of our human perceptual study. The second column shows the number of images that each model received a majority of votes.

Input Zdenek et al.[7] Ours

jumio

jumio.com

Thank You

jumio.com

References

[1] Baek, Youngmin, et al. "Character region awareness for text detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[2] Nazeri, Kamyar, et al. "Edgeconnect: Generative image inpainting with adversarial edge learning." arXiv preprint arXiv:1901.00212 (2019).
[3] Shi, Wenzhe, et al. "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[4] Zhang, Shuaitao, et al. "Ensnet: Ensconce text in the wild." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.
[5] Tursun, Osman, et al. "Mtrnet: A generic scene text eraser." 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019.
[6] Tursun, Osman, et al. "MTRNet++: One-stage mask-based scene text eraser." Computer Vision and Image Understanding 201 (2020): 103066.
[7] Zdenek, Jan, and Hideki Nakayama. "Erasing scene text with weak supervision." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020.