Don’t Miss the Fine Print! An Enhanced Framework to Extract Text from Low Resolution Images

Pranay Dugar[1], Aditya Vikram[2,*], Anirban Chatterjee[1], Kunal Banerjee[1][a] and Vijay Agneeswaran[1]
[1] *Walmart Global Tech, Bangalore, India*
[2] *Flipkart, Bangalore, India*

Keywords: Scene Text Recognition, Super-resolution, Text Extraction, Convolution Neural Network.

Abstract: Scene Text Recognition (STR) enables processing and understanding of the text in the wild. However, roadblocks like natural degradation, blur, and uneven lighting in the captured images result in poor accuracy during detection and recognition. Previous approaches have introduced Super-Resolution (SR) as a processing step between detection and recognition; however, post enhancement, there is a significant drop in the quality of the reconstructed text in the image. This drop is especially significant in the healthcare domain because any loss in accuracy can be detrimental. This paper will quantitatively show the drop in quality of the text in an image from the existing SR techniques across multiple optimization-based and GAN-based models. We propose a new loss function for training and an improved deep neural network architecture to address these shortcomings and recover text with sharp boundaries in the SR images. We also show that the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) scores are not effective metrics for identifying the quality of the text in an SR image. Extensive experiments show that our model achieves better accuracy and visual improvements against state-of-the-art methods in terms of text recognition accuracy. We plan to add our module on SR in the near future to our already deployed solution for text extraction from product images for our company.

1 INTRODUCTION

Textual information contained in images can bolster the semantic understanding of real-world data. Extracting text from an image has many applications, especially in the retail industry, such as, determining brand name, ingredients, price and country of origin of a product and detecting profanity. Generally, this task follows a two-step procedure. First, localize the text contained in an image using either a character-based or a word-based model. Second, identify the text in the localized region using a sequence-to-sequence model. These tasks are challenging due to the image degradation, image complexities, and diversity in sizes, shapes, and orientations of texts. Recent text extraction models have performed impressively on clear text but show a significant decline in accuracy when recognizing text in low-resolution images (Ye et al., 2020; Feng et al., 2019; Baek et al., 2019).

Over the years, various deep learning models have been designed to improve the quality of the images, and the items present in these images based on the use cases. Super-Resolution (SR) is one such technique used to improve the quality of an image by increasing its resolution while retaining edge consistency, creating a High-Resolution (HR) image from its Low Resolution (LR) counterpart. Various SR methods have been suggested based on deep neural architectures which show great promise. However, on attempting to utilize these models on the task of text extraction, it was observed that the image lost the clarity of text even though the overall image became sharper than the original.

In this paper, we attempt to address some of these problems. The significant contributions of our work are:

- An approach to generate synthetic LR-HR paired data that is generalizable to real case scenarios for product images.

[a] https://orcid.org/0000-0002-0605-630X

[*] Work done during internship at Walmart Global Tech, India, when the author was a student at Indian Institute of Science.

Dugar, P., Vikram, A., Chatterjee, A., Banerjee, K. and Agneeswaran, V.
Don’t Miss the Fine Print! An Enhanced Framework to Extract Text from Low Resolution Images.
DOI: 10.5220/0010971100003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 664-671
ISBN: 978-989-758-555-5; ISSN: 2184-4321

- A variation of perceptual loss termed recognition loss that effectively deblurs and sharpens the boundaries of the texts in the image while preserving textual characteristics.
- An improvised multi-loss function composed of detection and recognition losses as well as image features.
- Qualitative and quantitative view of how PSNR and SSIM (Horé and Ziou, 2010) are not good measures of image quality post super-resolution for textual details.
- Visually and analytically superior results for text super-resolution as compared to existing approaches.

It is worth noting that we plan to add our super-resolution solution shortly into our current deployment for text extraction from product images (Dugar et al., 2021), which has been in production for a year within Walmart.

The paper is organized as follows. In Section 2, we cover related work along with our motivation. Section 3 presents our methodology. The experimental results can be found in Section 4. Some of the additional application areas of our method are described in Section 5. The paper is concluded in Section 6.

2 RELATED WORK AND MOTIVATION

Text extraction from scene images is a widely studied topic. Many accurate and efficient methods that extract textual information from scene images have been proven effective in different constrained scenarios. The focus of many of the recent works (Wei Liu and Han, 2016; Liu et al., 2018; Luo et al., 2019) has been on natural scenes, which address challenges due to the high diversity of texts in blur, orientation, shape, and low-resolution. Traditionally, the problem to extract text from a low-resolution image is thought to have two primary aspects: super-resolution and text recognition.

Super-resolution aims to output a high-resolution image that exhibits consistency with the corresponding low-resolution image. Traditional approaches, such as bilinear, bicubic or designed filtering, are based on the assumption that the neighbouring pixels exhibit similar colours and produce the output by interpolating colours between neighbouring pixels. In the deep learning era, one of the most common approaches to address this problem is to map it to a regression problem, where we design a complex non-linear function that outputs the high-resolution image on being fed the low-resolution image as an input (Dong et al., 2016; Kim et al., 2016; Ledig et al., 2017). Then the textual information is extracted from the high-resolution image.

As far as the text recognition is concerned, there is literature that adopts a bottom-up fashion (Jaderberg et al., 2014) that detects individual characters first and then combines these into a word, or a top-down fashion (Jaderberg et al., 2015a) that treats the word image region as a whole and addresses it as a multi-class image classification problem. Based on the fact that the scene texts generally appear in character sequences in scene text images, CRNN (Shi et al., 2017) maps it to a sequence recognition problem and leverages the Recurrent Neural Network (RNN) to model the sequential features. Recently, attention mechanism has gained importance in text recognition literature (Luo et al., 2019). ASTER (Shi et al., 2019) addresses the problem with oriented or curved texts using Spatial Transformer Network (STN) (Jaderberg et al., 2015b), which is followed by text recognition using an attentional sequence-to-sequence model.

However, the main difficulty of recognising LR text is that the optical degradation blurs the characters' shape, which impedes the methods mentioned above to exhibit optimal performance while extracting text from many low-resolution images. In this work, we experiment with different kinds of loss functions, such as a variant of the perceptual loss, and an improvised multi-loss function combining both detection and recognition losses to get over the problem of blurring of character shapes.

3 METHODOLOGY

3.1 Data Collection and Annotation

The efficacy of the neural networks to approximate any function depends heavily on the dataset used to train the model. Previous approaches have generated a paired LR-HR dataset by downsampling the HR images using the existing interpolation methods such as linear, bicubic, and nearest-neighbour interpolation. However, we cannot take such a dataset as a sample representative of the natural scene text datasets. A single down-sample formulation generates all the LR images, and the model only learns the inverse of the downsampling function to generate the SR images. Recently, the authors of (Wang et al., 2020; Cai et al., 2019; Zhang et al., 2019) have suggested using images taken by a digital camera at different focal lengths to create an ideal paired LR-HR dataset for

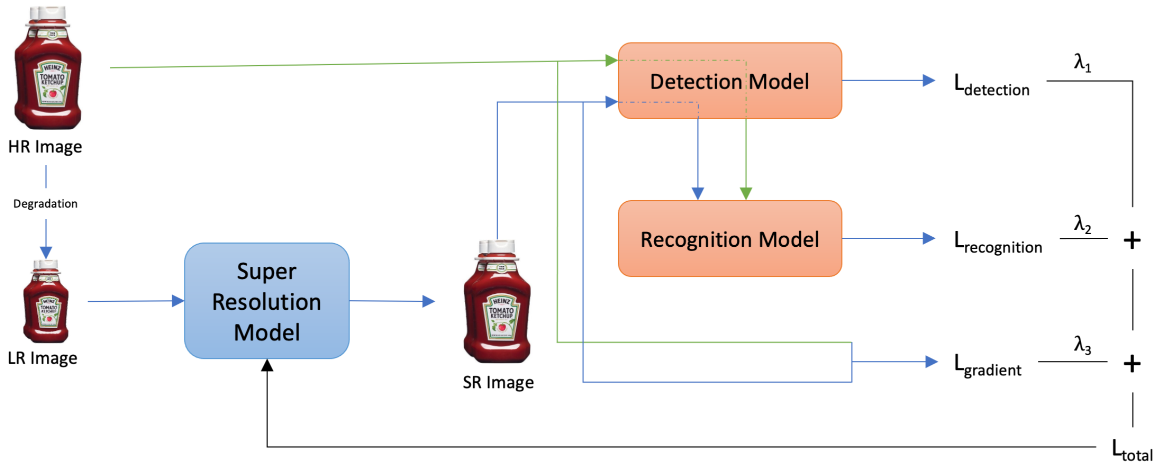

Figure 1: Architecture diagram representing the complete flow of model training.

image super-resolution. However, this is not a feasible approach to generate large-scale datasets required to train models.

We devised an approach to generate a suitable LR-HR pair for any large dataset to circumvent these challenges. Our proposal involves a two-stage interpolation method to generate a synthetic dataset that can mimic the natural scene text datasets. For a paired 2× LR-HR, we first downsample the original image to one-fourth of its original dimensions, followed by its upsampling to one-half of its original dimensions. Different interpolation techniques were randomly chosen for both downsampling and upsampling to introduce more randomness in the dataset. We use in-built interpolation methods in the Torchvision library (i.e. linear, bicubic, nearest, box, Hamming, and Lanczos) for both downsampling and upsampling of the images. Further, for training the model in batch mode, we create image patches of size 400 pixels × 400 pixels of the HR image and 200 pixels × 200 pixels for the LR image.

3.2 Loss Functions

Despite high PSNR values, pixel value based loss functions like Mean Squared Error (MSE), Mean Absolute Error (MAE) fail to generate images with high-level attributes, such as, textures. However, the existing perceptual loss function by (Johnson et al., 2016) uses a pre-trained model to calculate the differences between the target and the output image in the feature space of the neural network, and generates high texture quality images, but fails to do justice with the reconstruction of the texts in the generated SR image.

Recognition Loss. We add this new loss to the family of perceptual losses that focuses entirely on reconstructing high-quality texts with sharp boundaries and fine edges in the SR image. We leverage the feature maps generated by the fourth convolutional block of the pre-trained encoder of the text recognition ASTER model (Shi et al., 2019). We define Recognition Loss as the MSE between these feature maps of the generated SR image and the original HR image. Note that our experiments have confirmed that the recognition loss adapts well with various text extraction use-cases, and the reconstructed text is of better quality than all other existing techniques.

$$L_{rec} = ||\Psi_n(I^{HR}) - \Psi_n(I^{SR})||_2 \qquad (1)$$

where Ψ is the feature map obtained as an output of the n-th block of the ASTER's encoder model. Through multiple iterations, we found that output of the 4-th block works the best for text recognition related purposes.

Gradient Loss. Taking inspiration from HOG (Dalal and Triggs, 2005), we propose Gradient Loss to ensure that the model can better detect edges and corners in the images. The gradient is calculated along each channel, followed by the mean across channels to negate abnormalities across different image channels. Finally, MAE was used to calculate the gradient loss between the SR and the HR image pairs.

$$L_{grad} = ||\Delta I^{HR} - \Delta I^{SR}||_1 \qquad (2)$$

Here, Δ represents the cumulative gradient of the image and is calculated as shown below in equation 3.

$$\Delta I = \frac{1}{2 \times channels} \sum_{channels} (\delta I_{width} + \delta I_{height}) \qquad (3)$$

where, δ is the gradient of the image along its height/width, and is calculated as per the (Dalal and

Triggs, 2005). Please note that δ has the same dimensions as the image on which it is calculated, i.e., (width, height, channels) while Δ has the dimension (width, height, 1).

Detection Loss. This loss is proposed to ensure that the model can detect the precise locations of all the texts in an image with higher accuracy. A pre-trained CRAFT model (Baek et al., 2019) was used to generate the locations of the texts in the SR image generated by the model and the original HR. We use the predicted coordinates of the SR and the HR images to create two mask images consisting of detected regions being masked out using equation 4 for each pixel. An MSE across the two masks is taken as the final loss value. Thus, the loss is a pixel-wise MSE where each location represents if that pixel is part of text or not.

$$img\_mask(p) = \begin{cases} 1, & \text{if p in detected box} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$L_{det} = \frac{1}{P}\sum_{p} ||HR\_mask(p) - SR\_mask(p)||^2 \tag{5}$$

where, P is the total number of pixels in the image, summation taken over every individual pixel p, HR_mask is the detection mask created for HR image and SR_mask is the detection mask created for SR image.

Overall loss for the task is defined as:

$$\begin{aligned} TotalLoss = \lambda_1 L_{rec} + \lambda_2 L_{det} + \lambda_3 L_{grad} \\ + \lambda_4 L_{tv} + \lambda_5 L_{vgg} + \lambda_6 L_{mse} \end{aligned} \tag{6}$$

where, λ values are $[1e-2,\ 6e-5,\ 1e-4,\ 2e-4,\ 6e-3,\ 1e-0]$ in the same order. Except Total Variation (TV) Loss, which is measured only on the output SR image, every other loss functions takes into account both the HR image and the SR image. L_{vgg} is the perceptual loss calculated on VGG19.

Our high-level architecture diagram is shown in Figure 1. We start with a HR image from the dataset that we down-sample using the in-built methods in Torchvision library to obtain its corresponding LR image. This LR image is then fed into our super-resolution model to generate the SR image. The HR and the SR images are passed as inputs to the detection model followed by the recognition model. We collect the losses $L_{detection}$, $L_{recognition}$ and $L_{gradient}$, compute their weighted sum termed as L_{total} and use it to train our super-resolution model.

4 EXPERIMENTAL RESULTS

As in any super-resolution framework, there are two ways to gauge the performance of a model: visual perception and analytical scores. Through the following sections, we will cover these two aspects of our model in detail.

4.1 Dataset

As the focus of our model is to improve the text in an image, we perform experiments on datasets designed for the task of text extraction from images. These are open-source datasets such as ICDAR2013 (Karatzas et al., 2013), ICDAR2015 (Karatzas et al., 2015) and SVT (Wang et al., 2011). These three datasets provide word-level ground truth boxes of text in an image. We use these ground truth boxes as the area of consideration for our model in terms of visual perception and the ground truth text for analytical scoring metrics. A small caveat, though, is that the ground truth provided does not comprise all the words in the image but only the significant ones that are more clearly visible. The design of our model is such that it improves not only these significant words but also the non-significant words (small/slightly blurred). However, due to the lack of ground truth, we will see the improvement for these non-significant words only through visual perception. We downsample the images from the three datasets for creating a LR image dataset, and the original images act as the HR ground truth images. We compare our model against some state-of-the-art super-resolution models such as DNCNN (Zhang et al., 2017), IMDN (Hui et al., 2019) and ESRGAN (Wang et al., 2018).

4.2 Visual Perception

A super-resolution model is only as good as the amount of finer details that it can improve. The existing approaches perform effectively in terms of improving the quality of the overall image. However, as seen in Figure 2, the character boundaries get blurred after super-resolution in these models. Standard metrics used to verify the quality of super-resolution models are the PSNR and the SSIM scores (Horé and Ziou, 2010). However, as shown in Table 1, these metrics do not do justice in terms of the quality of the characters in the image. Some existing models give a higher value for PSNR and SSIM scores, but the images tell a different story. Of the six PSNR and SSIM scores for the three datasets, our model performs best only for the SSIM score for SVT dataset.

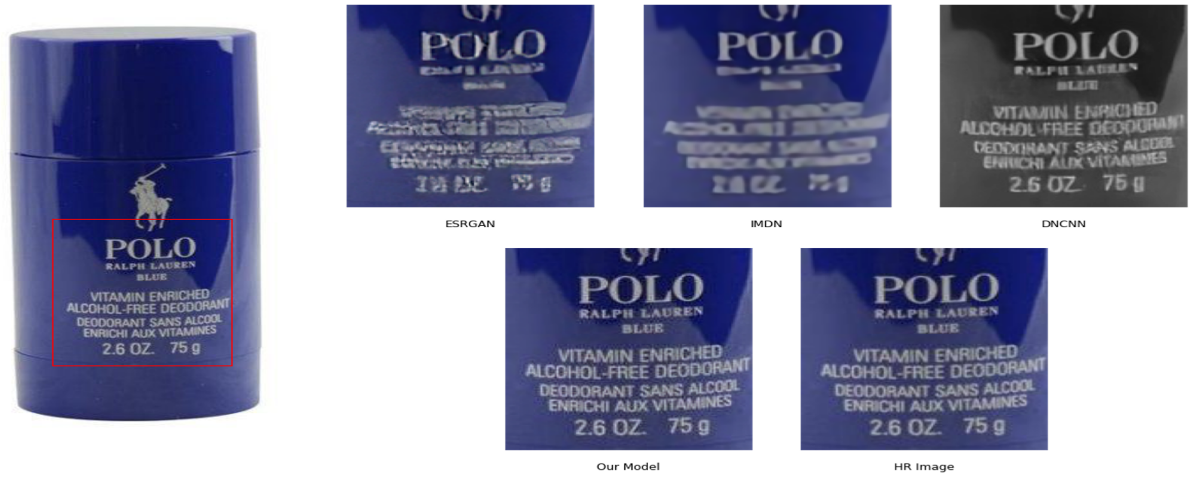

Figure 2: Super-resolution outputs for a product image with *clear* text by various models with the given HR reference image as input. The models names have been specified below the images.

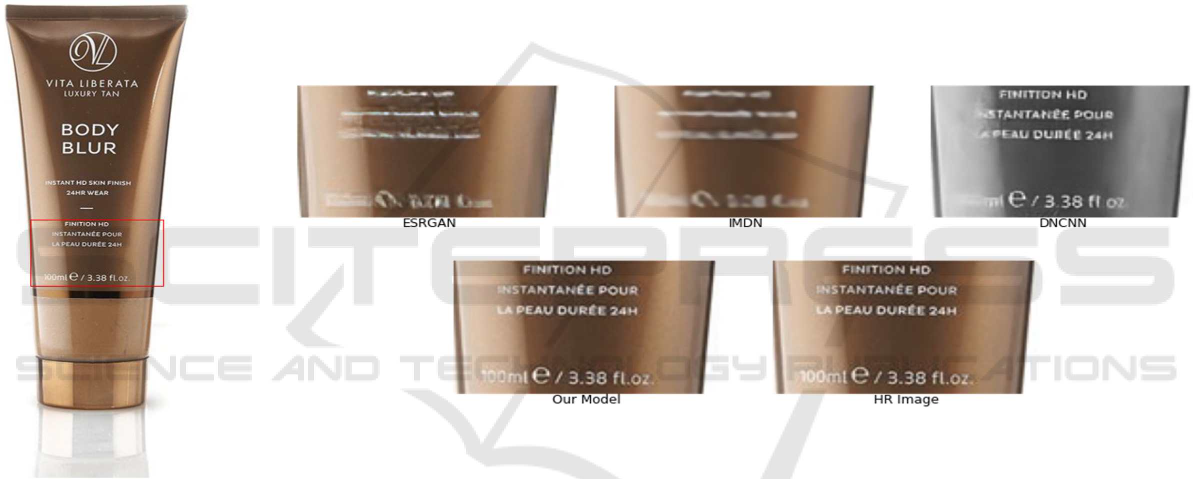

Figure 3: Super-resolution outputs for a product image with *small* text by various models with the given HR reference image as input. The models names have been specified below the images.

The quality of the text drops even further while considering the words that are not significant. This drop can be seen clearly in Figure 3. Though not entirely accurate, our model gives much better character boundaries than the existing models. Since visually it is clear that the model is performing significantly better and that PSNR and SSIM scores are not effective measures, we performed a more rigorous analysis to show that the model is significantly better in terms of text recognition.

To reduce the chance of misinterpretation, we had asked three annotators to independently check the images produced by the competition (ESRGAN (Wang et al., 2018), IMDN (Hui et al., 2019), DNCNN (Zhang et al., 2017)) and ours to identify the one from which understanding the text was the easiest – this was a blind process, i.e., the annotators did not know which method produced which output. For this experiment, we had chosen 20 images from each of the datasets: ICDAR2013, ICDAR2015 and SVT. We found that in ∼75% of the cases, the images produced by our method was declared the winner in spite of having lower SSIM and PSNR scores, as mentioned in Table 1. Kindly, note that the images in Figure 2 and Figure 3 are sample images which depict these results.

4.3 Text Recognition Analysis

From the SR images generated from different models, using the ground truth boxes provided in the dataset, the text areas are cropped and sent through the text recognition model defined in (Dugar et al., 2021). First, we compare the accuracy – a direct

Table 1: PSNR and SSIM scores of various models compared against our model. Note that the scores are averaged over only the regions of the ground truth boxes used in text recognition as these represent our areas of concern.

Model	ICDAR2013		ICDAR2015		SVT	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ESRGAN	29.432	0.827	29.338	0.826	30.458	0.839
IMDN	**32.266**	0.881	**32.170**	0.881	**33.383**	0.895
DNCNN	32.022	**0.897**	32.017	**0.897**	32.464	0.910
Our Model	29.236	0.882	29.122	0.881	32.545	**0.928**

Table 2: Normalised Edit Distance (Norm ED) of text and accuracy of an exact match for images generated from the two models (our model: Text SR Image and generic model: IMDN SR Image); we also provide these scores for the High-Resolution (HR) image for reference. Note that we use the same text recognition model in all three cases.

Dataset	Score Type	HR Image	Text SR Image	IMDN SR Image
ICDAR2013	Norm ED	0.954	**0.928**	0.919
	Accuracy	0.903	**0.876**	0.833
ICDAR2015	Norm ED	0.972	**0.958**	0.938
	Accuracy	0.908	**0.890**	0.836
SVT	Norm ED	0.930	**0.921**	0.848
	Accuracy	0.827	**0.821**	0.721

match of ground truth word, and normalised edit distance (Marzal and Vidal, 1993) – a character level comparison, of the backbone IMDN model against the model trained by our approach. For reference, these were both compared against the accuracy score on HR images, and we present the results in Table 2. The model trained by our approach gets closer to the accuracy score for the HR images.

The results motivated us to compare our model against other state-of-the-art super-resolution models. Table 3 shows the performance of various models on the given datasets. On all the datasets, our model performs significantly better than these models in terms of text recognition.

Though the PSNR and the SSIM scores of our model are lower than that of the existing models, it still achieves a better result in both visual and analytical terms.

5 ADDITIONAL APPLICATION AREAS

The technology described here is generic enough to be applied to various other application areas beyond what we report on product images here and in (Dugar et al., 2021) albeit with some domain-specific finetuning. We note a couple of such application areas here.

5.1 Healthcare

Walmart is devoted to serving its customers by delivering goods and merchandise at affordable prices and by facilitating healthier lifestyles. In addition to selling medicines at our stores, *Walmart Health* (Staff, 2019a) now provides primary, urgent and preventive healthcare services in some of our supercenters. While selling or administering medicine, one must be very careful about its expiry date to prevent harmful effects. Moreover, getting the ingredients wrong for a medicine may also endanger human lives. Therefore, unlike standard product images, the tolerance level of making a false prediction is close to zero in healthcare. Extracting the dates, especially, can be much more challenging because these are added to the labels at a later stage and are often more obscure than the rest of the text; an example of the same can be found in Figure 4. Our solution can be helpful in this domain with some small improvements, such as, adding the names of the drugs and their constituents

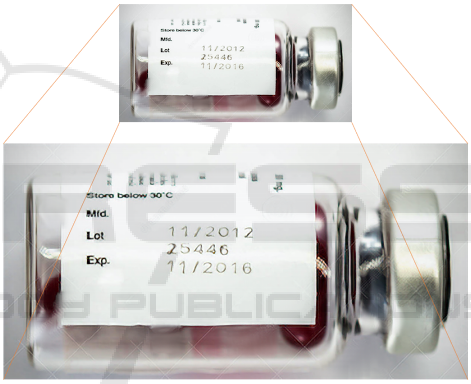

Figure 4: Extracting manufacturing and expiry dates from a medicine bottle. Note that the font, size and color of these dates are blurrier from the rest of the label, and also not aligned.

Table 3: Accuracy and Normalised Edit Distance for text recognition from images generated by our model against images from other state-of-the-art super-resolution models.

Model	ICDAR2013		ICDAR2015		SVT	
	Accuracy	NormED	Accuracy	NormED	Accuracy	NormED
ESRGAN	0.808	0.881	0.814	0.905	0.684	0.817
IMDN	0.833	0.919	0.836	0.938	0.721	0.848
DNCNN	0.853	0.919	0.863	0.945	0.726	0.853
Our Model	**0.876**	**0.928**	**0.890**	**0.958**	**0.821**	**0.921**

into our dictionary because these names do not appear in regular text.

5.2 Edge Devices

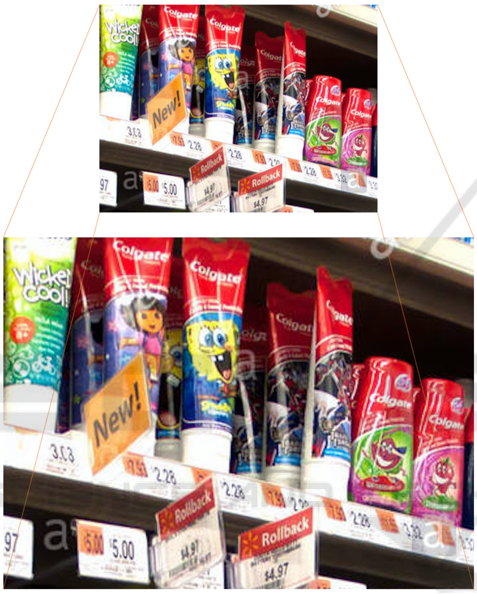

Figure 5: Extracting information about products on display. This information may help in identifying low or out-of-stock products, and/or notifying damaged products.

Recently, Walmart has given away smartphones with built-in apps to 740K associates to help them in their day to day activities in various ways (Staff, 2021b). We can further leverage these devices for inventory management and quality checks; for example, an associate may take a picture and notify the warehouse administration upon detecting a damaged product. However, the cameras mounted on the smartphones may not be of high definition, or the pictures may be taken from a distance, or there can be jerky hand movements – all of which may lead to low quality, tiny or blurry images. Similarly, the surveillance cameras placed on top of the aisles in Walmart stores and clubs may also be re-purposed to additionally gather information on products that are low or out-of-stock and identify damaged goods (Staff, 2019b); however, these images may again be of low quality. Our SR based solution may also contribute in such cases as shown in Figure 5. Another potential use case can be reading road signs for autonomous cars; Walmart has been looking in this space for supply chain management (Staff, 2020), especially for the last mile delivery (Staff, 2021a).

6 CONCLUSION

This paper proves the importance of the scene text image super-resolution for text detection and recognition. We have proposed an alternative way to generate the synthetic paired LR-HR dataset that mimics the actual data compared to the simple bicubic downsampling of the HR images. We have demonstrated that the model trained on our dataset is superior to the models trained on images generated by bicubic downsampling to handle scene text images in the wild through a series of experiments. To handle scene text image super-resolution, we have proposed Recognition Loss and an improvised architecture that enables the model to reconstruct the texts with clear boundaries and sharp edges in real-time. Our method outperforms multiple SR methods by a significant margin. However, it also shows that we are still far from decoding the highly degraded low-resolution scene texts, and the field requires more effort to solve the same.

In the future, we plan to include more diverse scene text image datasets across multiple languages and with different alignments to train the model better. We will also try to develop an improved loss function that will possibly outperform our current benchmarks. Introducing vision transformers into the scene text super-resolution domain may further push the performance, and hence we aim to investigate these models as well.

REFERENCES

Baek, Y., Lee, B., Han, D., Yun, S., and Lee, H. (2019). Character region awareness for text detection. In *CVPR*, pages 9365–9374.

Cai, J., Zeng, H., Yong, H., Cao, Z., and Zhang, L. (2019). Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, pages 3086–3095.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893.

Dong, C., Loy, C. C., He, K., and Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:295–307.

Dugar, P., Bhat, R. S., Tarsode, A. S., Dutta, U., Banerjee, K., Chatterjee, A., and Agneeswaran, V. S. (2021). From pixels to words: A scalable journey of text information from product images to retail catalog. In *CIKM*, pages 3787–3795.

Feng, W., He, W., Yin, F., Zhang, X.-Y., and Liu, C.-L. (2019). Textdragon: An end-to-end framework for arbitrary shaped text spotting. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9075–9084.

Horé, A. and Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369.

Hui, Z., Gao, X., Yang, Y., and Wang, X. (2019). Lightweight image super-resolution with information multi-distillation network. In *MM*, pages 2024–2032.

Jaderberg, M., Simonyan, K., Vedaldi, A., and Zisserman, A. (2015a). Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116:1–20.

Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. (2015b). Spatial transformer networks. In *NIPS*, pages 2017–2025.

Jaderberg, M., Vedaldi, A., and Zisserman, A. (2014). Deep features for text spotting. In *ECCV*, pages 512–528.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711.

Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V. R., Lu, S., et al. (2015). Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE.

Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., Bigorda, L. G. i., Mestre, S. R., Mas, J., Mota, D. F., Almazàn, J. A., and de las Heras, L. P. (2013). Icdar 2013 robust reading competition. In *ICDAR*, pages 1484–1493.

Kim, J., Lee, J. K., and Lee, K. M. (2016). Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 105–114.

Liu, Z., Li, Y., Ren, F., Goh, W., and Yu, H. (2018). Squeezedtext: A real-time scene text recognition by binary convolutional encoder-decoder network. In *AAAI*, pages 7194–7201.

Luo, C., Jin, L., and Sun, Z. (2019). Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90:109–118.

Marzal, A. and Vidal, E. (1993). Computation of normalized edit distance and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):926–932.

Shi, B., Bai, X., and Yao, C. (2017). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304.

Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., and Bai, X. (2019). Aster: An attentional scene text recognizer with flexible rectification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9):2035–2048.

Staff, W. (2019a). Walmart health. Accessed: 2021-17-09.

Staff, W. (2019b). Walmart's new intelligent retail lab shows a glimpse into the future of retail, irl. Accessed: 2021-17-09.

Staff, W. (2020). Walmart and gatik go driverless in arkansas and expand self-driving car pilot to a second location. Accessed: 2021-17-09.

Staff, W. (2021a). Walmart invests in cruise, the all-electric self-driving company. Accessed: 2021-17-09.

Staff, W. (2021b). Walmart unveils all-in-one associate app, me@walmart, and gives 740,000 associates a new samsung smartphone. Accessed: 2021-17-09.

Wang, K., Babenko, B., and Belongie, S. (2011). End-to-end scene text recognition. In *2011 International Conference on Computer Vision*, pages 1457–1464.

Wang, W., Xie, E., Liu, X., Wang, W., Liang, D., Shen, C., and Bai, X. (2020). Scene text image super-resolution in the wild. In *ECCV*, pages 650–666.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Loy, C. C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*.

Wei Liu, Chaofeng Chen, K.-Y. K. W. Z. S. and Han, J. (2016). Star-net: A spatial attention residue network for scene text recognition. In *BMVC*, pages 43.1–43.13.

Ye, J., Chen, Z., Liu, J., and Du, B. (2020). Textfusenet: Scene text detection with richer fused features. In *IJCAI*, pages 516–522.

Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155.

Zhang, X., Chen, Q., Ng, R., and Koltun, V. (2019). Zoom to learn, learn to zoom. In *CVPR*, pages 3762–3770.