Rooms with Text: A Dataset for Overlaying Text Detection

Oleg Smirnov
Amazon
osmirnov@amazon.com

Aditya Tewari
Amazon
aditewar@amazon.com

Abstract

In this paper, we introduce a new dataset of room interior pictures with overlaying and scene text, totalling to 4836 annotated images in 25 product categories. We provide details on the collection and annotation process of our dataset, and analyze its statistics. Furthermore, we propose a baseline method for overlaying text detection, that leverages the character region-aware text detection framework [1] to guide the classification model. We validate our approach and show its efficiency in terms of binary classification metrics, reaching the final performance of 0.95 F1 score, with false positive and false negative rates of 0.02 and 0.06 correspondingly.

1. Introduction

Text detection and recognition task in images or videos typically do not distinguish between overlaying and scene text. Overlaying text is artificially superimposed on the image at the time of editing, and scene text is a text captured by the recording system. Distinguishing the overlaying text from the scene text is an important computer vision problem with multiple practical applications.

In video processing, locating and labeling superimposed captions is a pre-text stage for semantic video indexing, summarization, video surveillance and security, and multilingual video information access. In the e-commerce and social media domains, detecting overlaying text is paramount for identifying fraudulent and forbidden content, such as spam and unsolicited advertising in user generated imagery. Overlaying text recognition also finds important applications in forensic science [2].

The problem is particularly prominent in scenes with complex background, where one can expect large amounts of organic text. Those examples include but are not limited to urban street landscapes with outdoor signs, room interiors with wall decor, books in shelves, clocks, and similar objects.

Despite of recent progress in robust text in scene detection, state-of-the-art approaches are not applicable to the problem at hand, since conventional detectors typically specialize on one kind of text only. From the practical point of view, non-specialized models can not be deployed for fraudulent and forbidden content detection scenarios, which are highly sensitive to false negative rates.

In this paper, we present an annotated database that consists of real-world e-commerce images with overlaying and scene text in various combinations. We describe the collection and annotation process of our dataset, and analyze its statistics. Further, we propose a baseline method for overlaying text identification, which incorporates text localization to attend to relevant image regions for classification. We benchmark the performance of our method against an off-the-shelf logo and watermark detection algorithm to verify its effectiveness.

The rest of the paper is organized as follows. In Section 2 we review relevant work in the field. In Section 3 we introduce and analyze the Rooms with Text dataset consisting of room interior images annotated with text type labels. Further, in Section 4 we propose a simple, yet competitive baseline for distinguishing overlaying text from scene text in images. In Section 5 we discuss results of experiments and ablation studies. We conclude and outline future work in Section 6.

2. Related Work

Most of the literature in overlaying text detection is motivated by the task of recognizing video subtitles, also known as captions in video streaming.

Conventional methods operate under assumption that superimposed text regions have distinctive visual characteristic that can be identified by analyzing color [40] and texture information [17]. Edge-based approaches are also considered useful since text areas contain rich edge information. [31] proposed separating the instances based on the distribution of outputs produced by Canny and Sobel edge detectors. [27] trained a classifier on a linear combination of the histogram of oriented gradients features. [18] developed a multi-resolution algorithm that identifies regions of interest on a color histogram curve, followed by connected component analysis on candidate text areas.

Another stream of work studied regions of high contrast that can be characterized by saliency methods [22], as well as more subtle features like high frequencies determined from wavelet coefficients [9], and compression algorithm artefacts [2]. Among deep learning models, [32] proposed an end-to-end pipeline that combines TextBoxes [21] text detection algorithm with a convolutional recurrent neural network for extracting overlays from social media videos.

However, video subtitles recognition constitutes a relatively simple subset of a more generic overlaying text detection task. Video overlays are located in a designated position in the image, have strictly horizontal orientation, and often depicted on a contrasting background to optimize for readability. Moreover, above mentioned approaches are commonly evaluated on synthetic data. In absence of standardized benchmarks and labeled datasets, it is challenging to compare and evaluate their performance for real life applications.

Besides the direct methods, the task of overlaying text detection can be tackled by re-purposing approaches in logo and watermark detection, and generic scene text recognition.

Watermark and logo detection from images has been extensively studied in computer vision and pattern recognition literature. To this end, a watermark is understood as a semi-transparent logo or inscription that is added to the image at post-processing stages with a purpose of intellectual property protection, product brand management on social media, and others.

The task of watermark detection is similar to overlaying text detection in the sense of identifying image regions that do not organically belong to the scene. However, it is different in that image watermarks make use of semi-transparent alpha compositing, that enables usage of strong image priors [34] for detection and removal.

Identifying and removing a watermark from a single image without a-priori information is a daunting task. Common approaches frame it as a multi-image problem, that assumes that watermarks are added in a consistent manner to many images. Classical methods leverage gradient information [36, 38] and iterative subsequent matching [6] algorithm. More recently, [7, 30] proposed consistency-based approaches for discovering repeated patterns of low variability in image collections. SplitNet method [5] builds upon a multi-task network to predict a mask and coarser restored image for further refinement with a mask-guided spatial attention.

Contrasting to the watermark detection, logo detection rely neither on the semi-transparency nor on multi-image assumptions. Furthermore, logo detection formulation assumes availability of a dataset of well-known brand logotypes, making it a closed-vocabulary problem. Recent families of methods make use of end-to-end image recognition [14, 20] and object detection [12, 13] approaches correspondingly. Research in this field leverages publicly available logo databases [12, 28, 37] as well as synthesized datasets [29, 33].

Scene text recognition. The problem of accurate text recognition and localization was solved with the abundant availability of datasets such as MSRA-TD500 [39], ICDAR 2015 [15], COCO-Text [35], and Total-text [3]. Recent challenges [4] include text instances arranged in curved and other irregular shapes, that further contribute to robustness of detection algorithms in real world scenarios.

Unlike objects in general, texts are often presented in irregular shapes with various aspect ratios. To handle this problem, EAST [41] directly predicts geometry maps that combine rotated boxes with quadrangle coordinates. Another common approach is based on works dealing with segmentation, which aims to seek text regions at the pixel level. For example, Pixellink [8] detects texts by estimating word bounding areas. End-to-end approaches, e.g. FOTS [24], concatenates and trains the detection and recognition modules simultaneously so as to enhance detection accuracy by leveraging the recognition result. State-of-the-art character-level text detectors, such as CRAFT [1] (Character Region Awareness For Text detection), use text block candidates and produce probability maps to identify individual characters.

3. Dataset

In this paper, we propose Rooms-with-Text – a fully annotated dataset of room interior images with overlaying and scene text. Dataset construction comprises of three steps, namely candidate image selection, data labeling, and manual vetting.

Candidate image selection. The dataset contains of publicly available images collected from a large online retailer. Candidate images were identified by the means of the CRAFT [1] text detector. CRAFT exhibits an impressive performance in reliably detecting challenging and indistinct test regions. However, due to the objective of a character-level approach, the model tends to hallucinate in presence of letter-like shapes and patterns, which are common in indoor interior design. As a result, running CRAFT inference without subsequent filtering leads to a vast amount of false positives. To address this challenge, we devise a heuristic rule:

$$G(x) = \frac{4}{h \cdot w} \sum_{\substack{0 \le i < h/2 \\ 0 \le j < w/2}} F^{\text{rg}}_{i,j}(x)\{F^{\text{rg}}_{i,j}(x) > T\} \qquad (1)$$

where x is an input image of size $h \times w$, $F^{\text{rg}} \in R^{\frac{h}{2} \times \frac{w}{2}}$ is the output region scores matrix of a trained CRAFT model, and $T = 0.8$ is the region threshold constant proposed in

the paper. We consider images for annotation, if $G(x) > 5 \times 10^{-4}$

We provide sample pictures in Appendix B. Notable examples contain text in various languages and writing systems, including Chinese characters. Overlaying text is represented by logotypes, inscriptions, and watermarks with different levels of transparency. Whereas organic scene text can occur is wall posters, book titles, and other real objects in fonts of various size. The second from the left image in the first row exemplifies erroneous bounding boxes produced by the CRAFT algorithm.

Data labeling. Annotation labels were collected from 5 qualified workers per image, recruited using Amazon Mechanical Turk platform.[1] Workers' qualification was determined by the MTurk Masters designation as well as their performance on previous image labeling tasks.

The annotators were presented with a room image, and asked to categorize it into 4 non-overlapping classes: "Overlaying", "Organic", "Both", and "None". Each vote option was illustrated with 2-3 sample images, manually curated to be representative of the corresponding class.

We performed 3 rounds of annotation collection, starting with small size data batches and analyzing label consistency after each stage. We noticed that the workers often mislabel examples where scene text is hard to notice due to factors such as tiny font size, visual occlusion, and peripheral location. For the annotators' convenience, we additionally decorated query images with text bounding boxes, obtained by the CRAFT text detector.

Annotation vetting. After finishing the previous steps, each image was manually examined and reviewed to guarantee the quality of data after selection and annotation. If a labeled example did not meet the quality requirements, the image was rejected and re-annotated.

By inspecting the distribution of voting times for the pool of annotators in the initial batch, we observed that the 5^{th} percentile of voting time is 7 seconds, which we select as a threshold on the minimum time to consider a label valid. On subsequent stages, we discarded the annotations corresponding to the lowest 5% of voting time as possible outliers.

3.1. Dataset Statistics

The images in the database depict products from 25 Home and Furniture categories in lifestyle context. Figure 1b shows the distribution of the numbers of images per product category. Figure 1a depicts the distribution of the numbers of identified text regions on a logarithmic scale. The distribution roughly follows the Zipf's law.

On the final processing stage, resulting per-example labels were determined by majority voting across 5 annotations. The histogram of the number of assigned labels per example is shown in Figure 1a. We found that the overall agreement in votes is fairly consistent across the whole dataset. In 60.2% of cases all 5 annotators have agreed on the single label. Moreover, another 34.2% of examples have 3 or 4 votes for the same label. In total, we found 65 ambiguous examples with no clear voting majority, that constitutes 1.3% of the dataset. The labels for those examples were manually reviewed and corrected during the final cleaning stage.

The final cleaned dataset includes 4836 images, split into 3627 training and 1209 validation examples.

4. Method

We cast the problem of overlaying text detection as a binary image classification task. For this purpose, all images with only scene text or no text at all are considered as negative examples. Any image with a logo, watermark, caption, or an overlaid text of any other form is considered to belong to the positive class. This mapping corresponds to a practical scenario of detecting defective images corrupted with unwanted text inscriptions.

Our overlaying text detection pipeline comprises of two stages. The first component adapts the state-of-the-art text detector backbone for text localization. On this stage, the CRAFT model scores are used to generate a proposal mask. The second component utilizes the mask to attend to the regions of interest in the input image, for subsequent classification a convolutional neural network. Both components are trained together in an end-to-end manner.

The CRAFT model score is represented as a tensor $F^{\{\text{rg, af}\}}(x) \in R^{\frac{h}{2} \times \frac{w}{2} \times 2}$ with the region score rg and affinity score af channels:

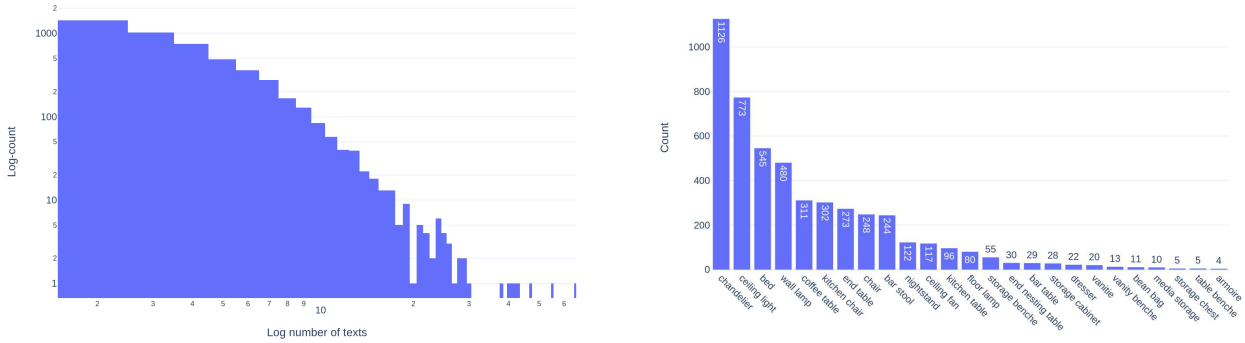
$$y = x \odot \underset{\times 2}{\text{UpSample}}(H(F^{\{\text{rg, af}\}}(x))). \tag{2}$$

The tensor $F^{\{\text{rg, af}\}}(x)$ is processed with a convolutional layer H with ReLU activation to produce a single channel $\frac{h}{2} \times \frac{w}{2} \times 1$ mask. The resulting mask is upsampled with bilinear interpolation by the factor of 2 to match the dimensions, and combined with the input $h \times w$ image x with element-wise multiplication. Finally, the masked image y is fed into a ResNet-18 [11] network for classification.

The network architecture is schematically illustrated in Figure 2.

We found that initializing weights of the convolutional layer H with a Gaussian kernel with a standard deviation of 8 pixels facilitates model convergence. Gaussian blurring creates smoother information about text neighborhoods, and also increases the size of the region where the text pixels interact with the rest of the image pixels.

[1] http://www.mturk.com



(a) Log-log histogram of the number of text regions per image

(b) Histogram of the number of image per product category

Figure 1. Rooms-with-Text dataset statistics

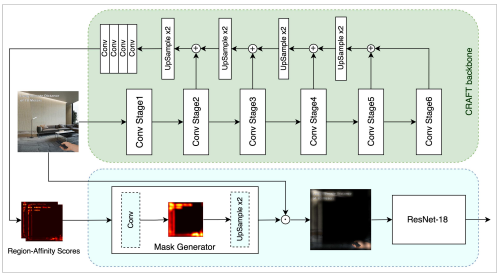


Figure 2. Diagram of the model architecture

5. Experiments

We compare the proposed CRAFT-masked ResNet approach with alternative methods. As a simple baseline, we adapt the reference text detector approach for binary classification task. To this end, the region and affinity score maps from CRAFT model are flattened and fed into a linear binary classifier algorithm. We refer to this method as "Binarized CRAFT". We also benchmark a recent watermark and logo detection method SplitNet [5] for overlaying text detection. Similarly to scene text detection, we build "Binarized SplitNet" and "SplitNet-masked" algorithms upon the watermark mask matrix, extracted from a SplitNet model. The former method uses the SplitNet mask directly as features for a binary classifier. Whereas the latter leverages SplitNet backbone instead of CRAFT in our method to attend to relevant image regions for classification with ResNet-18.

Finally, we perform ablation study to validate the impact of region attention masks by training a vanilla ResNet-18 model directly on unmasked dataset examples.

5.1. Results

We report overlaying text detection performance in binary classification metrics, including precision, recall, the area under the ROC-curve (AUC), F1 score, as well as true positive and false positive rates (TPR and FPR, correspondingly). Table 1 summarizes performance on the validation subset of Rooms-with-Text dataset among different detection methods.

Our method outperforms alternative approaches in all metrics. We observe that the watermark detection backbone pre-trained on synthetic data demonstrates an inferior performance compared to text detection. Interestingly, we find that a vanilla ResNet-18 network produces competitive results. However, its significantly higher FPR and FNR make it impractical for real-world use cases.

Method	AUC↑	Precision↑	Recall↑	F1 score↑	FPR↓	FNR↓
Binarized SplitNet	0.70	0.59	0.43	0.53	0.20	0.53
SplitNet-masked	0.76	0.62	0.57	0.59	0.21	0.43
Binarized CRAFT	0.84	0.72	0.63	0.67	0.15	0.37
CRAFT-masked	**0.99**	**0.97**	**0.94**	**0.95**	**0.02**	**0.06**
ResNet-18	0.93	0.82	0.80	0.81	0.11	0.20

Table 1. Overlaying text detection performance

To better understand the failure modes, we provide qualitative results of the proposed method in Appendix B.

6. Conclusions

In this work we proposed an annotated database of real-world interior images, designed for the task of distinguishing overlaying text from scene text.

Motivated by characteristics of our dataset, we introduce a simple baseline method for overlaying text detection. We validated and demonstrated the efficiency of the character region-aware text detection backbone for identifying relevant image areas for classification.

In future work, we plan to investigate recent advances in the field of image statistics prior for extending the process from overlaying text detection to image restoration and overlaying text removal.

References

[1] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019. 1, 2, 7

[2] Dinesh Bhardwaj and Vinod Pankajakshan. Image overlay text detection based on jpeg truncation error analysis. *IEEE Signal Processing Letters*, 23(8):1027–1031, 2016. 1, 2

[3] Chee Kheng Ch'ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 935–942. IEEE, 2017. 2

[4] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1571–1576. IEEE, 2019. 2

[5] Xiaodong Cun and Chi-Man Pun. Split then refine: stacked attention-guided resunets for blind single image visible watermark removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1184–1192, 2021. 2, 4, 6

[6] Maryam Dashti, Reza Safabakhsh, Mohammadreza Pourfard, and Mohammadjavad Abdollahifard. Video logo removal using iterative subsequent matching. In *2015 The International Symposium on Artificial Intelligence and Signal Processing (AISP)*, pages 84–88. IEEE, 2015. 2

[7] Tali Dekel, Michael Rubinstein, Ce Liu, and William T Freeman. On the effectiveness of visible watermarks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2146–2154, 2017. 2

[8] Dan Deng, Haifeng Liu, Xuelong Li, and Deng Cai. Pixellink: Detecting scene text via instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[9] Julinda Gllavata, Ralph Ewerth, and Bernd Freisleben. Text detection in images based on unsupervised classification of high-frequency wavelet coefficients. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 1, pages 425–428. IEEE, 2004. 2

[10] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2315–2324, 2016. 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[12] Steven CH Hoi, Xiongwei Wu, Hantang Liu, Yue Wu, Huiqiong Wang, Hui Xue, and Qiang Wu. Logo-net: Large-scale deep logo detection and brand recognition with deep region-based convolutional networks. *arXiv preprint arXiv:1511.02462*, 2015. 2

[13] Rahul Kumar Jain, Taro Watasue, Tomohiro Nakagawa, SATO Takahiro, Yutaro Iwamoto, RUAN Xiang, and CHEN Yen-Wei. Logonet: Layer-aggregated attention centernet for logo detection. In *2021 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–6. IEEE, 2021. 2

[14] Alexis Joly and Olivier Buisson. Logo retrieval with a contrario visual query expansion. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 581–584, 2009. 2

[15] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 2

[16] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1484–1493. IEEE, 2013. 6

[17] Wonjun Kim and Changick Kim. A new approach for overlay text detection and extraction from complex video scene. *IEEE transactions on image processing*, 18(2):401–411, 2008. 1

[18] Lalita Kumari, Vidyut Dey, and JL Raheja. A three-layer approach for overlay text extraction in video stream. In *Soft Computing: Theories and Applications*, pages 79–87. Springer, 2018. 1

[19] Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-Tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A system for massively parallel hyperparameter tuning. *Proceedings of Machine Learning and Systems*, 2:230–246, 2020. 6

[20] Zhe Li, Matthias Schulte-Austum, and Martin Neschen. Fast logo detection and recognition in document images. In *2010 20th International Conference on Pattern Recognition*, pages 2716–2719. IEEE, 2010. 2

[21] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 2

[22] Rainer Lienhart and Axel Wernicke. Localizing and segmenting text in images and videos. *IEEE Transactions on circuits and systems for video technology*, 12(4):256–268, 2002. 2

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[24] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5676–5685, 2018. 2

[25] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1454–1459. IEEE, 2017. 6

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[27] Bernhard Quehl, Haojin Yang, and Harald Sack. Improving text recognition by distinguishing scene and overlay text. In *Seventh International Conference on Machine Vision (ICMV 2014)*, volume 9445, page 944509. International Society for Optics and Photonics, 2015. 1

[28] Stefan Romberg, Lluis Garcia Pueyo, Rainer Lienhart, and Roelof Van Zwol. Scalable logo recognition in real-world images. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, pages 1–8, 2011. 2

[29] Alexander Sage, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Logo synthesis and manipulation with clustered generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5879–5888, 2018. 2

[30] Xi Shen, Ilaria Pastrolin, Oumayma Bounou, Spyros Gidaris, Marc Smith, Olivier Poncet, and Mathieu Aubry. Large-scale historical watermark recognition: dataset and a new consistency-based approach. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6810–6817. IEEE, 2021. 2

[31] P Shivakumara, N Vinay Kumar, DS Guru, and Chew Lim Tan. Separation of graphics (superimposed) and scene text in video frames. In *2014 11th IAPR International Workshop on Document Analysis Systems*, pages 344–348. IEEE, 2014. 1

[32] Adam Słucki, Tomasz Trzciński, Adam Bielski, and Paweł Cyrta. Extracting textual overlays from social media videos using neural networks. In *International Conference on Computer Vision and Graphics*, pages 287–299. Springer, 2018. 2

[33] Hang Su, Xiatian Zhu, and Shaogang Gong. Deep learning logo detection with data expansion by synthesising context. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 530–539. IEEE Computer Society, 2017. 2

[34] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018. 2

[35] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 2

[36] Jinqiao Wang, Qingshan Liu, Lingyu Duan, Hanqing Lu, and Changsheng Xu. Automatic tv logo detection, tracking and removal in broadcast video. In *International Conference on Multimedia Modeling*, pages 63–72. Springer, 2007. 2

[37] Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, and Shuqiang Jiang. Logodet-3k: A large-scale image dataset for logo detection. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1):1–19, 2022. 2

[38] Wei-Qi Yan, Jun Wang, and Mohan S Kankanhalli. Automatic video logo detection and removal. *Multimedia Systems*, 10(5):379–391, 2005. 2

[39] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1083–1090. IEEE, 2012. 2

[40] DongQing Zhang, Belle L Tseng, and S-F Chang. Accurate overlay text extraction for digital video analysis. In *International Conference on Information Technology: Research and Education, 2003. Proceedings. ITRE2003.*, pages 233–237. IEEE, 2003. 1

[41] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017. 2

A. Appendix: Experimental Setup

We use PyTorch [26] framework to implement our models. We perform training and evaluation on Rooms-with-Text dataset. Examples with labels "Overlaying" and "Both" were mapped to the positive class, whereas examples labeled as "Scene" and "None" were considered as the negative class. The input images were resized and padded to the dimensions as required by corresponding models with aspect ratio preservation.

We use a mini-batch size of 32, and stochastic gradient descent optimizer with an initial learning rate of 0.015 and momentum 0.9. The learning rate was annealed by the factor of 0.5 after each epoch, when a plateau was encountered. The optimal combination of hyperparameters was found by tuning on the validation set with the asynchronous Hyperband [19] algorithm.

All networks were trained until convergence with the binary cross-entropy objective. To combat overfitting, a weight decay penalty of 1×10^{-5} was added to all model parameters. We have not used any additional data augmentation techniques during model training.

Following the training protocol in [5], SplitNet backbone was pre-trained on synthesized datasets derived from background images from MSCOCO [23]. The CRAFT backbone was pre-trained on SynthText [10], and then fine-tuned on ICDAR2013 [16] and ICDAR2017 [25].

B. Appendix: Dataset Examples

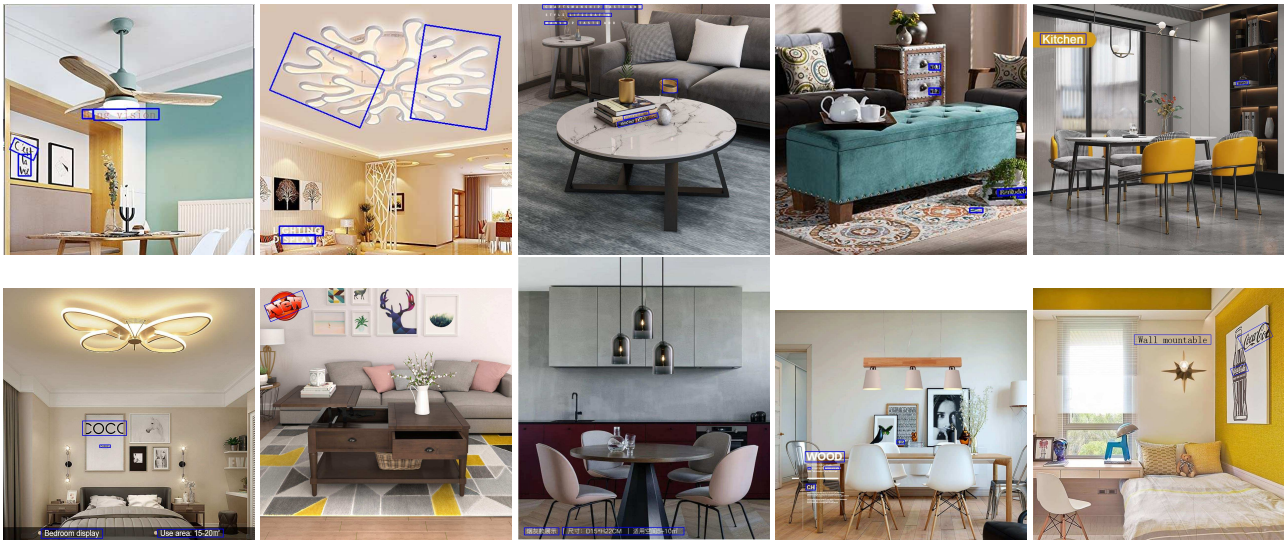


Figure 3. Examples of real-world product images with scene and overlaying text. Bounding boxes are generated with the CRAFT [1] text detector for illustration purposes. Best viewed on screen

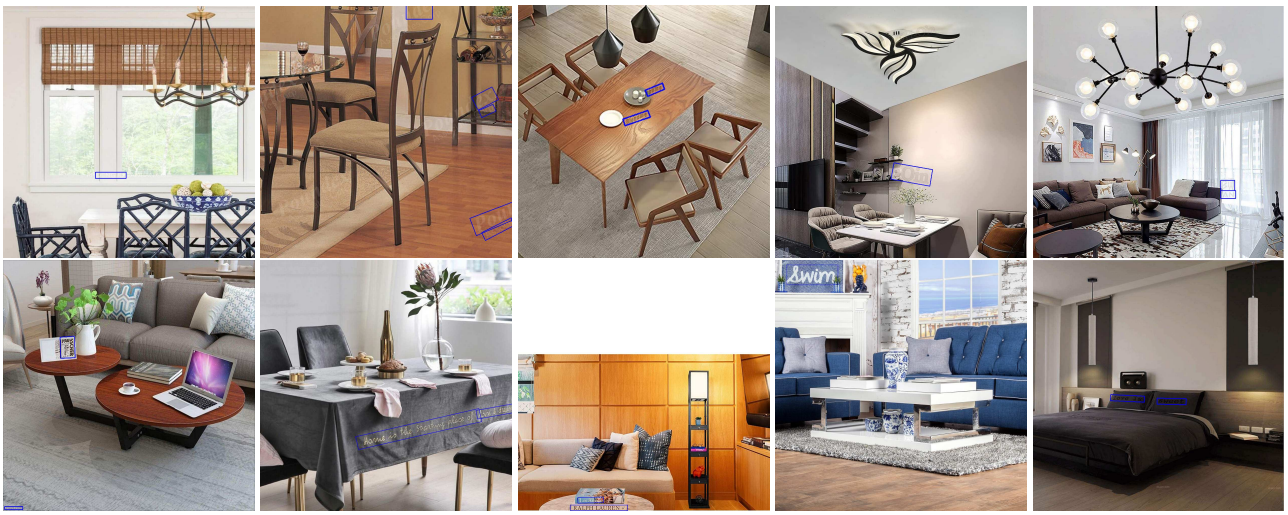


Figure 4. Qualitative results of the proposed method. The top row and the bottom row depict false negative and false positive predictions correspondingly, along with bounding boxes generated by the CRAFT [1] text detector. Best viewed on screen