Efficient and Effective Scene Text Synthesis Supplementary Materials

Fangneng Zhan

1 Overview

Realistic image composition by inserting foreground objects into a background image has been studied extensively. The target is to achieve composition realism by adjusting object geometry and object appearance automatically and adaptively with respect to the surrounding background. With the advances of deep neural network research, image synthesis has been investigated as an image augmentation approach when only a limited amount of annotated images is available. For example, [7], [11], [45] study synthesis of scene text images for training better scene text detection and recognition models, [5] uses synthetic chair images for training better optical flow networks. In recent years, a number of GAN-based techniques have been reported which synthesize new images by generation from random noises [3], [25], [51], appearance transfer [20], [46], [47], [55] or image composition [48], [49], [50].

Our proposed technique adopts the image composition approach for the synthesis of verisimilar scene text images. Beyond training GANs for realistic text appearance, it is capable of identifying suitable text embedding locations within the background image according to the semantic coherence. In addition, it exploits local contexts and is capable of aligning the foreground texts with the contextual structures within the background image realistically.

Automatic detection and recognition of various texts in scene images has attracted increasing research interests in recent years due to many relevant applications in practice [15], [30]. Different detection techniques have been proposed from those earlier using hand-crafted features [22], [24], [31] to the recent using DNNs [13], [36], [38], [43], [53]. Different detection approaches have also been explored including character based detection [9], [10], [14], [35], word-based detection [8], [13], [19], [21], [54] and the recent line-based detection [41], [52]. Meanwhile, different scene text recognition techniques have been developed from the earlier recognizing characters directly [1], [6], [11], [12], [27], [42] to the recent recognizing words or text lines using recurrent neural network (RNN), [28], [29], [33], [34], [44] as well as various attention models [4], [18].

1.1 Datasets

The proposed image synthesis technique has been evaluated over the following public datasets:

ICDAR 2013 [16] dataset contains 229 images for training and 233 images for testing with word-level annotations. For recognition task, there are 848 word images for training recognition models and 1095 word images for evaluation.

ICDAR 2015 [15] is a dataset that consists of 1,670 images (17,548 annotated incidental scene text regions) acquired using the Google Glass. Incidental scene text refers to text that appears in the scene without the user taking any action to rectify the position and quality of the text regions.

MSRA-TD500 [40] dataset consists 300 natural images for training, 200 images for testing with diverse visual contents and resolutions, which are taken from cluttered indoor and outdoor scenes using a pocket camera.

IIIT5K [23] dataset consists of 2000 training images and 3000 test images with cropped scene texts and born-digital. For each image, there are two lexicons: one with 50-word and the other with 1000-word.

SVT [37] dataset consists of 249 street view images from which 647 words images are cropped. Each word image has a 50-word lexicon.

CUTE [26] has 288 curved word images cropped from the CUTE dataset that are originally collected for scene text detection research.

TABLE 1
Scene text detection performance on the ICDAR2013 dataset by using the EAST model as described in Section 4.2.1, where "Synth", "Gupta" and "Zhan" denote the training images that were synthesized by our proposed method, Gupta's [7] and Zhan's [45], respectively. "1K" denotes the number of synthesized images used, "Random" denotes embedding texts with random locations and colors, "RD" and "TE" denote the proposed region detection and text embedding techniques.

Training Data	R	P	F
1K Synth (Random)	69.13	64.82	66.91
1K Synth (RD)	71.64	66.46	68.95
1K Synth (TE)	69.76	65.71	67.67
1K Synth (RD+TE)	**72.32**	**67.57**	**69.86**
1K Gupta [7]	68.68	67.50	68.09
1K Zhan [45]	70.96	66.87	68.85

1.2 Scene Text Detection

1.2.1 Implementation

For the scene text detection task, we adopt an adapted EAST model [54] to train all text detectors. EAST is a fully convolutional network (FCN) which can directly localize texts of arbitrary orientations at word or text-line level. It allows

TABLE 2
Scene text recognition performance over the datasets ICDAR2013, ICDAR2015, IIIT5K, SVT and CUTE, where "50" and "1K" in the second row denote the lexicon size and "None" means no lexicon used. ASTR denotes the recognition model as described in Section 4.3.1.

Methods	ICDAR2013	ICDAR2015	IIIT5K			SVT		CUTE
	None	None	50	1k	None	50	None	None
Yao [42] [-]	-	-	80.2	69.3	-	75.9	-	-
Almazan [1] [-]	-	-	91.2	82.1	-	89.2	-	-
Gordo [6] [-]	-	-	93.3	86.6	-	91.8	-	-
Jaderberg [12] [Jaderberg 8M]	81.8	-	95.5	89.6	-	93.2	71.7	-
Jaderberg [13] [Jaderberg 8M]	**90.8**	-	97.1	92.7	-	95.4	80.7	-
Shi [28] [Jaderberg 8M]	89.6	-	97.8	95.0	81.2	**97.5**	82.7	-
Shi [29] [Jaderberg 8M]	88.6	-	96.2	93.8	81.9	95.5	81.9	59.2
Yang [39] [Private]	-	-	97.8	96.1	-	95.2	-	**69.3**
Lee [18] [Jaderberg 8M]	90.0	-	96.8	94.4	78.4	96.3	80.7	-
ASTR [Jaderberg 5M] [14]	86.6	64.1	96.8	93.2	81.0	96.1	81.5	55.8
ASTR [Gupta 5M] [7]	87.0	66.6	97.6	94.8	81.3	95.2	80.1	55.9
ASTR [Zhan 5M] [45]	87.7	67.4	97.9	95.4	82.1	96.9	82.2	56.8
ASTR [Ours 5M]	89.4	**68.1**	**98.7**	**96.3**	**84.1**	97.2	**82.9**	58.6

end-to-end training and optimization without unnecessary intermediate components and steps, and achieves superior detection accuracy and efficiency as compared with state-of-the-art methods. It uses the PVANET [17] as the backbone in its original implementation. We instead use the ResNet-152 in our implementation as PVANET is not publicly available.

The proposed image synthesis technique is evaluated by using the dataset ICDAR2013 that has been widely used in scene text detection study for years. Two experiments were performed to demonstrate the effectiveness of our synthesized images in training deep detection networks. In the first experiment, we employ 400K images synthesized by our proposed method to train an EAST model and compare the trained model with the state-of-the-art as shown in Table 2. The purpose is to show that our synthesized images can perform similarly or even better than real images while applied for training deep detection models. In the second experiment, we carry out an ablation study that uses 1K synthesized images to evaluate the performance of our proposed region detection and text embedding techniques. This experiment also compares our image synthesis technique with two state-of-the-art image synthesis techniques as shown in Table 3. Note we use 5000 images without containing any scene texts as the background images and select the foreground texts from publicly available corpses. The number of embedded words or text lines is limited to a maximum of 15 for each background image.

1.2.2 Result Analysis

Table 2 shows experimental results when 400K synthesized images are used to train the adapted EAST model. As Table 2 shows, the trained EAST model achieves similar performance as compared with state-of-the-art models that use either real images or synthesized images in training. Though a much larger amount of synthesized images is used as compared with those using real images, the proposed image synthesis technique generates new images by machines which removes the complicated image collection and selection process as required by real images. More importantly, the proposed image synthesis technique produces object annotations automatically which removes the time-consuming object annotation process. The use of a larger amount of synthesized images will introduce a longer training time but this is far more manageable as compared with manual collection, selection, and annotation of a large amount of real images. Note Gupta, et al also used their synthesized images in training, but they used 800K synthesized images in training and the achieved f-score is clearly lower than ours using 400K synthesized images.

Table 3 shows ablation study results as well as comparisons with other image synthesis techniques. For fair comparisons, the adapted EAST and 1K synthesized images (by different synthesis methods) are used in detection model training consistently. As Table 3 shows, random placement of source texts into background images (no control of embedding locations and text appearance) is capable of producing useful training images (with a detection f-score of 66.91%). The including of either our proposed region detection (RD) or text embedding (TE) clearly improves the detection f-score by up to 2%, and the including of both further improves the detection f-score by up to 3%. In addition, we can see that our synthesized images perform clearly better than the synthesized images in [7] and [45], with an up to 2% f-score improvement. The better performance is largely due to the semantic coherence, geometry alignment, and realistic appearance within our proposed image synthesis technique as illustrated in Fig. 5.

1.3 Scene Text Recognition

1.3.1 implementation

For the scene text recognition task, we adopt an attention-based scene text recognition model (ASTR) which is a sequence-to-sequence learning method [2], [32]. The ASTR consists of an encoder and a decoder, where the encoder extracts a sequential representation from the input image and the decoder recurrently generates a sequence conditioned on the sequential representation. The text recognition model covers 68 characters including 10 digits, 26 lowercase letters and 32 ASCII punctuation marks. In evaluations, only digits and letters are counted and the rest is directly discarded. If a lexicon is provided, the lexicon word that has the minimum edit distance with the predicted word is selected. In addition, evaluations are based on the correctly recognized words (CRW) which can be determined based on the ground truth transcription.

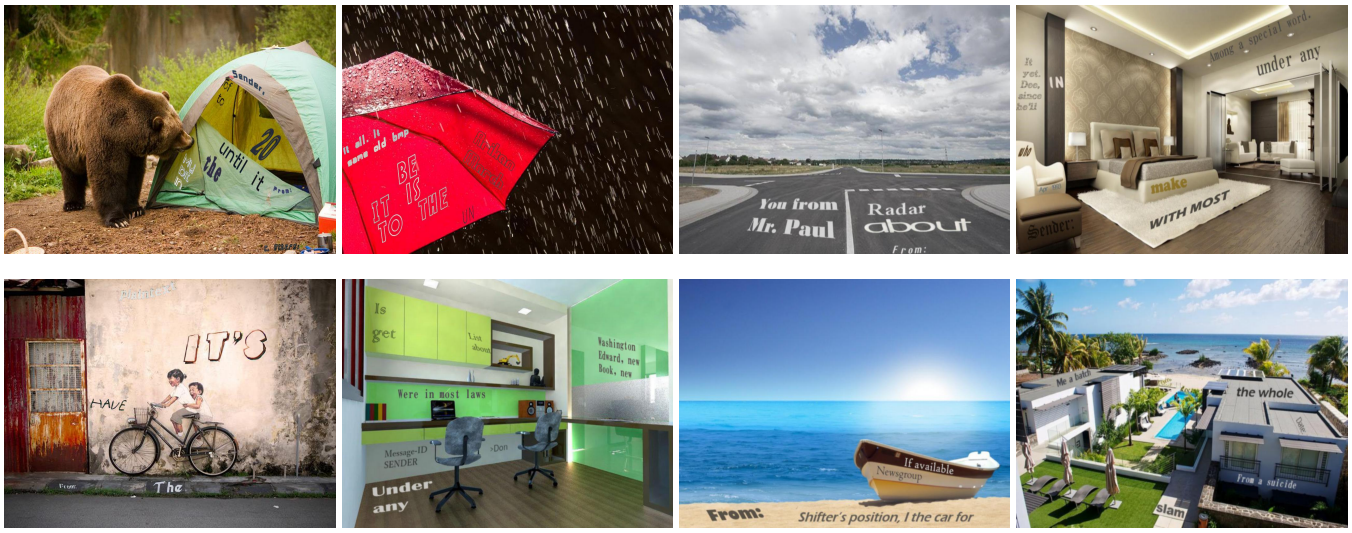

Fig. 1. A number of sample images by our proposed synthesis technique that show how the proposed semantic coherence, geometry alignment and realistic text appearance work together for automatic and verisimilar scene text synthesis.

The proposed image synthesis technique is evaluated over five public datasets including ICDAR2013, ICDAR2015, IIIT5K, SVT and CUTE as shown in Table 4. Two sets of benchmarking evaluations were performed. The first compares our proposed image synthesis technique with a list of state-of-the-art scene text recognition techniques that use different amounts of synthesized images (most used a much larger amount as shown in Table 4) as well as different recognition models as proposed in the respective research. The purpose is to demonstrate how our synthesized images perform as compared with state-of-the-art recognition methods. The second compares our proposed image synthesis technique with a number of state-of-the-art image synthesis techniques, where the same amount of synthesized image (5M) and the same recognition model ASTR is used for all model training consistently. It provides a more direct comparison by using the same recognition model and the same amount of synthesized images.

1.3.2 Result Analysis

Table 4 shows the scene text recognition results. As Table 4 shows, our ASTR model (ASTR [Ours 5M]) achieves state-of-the-art scene text recognition accuracy when 5M images synthesized by our proposed method are used in training. Though the accuracy by the ASTR [Ours 5M] is not always the highest, it performs better than other state-of-the-art techniques for most evaluated datasets under different scenarios with or without using lexicons. The slighter lower accuracy for some dataset such as ICDAR2013 and SVT (using 50 lexicon) is largely due to a larger amount of training images, e.g. 8M in [28] or a constrained-output recognizer [13]. In addition, the ASTR was common model for text recognition without specific design whereas the state-of-the-art recognition models usually used the latest networks with proposed tricks. It should be noted that the dataset CUTE contains a large amount of curved texts that cannot be recognized properly by most state-of-the-art methods (whereas the methods in [29], [39] were specially designed to recognized curved/irregular texts).

Table 4 also shows the comparison between our proposed image synthesis methods and three state-of-the-art image synthesis methods as reported in [7], [14], [45]. For fair comparison, the same amounts of synthesized images (5 million) were taken from our proposed method and the three state-of-the-art methods and the same ASTR model is used consistently. The trained recognition models are labelled by "ASTR [Ours 5M]", "ASTR [Jaderberg 5M]", "ASTR [Gupta 5M]" and "ASTR [Zhan 5M]", respectively. As Table 4 shows, the ASTR trained by using our synthesized images outperforms the ASTR trained by using the "Jaderberg 5M", "Gupta 5M" and "Zhan 5M" consistently across all 5 evaluated datasets. The outstanding recognition performance of our "ASTR [Ours 5M]" is largely due to the semantic coherence, geometry alignment and realistic text appearance in our proposed image synthesis method.

REFERENCES

[1] J. Almazan, A. Gordo, A. Fornes, and E. Valveny. Word spotting and recognition with embedded attributes. *In TPAMI*, (12):2552–2566, 2014.

[2] O. Alsharif and J. Pineau. End-to-end text recognition with hybrid hmm maxout models. *In ICLR*, 2014.

[3] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *In NIPS*, 2016.

[4] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou. Focusing attention: Towards accurate text recognition in natural images. *In ICCV*, pages 5076–5084, 2017.

[5] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. *Proc. ICCV*, 2015.

[6] A. Gordo. Supervised mid-level features for word image representation. *In CVPR*, 2015.

[7] A. Gupta., A. Vedaldi., and A. Zisserman. Synthetic data for text localisation in natural images. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[8] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li. Single shot text detector with regional attention. *arXiv:1709.00138*, 2017.

[9] T. He, W. Huang, Y. Qiao, and J. Yao. Text-attentional convolutional neural network for scene text detection. *In TIP*, (6):2529–2541, 2016.

[10] W. Huang, Y. Qiao, and X. Tang. Robust scene text detection with convolution neural network induced mser trees. *In ECCV*, pages 497–511, 2014.

[11] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014.

[12] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Deep structured output learning for unconstrained text recognition. *In ICLR*, 2015.

[13] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. *In IJCV*, (1):1–20, 2016.

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

[15] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, and F. Shafait. Icdar 2015 competition on robust reading. *In ICDAR*, pages 1156–1160, 2015.

[16] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, L. P. de las Heras, and et al. Icdar 2013 robust reading competition. *In Proc. ICDAR*, pages 1484–1493, 2013.

[17] K. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park. Pvanet: Deep but lightweight neural networks for real-time object detection. *arXiv:1608.08021*, 2016.

[18] C.-Y. Lee and S. Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. *In CVPR*, pages 2231–2239, 2016.

[19] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu. Textboxes: A fast text detector with a single deep neural network. *In AAAI*, pages 4161–4167, 2017.

[20] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. *In NIPS*, 2017.

[21] Y. Liu and L. Jin. Deep matching prior network: Toward tighter multi-oriented text detection. *In CVPR*, 2017.

[22] S. Lu, T. Chen, S. Tian, J.-H. Lim, and C.-L. Tan. Scene text extraction based on edges and support vector regression. *In IJDAR*, (2):125–135, 2015.

[23] A. Mishra, K. Alahari, and C. Jawahar. Scene text recognition using higher order language priors. *In BMVC*, 2012.

[24] L. Neumann and J. Matas. Real-time scene text localization and recognition. *In CVPR*, pages 3538–3545, 2012.

[25] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015.

[26] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan. A robust arbitrary text detection system for natural scene images. *In Expert Syst. Appl.*, 41(18):8027–8048, 2014.

[27] J. A. Rodrguez-Serrano, A. Gordo, and F. Perronnin. Label embedding: A frugal baseline for text recognition. *In IJCV*, 2015.

[28] B. Shi, X. Bai, and C. Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *In TPAMI*, 39(11):2298–2304, 2017.

[29] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai. Robust scene text recognition with automatic rectification. *In CVPR*, 2016.

[30] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, and X. Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). *In ICDAR*, 01:1429–1434, 2017.

[31] P. Shivakumara, R. P. Sreedhar, T. Q. Phan, S. Lu, and C. L. Tan. Multioriented video scene text detection through bayesian classification and boundary growing. *IEEE Transactions on Circuits and systems for Video Technology*, (8):1227–1235, 2012.

[32] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034*, 2013.

[33] B. Su and S. Lu. Accurate scene text recognition based on recurrent neural network. *In ACCV*, 2014.

[34] B. Su and S. Lu. Accurate recognition of words in scenes without character segmentation using recurrent neural network. *In PR*, 2017.

[35] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. L. Tan. Text flow: A unified text detection system in natural scene images. *In ICCV*, pages 4651–4659, 2015.

[36] Z. Tian, W. Huang, P. H. T. He, and Y. Qiao. Detecting text in natural image with connectionist text proposal network. *In ECCV*, pages 56–72, 2016.

[37] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. *In ICCV*, 2011.

[38] C. Xue, S. Lu, and F. Zhan. Accurate scene text detection through border semantics awareness and bootstrapping. In *Proceedings of the European conference on computer vision (ECCV)*, pages 355–372, 2018.

[39] X. Yang, D. He, Z. Zhou, D. Kifer, and C. L. Giles. Learning to read irregular text with attention mechanisms. *In IJCAI*, pages 3280–3286, 2017.

[40] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu. Detecting texts of arbitrary orientations in natural images. *In CVPR*, pages 1083–1090, 2012.

[41] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou, and Z. Cao. Scene text detection via holistic, multi-channel prediction. *arXiv:1606.09002*, 2016.

[42] C. Yao, X. Bai, B. Shi, and W. Liu. Strokelets: A learned multi-scale representation for scene text recognition. *In CVPR*, 2014.

[43] X. C. Yin, W. Y. Pei, J. Zhang, and H. W. Hao. Multiorientation scene text detection with adaptive clustering. *In TPAMI*, (9):1930–1937, 2015.

[44] F. Zhan and S. Lu. Esir: End-to-end scene text recognition via iterative image rectification. In *CVPR*, pages 2059–2068, 2019.

[45] F. Zhan, S. Lu, and C. Xue. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–266, 2018.

[46] F. Zhan, C. Xue, and S. Lu. Ga-dan: Geometry-aware domain adaptation network for scene text detection and recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9105–9115, 2019.

[47] F. Zhan, Y. Yu, K. Cui, G. Zhang, S. Lu, J. Pan, C. Zhang, F. Ma, X. Xie, and C. Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15028–15038, 2021.

[48] F. Zhan, C. Zhang, W. Hu, S. Lu, F. Ma, X. Xie, and L. Shao. Sparse needlets for lighting estimation with spherical transport loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12830–12839, 2021.

[49] F. Zhan, C. Zhang, Y. Yu, Y. Chang, S. Lu, F. Ma, and X. Xie. Emlight: Lighting estimation via spherical distribution approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[50] F. Zhan, H. Zhu, and S. Lu. Spatial fusion gan for image synthesis. In *CVPR*, pages 3653–3662, 2019.

[51] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *In ICCV*, 2017.

[52] Z. Zhang, W. Shen, C. Yao, and X. Bai. Symmetry-based text line detection in natural scenes. *In CVPR*, pages 2558–2567, 2015.

[53] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai. Multi-oriented text detection with fully convolutional networks. *In CVPR*, pages 4159–4167, 2016.

[54] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. East: An efficient and accurate scene text detector. *In CVPR*, 2017.

[55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *In ICCV*, 2017.