DUET: Detection Utilizing Enhancement for Text in Scanned or Captured Documents

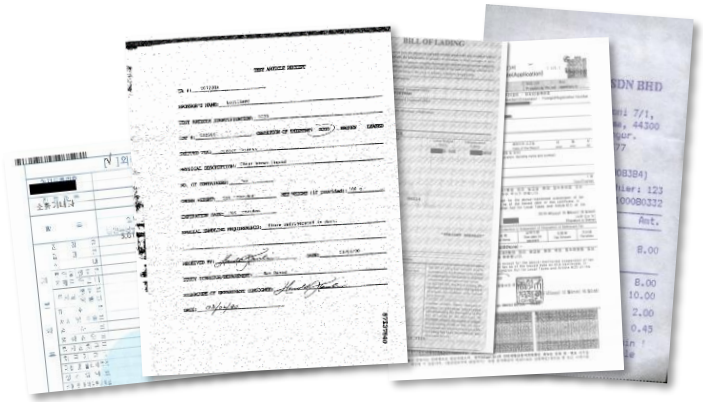
Eun-Soo Jung*, HyeongGwan Son*, Kyusam Oh,
Yongkeun Yun, Soonhwan Kwon, and Min Soo Kim

** Equal contributions*

SAMSUNG SDS

Background and Motivation

- Most of the previous studies on text detection focus on text in the wild[1-6]
- Text in the wild (or scene text)
  - (Relatively) More labeled data[7-11]
- Text in documents
  - Very few labeled data[12]
    → insufficient to train deep neural model
- Text detection models trained with scene text
  : *Limited* to cover features of document images

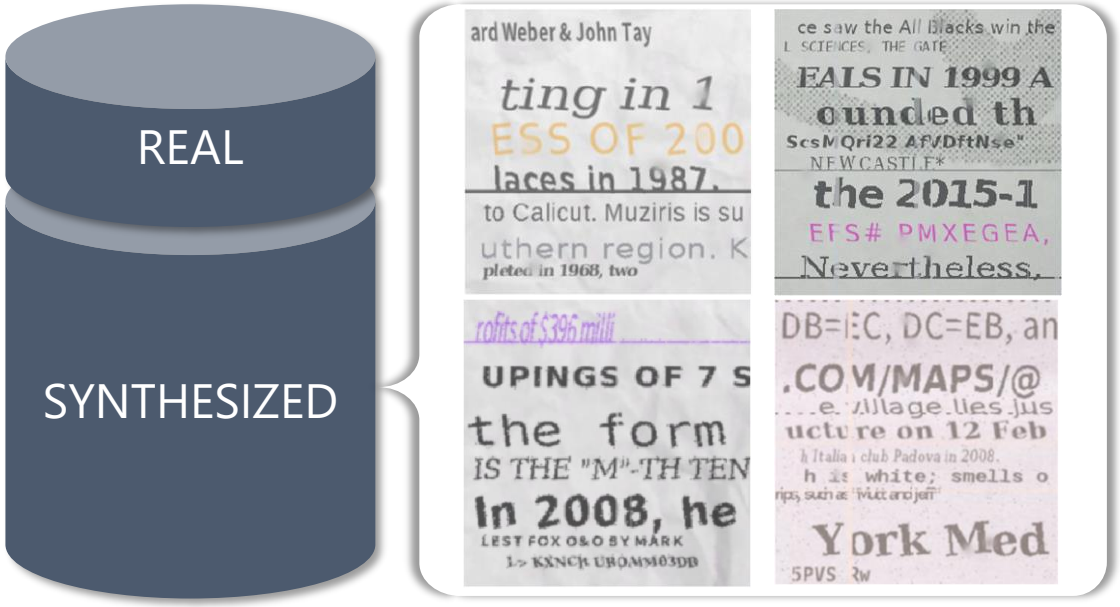


Goal: Deep neural text detector with improved accuracies for document images

Contributions

Training data

- Data synthesizing
- *Overcome* the shortage of labeled document image data



Learning strategy

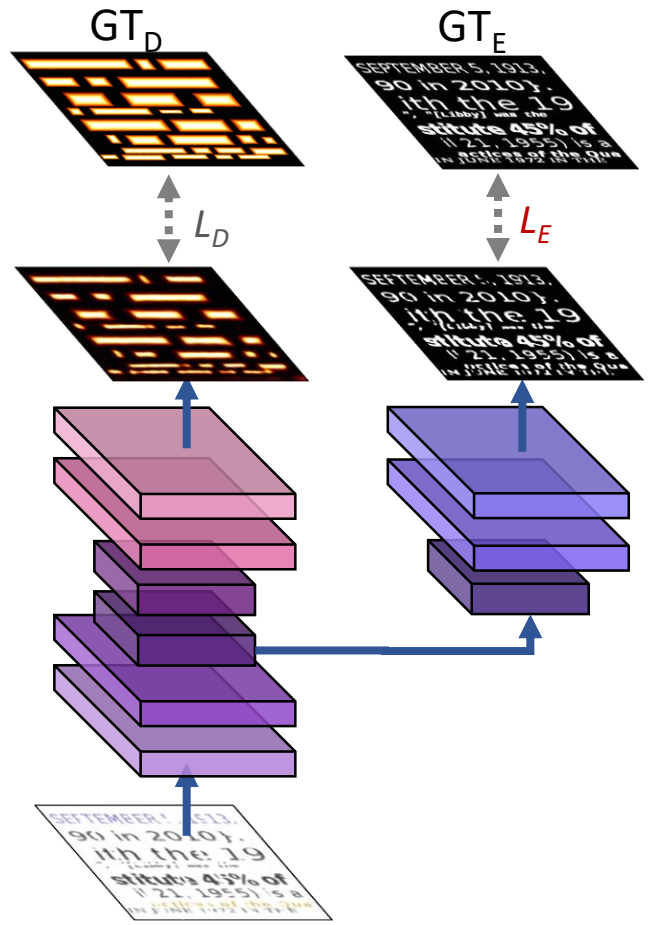
- Multi-task learning[13-16]
- Weakly-supervised learning[17-18]
- *Overcome* various types of noise in document images

Network Architecture

- Fully convolutional encoder-decoder structure
  - Feature Pyramid Network (FPN)[19]
  - ResNeXt101[20]

- Multi-task learning
  - Branch for text detection
    → detection loss (L_D)
  - Branch for text enhancement
    : pixel-level binary classification (text/non-text)
    → enhancement loss (L_E)
  - Multi-task loss
    → $L = \lambda_1 L_D + (1 - \lambda_1) L_E$

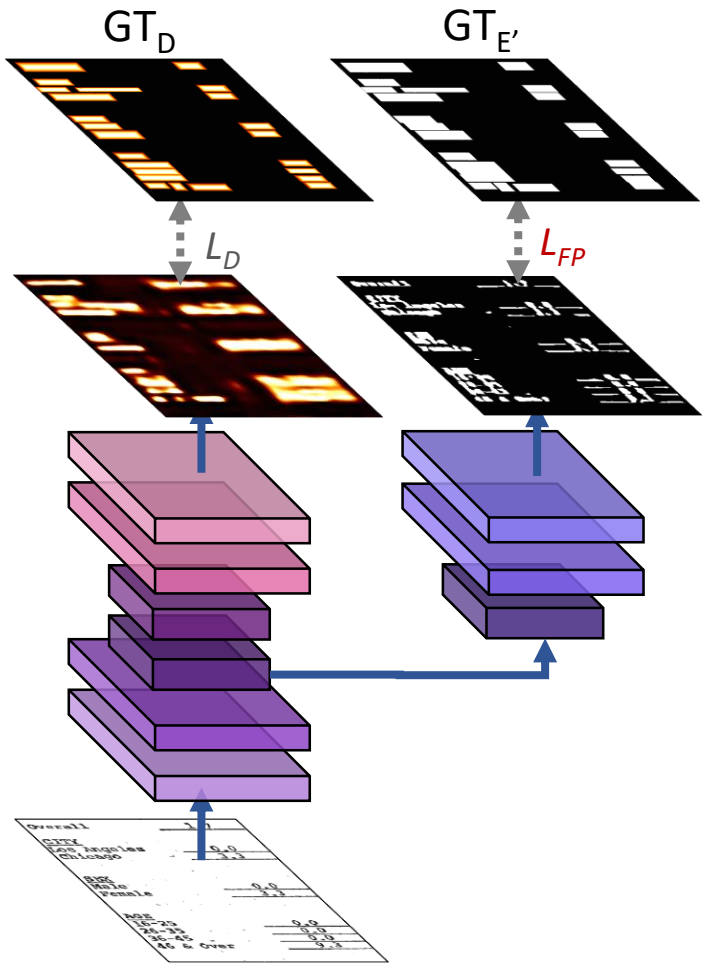
(λ_1: balancing parameter between L_D and L_E)



GT: ground truth

Training

- Phase 1
  - Training **synthesized** data
  - Fully-supervised learning
- Phase 2
  - Training **synthetic** data and **real** data
  - Detection GTs: given
  - Enhancement GTs: ??
    → weakly-supervised learning
    - Binaraized GT_D ($GT_{E'}$)
      → false positive loss (L_{FP})

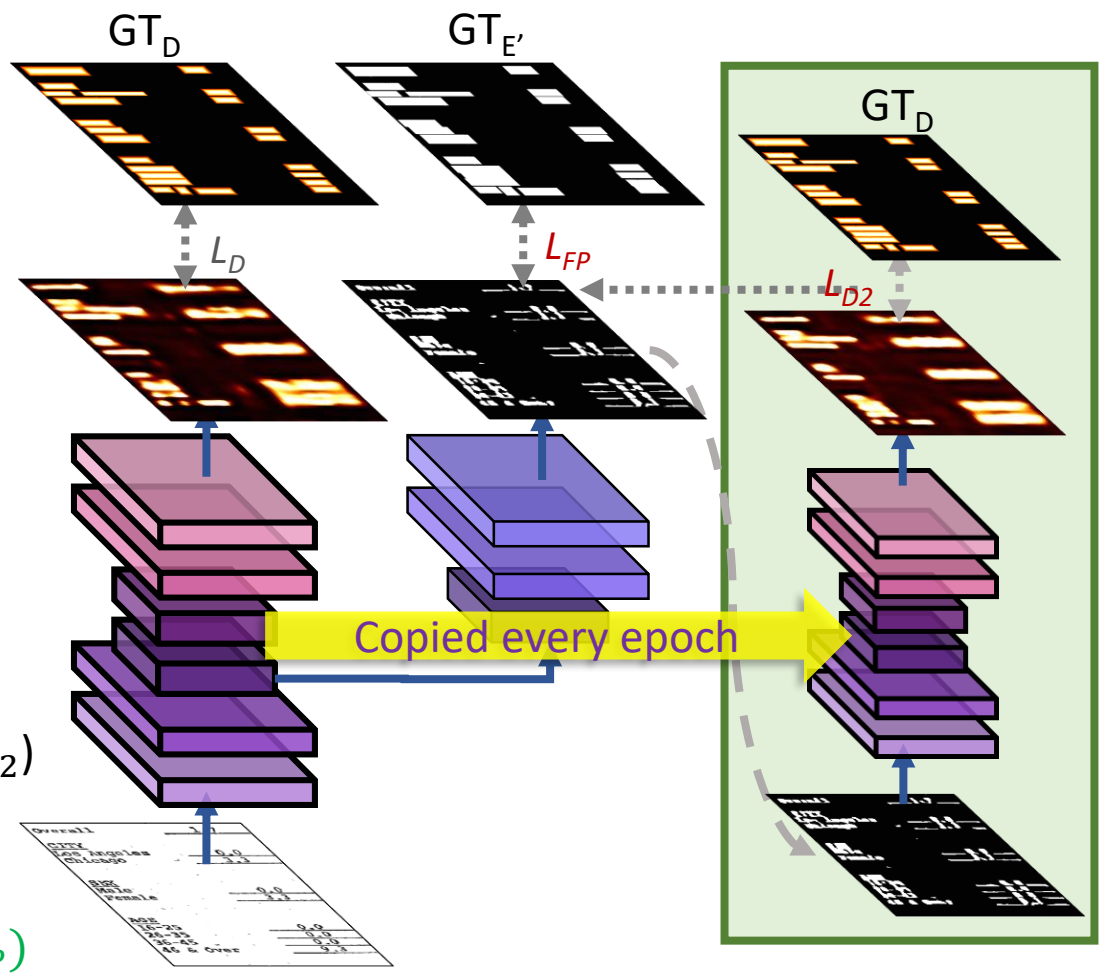


Training

- Phase 1
  - Training **synthesized** data
  - Fully-supervised learning
- Phase 2
  - Training **synthetic** data and **real** data
  - Detection GTs: given
  - Enhancement GTs: ??
    → weakly-supervised learning
    - Binaraized GT_D ($GT_{E'}$)
      → false positive loss (L_{FP})
    - Using interim trained detector
      → detection loss for enhanced output (L_{D2})
  - Multi-task loss
    → $L = \lambda_1 L_D + (1 - \lambda_1) L_E$

    $L = \lambda_1 L_D + (1 - \lambda_1)(\lambda_2 L_{D2} + (1 - \lambda_2) L_{FP})$

    (λ_2: balancing parameter between L_{D2} and L_{FP})



Experiment and Results

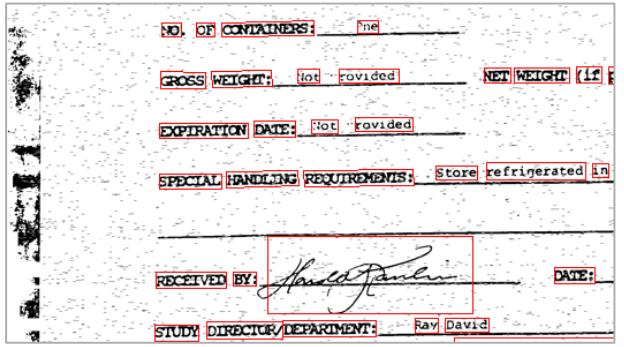
- Benchmark database
  : Form Understanding in Noisy Scanned Documents (**FUNSD**) dataset[12]
  (120 train, 29 validation, 50 test)

- Comparisons with previous studies

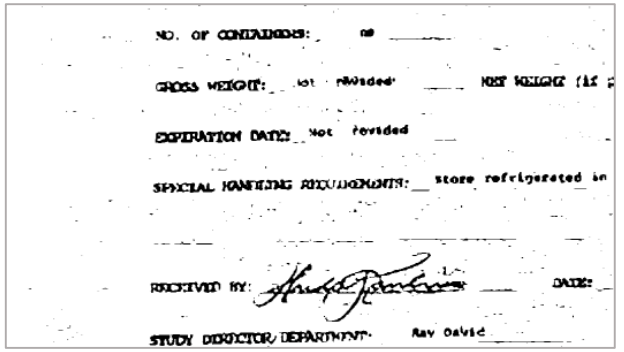
Method	Precision	Recall	F-score
Tesseract[21]	45.4	68.0	54.4
Google Vision (API)[22]	79.8	62.0	69.8
Faster R-CNN[23]	70.4	84.8	76.9
EAST[4]	51.6	84.0	63.9
CRAFT[5]	91.2	84.2	87.6
CharNet[6]	**95.1**	57.4	71.6
DUET *(proposed)*	93.1	**92.2**	**92.6**

- *IoU threshold @ 0.5*
- Results from [21-23] and [4] are provided by [12]
- For [5] and [6], the trained models and test codes from the original studies were used

- Output examples



Detection



Enhancement

Conclusion

- Text detector for document images
- Enhance robustness for noisy documents
  - Auxiliary task: text enhancement
- Overcome data insufficiency
  - Synthesized training data
  - Weak-supervision to train enhancement of the real training data

References

1. S. Long et al., “TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes,” in ECCV, 2018, pp. 20–36.
2. J. Ma et al., “Arbitrary-oriented Scene Text Detection via Rotation Proposals,” IEEE Trans. Multimed., vol. 20, no. 11, pp. 3111–3122, 2018.
3. D. He et al., “Multi-scale FCN with Cascaded Instance Aware Segmentation for Arbitrary Oriented Word Spotting in the Wild,” in CVPR, 2017, pp. 3519–3528.
4. X. Zhou et al., “EAST: An Efficient and Accurate Scene Text Detector,” in CVPR, 2017, pp. 5551–5560.
5. Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, “Character Region Awareness for Text Detection,” in CVPR, 2019, pp. 5551–5560.
6. L. Xing, Z. Tian, W. Huang, and M. R. Scott, “Convolutional Character Networks.” in ICCV, 2019, pp. 9126-9136.
7. D. Karatzas et al., “ICDAR 2013 Robust Reading Competition,” in ICDAR, 2013, pp. 1484–1493.
8. D. Karatzas et al., “ICDAR 2015 Rompetition on Robust Reading,” in ICDAR, 2015, pp. 1156–1160.
9. C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, “Detecting Texts of Arbitrary Orientations in Natural Images,” in CVPR, 2012, pp. 1083–1090.
10. A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, “Coco-text: Dataset and Benchmark for Text Detection and Recognition in Natural Images,” arXiv preprint arXiv:1601.07140 , 2016.
11. C. K. Ch'ng, and C. S. Chan, “Total-text: A Comprehensive Dataset for Scene Ttext Detection and Recognition.” in ICDAR, vol. 1, pp. 935-942, 2017.
12. G. Jaume, H. Kemal Ekenel, and J.-P. Thiran, “FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents,” ICDARW, vol. 2, pp. 1–6, 2019.
13. B. Bakker, and T. Heskes, “Task Clustering and Gating for Bayesian Multitask Learning.” J. Mach. Learn. Res., vol. 4, pp. 83-99, 2003.
14. B. Romera-Paredes, A. Argyriou, N. Berthouze, and M. Pontil, “Exploiting unrelated tasks in multi-task learning.” in AISTATS, 2012, pp. 951-959.
15. L. Deng, G. E. Hinton, and B. Kingsbury, “New types of deep neural network learning for speech recognition and related applications: An overview.” in ICASSP, 2013, pp. 8599–8603.
16. K. He, G. Gkioxari, P. Dollár, and R. Girshick. "Mask R-CNN." in CVPR, 2017, pp. 2961-2969.
17. J. Dai, K. He, and J. Sun, “Boxsup: Exploiting Bounding Boxes to Supervise Convolutional Networks for Semantic Segmentation.” in CVPR, 2015, pp. 1635-1643.
18. J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, “Ficklenet: Weakly and Semi-supervised Semantic Image Segmentation using Stochastic Inference.” in CVPR, 2019, pp. 5267-5276.
19. T. Y. Lin et al., “Feature Pyramid Networks for Object Detection,” in CVPR, 2017, pp. 2117–2125.
20. S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated Residual Transformations for Deep Neural Networks,” in CVPR, 2017, pp. 1492–1500.
21. R. Smith, “An Overview of The Tesseract OCR Engine,” ICDAR 2007 0
22. https://cloud.google.com/vision/docs/pdf
23. S. Ren, et al., “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” NIPS 2015.