

DOCUMENT LAYOUT ANALYSIS VIA DYNAMIC RESIDUAL FEATURE FUSION

Xingjiao Wu^{1,2}, Ziling Hu², Xiangcheng Du^{2,3}, Jing Yang^{2*}, Liang He^{1,2}

¹ Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai, China

² School of Computer Science and Technology, East China Normal University, Shanghai, China

³ Videt Tech Ltd., Shanghai, China

{52184506007, 51184506059, 51194506062}@stu.ecnu.edu.cn, {jyang, lhe}@cs.ecnu.edu.cn

ABSTRACT

The document layout analysis (DLA) aims to split the document image into different interest regions and understand the role of each region, which has wide application such as optical character recognition (OCR) systems and document retrieval. However, it is a challenge to build a DLA system because the training data is very limited and lacks an efficient model. In this paper, we propose an end-to-end united network named Dynamic Residual Fusion Network (DRFN) for the DLA task. Specifically, we design a dynamic residual feature fusion module which can fully utilize low-dimensional information and maintain high-dimensional category information. Besides, to deal with the model overfitting problem that is caused by lacking enough data, we propose the dynamic select mechanism for efficient fine-tuning in limited train data. We experiment with two challenging datasets and demonstrate the effectiveness of the proposed module.

Index Terms— Semantic segmentation, Document Layout Analysis, Deep Learning

1. INTRODUCTION

Document layout analysis (DLA) is an important research area dedicated to extracting semantic information from the document image. As a critical preprocessing step of document understanding systems, DLA can provide information for several applications such as document retrieval, content categorization, and text recognition.

Traditional DLA methods are highly dependent on manually designed features and can not deal with complex layouts. Recently, with the rapid development of deep learning, significant progress has been made in DLA research. Inspired by Fully Convolutional Network (FCN) [1], many deep learning-based methods are proposed. These methods consider DLA as a special segmentation task [2]. However, it is still a challenge to design an efficient DLA system. Accurately predicting the category of regions in a document is limited by the

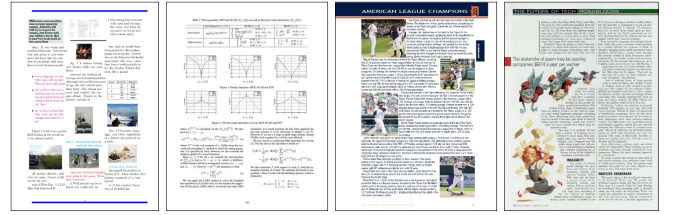


Fig. 1. Images from different sources. The image in column 1 is synthetic data and others are from the real dataset.

bottlenecks that lack enough training data and a suitable end-to-end model. This work focuses on implementing a universal and effective DLA framework in which an effective network and an appropriate update learning method are designed.

Because the current dataset size can not satisfy the training of the generalized model, many researchers dedicated to using synthetic data to help the training model [3]. However, it is still a difficult problem that there is a distribution gap between synthetic data and real data. Inspired by the work of [4, 5], we propose a dynamic select mechanism, an approach for efficient input-related fine-tuning. The core of the approach is that we split a layer into a pre-train path and a fine-tune path and aggregate information from both paths to generate two channel-wise select weights for aggregating the feature maps. With a slight increase in trainable parameter, the dynamic select mechanism can prevent the model from overfitting.

The dynamic selection mechanism provides a new model encoding method, but the final model also needs to be decoded. How to transfer the previous information most effectively is also an important issue. Many works have discussed high-dimensional semantic information and low-dimensional semantic information. High-dimensional features are abundant in category semantic information and low-dimensional features skilled in reconstructing original resolution detail. Both features are important to accurate pixel classification. To utilize all dimensional features, the FCN model directly superimposes features extracted from the encoder layers on the decoder layers. However, this mechanism is harmful to

*Corresponding author.

Xingjiao Wu and Ziling Hu contributed equally to this work.

the category of semantic information in high-dimensional features. In this work, we designed a dynamic residual feature fusion module that effectively fuses the different dimension features and generates robust fusional features. The core idea of the proposed dynamic residual feature fusion module is to extra global semantic information from high-dimensional features and weight fusional features as guidance. Then high-dimensional features are added to fusional features to maintain high-dimensional category semantic information flow.

Our contributions can be summarized as follows:

- We propose a dynamic residual feature fusion module which can effectively fuse the high-dimensional category semantic and low-dimensional detail information for document layout analysis.
- We propose a dynamic select mechanism that can prevent the model from overfitting when fine-tuning in limited data.
- The DRFN achieves comparable results on two benchmark datasets.

2. RELATED WORK

In the early stage, the classical document layout analysis (DLA) strategies can be roughly divided into two categories: bottom-up and top-down. Usually, the bottom-up method obtains document analysis result dynamically from a smaller granularity data level. In general, bottom-up analysis strategies use local features from documents such as pixel distributions and properties of connected components to detect individual pixels or words. Then, detected elements are grown up to form larger document regions and stop until they reach the presenting analysis level. The bottom-up strategies have attracted researchers because they have high accuracy and can deal with complex layouts. However, most methods are high time complexity and require large space. The classical bottom-up strategies can be subdivided into five core categories: connected component analysis [6], texture analysis [7], learning-based analysis [8], Voronoi diagrams [9], and Delaunay triangulation [10].

The top-down strategies iteratively divide whole pages into small document elements such as columns, blocks, textiles, and words. Similarly, the top-down strategies can be subdivided into four categories: Texture-based Analysis [11], Run Length Smearing Algorithm (RLSA) [12], DLA projection-profile [13], and With space analysis [14]. Usually, the top-down strategies work greatly on regular document layouts. However, it is limited by the requirement for clean and skews corrected document images. Though it can deal with regular common rectangular document layouts, the bottom-up strategies and the top-down strategies are not effective enough to handle complex layouts.

With the rapid development of deep learning, serious methods are proposed and achieve impressive performance in this field. The DLA task is considered as an especially semantic segmentation task [15], whose goal is performing a pixel-level understanding of the segmentation objects. Deep learning-based DLA methods are mainly inspired by Fully Convolutional Network (FCN) [1]. He *et al.* [16] uses an FCN with the multi-scale feature for semantic page segmentation with an auxiliary element contour detection task. Wick *et al.* [17] adopt a modified Fcn to deal with historical document segmentation. Li *et al.* [18] propose a novel end-to-end deep neural network named DeepLayout for page layout analysis.

The low-dimensional features contain more detailed information and the high-dimensional features contain more semantic information [19]. For generating more robust features, FCN fuse the low-dimensional features and high-dimensional features using a skip connection structure. However, the skip connection structure can not take advantage of the high-dimensional features, because the contained high-dimensional category semantic information is confused by low-dimensional features. Therefore, we propose a dynamic residual feature fusion module. This module achieves recovering images detail and meanwhile maintaining the category semantic information.

3. FRAMEWORK

In this section, we will introduce the details of our methods. Firstly, we elaborate on the overall architecture of the proposed network. Then, the dynamic residual fusion feature module is expounded. Next, we present the details of the dynamic select mechanism. Finally, we will introduce synthetic document data.

3.1. Network Architecture

The overview of the proposed network is shown in Fig 2. We adopt the idea that trains an FCN network extracting feature for pixel-wise segmentation with dynamic residual feature fusion module and dynamic select mechanism. For extracting features effectively, we adopt an encoder-decoder structure, which is widely used in general image semantic segmentation task. To build a powerful encoder, we apply ResNet-18 as the backbone and adopt dynamic select mechanism on each restock. Moreover, the dilated convolution with the rate of 2 is applied to DSBlock-4, so the output size of feature maps from the encoder is 1/16 of the input image. Then, we use spatial pyramid pooling proposed in Deeplabv3+ [20] to enhance the feature F output from the encoder. After that, we upsample F and feed it in the decoder part. We perform the dynamic residual feature fusion module as a fast and effective decoder structure, which can maintain high-dimensional category information while making full use of low-dimensional

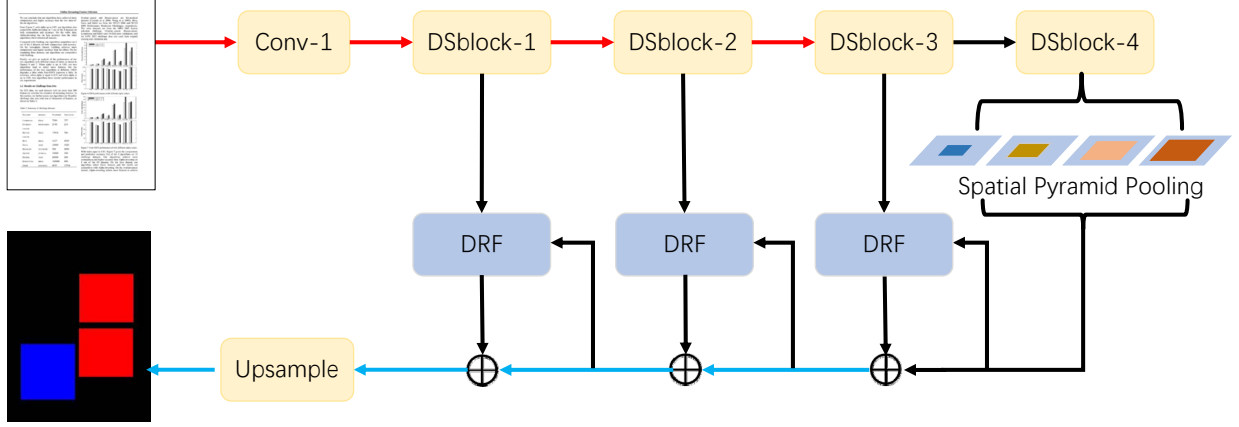


Fig. 2. Illustration of DRFN. We adopt the encoder-decoder architecture and use Resnet-18 as the backbone. The top half is the encoder part. The DSblock is a resblock with a dynamic select mechanism. We use a mechanism on each resblock. The lower half is the decoder part. We adopt the dynamic residual feature fusion module as the decoder block. The red and blue lines represent the downsample and upsample operators respectively.

detail information.

3.2. Dynamic Residual Feature Fusion Module

The encoder-decoder networks mainly consider using the encoder to extract category information and recover spatial information by the decoder. PSPNet [21] and Deeplab [20] directly use bilinearly upsample, which are disadvantageous to restoring the spatial position to the origin resolution because of lacking low-dimensional detail information. FCN [1] and PANet [22] gradually recover clear object boundary details by using different scale low-dimensional features. However, these methods directly combine different scale features. It may be harmful for category semantic information in high-dimensional features and cause inaccurate pixel classification. Therefore, we propose a dynamic residual feature fusion module which maintains the category semantic information by a residual struct. Besides, global semantics in high-dimensional can be a guidance weight to select precise detailed information.

Our dynamic residual feature fusion module is shown in Fig 3. In detail, we first concatenate the low-dimensional features and high-dimensional features. Then we deploy a 3×3 convolution layer to reduce the channel size and generate the fusion features. For reducing calculation overhead, we adopt the depthwise separable convolution [23]. Then the guidance weight is extracted from high-dimensional features through a global average pooling and two 1×1 convolutional layers with batch normalization and sigmoid activation. The first convolutional layer compact the size of guidance weight with a reduction ratio r (we set as 16 in our experiments), and the second one restore it to its original size. Then the guidance weight is multiplied by the fusion features. Finally, the high-dimensional features are added with the fusion features. This

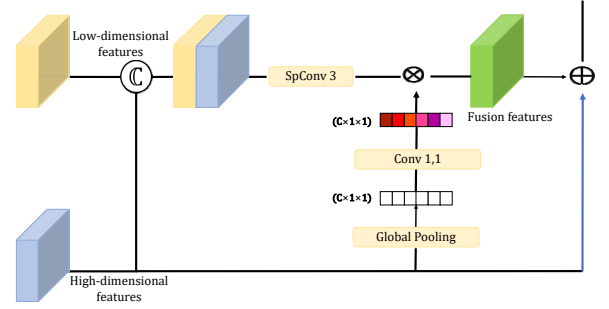


Fig. 3. Dynamic Residual Feature Fusion Module Structure. Where 3 represent 3×3 convolutional layer and 1 represents 1×1 Convolutional layer. SpConv represents depthwise separable convolution. The blue line represents upsample operators.

module maintains the category information flow in the decoder layer and deploys different dimension features.

3.3. Dynamic Select Mechanism

Deep learning-based vision model has shown remarkable success in many computer vision fields but often rely on large amounts of the labelled training examples. It is a challenge that trains a universal document layout analysis model because of lacking enough data. To deal with this problem, much research adopts a pre-trained model trained in the large synthetic dataset and fine-tune the model in the target dataset. However, fine-tuning in a small dataset may result in overfitting. Therefore, we proposed the dynamic select mechanism.

Given a pre-trained network model, our dynamic select mechanism split the encoder block as two parallel trainable fine-tune paths P_t and frozen pre-trained path P_f . Then our

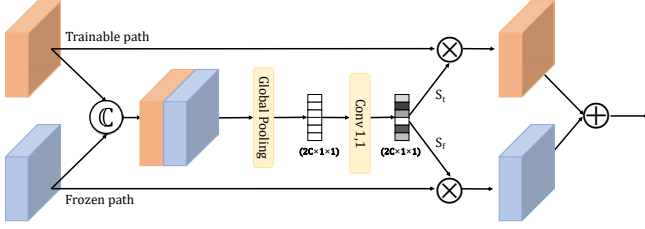


Fig. 4. Trainable path and frozen path in dynamic select mechanism structure.

mechanism dynamically generates two channel-wise select weights S_t and S_f . We implement our method on Resnet [24], the output of the l -th encoder block is computed as:

$$x_l = S_t(x_{l-1})P_t(x_{l-1}) + S_f(x_{l-1})P_f(x_{l-1}) + x_{l-1} \quad (1)$$

Where $S_t(x_{l-1}) \in \mathbb{R}^{C \times 1 \times 1}$ and $S_f(x_{l-1}) \in \mathbb{R}^{C \times 1 \times 1}$.

For generating the two select weights, we need to gather information from both parallel paths. We first fuse the results from two paths through a concatenate operation. Then global statistics are extracted via a global average pooling. Next, similar to the dynamic residual feature fusion module, two 1×1 convolutional layers are adapted to generate the global weight S . Then we reshape S to the size of $\mathbb{R}^{C \times 2 \times 1}$ and apply a softmax function. Finally, S_t and S_f are obtained by splitting S . On each channel, the sum of S_t and S_f is always equal to 1.

Our model dynamically decides a channel-wise feature select strategy for each image to fully utilize both towpaths during the whole training process.

4. EXPERIMENTS

To evaluate the effectiveness of the DRFN, we conducted several sets of experiments. We use DSSE-200 and CS-150 as benchmark datasets. We train DRFN on 4000 synthetic data for pre-train and fine-tune in two real benchmark datasets. Besides, we use DSSE-200 to verify the effect of the dynamic residual feature fusion module and dynamic select mechanism. To measure the performance on DLA task, we divide the classification criteria into four classification criteria: background, figure, text, table. We use accuracy, precision, recall and F1 metrics to evaluate the DRFN.

4.1. Datasets

DSSE-200. The DSSE-200 [3] is a complex document layout dataset including various dataset styles. The dataset contains 200 images from pictures, PPT, brochure documents, old newspapers and scanned documents.

CS-150. The CS-150 [25] was proposed by Clark. The dataset consists of 150 papers and includes 1175 images. The

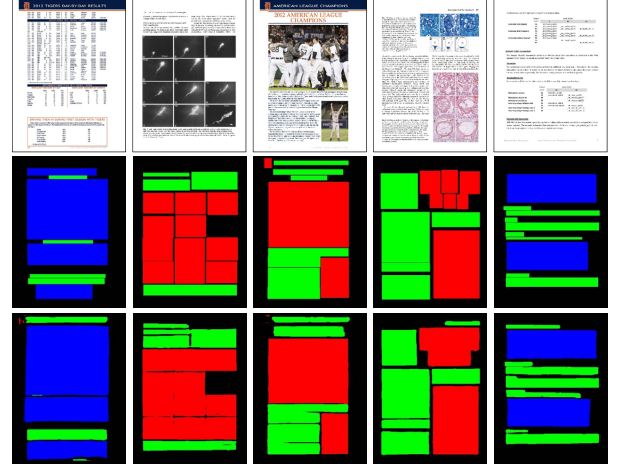


Fig. 5. The DSSE-200 real documents and their corresponding segmentation. Top: original. Middle: ground-truth. Bottom: predictions. Segmentation label colors are: figure (red), table (blue) and background (black).

classification criteria in this dataset are divided into three types: image, table, and others.

4.2. Implementation Details

The baseline is the fully convolutional neural network (FCN). We use Resnet-18 pre-trained on Imagenet [26] as the backbone of the DRFN because it has excellent performance and lightweight parameters. We use Adma as the optimizer, and the batch size is set to 8. For the data argument, we adopt several strategies. Firstly, the images are horizontally flipped with the probability of 0.5. Then we flip up the images with the probability of 0.5. Next, we randomly crop samples with a size ratio of 0.7 from the transformed images.

For training our model, we generate 4000 synthetic images as the per-train dataset. When fine-tuning, we adopt the dynamic select mechanism. We first pre-train our model on the per-train dataset. Then we use the pre-trained parameter to initialize the model. During the fine-tuning process, we use real data from each dataset. Because the DSSE-200 and CS-150 do not provide standard training datasets, we randomly select some images as the train data and others as test data. We selected 40 images on DSSE-200 and 940 images on CS-150.

4.3. Ablation Study

To verify the effectiveness of our approach, we do a series of comparative experiments on the DSSE-200 dataset. These experiments mainly focus on evaluating dynamic residual feature fusion module and dynamic select mechanism.

Dynamic Residual Feature Fusion For evaluating the effectiveness of dynamic residual feature fusion module, we com-

Table 1. The results on the DSSE-200 dataset. The baseline is FCN/Resnet-18.

Model	Method	A	P	R	F1
1	FCN/VGG16 [1]	80.6	71.0	77.6	74.1
2	FCN/Resnet-50	81.2	75.7	75.4	75.5
3	baseline	80.7	72.2	81.6	76.6
4	DRFN	82.8	77.1	83.7	80.3

Table 2. The result on dataset. The DV3+ means the DeeplabV3+ model with Xception backbone. The DV3+R18 means DV3+/Resnet-18. The Para represents the trainable parameter amount of the model.

Method	DSSE-200				CS-150				Para
	A	P	R	F1	A	P	R	F1	
Segnet [27]	87.0	82.5	86.0	84.2	99.4	93.8	94.1	94.0	29M
PSPnet [21]	88.4	87.7	83.5	85.5	99.3	94.0	93.5	93.8	46M
PANet [22]	87.9	86.0	82.4	84.2	99.3	94.2	94.0	94.1	168M
DV3+ [28]	87.0	83.3	81.7	82.4	-	-	-	-	53M
DV3+R18	90.2	87.6	87.1	87.3	99.4	94.6	94.5	94.6	15M
baseline	87.1	83.3	83.8	83.5	99.4	94.0	94.1	94.1	12M
DRFN	90.5	88.8	90.1	89.5	99.5	95.3	94.8	95.1	17M

pared our model with a baseline on DSSE-200 dataset. Here we directly deploy Resnet-18 on as the encoder, without dynamic select mechanism. And we only use synthetic data to train all models. We also evaluate the performance with a different backbone. The results are shown in Table1. Compared with the baseline, FCN [1] with Resnet-18, our model achieved a better performance in all metrics. Besides, an interesting first phenomenon can be seen by comparing model 2 and model 3. Though with a deeper and more parameters backbone, the model does not achieve better performance and even make a degradation. This is the reason why we chose Resnet-18 as the backbone.

Dynamic Select Mechanism To evaluate the effectiveness of the dynamic select mechanism, we fine-tune the model on DSSE-200 dataset and evaluate the performance. The results are shown in Table3. In our implementation, we initialize the model with the parameter pre-trained on synthetic data. Compared with the model directly fine-tune with Resnet-18, dynamic select mechanism improved model performance on all metrics.

4.4. Comparisons with Prior Arts

Comparisons on DSSE-200 dataset (Table2). Following the setting mentioned above, we experiment on the DSSE-200. Compared with other classical semantic segmentation methods, the experiment shows that our models are more powerful with much less parameters. Moreover, as shown in

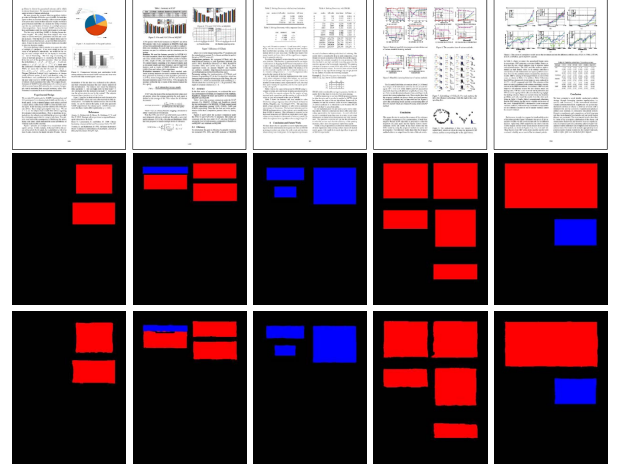


Fig. 6. The CS-150 real documents and their corresponding segmentation. Top: original. Middle: ground-truth. Bottom: predictions. Segmentation label colors are: figure (red), table (blue) and background (black).

Table 3. The fine-tune results on the DSSE-200 dataset. The DSM represents the dynamic select mechanism.

Method	A	P	R	F1
fine-tune with Resnet-18	90.3	88.3	88.5	88.4
fine-tune with DSM	90.7	88.8	90.1	89.5

Fig 5, our model can correctly identify text pixels which are unlabeled in the ground-truth. This result proves our model adapts the target data while retaining the knowledge learned in the source synthetic data.

Comparisons on CS-150 dataset (Table2). We also evaluate our model on the CS-150 dataset. Because only the image area and the table area are marked in CS-150 dataset, we process the label to classify the image as background, figure, and table. The results are shown in Fig 6. Observing the experimental results, our results reach the state-of-the-art.

5. CONCLUSIONS

In this paper, we propose a novel solution for constructing a model of universal document layout analysis. To make full use of low-dimensional information and maintain high-dimensional semantic information, we propose the dynamic residual feature fusion module. Furthermore, we propose the dynamic select mechanism for efficient input-related fine-tuning. Extensive experiments on document layout analysis benchmarks show the superior performance of the proposed method.

6. REFERENCES

- [1] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *CVPR*, 2015, pp. 3431–3440.
- [2] K. Chen, M. Seuret, J. Hennebert, and R. Ingold, “Convolutional neural networks for page segmentation of historical document images,” in *ICDAR*, vol. 1, 2017, pp. 965–970.
- [3] X. Yang, E. Yumer, P. Asente, M. Kralej, D. Kifer, and C. Lee Giles, “Learning to extract semantic structure from documents using multimodal fully convolutional neural networks,” in *CVPR*, 2017, pp. 5315–5324.
- [4] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, “Spottune: transfer learning through adaptive fine-tuning,” in *CVPR*, 2019, pp. 4805–4814.
- [5] X. Li, W. Wang, X. Hu, and J. Yang, “Selective kernel networks,” in *CVPR*, 2019, pp. 510–519.
- [6] L. O’Gorman, “The document spectrum for page layout analysis,” *IEEE Trans. PAMI*, vol. 15, no. 11, pp. 1162–1173, 1993.
- [7] N. Journet, V. Eglin, J.-Y. Ramel, and R. Mullot, “Text/graphic labelling of ancient printed documents,” in *ICDAR*, 2005, pp. 1010–1014.
- [8] S. Marinai, M. Gori, and G. Soda, “Artificial neural networks for document analysis and recognition,” *IEEE Trans. PAMI*, vol. 27, no. 1, pp. 23–35, 2005.
- [9] K. Kise, A. Sato, and M. Iwata, “Segmentation of page images using the area voronoi diagram,” *Computer Vision and Image Understanding*, vol. 70, no. 3, pp. 370–382, 1998.
- [10] S. Eskenazi, P. Gomez-Krämer, and J.-M. Ogier, “The delaunay document layout descriptor,” in *Proceedings of the 2015 ACM Symposium on Document Engineering*, 2015, pp. 167–175.
- [11] R. Saabni, A. Asi, and J. El-Sana, “Text line extraction for historical document images,” *Pattern Recognition Letters*, vol. 35, pp. 23–33, 2014.
- [12] A. Alaei, U. Pal, and P. Nagabhushan, “A new scheme for unconstrained handwritten text-line segmentation,” *Pattern Recognition*, vol. 44, no. 4, pp. 917–928, 2011.
- [13] F. Shafait and T. M. Breuel, “The effect of border noise on the performance of projection-based page segmentation methods,” *IEEE Trans. PAMI*, vol. 33, no. 4, pp. 846–851, 2010.
- [14] F. Shafait, J. Van Beusekom, D. Keysers, and T. M. Breuel, “Background variability modeling for statistical layout analysis,” in *ICPR*, 2008, pp. 1–4.
- [15] D. Ravì, M. Bober, G. M. Farinella, M. Guarnera, and S. Battiato, “Semantic segmentation of images exploiting dct based features and random forest,” *Pattern Recognition*, vol. 52, pp. 260–273, 2016.
- [16] D. He, S. Cohen, B. Price, D. Kifer, and C. L. Giles, “Multi-scale multi-task fcn for semantic page segmentation and table detection,” in *ICDAR*, vol. 1, 2017, pp. 254–261.
- [17] C. Wick and F. Puppe, “Fully convolutional neural networks for page segmentation of historical document images,” in *DAS*, 2018, pp. 287–292.
- [18] Y. Li, Y. Zou, and J. Ma, “Deeplayout: A semantic segmentation approach to page layout analysis,” in *International Conference on Intelligent Computing*, 2018, pp. 266–277.
- [19] D. Chen, X. Cao, F. Wen, and J. Sun, “Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification,” in *CVPR*, 2013, pp. 3025–3032.
- [20] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE Trans. PAMI*, vol. 40, no. 4, pp. 834–848, 2017.
- [21] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *CVPR*, 2017, pp. 2881–2890.
- [22] H. Li, P. Xiong, J. An, and L. Wang, “Pyramid attention network for semantic segmentation,” *arXiv preprint arXiv:1805.10180*, 2018.
- [23] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *CVPR*, 2017, pp. 1251–1258.
- [24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.
- [25] C. A. Clark and S. Divvala, “Looking beyond text: Extracting figures, tables and captions from computer science papers,” in *Workshops at the AAAI*, 2015.
- [26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR. Ieee*, 2009, pp. 248–255.

- [27] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Trans. PAMI*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *ECCV*, 2018, pp. 801–818.